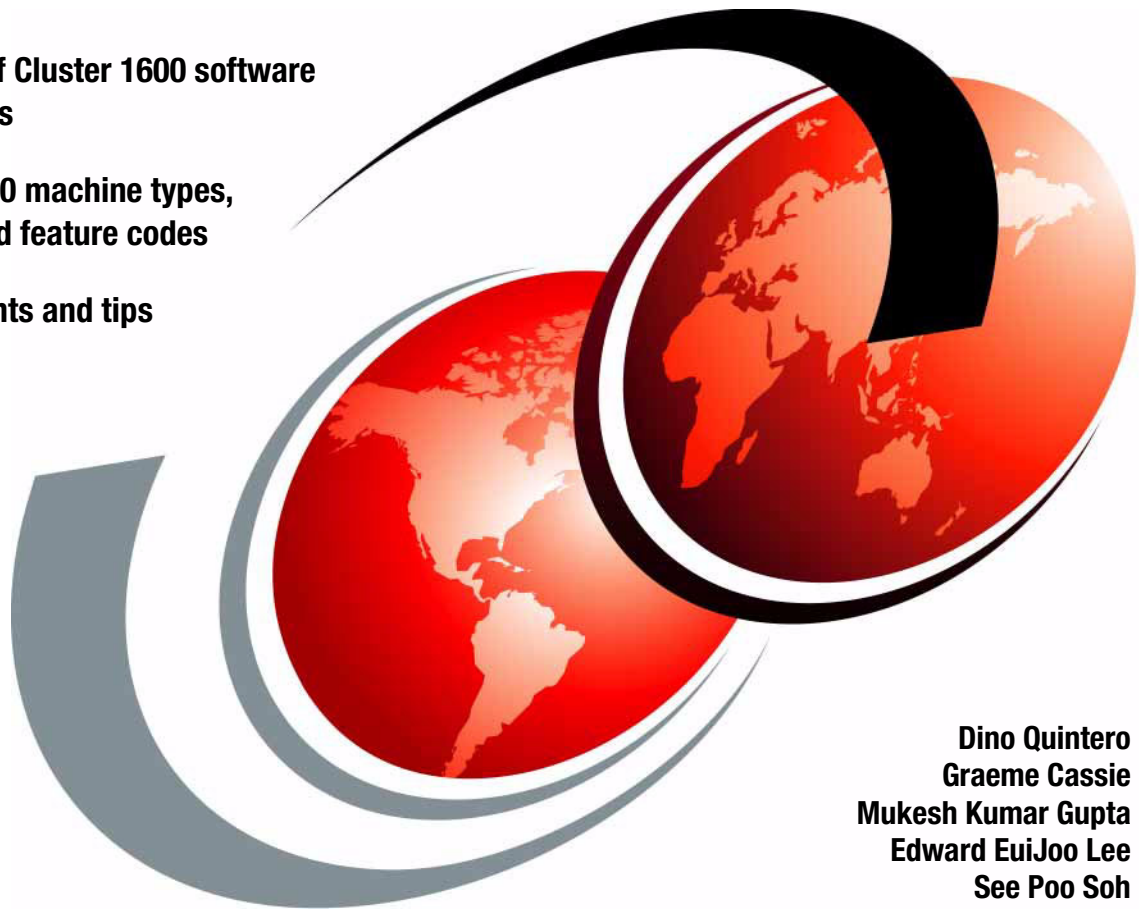


IBM **@**server pSeries Cluster Systems Handbook

Overview of Cluster 1600 software
components

Cluster 1600 machine types,
models, and feature codes

Solution hints and tips



Dino Quintero
Graeme Cassie
Mukesh Kumar Gupta
Edward EuiJoo Lee
See Poo Soh



International Technical Support Organization

IBM @server pSeries Cluster Systems Handbook

October 2003

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

First Edition (October 2003)

This edition applies to Version 3, Release 5, of Parallel System Support Programs, and Version 1, Release 3, Modification 2, of Cluster Systems Management for use with the AIX Operating System Version 5, Release 2, Modification 1.

© Copyright International Business Machines Corporation 2003. All rights reserved.

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

Contents

Notices	ix
Trademarks	x
 Preface	xi
Become a published author	xiii
Comments welcome	xiv
 Chapter 1. Introduction	1
1.1 Overview of Cluster 1600	2
1.2 Choosing PSSP or CSM	3
1.2.1 Cluster management with PSSP	3
1.2.2 Cluster management with CSM	5
 Chapter 2. Cluster 1600 hardware	11
2.1 Overview	12
2.2 Cluster 1600 hardware components	13
2.2.1 Nodes	15
2.2.2 Frames	15
2.2.3 Switches	16
2.2.4 PSSP control workstations	16
2.2.5 CSM for AIX management server	16
2.2.6 Hardware Management Console (HMC)	17
2.3 CSM and PSSP hardware support	17
2.3.1 CSM-managed node requirements	18
2.4 PSSP control workstation	20
2.4.1 Control workstation requirements	20
2.4.2 Supported control workstations	21
2.4.3 High Availability Control Workstation	22
2.4.4 HACWS limitations	25
2.5 CSM management server	26
2.5.1 Memory and disk space	26
2.5.2 Network requirements	26
2.5.3 Asynchronous card requirements	27
2.5.4 Using a Logical Partition (LPAR) as a CSM management server ..	27
2.6 Cluster 1600 server concepts	28
2.6.1 pSeries architecture	32
2.6.2 Cluster 1600 and the HMC	38
2.6.3 Firmware	44
2.6.4 Electronic Service Agent	48

2.6.5 Planning for Cluster 1600 servers	50
2.7 Hardware supported and currently marketed	52
2.8 pSeries servers	52
2.8.1 pSeries 615 server (7029-6C3 and 6E3 deskside)	52
2.8.2 pSeries 630 server (7028-6C4 and 6E4 deskside)	57
2.8.3 pSeries 650 server (7038-6M2)	63
2.8.4 pSeries 655 server (7039-651)	73
2.8.5 670 Server (7040-671)	82
2.8.6 690 server (7040-681)	91
2.8.7 xSeries servers	100
2.9 Switches	100
2.9.1 9076 model 555	101
2.9.2 9076 model 556	101
2.9.3 9076 model 557	102
2.9.4 9076 model 558	103
2.9.5 7045-SW4 pSeries HPS (High Performance Switch)	103
2.10 Switch adapters	108
2.10.1 Switch adapter placement restrictions	109
2.10.2 pSeries HPS switch network interface cards (SNI)	110
2.11 Legacy hardware supported but no longer marketed	112
2.11.1 pSeries 660 Model 6M1 (7026-6M1)	112
Chapter 3. Network configuration	121
3.1 SP LAN Ethernet	122
3.1.1 Supported Ethernet adapters and their placement	122
3.1.2 Ethernet network topology	124
3.1.3 IP label convention	124
3.2 Switch network	125
3.2.1 Benefits of a Switch network	125
3.2.2 SP Switch2	126
3.2.3 Switch IP network and addressing	128
3.3 Other networks	128
3.4 Network considerations	129
3.4.1 The RS-232 connection	129
3.4.2 System topology considerations	130
3.4.3 Boot/install server requirements	130
3.4.4 The SP Ethernet administrative LAN	134
3.4.5 Additional LANs - considerations	137
3.4.6 IP over the switch - considerations	137
3.4.7 Subnetting - considerations	138
3.4.8 HMC trusted network - considerations	139
3.4.9 Network router node considerations	144
3.4.10 Clustered server configuration considerations	145

3.4.11 SP-attached server considerations	146
3.5 Sample scenarios of Cluster 1600 managed by PSSP	147
3.5.1 CWS with two HMCs and four pSeries	147
3.5.2 One CWS, one HMC and one pSeries	148
3.5.3 CWS, two HMCs, two 9076s, with one pSeries and SP Switch2	149
3.5.4 CWS, HMC, 9076 frame and pSeries with SP Switch	150
3.6 Networking for Cluster Systems Management (CSM).	151
3.6.1 CSM hardware control	151
3.6.2 Hardware and network requirements	151
3.6.3 Virtual LANs (VLANs)	153
3.6.4 Conceptual diagram for pSeries cluster	154
3.6.5 pSeries HPS switch network overview	156
3.6.6 Switch Network Manager (SNM).	158
3.6.7 Considerations for Cluster 1600 managed by CSM network	160
3.6.8 Examples	161
3.6.9 Redundant HMC Layout for pseries HPS in Cluster 1600	163
3.6.10 Redundant layout for pSeries in the Cluster 1600	164
3.6.11 Conceptual Cluster 1600 without a pSeries HPS	165
3.6.12 Management Server with two HMCs and four pSeries	166
Chapter 4. Software support	169
4.1 Software components of the Cluster 1600	170
4.2 Parallel System Support Programs (PSSP)	172
4.2.1 Administration and operation	173
4.2.2 Reliable Scalable Cluster Technology (RSCT)	174
4.2.3 IBM Virtual Shared Disk (VSD)	175
4.2.4 Security	175
4.2.5 Communication subsystem	176
4.2.6 Network Time Protocol (NTP)	176
4.2.7 System availability	176
4.2.8 Other PSSP services	177
4.2.9 New in PSSP 3.5.	177
4.2.10 Software requirements	179
4.2.11 Software compatibility matrix	179
4.2.12 Documentation references - PSSP	179
4.3 Cluster Systems Management (CSM).	180
4.3.1 Administration and operation	181
4.3.2 Reliable Scalable Cluster Technology (RSCT)	188
4.3.3 New in CSM 1.3.2 for AIX	189
4.3.4 Supported platform	190
4.3.5 PSSP-to-CSM transition	191
4.3.6 Documentation references - CSM.	191
4.4 General Parallel File System (GPFS)	192

4.4.1	Architecture	192
4.4.2	Administration and operation	196
4.4.3	Higher performance/scalability	196
4.4.4	Recoverability	197
4.4.5	Migration	198
4.4.6	New in GPFS 2.1 for AIX 5L	198
4.4.7	Software requirements	198
4.4.8	Documentation references	199
4.5	LoadLeveler	200
4.5.1	Administration and operations	201
4.5.2	Capabilities	203
4.5.3	New in LoadLeveler 3.2	205
4.5.4	New in LoadLeveler 3.1	206
4.5.5	Software requirement	207
4.5.6	LoadLeveler configuration suggestions	208
4.5.7	Documentation references - LoadLeveler	208
4.6	Scientific subroutine libraries	208
4.6.1	Engineering and Scientific Subroutines Library (ESSL) family of products	208
4.6.2	Operations	209
4.6.3	New in ESSL 4.1	210
4.6.4	New in Parallel ESSL 3.1	210
4.6.5	Software requirements	211
4.6.6	Documentation references - ESSL and PESSL	213
4.6.7	Mathematical Acceleration Subsystem (MASS)	214
4.7	Parallel Environment (PE)	216
4.7.1	Parallel Programming support	216
4.7.2	Operation	217
4.7.3	New in PE 4.1	218
4.7.4	Software requirements	219
4.7.5	Documentation references - Parallel Environment (PE)	220
4.8	IBM High Availability Cluster Multi-Processing for AIX (HACMP)	220
4.8.1	HACMP operations	221
4.8.2	Administration and operation	223
4.8.3	New in HACMP 5.1	223
4.8.4	Software requirements	224
4.8.5	Documentation references - HACMP	224
4.9	Performance Toolbox (PTX) and Performance AIDE (PAIDE)	225
4.9.1	Administration and operation	225
4.9.2	Platform requirements	226
4.10	Software ordering and configuration	226
Chapter 5. Solutions and offerings “best practices”		229

5.1 High Performance Computing (HPC) environment	230
5.1.1 A hypothetical solution	230
5.1.2 Solution architecture	230
5.1.3 Solution discussion	232
5.1.4 Cluster management considerations.	236
5.2 Transition from SP nodes to LPARs	237
5.2.1 System migration by utilizing alternate disk migration	238
5.3 Virtual Serial port implications with LPARS.	240
5.3.1 Console device	240
5.3.2 Serial port implication in an LPAR/SP environment	242
5.3.3 Virtual terminal window	242
5.4 HMC considerations	244
5.4.1 Redundant HMC	244
5.5 Web-based System Manager client solutions	245
5.5.1 Web-based System Manager functionality through the firewall	246
Appendix A. Performance	251
A.1 Switch performance	252
A.2 Adapter performance	253
A.3 Node I/O slot performance	254
A.4 Application performance	255
A.5 MPI/user space	257
A.6 TCP/IP	260
Related publications	265
IBM Redbooks	265
Other publications	265
Online resources	266
How to get IBM Redbooks	267
Help from IBM	267
Abbreviations and acronyms	269
Index	273

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:

@server™

@server™

AIX®

Enterprise Storage Server®

IBM®

IntelliStation®

LoadLeveler®

Lotus®

Micro Channel®

NetView®

PowerPC®

POWER2™

POWER3™


POWER4™

POWER4+™

PTX®

pSeries®

Redbooks™

(logo) ™

RS/6000®

SAA®

SP1®

Tivoli®

VisualAge®

The following terms are trademarks of other companies:

Intel, Intel Inside (logos), MMX, and Pentium 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.

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

The IBM @server™ Cluster 1600 server, which was introduced to meet the rigorous demands of mission-critical enterprise applications, continues to offer outstanding performance, scalability, reliability, availability, serviceability, and management capabilities. In this IBM® Redbook, we highlight the benefits of using a Cluster 1600, and describe which hardware components can be managed by either Parallel System Support Programs (PSSP) Version 3, Release 5, or Cluster Systems Management (CSM) Version 1, Release 3, Modification 2.

This publication contains the following information on the Cluster 1600:

- ▶ Cluster 1600 hardware components
- ▶ Networking components and considerations
- ▶ Cluster 1600 software components
- ▶ Scalability of the Cluster 1600
- ▶ Solutions and offerings scenarios

The Cluster 1600 helps to reduce the complexities and costs of system management, thus lowering the total cost of ownership and allowing simplification of application service level management. It also provides the infrastructure that supports availability, data sharing, and response time. This redbook will be useful for IT professionals seeking to implement Cluster 1600 mission-critical solutions to address business intelligence applications, server consolidation, and collaborative computing. The team that wrote this redbook

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

Dino Quintero is a Project Leader and IBM Certified Senior IT Specialist at the International Technical Support Organization, Poughkeepsie Center. His responsibilities include assessing, designing, and implementing pSeries®/AIX technical solutions for various customer sets including those in clustered environments, and writing Redbooks™ and teaching ITSO workshops.

Graeme Cassie is a pre-sales Technical Consultant in Scotland, currently working for Morse Group Ltd., one of Europe's largest technology integrators. Previously, Graeme spent almost 15 years with IBM as a Customer Engineer specializing in mid-range systems. He has 20 years of experience in the IT field. He holds an HNC in Electronic Engineering, and holds IBM Certifications for RS/6000® SP, HACMP, p690, mid-range storage, Enterprise Disks, Enterprise Tape and e-business solutions design. He is also a Certified AIX Technical

Expert (CATE). His areas of expertise include solution design and implementation around IBM pSeries and IBM storage systems.

Mukesh Kumar Gupta is a Senior IT Specialist (Integrated Technology Services) with IBM, India. He has eight years of experience in the Support and Services field, and has worked at IBM for five years. His areas of expertise include AIX®, PSSP, RS/6000, RS/6000 SP, HACMP, Cluster 1600 and Storage (administration, support and implementation), and he works closely with pre-sales in solution design.

Edward EuiJoo Lee works as a Senior Technology Architect for General Electric (GE-ITS) in New York, USA, and has written many internal technical papers for GE. Edward has 12 years of experience in Open Systems technology and solution development. He participates in speaking engagements on technical subjects, and served as a speaker for IBM's SP World in 1999. He is an IBM Certified Advanced Technical Expert (CATE), and is also a Certified Professional in Information Technology Services Management (ITSM). He holds a B.S. degree in Information Technology Engineering from S.U.N.Y., and is currently working towards a Master's Degree in Management.

See Poo Soh is an Advisory Technical Specialist in Singapore with the IBM Systems Group, and has nine years of experience in the IT field. He holds a Bachelor of Engineering degree from the National University of Singapore, and a Graduate Diploma in Business Administration from the Singapore Institute of Management. His areas of expertise include High Performance Computing, where he spent 3 1/2 years working with HPC systems in a customer environment, and 3 1/2 years as a Technical Specialist in IBM, architecting HPC solutions and supporting HPC engagements. He is also skilled in SAP and Lotus® Domino sizing on the pSeries platform, and has worked on infrastructure solutions in the Data Center.



Team members (left to right): Graeme Cassie, Mukesh Kumar Gupta, Edward EuiJoo Lee, See Poh Soh, Dino Quintero (project leader)

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

Margarita Hunt, Keigo Matsubara
International Technical Support Organization, Austin Center

Ramesh Goel, Umang Mathur
IBM, India

Dave Delia, Duane Witherspoon, Paul Swiatocha Jr., Janet Ellsworth, John Simpson, Skip Russel, Robin Hanrahan, Paula Trimble, Gary Mincher, Brian Herr, Robert Curran, Joan McComb, Waiman Chan, Patrick Caffrey, Gordon Mcpheeters, Elaine Krakower, Laura Murphy
IBM Poughkeepsie

Bernard King-Smith
IBM Poughkeepsie Development Lab

Patricia Curry
IBM Fishkill

Simon Robertson
IBM UK

Bruno Blanchard
IBM France

Alvin Hua Juay Teng
IBM Singapore

Joseph Skovira
IBM Center, Parallel Computing Cornell Theory Center

EunYoung Chung, YoungSang Kim, SeckJun Song
HJS Inc., Paramus, New Jersey

Become a published author

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

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

ibm.com/redbooks

- Send your comments in an Internet note to:

redbook@us.ibm.com

- Mail your comments to:

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



Introduction

The IBM eServer Cluster 1600 high performance system uses the power of parallel processing to expand your applications. Designed and extended as a clustered IBM eServer pSeries system for performance, scalability, reliability, availability, serviceability, and management, this system makes feasible the processing of applications characterized by large scale data handling and compute-intensive applications.

The Cluster 1600 system is the IBM pSeries family of parallel computing solutions. It provides a state-of-the-art parallel computing system and industry-leading application enablers.

This redbook details the hardware and software components that make up the Cluster 1600. It also details the network connectivity required to join the components together and provides useful scenarios showing how you can utilize a Cluster 1600 system.

We cover the following topics in this chapter:

- ▶ 1.1, “Overview of Cluster 1600” on page 2
- ▶ 1.2, “Choosing PSSP or CSM” on page 3

1.1 Overview of Cluster 1600

Cluster 1600 builds on the success of RS/6000 SP technology, extends the benefits to more hardware building blocks, and provides flexibility for creating a new cluster configuration. Yet it reduces the complexity and cost of developing and managing servers in enterprise data centers. Cluster 1600 further exploits scalability, incrementally and non-disruptively, to the demands of your enterprise.

Enterprise applications and databases often exceed the capacity of even the largest single systems, and some enterprises are outgrowing systems faster than they can capitalize them. Cluster 1600 addresses this rapid growth and change, while controlling costs. Organizations need to move away from the model of buying dedicated servers for particular applications. Enterprises must find a non-disruptive solution to enable clustering and allocation of resources as demand rises and falls in different application environments.

The Cluster 1600 addresses the business needs of enterprises with a technical solution for both current customers and new customers. It further expands modularity by being able to support both AIX and Linux offerings.

The Cluster 1600 is a collection of interconnected computers used as a unified computing resource with a single administrative domain. The Cluster 1600 basic configuration starts with minimum of two supported servers interconnected by a network and interprocessor communications for I/O.

The Cluster 1600 has evolved from the RS/6000 SP through the addition of standalone “SP-attached servers” to where it is today. The Cluster 1600 looks very different from the traditional RS/6000 SP, but it essentially performs the same task with more flexibility than ever before.

Previously, we saw a change in the hardware associated with Cluster 1600, and we are now seeing a substantial change in the *software* that manages the cluster. The Cluster 1600 is in a transition period at the moment, as the management software is changing from Parallel System Support Programs (PSSP) to Cluster Systems Management (CSM). During this transition period, both PSSP and CSM are available to manage Cluster 1600 systems. PSSP will continue to be supported for a number of years in order to enable customers to make the move to CSM when they are ready.

The change to CSM has been driven by the need for a more flexible approach to clustering than PSSP is able to offer. PSSP is based on the “SP frame” concept, which is now less relevant with the new POWER4™ servers. CSM, on the other hand, is based on a *node* concept, rather than a frame concept.

CSM and PSSP cannot coexist in the same Cluster 1600 system, so you need to make careful choices when planning a new Cluster 1600 system, upgrading an existing Cluster 1600, or deciding to move to CSM. Sometimes the choices are easy, as some hardware is only supported on CSM and some hardware is only supported on PSSP. Likewise, some versions of application software are only supported on CSM and others are only supported on PSSP.

In other cases, where the hardware and application software are supported on both PSSP and CSM, the choice may not be so straightforward. In the following chapters, we help you to decide on the correct management software for your situation. For further information on choosing the most appropriate management software, contact your IBM sales representative or your IBM Business Partner.

In this redbook, we concentrate on the new Cluster 1600 enhancements. For information on existing RS/6000 SP and CES systems, refer to *RS/6000 SP and Clustered IBM @serverpSeries Systems Handbook*, SG24-5596.

1.2 Choosing PSSP or CSM

In the following sections, we describe the characteristics of both PSSP and CSM to help you to decide which software management product is most appropriate for your environment.

1.2.1 Cluster management with PSSP

PSSP enables a single point of control of all cluster nodes and resources. It simplifies system management and reduces the administrative and operational costs of managing distributed servers.

PSSP has always been the main system management software in SP implementations. Its function and usability has evolved over the years. Now, PSSP is enhanced to support a cluster of standalone pSeries servers, with or without standard SP frames and nodes.

PSSP management highlights include single point of control, system monitoring and management, system administration, and cluster security. In the following sections, we describe each item in more detail.

Single point of control

PSSP single point of control provides:

- ▶ Installation and configuration
- ▶ Installation of cluster nodes

- ▶ Simultaneous installation of nodes with tiered processes using “boot/install servers”
- ▶ Ability to upgrade one node at a time
- ▶ Hardware control
- ▶ Parallel tools, such as **dsh**, **sysctl**, and node groups
- ▶ Hardware and switch control, such as power on/off, reset, and so on
- ▶ Visual monitoring
- ▶ Control available through the command line, SMIT, and Perspectives GUI
- ▶ Base clustering technology
- ▶ Use of base clustering services such as Topology Services and Group Services
- ▶ Event Management and Event monitoring - provided by Reliable Scalable Cluster Technology (RSCT), the framework for clustering availability, scalability, and manageability

System monitoring and management

With PSSP system monitoring and management, you can do the following:

- ▶ Easily monitor system resources
- ▶ Monitor user-defined events
- ▶ Monitor and act on particular events (such as /tmp or /var filling up, daemon status, switch events, node status, and other events)
- ▶ Allow user-defined actions when events occur (such as SNMP traps, AIX errors, and so on)
- ▶ Use commands to control event management; resource monitors can be scripts
- ▶ Use SP and Tivoli® integration for users utilizing Tivoli
- ▶ Use a GUI and API interface for associating actions with a monitored event

System administration

PSSP provides tools to help you to simplify system management:

- ▶ Node Groups.
- ▶ File collection technology.
- ▶ Keep specified files on each cluster node identical by using the Software Update Protocol (SUP).
- ▶ SP log management.

- ▶ Log tracing - keep the trace of logs from each cluster node on the Control Workstation by using the splm tool.
- ▶ Time synchronization - use the Network Time Protocol (NTP) to synchronize the time all over the SP system.
- ▶ Automounter - The automounter is supported by AIX, and PSSP utilizes it. Users on each node can use their own home directory from the server on the request base.
- ▶ Network tunables - The node-specific network tunables are set using custom files and scripts, during the initial installation and also during regular operations.

Cluster security

Security administration is an important management task in a cluster environment. PSSP provides these tools for managing security in a cluster environment:

- ▶ AIX authentication
- ▶ Kerberos authentication
- ▶ Distributed Computing Environment (DCE) support
- ▶ Secure remote command support
- ▶ Firewall configuration inside a cluster

1.2.2 Cluster management with CSM

IBM Cluster Systems Management (CSM) provides a robust, powerful, and centralized way to manage large numbers of pSeries and xSeries servers in a Cluster 1600. CSM is comprised of a modular architecture, so that it integrates both IBM software and open source software into a complete systems management solution, but also allows administrators to decide which parts of CSM to use.

CSM leverages the rich heritage and proven technology of the IBM RS/6000 SP by utilizing and deriving various software from the PSSP systems management software product (such as **dsh** and **RSCT**).

There are command line interfaces available for all CSM functions. A high priority was placed on making the CSM command line interface consistent and streamlined. The CSM command line interface allows a user with the appropriate permissions to access all the resources in the system, their attributes, and state values. Query and control of nodes, file systems, CPU and memory statistics, global cluster parameters, and so on can all be accessed via the command line, and scripts can be written to take advantage of this.

On both Linux and AIX, a Distributed Command Execution Manager (DCEM) GUI is available to invoke remote commands across a CSM cluster.

On AIX there are SMIT panels available for install, setup, and hardware control functions in CSM. Also on AIX, an intuitive graphical interface available from the Web-Based System Manager GUI provides a powerful way for an administrator to manage and visually monitor their CSM cluster.

Some of the CSM management highlights include single point of control, CSM event management, system administration, Configuration File Management (CFM), and cluster security. In the following sections, we discuss these items in more detail.

Single point of control

With CSM single point of control, you can do the following:

- ▶ Install and update machines remotely
- ▶ Remotely power on, off and reboot nodes in the cluster
- ▶ Continuously monitor across all the machines in the cluster
- ▶ Automate responses that can be run any time a problem occurs, and which can provide notification or take corrective action
- ▶ Run probes across nodes in the cluster to diagnose problems
- ▶ Change files in one spot and distribute them to all the machines or a set of machines in the cluster
- ▶ Manage systems across multiple hardware domains
- ▶ Manage multiple high availability clusters inside a given CSM domain
- ▶ Handle large scaling, and handle distributed operations in parallel (such as remote command execution, monitoring and automated responses, installing and updating nodes)

In addition, CSM is independent of any switch topology or other types of domains, such as the HATS/HAGS domains used in PSSP. Also, CSM architecture is designed to have very few dependencies on the hardware platforms that it runs on, so it can be brought easily to other platforms

CSM event management

CSM event management provides the following:

- ▶ Monitoring for various conditions across nodes or node groups in the cluster, and running actions in response to events that occur in the cluster.
- ▶ Monitoring of network health, power status, state of applications and daemons, CPU, memory and file system utilization.
- ▶ Execution of conditional commands that can be run on the management server or on any node of the cluster or notification actions such as logging, e-mailing or paging.
- ▶ Generation of SNMP traps in response to events in the cluster.
- ▶ Predefined “Conditions” for many types of information.
- ▶ Predefined “Responses” for e-mail notification, SNMP traps, logging and displaying a message to a console
- ▶ The ability for a user to take a condition, quickly associate it with a response, and start monitoring. The administrator can easily customize these conditions and responses.
- ▶ Administrator-defined recovery actions, including cleaning up file systems that are filling up, taking actions to help restart a critical application that went down, and so on.

System administration

Following are CSM system administration highlights:

- ▶ The administrator installs the management server, defines the nodes to be in the cluster and then CSM can remotely do a parallel network AIX Operating System (OS) install of the nodes.
- ▶ CSM updates software on the nodes for new CSM versions and for new open source updates. If a node is down during an update, CSM automatically performs the update when the node comes back up.
- ▶ CSM automatically sets up the security configuration for the underlying cluster infrastructure. It can also do the necessary set up for rsh or ssh (exchange of ssh keys).
- ▶ An administrator can run commands in parallel across nodes or node groups in the cluster and gather the output using the **dsh** (distributed shell) command.
- ▶ The **dsh** command can use rsh (UNIX® basic remote shell) or ssh (secure shell). The administrator decides which one to use.

- ▶ The **dshbak** command can format the output returned from **dsh** if needed (for example, collapsing identical output from more than one node so that it is displayed only once).
- ▶ The administrator can run diagnostic probes provided by CSM to automatically perform “health checks” of particular software functions if a problem is suspected. These probes can also be run periodically or automatically, as a response to a condition occurring in the system.
- ▶ Current probes shipped with CSM include probes to diagnose network connectivity, NFS health, and the status of daemons that CSM runs.
- ▶ Administrators can also write their own probes to add into the existing CSM probe infrastructure.

Configuration File Management (CFM)

Configuration File Manager (CFM) is provided to synchronize and maintain the consistency in files across nodes in the cluster. This prevents the administrator from having to copy files manually across the nodes in the cluster. The particular files can be changed once on the management server, and then distributed to all the nodes or node groups in the cluster.

CFM can use make use of meta variables for IP address or hostname substitution in files being transferred. Scripts can also be run for processing before and after a file is copied (for example, to stop or start daemons and so on).

CFM makes use of **rdist** for file transfer; **rdist** can use **rsh** (UNIX basic remote shell) or **ssh** (secure shell). The administrator decides which one to use.

Cluster security

Distributed command execution (**dsh**) and CFM make use of either **rsh** or **openSSH**. CFM uses a push mechanism to distribute updates to the nodes.

The underlying cluster infrastructure uses an “out-of-the-box”, host-based authentication mechanism. The security is designed to be pluggable in order to more easily allow other security mechanisms to be used in the future.

Hardware control uses an encrypted id and password to communicate with the hardware control points. We recommend that you put the hardware control communication to the hardware control points and terminal servers on a VLAN that is separate from the VLAN used to install and manage the nodes.

The cluster security infrastructure provides an “out-of-the-box” security authentication mechanism based on hostnames and a public key/private key

setup. The security infrastructure allows for pluggable security mechanisms; Kerberos V5 can be supported in the future.

CSM automatically sets up and exchanges public keys between the management server and the managed nodes during a full install. Access to particular resources in the cluster are determined by ACL files that CSM assists with setting up. The managed nodes are not given access to any resources on the management server, thus preventing the management server from having to trust the managed nodes.



Cluster 1600 hardware

In this chapter, we describe the hardware building blocks that fit together to build a Cluster 1600 system. Because Parallel System Support Programs (PSSP) and Cluster Systems Management (CSM) support different hardware, we state which hardware is supported by CSM within a Cluster 1600, and which hardware is supported by PSSP within a Cluster 1600. We first look at the generic building blocks, and then look at the individual machines that make up a Cluster 1600 system.

Important: All cluster scalability limits shown in this chapter are subject to change without notice. For the latest scalability limits, refer to the IBM Systems Sales Web site:

www.ibm.com/servers/eserver/pseries

2.1 Overview

Cluster 1600 is a collection of interconnected computers used as a unified computing resource with a single administrative domain. The Cluster 1600 basic configuration starts with a *minimum* of:

- ▶ Two supported servers
- ▶ A cluster management console, and management software
- ▶ The cluster interconnected by network and interprocessor communications for input/output (I/O)

Traditional clustered server environments were most widely found in academic, research, and vast computational environments. Lately, however, server clustering is being utilized in commercial environments to achieve and maximize parallel application environments (for example, parallel database, server/workload consolidations, and redeployment of servers).

Cluster 1600 unifies the existing offerings of RS/6000 SP and Cluster Enterprise Servers (CES), and is managed by IBM cluster management software, the Parallel System Support Programs (PSSP) or Cluster Systems Management (CSM). A collection from the following systems can be part of this Cluster 1600. The intermix possibilities depend on the management software, including:

- ▶ Legacy RS/6000 SP
- ▶ Standalone servers such as RS/6000 S70/S7A and pSeries p680
- ▶ Rack-mounted servers such as pSeries 660 model 6H1/6M1/6H0
- ▶ All POWER4 servers
- ▶ All POWER4+ servers

Important: Refer to Table 2-2 on page 17 for server types supported by PSSP and CSM.

With the ability to “mix and match” cluster building blocks, ranging from low-cost pSeries servers running AIX to high-end pSeries 690 running AIX, the Cluster 1600 offers maximum flexibility in matching applications and services to computing resources.

Configuring a Cluster 1600 offers the following benefits and more:

- ▶ Single administrative domain
- ▶ Centralized topology
- ▶ Broader range of building blocks
- ▶ Functional enhancements
- ▶ Increased scalability
- ▶ Denser packaging
- ▶ Hybrid clusters (intermix of Linux and/or AIX environments)

- ▶ Grid-enabled
- ▶ Autonomic computing
- ▶ Reliability, Availability, Serviceability (RAS), performance and flexibility
- ▶ SMP scalability to extend outside of the SMP server, 64-bit exploration, broader range of cluster building blocks, mature software environment

The hardware components that make up a Cluster 1600 are described in 2.2, “Cluster 1600 hardware components” on page 13. We detail hardware that is supported on both CSM 1.3.2 managed clusters and PSSP 3.5 managed clusters.

Important: Not all hardware components are supported on both CSM and PSSP. Some components are only supported on CSM, and some components are only supported on PSSP.

Because PSSP and CSM cannot coexist in the same Cluster 1600, it is important to check the support for all components. Table 2-2 on page 17 references the components with support available from CSM and PSSP. Note that this support may change with later versions.

2.2 Cluster 1600 hardware components

A Cluster 1600 can be built from a collection of pSeries, RS/6000 and xSeries servers (CSM-managed clusters only), which form the building blocks of the cluster. The servers within the cluster are all networked to and managed by the control workstation (PSSP) or the management server (CSM). The servers are often connected together by a high bandwidth/low latency network. This network is referred to as “the switch network”.

The Cluster 1600 is very flexible, and every cluster can be very different from each other. To protect the investment, the Cluster 1600 can incorporate older servers, and in particular RS/6000 SP clusters. These older servers, which are not currently marketed, are referred to as *legacy hardware*.

Although the building blocks retain their own model type and serial number, when they are part of a Cluster 1600, they collectively take on another model type and serial number. Individually, each server has a feature code related to the Cluster 1600 that it is a member of. The Cluster 1600 model type is a 9078-160. This model type and the features attached to it are reflected in IBM’s administration records.

The Feature codes for Cluster 1600 servers are shown in Table 2-1.

Table 2-1 Cluster 1600 Feature codes

Description	M/T	Model	Feature
7017 server type	9078	160	0001
7026 server type	9078	160	0002
9076-555/557 type (SP Switch)	9078	160	0003
9076-556/558 type (SP Switch2)	9078	160	0004
9076-SP	9078	160	0005
9076 SP expansion frame	9078	160	0006
Control Workstation	9078	160	0007
7040 server type	9078	160	0008
LPAR 7040 (switched)	9078	160	0009
7039 server type	9078	160	0010
LPAR 7039 (switched)	9078	160	0011
7028 server type	9078	160	0012
LPAR 7028 (switched)	9078	160	0013
7038 server type	9078	160	0014
LPAR 7038 (switched)	9078	160	0015
7029 server type	9078	160	0016
7045-SW4 HPS pSeries HPS (High Performance Switch)	9078	160	0017

Note: CSM can manage the following servers within a single cluster:

- ▶ pSeries servers running AIX 5L
- ▶ Selected xSeries servers
- ▶ Selected IntelliStations and Blade Center cards

Cluster 1600 only supports the pSeries nodes shown in Table 2-1.

2.2.1 Nodes

Servers, or “nodes”, as they are known when clustered, are either pSeries, RS/6000 servers, SP nodes, xSeries servers, IntelliStation® or Blade Center nodes. These provide the computing power for the cluster. A Cluster 1600 can contain between 2 and 128 nodes. Larger configurations are available with the IBM special order process.

There are five types of Cluster 1600 nodes:

- ▶ High Nodes - RS6000/SP - legacy
- ▶ Wide Nodes - RS6000/SP - legacy
- ▶ Thin Nodes - RS6000/SP - legacy
- ▶ SP-attached servers and Clustered Enterprise Servers - legacy
- ▶ Selected pSeries servers - currently available

Attention: Not all nodes are supported on both CSM and PSSP. Table 2-2 on page 17 lists the supported nodes.

2.2.2 Frames

The IBM RS/6000 SP system frames contain and provide power for processor nodes, switches, hard disk drives and other hardware. A frame feature code provides an empty frame with its integral power subsystem and AC power cable. The nodes and the other components have different specific feature codes.

The frames are offered in a list of five options:

- ▶ Tall (1.93 m) model 550 frames - legacy
- ▶ Tall expansion frames (F/C 550) - legacy
- ▶ Short (1.25 m) model 500 frames - legacy
- ▶ Short expansion frames (F/C 1500) - legacy
- ▶ SP switch frames (F/C 2031 for SP Switch and F/C 2032 for SP Switch2) - currently available
- ▶ 7040-W42 p655 racks - currently available

Frames have locations known as *drawers* into which the processor nodes are mounted. Tall frames have eight drawers and short frames have four.

Each drawer location is further divided into two *slots*. A slot has the capacity of one thin node or SP Expansion I/O unit. A wide node occupies one full drawer, while a high node occupies two full drawers. The maximum number of SP frames supported in an SP system is 128.

2.2.3 Switches

The SP switches provide a message-passing network that connects all processor nodes with a minimum of four paths between any pair of nodes. The switch provides a high bandwidth, low latency network which can use either TCP/IP or MPI protocol to pass messages. With high performance parallel applications, normally the MPI protocol is used. Other typical uses are for performing high speed network backups.

The SP series of switches can also be used to connect the SP system with optional external devices when a switch router is used. A switch feature code provides you with a switch assembly and the cables to support node connections. The number of cables you receive depends on the type of switch you order.

The SP switches available with PSSP 3.5 support are the SP Switch or the SP Switch2:

- ▶ SP Switch (F/C 4011), 16-port switch (available until December 31, 2003)
- ▶ SP Switch2 (F/C 4012), 16-port switch

The switch available with CSM 1.3.2 support is the pSeries HPS (High Performance Switch):

- ▶ pSeries High Performance Switch (HPS), 7045-SW4, 32-link switch

Important: PSSP does not manage the High Performance Switch, and CSM does not manage SP switches.

2.2.4 PSSP control workstations

You can view the control workstation (CWS) as the control point of the Cluster 1600 system managed by PSSP. The subsystems running on the CWS are for controlling and monitoring the Cluster 1600. The CWS subsystems provide configuration data, security, hardware monitoring, diagnostics, a single point of control service, and optionally, job scheduling data and a time source. The CWS must be a pSeries or an RS/6000 system running AIX.

2.2.5 CSM for AIX management server

You can think of the management server as the control point of the Cluster 1600 system managed by CSM. The management server is very similar to the control workstation on a PSSP Cluster 1600. The management server provides configuration data, hardware monitoring, diagnostics, a single point of control service, and node installation. The CSM management server provides the

interface for the remote hardware control capability. Remote hardware control allows you to power on and off, reboot, bring up a console through a tty connection and query nodes from the management server. The management server for a Cluster 1600 can be a pSeries or an RS/6000 system running AIX.

2.2.6 Hardware Management Console (HMC)

The Hardware Management Console (HMC) is a dedicated desktop workstation that gives you a GUI for configuring and operating multiple pSeries servers in SMP, or LPAR, or clustered environments. First introduced with the p690, the HMC can now provide hardware control and console facilities for the full POWER4 and later range of pSeries servers.

The HMC is used to configure and manage LPARs, dynamic LPARs, and Capacity Upgrade on Demand (CUoD), as well as the basic console and hardware control functions.

The HMC is the interface between the CWS or management server and the POWER4 or later nodes, and is mandatory with Cluster 1600 systems.

Note: The HMC cannot be used as either the CWS or a CSM management server.

2.3 CSM and PSSP hardware support

Hardware support varies between CSM and PSSP. Hardware which is supported by PSSP may not be supported on CSM, and vice versa. The general direction is that new hardware being released will be supported on CSM but not PSSP.

Table 2-2 shows the hardware supported by PSSP 3.5 and CSM 1.3.2 within a Cluster 1600.

Table 2-2 CSM 1.3.2 and PSSP 3.5 Cluster 1600 supported hardware

Hardware description	PSSP 3.5	CSM 1.3.2
Nodes		
9076 SP nodes PWR_1, PWR_2, P2SC, thin and wide	Yes	No
9076 SP nodes 601, 604, 604e high nodes	yes	No
9076 SP nodes 332 MHz SMP thin and wide (F/C 2050, F/C 2051)	Yes	Yes
9076 SP nodes POWER3™ thin and wide (F/C 2052, F/C 2053)	Yes	Yes

Hardware description	PSSP 3.5	CSM 1.3.2
9076 SP nodes 375 MHz/450 MHz thin and wide (F/C 2056, F/C 2057)	Yes	Yes
9076 SP nodes POWER3 and 375 MHz POWER3 high (F/C 2054, F/C 2058)	Yes	Yes
7017 S70, S7A, S80, S85	Yes	No
7026 H80, M80, 6H0, 6H1, 6M1	Yes	Yes
7028 6C4 (p630)	Yes	Yes
7029 6C3 (p615)	No	Yes
7038 6M2 (p650)	Yes	Yes
7039 651 (p655)	Yes	Yes
7040 671, 681 (p670, p690)	Yes	Yes
Switches		
SP Switch	Yes	No
SP Switch2	Yes	No
7045-SW4 pSeries HPS (High performance switch)	No	Yes

2.3.1 CSM-managed node requirements

The following section describes the hardware requirements for supported CSM-managed nodes.

There is a distinction to be made between a Cluster 1600 and a CSM-managed cluster:

- ▶ A Cluster 1600 is restricted to the nodes identified in Table 2-2 on page 17.
- ▶ A CSM-managed Cluster 1600 supports all the nodes that are supported in the Cluster 1600 and in addition, the nodes listed in Table 2-3 on page 19.

We support both AIX and Linux as managed nodes, up to 128, in the CSM for AIX cluster environment.

This means that you can have a maximum of 128 nodes, including both AIX and Linux nodes.

Table 2-3 CSM supported nodes

Hardware description	PSSP 3.5	CSM 1.3.2
xSeries 330, 342, 335, 345, 360, 440, running Linux	No	Yes
Blade servers running linux (8677 and 8678)	No	Yes
IntelliStation 6221(Hardware is controlled for the IntelliStation 6221 through the APC Master Switch)	No	Yes
eServer 325 running Linux (AMD system)	No	Yes

Tip: Each node type is required to have at least 128 MB of RAM.

AIX-managed nodes

AIX-managed nodes must be one of the following servers, running AIX 5L Version 5.1 with Recommended Maintenance Package 5100-03, or later, plus the latest APARs:

- ▶ pSeries server
- ▶ RS/6000
- ▶ A partition of pSeries server (LPAR)
- ▶ A CSM-supported SP node (see Table 2-2 on page 17)

Tip: You may configure the management server as a managed node.

Linux-managed nodes

The following servers are supported as Linux-managed nodes:

- ▶ CSM-supported xSeries servers. For a list of supported servers, refer to Table 2-2 on page 17.

Important:

- ▶ Supported servers must run Red Hat 7.2, 7.3, 8.0, AS 2.1 or SUSE 8.0, 8.1, SLES 7 or SLES 8 Linux. Contact your IBM sales representative or IBM Business Partner for the list of latest Linux distributions supported.
- ▶ Only selected xSeries nodes, capable of hardware control, are supported by CSM.

- CSM-supported pSeries servers.

Note: These servers must run Red Hat AS 2.1, or SUSE SLES 7 or SLES 8 Linux.

2.4 PSSP control workstation

A Cluster 1600 system requires a customer-supplied pSeries or RS/6000 system known as a *control workstation*. The control workstation serves as a point of control for managing, monitoring, and maintaining the Cluster 1600 system frames and individual processor nodes. A system administrator can perform these control tasks by logging into the control workstation from any other workstation on the network.

The control workstation can also act as a boot/install server for other servers in the Cluster 1600 system. In addition, the control workstation can be set up as an authentication server using Kerberos. The control workstation can be the Kerberos primary server, with the master database and administration service, as well as the ticket-granting service. As an alternative, the control workstation can be set up as a Kerberos secondary server, with a backup database, to perform ticket-granting service.

Kerberos is no longer the only security method. The Distributed Computing Environment (DCE) can be used with Kerberos V4, or by itself, secure remote command method (SSH) or no security (AIX standard security).

A high availability solution, named the High Availability Control Workstation (HACWS), may be implemented for the control workstation.

Note: A CWS is required on all PSSP-managed Cluster 1600 systems, even when POWER4 nodes are used with an HMC.

2.4.1 Control workstation requirements

The control workstation has the following requirements:

- pSeries or RS6000 server (from the supported list on Table 2-4 on page 22)
- RS-232 port available for each frame/server

Note: An RS-232 connection is not required between the CWS and the HMC. However, it *is* required between the HMC and the POWER4 server.

Restriction: The native RS-232 ports on the system planar cannot be used as tty ports for the hardware controller interface. Either an 8-port or 128-port async adapter must be ordered if RS-232 ports are required.

- ▶ Two RS-422 ports available for each p655 frame (8-port async card)
- ▶ Ample disk space for mksysbs and NIM requirements
- ▶ Suitable Ethernet network adapters
- ▶ Graphic adapter and graphical screen
- ▶ AIX, PSSP, C++ or C for AIX

Tip: The CWS must have at least 128 MB of main memory. An extra 64 MB of memory should be added for each additional system partition. For Cluster 1600 systems with more than 80 nodes, 256 MB is required and 512 MB of memory is recommended.

The CWS must have at least 9 GB of disk storage. If the SP is going to use an HACWS configuration, you can configure 9 GB of disk storage in the rootvg volume group and 9 GB for the spdata in an external volume group.

Because the control workstation is used as a Network Installation Manager (NIM) server, the number of unique filesets required for all the nodes in the SP system might be larger than a normal single system. You should plan to reserve 6 GB of disk storage for the file sets and 2 GB for the operating system. This will allow adequate space for future maintenance, system mksysb images, and LPP growth.

Keep in mind that if you have nodes at different levels of PSSP or AIX, each node requires its own LPP source, which takes up extra space. A good rule of thumb to use for disk planning for a production system is 4 GB for the rootvg to accommodate additional logging and /tmp space, plus 4 GB for each AIX release and modification level for lppsource files. Additional disk space should be added for mksysb images for the nodes.

If you plan on using rootvg mirroring, then for one mirror, double the number of physical disks you estimated so far. For two mirrors, triple the estimate.

2.4.2 Supported control workstations

Table 2-4 on page 22 lists the currently available supported pSeries control workstations for PSSP 3.5.

Table 2-4 Supported control workstations on PSSP 3.5

Machine type	Models
7028	p630 - 6C4, 6E4
7038	p650 - 6M2
7044	170

Note: For supported legacy CWS, refer to *RS/6000 SP and Clustered IBM eServer pSeries System Handbook*, SG24-5596.

A typical control workstation configuration for a large Cluster 1600 is shown in Example 2-1. Smaller clusters may require less disk space and less CPU, although the network requirements remain the same.

Example 2-1 Typical control workstation configuration

Product	Description	Qty
7028-6E4	Desktop Server 1:pSeries 630	1
2633	IDE CD-ROM Dr.(Black bezel)	1
2848	POWER GXT135P GRAPHICS ACC	1
2943	8-Port Ad.EIA-232/RS-422	1
3158	36.4 GB 10K RPM U3 SCSI DDA	2
3159	73.4 GB 10K RPM U3 SCSI DDA	2
3628	Color Monitor,St.Black and C.	1
4254	SCSI Connector Cable	1
4451	1024MB DIMMs, 208-pin, 8NS DDR	1
4961	Univ 4-Port 10/100 Enet Ad.	2
5005	Software Preinstall	1
5134	2way 1.2GHz POWER4+ Proc Card	2
6273	Power Supply, 645 Watt AC, HS	1
6568	U3 SCSI Backplane for HS Disks	1
8700	Quiet Touc Keyb.,US Engl.	1
8741	3-Button Mouse-Business Black	1
9300	Lang.Group Spec.-US English	1
9556	6 Slot PCI Riser Init ord only	1
9825	Power Cord - U.K.	1

2.4.3 High Availability Control Workstation

The High Availability Control Workstation (HACWS) is a component that is used to reduce the possibility of single point of failure in the SP system. Although there

already are redundant power supplies and replaceable nodes, there are also many elements of hardware and software that could fail on a control workstation.

With a HACWS, your SP system has the added security of a backup control workstation. Also, HACWS allows your control workstation to be powered down for maintenance or updating without affecting the entire SP system.

The design of the HACWS is modeled on the High Availability Cluster Multi-Processing for the AIX (HACMP) Licensed Program Product. HACWS utilizes HACMP running on two RS/6000 control workstations in a two-node rotating configuration.

HACWS utilizes an external disk that is accessed non-concurrently between the two control workstations for storage of SP-related data. There is also a dual RS-232 frame supervisor card with a connection from each control workstation to each SP frame in your configuration.

This HACWS configuration provides automated detection, notification, and recovery of control workstation failures. Figure 2-1 on page 24 shows the HACWS overview.

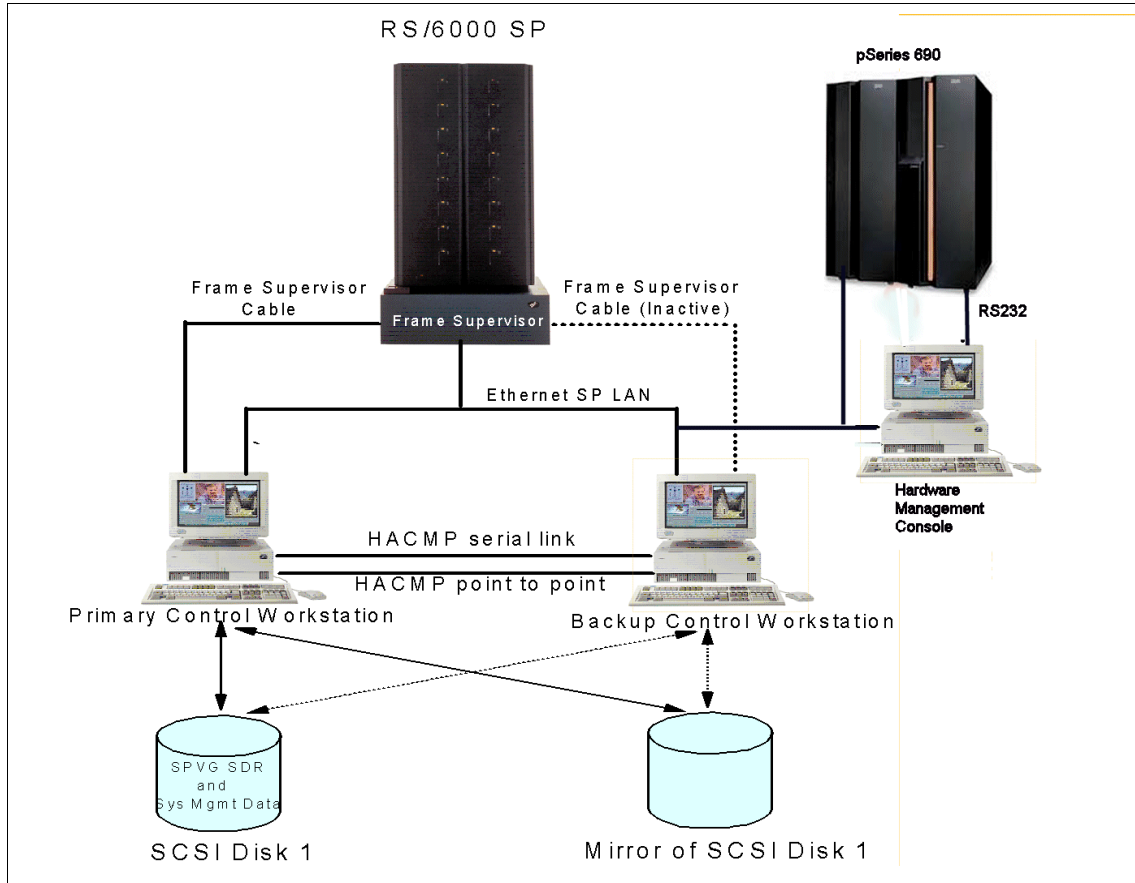


Figure 2-1 HACWS overview

HACWS provides a high availability solution on the control workstation, which serves as a single point of control for managing and maintaining the SP nodes using Parallel System Support Programs (PSSP).

Depending on your system environment and the type of applications that are running on your CWS, the impact of a failure in the CWS may not affect the operation of a Cluster environment. However, if your needs are such that you choose to run critical applications on the CWS, then you should consider protecting it with the HACWS program.

Users who do not run any critical applications on the CWS, or who are unconcerned about spending time restoring system backup and rebuilding the control workstation after its failure, might not require the high availability configuration. For such users, however, HACWS still provides minimum downtime for maintenance of the CWS.

The loss or failure of the CWS has the following effects on the management of an Cluster 1600 system, but should have little or no impact on the active jobs on the cluster nodes:

- ▶ It is not possible to control the cluster hardware.
- ▶ The System Data Repository (SDR) is unavailable.
- ▶ Existing jobs continue to completion, but new parallel jobs cannot be started.
- ▶ No configuration changes can be made.
- ▶ Software installations can not be done from the CWS.
- ▶ Should a switch fault occur, reset processing cannot be completed.
- ▶ Error logging of alerts raised by SP/Cluster nodes are lost (although the information is still logged on the individual nodes).
- ▶ While the CWS is unavailable, some administrative tasks that use Parallel System Support Programs (PSSP) are not able to proceed.

The HACWS configuration requires external disks that provide a non-concurrent access feature for both primary and backup control workstations. In a typical configuration, the cluster management data, as well as the AIX system images, PSSP, and related software install file sets, NIM configuration data files, and any other software installp filesets should reside on the external shared disks. Any disks supported by HACMP and RS/6000 models for both the primary and backup control workstations can be used.

Each CWS requires the same number of connections to the SP Ethernet on the same LAN segments. Each SP Ethernet LAN segment must be cabled to the same Ethernet (enx) adapter. Standby network adapters are optional in the HACWS configuration. However, the presence of a standby adapter may avoid the need to fail over (switch to backup) to the inactive CWS in case of a single LAN adapter failure.

2.4.4 HACWS limitations

Be aware of the following limitations before implementing HACWS:

- ▶ There is no hardware control to SP-attached nodes or CES when the CWS fails over. The limitation is due to the serial connection from the CWS to the attached servers.

- ▶ The SP-attached servers do not have a secondary serial port to allow an RS-232 connection to the secondary CWS.
- ▶ These limitations do not apply to HMC-attached servers.

Note: HACWS configuration works with either HACMP “classic” or HACMP Enhanced Scalability.

2.5 CSM management server

The management server is very similar to the control workstation on a PSSP Cluster 1600. The management server provides configuration data, hardware monitoring, diagnostics, a single point of control service, and node installation. The management server can be any RS/6000, pSeries server or pSeries LPAR that is capable of running AIX 5.2. It can even be one of the managed nodes.

Restriction: A p655 cannot be a management server, as there are no media drawers.

The following are the hardware requirements for the CSM management server:

- ▶ The machine you use for your management server must have a CD-ROM drive.
- ▶ The management server must be a workstation capable of running AIX 5L Version 5.2.

2.5.1 Memory and disk space

On the management server, a minimum of 1024 MB of memory and 120 MB of disk space is required for installing CSM. An additional 2.0 GB of disk space is required for installing the AIX operating system and CSM. In addition, 20 MB is required if you are managing Linux nodes, at least 1 GB for NIM and at least 1 GB if storing backups from the managed nodes.

2.5.2 Network requirements

Ethernet adapters are required. When configuring a CSM cluster, give particular attention to secure hardware control functions. We suggest that you define and isolate the management server on a private Ethernet network when connecting to a management VLAN (a virtual LAN that connects the management server to the cluster hardware control points). One Ethernet connection is also required for

each cluster VLAN (virtual LAN that connects the nodes to each other and to the management server). You may also want to have a separate network connection that can connect to a public LAN.

2.5.3 Asynchronous card requirements

An asynchronous card will only be required if you are controlling non-HMC attached nodes, such as p660s or legacy SP nodes. Both 8-port and 128-port asynchronous adapter cards are available.

2.5.4 Using a Logical Partition (LPAR) as a CSM management server

CSM 1.3 for AIX supports a logical partition (LPAR) as a CSM management server with the limitations and considerations listed below. Depending on your cluster configuration and needs, you may choose to use an LPAR as the CSM management server. However, it is important to understand the limitations and considerations to decide if an LPAR CSM management server is appropriate for your cluster.

Considerations for an LPAR management server

1. The CSM management server can be brought down inadvertently by someone on the HMC deactivating that LPAR. Even if someone does not have access to the CSM management server, if they still have access to the HMC, they can power off the management server. It can also be effected by someone moving resources such as CPU or I/O from that LPAR.
2. If the firmware needs to be upgraded, the LPAR management server may go down along with the rest of the CEC in order to upgrade the firmware. However, upon bringing the CEC back up, the system will return to normal.
3. There is no direct manual hardware control of the CSM management server. The administrator must go through the HMC for power control of the management server.
4. In many cases a physically separate CSM management server can be safer, since an LPAR management server belongs to a CEC that is down for a hardware or power failure, you will lose access to your management server.
5. An LPAR management server may not have an attached display. This may affect the performance of your CSM GUIs.
6. An LPAR management server may not contain media devices such as CD, tape, or diskette drives. This can affect your back-up strategy. In machines such as the p690, you may be able to assign a CD-ROM drive to the management server LPAR.
7. An LPAR management server should not also be defined as a managed node.

Note: A cluster that is installed and configured can still function even if the management server goes down. For example, cluster applications can continue to run, and nodes in the cluster can be rebooted, even if the management server is down.

However, tasks including monitoring, automated responses for detecting problems in the cluster, and scheduled file and software updates will not occur while the management server is down. Depending on your cluster configuration and use, this may or may not be a concern.

2.6 Cluster 1600 server concepts

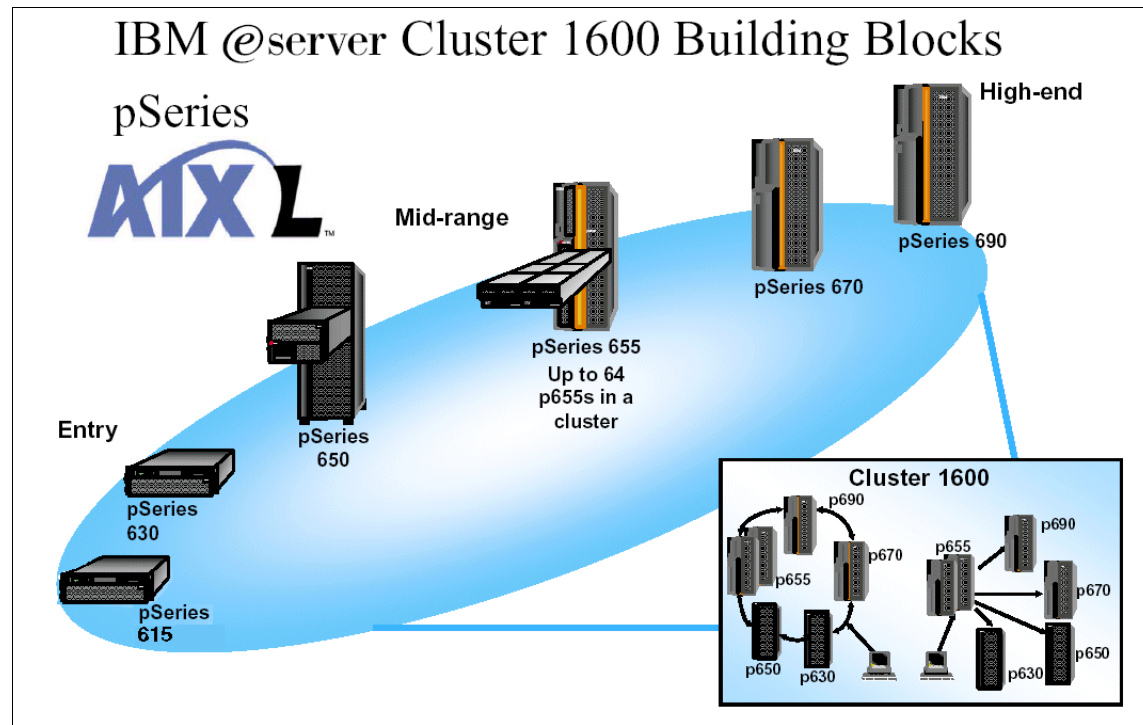


Figure 2-2 Cluster 1600 servers

Servers currently available for the Cluster 1600 are from the pSeries range. The servers shown in Figure 2-2 are the cluster-enabled pSeries servers. The currently marketed pSeries servers which can be attached to a Cluster 1600 all contain POWER4 or later technology.

As a result of the LPAR technology introduced with POWER4, a node can either be a physical server or an LPAR within a server. Logical partitions (LPARs) run separate instances of AIX or Linux within the same server. LPARs are completely isolated from each other and use separate resources allocated from the server. CPU, memory and I/O can be allocated to any LPAR with the granularity of a single processor, 256 MB of memory and any single PCI/PCI-X adapter.

Figure 2-3 displays the concepts of LPAR. With AIX version 5.2, these LPARs can be dynamic, which means that you can change resources without the requirement to reboot.

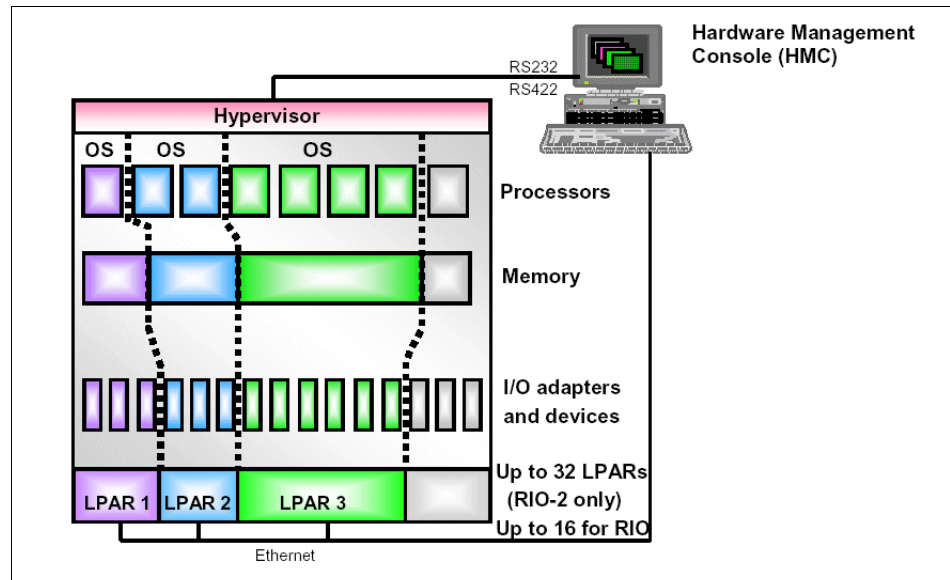


Figure 2-3 LPAR concept diagram

For more information on LPARs, refer to *The Complete Partitioning Guide for IBM pSeries Servers*, SG24-7039.

Because of LPARs, we now talk about *operating system images* as opposed to *servers*. All current pSeries servers can only be attached to a Cluster 1600 when they are controlled by a HMC. A Cluster 1600 can consist of 2 to 128 AIX operating system images. An operating system image, or logical node, can be one of the following:

- ▶ A 7040 server (p670, p690) running as a full system partition
- ▶ An LPAR of a 7040 server (p670, p690)
- ▶ A 7039 server (p655) running as a full system partition
- ▶ An LPAR of a 7039 server (p655)
- ▶ A 7038 server (p650)

- An LPAR of a 7038 server (p650)
- A 7028 server (p630)
- An LPAR of a 7028 server (p630)
- A 7026 server (H80, M80, p660, 6H0/6H1, 6M1)
- A 7017 server (S70, S7A, S80, p680)
- A 9076 SP node

Logical nodes are limited in a cluster running PSSP as described in Table 2-5. Any cluster system that exceeds any of the supported logical node limits requires a special bid or request for price quotation (RPQ).

Table 2-5 Cluster limits for a Cluster 1600 managed by PSSP

Nodes	SP Switch	SP Switch2	SP Switch2 (total switched plus non-switched)	Industry std interconnect (non-switched)
p690/p670 servers per cluster	32	32	32	32
p655 servers per cluster		64	64	64
p650 servers per cluster		64	64	64
p630 servers per cluster		64	64	64
LPARs per p690 server	8	32	32	32
LPARs per p670 server	4	8	16	16
LPARs per p655 server		2	4	4
LPARs per p650 server		2	8	8
LPARs per p630 server		2	4	4
LPARs per cluster	128	128	128	128

Note: With the SP Switch, switched and non-switched logical nodes cannot be mixed.

A Cluster 1600 with PSSP must meet *all* of the limits listed in Table 2-6; otherwise, a special bid order is required.

Table 2-6 A Cluster 1600 with PSSP must meet all of the following limits

No more than 128 logical nodes from the set (7040, 7039, 7038, 7028, 7026, 7017, 9076)
No more than 32 servers from the set (7040)
No more than 64 servers from the set (7039)
No more than 32 servers from the set (7038)
No more than 64 servers from the set (7040, 7039, 7038, 7028, 7026, 7017)
No more than 16 servers from the set (7017)
No more than 128 9076 SP Nodes

For example:

- ▶ 32 p690s with 4 LPARs each or 16 p690s with 8 LPARs each
- ▶ 32 p660s, 12 p690s with 4 LPARs each, 4 p690s with 8 LPARs each and 16 p630s
- ▶ 16 p690s with 4 LPARs each, 32 p660s and 32 SP nodes
- ▶ 12 p690s with 4 LPARs each, 16 p680s, 16 p660s and 48 SP nodes

Logical nodes are limited in a cluster running CSM, as described in Table 2-7. Any cluster system that exceeds any of the supported logical node limits requires a special bid or RPQ.

Table 2-7 Cluster limits for CSM

Node maximums	IBM eServer pSeries High Performance Switch (HPS)	HPS (total switched plus non-switched)	Industry std interconnect (non-switched)
p690 servers per cluster	16	32	32
p670 servers per cluster	N/A	N/A	32
p655 servers per cluster	16	64	64
p650 servers per cluster	N/A	N/A	64
p630 servers per cluster	N/A	N/A	64
LPARs per p690 server	16	16	16

Node maximums	IBM eServer pSeries High Performance Switch (HPS)	HPS (total switched plus non-switched)	Industry std interconnect (non-switched)
LPARs per p670 server	N/A	N/A	16
LPARs per p655 server	4	4	4
LPARs per p650 server	N/A	N/A	8
LPARs per p630 server	N/A	N/A	4
LPARs per cluster	128	128	128

A Cluster 1600 with CSM must meet all of the limits in Table 2-8; otherwise, a special order bid is required.

Table 2-8 A Cluster 1600 with CSM must meet all of the following limits

No more than 128 logical nodes from the set (7040, 7039, 7038, 7028, 7026, 9076)
No more than 32 servers from the set (7040)
No more than 64 servers from the set (7039)
No more than 64 servers from the set (7038)
No more than 64 servers from the set (7040, 7039, 7038, 7028, 7026)

For example:

- ▶ 32 p690s with 4 LPARs each or 16 p690s with 8 LPARs each
- ▶ 32 p660s, 12 p690s with 4 LPARs each, 4 p690s with 8 LPARs each and 16 p630s
- ▶ 16 p690s with 4 LPARs each, 32 p660s and 32 SP nodes
- ▶ 12 p690s with 4 LPARs each, 16 p680s, 16 p660s and 48 SP nodes

Attention: The cluster scalability for Linux on pSeries is the same as AIX on pSeries, which is 128 nodes.

2.6.1 pSeries architecture

pSeries servers that are eligible for cluster attachment are all POWER4 or later architecture. The larger servers (p655, p670, and p690), all use Multiple Chip

Modules (MCM). The mid-range servers (p630 and p650) use Single Chip Modules (SCM). The POWER4 architecture and both SCM and MCM packaging is described in the following section.

POWER4 chip

The components of the POWER4 chip are shown in Figure 2-4. The chip has two processors on board. Included in what we are referring to as “the processor” are the various execution units and the split first level instruction and data caches.

The two processors share a unified second level cache, also on board the chip, through a Core Interface Unit (CIU), as shown in the figure. The CIU is a crossbar switch between the L2 (implemented as three separate, autonomous cache controllers), and the two processors. Each L2 cache controller can operate concurrently and feed 32 bytes of data per cycle.

The CUI connects each of the three L2 controllers to either the data cache or the instruction cache in either of the two processors. Additionally, the CUI accepts stores from the processors across 8-byte wide buses and sequences them to the L2 controllers.

Each processor has associated with it a non-cacheable (NC) Unit (the NC Unit in Figure 2-4) that is responsible for handling instruction serializing functions and performing any non-cacheable operations in the storage hierarchy. Logically, this is part of the L2.

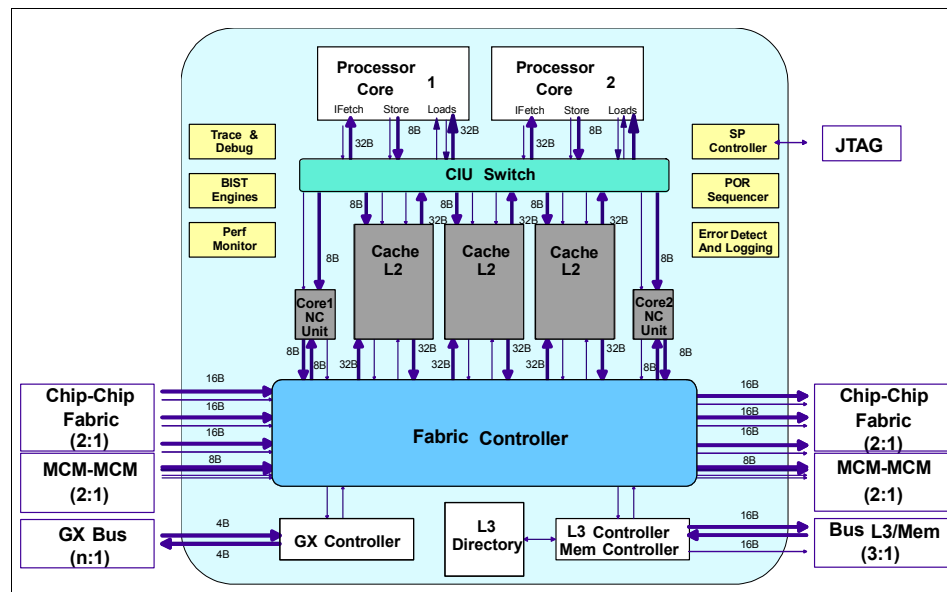


Figure 2-4 POWER4 logical description

The directory for a third level cache, L3, and logically its controller, are also located on the POWER4 chip. The actual L3 is on a separate 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.

The GX controller is responsible for controlling the flow of information in and out of the system. Typically, this would be the interface to an I/O drawer attached to the system. However, with the POWER4 architecture, this is also where we would natively attach an interface to a switch for clustering multiple POWER4 nodes together.

Also included on the chip are functions we logically call *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 (SP) used to control the overall system, power-on Reset (POR) Sequencing logic, and Error Detection and Logging circuitry.

Multi-Chip Modules (MCM)

The POWER4 chips are packaged on a single module called Multi-Chip Modules (MCM). 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, as shown in Figure 2-6 on page 36.

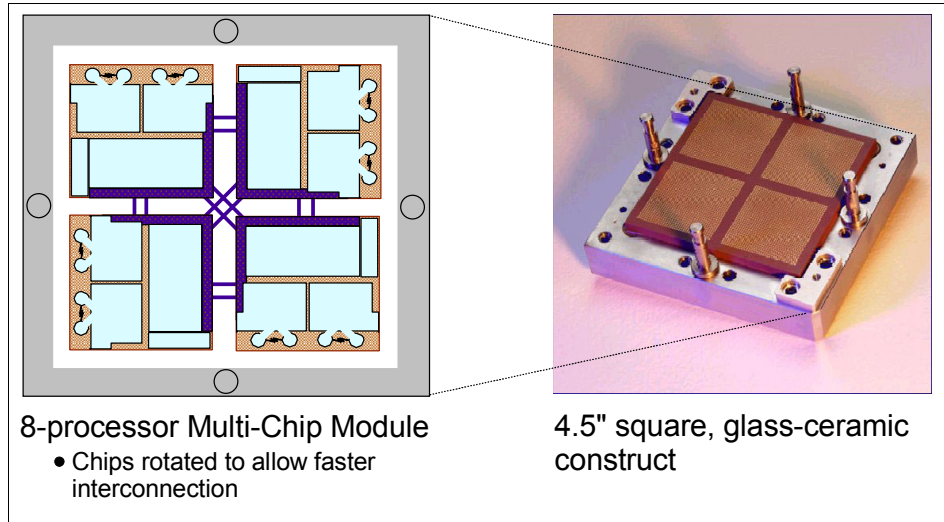


Figure 2-5 POWER4 Multi-Chip Module (MCM)

An internal representation of the MCM is shown in Figure 2-6 on page 36, 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

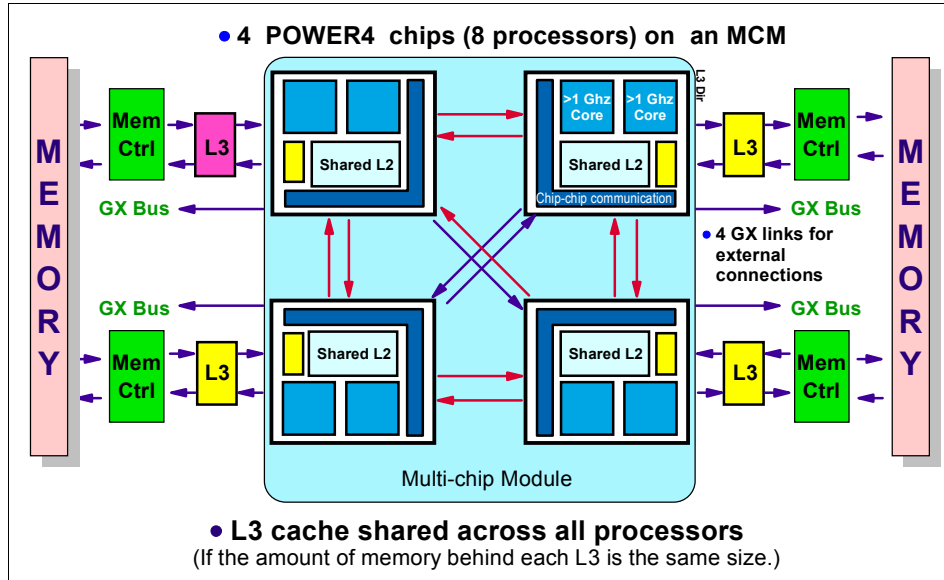


Figure 2-6 Multi-Chip Module with L2, L3, and memory

The MCM is a proven technology that IBM has been using for many years in the mainframe systems (now IBM eServer 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

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.

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.

For more information on MCMs and POWER4 architecture, refer to the white paper “POWER4 System Micro architecture”, which can be found at:

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

You can also refer to the IBM Redbook *IBM @server pSeries 690 and pSeries 670 System Handbook*, SG24-7040.

Single Chip Module (SCM)

Single Chip Modules are used in the p630 and p650 mid-range servers and contain one single POWER4+ chip with either one or two processor cores (CPUs). The SCM is permanently mounted on a processor card. One key difference is that the chip-to-chip fabric bus (which was used between chips on the same MCM) is no longer relevant in the SCM. It is replaced by a module-to-module fabric.

Each SCM is a Ceramic Column Grid Array (CCGA) package where the chip carrier is raised slightly from its board mounting by small metal solder columns that provide the required connections and improved thermal resilience characteristics. Similar to the SCM, each processor card also contains the L3 cache and the memory DIMMs, as shown in Figure 2-7 on page 38. The processor card is mounted in a rugged metal enclosure (book) that protects and secures the card (both in and out of the server), and helps manage airflow used for cooling.

The POWER4+ storage subsystem consists of three levels of cache and the memory subsystem. The first two levels of cache are on board the POWER4+ chip. The first level is 64 KB of Instruction (I) and 32 KB of Data (D) cache per processor core. The second level is 1.5 MB of L2 cache on the POWER4+ and 1.44 MB on the POWER4 chip.

All caches have either full ECC₁ or parity protection on the data arrays, and the L1 cache has the ability to re-fetch data from the L2 cache in the event of soft errors detected by parity checking. A 2-way configuration using two 1-way processor cards offers better performance than a configuration using a single 2-way processor card, because the maximum capacity of memory is doubled from 16 GB to 32 GB and the Level 2 (L2) and Level 3 (L3) cache on each card is dedicated to a single processor. A 2-way configuration using a 2-way processor card shares the L2 and L3 caches.

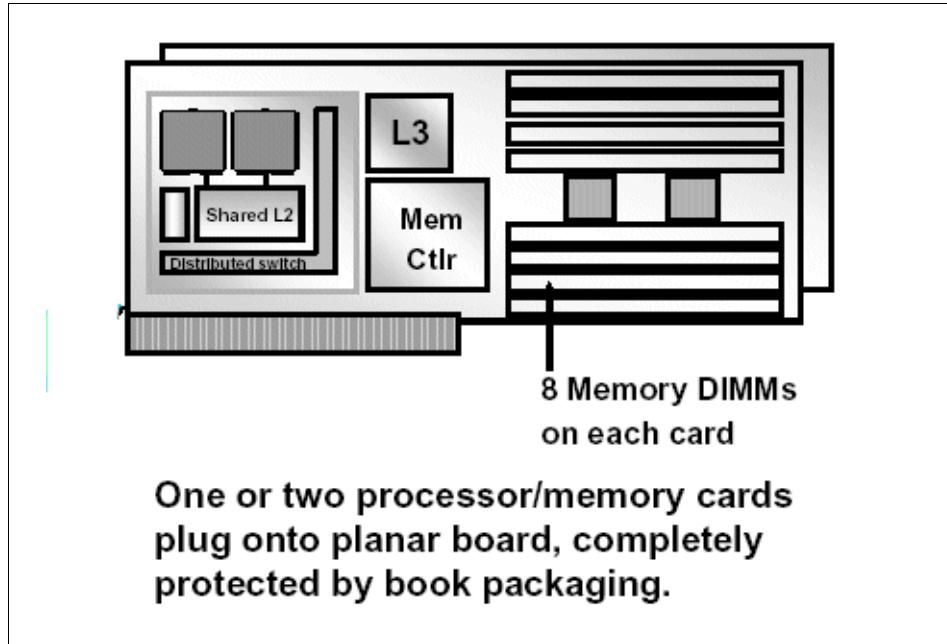


Figure 2-7 POWER4+ processor card layout

2.6.2 Cluster 1600 and the HMC

A prerequisite to attaching a POWER4 or POWER4+ pSeries server to a Cluster 1600 is the Hardware Management Console (HMC). The current model of the HMC is the IBM 7315-C02. The IBM 7315-C01 with feature codes (F/C 7315) and (F/C 7316) can also be used for Cluster 1600 attachment.

The IBM 7315-C02 HMC is a dedicated desktop workstation that gives you a GUI for configuring and operating multiple pSeries servers in SMP, LPAR, or clustered environments. It comes with a hardware management application for configuring and partitioning the server.

The IBM 7315-C02 HMC delivers functions necessary to manage LPAR configurations:

- ▶ Creating and storing LPAR profiles that define the processor, memory, and I/O resources allocated to an individual partition
- ▶ Starting, stopping, and resetting a system partition
- ▶ Booting a partition or system by selecting a profile
- ▶ Displaying system and partition status

- Displaying a virtual operator panel of the contents for each partition or controlled system

The HMC offers a service focal point for the systems it controls. It is connected to the service processor of the system via a dedicated serial link. The HMC provides tools for problem determination and service support, such as call home and error log notification through an analog phone line.

Note: The IBM 7315-C02 HMC is a dedicated function device. It is utilized only for the control and service functions of the pSeries servers it serves. It is not available for use as a general purpose computing resource.

The HMC can control servers of multiple machine types. When combining multiple machine types on an HMC, substitutions of other server types are based upon the relative weighting factor of the server. For example, a p690 is twice the weighted factor of a p630. A supported HMC configuration could control four p690 servers and eight p630 servers. For the number of servers supported per HMC, refer to Table 2-9.

Table 2-9 Number of servers per HMC

Server	SP Switch	SP Switch2	SP Switch2 (total switched plus non-switched, subject to note 2 limits)	Industry std interconnect (non - switched)
p690/p670	8	8	8	8
p655		16	16	16
p650		16	16	16
p630		16	16	16
Number of LPARs per HMC	32	32	32	32

Note: With the SP Switch, switched and non-switched logical nodes cannot be mixed.

Redundant HMC

Each system that supports a Hardware Management Console (HMC) has two HMC port RS-232 connections, so that you may optionally attach a second HMC to the same system. The benefits of using two HMCs are:

- ▶ It ensures that access to the HMC management function capabilities are not interrupted.
- ▶ It ensures access if the network is down.

In configurations with two HMCs, both HMCs are fully active and accessible at all times, enabling you to perform management tasks from either HMC at any time. There is no primary or backup designation.

To avoid basic conflicts, mechanisms in the communication interface between HMCs and the managed systems allow an HMC to temporarily take exclusive control of the interface, effectively locking out the other HMC. Usually this locking is done only for the duration of time it takes to complete an operation, after which the interface is available for further commands.

HMCs are also automatically notified of any changes that occur in the managed systems, so the results of commands issued by one HMC are visible in the other. For example, if you select to activate a partition from one HMC, you will observe the partition going to the Starting and Running states on both HMCs.

However, the locking between HMCs does *not* prevent users from running commands that might seem to be in conflict with each other. For example, if the user on one HMC selects to activate a partition, and a short time later a user on the other HMC selects to power the system off, the system will power off. Effectively, any sequence of commands that you can do from a single HMC is also permitted when it comes from redundant HMCs.

For this reason, it is important to carefully consider how you want to use this redundant capability to avoid such conflicts. You might choose to use them in a primary and backup role, even though the HMCs are not restricted in that way.

Because authorized users can be defined independently for each HMC, determine whether the users of one HMC should be authorized on the other. If so, the user authorization must be set up separately on each HMC.

Important: Because both the HMCs provide Service Focal Point and Service Agent functions, connect a modem and phone line to only one of the HMCs, and enable its Service Agent. To prevent redundant service calls, do not enable Service Agent on both HMCs.

Perform HMC software maintenance separately on each HMC, at separate times, so that there is no interruption in accessing HMC function. This allows one HMC to run at the new fix level, while the other HMC can continue to run at the previous fix level. However, the best practice is to move both HMCs to the same fix level as soon as possible.

HMC considerations

When attaching HMCs to a Cluster 1600, we recommend that you configure a separate Ethernet adapter for the “administrative LAN”. This LAN is limited to 10/100Mb/s. If multiple servers are to be controlled from a single HMC, either an 8-port asynchronous adapter (F/C 2943) or a 128-port asynchronous adapter (F/C 2944) is required.

The following list provides a summary of HMC requirements for a Cluster 1600:

- ▶ Machine type (M/T) 7040, 7039, and 7028 require a model 7315-C01 or later as the HMC.
- ▶ The HMC requires an RS-232 connection to each server.
- ▶ The RS-232 cable connects to the first HMC port on both the Hardware Management Console and the p655 or p630 servers.
- ▶ The second HMC port is used for redundant HMC configurations.
- ▶ Machine type 7039 also requires RS-422 connections to each Bulk Power Controller installed in the server’s M/T 7040-W42 frame.

Note: Machine type 7040, 7039, and 7028 servers do not require any RS-232 connections to the CWS.

Tip for the M/T 7039: When used in the HMC for a M/T 7039 cluster, the 8-port asynchronous adapter (F/C 2943) can be configured for simultaneous RS-232 and RS-422 communication.

- ▶ For smaller systems, this allows one adapter to communicate with the two BPC units in the frame (RS-422) and with six p655 servers (RS-232).
- ▶ For larger systems, a 128-port adapter may be required for server communications and the 8-port adapter would be dedicated to RS-422 frame communications.

For additional information, refer to the server-specific hardware documentation.

Tip for LPARs:

- ▶ For systems configured with an SP Switch2, the LAN adapters must be placed in I/O subsystem slot 8 using the same respective LPAR as the switch adapter. Place a second LAN adapter in slot 9.
- ▶ For systems configured with an SP Switch, the LAN adapter must be placed in the same respective LPAR as the switch adapter, but does not need to be in the same I/O subsystem.

Tip for Dynamic LPAR: An Ethernet connection between the HMC and each active partition on the pSeries server is recommended for all installations—and is *required* for dynamic LPAR configurations. This connection is used to provide:

- ▶ Additional systems management such as using Web-based System Manager for AIX in the individual partitions.
- ▶ Collection and passing of hardware service events to the HMC for automatic notification of error conditions to IBM.
- ▶ Total system inventory collection.

For full connectivity options, see Chapter 3, “Network configuration” on page 121.

HMC standard hardware

The IBM 7315-C02 HMC is a dedicated desktop workstation use for system and partition control of pSeries servers. It also provides a service focal point function for these servers. The HMC has the following fixed hardware attributes:

- ▶ Intel® Pentium®-based desktop workstation
- ▶ 1 GB of system memory
- ▶ 40 GB minimum hard disk
- ▶ DVD-RAM for backup
- ▶ Two integrated serial ports
- ▶ One graphics port
- ▶ One integrated Ethernet port
- ▶ Six USB ports
- ▶ Three PCI slots

Alternative rack mount HMC hardware

The 7315-CR2 is a rack-mounted HMC that mounts into a 19-inch rack.

The HMC has the following fixed hardware attributes:

- ▶ Intel XEON-based rack-mounted system unit
- ▶ 1U tall
- ▶ 1 GB of system memory
- ▶ 40 GB minimum hard disk
- ▶ DVD-RAM for backup
- ▶ One integrated serial port
- ▶ One graphics port
- ▶ Two integrated 10/100 Mbps Ethernet ports
- ▶ Three USB ports
- ▶ Two PCI slots

HMC hardware options

Table 2-10 details the optional feature codes available for order with, or for, an HMC.

Table 2-10 HMC hardware options

F/C 2943	8-port asynchronous adapter PCI BUS EIA-232/RS-422
F/C 2944	128-port asynchronous controller PCI bus
F/C 8120	Attachment cable HMC to host 6 meters (RS-232)
F/C 8121	Attachment cable HMC to host 15 meters (RS-232)
F/C 8122	Attachment cable HMC to W42 rack 6 meters (RS-422)
F/C 8123	Attachment cable HMC to W42 rack 15 meters (RS-422)
F/C 8137	2.4 MB/sec enhanced remote asynchronous node (RAN) 16-port EIA-232
F/C 8131	128-port asynchronous controller cable 4.5 m (1.2 MB/sec transfers)
F/C 8132	128-port asynchronous controller cable 23 cm (1.2 MB/sec transfers)
F/C 8133	RJ-45 to DB-25 converter cable
F/C 8136	1.2 MB/sec rack-mountable remote asynchronous node (RAN) 16-port EIA-232
F/C 2934	Asynchronous terminal/printer cable EIA-232 (2.4 MB/sec transfers)
F/C 3124	Serial port to serial port cable for drawer-to-drawer connections (2.4 MB/sec transfers)
F/C 3125	Serial port to serial port cable for rack-to-rack connections (2.4 MB/sec transfers)
F/C 4962	10/100 Mb/s Ethernet PCI adapter II

F/C 3628	IBM P260/275 Color monitor, Business Black and cable
F/C 3636	L200P Flat panel monitor

HMC standard software

The HMC is preloaded with dedicated Linux operating software.

Tip: HMC recovery software for pSeries (5639-N47) must be ordered in conjunction with every 7315-C02 system.

For more information on HMC and configuration, refer to *Effective System Management Using the IBM Hardware Management Console for pSeries*, SG24-7038.

2.6.3 Firmware

Firmware is a critical component in all modern systems; it is the system- or device-specific microcode that interfaces the hardware to the software or other pieces of hardware. It is critical that the firmware is kept up to date. Occasionally, new versions of system firmware and microcode for devices are introduced for the following reasons:

- ▶ New features, connectivity, security, and resource support are included to improve system operation.
- ▶ Serviceability is also an issue, and new levels may include improved problem determination and fault isolation, resulting in accurate error codes.
- ▶ Performance enhancements in the system: firmware may improve or resolve problems in response times and throughput.
- ▶ Finally, new levels of system firmware may improve the user interface with easier-to-understand messages.

For these reasons, it is important to know your current level and be able to check it against the latest available version. In this section, we explain how to check what your installed version of system firmware is, what the latest available version is, and how to install it. It also explains how to check your microcode for devices on some of your devices. Note that firmware and microcode for devices, as well as their installation procedures, are different from machine to machine and device to device. The installation instructions must be followed exactly.

It is now the customer's responsibility to keep the firmware at a current level. In order to help customers to easily install, update, and manage their own microcode (also called machine code or firmware) pSeries and RS/6000 systems

and associated I/O adapters and devices are using new software tools and existing services.

The following tools and existing services reiterate and support the Customer-Managed Microcode concept, which emphasizes customer responsibility for managing their own microcode updates:

- ▶ Microcode update application for AIX 5L V5.1 and V5.2
- ▶ Microcode discovery service Web site:
<http://techsupport.services.ibm.com/server/aix.invscountMDS>
- ▶ pSeries and RS/6000 microcode updates Web site
<http://techsupport.services.ibm.com/server/mdownload>
- ▶ Microcode update files and discovery tool V1.1 on CD-ROM

Note: All product microcode installation is also available with IBM service (installation may be fee-based).

The microcode update application, files, tools, and services make microcode available at your fingertips. Simple process steps help you analyze the system, acquire microcode, and install it without assistance from IBM.

These tools and services also support “secure” accounts that do not have immediate access to the Internet and customer systems without a diskette drive. They enable efficient microcode-level surveying for new code.

Two new tools and two existing services are available to assist you in managing and installing your microcode.

- ▶ Microcode update application for AIX 5L V5.1 and V5.2
- ▶ Microcode discovery service Web site at:
<http://techsupport.services.ibm.com/server/aix.invscountMDS>
- ▶ pSeries and RS/6000 microcode updates Web site at:
<http://techsupport.services.ibm.com/server/mdownload>
- ▶ Microcode update files and discovery tool V1.1 on CD-ROM

Microcode update application

The microcode update application is the primary microcode management tool for customers using AIX 5L V5.1 and V5.2. This application exists on standalone systems (those with no HMC attached) and HMC-controlled systems.

The microcode update application executes a survey process that lets you and IBM determine whether there is a microcode update for your system. This

process can now be done without IBM's intervention; there is no need to contact an IBM service support representative.

Microcode update application distribution

The distribution methods for microcode are now more flexible. The Microcode Update Application tool for AIX 5L V5.1 and V5.2, or later, allows you to get the latest microcode from three different repositories:

- ▶ If you have access to the Internet, the tool accesses the IBM microcode Web site:
<http://techsupport.services.ibm.com/server/mdownload>
- ▶ You can order the microcode update files and discovery tool V1.1 on CD, which lets you survey, report, and retrieve microcode without having to connect to the Internet.
- ▶ You can FTP to and from another machine or site to which you may have downloaded the latest level of microcode for temporary storage or test purposes.

Microcode discovery service Web site

You can determine if your pSeries or RS/6000 system is at the latest microcode level. Microcode discovery service gives you two ways to generate a real-time comparison report showing subsystems that may need to be updated. Before using this service, you will need to install a tool known as Inventory Scout onto the server you want to survey for microcode information.

Using a secure Internet connection and a signed, trusted Java™ applet running on your Internet-connected workstation, IBM can connect to Inventory Scout, running on your server, and capture hardware and microcode data. The workstation or PC running the browser must be able to connect to your server. If your server is disconnected from the Internet, you can still create a data file on that system and then upload the data file to IBM from an Internet-connected system.

pSeries and RS/6000 microcode updates Web site

The pSeries and RS/6000 microcode updates Web site provides all available product microcode. Use this page to keep your microcode current with the latest available microcode updates. You can view “what's new” to see a list of microcode releases for systems, adapters, and devices sorted by release date, with the most recent releases on top. You can select the microcode updates and download the file to upgrade your system. You can stay informed of future microcode updates by subscribing to the microcode update mailing list to receive the Hardware Microcode Bulletins at:

<http://techsupport.services.ibm.com/server/mdownload/download.html>

Microcode update files and discovery tool V1.1

The microcode update files and discovery tool V1.1 complements the microcode update application for AIX 5L V5.1 and V5.2 and the pSeries and RS/6000 microcode updates Web site.

The microcode update files and discovery tool V1.1 expands the availability of microcode onto a CD-ROM and provides simple process steps to analyze the system, acquire microcode, and install it without the need for an IBM Service Support Representative (SSR) or the Internet. It supports “secure” accounts that do not have access to the Internet and systems without a diskette drive. It enables efficient microcode-level surveying for new code.

With this CD-ROM, you can perform microcode updates when it is most convenient to you. This comprehensive source of the latest code eliminates having to access the Internet in multiple sessions in order to create a diskette of the latest code.

Microcode formats

The latest microcode levels are available in several formats for supported IBM systems and devices as of the date the CD was created (reference the date on the CD label). Use the microcode discovery tool to review installed microcode and download the latest microcode to your local disk. You can choose from one of these formats:

- ▶ A Java applet that runs in a Web browser GUI environment
- ▶ A script that runs in a command line AIX window text environment

Alternative microcode check

To quickly check the microcode level on your Cluster 1600 system, from AIX 4.3.3 ML08 onward, use the **lsmcode** command. This command, when run with no flags, displays the platform system firmware microcode level and the service.

Use the command **lsmcode -A** to display all system and device microcode on the system. Example 2-2 shows typical output from this command, run on a 7040-681 LPAR.

Example 2-2 lsmcode -A output

```
{node1:root}/-> lsmcode -A
sys0!system:RH021114      |System Firmware:RG021019_GA3_H|SPCN Firmware:0000RH
06182
fcs0!df1000f9.382101
ent0!1410ff01.SCU002
fcs1!df1000f9.382101
ent1!1410ff01.SCU002
ent2!1410ff01.SCU002
```

```
ent3!1410ff01.SCU002
hdisk0!ST33660.41543033.43353039
ses0!0011
hdisk1!ST33660.41543033.43353039
ses1!0011
hdisk2!ST33660.41543033.43353039
ses2!0011
hdisk3!ST33660.41543033.43353039
ses3!0011sys0!service:CL000628
```

2.6.4 Electronic Service Agent

Electronic Service Agent and Service Director are IBM software applications supplied with attached servers. Electronic Service Agent is a replacement for the previously-supplied Service Director. Both applications perform the same function and are referred to as service agent in this redbook.

Service Agent runs on AIX and also on the HMC. Service Agent monitors the “health” of your pSeries, RS/6000 and Cluster 1600 systems. When a system fault is detected, the severity of the fault is analyzed and, if required, Service Agent notifies the IBM support center.

In addition, you can also configure Service Agent to send an automated e-mail message containing the fault information to your system administrator. This requires mail to be active on each node. Upon receiving the fault notification, IBM automatically dispatches a service representative, along with parts if needed, to correct the fault.

Service Focal Point

Service Focal Point is an application, specific to HMC-controlled systems, that is installed by the customer engineer at installation, or shortly after, when the LPAR configuration has been established. The Service Focal Point works hand-in-hand with the Service Agent. The application runs on the HMC of POWER4 and later nodes to accommodate error reporting, analysis and repair in the LPAR environment.

Service Focal Point (SFP) leverages the design capabilities of the HMC to provide equivalent virtual function to the current capabilities presented by physical op panels, service processor TTY menu interfaces and system firmware interfaces, as well as the capability for configuring/reconfiguring building block hardware into partitions.

SFP is a system infrastructure which manages serviceable event information for the system building blocks. It consists of resource managers which monitor and

record information about different objects in the system. It filters and correlates events from the resource managers and initiates calls home, when appropriate. It also provides a user interface that allows a user to view the events and perform problem determination. When a problem is corrected, the user can record actions that have been taken to resolve the hardware problem. These features of SFP support the overall problem management strategy in a complex system.

The SFP application receives Service Action Events from the service processor for critical system down situations, and from the Service Agent client application programs running on the individual logical partitions for system recoverable or predictive events, as well as operating system- or device driver-detected events.

Service Action Event (SAE) log

The SFP collects the serviceable events from different building blocks and logs them in a Service Action Event (SAE) log. The log entries are generated by analysis routines that run on an error that has occurred in a building block. The resource manager for the building block forwards information about the event to the Service Focal Point and the information is placed in the SAE log. The particular content of the error data depends upon the type of the error and on the system configuration itself.

The SAE log on the SFP also contains pointers to extended information that may have been recorded at the time of a serviceable event by the building block. Extended error collection includes not only the collection of first failure data capture, but also vital product data, partition information, operating system error logs, service processor error logs, error register data, and so on.

When the SFP receives a new log entry, filtering is performed to determine if this was a unique event. The filtering is done because sometimes an event notification can come from more than one resource manager for the same event, or a resource manager may forward a notification for an event which previously occurred but has not been corrected yet.

Service Agent component

When a Service Action Event is logged in the SFP, the system needs to communicate the failure back to IBM. During this call home function, particular error data and system configuration information needs to be sent to IBM to drive the service delivery infrastructure.

The SFP utilizes the Service Agent Focal Point application residing on the HMC, along with the HMC modem, to initiate the call home and transfer the pertinent error information to IBM Service. When a call home is required, Service Agent manages the connection to IBM which is used to open a problem record. The problem record is used by the service delivery team to determine whether or not

to dispatch a customer engineer (CE) with the appropriate service parts to the system to perform a repair.

When service personnel perform a repair on the system, the SFP is used to identify the source of the problem and record information relating to the repair. When the repair has been completed, the service representative can update the SAE log with Field Replacement Unit (FRU) replacement information and any comments that the service representative has. The information stored by the SFP represents the system's service history and can be used to ensure proper maintenance over the life of the system.

Service Agent Gateway Server (SAS)

Where a customer has multiple pSeries servers or Cluster 1600s, a Service Agent Gateway Server can be set up. This will be the server in the organization which has a modem attached to it. Other servers that have Service Agent running on them will contact the Service Agent Gateway Server when they need to log a service call. The Service Agent Gateway Server will then dial IBM and log the service call for the client server.

The client Service Agent Servers must be registered with the Service Agent Gateway Server before calls can be logged. The client Service Agent Servers contact the Service Agent Gateway Server through the LAN, therefore only the Service Agent Gateway Server requires a modem and analog phone line.

2.6.5 Planning for Cluster 1600 servers

Planning for Cluster 1600 servers should be done with consultation with the following documentation:

- ▶ *IBM eServer Cluster 1600 Planning Volume 1, Hardware and Physical Environment, GA22-7280*
- ▶ *IBM eServer Cluster 1600 Planning Volume 2, Control Workstation and Software Environment, GA22-7281*
- ▶ *pSeries Systems Handbook 2003 Edition, SG24-5120*

In the following sections, we summarize the key points for planning for Cluster 1600 servers.

Network communication

Network communication is a very complex aspect of planning large clustered systems. Cluster 1600 systems have several communication requirements, including the following:

- ▶ All SP systems require an SP Ethernet LAN for system administration.

- ▶ Switch-configured systems require a frame-to-frame switch cable network.
- ▶ SP systems connected to external networks (or with networks between SP system partitions) require additional communication adapters.

The required SP Ethernet LAN that connects all nodes to the control workstation is needed for system administration—and it is to be used for that purpose exclusively. If you attempt to route non-administrative traffic over the SP Ethernet and it interferes with administrative traffic, you have to reroute the non-administrative traffic.

Further network connectivity is supplied by various adapters (some of which are optional) that provide connection to I/O devices, networks of workstations, and mainframe networks. Ethernet, FDDI, Token-Ring, HIPPI, SCSI, FCS, and ATM are examples of adapter types that can be used as part of an RS/6000 SP system.

Important: The establishment of a trusted network between the control workstation (CWS) and the Hardware Management Console (HMC) is recommended. For more information, refer to the PSSP V3.4 Read This First document, or to the PSSP V3.5 documentation at:

http://www.rs6000.ibm.com/resource/aix_resource/sp_books/

For more details on network configurations, see Chapter 3, “Network configuration” on page 121.

Security

Security should be planned very carefully because there are many options available to secure your system. Physical, network, login, and file access should be policy-driven on a “need to have” access basis.

Kerberos should be considered for authentication purposes and for PSSP daemon use. Additionally, ssh should be considered if encryption is also required. Restricted root access should also be considered for individual nodes administrators. For more information on security, refer to the following documentation:

- ▶ *Exploiting RS/6000 SP Security: Keeping It Safe*, SG24-5521
- ▶ *IBM @server Certification Study Guide: Cluster 1600 Managed by PSSP*, SG24-7013
- ▶ *Chapter 2, PSSP 3.5 Administration Guide*, SA22-7348
- ▶ *Chapter 6, IBM eServer Cluster 1600 Planning Volume 2, Control Workstation and Software Environment*, GA22-7281

Environmental considerations

When planning to install the servers, consideration should be given to the following items:

- ▶ Electrical requirements (some systems are 3-phase)
- ▶ Floor space (including engineering access space)
- ▶ Weight distribution (some systems are very heavy)
- ▶ Heat output (check air conditioning requirements)
- ▶ Delivery access (check if either height or weight reduction is required)

Server-specific documentation includes environmental requirements. You should also contact your IBM Customer Engineer or IBM Business Partner to help with installation planning.

2.7 Hardware supported and currently marketed

This section describes the Cluster 1600 hardware components, including servers, switches and control workstations, that are currently supported in the Cluster 1600 and available from IBM.

2.8 pSeries servers

In this section, we present the pSeries servers that are building blocks for a Cluster 1600 system; we describe each server and available features and options.

2.8.1 pSeries 615 server (7029-6C3 and 6E3 deskside)

The 7029 pSeries 615 Model 6E3 deskside server, shown in Figure 2-8, gives you the tools for managing e-business, greater application flexibility, and innovative technology, all designed to help you capitalize on the e-business revolution. Each model is available in 1-way or 2-way configurations.



Figure 2-8 The p615 6E3 (left) and 6C3 (right)

The symmetric multiprocessor (SMP) uses 1-way or 2-way, state of the art, 64-bit, copper-based, POWER4+ microprocessors running at 1.2 GHz. Each processor includes 8 MB of Level 3 (L3) cache. The base 1 GB of main memory can be expanded to 16 GB for faster performance and exploitation of 64-bit addressing, as used in large database applications.

The Model 6E3 contains up to 11 bays. There are four front-accessible, hot-swap-capable DASD bays in a minimum configuration with an additional four hot-swap-capable DASD bays optional.

The eight DASD bays can accommodate up to 1174.4 GB of disk storage. Two of the remaining three bays are used for a diskette drive and a DVD-ROM, and the third bay can contain a DVD-RAM or tape drive.

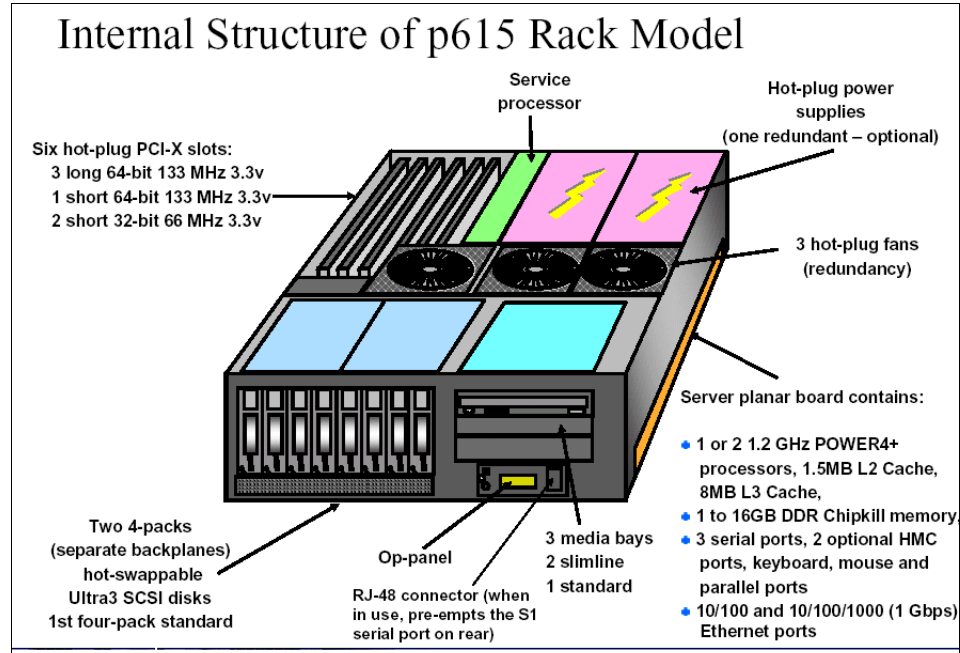


Figure 2-9 7029-6C3 internal structure

Note: Support for logical partitioning (LPAR) is not available on the Models 6E3 and 6C3.

Summary of standard and optional features

Table 2-11 lists the standard and optional features available for the pSeries 615 model 6C3 and model 6E3.

Table 2-11 The fp615 standard and optional features

Description	Quantity	Type
Processors	1 - 2	POWER+ 1.2GHz 8 MB L3 Cache (F/C 5133)
Memory	1 - 16GB	DDR SDRAM
Internal disk bays	8	Hot swap 18.2 GB - 146.8 GB per bay
Media bays	3	2 slimline and 1 full media bay
PCI slots	4 2	64-bit hot swap PCI-X (long) 32-bit hot swap PCI-X (short)

Description	Quantity	Type
Standard ports	1 1 3 1 2	Keyboard Mouse Serial Parallel HMC ports
Integrated SCSI adapters	2	Ultra 3 SCSI controllers (internal)
Integrated LAN adapter ports	2	1 x 10/100 Ethernet port and 1 x 10/100/1000 Ethernet port
I/O Drawers	N/A	N/A
Redundant power supply	Optional	Redundant AC power (F/C 6266)
Operating System version	AIX Linux	Version 5.1 with 5100-04 ML or later SLES 8 for pSeries
Physical specifications		Rack Mount: Width: 437 mm (17.2 in) Depth: 508 mm (20.0 in) Height: 178 mm (07.0 in) Weight: 35.5 kg (78.0 lb) - minimum configuration Weight: 43.1 kg (94.8 lb) - maximum configuration
Rack 7014-T00 Rack 7014-T42	9 10	Maximum is nine per rack Maximum is ten per rack Racks may be shared with peripherals
Power requirements		Operating voltage: 100 to 127 or 200 to 240 V ac (auto-ranging) Operating frequency: 47/63 Hz Power requirements: 300 watts (typical); 450 watts (maximum) Power source loading: 0.30 kVA (typical configuration) 0.50 kVA (maximum configuration)

Restriction: PCI-X slot 1 only allows short cards. 64-bit cards are not allowed in the 32-bit slots (slot 2 and 3).

Cluster considerations

The Cluster 1600 is enhanced to include the p615 server within the available cluster building blocks. This server is now available as both an initial option and a field addition to the Cluster 1600.

Restriction: The 615 is only supported in a CSM-managed Cluster 1600.

When ordering a p615 for attachment to a Cluster 1600, the features listed in Table 2-12 should be considered.

Table 2-12 Cluster features for the p615

Feature code	Description
7315-C02	Hardware Management Console (HMC)
7315 F/C 8120	Attachment cable HMC to host 6 meters
7315 F/C 8121	Attachment cable HMC to host 15 meters
9078-160 F/C 0012	Cluster 1600 7029 type server

Tip: An HMC must be available, or ordered with the p615. If an HMC is already available, then the console attachment cable must be ordered. An asynchronous card may also be required. See Table 2-10 on page 43 for the HMC options.

CSM/PSSP support statement

The following support statement details the CSM and PSSP support for the p615:

- ▶ The p615 is supported on CSM V1.3.1 with APAR IY42369 with AIX 5.1 ML04, AIX 5.2 ML01 or later.

Note: The CSM management server must be running AIX 5L V5.2 with Recommended Maintenance package 5200-01 (APAR IY39795). The other machines within the cluster are referred to as managed nodes and can be running AIX 5L V5.2 with 5200-01, or V5.1 with the 5100-04 Recommended Maintenance package (APAR IY39794). They can also be xSeries machines running CSM for Linux V1.3.

For all AIX servers, the following CSM for AIX 5L 1.3.1 service is required:

- ▶ Added Service for RSCT on AIX 5.1 IY42782
- ▶ Added Service for RSCT on AIX 5.2 IY42783
- ▶ CSM for AIX 5L added service IY42353
- ▶ CSM Support for p670 and p690 POWER4+ IY42356
- ▶ CSM Support for 7026 servers, 9076 SP nodes and p650 IY42379
- ▶ CSM 1.3.1 support for 9076 SP Node Features 2054/2058 IY42847
- ▶ CSM Support of p655 POWER4+ IY42377

- ▶ The p615 is *not* supported on PSSP 3.5.

For more information on the pSeries 615, refer to the specific server documentation or to *pSeries Systems Handbook 2003 Edition*, SG24-5120.

2.8.2 pSeries 630 server (7028-6C4 and 6E4 deskside)

The 7028 pSeries 630 Model 6C4, as shown in Figure 2-10 on page 58, is a rack-mounted server, and the 7028 IBM eServer pSeries 630 Model 6E4 is the deskside model as shown in the figure. The Model 6C4 provides the power, capacity, and expendability required for e-business computing. It offers 64-bit scalability via the 64-bit POWER4 or POWER4+ processor packaged as 1-way and 2-way cards.



Figure 2-10 The p630 6C4 (left) and 6E4 (right) picture

With its two-processor card positions, the model 6C4 can be configured into 1-way, 2-way, or 4-way configurations. The processor cards operate at 1.0 GHz with 32 MB of L3 cache per processor card, or 1.2 and 1.45 GHz with 8 MB of L3 cache per processor. The pSeries 630 Model 6C4 memory DIMMs are mounted on the CPU card and can contain up to 32 GB of memory. The p630's processors are packaged on Single Chip Modules (SCM). There is one POWER4 or POWER4+ chip on the module with either one or two CPU cores within the chip. The model 6C4 contains six bays. The four front-accessible, hot-swap capable bays can accommodate up to 587.2 GB of disk storage.

Internal Structure of Rack Model

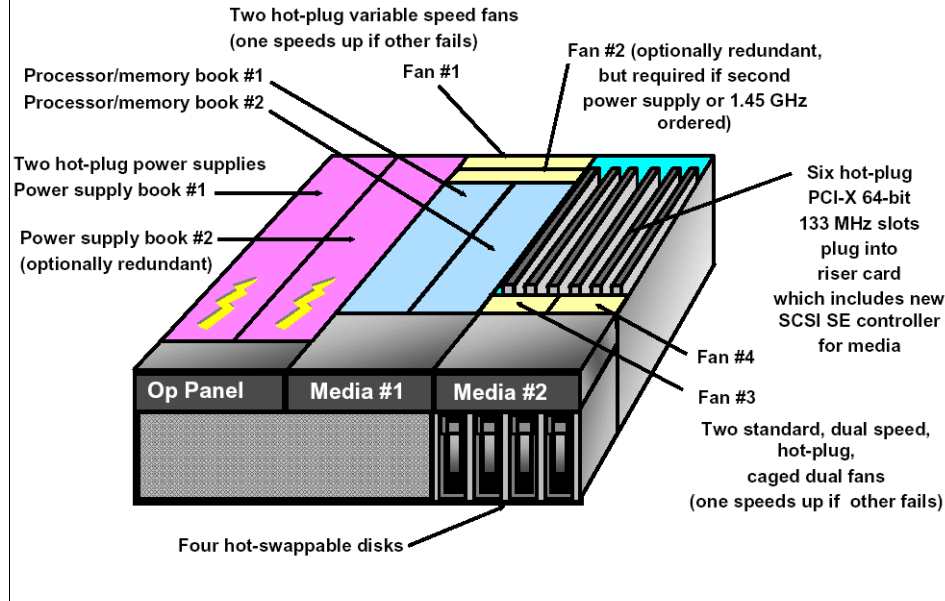


Figure 2-11 7028-6C4 internal structure

LPAR considerations

Support for logical partitioning (LPAR) is available on the models 6E4 and 6C4. LPAR enables you to divide a Model 6E4 system into two LPARs and the Model 6C4 into as many as four LPARs. Each LPAR functions as a “computer within a computer” with its own instance of the operating system.

For more information on LPARs on pSeries refer to the IBM Redbook *The Complete Partitioning Guide for IBM eServer pSeries Servers*, SG24-7039.

p630 I/O drawers

The IBM eServer 630-6C4 can attach two 7311-D20 I/O drawers to increase the number of I/O adapters that can be inserted into the system. I/O drawers are only supported on the rack mounted model of p630.

7311-D20 I/O drawer

The IBM 7311 Model D20 is a rack-mounted, high density expansion drawer that attaches to the pSeries 630 Model 6C4 to provide remote I/O. There are seven hot-swap PCI-X I/O slots in the 7311-D20 and twelve optional hot-swap disk drive bays.

The 7311-D20 occupies 4U (7.00-inches) of space in a rack and mounts in a 19-inch standard rack drawer. The Model D20 is 24-inches deep. The fans, power supplies, and PCI adapters, are top-accessible while the disk drives are front-accessible for easy service and maintenance. Figure 2-12 shows the internal structure of a 7311-D20.

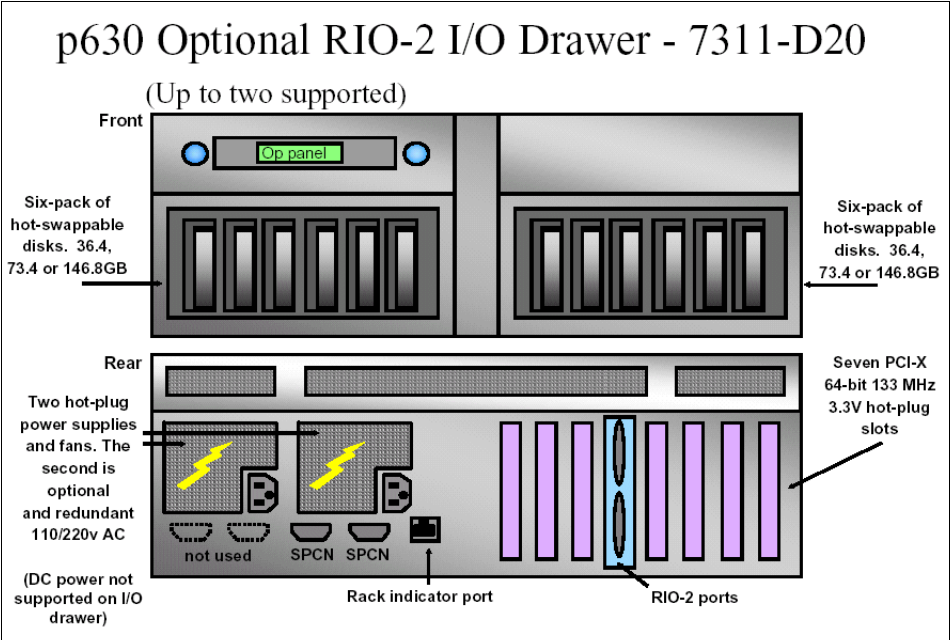


Figure 2-12 7311-D20 I/O drawer (p630)

Tip: If the disk drives in the Model D20 are being used to support two LPARs, then two SCSI adapters must be installed in the Model D20; one adapter is needed to support each 6-pack of DASD.

Summary of standard and optional features

Table 2-13 shows the standard and optional features available for the pSeries 630 model 6C4 and 6E4.

Table 2-13 Standard and optional features

Description	Quantity	Type
Processors	1 - 4 1 - 4 1 - 4	POWER4 1.0GHz, 32MB L3 Cache (F/C 5131) POWER4+ 1.2GHz 8MB L3 Cache (F/C 5133) POWER4+ 1.45GHz, 8MB Cache (F/C 5126)
Memory	1 - 32GB	DDR SDRAM
Internal disk bays	4	Hot swap 18.2GB - 146.8GB per bay
Media bays	2	Bay 1 for CD or DVD Bay 2 for CD, DVD, diskette or tape (optional)
PCI slots	4 6	64 bit hot swap PCI-X with 1.0GHz proc 64 bit hot swap PCI-X with 1.2 and 1.4GHz
Standard ports	1 1 3 1 2 2	Keyboard Mouse Serial Parallel HMC ports Model 6C4 only RIO ports, Rio-2 ports with F/C 9581
Integrated SCSI adapters	2	Ultra 3 SCSI controllers (1 int & 1 ext VHDC I AMP half pitch 68-pin)
Integrated LAN adapter ports	2	10/100 Ethernet controller with 2 ports
I/O Drawers	Optional	2 x 7311-D20 can be attached to model 6C4
Redundant power supply	Optional	Redundant AC power (F/C 6273)
Operating System Version	AIX	Version 5.1 with 5100-02 ML or later
Physical specifications		Width: 444.4mm (17.5in) Depth: 609.6mm (24.0in) Height: 172.8mm (06.8in)(4 EIA units) weight: 32.0 kg (70.4 lb) - Minimum Configuration 47.3 kg (104 lb) - Maximum Configuration with rails
Rack 7014-T00 Rack 7014-T42	9 10	Maximum is nine per rack Maximum is ten per rack Racks may be shared with peripherals

Description	Quantity	Type
Power requirements		Operating voltage: (auto-ranging) 100 to 127 or 200 to 240 V ac Operating frequency: 50/60 Hz Power requirements: 2-way (typical configuration): 330 watts 2-way (maximum configuration): 500 watts 4-way (typical configuration): 500 watts 4-way (maximum configuration): 750 watts Power source loading: 2-way (typical configuration): 0.348 kVA 2-way (maximum configuration): 0.522 kVA 4-way (typical configuration): 0.522 kVA 4-way (maximum configuration): 0.783 kVA Maximum altitude: 2,135 m (7,000 ft)

Cluster considerations

The Cluster 1600 is enhanced to include the p630 server within the available cluster building blocks. This server is now available as both an initial option and a field addition to the Cluster 1600.

When ordering a p630 for attachment to a Cluster 1600, the features in Table 2-14 should be considered.

Table 2-14 Cluster features for p630

Feature code	Description
7315-C02	Hardware Management Console (HMC)
7315 F/C 8120	Attachment cable HMC to host 6 meters
7315 F/C 8121	Attachment cable HMC to host 15 meters
7028 F/C 8398	SP Switch2 PCI-X attachment adapter
7028 F/C 8397	SP Switch2 PCI attachment adapter
9078-160 F/C 0012	Cluster 1600 7028 type server
9078-160 F/C 0013	Cluster 1600 7018 LPAR (switched)

Tip: An HMC must be available, or ordered with the p630. If an HMC is already available, then the console attachment cable must be ordered. An asynchronous card may also be required. See Table 2-10 on page 43 for the HMC options.

CSM/PSSP support statement

The following support statement details the CSM and PSSP support for the p630:

- ▶ The IBM @server p630 is supported on CSM V1.3.1 with APAR IY42369 with AIX 5.1 ML04, 5.2 ML01 or later.

Note: The CSM management server must be running AIX 5L V5.2 with Recommended Maintenance package 5200-01 (APAR IY39795). The other machines within the cluster are referred to as managed nodes and can be running AIX 5L V5.2 with 5200-01, or V5.1 with the 5100-04 Recommended Maintenance package (APAR IY39794). They can also be xSeries machines running CSM for Linux V1.3.

For all AIX servers, the following CSM for AIX 5L 1.3.1 service is required:

- ▶ Added Service for RSCT on AIX 5.1 IY42782
- ▶ Added Service for RSCT on AIX 5.2 IY42783
- ▶ CSM for AIX 5L added service IY42353
- ▶ CSM Support for p670 & p690 POWER4+ IY42356
- ▶ CSM Support for 7026 servers, 9076 SP nodes and p650 IY42379
- ▶ CSM 1.3.1 support for 9076 SP Node Features 2054/2058 IY42847
- ▶ CSM Support of p655 POWER4+ IY42377

- ▶ The p630-6C4 is supported on PSSP 3.5 with AIX 5.1 ML03 or later.

For more information on the pSeries 630 and features, refer to the specific server documentation or to *pSeries Systems Handbook 2003 Edition*, SG24-5120.

2.8.3 pSeries 650 server (7038-6M2)

The pSeries 650 model 6M2, as shown in Figure 2-13 on page 64, consists of a single rack-mounted drawer containing the processors, memory, disk drives, and hot-plug I/O slots. The p650 server and optional attached 7311-D10 or D20 I/O drawers implement redundant power and redundant cooling.



Figure 2-13 The p650 6M2

The p650 system can accommodate one, two, three, or four 2-way processor cards to create configurations of 2, 4, 6, or 8 processors as required, offering outstanding scalability as your business needs dictate.

The p650 rack-mounted drawer configuration offers flexibility regarding the number of server and I/O drawers that can be mounted in the rack, providing more compute and I/O power per square foot of valuable floor space. The p650 requires 8U (EIA Units) of rack space and up to eight 7311-D10 I/O drawers can be attached, requiring 4U of rack space for each pair of I/O drawers. A minimum p650 configuration requires only 8U of rack space, while a system configured with eight 7311-D10 drawers fits into a 24U space.

(The maximum configuration, which consists of one p650 drawer and eight 7311-D20 I/O drawers, requires 40U of rack space and provides 63 available hot-plug, PCI-X slots with up to eight processors and 64 GB of memory.) Depending on the number of I/O drawers attached, up to four p650 systems can be installed in a 7014-T00 rack, with room remaining to install external data storage drawers.

Figure 2-14 on page 65 shows the internal structure of pSeries 650.

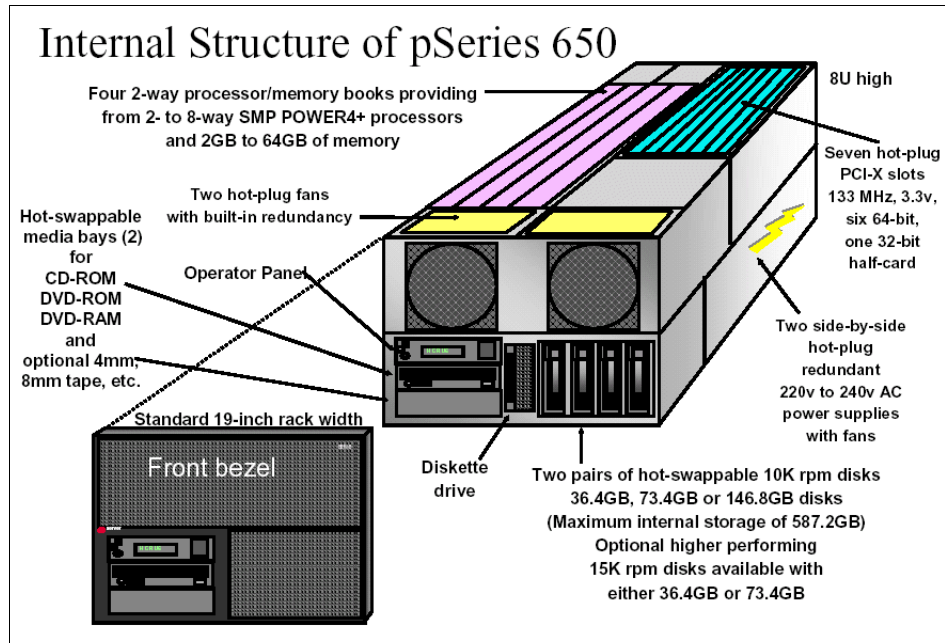


Figure 2-14 7038-6M2 internal structure

Table 2-15 on page 67 shows the standard and optional features available for the pSeries 650 model 6M2.

p650 I/O drawers

The 650 has a choice of I/O drawers available to it: the 7311-D10, and the 7311-D20. The primary difference is that the D20 contains internal SCSI disks which can be used to boot LPARs from, and the D10 only contains adapter slots. In the following sections, we discuss each drawer in more detail.

7311-D10 I/O drawer

The IBM 7311 model D10 I/O drawer is a rack-mountable expansion cabinet that can be attached to a pSeries 650 server. Each model D10 drawer gives you six full-length adapter slots. Up to eight model D10 drawers are supported by the pSeries 650, with connections provided by remote I/O adapters and cables.

The model D10 requires 4U of vertical space in a 19-inch rack, such as the IBM 7014-T00 or 7014-T42. Two D10 drawers can fit side by side within the enclosure that provides rack mounting hardware.

The model D10 offers a modular growth path for pSeries 650 systems with increasing I/O requirements. When a pSeries 650 is fully configured with eight

attached model D10 drawers, the combined system supports up to 55 PCI adapters. Figure 2-15 shows the structure of a 7311-D10 I/O drawer.

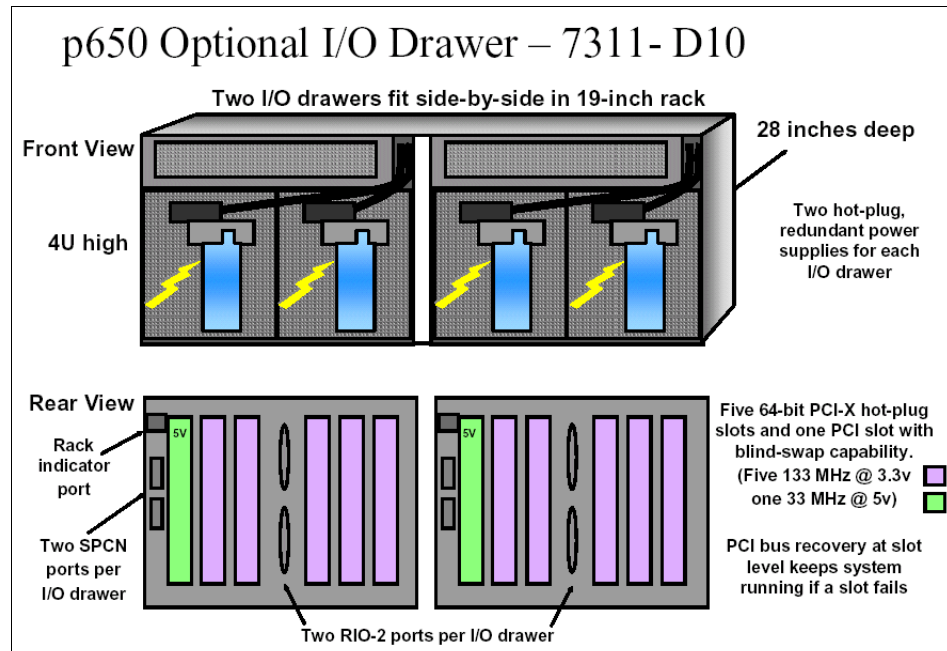


Figure 2-15 7311-D10 internal structure

7311-D20 I/O drawer

The IBM 7311 Model D20 is a rack-mounted, high-density expansion drawer that attaches to the pSeries 630 Model 6C4 to provide remote I/O. There are seven hot-swap PCI-X I/O slots in the 7311-D20 and twelve optional hot-swap disk drive bays.

The 7311-D20 occupies 4U (7.00-inches) of space in a rack and mounts in a 19-inch standard rack drawer. The model D20 is 24 inches deep. The fans, power supplies, and PCI adapters are top-accessible, and the disk drives are front-accessible for easy service and maintenance. Figure 2-16 on page 67 shows the internal structure of a 7311-D20.

p650 Optional I/O Drawer - 7311-D20

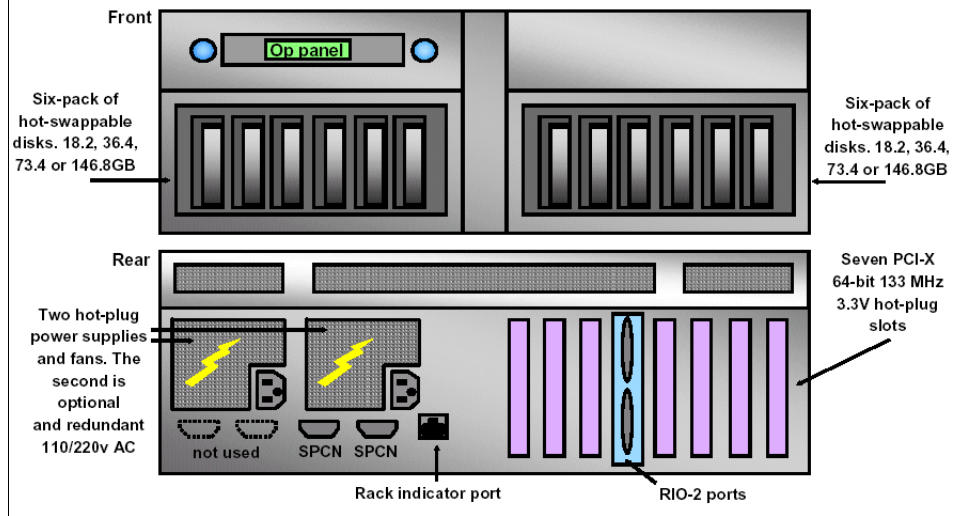


Figure 2-16 7311-D20 internal structure (p650)

Tip: If the disk drives in the model D20 are being used to support two LPARs, then two SCSI adapters must be installed in the model D20 (one to support each 6-pack of DASD).

Summary of standard and optional features

Table 2-15 p650 standard and optional features

Description	Quantity	Type
Processors	2, 4, 6, 8 2, 4, 6, 8	POWER4+ 1.2GHz 8 MB L3 Cache (F/C 5122) POWER4+ 1.45GHz, 32 MB Cache (F/C 5208)
Memory	1 - 64GB	DDR SDRAM
Internal disk bays	4	Hot swap 18.2 GB - 146.8 GB per bay
Media bays	2	Auto docking bays for CD, DVD, diskette or tape (optional)
PCI slots	6 1	64-bit hot swap 133MHz PCI-X full length 32-bit hot swap 66MHz PCI-X half length

Description	Quantity	Type
Standard ports	1 1 4 1 2 1 2 - 8	Keyboard Mouse Serial Parallel HMC ports Status beacon port RIO ports or RIO-2 ports. Number of ports is dependent on proc cards and GX RIO cards
Integrated SCSI adapters	2	Ultra 3 SCSI controllers (1 int & 1 ext 68-pin)
Integrated LAN adapter ports	1	10/100 Ethernet controller
I/O drawers	Optional	8 x 7311-D10 or D20 drawers can be attached
Redundant power supply	Standard	AC power supplies (F/C 9172)
Operating system version	AIX	Version 5.1 with 5100-02 ML or later
Physical specifications		Width: 444.4mm (17.5 in) Depth: 770mm (29.9 in) Height: 351mm (13.8 in) (8 EIA units) Weight: 93.0 kg (205.4 lb) - maximum configuration
Rack 7014-T00 Rack 7014-T42 see following notes	4 5	Maximum is four per rack Maximum is five per rack, 5 PDUs required Racks may be shared with peripherals
Power requirements		Operating voltage: 200 to 240 V ac 50/60 Hz Electrical output: 1,070 watts (typical); 1,600 watts (maximum) Power source loading: 1.126 kVA (typical configuration) 1.684 kVA (maximum configuration) Thermal Output: 1,070 joules/sec (3,652 Btu/hr, typical configuration) 1,600 joules/sec (5,461 Btu/hr, maximum configuration)

Note: An IBM 7014-T00 or T42 rack must have at least one Power Distribution Unit (PDU) per p650 system installed. It is recommended that each power supply in a p650 system be connected to a different PDU. No more than two p650 power supplies may be connected to the same PDU.

Each PDU has nine C13 power connectors, in groups of three. When a p650 power supply is connected to one PDU power outlet, the other two connectors in that group of three m

Important: The 7014 PDUs must be one of the following feature codes: F/C 7176, F/C 7177, F/C 7178, F/C 9176, F/C 9177, or F/C 9178. If you are using an existing rack, then an upgrade to the rack may be required.

Capacity on Demand (CuOD)

Capacity on Demand (CuOD) is a way of planning for future growth within your server. It means that you can have extra resources available which can be utilized at short notice and with no disruption to service. Payment for the extra resources is only made when the resources are used.

Capacity on Demand is available for both CPU resource and memory resource. It also has the added advantage that if one of the active CPUs fails, it can be substituted automatically and without disruption for one of the inactive CPUs. This action is non-chargeable.

The pSeries 650 systems can be shipped with non-activated resources (processors and/or memory), which may be purchased and activated at a certain point in time without affecting normal machine operation.

The following methods are available:

- ▶ Capacity Upgrade on Demand
- ▶ Memory Capacity Upgrade on Demand
- ▶ On/off Capacity on Demand
- ▶ Trial Capacity Upgrade on Demand

For further information about Capacity on Demand, refer to specific server documentation or to *pSeries 650 Model 6M2 Technical Overview and Introduction*, REDP0194.

LPAR considerations

Support of logical partitioning (LPAR) is available on the pSeries 650. LPAR enables you to split a Model 6M2 system into eight LPARs. Each LPAR functions as a “computer within a computer”, with its own instance of the operating system. Note the following points:

- ▶ Static LPAR is supported with AIX V5.1 or V5.2. Dynamic LPAR requires AIX V5.2.
- ▶ A 7315-C01 Hardware Management Console is required for all LPAR configurations.

- ▶ A fully configured pSeries 650 can support up to eight partitions. A partition must have at least a bootable SCSI or SSA or Fibre Channel adapter, and an Ethernet adapter/controller. Since all the PCI slots in p650 can be independently assigned to partitions, this means that four partitions with minimal I/O can be configured without requiring a separate 7311-D10 I/O drawer.

A model D10 I/O drawer can provide up to three additional minimal I/O partitions. To configure the maximum of eight partitions would require at least two model D10 I/O drawers. The actual number of D10 I/O drawers required depends on the customer's I/O requirements for each partition.

Note: Both the integrated internal SCSI bus and the integrated external SCSI bus must be assigned to the same partition.

- ▶ When the system is configured with the (F/C 6578) Ultra3 SCSI Backplane, all of the internal disk drives and media share the same SCSI bus and must be assigned to the same partition. That means that all but the first partition must be NIM-installed and use network diagnostics, or can only be CD-ROM-installed or use network diagnostics when the first partition is shut down.

The user should not use the internal disk drives during install because drives could have data for the first partition and may be reassigned back to the first partition. We recommend that you use NIM and network diagnostics for all partitions except the first partition—or that internal disk drives not be used for any partition.

- ▶ Two independent pairs of hot-swappable drives are optionally available by using the (F/C 6579) Split SCSI Backplane. This configuration is available for systems that require two logical partitions without using any I/O drawers or external boot devices. When configured using a single SCSI card and the external SCSI port, the internal media bays will be grouped with the two drives controlled by the external SCSI port.

Tip: To free the media devices for assignment to LPARs, it is recommended to use two SCSI adapters to support the split SCSI backplane. In this configuration, the media devices will be attached to the internal SCSI controller and be free to be assigned to any partition, and not affect a configuration using the internal disks.

- ▶ When the system is configured with the (F/C 6578) Ultra3 SCSI Backplane, external disk drives must be provided for all but the first partition. When the system is configured with the (F/C 6579) Split SCSI Backplane, external disk

drives must be provided for all but the first two partitions. AIX does not support diskless LPAR operation.

For more information on LPARs on pSeries, refer to *The Complete Partitioning Guide for IBM eServer pSeries Servers*, SG24-7039.

Note When running LPAR with AIX V5.1, fully populated processor cards cannot mix memory DIMMs with different capacities on the same card.

Cluster considerations

The model 6M2 is supported in either a non-switched IBM Cluster 1600 or a switched Cluster 1600 system using the SP Switch2 adapter (F/C 8398). Up to 32 p650s running in non-LPAR (full system partition) mode are supported in a cluster. A Cluster 1600 can scale up to 128 LPARs. PSSP Version 3.5 is required to support SP Switch2 adapter (F/C 8398) with AIX 5L Version 5.2 or Version 5.1.

If a model 6M2 configured in LPAR mode is part of the cluster, all LPARs must be part of the cluster. It is not possible to use selected LPARs as part of the cluster and use others for non-cluster use.

The HMC uses a dedicated connection to the model 6M2 to provide the functions needed to control the server, such as powering the system on and off. The HMC must have an Ethernet connection to the CWS. Each LPAR in the model 6M2 must have an Ethernet adapter to connect to the CWS trusted LAN.

When ordering a p650 for attachment to a Cluster 1600, the features in Table 2-16 should be considered.

Table 2-16 Cluster features for p650

Feature code	Description
7315-C02	Hardware Management Console (HMC)
7315 F/C 8120	Attachment cable HMC-to-host 6 meters
7315 F/C 8121	Attachment cable HMC-to-host 15 meters
7038 F/C 8398	SP switch2 PCI-X attachment adapter
9078-160 F/C 0014	Cluster 1600 7038-type server
9078-160 F/C 0015	Cluster 1600 7038 LPAR (switched)
PSSP	<p>The 7038-6M2 with F/C 8398 requires PSSP V3.5 with the following APARs:</p> <ul style="list-style-type: none"> – IY42352 PSSP V3.5 support for p650 with SP Switch2 PCI-X in RIO mode <p>The 7038-6M2 with F/C 8398 requires PSSP V3.4 or PSSP V3.5 with the following APARs:</p> <ul style="list-style-type: none"> – IY42359 PSSP V3.4 support for p650 with SP Switch2 PCI-X in RIO-2 mode – Y42358 PSSP V3.5 support for p650 with SP Switch2 PCI-X in RIO-2 mode

Tip: An HMC must be available, or ordered with the p650. If an HMC is already available, then the console attachment cable must be ordered. An Asynchronous card may also be required. See Table 2-10 on page 43 for the HMC options.

CSM/PSSP support statement

The following support statement details the CSM and PSSP support for the p650:

- The IBM @server p650 is supported on CSM V1.3.1 with APAR IY42369 with AIX 5.1 ML04, 5.2 ML01 or later.

Note: The CSM management server must be running AIX 5L V5.2 with Recommended Maintenance package 5200-01 (APAR IY39795). The other machines within the cluster are referred to as managed nodes and can be running AIX 5L V5.2 with 5200-01, or V5.1 with the 5100-04 Recommended Maintenance package (APAR IY39794). They can also be xSeries machines running CSM for Linux V1.3.

For all AIX servers, the following CSM for AIX 5L 1.3.1 service is required:

- ▶ Added Service for RSCT on AIX 5.1 IY42782
- ▶ Added Service for RSCT on AIX 5.2 IY42783
- ▶ CSM for AIX 5L added service IY42353
- ▶ CSM Support for p670 and p690 POWER4+ IY42356
- ▶ CSM Support for 7026 servers, 9076 SP nodes and p650 IY42379
- ▶ CSM 1.3.1 support for 9076 SP Node Features 2054/2058 IY42847
- ▶ CSM Support of p655 POWER4+ IY42377

- ▶ The IBM @server p650-6M2 is supported on PSSP 3.5 with AIX 5.1 ML03 or later.

For more information on the pSeries 650 and its features, refer to the specific server documentation manual or to *pSeries Systems Handbook 2003 Edition*, SG24-5120.

2.8.4 pSeries 655 server (7039-651)

The pSeries 655 (7039-651), as shown in Figure 2-17 on page 74, is the latest in a family of products within IBM's UNIX, 64-bit, symmetric multiprocessing (SMP) servers. The pSeries 655 offers a super-dense package (up to 128 processors in a single frame), ideal for many high-performance computing and business intelligence applications. Based on IBM's leadership POWER4 Technology, the p655 is designed for medium to large customers whose workloads are best managed with a clustered server solution.

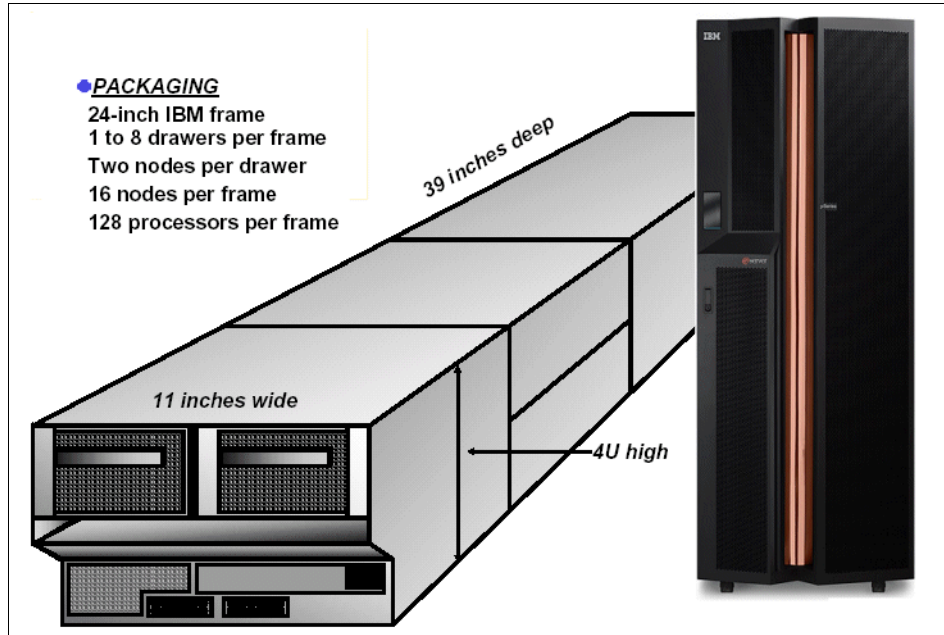


Figure 2-17 The p655

Packaged in scalable, single multi-chip module (MCM)-based *thin* nodes, within a half-drawer, 4U form factor, up to sixteen p655s can be installed within a 7040-W42 frame. The p655 server can be interconnected to the SP Switch2 fabric using the SP Switch2 PCI-X adapter as a standalone cluster, or as an additional building block of the Cluster 1600. The p655 can be partitioned into four logical partitions (LPARs), either switched or non-switched, for greater flexibility.

Note: The p655 must be mounted in a 7040-W42 24-inch rack.

Within the 7040-W42 frame for the p655, the 7040-61D I/O drawer offers expansion of up to 20 blind-swap PCI bus slots, and up to 16 additional Ultra3 SCSI disks per server that can be hot-plugged.

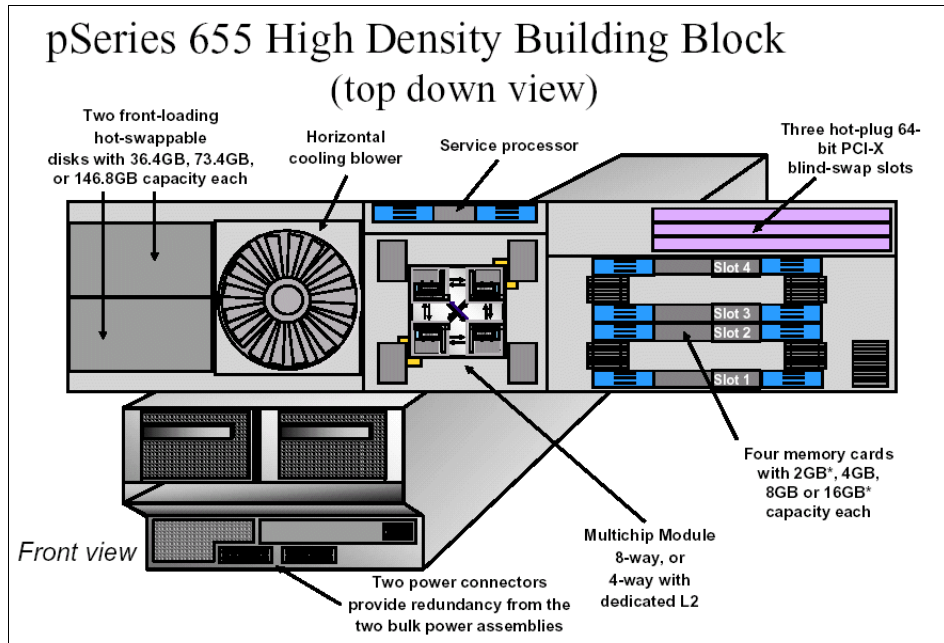


Figure 2-18 7039-651 internal structure

The ultra-powerful pSeries 655 sets the standard for supercomputing class UNIX servers with its mainframe-heritage LPAR and self-managing features. The p655 is the latest system to offer the POWER4+ microprocessor. The POWER4+ chip is a dual processor SMP on a single piece of silicon. It owes its advanced performance to the IBM SOI and copper fabrication technology.

The pSeries 655 can be configured with a 7040-61D I/O Drawer. The I/O drawer can contain up to 16 hot-swap-capable disks structured into four 4-packs of 18.2 GB, 36.4 GB, 73.4 GB, or 146.8 GB disks.

The drawer contains two planars with 10 slots/planar, 20 slots/drawer. The slots are hot-plug-capable to permit PCI adapters to be added or replaced without extending the I/O drawer, all while the system remains available to the customer. PCI-X planars in the 7040-61D I/O Drawer, with RIO-2 connectivity, are supported on the POWER4+ version of the 7039-651.

7040-61D I/O drawer

The I/O drawers provide internal storage and I/O connectivity to the system. Figure 2-19 on page 76 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.

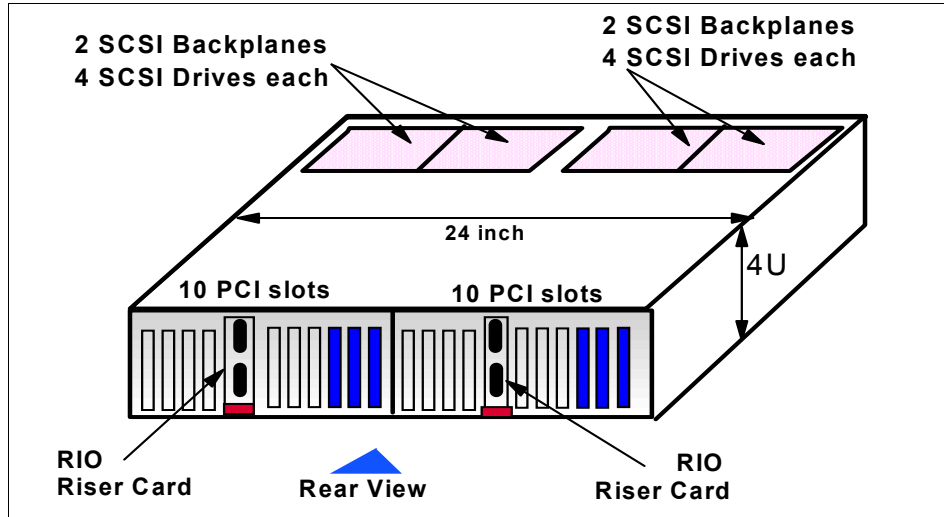


Figure 2-19 I/O drawer rear view

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.

Tip: Both planars within the 7040-61D I/O drawer can be connected to the same p655 CEC, or each individual planar can be attached to a separate p655 CEC.

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 2-20 on page 77).

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

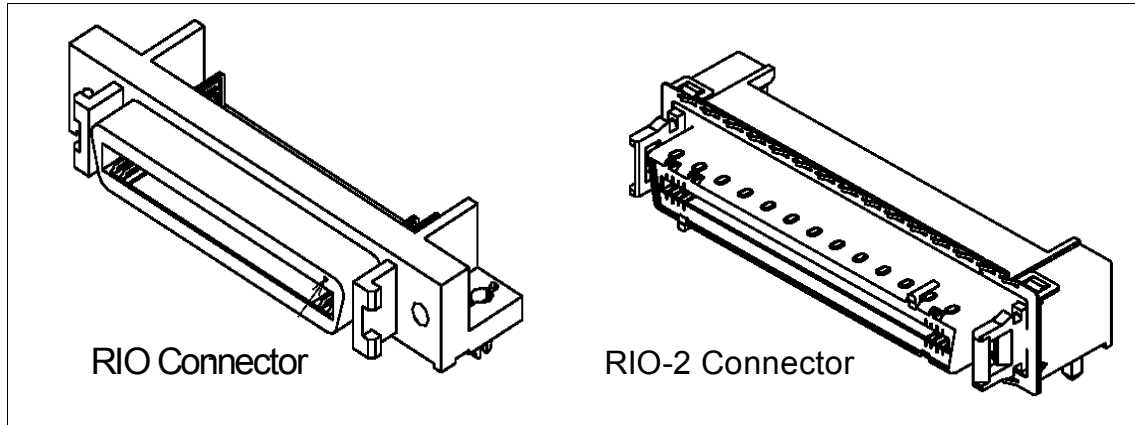


Figure 2-20 Difference between RIO and RIO-2 connectors

The I/O Drawer for pSeries 670, pSeries 655 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 (F/C) of the configured I/O planar:

- ▶ RIO drawer uses F/C 6563, I/O Drawer PCI Planar, 10 slots, 2 Integrated Ultra3 SCSI Ports.
- ▶ RIO-2 drawer is configured with F/C 6571, I/O Drawer PCI-X Planar, 10 slots, 2 Integrated Ultra3 SCSI Ports.

Summary of standard and optional features

Table 2-17 shows the standard and optional features available for the pSeries 655 model 651.

Table 2-17 p655 standard and optional features

Description	Quantity	Type
Processors	8 4 8 4	POWER4 1.1GHz 128 MB L3 Cache (F/C 5511) POWER4 1.3GHz 128 MB L3 Cache (F/C 5513) POWER4+ 1.5GHz, 128 MB Cache (F/C 5515) POWER4+ 1.7GHz, 128 MB Cache (F/C 5518)
Memory	4 - 32GB	DDR SDRAM (64 GB with RPQ)
Internal disk bays	2	18.2GB - 146.8 GB per bay

Description	Quantity	Type
Media bays	0	N/A
PCI slots	3	64-bit hot swap PCI-X
Standard ports	2 2	HMC ports RIO ports, RIO-2 ports feature-dependent
Integrated SCSI adapters	1	Internal for internal disks
Integrated LAN adapter ports	2	10/100 Ethernet controller with 2 ports
I/O drawers	Optional	1 x 7040-61D I/O drawer can be attached
Redundant power supply	Standard	Redundant AC power within the rack
Operating System Version	AIX	Version 5.1 with 5100-02 ML or later
Physical specifications: Rack		<ul style="list-style-type: none"> ▶ Width: 78.5 cm (30.9 in) (with acoustic doors) 78.5 cm (30.9 in) (with slim-line doors) ▶ Depth: 179.9 cm (70.8 in) (with acoustic doors) 144.3 cm (56.8 in) (with slim-line doors) ▶ Height: 202.5 cm (79.7 in) (with acoustic doors) 202.5 cm (79.7 in) (with slim-line doors) ▶ Weight: 1652 kgs (3,635 lb maximum) (with acoustic doors) 1,642 kgs (3,613 lb maximum) (with slim-line doors)
Rack 7040-W42	16	Maximum is sixteen per rack
Power requirements		<p>Operating voltage @ 50/60 Hz: (3-phase): 200 to 240 V ac 380 to 415 V ac 480 V ac</p> <p>Electrical output: 31,818 watts (maximum, non-redundant line cords). Power source loading: 32.1 kVA (maximum non-redundant line cord configuration). Thermal output: 30,012 joules/sec (102,000 Btu/hr) (maximum configuration, non-redundant line cords)</p>

LPAR considerations

The pSeries 655 can be partitioned into a maximum of four LPARs, switched or non-switched. Each logical partition operates under the control of its own copy of the operating system. PCI-X/PCI cards may be assigned to an LPAR on a slot-by-slot basis, if required. The integrated Ethernet ports can be individually assigned to an LPAR. Internal disk can be assigned to only one LPAR.

To achieve the minimum resources per LPAR, additional I/O expansion for disk is required. Information on resources required to enable LPAR can be found in *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590.

Racking considerations

The IBM 7040-W42 system rack provides space for mounting system drawer components for pSeries 655 servers and 7040-61D I/O drawers. The Model W42 is a 24-inch rack that provides 42U of rack space.

The 7040-W42 utilizes a 350 V DC bulk power subsystem to provide power for the components within the rack. The bulk power subsystem incorporates redundant bulk power assemblies mounted in the front and rear sections of the top 8U of the rack.

Tip: The rack must have RS-422 connections from the HMC to each of the bulk power supply controllers. These cables require an asynchronous card to be present in the HMC. If the 128 port card (F/C 2944) is used, then RS-232 - RS-422 connectors are also required.

The racking considerations for the pSeries 655 can be quite complex, especially when planning for future expansion. When a p655 is configured using the IBM econfig tool, the rack position is specified with a feature code that designates the EIA position within the rack. The power cable is also specified, which designates either the right or left position within the EIA position. Figure 2-21 on page 80 shows the physical location code, EIA units, feature codes and filling order of the 7040-W42 rack.

The eConfig tool tries to optimize your rack placements and will place them by default. The p655 should be placed by power cable feature (38xx), and the 7040-61D I/O drawers should be placed by EIA feature (46xx). The 7040-61D I/O drawer can go in any position, but should be placed nearest to the server which it is attached to. RIO and RIO2 cables should be ordered with the I/O drawer and not the p655. Cables ordered against the CEC are not validated. For more information on racking rules, see Appendix D of *pSeries 655 Installation Guide*, SA38-0616.

Note: On multiple rack configurations, rack-to-rack cables are not supported between the p655 CEC and the 7040-61D I/O drawer.

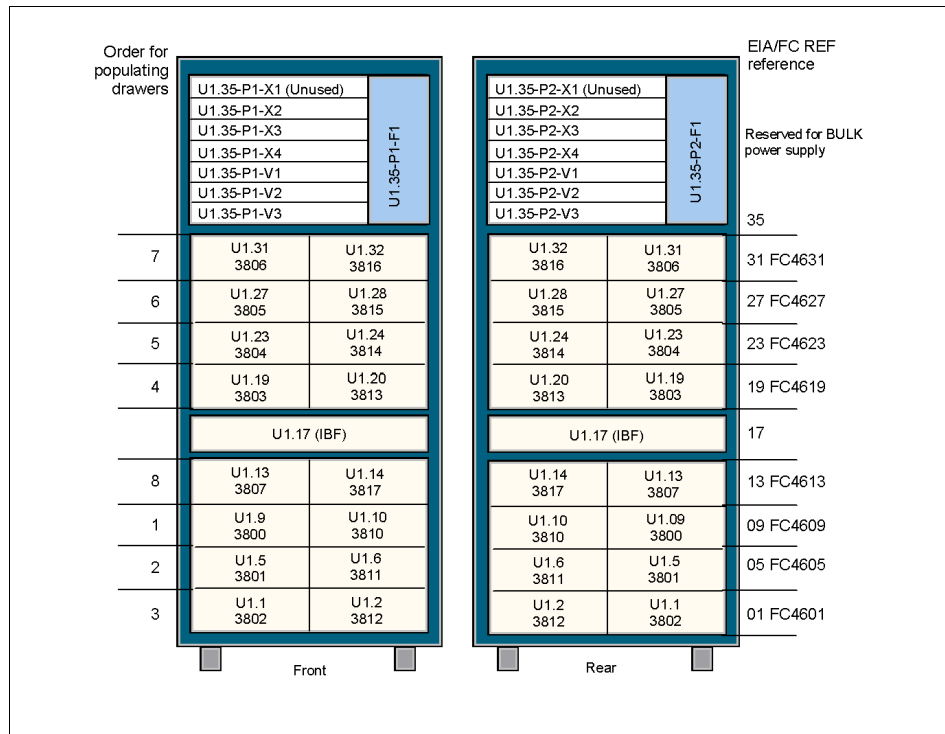


Figure 2-21 p655 racking considerations

Cluster considerations

Following are considerations to take into account when ordering a p655 for attachment to a Cluster 1600. One or more of the following APARs will need to be installed on your Cluster 1600 system in order to support a p655 as a PSSP-managed server:

- ▶ APAR IY37884 needs to be installed on all p655 servers that will run PSSP 3.4, and on all control workstations running PSSP 3.4 that manage clusters containing p655 servers.
- ▶ APAR IY37885 needs to be installed on all p655 servers that will run PSSP 3.5, and on all control workstations running PSSP 3.5 that manage clusters containing p655 servers.
- ▶ APAR IY37884 provides support for the switchless attachment of p655 servers running PSSP 3.4 to Cluster 1600 systems managed by PSSP 3.4 or

PSSP 3.5. With APAR IY37884, p655 servers can be included in clusters containing an SP Switch2 switch, but may not be attached directly to the SP Switch2.

- ▶ APAR IY37885 provides support for the switchless or SP Switch2 attachment of p655 servers running PSSP 3.5 to Cluster 1600 systems managed by PSSP 3.5. SP Switch2 attachment is supported via the SP Switch2 PCI-X Attachment Adapter.
- ▶ APAR IY37885 to allow for current hardware installation, configuration, bring-up and application enablement. Software updates are planned to be made available in the future for production support. In the interim, e-fixes are planned to be provided as they become available.

Important: Problems may be encountered with the s1term whenever Kerberos 4 keyfiles or supper file collection passwords are being passed to a p655 server in SMP (full system partition) mode which does not yet have APAR IY36001 (PSSP 3.4) or APAR IY36002 (PSSP 3.5) installed on all control workstations and/or p655 servers.

PSSP service can be installed automatically during an installation or migration using function introduced in IY42352 for PSSP 3.5 and IY41696 for PSSP 3.4. After installing APAR IY42352 or IY41696 on your control workstation, place PSSP service in directory /spdata/sys1/install/pssplpp/PSSP-3.5 or /spdata/sys1/install/pssplpp/PSSP-3.4 and it will be automatically installed during an installation or migration. For details, refer to *PSSP Installation and Migration Guide*, GA22-7347.

When ordering a p655 for attachment to a Cluster 1600, the features in Table 2-18 should be considered.

Table 2-18 Cluster features for p655

Feature code	Description
7315-C02	Hardware Management Console (HMC)
7315 F/C 8120	Attachment cable HMC-to-host 6 meters (RS-232)
7315 F/C 8121	Attachment cable HMC-to-host 15 meters (RS-232)
7315 F/C 8122	Attachment cable HMC-to-W42 rack 6 meters (RS-422)
7315 F/C 8123	Attachment cable HMC-to-W42 rack 15 meters (RS-422)
7039 F/C 8398	SP Switch2 PCI-X attachment adapter
7039 F/C 6420	pSeries HPS SNI GX bus adapter card

Feature code	Description
9078-160 F/C 0010	Cluster 1600 7039 type server
9078-160 F/C 0011	Cluster 1600 7039 LPAR (switched)

Restrictions:

- ▶ p655 servers cannot be attached to clusters containing an SP Switch.
- ▶ When a pSeries HPS SNI GX bus adapter is installed, there are only two PCI-X slots available in the p655 CEC.

CSM/PSSP support statement

The following support statement details the CSM and PSSP support for the p655:

- ▶ The p655 is supported on CSM V1.3.1 with APAR IY42369 with AIX 5.1 ML04, 5.2 ML01 or later.

Note: The CSM management server must be running AIX 5L V5.2 with Recommended Maintenance package 5200-01 (APAR IY39795). The other machines within the cluster are referred to as “managed nodes” and can be running AIX 5L V5.2 with 5200-01, or V5.1 with the 5100-04 Recommended Maintenance package (APAR IY39794). They can also be xSeries machines running CSM for Linux V1.3.

For all AIX servers, the following CSM for AIX 5L 1.3.1 service is required:

- ▶ Added Service for RSCT on AIX 5.1 IY42782
- ▶ Added Service for RSCT on AIX 5.2 IY42783
- ▶ CSM for AIX 5L added service IY42353
- ▶ CSM Support for p670 and p690 POWER4+ IY42356
- ▶ CSM Support for 7026 servers, 9076 SP nodes and p650 IY42379
- ▶ CSM 1.3.1 support for 9076 SP Node Features 2054/2058 IY42847
- ▶ CSM Support of p655 POWER4+ IY42377

- ▶ The p655-651 is supported on PSSP 3.5 with AIX 5.1 ML03 or later.

For more information on the pSeries 655 and features, refer to the specific server documentation or to *pSeries Systems Handbook 2003 Edition*, SG24-5120.

2.8.5 670 Server (7040-671)

The IBM 7040 pSeries 670 Model 671, as shown in Figure 2-22 on page 83, brings advanced POWER4 and POWER4+ technology to the midrange server environment. The p670 yields increased flexibility through logical partitioning,

Linux support, and a wide variety of application availability. The p670 meets the requirements of customers who need lower life cycle costs with flexibility for multiple workloads, but who do not require scalability to very large systems. The attributes of the pSeries 670 set the stage to grow as your needs change: high availability, scalability, reliability, and flexibility.



Figure 2-22 p670 picture

The p670 can be configured as a 4-way, 8-way, or 16-way computer system, with the CPUs packaged on two Multi-Chip Modules (four or eight POWER4 processors per MCM). The CPUs are either 1.1 GHz POWER4 processors or 1.5 GHz POWER4+ processors with 128 MB of L3 ECC cache per MCM. The memory within the system is DDR-expandable from 4 GB to 256 GB. You can have up to three 7040-61D I/O drawers per server; each I/O drawer supports 20 blind-swap PCI or PCI-X bus slots, and up to 16 Ultra3 SCSI disks that can be hot-plugged. Hardware can be split into as many as sixteen LPARs, each functioning as a “computer within a computer” with its instance of the operating system. Figure 2-23 shows the internal structure of the p670.

pSeries 670 System Frame Organization

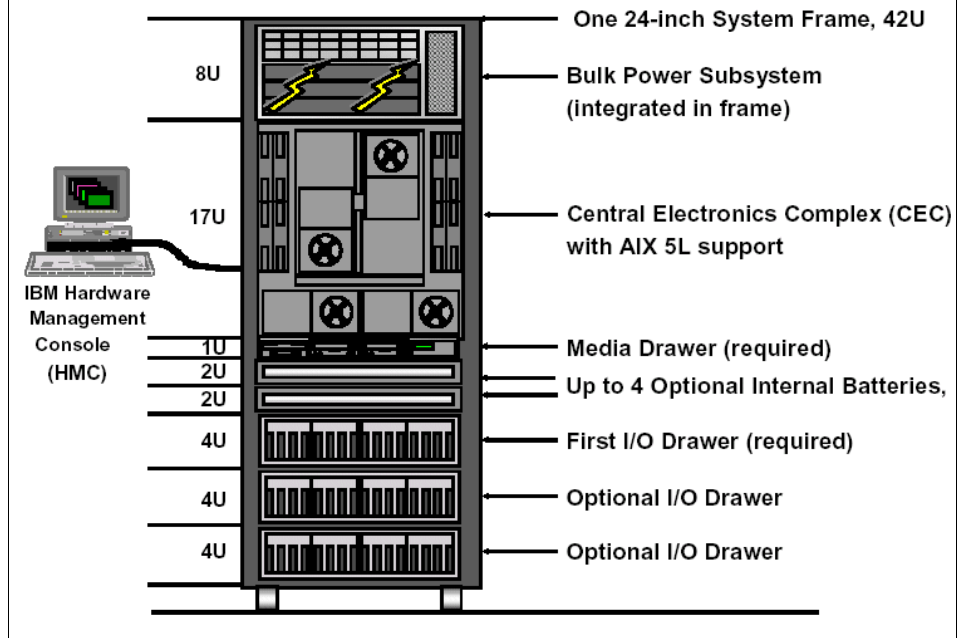


Figure 2-23 7040-671 internal structure

7040-61D I/O drawer

The I/O drawers provide internal storage and I/O connectivity to the system. Figure 2-24 on page 85 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.

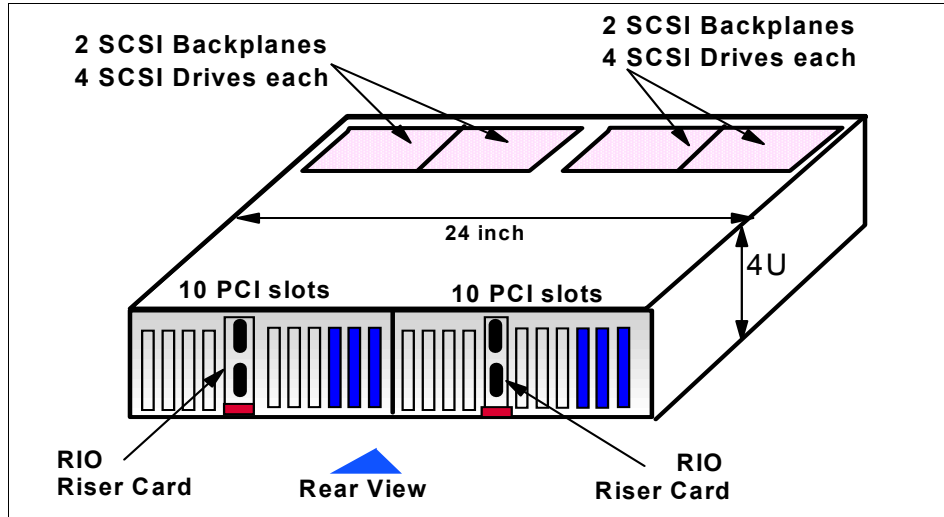


Figure 2-24 p670 I/O drawer rear view

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.

Tip: 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.

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 2-25 on page 86):

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

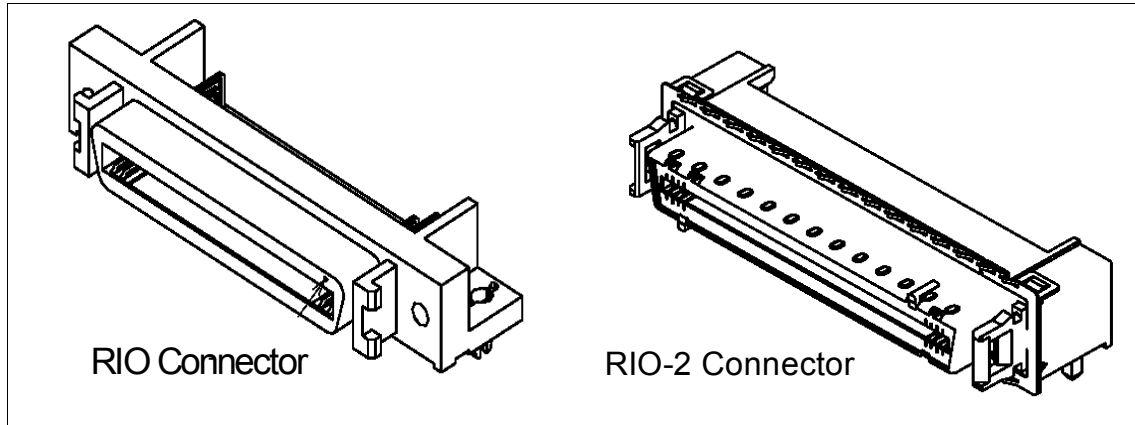


Figure 2-25 Difference between RIO and RIO-2 connectors

The I/O Drawer for pSeries 670, pSeries655 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 F/C 6563, I/O Drawer PCI Planar, 10 slots, 2 Integrated Ultra3 SCSI Ports.
- ▶ RIO-2 drawer is configured with the F/C 6571, I/O Drawer PCI-X Planar, 10 slots, 2 Integrated Ultra3 SCSI Ports.

Summary of standard and optional features

Table 2-19 shows the standard and optional features available for the pSeries 670 model 671.

Table 2-19 Standard and optional p670 features

Description	Quantity	Type
Processors	4, 8, 16 4, 8, 16	POWER4 1.1GHz, 128 MB L3 Cache POWER4+ 1.5GHz 128 MB L3 Cache
Memory	4 - 256 GB	DDR SDRAM
Internal disk bays	16 - 48	16 per 7040-61D I/O drawer
Media bays	4	Located in the media drawer (SCSI adapter required)
PCI slots	20 - 60	64-bit hot swap PCI or PCI-X within 7040-61D I/O drawer

Description	Quantity	Type
Standard ports	2 2 2 - 4	Serial HMC ports 2 - 4 RIO or RIO2 loops supported
Integrated SCSI adapters	4 - 12	Ultra3 SCSI controllers built into the 7040-61D I/O drawer planar
Integrated LAN adapter ports	0	N/A
I/O drawers	1 - 3	7040-61D I/O drawers
Redundant power supply	Standard	Redundant AC bulk power supplies
Operating System Version	AIX	AIX 5L V5.1 with the 5100-02 Recommended Maintenance package (APAR IY28102) or later, or AIX 5L V5.2 or later
Physical specifications		Width: 79 cm (30.3 in) Depth: 149 cm (58.8 in) (Acoustic) Depth: 134 cm (52.8 in) (slimline) Height: 202 cm (79.7 in) Weight: 1,184kgs (2606lbs)max cfg (acoustic) Weight: 1,170kgs (2574lbs)max cfg (slimline)
Power requirements		Operating voltage @ 50/60 Hz: (3-phase) 200 to 240 V AC 380 to 415 V AC 480 V AC Operating voltage @ 50/60Hz (Single-phase): 200 to 415V AC Electrical output: 15,400 watts (maximum) Power source loading: 6.7 kVA (maximum configuration) Thermal Output: 6,670 joules/sec (22,800 Btu/hr) (maximum configuration, 16-way CEC)

Note: The pSeries 670 can be field upgraded with a CEC change to a pSeries 690.

Capacity on Demand (CuOD)

Capacity on Demand (CuOD) is a way of planning for future growth within your server. It means that you can have extra resources available which can be utilized at short notice and with no disruption to service. Payment for the extra resources is only made when the resources are used. Capacity on Demand is available for both CPU resource and memory resource. It also has the added advantage that if one of the active CPUs fail it can be substituted automatically and without disruption for one of the inactive CPUs. This action is non-chargeable. The pSeries 670 systems can be shipped with non-activated resources (processors and/or memory), which may be purchased and activated at a certain point in time without affecting normal machine operation.

The following are the methods available:

- ▶ Capacity Upgrade on Demand
- ▶ Memory Capacity Upgrade on Demand
- ▶ On/off Capacity on Demand
- ▶ Trial Capacity Upgrade on Demand

For further information on capacity on demand, refer to the specific server documentation or to *pSeries Systems Handbook 2003 Edition*, SG24-5120 or to *IBM @server pSeries 690 and pSeries 670 System Handbook*, SG24-7040.

LPAR considerations

The pSeries 670 can be divided into as many as 16 logical partitions. System resources can be dedicated to each LPAR. p670 LPARs have following considerations:

- ▶ LPAR allocation, monitoring, and control is provided by the Hardware Management Console.
- ▶ Each LPAR functions under its own instance of the operating system.
- ▶ A minimum of one processor is required per LPAR.
- ▶ A minimum of 1 GB of system memory is required per LPAR. While it is not required, 4 GB of system memory is recommended per LPAR to ensure adequate memory is available.
- ▶ I/O adapters in PCI slots may be allocated to an LPAR on an individual slot basis.
- ▶ Integrated Ultra3 SCSI controllers located in the 7040-61D I/O drawers may be individually allocated to an LPAR. These integrated SCSI adapters each support one 4-pack Disk Backplane.

Tip: While it is not mandatory, consideration should be given to allocating one half of a 7040-61D I/O drawer for each LPAR. This would provide one 10-slot PCI or PCI-X planar, two integrated SCSI controllers, and two 4-Pack SCSI Disk Backplanes to the LPAR. This will help to ensure balanced I/O bandwidth for the LPAR configurations.

Cluster considerations

In this section, we discuss considerations you need to take into account when ordering a p670 for attachment to a Cluster 1600. Each pSeries 670 clustered server interfaces to the control workstation via Ethernet twisted pair connections. An Ethernet connection to the control workstations is required for each clustered server (or LPAR operating as a clustered server), as well as on the HMC controlling the p670 system. The 10/100 Mbps Ethernet PCI Adapter II (F/C 4962) should be utilized for this connection for the p670 server.

Important: The establishment of a trusted network between the control workstation (CWS) and the Hardware Management Console (HMC) is recommended. Refer to the PSSP V3.4 Read This First document or to the PSSP V3.5 documentation located at:

http://www.rs6000.ibm.com/resource/aix_resource/sp_books/

Restriction: The Ethernet adapters supporting the LPARs and clustered servers are allowed in slots 8, 9, 18, and 19 of the 7040-61D I/O drawer when utilizing the SP Switch2 Attachment Adapter (F/C 8397) or Switch2 PCI-X Attachment Adapter (F/C 8398). It can be installed in slots 1 through 7 or in slots 11 through 17 with the SP Switch Attachment Adapter (F/C 8396). In its basic configuration, the HMC incorporates an Ethernet connection, which is used for this connection.

Four SP Switch Attachment Adapters (F/C 8396) are allowed per 7040-671 server. These adapters are located in the 7040-61D I/O drawers. A maximum of one adapter is allowed per F/C 6563 I/O planar, and two adapters per 7040-61D I/O drawer.

Eight SP Switch2 attachment adapters (F/C 8397) or Switch2 PCI-X attachment adapters (F/C 8398) are allowed per 7040-671 server. These adapters are located in the 7040-61D I/O drawers. A maximum of two F/C 8397 adapters are allowed per F/C 6563 PCI I/O planar. A maximum of two F/C 8398 adapters are allowed per F/C 6571 PCI-X I/O planar.

Restriction: F/C 8397 adapters are not allowed with F/C 6571 PCI-X planars. F/C 8398 adapters are not allowed with F/C 6563 PCI planars.

When ordering a p670 for attachment to a Cluster 1600, you should consider the features listed in Table 2-20.

Table 2-20 Cluster features for p670

Feature code	Description
7315-C02	Hardware Management Console (HMC)
7315 F/C 8120	Attachment cable HMC-to-host 6 meters
7315 F/C 8121	Attachment cable HMC-to-host 15 meters
7040 F/C 8396	SP Switch PCI Attachment Adapter (withdrawn 12/31/03)
7028 F/C 8398	SP Switch2 PCI-X Attachment Adapter
7018 F/C 8397	SP Switch2 PCI Attachment Adapter
9078-160 F/C 0008	Cluster model 1600 7028-type server
9078-160 F/C 0009	Cluster model 1600 7018 LPAR (switched)

Note: The SP Switch (F/C 8396), and Switch2 (F/C 8397) adapters are “double wide” PCI cards and take up the space of two PCI slots. They must be mounted in the optional PCI Blind-Swap Cassette Kit for Double Wide Adapters (F/C 4598).

CSM/PSSP support statement

The following support statement details the CSM and PSSP support for the p670:

- The p670 is supported on CSM V1.3.1 with APAR IY42369 with AIX 5.1 ML04, 5.2 ML01, and later.

Note: The CSM management server must be running AIX 5L V5.2 with Recommended Maintenance package 5200-01 (APAR IY39795). The other machines within the cluster are referred to as “managed nodes” and can be running AIX 5L V5.2 with 5200-01, or V5.1 with the 5100-04 Recommended Maintenance package (APAR IY39794). They can also be xSeries machines running CSM for Linux V1.3.

For all AIX servers, the following CSM for AIX 5L 1.3.1 service is required:

- ▶ Added Service for RSCT on AIX 5.1 IY42782
- ▶ Added Service for RSCT on AIX 5.2 IY42783
- ▶ CSM for AIX 5L added service IY42353
- ▶ CSM Support for p670 and p690 POWER4+ IY42356
- ▶ CSM Support for 7026 servers, 9076 SP nodes and p650 IY42379
- ▶ CSM 1.3.1 support for 9076 SP Node Features 2054/2058 IY42847
- ▶ CSM Support of p655 POWER4+ IY42377

- ▶ The p670-671 is supported on PSSP 3.5 with AIX 5.1 ML03 or later.

For more information on the pSeries 670 and features, refer to the specific server documentation or to *pSeries Systems Handbook 2003 Edition*, SG24-5120.

2.8.6 690 server (7040-681)

The p690 was the first system to offer the POWER4 microprocessor. The p690 is shown in Figure 2-26 on page 92. The POWER4 chip was the first ever dual processor SMP on a single piece of silicon. The balanced performance of the POWER4 and POWER4+ chips are equally at home with mission-critical commercial or compute-intensive applications.



Figure 2-26 The IBM @server690

The p690 can be configured as a 4-way, 8-way, 16-way, 24-way or 32-way computer system, with the CPUs packaged on two Multi-Chip Modules (four or eight POWER4 processors per MCM). The CPUs are either 1.1GHz, 1.3GHz POWER4 processors or 1.5GHz, 1.7GHz POWER4+ processors with 128 MB of L3 ECC cache per MCM. The memory within the system is DDR-expandable from 8 GB to 512 GB. You can have up to eight 7040-61D I/O drawers per server; each I/O drawer supports 20 blind-swap PCI or PCI-X bus slots and up to 16 hot-swap-capable disks structured into four 4-packs of 18.2 GB, 36.4 GB, 73.4 GB or 146.8 GB disks. This delivers up to 2.34 terabytes maximum per I/O Drawer, with up to 8 drawers and 18.79 terabytes of internal disk for the system.

Hardware can be split into as many as thirty-two LPARs, each functioning as a “computer within a computer” with its instance of the operating system. Figure 2-27 on page 93 shows the internal structure of the p690.

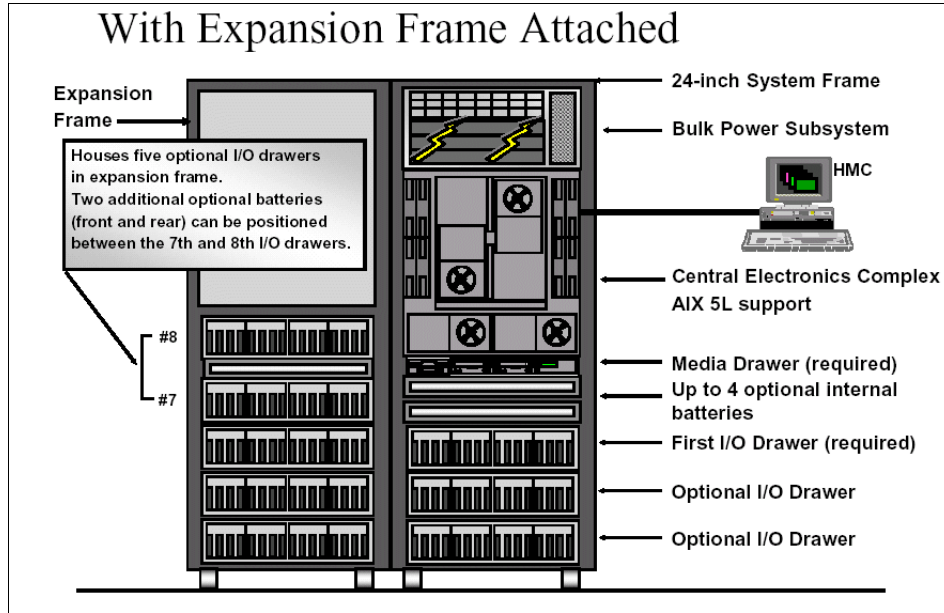


Figure 2-27 7040-681 pSeries 690 internal structure

Complementing the power of the p690 are extraordinary self-managing capabilities. The system can detect faulty memory, processors, cache, and PCI buses and dynamically take them offline; it even makes the service call. It performs all of these tasks without requiring administrative action or interrupting operations. Growth is easy with the hot-swappable disk drive and hot-plug PCI and PCI-X. For the ultimate in availability, the systems can be clustered together with High Availability Cluster Multiprocessing software, the leader in UNIX availability solutions.

7040-61D I/O drawer

The I/O drawers provide internal storage and I/O connectivity to the system. Figure 2-28 on page 94 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. Up to eight drawers can be connected to a pSeries 690.

Note: When utilizing RIO attachment, the following number of I/O drawers are allowed per MCM:

- ▶ One MCM allows attachment of up to two I/O drawers.
- ▶ Two MCMs allow attachment of up to four I/O drawers.
- ▶ Three MCMs allow attachment of up to six I/O drawers.
- ▶ Four MCMs allow attachment of up to eight I/O drawers.

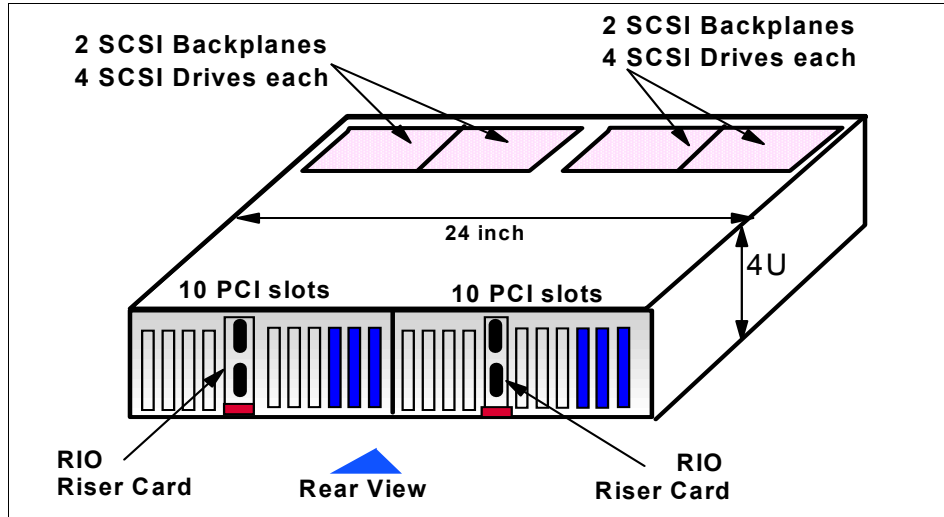


Figure 2-28 p690 I/O drawer rear view

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 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 2-29 on page 95):

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

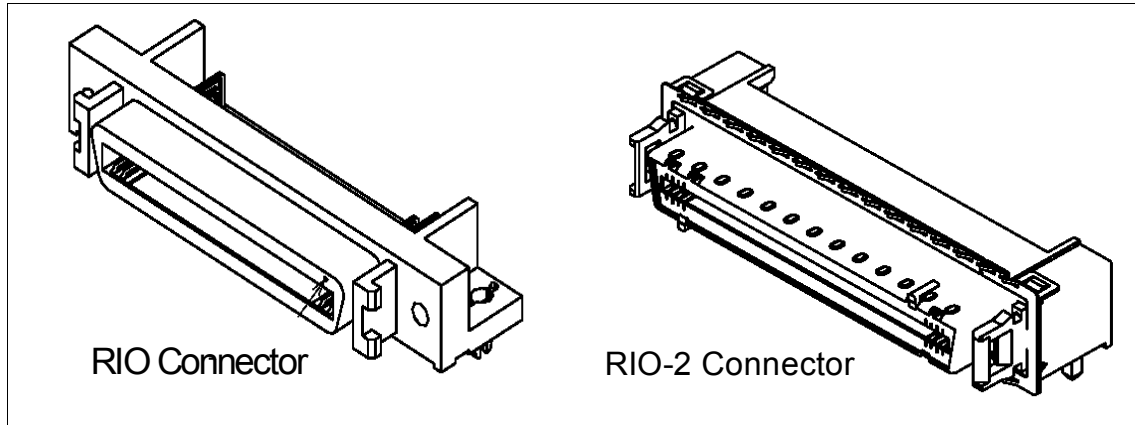


Figure 2-29 Difference between RIO and RIO-2 connectors

The I/O drawer for pSeries 670, pSeries655, 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 F/C 6563, I/O Drawer PCI Planar, 10 slots, 2 Integrated Ultra3 SCSI Ports.
- ▶ RIO-2 drawer is configured with F/C 6571, I/O Drawer PCI-X Planar, 10 slots, 2 Integrated Ultra3 SCSI Ports.

Summary of standard and optional features

Table 2-21 shows the standard and optional features available for the pSeries 690 model 7040-681.

Table 2-21 Standard and optional p690 features

Description	Quantity	Type
Processors	8, 16, 24, 32 8, 16, 24, 32 8, 16, 24, 32 8, 16, 24, 32	POWER4 1.1GHz, 128 MB L3 Cache POWER4 1.3GHz 128 MB L3 Cache POWER4+ 1.5GHz 128 MB L3 Cache POWER4+ 1.7GHz 128MB L3 Cache
Memory	8 - 512GB	DDR SDRAM
Internal disk bays	16 - 128	16 per 7040-61D I/O drawer
Media bays	4	Located in the media drawer (SCSI adapter required)

Description	Quantity	Type
PCI slots	20 - 160	64-bit hot swap PCI or PCI-X within 7040-61D I/O drawer
Standard ports	2 2 2 - 14 loops	Serial HMC ports 2 - 14 RIO or RIO2 loops supported
Integrated SCSI adapters	4 - 32	Ultra3 scsi controllers built into the 7040-61D I/O drawer planar
Integrated LAN adapter ports	0	
I/O drawers	1 - 8	7040-61D I/O drawers
Redundant power supply	Standard	Redundant AC bulk power supplies
Operating system version	AIX	AIX 5L V5.1 with the 5100-01 Recommended maintenance package (APAR IY21957) or later, or AIX 5L V5.2 or later
Physical specifications		Width:154 cm (60.6 in) Depth: 149 cm (58.8 in) (Acoustic) Depth: 134 cm (52.8 in) (slimline) Height:202 cm (79.7 in) Weight: 1,850kgs (2606lbs)max cfg (acoustic) Weight: 1,823kgs (2574lbs)max cfg (slimline)
Power requirements		Operating voltage @ 50/60 Hz: (3-phase) 200 to 240 V AC 380 to 415 V AC 480 V AC Electrical output: 15,400 watts (maximum) Power source loading: 15.0 kVA (maximum configuration)

Capacity on Demand (CuOD)

The pSeries 690 systems can be shipped with non-activated resources (processors and/or memory), which may be purchased and activated at a certain point in time without affecting normal machine operation.

The following are the methods available:

- ▶ Capacity Upgrade on Demand
- ▶ Memory Capacity Upgrade on Demand
- ▶ On/off Capacity on Demand
- ▶ Trial Capacity Upgrade on Demand
- ▶ For further information on capacity on demand, refer to the specific server documentation or to *pSeries Systems Handbook 2003 Edition*, SG24-5120 or *IBM @server pSeries 690 and pSeries 670 System Handbook*, SG24-7040.

LPAR considerations

The pSeries 690 can be divided into as many as 32 logical partitions. System resources can be dedicated to each LPAR. p690 LPARs have the following considerations.

- ▶ LPAR allocation, monitoring, and control is provided by the Hardware Management Console.
- ▶ Each LPAR functions under its own instance of the operating system.
- ▶ A minimum of one processor is required per LPAR.
- ▶ A minimum of 1 GB of system memory is required per LPAR. While it is not required, 4 GB of system memory is recommended per LPAR to ensure adequate memory is available.
- ▶ I/O adapters in PCI slots may be allocated to an LPAR on an individual slot basis.
- ▶ Integrated Ultra3 SCSI controllers located in the 7040-61D I/O drawers may be individually allocated to an LPAR. These integrated SCSI adapters each support one 4-pack Disk Backplane.

Note: The following restrictions apply:

- ▶ Servers configured with the support processor with remote I/O loop attachment (F/C 6404) can support up to a maximum of 16 LPARs.
- ▶ Servers configured with the support processor with remote I/O-2 (RIO-2) loop attachment (F/C 6418) can support up to a maximum of 32 LPARs.

Tip: While not mandatory, consideration should be given to allocating one half of a 7040-61D I/O drawer for each LPAR. This would provide one 10-slot PCI or PCI-X planar, two integrated SCSI controllers, and two 4-Pack SCSI disk backplanes to the LPAR. This will help to ensure balanced I/O bandwidth for the LPAR configurations.

Cluster considerations

In this section, we discuss considerations you should take into account when ordering a p690 for attachment to a Cluster 1600. Each p690 clustered server

interfaces to the control workstation via Ethernet twisted pair connections. An Ethernet connection to the control workstations is required for each clustered server (or LPAR operating as a clustered server), as well as on the HMC controlling the p690 system. The 10/100 Mbps Ethernet PCI Adapter II (F/C 4962) should be utilized for this connection for the p690 server.

Note: The establishment of a trusted network between the control workstation (CWS) and the Hardware Management Console (HMC) is recommended. Refer to the PSSP V3.4 Read This First document or to the PSSP V3.5 documentation at the following site:

http://www.rs6000.ibm.com/resource/aix_resource/sp_books/

Restriction: The Ethernet adapters supporting the LPARs and clustered servers are allowed in slots 8, 9, 18, and 19 of the 7040-61D I/O drawer when utilizing the SP Switch2 Attachment Adapter (F/C 8397) or Switch2 PCI-X Attachment Adapter (F/C 8398). It can be installed in slots 1 through 7, or in slots 11 through 17 with the SP Switch Attachment Adapter ((F/C 8396).

Eight SP Switch Attachment Adapters (F/C 8396) are allowed per 7040-681 server. These adapters are located in the 7040-61D I/O drawers. A maximum of one adapter is allowed per (F/C 6563) I/O planar and two adapters per 7040-61D I/O drawer.

Thirty two SP Switch2 Attachment Adapters (F/C 8397) or Switch2 PCI-X Attachment Adapters (F/C 8398) are allowed per 7040-681 server. These adapters are located in the 7040-61D I/O drawers. A maximum of two F/C 8397 adapters is allowed per (F/C 6563) PCI I/O planar. A maximum of two F/C 8398 adapters is allowed per (F/C 6571) PCI-X I/O planar.

Restriction: Feature (F/C 8397) adapters are not allowed with (F/C 6571) PCI-X planars. Feature (F/C 8398) adapters are not allowed with (F/C 6563) PCI planars.

When ordering a p670 for attachment to a Cluster 1600, the features listed in Table 2-22 should be considered.

Table 2-22 Cluster features for p670

Feature code	Description
7315-C02	Hardware Management Console (HMC)
7315 F/C 8120	Attachment cable HMC to host 6 meters
7315 F/C 8121	Attachment cable HMC to host 15 meters

Feature code	Description
7040 F/C 8396	SP Switch PCI Attachment Adapter (withdrawn 12/31/03)
7028 F/C 8398	SP Switch2 PCI-X Attachment Adapter
7018 F/C 8397	SP Switch2 PCI Attachment Adapter
7040 F/C 6432	pSeries HPS SNI 2-link attachment book
7040 F/C 6434	pSeries HPS SNI 4-link attachment book
9078-160 F/C 0008	Cluster 1600 7028-type server
9078-160 F/C 0009	Cluster 1600 7018 LPAR (switched)

Note: The SP Switch (F/C 8396), and Switch2 (F/C 8397) adapters are “double wide” PCI cards and take up the space of two PCI slots. They must be mounted in the optional PCI Blind-Swap Cassette Kit for Double Wide Adapters (F/C 4598).

CSM/PSSP support statement

The following support statement details the CSM and PSSP support for the p690:

- ▶ The p690 is supported on CSM V1.3.1 with APAR IY42369 with AIX 5.1 ML04, 5.2 ML01, and later.

Note: The CSM management server must be running AIX 5L V5.2 with Recommended Maintenance package 5200-01 (APAR IY39795). The other machines within the cluster are referred to as “managed nodes” and can be running AIX 5L V5.2 with 5200-01, or V5.1 with the 5100-04 Recommended Maintenance package (APAR IY39794). They can also be xSeries machines running CSM for Linux V1.3.

For all AIX servers, the following CSM for AIX 5L 1.3.1 service is required:

- ▶ Added Service for RSCT on AIX 5.1 IY42782
- ▶ Added Service for RSCT on AIX 5.2 IY42783
- ▶ CSM for AIX 5L added service IY42353
- ▶ CSM Support for p670 and p690 POWER4+ IY42356
- ▶ CSM Support for 7026 servers, 9076 SP nodes and p650 IY42379
- ▶ CSM 1.3.1 support for 9076 SP Node Features 2054/2058 IY42847
- ▶ CSM Support of p655 POWER4+ IY42377

- The p690-681 is supported on PSSP 3.5 with AIX 5.1 ML03 and later.

For more information on the pSeries 690 and its features, refer to the specific server documentation or to *pSeries Systems Handbook 2003 Edition*, SG24-5120, or to *Configuring a p690 in an IBM Cluster 1600*, REDP0187.

2.8.7 xSeries servers

With CSM for AIX-managed clusters, selected xSeries nodes running Linux can be included in the cluster. These servers are not available from the Cluster 1600, although they can be attached to a CSM for AIX-managed cluster. For more information on xSeries servers, see:

<http://www.ibm.com/servers/eserver/education/xseries/xref.html>

2.9 Switches

The SP switches provide a message-passing network that connects all processor nodes with a minimum of four paths between any pair of nodes. The switches provide a high bandwidth, low latency network that can use either TCPIP or MPI protocol to pass messages.

Currently there are three types of switches available for the Cluster 1600. Two are supported in clusters managed by PSSP, and one is supported in clusters managed by CSM. The two switch types for PSSP managed clusters are the SP Switch and the SP Switch2. There are two models available for each switch type. Each switch type has a 23-inch rack version (SP frame) and a 19-inch rack version (7014 pSeries rack).

Attention: The SP Switch and all associated features will be withdrawn from marketing on December 31, 2003.

For more information on the PSSP switches, refer to *RS/6000 SP and Clustered IBM eServer pSeries System Handbook*, SG24-5596.

The switch that is supported in CSM-managed clusters is the 7045-SW4 pSeries HPS. There is one model of this switch available, and it is currently supported on p690 and p655 nodes. It is packaged in a 23-inch form and can be mounted in either a 7040-61R or a 7040-W42 rack.

The frame model types and descriptions for the switches are described in the following sections.

2.9.1 9076 model 555

The SP Switch provides high bandwidth, low latency communication between nodes, supplying a minimum of four paths between any pair of nodes. The SP Switch can be used in conjunction with the SP Switch Router to dramatically increase the speed for TCP/IP, file transfers, remote procedure calls, and relational database functions.

The required SP Switch Adapter (F/C 4020), SP Switch MX Adapter (F/C 4022), or SP Switch MX2 Adapter (F/C 4023) connects each SP node to the SP Switch subsystem. One adapter of the required type must be ordered for each node in a switch-configured SP system. If you are using switch expansion frames, the SP Switch subsystem will allow you to scale your SP system up to 128 nodes.

When you order F/C 4011, you receive one 16-port SP Switch and all of the switch-to-node cables you need to connect the switch ports to up to sixteen nodes, both within the switch-equipped frame and in any non-switched expansion frames. You must specify the length of all switch-to-switch cables that make the connections between switch-equipped frames and to an SP Switch Frame.

The IBM 9076 Tall 24-inch Frame Model 555 requires an SP Switch and a minimum of two clustered RS/6000 or pSeries servers, along with a control workstation to create a functional system. An additional SP Switch is supported in the Model 555 expansion frame feature 1555. The 9076 Model 555 has a three-phase power system that includes N+1 power capability. This frame accepts power input 200-240V 50/60Hz 3-phase 50 A or 380-415V 50/60Hz 3-phase 30 A. Frame Redundant Power Input Features are available.

SP nodes are supported on the SP Switch ports within the Model 555. SP Switches can connect to other SP Switches in other 9076 Models (for example, switch-to-switch connections between 9076 Model 555, and Model 550 and Model 557, are supported. Special attention to cluster system planning for Model-to-Model cabling is required).

Important: The SP Switch and all associated features will be withdrawn from marketing on December 31, 2003.

2.9.2 9076 model 556

The SP Switch2 is the next step in the evolution of the SP interconnection fabric and offers enhanced performance and RAS (Reliability, Availability, and Serviceability) over the SP Switch. The performance gains are significant increases in bandwidth and reductions in latency.

SP Switch2 provides a one-way bandwidth of up to 500 MB per second between nodes (1 GB bidirectional) for interconnecting POWER3 SMP high nodes.

Because the SP Switch2 design improvements are evolutionary steps on the SP and high performance switches, the SP Switch2 is fully compatible with applications written for the older switches. The SP Switch2 software support continues to emphasize usage of Message Passing Interface (MPI), Low level Application Interface (LAPI), and Internet Protocol (IP), but the new switch subsystem enhances message handling so that the switch can keep pace with the increased performance of the newer nodes.

The IBM 9076 Tall Frame Model 556 requires an SP Switch2 and a minimum of two clustered RS/6000 or pSeries servers, along with a control workstation, to create a functional system. Additional SP Switch2s and features are supported for each Model 556, to scale the frame up to a maximum of eight SP Switch2s.

Both single-plane and two-plane SP Switch2 environments are supported within the server scaling limits. The 9076 Model 556 has a three-phase power system which includes N+1 power capability. This frame accepts power input 200-240V 50/60Hz 3-phase 50 A, or 380-415V 50/60Hz 3-phase 30 A. Frame-redundant power input features are available.

SP nodes are supported on the SP Switch2 ports within the Model 556. SP Switch2s can connect to other SP Switch2s in other 9076 Models (for example, switch-to-switch connections between 9076 Model 556, and Model 550 and Model 558, are supported. Special attention to cluster system planning for Model-to-Model cabling is required.)

2.9.3 9076 model 557

The IBM 9076 Model 557 requires an SP Switch and a minimum of two clustered RS/6000 or pSeries servers, along with a control workstation, to create a functional system. The Model 557 is used for mounting the SP Switch and power subsystem in IBM 19-inch racks, enhancing the available switch building blocks of the Cluster 1600. This model is field-installed into a new or existing 7014-T00 or 7014-T42 rack, requiring 16 EIA positions.

The 9076-557 can have up to one SP Switch. The Model 557 requires two of the 7014 PDUs, or power distribution units, for dual power input. The two power cords for Model 557 plug into one receptacle of each of the PDU units.

SP nodes are supported on the SP Switch ports within the Model 557. The SP Switch within Model 557 can connect to other SP Switches in other 9076 Models (for example, switch-to-switch connections between 9076 Model 557, and Model

550 and Model 555, are supported). Special attention to cluster system planning for Model-to-Model cabling is required.

Attention: The SP Switch and all associated features will be withdrawn from marketing on December 31, 2003.

2.9.4 9076 model 558

The IBM 9076 Model 558 requires an SP Switch2 and a minimum of two clustered RS/6000 or pSeries servers, along with a control work station, to create a functional system. The Model 558 is used for mounting the SP Switch2 and power subsystem in IBM 19-inch racks, enhancing the available switch building blocks of the Cluster 1600. This model is field-installed into a new or existing 7014-T00 or 7014-T42 rack, requiring 16 EIA positions.

The 9076-558 can have up to two SP Switch2s. The Model 558 requires two of the 7014 PDUs, or power distribution units, for dual power input. The two power cords for Model 558 plug into one receptacle of each of the PDU units. Both single-plane and two-plane SP Switch2 environments are supported within the server scaling limits.

The SP Switch2s within Model 558 can connect to other SP Switch2s in other 9076 Models (for example, switch-to-switch connections between 9076 Model 558, and Model 550 and Model 556, are supported. Special attention to cluster system planning for Model-to-Model cabling is required).

2.9.5 7045-SW4 pSeries HPS (High Performance Switch)

The pSeries High Performance Switch (HPS) communication subsystem is an evolutionary step in network data transfers using technology based on the proven architecture of SP Switch and SP Switch2. The technology driving the pSeries HPS is designed to augment the latest offerings of pSeries 690 and 655 clustered servers by increasing the communication bandwidth between servers and partitions within the cluster. The benefits of implementing this new communication subsystem include many enhancements over previous SP switch offerings, including:

- ▶ Parallel, interconnected communication channels that form a unified switch network
- ▶ Significantly improved communication bandwidth and reductions in latency
- ▶ The option to use either fiber optic or copper cables for switch-to-switch network connections
- ▶ Improved reliability, availability, and serviceability (RAS)

Restriction: The pSeries HPS is the first offering in a new generation of switch technology. As such, it does not support migration from, nor coexistence with, previous versions of SP switch networks. It is only supported on CSM-managed Clusters and is *not* supported with PSSP.

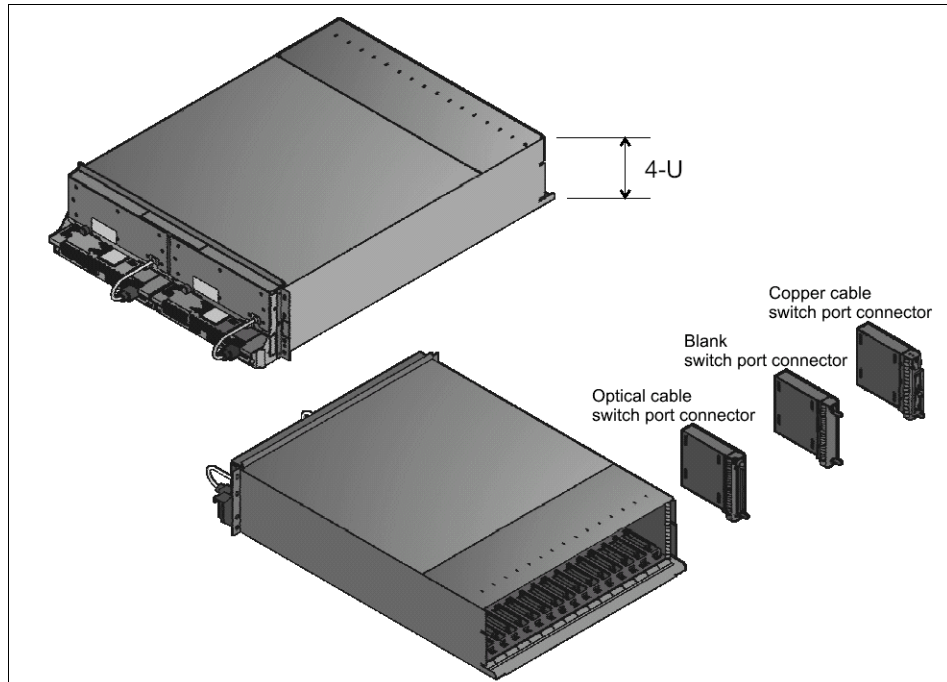


Figure 2-30 HPS switch internal structure

The pSeries HPS (M/T 7045-SW4) is packaged in a 4U-high 23-inch chassis with optional slot in switch port connector books. Figure 2-30 shows the internal structure of the HPS switch. These books can be copper, optical or blanks. Server connections are copper, and inter-switch connections can be either copper or fibre (optical).

These books have two links on each book. The server attachment books can be located in slots C3, C4, C7, C8, C11, C12, C15, C16, as shown in Figure 2-31 on page 105. Inter-switch link books can be located in the remaining 8 slots.

The switch supports 16 servers and 16 inter-switch links. The switch can be mounted in either a 7040-61R or a 7040-W42 rack. A p690 or p655 rack can only support one 7045-SW4 switch. Additional switches must be mounted in a separate 7040-W42 rack.

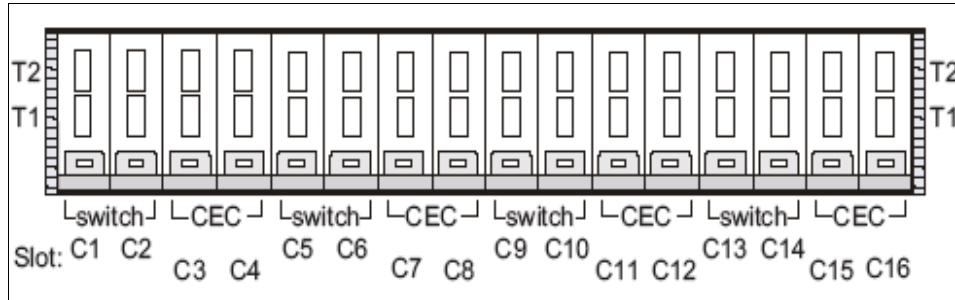


Figure 2-31 pSeries HPS book locations

The pSeries HPS switch is supported on 7040-681 and 7039 servers only. It is connected to the servers using either 3-meter or 10-meter copper cables. The server interface is either a 2-link or a 4-link switch network interface (SNI) card which plugs directly onto the GX bus on the server. The switch network interface card can be shared with multiple LPARs.

Both links within the SNI card allow message passing across direct internal connections. This means that both links can be on the same unified single switch network, which increases the resilience over the previous SPswitch2 technology. If one link fails, traffic can be directed over the other link seamlessly. Each switch link can be shared across 16 LPARs, and a single LPAR can also be connected to up to 8 SNI cards for increased bandwidth.

Tip: pSeries HPS switched nodes can be intermixed with non-switched nodes.

As with the SP switch technology, switches can be server switch boards (SSB) or inter-switch boards (ISB). Server switch boards are connected directly to servers, and inter-switch boards are only connected to other switches. You can have up to three SSBs in a switch network before you require ISBs.

The switch is not required to be controlled by a specific node but instead is controlled by a program running on the HMC called Switch Network Manager (SNM). This program is part of the HMC software build and controls, and it monitors the switch network. The human interface with this program is integrated through the HMC Web-Based System Management GUI. This allows you to configure, power on/off, run diagnostics and check the status of the HPS switch. As prerequisites, the Cluster 1600 must have the following software installed:

- ▶ AIX 5.2 latest Recommended Maintenance Level
- ▶ CSS file sets for AIX
- ▶ SNM switch network manager (HMC)
- ▶ SFP Service focal point (HMC)

- ▶ Service Agent (Service Gateway between IBM and customer)
- ▶ Inventory Scout
- ▶ Cluster Systems Management

The hardware connection to the HMC is by two RS-422 connections between the HMC and each of the 7040 racks that contain servers attached to the pSeries HPS. The RS-422 cable plugs into the bulk power controller (BPC) at the top of the 7040 racks. The switch cable is the only direct connection between the node and the HPS switch.

Note: Four RS-422 cables are required per rack if redundant HMC is being used.

For more information on HPS switch cabling, refer to Chapter 3, “Network configuration” on page 121.

Summary of standard and optional features

Table 2-23 lists the standard and optional features available for the 7045-SW4 pSeries HPS.

Table 2-23 Standard and optional 7045-SW4 HPS switch features

Description	Feature code
pSeries HPS ISB indicator	F/C 9049
pSeries HPS SSB indicator	F/C 9047
RS-422 cable 15 m (HMC-to-frame)	F/C 8123
RS-422 cable 6 m (HMC-to-frame)	F/C 8122
Dual SNI link-to-single SNI link convertor	F/C 6437
Switch Port connection card, optical	F/C 6436
Switch Port connection card, blank	F/C 6435
Switch network interface, 4-link (book) (p690)	F/C 7040-6434
Switch Port connection card, copper	F/C 6433
Switch network interface, 2-link (book) (p690)	F/C 7040-6432
Switch network interface, 2-link (GX bus-mounted card) (p655)	F/C 7039-6420
Frame extender	F/C 6234
Network diagnostic tool kit	F/C 3756

Description	Feature code
Switch cable, 40 m (fiber optic)	F/C 3257
Switch cable, 20 m (fiber optic)	F/C 3256
Switch cable, 10 m (copper)	F/C 3167
Switch cable, 2.2 m (copper)	F/C 3166

Racking considerations

The 7040-W42 or 7040-61R System Racks provide space for mounting the pSeries HPS. The Model W42 is a 24-inch rack that provides 42U of rack space. The 7040-W42 utilizes a 350 V DC bulk power subsystem to provide power for the components within the rack.

The bulk power subsystem incorporates redundant bulk power assemblies mounted in the front and rear sections of the top 8U of the rack. Note the following:

- ▶ The 7040-61R rack can contain 1 pSeries HPS with a p690 CEC (681 only).
- ▶ The 7040-W42 rack can contain 1 pSeries HPS with p655 CECs.
- ▶ The 7040-W42 rack with no p655 CECs can contain up to 8 pSeries HPSs and up to 16 pSeries HPSs with expansion frame (F/C 8691).

Tip: The rack must have RS-422 connections from the HMC to each of the bulk power supply controllers.

These cables require an async card to be present in the HMC. If the 128 port card (F/C 2944) is used RS-232 - RS-422 connectors are also required.

CSM/PSSP support statement

The following support statement details the CSM and PSSP support for the 7045-SW4 pSeries HPS:

- ▶ The 7045-SW4 pSeries HPS is supported on CSM V1.3.2 with 5.2 ML 02 and later.

Note: The CSM management server must be running AIX 5L V5.2 with Recommended Maintenance package 5200-02. The other machines within the cluster are referred to as “managed nodes” and can be running AIX 5L V5.2 with 5200-02.

- ▶ Added Service for RSCT on AIX 5.2 IY42783
- ▶ CSM 1.3.2 for AIX 5L
- ▶ Service Focal point
- ▶ Service Agent
- ▶ Inventory Scout

- ▶ The 7045-SW4 HPS switch is *not* supported on PSSP 3.5.

2.10 Switch adapters

There are currently four types of switch adapter available: MX, two PCI types, and PCI-X. The MX adapters are only available for SP nodes with an MX slot available. The PCI adapters and PCI-X adapters are available for all attached servers. The available switch adapters are:

- ▶ PCI SP Switch adapter (F/C 8396) (withdrawn from marketing 12/31/2003)
- ▶ PCI SP Switch2 adapter (F/C 8397)
- ▶ PCI-X SP Switch2 adapter (F/C 8398)
- ▶ HPS Switch network interface card (F/C 6434,6432, 6420)

Restrictions:

- ▶ PCI SP Switch Attachment Adapter (F/C 8396) uses 5 volt switching, and are only supported in 5 volt PCI slots. No PCI-X planars support the PCI SP Switch Attachment Adapter.
- ▶ SP Switch Attachment Adapter (F/C 8396) and PCI-X SP Switch2 Attachment Adapter (F/C 8398) *cannot* be hot-plugged.
- ▶ The PCI SP Switch2 Attachment Adapter (F/C 8397) is a double wide adapter.

For each adapter, you must also order one of the following cables (depending on your floor plan layout) to connect the adapter to a valid switch port:

- ▶ 2.6 m (9 ft.) switch cable (F/C 9302)
- ▶ 5 m (16 ft.) switch cable (F/C 9305)
- ▶ 10 m (33 ft.) switch cable (F/C 9310)
- ▶ 15 m (49 ft.) switch cable (F/C 9315)
- ▶ 20-m (66 ft.) switch cable (F/C 9320)

2.10.1 Switch adapter placement restrictions

The following are placement limitations for the SP Switch and the SP Switch2 Attachment Adapters within the current server range.

Note: Each instance of an SP Switch (F/C 8396), SP Switch2 (F/C 8397), or SP Switch2 PCI-X (F/C 8398) adapter should be counted as two adapters when calculating the maximum adapter limitation.

M/T 7040 models

Placement considerations for the SP Switch and the SP Switch2 Attachment Adapters for the machine type 7040:

- ▶ For the SP System Attachment Adapter (F/C 8396) – install in I/O subsystem slot 8 only (one adapter per LPAR). (This is withdrawn from marketing December 31, 2003.)
- ▶ For SP Switch2 Attachment Adapter (F/C 8397).
 - Single-plane – install in I/O subsystem slot 3 or 5, or both if on separate LPARs (one adapter per LPAR).
 - Two-plane – install in I/O subsystem slot 3 for css0 and slot 5 for css1 (one adapter per LPAR).

Note: The SP Switch Attachment Adapter (F/C 8396) and the SP Switch2 Attachment Adapters (F/C 8397 and F/C 8398) are mutually exclusive on a pSeries 690 system.

M/T 7039 models

Placement considerations for the SP Switch and the SP Switch2 Attachment Adapters for the machine type 7039:

- ▶ For SP Switch2 PCI-X Attachment Adapter (F/C 8398):
 - Single-plane – install in server slot 1 or 3 or both if on separate LPARs (one adapter per LPAR, 2 maximum per server).
 - Two-plane – install in server slot 1 for css0 and server slot 3 for css1.

Note: M/T 7039 LPAR and multi-plane switch restrictions

- ▶ If the servers are going to be configured with a two-plane switch fabric, they cannot be configured with LPARs.
- ▶ If the servers are configured with LPARs, the system is restricted to single-plane switch configurations.
- ▶ Servers configured with two LPARs require one adapter for each LPAR connected to the switch. However, the 7039 can be configured with one LPAR attached to the switch and the other LPAR off the switch.
- ▶ Additional switch adapters (F/C 8396, F/C 8397, or F/C 8398) are not permitted in RIO drawers attached to these servers.

M/T 7028 models

Placement considerations for the SP Switch and the SP Switch2 Attachment Adapters for the machine type 7028:

- ▶ The adapter must be installed in slot 1 of the server.
- ▶ Server slot 2 must be left empty.

2.10.2 pSeries HPS switch network interface cards (SNI)

The SNI card is either a 2-link or a 4-link that plugs directly onto the GX bus on the server. The switch network interface card can be shared with multiple LPARs. Both links within the SNI card allow message passing across direct internal connections. This means that both links can be on the same unified single switch network, which increases the resilience over the previous SPswitch2 technology. If one link fails, traffic can be directed over the other link seamlessly. Each switch link can be shared across 16 LPARs, and a single LPAR can also be connected to up to 8 links for increased bandwidth.

Type 7040 servers

Type 7040 servers can support up to two 2-link or 4-link SNI books (F/C 6434, F/C 6432).

Restriction: The number of links supported is dependent on the number of MCMs installed in the p690.

- ▶ 1 MCM supports 2 links.
- ▶ 2 MCM support 4 links.
- ▶ 3 MCM support 6 links.
- ▶ 4 MCM support 8 links.

The SNI cards (F/C 6434, F/C 6432) plug into the GX bus in the P690 CEC. There are two slots available for the SNI books. Figure 2-32 on page 111 shows the SNI GX bus slots on the p690 CEC.

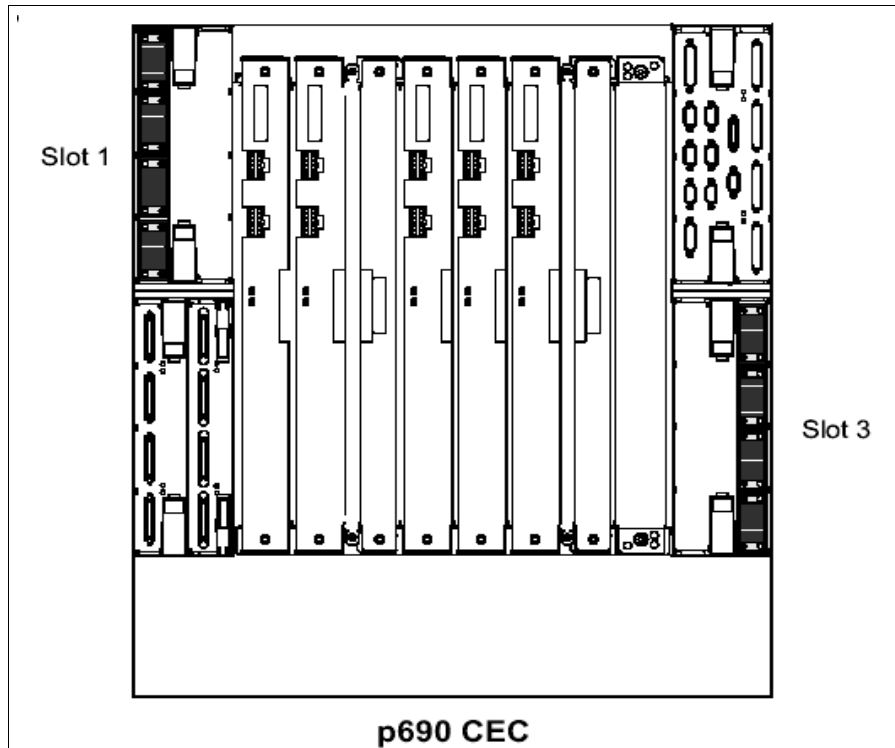


Figure 2-32 p690 SNI GX bus slots location

Note: The two links on each SNI card are paired. A single link convertor wrap is required to use them as a single link.

Type 7039 servers

Type 7039 servers can support one 2-link card (F/C 6420). The GX card plugs into slot 1 in the p655 CEC. The SNI card plugs in instead of the first PCI-X card.

Restriction: When a SNI card is in the p655, there are only two PCI-X slots available for PCI-X cards within the CEC.

Figure 2-33 on page 112 shows the SNI location on a p655 server.

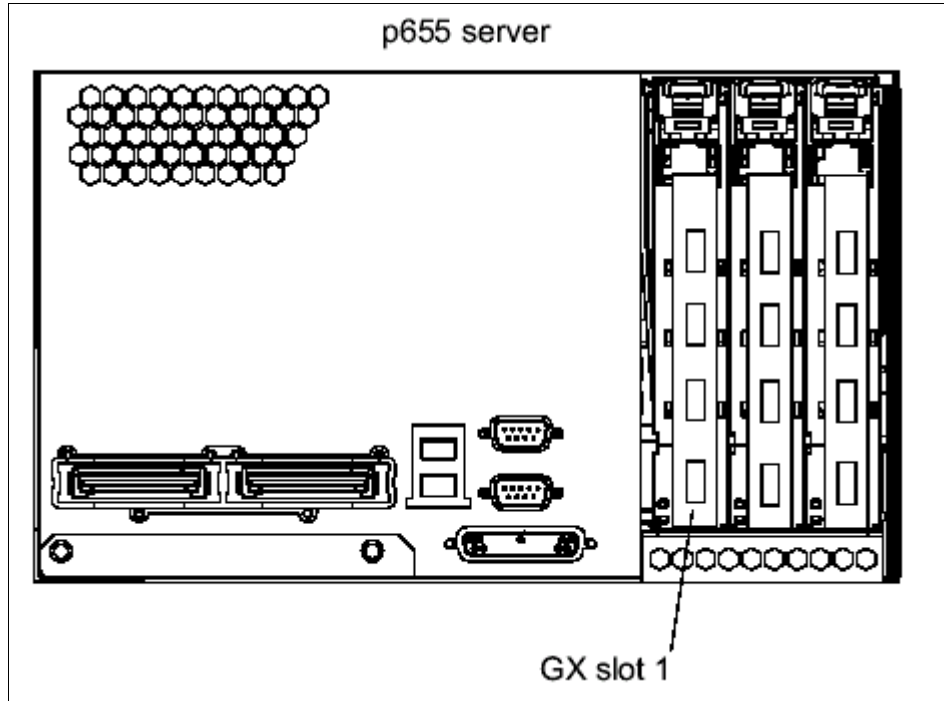


Figure 2-33 p655 SNI GX bus slot location

Note: The two links on each SNI card are paired. A single-link convertor wrap is required to use them as a single link.

2.11 Legacy hardware supported but no longer marketed

This section only details changes since the IBM Redbook *RS/6000 SP and Clustered IBM @server pSeries System Handbook* was published. For further details about legacy hardware, refer to *RS/6000 SP and Clustered IBM @server pSeries System Handbook*, SG24-5596.

2.11.1 pSeries 660 Model 6M1 (7026-6M1)

The pSeries 660 Model 6M1 is a mid-range member of the 64-bit family of symmetric multiprocessing (SMP) enterprise servers from IBM. Positioned between the Model 6H1 and the powerful Model S85, the Model 6M1 provides the power, capacity, and expandability required for e-business, mission-critical

computing. The Model 6M1 can assist you in managing the evolution of your business to incorporate the power of the Web and 64-bit computing into your environment, while still supporting existing 32-bit applications. Figure 2-34 shows the 7026-6M1 and its primary I/O drawer.

The p660 model 6M1 is the follow-on server to the RS/6000 7026-M80.



Figure 2-34 The 7026-6M1 and I/O drawer

The Model 6M1 delivers 64-bit scalability via the 64-bit RS64 IV processor packaged as 2-way and 4-way cards. With its two-processor positions, the model 6M1 can be configured into 2-, 4-, 6-, or 8-way SMP configurations.

The model 6M1 also incorporates an I/O subsystem supporting 32-bit and 64-bit standard PCI adapters. The Model 6M1 has 2-way and 4-way processor cards that operate at 750 MHz and incorporate 8 MB of L2 cache per processor.

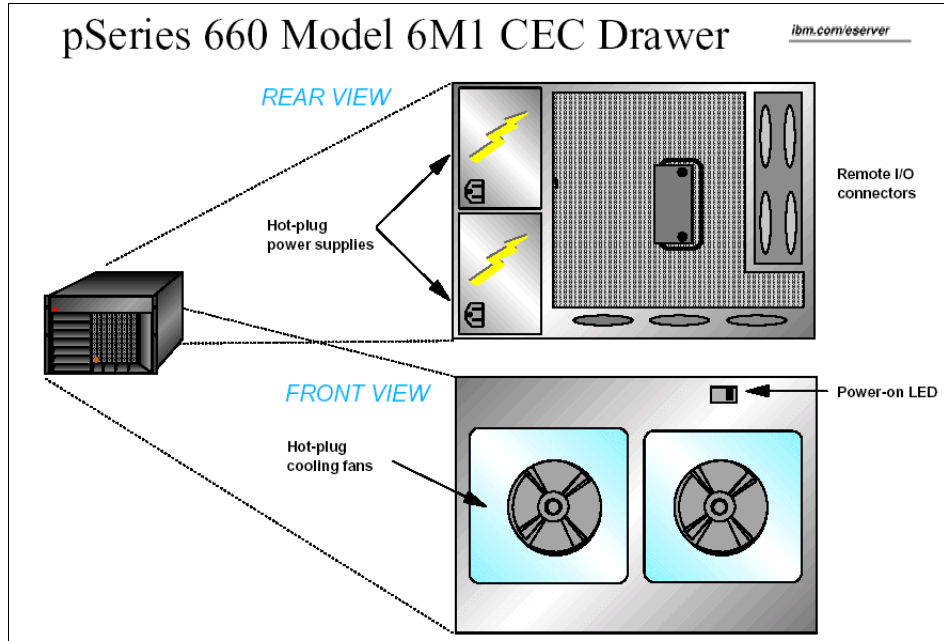


Figure 2-35 p660-6M1 internal structure

The Model 6M1 is packaged as a rack-mounted Central Electronics Complex (CEC) drawer, cable-attached to rack-mounted Remote I/O (RIO) drawers. The CEC and I/O drawers offer redundant power and redundant cooling. The CEC drawer incorporates the system processors, memory, and supporting systems logic. The primary IO drawer contains the PCI cards, boot disks and media bays. The internal structure is shown in Figure 2-36 on page 115.

Up to three secondary I/O drawers can be added to give you up to a total of 56 PCI slots and eight media bays. System storage is added via remotely attached SCSI, SSA, or Fibre Channel storage subsystems.

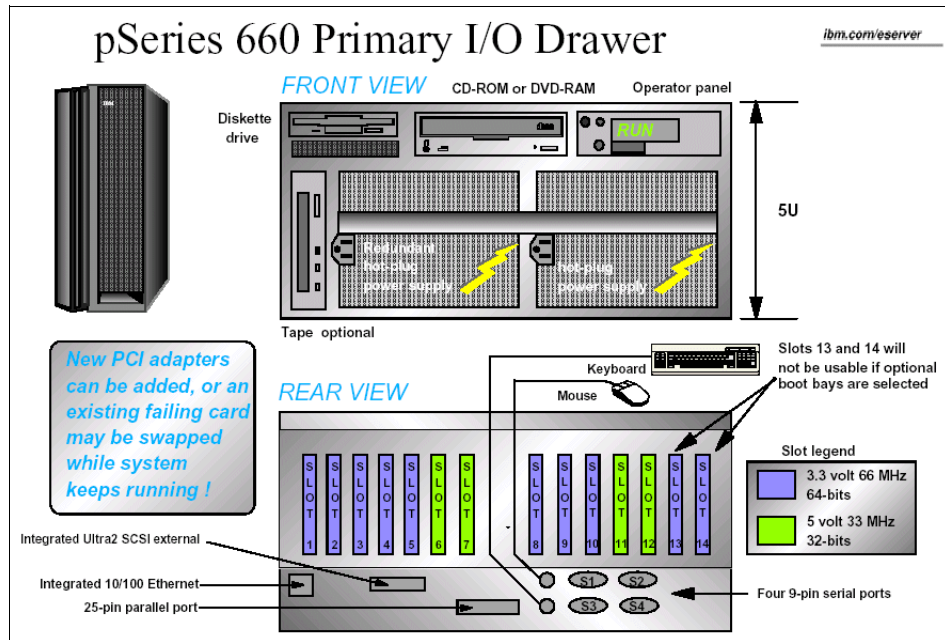


Figure 2-36 p660 IO drawer internal structure

Summary of standard and optional features

Table 2-24 lists the standard and optional features available for the pSeries 660 model 6M1.

Table 2-24 p660-6M1 standard and optional features

Description	Quantity	Type
Processors	2, 4, 6 or 8 2 or 4	750 MHz RS64 IV with 8 MB L2 cache per processor 500 MHz RS64 III with 4 MB L2 cache per processor
Memory	2 - 64GB	DIMMs
Internal disk bays	2	Hot swap 18.2 GB - 146.8 GB per bay
Media bays	3	In I/O drawer
PCI slots	10 4	64-bit hot swap PCI 32-bit hot swap PCI

Description	Quantity	Type
Standard ports	1 1 4 1 4	Keyboard Mouse Serial Parallel RIO ports
Integrated SCSI adapters	2	Ultra 3 SCSI controllers (1 int and 1 ext VHDC I AMP half pitch 68-pin)
Integrated LAN adapter ports	1	10/100 Ethernet controller
I/O drawers	1 - 4	
Redundant power supply	standard	Redundant AC power
Operating system version	AIX	Version 4.33
Physical specifications		Central Electronics Complex: Width: 445 mm (17.5 in) Depth: 826 mm (32.5 in) Height: 356 mm (14.0 in) Weight: 59.7 kg (132 lb) (minimum configuration) I/O Drawer: Width: 445 mm (17.5 in) Depth: 820 mm (32.3 in) Height: 218 mm (8.6 in) Weight: 41 kg (90 lb) (minimum configuration) 52 kg (115 lb) (maximum configuration)
Rack 7014-T00 Rack 7014-T42	2 3	Maximum is two per rack Maximum is three per rack Racks may be shared with peripherals

Description	Quantity	Type
Power requirements		Central Electronics Complex: Operating voltage: 200 to 240 V ac 50/60 Hz Electrical output: 370 watts (typical); 550 watts (maximum) Power source loading: 0.39 kVA (typical configuration) 0.60 kVA (maximum configuration) Thermal Output: 370 joules/sec (1,265 Btu/hr, typical configuration) 550 joules/sec (1,877 Btu/hr, maximum configuration) I/O Drawer: Operating voltage: 200 to 240 V ac 50/60 Hz Electrical output: 220 watts (typical); 515 watts (maximum) Power source loading: 0.23 kVA (typical configuration) 0.54 kVA (maximum configuration) Thermal Output: 220 joules/sec (750 Btu/hr, typical configuration) 511 joules/sec (1,750 Btu/hr, maximum configuration)

Cluster considerations

The Cluster 1600 is enhanced to include the 7026-6M1 server within the available cluster building blocks.

SP Switch2 attached servers utilize one or two SP Switch2 Attachment adapters (F/C 8397) for high-speed interconnection to the SP Switch2 mounted in a 9076 frame. This environment does not require a 9076 SP Node in the configuration:

- ▶ SP Switch-attached servers utilize the SP Switch Attachment adapter (F/C 8396) for high-speed interconnection to the SP Switch mounted in a 9076 frame.
- ▶ A non-switch attached environment utilizes an Ethernet connection to pass data between the clustered servers. This environment does not require a switch, 9076 Frame, or 9076 SP node in the configuration.
- ▶ The SP Switch2 supports two configurations allowing communications over either one or two switch planes. Two switch plane configurations must incorporate two SP Switch2 Attachment Adapters (F/C 8397) in each attached server. Implementations utilizing a one-switch plane configuration are limited to a maximum of one Switch2 Attachment Adapter (F/C 8397) per

attached server. Each clustered server interfaces to the control workstation and SP switch via a combination of cables, depending on environment.

Restrictions: The cable and PCI Card for SP Control Workstation Attachment (F/C 3154) is required in either the non-switch-attached or switch-attached environments. Feature (F/C 3154) provides internal connection from an internal connector on the planar of the primary I/O drawer to a PCI slot location on the rear bulkhead of the I/O drawer. The PCI card associated with feature (F/C 3154) must be located in the primary I/O drawer in the following slots:

- ▶ Slot #7 for non-switched clustered server environment.
- ▶ Slot #7 for SP Switch Attachment Adapter (F/C 8396) environment.
- ▶ Slot #6 for SP Switch2 Attachment Adapter (F/C 8397) environment.
- ▶ Feature (F/C 3154) must be ordered with each 7026 server in the cluster.

The non-switch attached cluster environment requires each server to be connected to the control workstation via a Clustered Server Control Panel-to-Control Workstation Cable (F/C 3151). This cable may be ordered with either the clustered server (recommended) or the control workstation, as desired. This cable is not required for switch-attached servers.

SP switch-attached servers interface to the control workstation and SP frame via nomenclature and cables ordered from the 9076 SP. The appropriate features for this interface are provided by the SP-attached Server Frame Identification feature (F/C 9123) and SP-attached Server Cable for 7026 feature (F/C 9125).

An appropriate length switch cable is required to attach each switch-attached server. This cable is ordered as a feature of the 9076 SP.

A minimum of one Ethernet adapter is required for each clustered server. This adapter must be recognized by the clustered server as EN0 and must reside in slot 1 of the primary I/O drawer. The supported adapters are:

- ▶ 10/100 Ethernet 10BaseTX adapter (F/C 2968)
- ▶ 10/100 Mbps Ethernet PCI Adapter II (F/C 4962)

The SP Switch Attachment Adapter (F/C 8396) is a 5 V adapter and must be located in slot #6 of the primary I/O drawer. Only one (F/C 8396) adapter is allowed per server.

The SP Switch Attachment Adapter (F/C 8397) is a universal PCI adapter. Two are allowed per 7026 server. The first (F/C 8397) adapter is always located in slot #5 of the primary I/O drawer. If two (F/C 8397) adapters are utilized, the second adapter must be placed in slot #3 and requires slot #4 to remain empty.

Graphics adapters and natively-attached displays are not supported in the clustered server environment. An ASCII terminal may be attached to servers in these environments, but is not required. System control functions are provided by the SP control workstation.

Additional information regarding server clustering and control workstations is available at the following IBM PSSP Web site:

http://www.ibm.com/servers/eserver/pseries/library/sp_books/planning.html

CSM/PSSP support statement

The following support statement details the CSM and PSSP support for the p660-6M1:

- ▶ The IBM @server 7026-6M1 is supported on CSM V1.3.1 with APAR IY42369 with AIX 5.1 ML04, 5.2 ML01, and later.

Note: The CSM management server must be running AIX 5L V5.2 with Recommended Maintenance package 5200-01 (APAR IY39795). The other machines within the cluster are referred to as “managed nodes” and can be running AIX 5L V5.2 with 5200-01, or V5.1 with the 5100-04 Recommended Maintenance package (APAR IY39794). They can also be xSeries machines running CSM for Linux V1.3.

For all AIX servers, the following CSM for AIX 5L 1.3.1 service is required:

- ▶ Added Service for RSCT on AIX 5.1 IY42782
 - ▶ Added Service for RSCT on AIX 5.2 IY42783
 - ▶ CSM for AIX 5L added service IY42353
 - ▶ CSM Support for p670 and p690 POWER4+ IY42356
 - ▶ CSM Support for 7026 servers, 9076 SP nodes and p650 IY42379
 - ▶ CSM 1.3.1 support for 9076 SP Node Features 2054/2058 IY42847
 - ▶ CSM Support of p655 POWER4+ IY42377
- ▶ The IBM @server p660-6M1 is supported on PSSP 3.5 with AIX 5.1 ML03 or later.



Network configuration

In this chapter, we look at the different networks in a cluster configuration, and explain the tasks relating to network devices and network configurations.

We cover the following topics:

Networking for Parallel System Support Programs (PSSP)

- ▶ SP LAN Ethernet
- ▶ Switch network
- ▶ Other networks
- ▶ Network considerations
- ▶ Sample scenarios for Cluster 1600 managed by PSSP

Networking for Cluster Systems Management (CSM)

- ▶ CSM hardware control
- ▶ Hardware and network requirements
- ▶ Virtual LANs (VLANs)
- ▶ pSeries HPS switch network overview
- ▶ Examples

3.1 SP LAN Ethernet

The SP LAN, or internal Ethernet, is a fundamental resource for the cluster system because it is involved in most system management operations. It is needed when nodes are being installed and customized and when their software components are being controlled. Without it, you cannot manage your nodes. Due to its importance, it is imperative that it remain functional and available.

The network topology for the SP/CES Ethernet mainly depends on the size of the system, and it should be planned on an individual basis. Since the SP/CES private Ethernet network is used for installation and administration purposes, we strongly recommend that additional network connectivity, such as the SP Switch, additional Ethernet, Token Ring, FDDI, or ATM networks should be used for the applications on the SP/CES, which performs significant communication among nodes. This can also avoid overloading the SP/CES Ethernet administration network with application traffic.

When the cluster is built, each node provides an Ethernet (en0) interface dedicated to PSSP system management operation. This is connected to the control workstation (CWS) and is used by PSSP for internal cluster communication. This normally is a private network, with nodes on this network having private addresses.

The PSSP software requires that the SP system is Internet protocol Version 4 (IPv4) for all interfaces. IPv6 is a version that extends addressing capability. PSSP components can tolerate IPv6 alias addresses coexistence with the required IPv4 network address for the Ethernet and Token Ring interfaces only. Understand and keep in mind that none of the PSSP software components actually uses IPV6 addresses.

Note: Do not add IPv6 aliases if your system uses SP switch, or DCE, HACWS, or HACMP licensed programs. Those products do not tolerate IPv6. It is important that you read and clearly understand the information concerning IPv6 in the publications for each IBM licensed program.

3.1.1 Supported Ethernet adapters and their placement

Ethernet adapters supported for cluster Ethernet communication are:

- ▶ Twisted-pair cable connection:
 - 10/100 Ethernet PCI adapter II F/C 4962)
 - 10 MB AUI/RJ-45 Ethernet adapter (F/C 2987)

- ▶ BNC cable connection:
 - 10 MB BNC/RJ-45 Ethernet adapter (F/C 2985)
- ▶ New I/O adapters
 - The IBM 2-Port 10/100/1000 Base-TX Ethernet PCI-X adapter
 - IBM 2-Port Gigabit Ethernet-SX PCI-X adapter

Note: The IBM 2-port Gigabit Ethernet-SX PCI-X adapter provides two full duplex Ethernet connections for Universal Twisted Pair (UTP) and short wave multimode optical cable, respectively.

Both adapters provide functions to support high performance communications, including Jumbo frames at 1000Mbps speed to help increase throughput with less processor utilization; Large Send, which offloads TCP segmentation from system software to the adapter; and Checksum offload, which allows the adapter to calculate TCP/UDP checksum, rather than system software

Attention: H80/M80 requires a 10/100 Ethernet PCI adapter II (F/C 4962).

Feature codes for the PCI Ethernet adapters are:

- ▶ F/C 4962 IBM 10/100 Mbps Ethernet PCI adapter
- ▶ F/C 2985 PCI Ethernet BNC/RJ-45 adapter
- ▶ F/C 2987 PCI Ethernet AUI/RJ-45 adapter
- ▶ F/C 2969 Gigabit Ethernet SX PCI adapter
- ▶ F/C 2975 10/100/1000 BASE-T Ethernet PCI adapter

Note: 10/100 Ethernet adapter (F/C 2968) is withdrawn. Twisted-pair cable connections require one F/C 9223 for each SP-attached server and BNC cable connection require F/C 9222 for each SP-attached server. It has to be ordered with the SP system.

The 2-Port Gigabit Ethernet and audio adapters are supported on AIX 5L v5.1 and v5.2.

Restriction: For the pSeries servers, the integrated Ethernet is not supported for the cluster LAN.

Placement

These adapters must be placed in the en0 position of the cluster nodes. That is the lowest-numbered Ethernet bus slot in the first I/O tower, so that it will always be en0. The following slots apply to 7017 and 7026:

- ▶ Slot 5 on a 7017 in the primary I/O drawer
- ▶ Slot 1 on a 7026 in the primary I/O drawer

If you plan to attach an RS/6000 server to your cluster system, you must place a cluster Ethernet adapter in the en0 position inside the RS/6000 server. Due to the fact that the Ethernet adapter in this slot must be configured for PSSP communications, any non-supported Ethernet adapter that is in the en0 slot must be removed.

For HMC-controlled pSeries, refer to 3.4.8, “HMC trusted network - considerations” on page 139.

3.1.2 Ethernet network topology

As mentioned in 3.1, “SP LAN Ethernet” on page 122, the network topology for the SP/CES Ethernet mainly depends on the size of the system. For that reason, it should be individually planned.

The SP/CES private Ethernet network is used for installation and administration purposes, so additional network connectivity should be used for applications on the SP/CES, which performs significant communication among nodes. In this way, you can also avoid overloading the SP/CES Ethernet administration network with application traffic.

For further information about this topology, refer to *IBM eServer Cluster 1600 Planning, Installation and Service Guide*, GA22-7863.

3.1.3 IP label convention

Each network interface within a cluster system requires a unique IP label. The label should include information such as frame number, node number, and interface type, in order to reduce management complexity.

The host name for each node within the cluster system should be set to the IP label for the cluster Ethernet LAN en0.

The naming convention can be follows:

sAFFNNnet

where:

- ▶ sA = RS/6000 system type and number
- ▶ FF = The frame number
- ▶ NN = the node number
- ▶ net = The network interface type (for example, en0 = Ethernet)

The cluster LAN is assigned from a private range of Class A or C network addresses reserved for use in private networks. For more information about IP addressing, refer to *IBM eServer Cluster 1600 Planning Volume 2, Control Workstation and Software Environment*, GA22-7821.

3.2 Switch network

In a Cluster configuration, you can have another type of network called the Switch network. This network requires a switch device that gets mounted in a SP frame.

Switches are used to connect processor nodes, and provide the message passing network through which they communicate with a minimum of four disjoint paths between any pair of nodes. In any cluster system, you can use only *one* type of switch, either the SP Switch2 or the SP Switch.

Cluster nodes are supported to be connected to the switch, but they require a tall SP frame for housing the switch. In order to connect to the switch, the cluster nodes require switch adapters. In the following section, we discuss the benefits of a switch network, relating to the SP Switch2 and the SP Switch.

3.2.1 Benefits of a Switch network

A switch provides low latency, high-bandwidth communication between nodes, supplying a minimum of four paths between any pair of nodes. Switches provide enhanced, scalable high performance communication for parallel job execution, and they dramatically speed up TCP/IP, file transfers, remote procedure calls, and relational database functions. Using a switch offers the following improved capabilities:

- ▶ Interframe connectivity and communication
- ▶ Scalability up to 256 node connections, including intermediate switch frames
- ▶ Constant bandwidth and latency between node pairs
- ▶ Support for Internet Protocol (IP) communication between nodes
- ▶ IP Address Resolution Protocol (ARP) support
- ▶ Support for dedicated or multi-user environments
- ▶ Error detection and retry

- ▶ High availability
- ▶ Fault isolation
- ▶ Concurrent maintenance for nodes
- ▶ Improved switch chip bandwidth

3.2.2 SP Switch2

The Switch2 feature code 4012 can be used to interconnect the cluster nodes to form a Switch network. This switch has 16 ports for node connections and 16 ports for switch-to-switch connections. Switch2, which has evolved from the traditional SP Switch, provides the following added features:

- ▶ Next generation of SP Switch and adapter
- ▶ Switch throughput increased (over a 3x increase in bandwidth)
 - 150 to 500 MB/sec one way
 - 300 to 1000 MB/sec for bi-directional
- ▶ 32 ports
 - 16 internal ports for connections to nodes within the switch-equipped frame, or to nodes in non-switched expansion frames
 - 16 external ports for switch-to-switch connections
- ▶ N+1 redundancy on
 - Hot-swappable power supplies
 - Hot-swappable cooling fans
- ▶ Hot-swappable switch supervisor card
- ▶ Hot-swappable switch interposers

Note: Any unused switch ports must have a blank interposer card installed to prevent contamination of the connector and ensure proper cooling air flow.

Figure 3-1 on page 127 shows the node and switch ports.

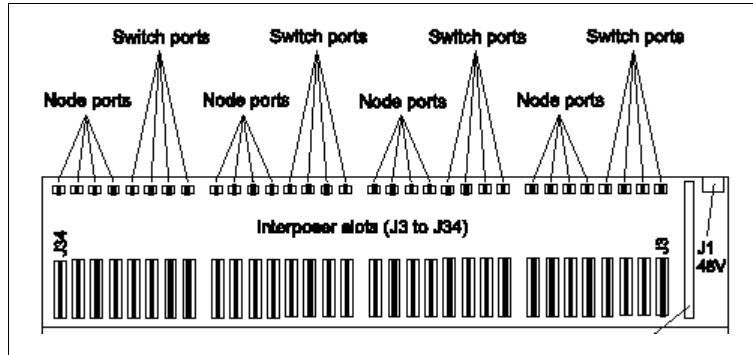


Figure 3-1 Switch2 allocation of node and switch ports (port side shown)

The functions of a switch are made available by the PSSP software. The functions of Switch2 were enhanced with the release PSSP 3.4. The following list describes these enhancements:

- The SP Switch2 supports two switch configurations, so that nodes can communicate over one or two switch planes.

By having two switch planes operational, you can improve communication performance and also achieve higher availability, since one switch plane can continue operating even when you take a switch down for maintenance.

- Optional switch connectivity.

A system using the SP Switch2 can have some nodes that are not connected to the switch. You can use the SP Switch2 and still keep older nodes in your system, but not connected to the switch.

- You can have up to 8 SP Switch2 node switch boards in a single SP frame.

This is advantageous for installations that require a large number of switch connections for SP-attached servers or clustered enterprise server configurations. This is a frame configured only with SP Switch2 switches.

- Relaxed node placement rules and optional connectivity.

Nodes can be placed anywhere allowed by the physical restrictions. The node switch port numbers are not predefined, so you can connect a node to any available switch port, and the numbers will be generated when the switch is started.

As a node is assigned to a CSS adapter, it is given the lowest available switch node number from zero (0) through 511. There is no correlation between the switch port number and any hardware connections cluster of two-plane Switch2 networks.

3.2.3 Switch IP network and addressing

If using IP for communication over the switch, each node needs to have an IP address and name assigned for the switch interface (the css0 adapter).

If you plan to use the Switch2 with two switch planes, you also need to have an IP address and name assigned for the css1 adapter, and you have the option to use the ml0 aggregate IP interface. If hosts outside the switch network need to communicate over the switch using IP with nodes in the cluster system, those hosts must have a route to the switch network through one of the cluster nodes.

If you use the Switch2 and all nodes are running PSSP 3.4 and PSSP 3.5, you have optional connectivity, so some nodes can be left off the switch.

Switch port numbering is used to determine the IP address of the nodes on the switch. If your system is not ARP-enabled on the css0 adapter and the css1 adapter in a two-plane Switch2 system, choose the IP address of the first node on the first frame. The switch port number is used as an offset added to that address to calculate all other switch IP addresses.

If ARP is enabled for the css0 and css1 adapters, the IP addresses can be assigned like any other adapter. That is, they can be assigned beginning and ending at any node. They do not have to be contiguous addresses for all the css0 and css1 adapters in the system. Switch port numbers are automatically generated by the Switch2, and so are not used for determining the IP address of the Switch2 adapters.

With the introduction of Switch2 and PSSP 3.4, you are able to build a switched cluster with some nodes that do not have to belong to the switch network. Configuring such a cluster is no different from constructing a standard switched or unswitched cluster.

Tip: Configure the *unswitched* nodes as if you were configuring the nodes *without* the switch. When adding the additional adapter information about the unswitched nodes, use the “Node list” option.

Configure the *switched* nodes as if you were configuring the nodes *with* the switch. When adding the additional adapter information about the switched nodes, use the “Node list” option.

3.3 Other networks

Once you have configured the cluster LAN and the optional Switch network, you can configure other networks that allow you to use the nodes. For example, you

could have high speed Ethernet adapters in the nodes. This network connection allows users to access the application running on the nodes.

The configuration of these additional adapters is quite straightforward; however, you need to ensure it is done in the CWS and propagated to the nodes, in order to keep the SDR consistent.

If you configure these adapters *within* the node, PSSP's SDR residing in the CWS will not know about it—and if you reconfigure or customize the node from the CWS in the future, you could end up deleting or overwriting the network configuration in the nodes.

To avoid this situation, keep the following points in mind when choosing to configure additional adapters in your cluster nodes:

- ▶ All configuration should be carried out through PSSP on the CWS.
- ▶ Set the right initial host name on the node so that applications can find this host.
- ▶ Use the `tuning.cust` script in PSSP to set the right routes for this additional adapter network.
- ▶ If you are installing an application on this node, it should be done *after* installing the external adapter and setting the initial host name.

3.4 Network considerations

In the following sections, we highlight considerations and other information that will help you when building the network for the Cluster 1600. These sections provide specific information on the Cluster 1600 managed by PSSP and its subsystems.

3.4.1 The RS-232 connection

In classic SP environments, the CWS has a direct connection to each SP frame and to each SP-attached server, to perform the following tasks:

- ▶ Controlling the hardware (for example, power on/power off)
- ▶ Transferring the Kerberos tickets and password files
- ▶ Communicating with the firmware (for example, node conditioning)

Note: In a pSeries-attached configuration, there is no direct RS-232 connection between the CWS and pSeries. The CWS is connected to the HMC via Ethernet, and the HMC is connected to the pseries server using an RS-232/RS-422 serial line.

3.4.2 System topology considerations

When configuring larger systems, you need to consider several areas when setting up your network: the Cluster 1600 Ethernet, the outside network connections, the routers, the gateways, and the switch traffic.

The Cluster 1600 Ethernet is the network that connects a control workstation to each of the nodes in the cluster that are to be operated and managed by that control workstation using PSSP. When configuring the cluster Ethernet, the most important consideration is the number of subnets you configure. Because of the limitation on the number of simultaneous network installs, the routing through the cluster Ethernet can be complicated. Usually the amount of traffic on this network is low.

If you connect the cluster Ethernet to your external network, you must make sure that the user traffic does not overload the SP Ethernet network. If your outside network is a high speed network like FDDI or HIPPI, routing the traffic to the cluster management Ethernet can overload it. For gateways to FDDI and other high speed networks, route traffic over the Switch network. Configure routers or gateways to distribute the network traffic so that one network or subnet is not a bottleneck. If the cluster management Ethernet is overloaded by user traffic, move the user traffic to another network.

If you expect a lot of traffic, then configure several gateways

3.4.3 Boot/install server requirements

When planning your management LAN topology, consider your network install server requirements. The network install process uses the admin/management Ethernet for transferring the install image from the install server to the cluster nodes. Running numerous, concurrent network installs can exceed the capacity of the admin LAN.

Following are suggested guidelines for designing the Ethernet topology for efficient network installs. Many of the configuration options will require additional network hardware, beyond the minimal node and control workstation requirements. There are also network addressing and security issues to consider.

HMC-controlled servers get network-connected and do not require you to use en0. For all other nodes, you must use the en0 adapter to connect each node to the SP Ethernet admin LAN. The following requirements pertaining to the cluster Ethernet admin LAN exist for all configurations:

- ▶ The NIM clients that are served by boot/install servers must be on the same subnet as the boot-install server's Ethernet adapter.
- ▶ NIM clients must have a route to the control workstation over the SP Ethernet.
- ▶ The control workstation must have a route to the NIM clients over the SP Ethernet.

Note: Keep the following points in mind regarding boot/install servers.

- ▶ Certain security options have limitations with multiple boot/install servers. See *Limitations in IBM eServer Cluster 1600 Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.
- ▶ Do not install cascading levels of boot/install servers. Every boot/install server node must have the control workstation as its boot/install server.

Single frame systems

For small systems, you can use the control workstation as the network install server. This means that the SP Ethernet admin LAN is a single network connecting all nodes to the control workstation. When installing the nodes, limit yourself to installing eight nodes at a time, because that is the limit of acceptable throughput on the Ethernet.

An alternate way to configure your system is to install a second Ethernet adapter in your control workstation, if you have an available I/O slot, and use two Ethernet segments to the SP nodes. Connect each network to half of the SP nodes. When network-installing the frame, you can install all 16 nodes at the same time.

Note: Set up your SP Ethernet admin LAN routing so nodes on one Ethernet can communicate to nodes on the other network. Set up your network mask so that each SP Ethernet is its own subnet within a larger network address.

Multiple frame systems

For multiple frame systems, you might want to spread the network traffic over multiple Ethernets, and keep the maximum number of simultaneous installs per network to eight.

You can use the control workstation to network-install specific SP nodes to be the network install servers for the rest of nodes. There are three ways to accomplish this:

- Use a control workstation with one Ethernet adapter for each frame of the system, and one associated SP Ethernet per frame.

For example, if you have a system with four frames (as in Figure 3-2 on page 132), the control workstation must have enough I/O slots for four Ethernet adapters. Each adapter connects one of the four SP frame Ethernet segments to the control workstation.

Using this method, you install the first eight nodes on a frame at a time—or up to 32 nodes, if you use all four Ethernet segments simultaneously. Running two installs will install up to 64 nodes.

Note: Set up your SP Ethernet routing so nodes on one Ethernet can communicate with nodes on another. Set up your network mask so that each SP Ethernet is its own subnet within a larger network address. This method is applicable up to the number of slots your control workstation has available.

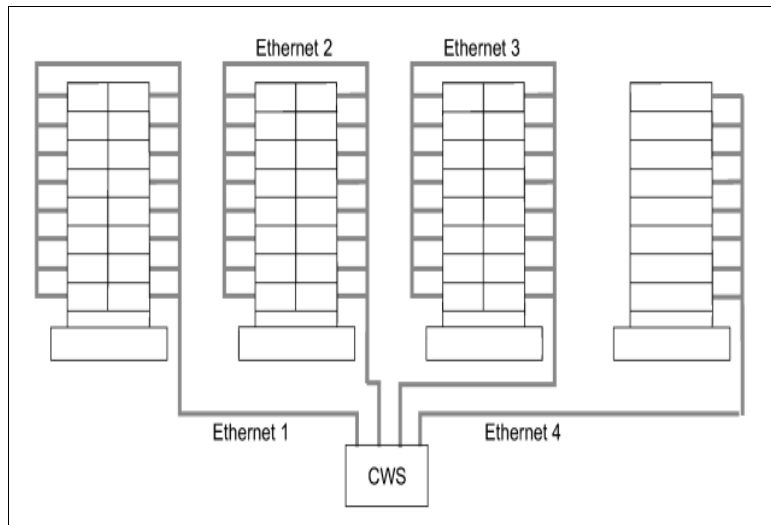


Figure 3-2 CWS with one Ethernet adapter for each frame

- A second approach designates the first node in each frame as a network install server, and then the remaining nodes of that frame are set to be installed by that node.

This means that, from the control workstation, you will have an SP Ethernet segment connected to one node on each frame. Then the network install

node in each frame has a second Ethernet card installed, which is connected to an Ethernet card in the rest of the nodes in the frame.

When using this method, as shown in Figure 3-3, installing the nodes requires that you first install the network install node in each frame. The second set of installs will install up to eight additional nodes on the frame.

The last install, if needed, installs the rest of the nodes in each frame. Be forewarned that this configuration usually brings performance problems due to two phenomena:

- All SP Ethernet admin LAN traffic (for example, installs, and SDR activity) is routed through the control workstation. The single control workstation Ethernet adapter becomes a bottleneck, eventually.

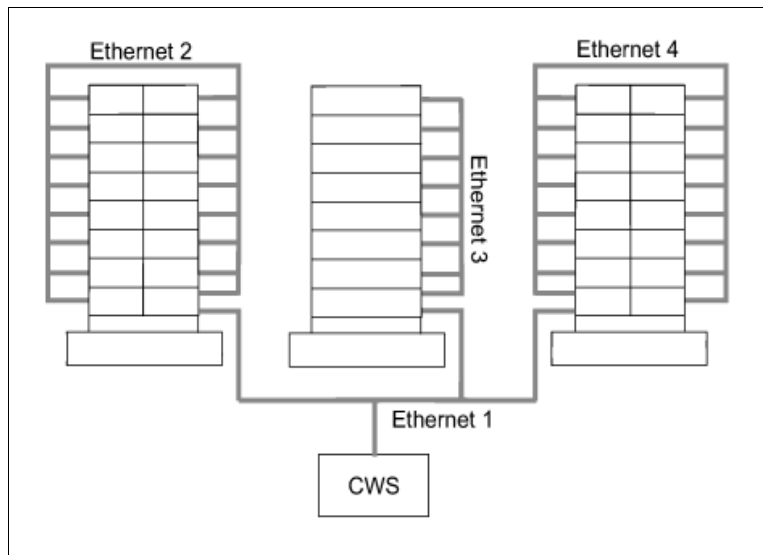


Figure 3-3 CWS with one Ethernet card; one node has two adapters

- An application running on a node which produces a high volume of SP Ethernet traffic (for example, LoadLeveler®) causes all subnet routing to go through the one control workstation Ethernet adapter. Moving the subject application to the control workstation can cut that traffic in half, but the control workstation must have the capacity to accommodate that application. You can improve the performance here by adding an external router, similar to that described in method 3.
- A third method adds an external router to the topology of the previous approach.

This router is made part of each of the frame Ethernets, so that traffic to the outside does not need to go through the control workstation. You can do this

only if the control workstation can also be attached externally, providing another route between nodes and the control workstation. Figure 3-4 shows this Ethernet topology for such a multi-frame system.

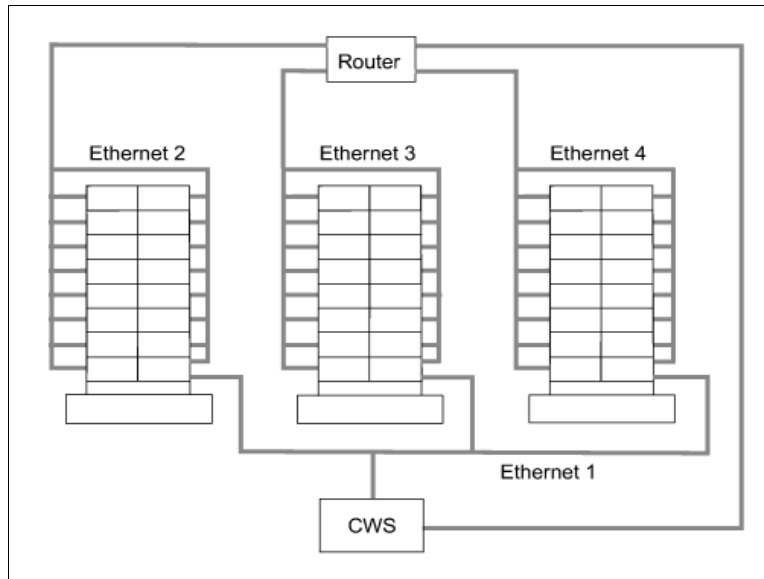


Figure 3-4 Ethernet topology for multi-frame system

3.4.4 The SP Ethernet administrative LAN

The SP Ethernet is the administrative LAN that connects all nodes in one system running PSSP to the control workstation. The SP LAN is essential for the Cluster 1600 and for the following PSSP software management tasks:

- ▶ RSCT communications (for example, hosts responds)
- ▶ NIM operations
- ▶ Systems management communications

Note: The I/O drawer of the pSeries server does not support the BNC-type network adapters. If you want to integrate the pSeries server into an existing Cluster 1600 with a BNC SP LAN, you will need a HUB in order for the SP LAN to provide BNC and Twisted pair (TP) cabling in one network.

For HMC-controlled servers, the SP LAN can also be used to connect the HMC to the control workstation for hardware control and monitoring; refer to 3.4.8, “HMC trusted network - considerations” on page 139, for more information.

For each node, ensure that the SDR `reliable_hostname` attribute is identical to the default host name returned by the `host` command for its SP Ethernet IP addresses. With the exception of HMC-controlled server nodes, the PSSP components expect that `en0` is the connection from the node to the SP Ethernet admin LAN for installs and other PSSP functions.

For an HMC-controlled server node, you can use any Ethernet adapter that is supported for connecting to the cluster Ethernet admin LAN, and identify it by name or by physical location. For all other nodes, you must connect the cluster Ethernet admin LAN to the Ethernet adapter in the node's lowest hardware slot of all the Ethernet adapters on that node.

When a node is network-booted, it selects the lowest Ethernet adapter from which to perform the install. This Ethernet adapter must be on the same subnet of an Ethernet adapter on the node's boot/install server. In nodes that have an integrated Ethernet adapter, it is always the lowest Ethernet adapter. Be sure to maintain this relationship when adding Ethernet adapters to a node. You can attach the cluster admin Ethernet to other site networks and use it for other site-specific functions. You assign all addresses and names used for the cluster Ethernet admin LAN.

You can make the connections from the control workstation to the nodes in one of three ways. The method you choose should be one that optimizes network performance for the functions required of the cluster admin Ethernet LAN by your site.

The three connection methods are:

- ▶ Single-subnet, single-stage cluster admin Ethernet, in which one interface on the control workstation connects to all cluster nodes.
- ▶ Multi-subnet, single-stage cluster admin Ethernet. In this method, there is more than one interface on the control workstation, and each connects to a subset of the cluster admin nodes.
- ▶ Multi-subnet, multi-stage cluster admin Ethernet. In this method, a set of nodes, acting as routers to the remaining nodes on separate subnets, connects directly to the control workstation.

The cluster's boot/install servers must be on the same subnet as their clients. In the case of a multi-stage, multi-subnet cluster Ethernet admin LAN, the control workstation is the boot/install server for the first node in each frame, and those nodes are the boot/install servers for the other nodes in the frames.

Also, when booting from the network, nodes broadcast their host request over their Ethernet admin LAN interface. Therefore, that interface must be the Ethernet adapter on the node that is connected to the boot/install network.

The IBM eServer pSeries introduces a new way to configure the SP LAN adapter. It is now recommended to configure the SP LAN adapter through its unique hardware location code. For further details on this subject, refer to *IBM eServer Cluster 1600: Planning, Installation and Service Guide*, GA22-7863.

Note: For HMC-controlled servers, you are no longer required to have the first Ethernet adapter (en0) configured as the SP LAN adapter. PSSP uses the hardware physical location codes to uniquely define the Ethernet adapter.

Tip: To avoid problems later on, we recommend that you use unique domain names not only for each HMC, but also in the Cluster 1600 environment.

Node Class

Since PSSP represents each LPAR as a node, the LPAR name is stored in the Node Class of the SDR. The Node Class is described in Table 3-1.

Table 3-1 List with new and enhanced attributes in the Node Class

Attribute name	Description	p690 values
LPAR_name	Logical partition identifier	Retrieved by hardmon. It is the same name as shown on the HMC in the Partition Management menu.

Adapter Class

The new feature to define an SP LAN adapter through its hardware location code requires the new `physical_location` attribute in the SDR Adapter Class. Since the SP LAN adapter can be different than en0, it needs to define explicitly which adapter belongs to the SP LAN. The Adapter Class is described in Table 3-2.

Table 3-2 List with new and enhanced attributes in the Adapter Class

Attribute name	Description	p690 values
<code>physical_location</code>	Physical location code for the adapter.	The value as shown by the <code>spadapttr_loc</code> or <code>lscfg -vl</code> command.
SP LAN	Indicates whether or not this adapter is connected to the SP admin network (SP LAN).	A value of 1 indicates that this is an SP LAN adapter. A value of 0 means that this adapter belongs to a different network. Only one adapter per node can have this value set to 1.

Important: There is no specific plug-in rule for the SP LAN Ethernet adapter placement. To see which Ethernet adapters are supported as SP LAN adapters, refer to *RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*, GA22-7280.

For further information about adapter placements, refer to *RS/6000 and eServer pSeries: PCI Adapter Placement Reference*, SA38-0538.

3.4.5 Additional LANs - considerations

The cluster admin LAN can provide a means to connect all nodes and the control workstation to your site networks. However, it is likely that you will want to connect your SP nodes to site networks through other network interfaces.

If the cluster admin LAN is used for other networking purposes, the amount of external traffic must be limited. If too much traffic is generated on the SP Ethernet admin LAN, the administration of the SP nodes might be severely impacted. For example:

- ▶ Problems might occur with network installs, diagnostic function, and maintenance mode access. In an extreme case, if too much external traffic occurs, the nodes will hang when broadcasting for the network.
- ▶ Additional Ethernet, Fiber Distributed Data Interface (FDDI), and Token-Ring networks can also be configured by the PSSP software. Other network adapters must be configured manually. These connections can result in increased network bandwidth.
- ▶ Performance in user file serving and other network-related functions may be reduced. You need to assign all the addresses and names associated with these additional networks.

Note: You can plan to run PSSP on a system with a firewall. For more information, see *Implementing a Firewalled RS/6000 SP System*, GA22-7874.

3.4.6 IP over the switch - considerations

If your SP has a switch and you want to use IP for communications over the switch, each node needs to have an IP address and name assigned for the switch interface, the css0 adapter. If you plan to use the SP Switch2 with two switch planes, you also need to have an IP address and name assigned for the css1 adapter, and you have the option to use the ml0 aggregate IP address. If hosts outside the SP switch network need to communicate over the switch using

IP with nodes in the SP system, those hosts must have a route to the switch network through one of the SP nodes.

If you are not enabling ARP on the switch, specify the switch network subnet mask and the IP address. Unlike all other network interfaces, which can have sets of nodes divided into several different subnets, the SP switch IP network must be one contiguous subnet that includes all nodes in the system. (If you use the SP Switch2 and all nodes are running PSSP 3.4 or greater, you have optional connectivity so some nodes can be left off the switch.)

- ▶ If you want to assign your switch IP addresses as you do your other adapters, you must enable ARP for the css0 adapter and, if you are using two switch planes, for the css1 adapter. If you enable ARP for those interfaces, you can use whatever IP addresses you wish, and those IP addresses do not have to be in the same subnet for the whole system. They must all be resolvable by the host command on the control workstation.

For more information about this subject, refer to *IBM eServer IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*, GA22-7280.

3.4.7 Subnetting - considerations

All but the simplest SP system configurations likely include several subnets. Thoughtful use of netmasks in planning your networks can economize on the use of network addresses. For more information about Internet addresses and subnets, refer to *AIX 5L Version 5.1 System Users Guide: Communications and Networks*, which can be found at:

http://publib16.boulder.ibm.com/pseries/en_US/aixuser/usrcomm/usrcommtfrm.htm

As an example, consider an SP Ethernet where none of the six subnets making up the SP Ethernet have more than 16 nodes on them. A netmask of 255.255.255.224 provides 30 discrete addresses per subnet, which is the smallest range that is usable in the wiring as shown. Using 255.255.255.224 as a netmask, we can then allocate the address ranges as follows:

- ▶ 129.34.130.1-31 to the control workstation to node 1 subnet
- ▶ 129.34.130.33-63 to the frame 1 subnet
- ▶ 129.34.130.65-96 to frame 2

In the same example, if we used 255.255.255.0 as our netmask, then we would have to use six separate Class C network addresses to satisfy the same wiring configuration (that is, 129.34.130.x, 129.34.131.x, 129.34.132.x, and so on).

3.4.8 HMC trusted network - considerations

Cluster 1600s that are managed by PSSP configurations which contain HMC-controlled servers require additional security planning and configuration to protect password data transferred from the control workstation to an HMC-controlled server. Cluster 1600 configurations managed by PSSP that do not contain HMC-controlled servers do not require additional security planning and configuration.

RMC LAN

The Resource Monitoring and Control (RMC) LAN is an Ethernet connection between the HMC and the pSeries server in an LPAR or SMP mode. Over this connection the HMC gathers data from active LPARs or SMP servers about the health of the system, and acts as a service focal point for service representatives to determine an appropriate service strategy. For more information about the RMC LAN, refer to *HMC Operations Guide*, SA38-0590.

Important: Use of the RMC LAN is not mandatory. It is possible to use the function of the RMC LAN over the SP LAN.

The trusted LAN

Since there is no direct RS-232/RS-422 connectivity between the HMC and the CWS, it is mandatory to have a trusted network connection between the HMC and the CWS to transfer Kerberos tickets and password files. A separate trusted LAN is recommended, but not necessary. If your SP LAN is secure, you do not need to set up an additional trusted LAN connection. You can send your secure sensitive data over the existing SP LAN.

A trusted network

In a *trusted network*, all hosts on the same network (LAN) are regarded as trusted, according to site security policies and procedures governing the hosts. Data on a trusted network can be seen by all trusted hosts and users on the trusted hosts, but the implied trust among and between the hosts assumes that the data will not be intercepted or modified. By way of implied mutual trust, traffic flowing across the trusted network is regarded as safe from unwanted or unintended interception or tampering. However, it does not imply that the data on the trusted network is itself private or encrypted.

You need a trusted network that best suits your environment:

- If you consider the SP Ethernet admin LAN to be a trusted network, no additional setup or configuration is needed to satisfy the requirement for security.

- If you do not consider the SP Ethernet admin LAN to be a trusted network, then you must establish another trusted network between the control workstation and the HMC, in order to satisfy your security requirement.

The remainder of this discussion will help you to plan a trusted network between the control workstation and the HMC.

A trusted network between the control workstation and an HMC requires that both hosts are connected via a network *other* than the SP Ethernet admin LAN. This other network is referred to as the *HMC trusted network*. We strongly suggest that you connect only the SP control workstation and HMC-controlled server systems to the HMC trusted network.

The HMC trusted network requires a set of IP addresses all in the same subnet. We suggest that you reserve the subnet for use only by hosts that will be connected to the HMC trusted network. If you plan on connecting multiple control workstations and multiple HMC systems to the HMC trusted network, then ensure that the subnet can accommodate the total number of trusted hosts (actual or expected).

In order for a control workstation to be connected to both the SP Ethernet admin LAN and the HMC trusted network, each control workstation requires the installation and configuration of an additional network adapter to be configured for the HMC trusted network.

Each HMC must have a network adapter configured for, and connected to, the HMC trusted network. Connecting an HMC to the HMC trusted network will require the reconfiguration of an existing HMC network adapter. How the existing HMC network adapter is reconfigured depends on whether your HMC-controlled server is a new SP-attached server install, or an existing SP-attached server.

New SP-attached server install

When your HMC-controlled server is already installed and configured, but not yet connected to your SP system, it is a new SP-attached server install. Therefore, you must reconfigure the existing network adapter in the HMC for the HMC trusted network.

Existing SP-attached server

When your HMC-controlled server is already an SP-attached server, the SP Ethernet admin LAN is used to connect the HMC to the control workstation. Specifically, the network adapter in the HMC is configured for the SP Ethernet admin LAN.

In this case, you must un-configure the network adapter in the HMC, disconnect it from the SP Ethernet admin LAN, connect it to the HMC trusted network, and then configure it for the HMC trusted network.

Configuration examples

In this section, we show different possible network connections. Figure 3-5 shows the RMC LAN with SP LAN and trusted network.

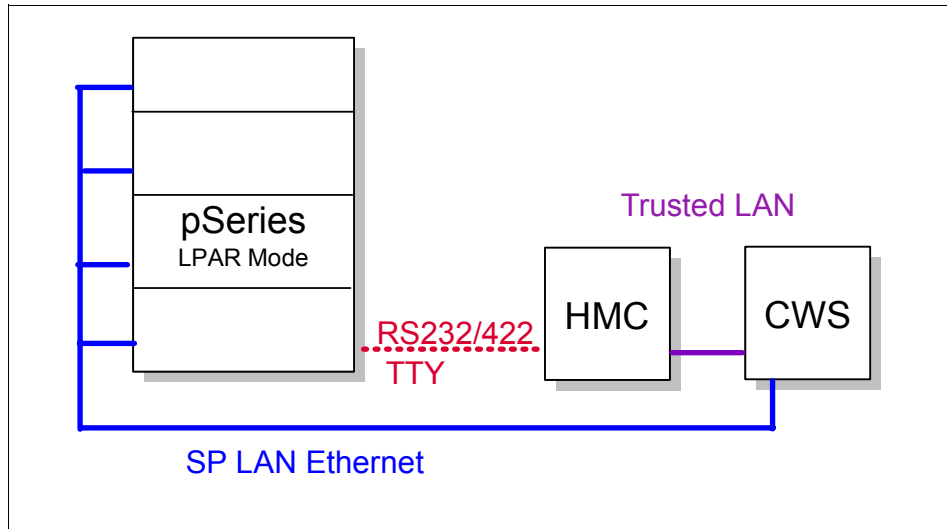


Figure 3-5 The RMC LAN uses the SP LAN and the trusted network

Figure 3-6 on page 142 shows a cluster with one physical network for the RMC, the SP LAN, and the trusted LAN.

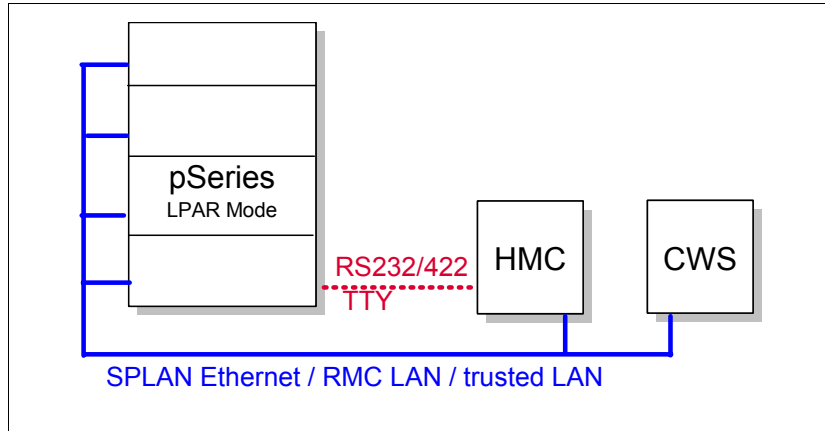


Figure 3-6 Cluster with only one physical network

Figure 3-7 shows two physical networks: one for the SP LAN, and one for the trusted network. The RMC LAN uses the SP LAN.

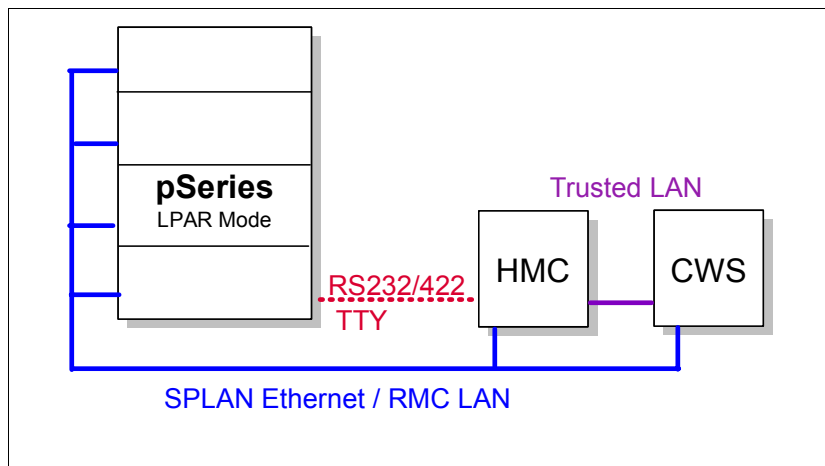


Figure 3-7 Cluster with two physical networks

Figure 3-8 on page 143 shows three separate physical networks for the SP LAN, the RMC LAN, and the trusted network.

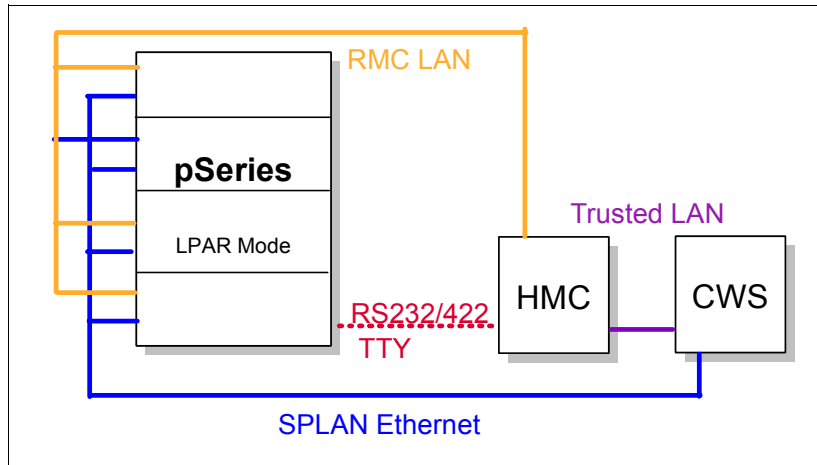


Figure 3-8 Each function has its own physical network

In Figure 3-9, you can attach the pSeries to the IBM eServer Cluster 1600 either in full system partition mode (also known as SMP mode), or in LPAR mode. Therefore, the configuration is different for each mode.

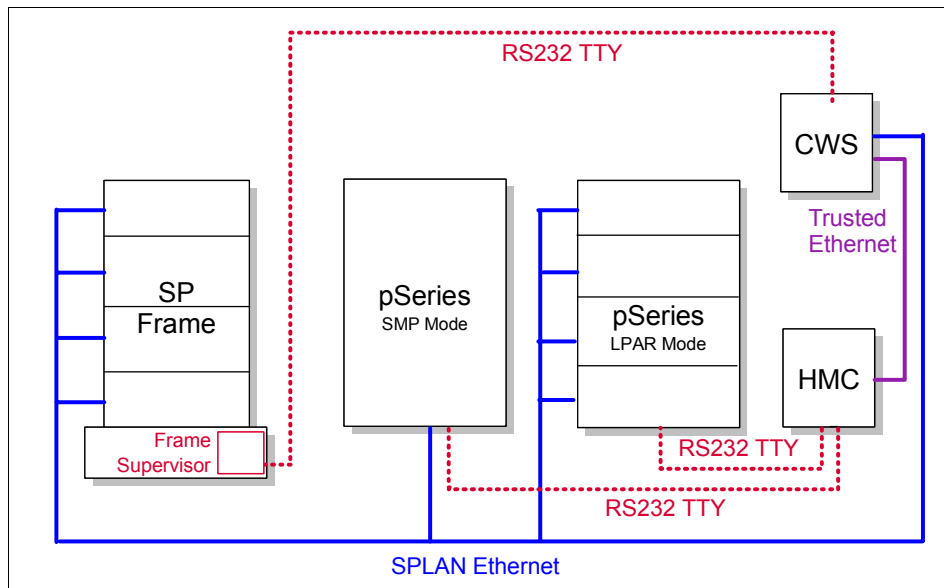


Figure 3-9 pSeries attached in Cluster 1600

3.4.9 Network router node considerations

If you plan to use an SP Switch Router and the SP Switch Router Adapter for routing purposes in your environment, the next few paragraphs on using standard nodes as a network router might not be applicable to your SP configuration. However, if you are not using the SP Switch Router, you might be interested in some considerations for using your nodes as network routers.

When planning router nodes on your system, several factors can help determine the number of routers needed and their placement in the SP configuration. The number of routers you need can vary, depending on your network type. (In some environments, router nodes might also be called “gateway nodes”.)

- ▶ For nodes that use low bandwidth networks (such as 10 Mb Ethernet or token ring) as the routed network, a customer network running at full bandwidth results in a lightly loaded CPU on the router node.
- ▶ For nodes that use high bandwidth networks (such as GB Ethernet or FDDI) as the customer routed network, a customer network running at or near maximum bandwidth results in high CPU utilization on the router node.

For this reason, you should not assign any additional role in the computing environment (such as a node in a parallel job) to a router using a high bandwidth network as the customer network. You also should not connect more than one high bandwidth network card to a router node.

Applications, such as POE, should run on nodes other than high bandwidth routers. However, low bandwidth gateways can run with these applications.

For systems that use low bandwidth routers, traffic can be routed through the SP Ethernet, but careful monitoring of the SP Ethernet will be needed to prevent traffic coming through the router from impacting other users of the SP Ethernet. For high bandwidth networks, traffic should be routed across the switch to the destination nodes. The amount of traffic coming in through the high bandwidth network can be up to 10 times the bandwidth the SP Ethernet can handle.

For more information about configuring network adapters and the various network tunables on the nodes, refer to *PSSP 3.5 Administration Guide*, SA22-7348.

SP Switch Router considerations

The SP Switch Router is something you can use in a system with the SP Switch. It is not supported with the SP Switch2. It is by type an extension node, more specifically a dependent node. The SP Switch Router gives you high speed access to other systems. Without the SP Switch Router, you would need to dedicate a standard node to performing external network router functions. Also,

because the SP Switch Router is external to the frame, it does not take up valuable processor space.

The SP Switch Router has two optional sizes. The smaller unit has four internal slots, and the larger unit has sixteen. One slot must be occupied by an SP Switch Router Adapter card, which provides the SP connection. The other slots can be filled with any combination of network connection cards, including the following types:

- ▶ Ethernet
- ▶ FDDI
- ▶ ATM
- ▶ SONET
- ▶ HIPPI
- ▶ HSSI

Additional SP Switch Router Adapters

Additional SP Switch Router Adapters are needed for communicating between system partitions and other SP systems. These cards provide switching rates of from 4 GB to 16 GB per second between the router and the external network.

To attach an extension node to an SP Switch, configuration information must be specified on the control workstation.

Communication of switch configuration information between the control workstation and the SP Switch Router takes place over the SP system's administrative Ethernet and requires use of the UDP port number 162 on the control workstation.

If this port is in use, a new communication port will have to be configured into both the control workstation and the SNMP agent supporting the extension node. You can improve throughput of data coming into and going out of the SP system by using the SP Switch Router.

The SP Switch Router can be connected with the SP Switch Router Adapter to an SP Switch, 8-port or 16-port. Each SP Switch Router Adapter in the SP Switch Router requires a valid, unused switch port in the SP system.

3.4.10 Clustered server configuration considerations

A “clustered” server is any of the AIX servers discussed in “Cluster 1600 hardware” on page 11. It is not mounted in an SP frame and has no SP frame or node supervisor, though some do have comparable function-enabling hardware control and monitoring. A clustered server is directly attached to the SP Ethernet

admin LAN and to the control workstation. The means of connection differ with the server hardware.

In a clustered server system configuration, you can assign frame numbers in any order. However, if you add SP frames with SP nodes or with SP switches, your system will then be subject to all the rules of an SP system, and these clustered servers become SP-attached servers. (Remember that those terms reflect only the system configuration in which the servers participate.)

If you might use the SP Switch2 or the SP Switch, you need to plan the respective switch network. If you might use the SP Switch, plan your system with suitable frame numbers and switch port numbers in advance so you can expand to an SP system without having to totally reconfigure existing servers.

For information on switch port numbering for a switchless system, refer to *IBM eServer Cluster 1600 Planning Volume 1, Hardware and Physical Environment*, GA22-7280.

3.4.11 SP-attached server considerations

An “SP-attached server” is any of the servers discussed in “Cluster 1600 hardware” on page 11. If the SP system has the SP Switch, an SP-attached server requires an available node slot within an SP frame to reserve a valid unused switch port on the switch in the same SP frame.

An SP-attached server is not supported with an SP Switch-8. You must connect the server to the SP Switch network with an adapter. That adapter connects to the valid unused switch port in the SP frame. For an HMC-controlled server node, each LPAR attaches to the switch. For more information on how to choose a valid port on the SP Switch, refer to *SP Switch Router Adapter Guide*, GA22-7310.

SP system partitioning is supported by default in switchless systems with at least one SP node frame. In that case, the number of SP-attached servers you can have is limited by the number of available switch ports.

But what if you have no need for multiple SP system partitions, but *do* need more SP-attached servers than are accommodated by the available switch ports in your system? Then you can force a switchless SP system to be non-partitionable. In this way, you can have more SP-attached servers because you can ignore the switch port numbering rules and assign sequential numbers. For more information about this topic, refer to *PSSP 3.5 Administration Guide*, SA22-7348.

If the system has the SP Switch2, an SP-attached server and each HMC-controlled server LPAR node can connect with an adapter to any available switch port in the SP frame. If all nodes are running PSSP 3.4 or greater, you can leave some nodes not connected to the switch.

You must assign a frame number to an SP-attached server. Be sure to read and understand the information regarding SP-attached servers in *IBM eServer Cluster 1600 Planning Volume 1, Hardware and Physical Environment*, GA22-7280.

3.5 Sample scenarios of Cluster 1600 managed by PSSP

In this section, we show a few sample Cluster 1600 scenarios using pSeries, HMC, and the CWS.

3.5.1 CWS with two HMCs and four pSeries

Figure 3-10 shows four pSeries servers in LPAR mode, with all nodes connected to the SP LAN. To achieve high availability, we use two HMCs. Each HMC is connected to all pSeries via RS-232 connections. The connection between the CWS and the HMC is via a trusted Ethernet connection.

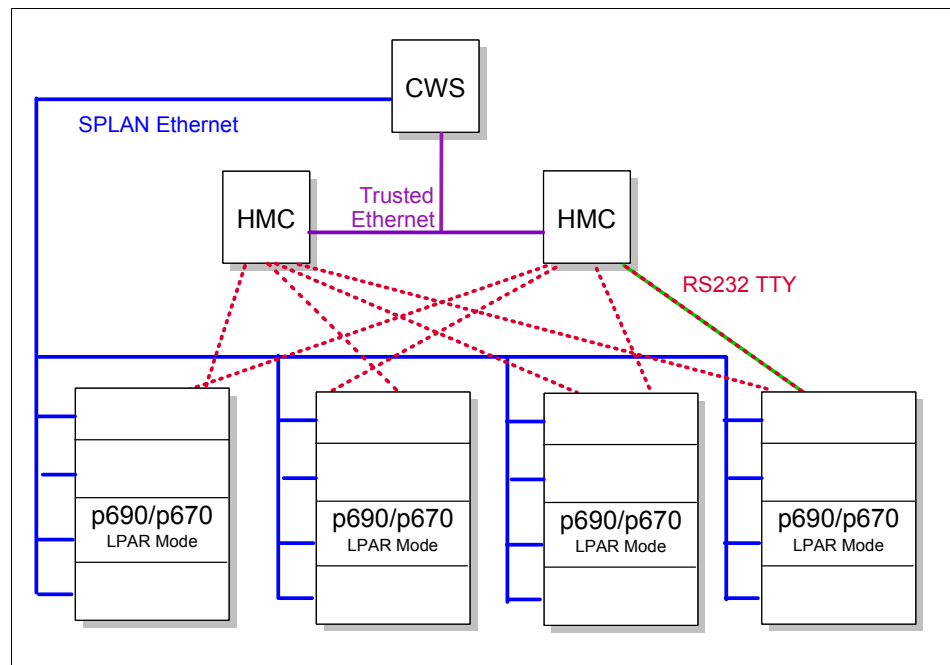


Figure 3-10 Four pSeries with two redundant HMCs and one CWS

To avoid having the HMC become, potentially, a single point of failure, we recommend that you connect each pSeries server to two different HMCs. That

way, in the event of an HMC failure, you are still capable of reaching the pSeries through the second HMC connection. Depending on how many pSeries servers you have, additional 8-port asynchronous adapters may be required.

Tip: For better performance when installing the nodes, make sure that when defining the pSeries frames (using the **spframe** command), you change the order of the HMC IP addresses.

3.5.2 One CWS, one HMC and one pSeries

Figure 3-11 shows the smallest possible Cluster 1600 based on one pSeries.

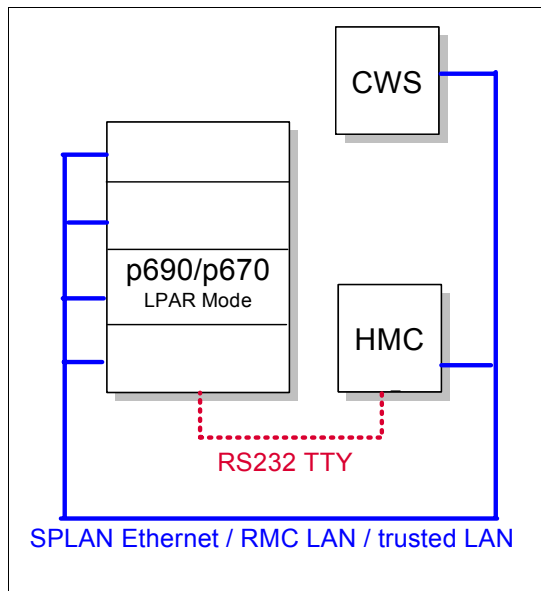


Figure 3-11 Cluster 1600 with one HMC, one pSeries and CWS

A useful upgrade for this scenario could be a trusted Ethernet between the HMC and the CWS.

Note: Keep the following in mind:

- Depending upon the pSeries type and model, the serial connectivity will be RS-232 or RS-422.
- In an SP Switch or a switchless environment, you must define a switch node number for each LPAR in the pSeries server. You can do that by manually editing the `/etc/switch.info` file. This file must include one line entry for each attached LPAR.

You can skip this configuration step if you are using SP Switch2 only, or you have a switchless Clustered Enterprise Server system (CES).

3.5.3 CWS, two HMCs, two 9076s, with one pSeries and SP Switch2

Figure 3-12 shows a Cluster 1600 after integrating a pSeries server in LPAR mode. Not all LPARs have a connection to the SP Switch2. This scenario shows a redundant HMC with a separate trusted LAN between the HMC and the CWS.

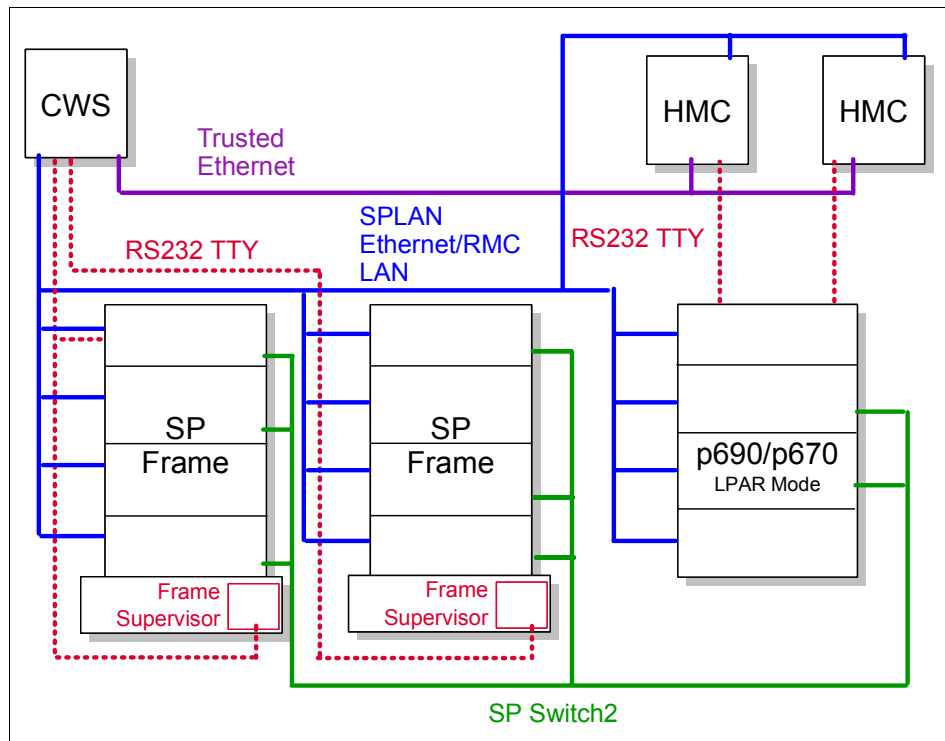


Figure 3-12 SP Frames with one pSeries, two LPARs on an SP Switch2

Note: Depending upon the pSeries type and model, the serial connectivity will be RS-232 or RS-422.

3.5.4 CWS, HMC, 9076 frame and pSeries with SP Switch

Figure 3-13 shows a pSeries in LPAR mode where all LPARs have an SP Switch connection. The RMC LAN traffic goes over the SP LAN.

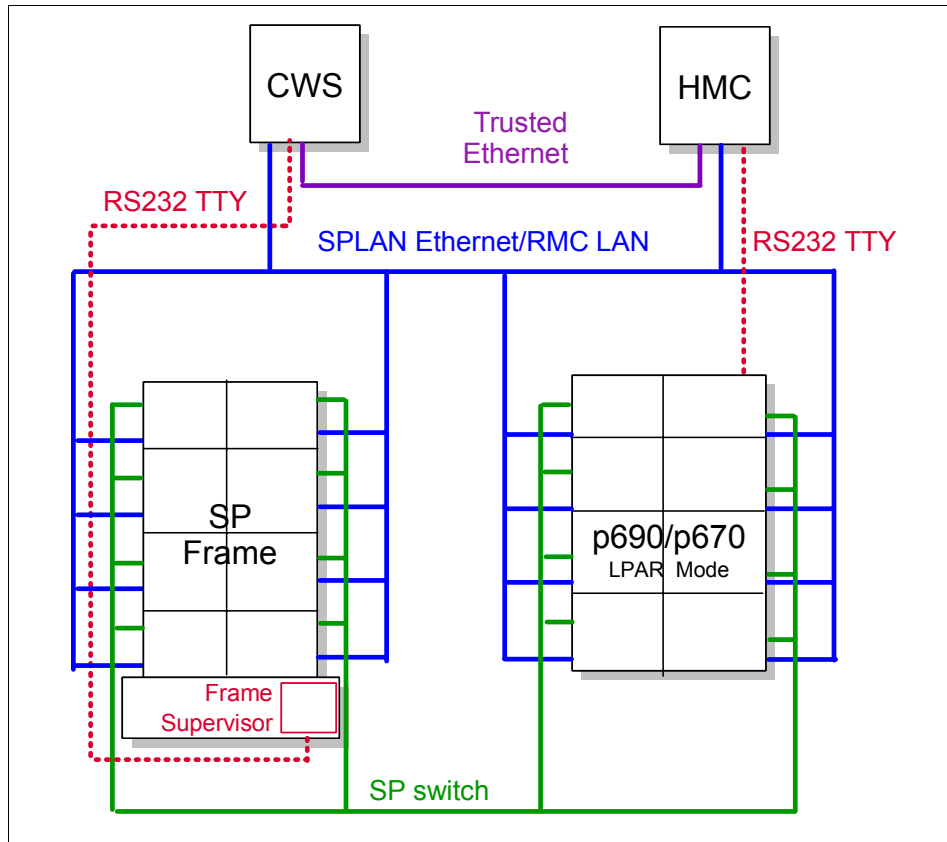


Figure 3-13 SP frame and pSeries with SP-Switched LPARs

Note: Ethernet and serial cables have limited length; keep those lengths in mind during configuration planning, because they limit the location of the servers.

3.6 Networking for Cluster Systems Management (CSM)

In the following sections, we highlight considerations and other information that will help you when building the network for the Cluster 1600. These sections provide specific information on the Cluster 1600 managed by Cluster Systems Management (CSM) and its subsystems, and on the pSeries High Performance Switch (HPS) used in Cluster 1600 managed by CSM.

3.6.1 CSM hardware control

IBM Cluster Systems Management (CSM) for AIX 5L hardware control software provides remote hardware control functions for CSM cluster nodes from a single point of control. CSM allows a system administrator to control cluster nodes remotely through access to the cluster management server. An administrator can run cluster management commands from the management server using the command line, Web-based System Manager graphical user interface (GUI), or System Management Interface Tool (SMIT) panels.

CSM hardware control functions depend on the specific hardware, software, network, and configuration requirements. The requirements for remote power are separate and distinct from the requirements for remote console. AIX 5L clusters without the hardware, software, network, or configuration required to use CSM hardware control can still have CSM installed on some or all cluster nodes. However, in such clusters the hardware control commands may be inoperable or provide only limited function.

CSM for AIX 5L supports cluster hardware control for both pSeries and xSeries nodes from an AIX management server. The hardware control commands can be run on the cluster management server to simultaneously control both node types in a mixed cluster. For a complete description of CSM mixed clusters, refer to *IBM CSM for AIX 5L: Administration Guide*, SA22-7918.

3.6.2 Hardware and network requirements

The management server for CSM for AIX 5L can be connected to cluster nodes and external networks using various configurations of IBM and non-IBM hardware and software that meet the CSM architecture requirements.

Note: To perform hardware control on HMC-attached pSeries machines, Hardware Management Console (HMC) for pSeries release 3 version 1.0 or later is required. The HMC can be used to partition physical pSeries servers into multiple logical partitions (LPARs), or nodes, each containing its own operating system image.

When configuring a CSM for AIX 5L cluster, give particular attention to secure hardware control functions.

For both performance and security reasons, it is important to understand and control the data that is transferred around your network. If the data you transmit over a network is sensitive, and is not encrypted, consider isolating that data on a dedicated network or LAN.

We recommend that you isolate the management server, Hardware Management Consoles (HMCs), console servers, and RSAs on a dedicated network. In a CSM environment, this network is known as a management virtual LAN (VLAN); see “Management VLAN” on page 153 for further details.

A VLAN hides the HMCs, console servers, and RSAs from the normal users and keeps the traffic between them from being visible on the public network. CSM management servers use Telnet, which transmits passwords in clear text, to communicate with console servers.

Many systems administrators choose to create a network for cluster administration data, such as node installations, backups, and monitoring. In a CSM environment, this network is known as a cluster VLAN; see “Cluster VLAN” on page 154. One Ethernet connection is also required for each cluster VLAN.

Many systems administrators also create a network for application-related traffic. In a CSM environment, this network is known as a public VLAN; see “Public VLAN” on page 154.

Considerations

It is possible to use private IP addresses (192.168.*.*, 172.16.*.*, or 10.*.*) for both the management VLAN and the cluster VLAN. Private addresses can provide a higher level of security than public addresses.

However, if your administrators’ workstations are not on the management VLAN, they will have no direct access to HMCs, RSAs, or console servers. They will have to access them by logging on to the management server.

If this is not acceptable, the only alternative is to use routing. However, routing can become complex if your administrators’ workstations are dispersed across multiple networks, or if they use DHCP for their workstations.

Incorrectly implemented routing can impact security, and thus defeat the purpose of creating multiple networks. For example, incorrect routing can cause secure data to be transmitted over the public VLAN.

Figure 3-14 shows a possible network configuration for a fairly simple environment. It contains three networks, a management VLAN, a cluster VLAN, and a public VLAN.

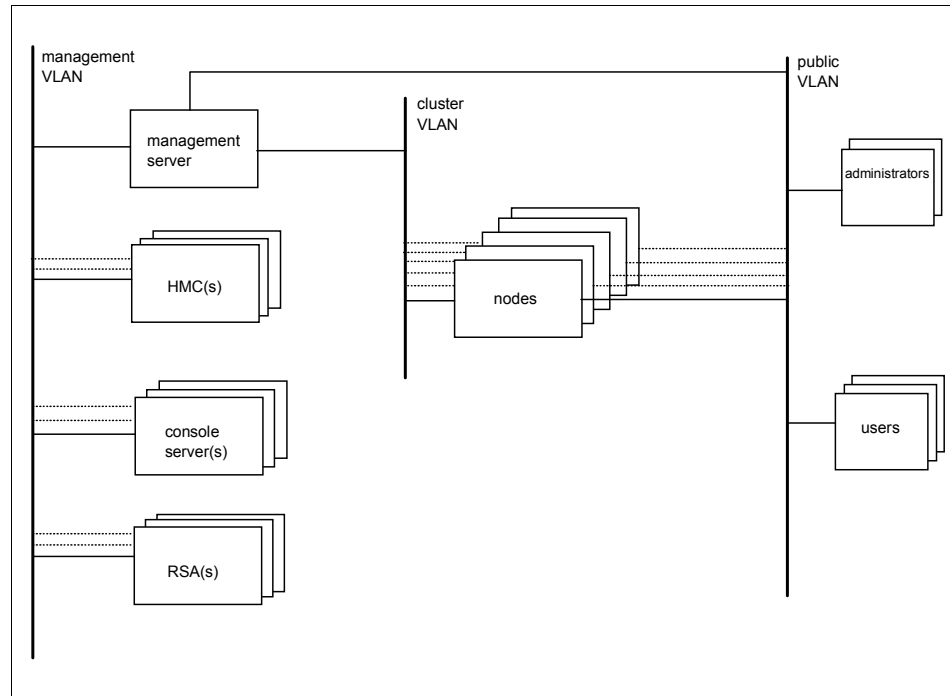


Figure 3-14 VLAN configuration

3.6.3 Virtual LANs (VLANs)

VLANs discussed in this redbook refer to VLANs as defined by IEEE standards. VLANs can be configured to control cluster security access. Figure 3-14 on page 153 shows a network partitioned into three virtual LANs, the management, cluster, and public VLANs. We explain these in the following sections.

Management VLAN

Hardware control commands such as **rpower** and **rconsole** are run on the management server and communicate to nodes through the management VLAN. The management VLAN connects the management server to the cluster hardware through an Ethernet connection.

For optimal security, the management VLAN must be restricted to hardware control points, remote console servers, the management server, and root users.

Routing between the management VLAN and cluster or public VLANs could compromise security on the management VLAN.

Note: The management VLAN is subject to the Remote Server Access (RSA) restriction of 10/100Mb/s.

Cluster VLAN

A cluster VLAN connects nodes to each other and to the management server through an Ethernet connection. Installation and CSM administration tasks such as running **dsh** and **ssh** are done on the cluster VLAN. Host names and attribute values for nodes on the cluster VLAN are stored in the CSM database.

Public VLAN

A public VLAN connects the cluster nodes and management server to the site network. Applications are accessed and run on cluster nodes over the public VLAN. The public VLAN can be connected to nodes through a second Ethernet adapter in each node, or by routing to each node through the Ethernet switch.

Note the following:

- ▶ The management VLAN contains only the CSM management server, HMCs, console servers, and RSAs.
- ▶ The cluster VLAN contains the management server and the nodes.
- ▶ The public network contains the nodes, the users' workstations, and the administrators' workstations. It may also contain other servers that are not members of the CSM cluster.
- ▶ The HMCs, console servers, and RSAs can be accessed only through the management server.

3.6.4 Conceptual diagram for pSeries cluster

Figure 3-15 on page 155 shows the hardware and networking configuration required for using CSM hardware control with IBM HMC-attached pSeries nodes. The p690 machines in the diagram are single pieces of hardware that have been partitioned by HMCs into 16 LPARs (nodes). The management server connects to the management and cluster VLANs through Ethernet adapters. The nodes must be connected to the cluster VLAN through their first Ethernet adapters (eth0), and directly or indirectly to an HMC. Configuration for a public VLAN is flexible and can be defined by the system administrator.

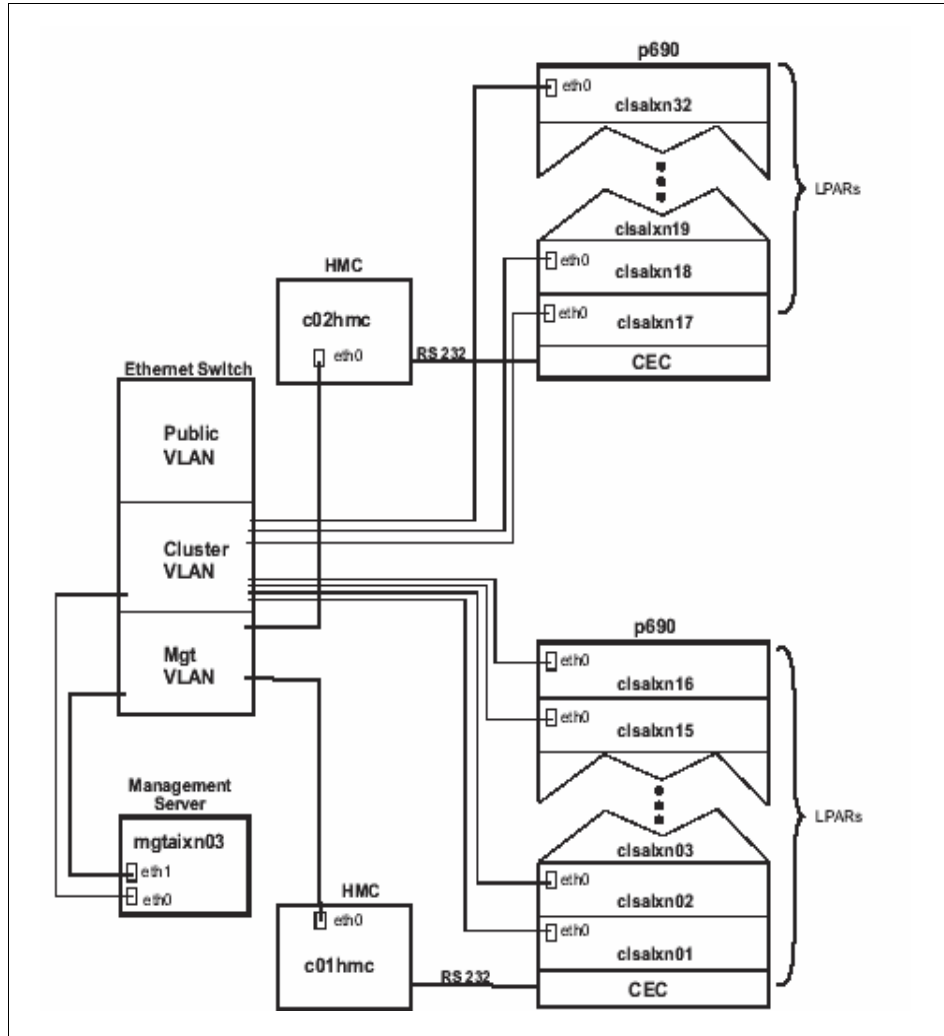


Figure 3-15 Conceptual diagram for series cluster

Redundant HMCs are supported providing automatic failover. The HMC can control multiple server machine types. When combining multiple machine types, substitutions of other server types are based upon the relative weighting factor of the server. For example, a p690 is twice the weighted factor of a p630 (eight p690s per HMC compared to sixteen p630s per HMC). Therefore, a supported HMC configuration could control four p690 servers and eight p630 servers.

Note: Ethernet and serial cables have limited length; keep those lengths in mind during configuration planning, because they limit the location of the servers in relation to the other system equipment.

3.6.5 pSeries HPS switch network overview

SP switch technology used several types of switch adapters to connect system components to the switch network. With the pSeries HPS, all network connections pass through a revolutionary new interface card packaged as either a 2-Link Switch Network Interface or a 4-Link Switch Network Interface (SNI).

Each link on the SNI allows message passing between the server bus and the switch. Therefore, each of the multiple links provides a function similar to a previous SP-type switch adapter. These architectural improvements in the SNI and the pSeries HPS create a distinct difference in switch networks when compared to previous adapters and switch networks.

For example, using the SP Switch2, a Cluster 1600 switch network could be configured with two SP Switch2 Adapters at each communication point. With the SP Switch2, the two switch adapters formed two independent switch networks. If one of the adapters developed a fault, the network experienced reduced performance.

In a similar manner, the pSeries HPS can also use two Switch Network Interfaces at each communication point. However, with the pSeries HPS, both the 2-Link Switch Network Interface and the 4-Link Switch Network Interface allow message passing across direct internal connections on each of the SNIs; refer to Figure 3-16 on page 157.

Using these direct internal connections, the interfaces create a single, unified switch network across the SNI pair. As a result, if one of the interfaces develops a fault, its load is automatically directed to the other member of the pair. Although there may be reduced performance at the failed interface component, the overall performance of the pSeries HPS network remains unchanged.

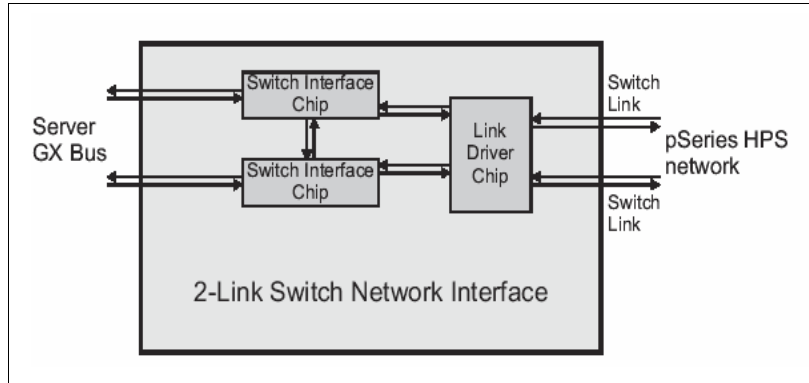


Figure 3-16 Conceptual view of a 2-Link SNI card

Topology overview

Each switch link may be shared by a maximum of sixteen independent logical partitions within a server (configured by the HMC). It is important to note that the ability to share links is a major enhancement to system capability. Since each link can be shared by up to sixteen LPARs, one SNI can now take the place of sixteen SP type switch adapters. This allows systems using the pSeries HPS to be configured for the maximum number of LPARs while using a minimum number of SNIs. Another way to look at link sharing is that a single LPAR can also be connected with up to eight SNIs for maximum bandwidth.

A pSeries HPS may be used as either a server switch board (SSB) or as an intermediate switch board (ISB). For small networks, the pSeries High Performance Switches are configured as SSBs and the switches are connected to each other. In this configuration, a single SSB can be connected to a maximum of sixteen links located on various configurations of Switch Network Interfaces. This allows the SSB to be connected with up to sixteen other switches.

In contrast, large network configurations use pSeries High Performance Switches configured as ISBs to connect the SSBs. If a network has three or fewer SSBs, an ISB is not needed. However, if a network has more than three SSBs, Intermediate Switch Boards are required to insure bandwidth availability. For more information, refer to *pSeries High Performance Switch Planning, Installation, and Service*, GA22-7951.

The pSeries HPS network consists of two interconnected SNIs at each communication point. In addition, each SNI has either one, two, or four external links (in a multi-link network, links are always added in pairs). The number and configuration of switch boards required to support these links depends on the total number of links in the network.

Note: Single-link SNIs require a converter wrap installed in multi-link SNIs. Adding single-link SNIs to a network places constraints on how the network is cabled.

With two interconnected SNIs at each communication point, the pSeries HPS network has parallel paths from the SNIs through the switch. While the SP Switch2 network could also be configured with one or two SP Switch2 Adapters per node, the two adapters created two independent networks between servers. Because those two networks were not interconnected, a message on one network could not be passed to the parallel network. If one of the networks (planes) went down, the system did not have redundancy available.

In comparison, the pSeries HPS supports up to eight communication paths between the switch and the server. However, the Switch Network Interfaces for the pSeries HPS are also interconnected within either the 2-Link or the 4-Link SNI package. With this design, should a fault occur on one link, the parallel adapter link will absorb the message load of the faulted adapter. With this architecture, each link on the pSeries HPS network provides full interconnection between all SNIs. Because the SNIs are completely interconnected, the parallel paths can be thought of as a unified network.

3.6.6 Switch Network Manager (SNM)

Switch Network Manager (SNM) is a new approach to switch management. SNM software resides on the Hardware Management Console (HMC), and manages networks consisting of pSeries High Performance Switches and 2-Link or 4-Link Switch Network Interfaces.

SNM is intended to be more LAN-like in its operation, meaning that the same level of control given in prior versions of switch management is no longer required. With SNM software, pSeries HPS networks do not require specific commands (like the SP “Ecommands”) to start or stop switch operations, bring servers partitions on and off the switch, or perform other functions previously driven by commands. Instead, the SNM software dynamically determines the network topology and initializes the active switch components. When new components are added or removed, SNM again manages this with little or no user intervention.

With the pSeries HPS, all servers or server logical partitions (LPARs) act as peers on the switch. In contrast to previous SP switch types, the cluster does not have a switch primary node. Instead of a primary node, the pSeries HPS network uses the interfaces provided by the HMC service network and SNM software to configure, initialize, control, and monitor the switch network.

SNM configuration functions

SNM dynamically sets up the internals of the network and defines its physical endpoints. This software also provides miswire detection, helps balance network bandwidth, offers limited capabilities to help plan network upgrades, and supplies service actions for non-operational components.

SNM initialization functions

SNM initialization functions allow you to set the operational parameters of the network. Initialization runs automatically once the network configuration is known. During initialization, SNM communicates with each active network component (switch and SNI), sets up their operational parameters, delivers route information to the endpoints, and signals to each server OS the availability of message passing over the network.

SNM control functions

SNM provides control and monitoring capabilities through the IBM Web-based System Manager Graphical User Interface (GUI) for function like:

- ▶ Power controls
- ▶ System status
- ▶ Diagnostics
- ▶ Network topology services
- ▶ Switch hardware control

SNM monitoring functions

SNM's monitoring functions allow the software to detect, recover, and report error conditions in the network. When errors are detected, SNM determines the severity of the error, takes the appropriate action to recover or disable failing component, and creates serviceable events when appropriate. In most network configurations, these errors will not result in network outages.

Service Focal Point (SFP)

In a partitioned environment, each partition runs independently and is not aware of other partitions on the system. If there is a problem with a shared resource such as a managed system power supply, all active partitions will report the same error. Service Focal Point recognizes that these errors repeat and filters them into one serviceable event for the service representative to review.

Note: Service Focal Point must be installed on the HMC.

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. With Service Agent installed on the HMC, the serviceable events captured by SFP are automatically sent to IBM (call-home support) and automatically generate a maintenance request.

Note: pSeries HPS can be used with pSeries 690 and pSeries 655 only.

3.6.7 Considerations for Cluster 1600 managed by CSM network

This section contains the following network hardware planning information:

- ▶ Administrative LAN Ethernet adapters and cabling
- ▶ RS-232 and RS-422 serial cabling requirements
- ▶ Switch interfaces and cabling

Administrative LAN Ethernet adapters and cabling

This includes the management server and all attached servers using customer-supplied twisted pair Ethernet cable. The LAN connection may be attached to the native Ethernet port on the HMC, or through additional PCI adapters placed in the HMC.

pSeries High Performance Switches are not directly connected to the administrative LAN. All LAN adapters and cables are server-mounted and server-connected.

RS-232 and RS-422 serial cabling requirements

This requires a serial network to monitor and control system variables such as heartbeat and power.

pSeries High Performance Switches do not require an RS-232 or RS-422 connections to the HMC. However, all switch-only frames housing the pSeries HPS and any server frame connected to the pSeries. HPS requires a pair of RS-422 connections to the frame BPC.

Note: Be aware of the following:

- ▶ RS-232 to RS-422 converters are included with the RS-422 cables shipped with F/C 8122 and F/C 8123.
- ▶ In contrast to other p690 installations that do not require RS-422 connections, all p690 (M/T 7040) servers connected to the pSeries HPS require RS-422 connections between the M/T 7040-61R frame BPCs and the HMC.

Switch interfaces and cabling

There are two types of switch connections (interfaces): server-to-switch, and switch-to-switch.

Server-to-switch

The server-to-switch interface requires three component types to connect each server or LPAR to the switch:

- ▶ Switch Network Interfaces on the server side of the connection
- ▶ Copper cable switch port connectors on the switch side of the interface
- ▶ Up to four copper cables connecting the server side to the switch side of the interface

Switch-to-switch

The switch-to-switch interface requires two component types to interconnect the switches on the network:

- ▶ Paired switch port connectors for each switch interconnection using either:
 - Copper cable option
 - Fiber optic cable option
- ▶ Cables to complete the switch-to-switch interface using either:
 - Two copper cables
 - Four fiber optic cables

3.6.8 Examples

In this section, we provide sample scenarios for the pSeries High Performance Switch (HPS) in a Cluster 1600, as shown in Figure 3-17 on page 163. In a Cluster 1600 with a pSeries HPS, one of the LPARs is the management server. The pSeries HPS is connected to the pSeries 655 and the pSeries 690 through switch interface cables.

The HMC is connected to the following:

1. The administrative LAN network
 - This includes the management server and all attached servers using customer-supplied twisted pair Ethernet cables.
 - The LAN connection may be attached to the native Ethernet port on the HMC, or through additional PCI adapters placed in the HMC.
2. The environmental control network
 - This requires a serial network used to monitor and control system variables such as heartbeat and power.
 - This requires RS-232 connections to the HMC 1 serial port on all servers (redundant HMC configurations also use the HMC 2 server port).
 - This requires RS-422 connections to both BPC ports on all cluster frames (including M/T 7040-61R frames and M/T 7040-W42 switch-only frames).

Note: Redundant HMCs are not required, but this optional configuration is available. When configured, redundant HMCs require separate cables from both HMCs to all required connection points. For additional information on redundant HMCs, refer to *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590.

For details about the limitation and restrictions on the HPS network, refer to “Cluster 1600 hardware” on page 11.

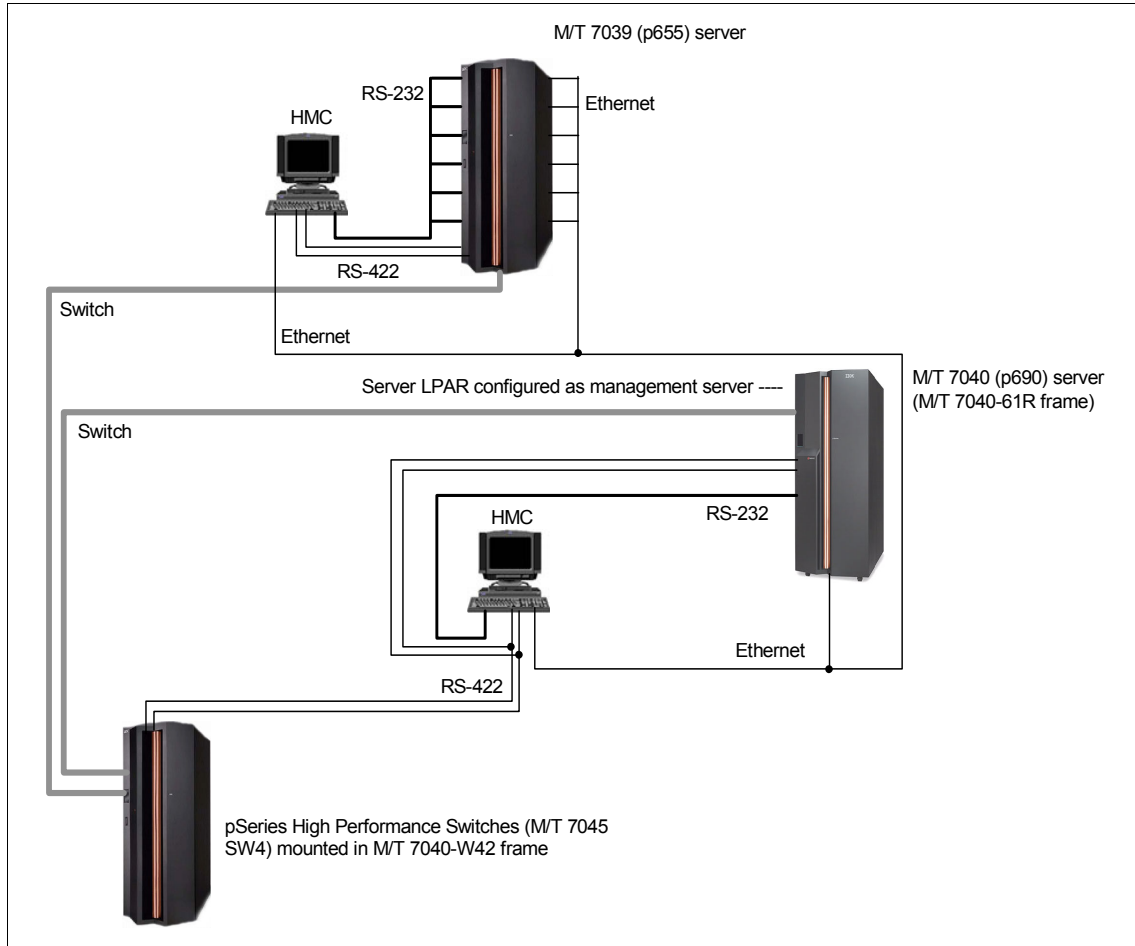


Figure 3-17 Conceptual pSeries HPS Cluster 1600

3.6.9 Redundant HMC Layout for pseries HPS in Cluster 1600

We describe a sample scenario as shown in Figure 3-18 on page 164 for the pSeries HPS in a Cluster 1600 with redundant HMCs. In a Cluster 1600 with a pSeries HPS with redundant HMCs, one of the pSeries server M/T 7028 is the management server. The pSeries HPS is connected to the pSeries 655 and the pSeries 690 through the switch interface cables, and the HMC is connected to the pSeries HPS via an RS/422 cable. The pSeries HPS requires a Model 7315-C01 as the HMC. When used as the part of Cluster 1600, the pSeries HPS requires a HMC.

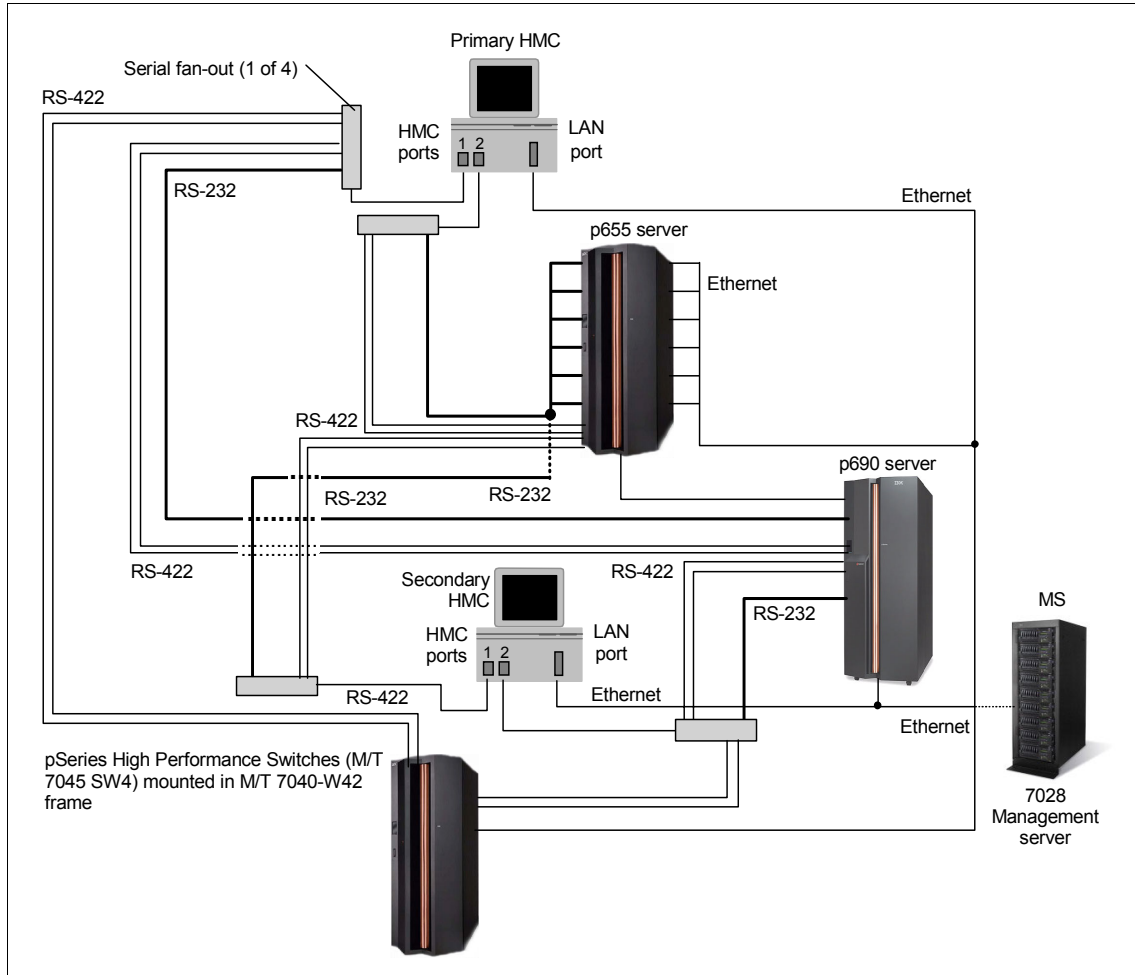


Figure 3-18 Redundant HMC layout for pSeries HPS in Cluster 1600

3.6.10 Redundant layout for pSeries in the Cluster 1600

Figure 3-19 on page 165 shows a cluster of pSeries servers, with all nodes connected to the Ethernet LAN. To achieve high availability, two HMCs are used. Each HMC is connected to all pSeries servers via RS-232/RS-422 connections. The connection between the management server and the HMC is via an Ethernet connection.

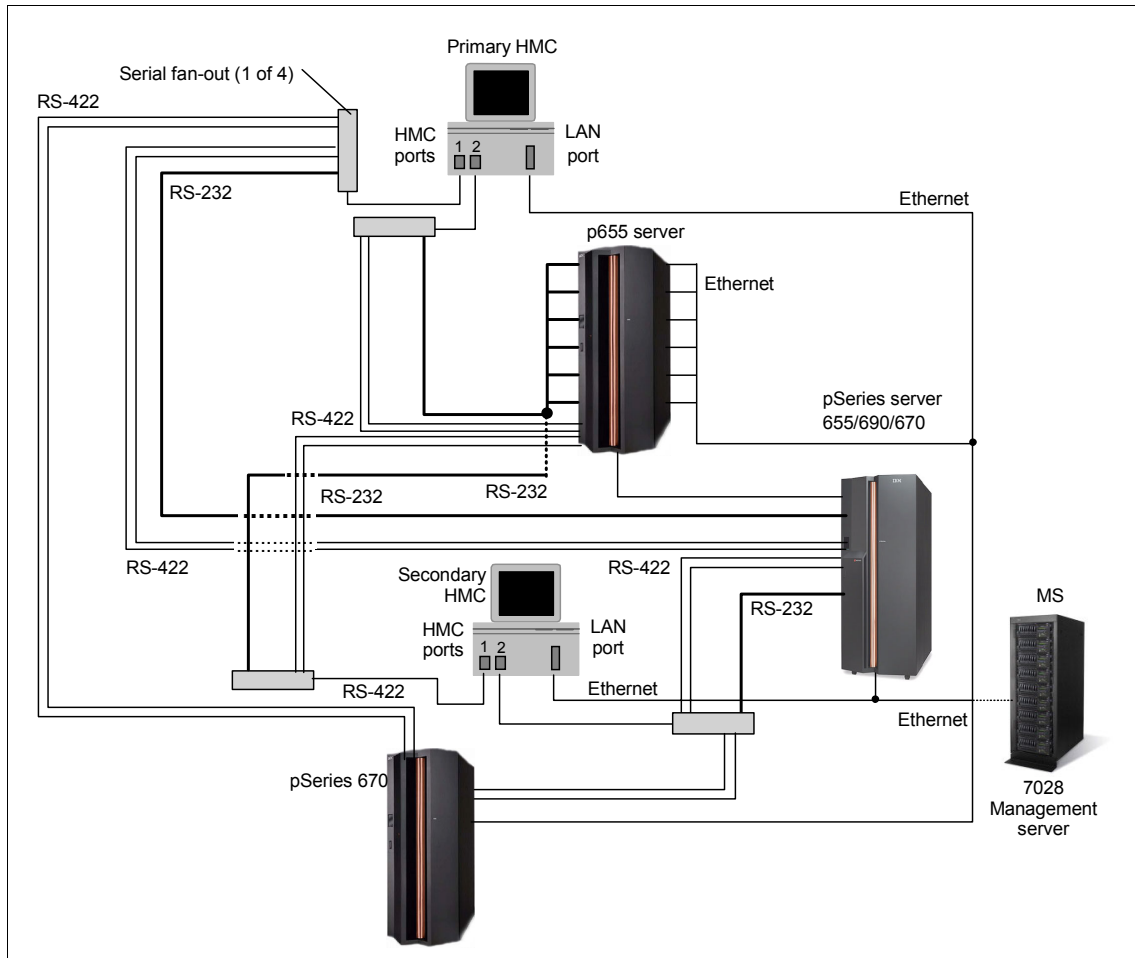


Figure 3-19 Redundant layout for pSeries in a Cluster 1600

3.6.11 Conceptual Cluster 1600 without a pSeries HPS

Figure 3-20 on page 166 shows classic SP nodes with frame and pSeries servers; all pSeries and the SP nodes are connected to the Ethernet LAN. One HMC is used to connect to all the pSeries servers, and the management server is connected to the HMC via the RS-232/RS-422. To achieve high availability, two HMCs can be used. Each HMC can be connected to all pSeries servers via RS-232/RS-422 connections. The management server is connected to the SP frame with RS-232 and Ethernet connections. The connection between the management server and the HMC is via an Ethernet connection.

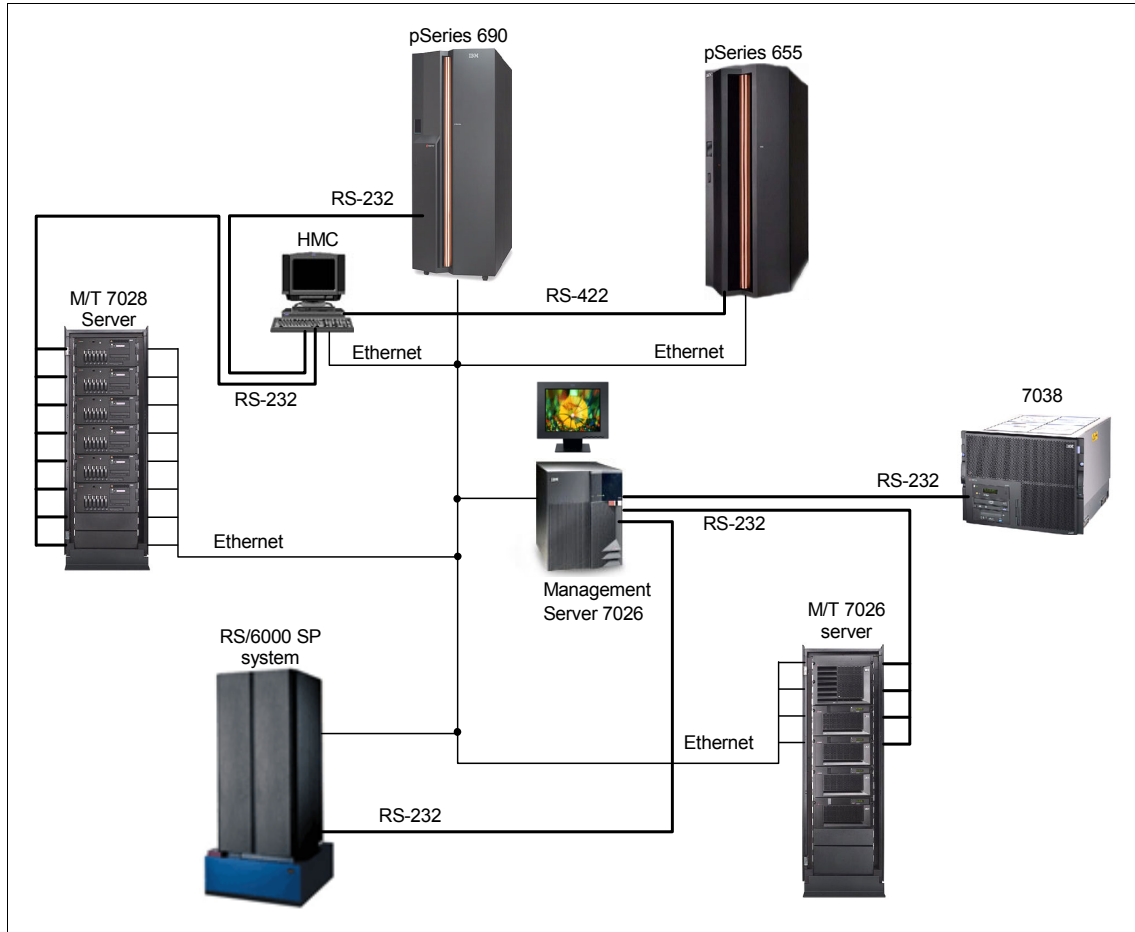


Figure 3-20 Conceptual Cluster 1600 without HPS

3.6.12 Management Server with two HMCs and four pSeries

Figure 3-21 on page 167 shows four pSeries servers in LPAR mode, with all nodes connected to the Ethernet LAN. To achieve high availability, two HMCs are used. Each HMC is connected to all pSeries servers via RS-232 connections. The connection between the management server and the HMC is via a trusted Ethernet connection.

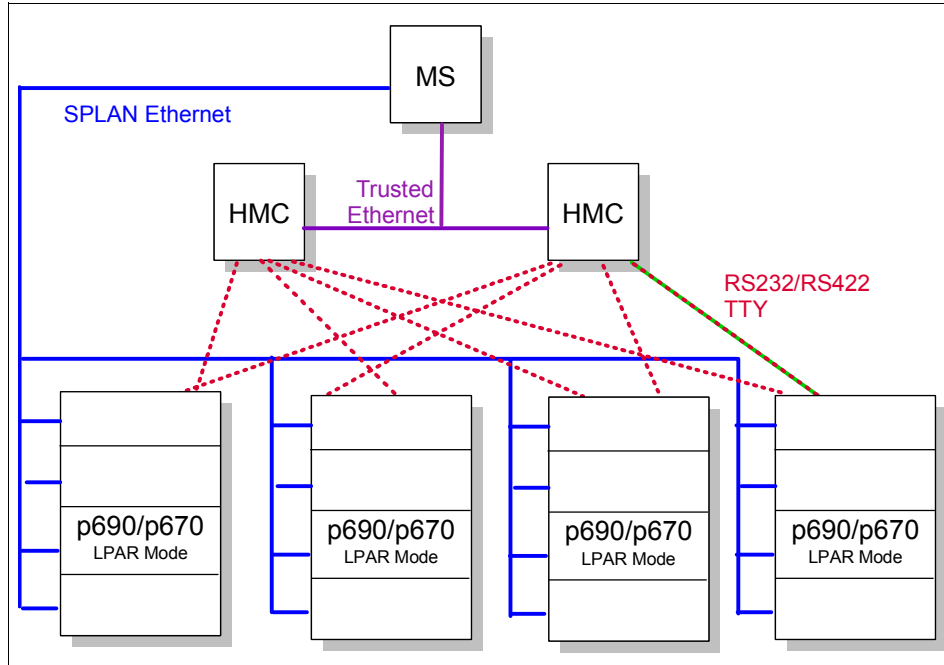


Figure 3-21 Four pSeries with two redundant HMCs and one management server

To avoid having the HMC become, potentially, a single point of failure, we recommend that you connect each pSeries to two different HMCs. That way, in the event of an HMC failure, you are still capable of reaching the pSeries through the second HMC connection. Depending on how many pSeries servers you have, additional 8-port asynchronous adapters may be required.



Software support

In this chapter, we describe the software available for Cluster 1600 systems:

- ▶ Parallel System Support Programs (PSSP)
- ▶ Cluster Systems Management (CSM)
- ▶ General Parallel File System (GPFS)
- ▶ LoadLeveler (LL)
- ▶ Scientific Subroutine Library (ESSL, PESSL, MASS)
- ▶ Parallel Environment (PE)
- ▶ Performance Toolbox (PTX®) and Performance AIDE (PAIDE)

Note: In this redbook, the terms “managed nodes” and “nodes” are used interchangeably to mean “managed servers”.

4.1 Software components of the Cluster 1600

The Cluster 1600 leverages and extends the software capabilities of the very successful RS/6000 SP. The full suite of software offered in Cluster 1600 is listed in Table 4-1.

Table 4-1 Software components on the Cluster 1600

Software component	Products
Cluster Management Software	<ul style="list-style-type: none">▶ Parallel System Support Programs (PSSP)▶ Cluster Systems Management (CSM)
Cluster File System	<ul style="list-style-type: none">▶ General Parallel File System (GPFS)
High Performance Computing Software Suite	<ul style="list-style-type: none">▶ LoadLeveler (LL)▶ Parallel Environment (PE)▶ Scientific Subroutines Libraries (SSL):<ul style="list-style-type: none">– Engineering and Scientific Subroutine Libraries (ESSL)– Parallel ESSL
High Availability Cluster Software	<ul style="list-style-type: none">▶ High Availability Cluster Multiprocessing (HACMP)

In this chapter, we discuss the software components in the different cluster management environments, as shown in Figure 4-1 on page 171 and Figure 4-2 on page 172.

Note: All software versions indicated in this chapter indicate the updated release only. Check with your IBM representative to determine the recommended maintenance package (APAR) required for the installation.

Figure 4-1 shows a functional diagram of the Cluster 1600 software that is tightly integrated with PSSP. This is implemented in typical RS/6000 SP installations, as well as in Cluster 1600 installations that use PSSP for cluster management.

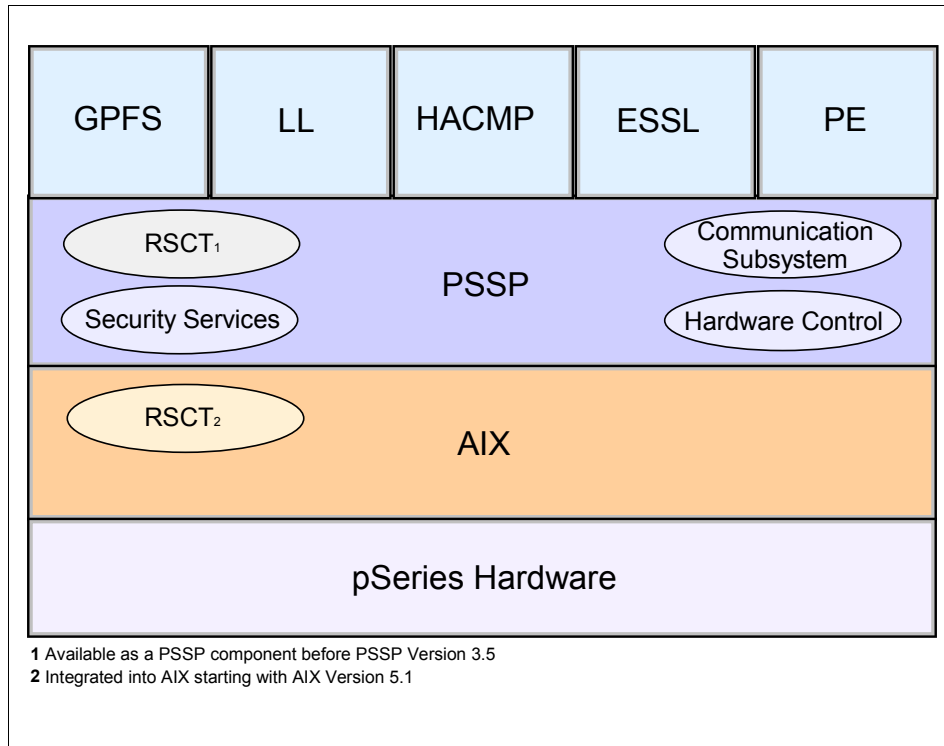


Figure 4-1 Functional diagram of Cluster 1600 software on PSSP

Figure 4-2 on page 172 shows a functional diagram of the Cluster 1600 software that operates within a CSM environment.

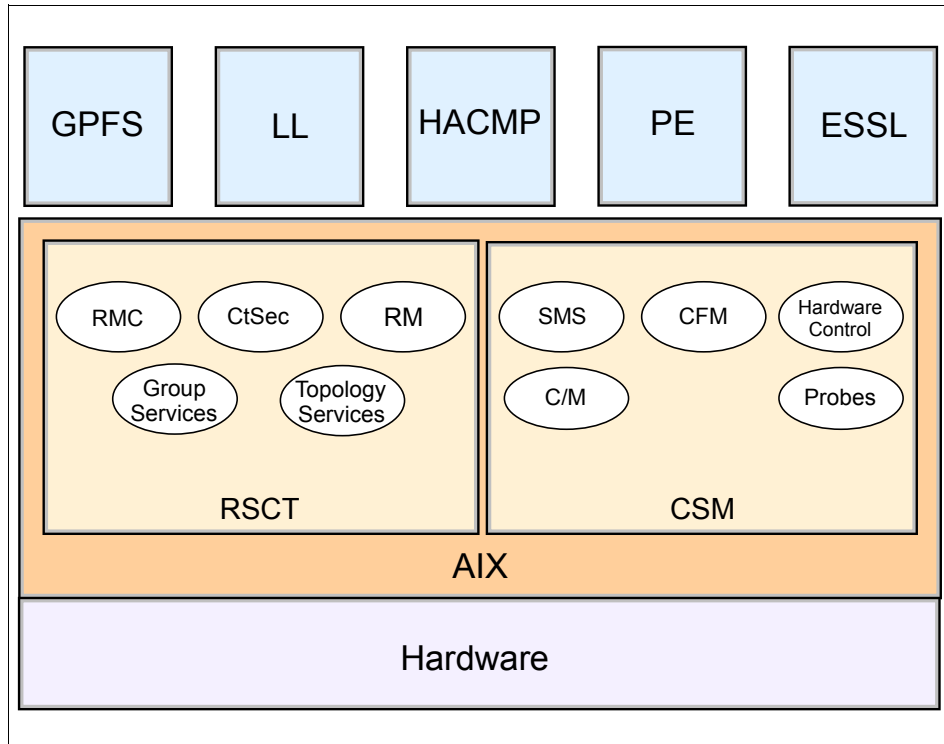


Figure 4-2 Functional diagram of Cluster 1600 software on CSM

4.2 Parallel System Support Programs (PSSP)

The PSSP software provides a comprehensive suite of applications for the installation, operation, management, and administration of the RS/6000 SP and Cluster 1600 system from a single point of control. The software is offered as program number 5765-D51 and consists mostly of base components and a set of optional components.

PSSP offers the following advantages:

- ▶ A full suite of system management applications with the unique functions required to manage the SP and Cluster 1600 system.
- ▶ Simplified installation, operation, and maintenance of all nodes in a Cluster 1600 system. You can operate from a single control workstation.
- ▶ Parallel system management tools for allocating Cluster 1600 resources across the enterprise.

- ▶ Advanced performance monitoring for consolidated analysis and reporting.
- ▶ Error detection and recovery features that reduce the impact and occurrence of unplanned outages.
- ▶ Coexistence is allowed for several releases within an SP partition, allowing for easier software migration.

4.2.1 Administration and operation

PSSP allows the system administrator/operator to perform all local and remote administrative functions from the control workstation (CWS). This makes the CWS a single point of control, as illustrated in Figure 4-3 on page 174.

The PSSP system administration component packages make operating and administering an SP system easy and efficient. This is accomplished through the tools and capabilities that PSSP offers:

- ▶ Installation facility

PSSP supports the installation of AIX and PSSP on nodes using the Network Installation Management (NIM) environment.
- ▶ Configuration management

The System Data Repository (SDR) stores node configuration information that can be retrieved across the workstation, file servers, and nodes.
- ▶ File management

Files may be grouped together such that any changes to these file collections can be effectively propagated to the appropriate nodes.
- ▶ User management

Administrators can easily add, change, and delete users and passwords.
- ▶ Consolidated accounting

This allows you to centralize records at the node level (for tracking use by wall clock time rather than processor time), and gather statistics on parallel jobs.
- ▶ System monitoring and control

PSSP provides system management tools that enable systems administrators to manage the Cluster 1600. PSSP allows authorized users to monitor and manipulate cluster hardware variables at node or frame levels. PSSP also provides for consolidation of error and status logs, to expedite problem determination.
- ▶ Other administrative tools
 - Parallel system management functions across multiple nodes.

- SP Perspectives, a set of graphical user interfaces (GUIs) that help to simplify administrative work.
- Centralized Management Interface, available in the form of a menu-driven interface for system management commands, as well as command line equivalents.

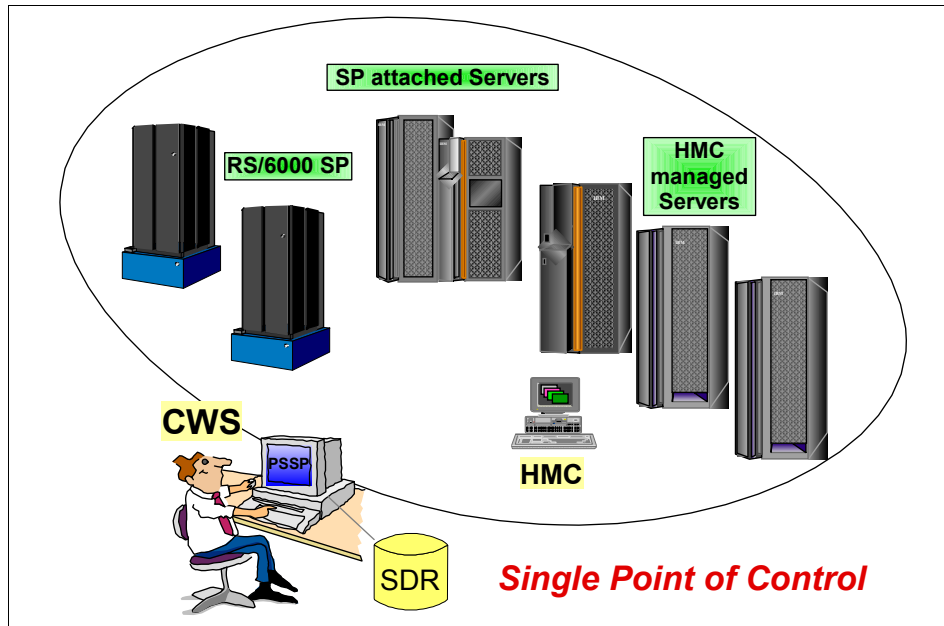


Figure 4-3 Single point of control from CWS in PSSP-managed clusters

4.2.2 Reliable Scalable Cluster Technology (RSCT)

RSCT is a set of software components that provides a comprehensive clustering technology on AIX and PSSP. It provides an infrastructure to achieve improved system availability, scalability, manageability, and ease of use for several IBM products.

Note: RSCT is an integrated component of PSSP up to PSSP 3.4. It is no longer shipped with the PSSP product set with the release of PSSP 3.5, as it is integrated into AIX 5L Version 5.1. Hence PSSP3.5 only runs on AIX 5L Version 5.1. AIX 5L Version 5.1 installs RSCT by default.

The RSCT design differences in PSSP and AIX are beyond the scope of this redbook. You can find these details in Chapter 3, *IBM @serverCluster 1600*

Managed by PSSP 3.5: What's New, SG24-6617. For more information about RSCT, also refer to IBM RSCT for AIX: Guide and Reference, SA22-7889.

4.2.3 IBM Virtual Shared Disk (VSD)

The IBM Virtual Shared Disk (IBM VSD) allows multiple nodes to access a disk as if the disk were attached locally to each node. This feature can enhance the performance of applications that provide concurrent control for data integrity, such as Oracle databases or the GPFS file system.

IBM Concurrent Virtual Shared Disk (CVSD)

The IBM Concurrent Virtual Shared Disk (VCVSD) takes advantage of the concurrent disk access environment supplied by AIX. VSD uses the services of Concurrent LVM (CLVM), which provides the synchronization of LVM and the management of concurrency for system administration services.

IBM Recoverable Virtual Shared Disk (RVSD)

The IBM Recoverable Virtual Shared Disk (IBM RVSD) provides recovery from failure of virtual shared disk server nodes and is designed to ensure continuous data and system access from anywhere in the cluster.

Hashed Shared Disk

This component works with the IBM Virtual Shared Disk component to offer data striping for your virtual shared disks.

Note: VSD, CVSD, and RVSD are optional components for PSSP installation.

For details on VSD and RVSD, see *Parallel System Support Programs for AIX, Managing Shared Disks*, SA22-7349.

4.2.4 Security

PSSP 3.5 offers the following security services:

- ▶ Standard AIX authentication
- ▶ Kerberos Version 4 method using either the PSSP or Andrew File System (AFS) components.
- ▶ Kerberos Version 5 method provided by the Distributed Computing Environment (DCE) software.

You can use them individually, or in combination, or choose not to use them at all.

4.2.5 Communication subsystem

The Communication subsystem software supports the SP Switch and SP Switch2, including the device drivers. It allows for switch configuration and initialization, provides adapter diagnostics and fault-handling, as well as parallel communications application programming interfaces (APIs).

4.2.6 Network Time Protocol (NTP)

NTP can be used to manage time synchronization time-of-day clocks on your control workstation and processor nodes in one of the following ways:

- ▶ Using an established NTP server on your network
- ▶ Using an NTP server from the Internet
- ▶ Run NTP on one of the node to generate a consensus time

4.2.7 System availability

To significantly improve system availability, PSSP also contains functions and interfaces to other products that can help reduce unplanned outages and minimize the impact of outages that do occur.

System partitioning

This allows the system to be organized into non-overlapping groups of nodes for various purposes, such as testing of new software or creating different production environments. However, system partitioning is only supported in switchless or SP Switch-attached SP systems. It is not supported in systems with the SP Switch2 or in switchless clustered server systems.

Coexistence can reduce scheduled maintenance time

PSSP can coexist with several releases within the Cluster 1600, thus allowing easier migration to new software levels.

Node isolation

This removes a node or server in the cluster from active duty and enables it to be reintegrated without causing an SP Switch fault or disrupting switch traffic. This isolation is useful for correcting an error condition or installing new hardware and software without impacting production.

Error recovery on an SP switch

This is managed by the active primary node. The primary node handles switch operations and process errors reported from the switch, and initializes the switch

fabric. A separate node can be designated to take over this function in order to eliminate the primary node as a single point of failure.

High Availability Control Workstation (HACWS)

This connects two control workstations (with HACMP installed) to an SP system to provide a backup control workstation in the event the primary one becomes unavailable (only one is active at any time).

HACWS does not automatically support SP-attached or clustered servers. However, if you want to use an SP-attached or clustered server in an HACWS configuration, see the chapter entitled “Planning for a high availability control workstation” in *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281, for limits and restrictions.

4.2.8 Other PSSP services

Other services offered in PSSP are as follows:

- ▶ The IBM eServer Cluster Information Center provides one simple interface for all softcopy SP documentation and information resources. It consists of HTML, Java, and Java script files, and it works with a Web browser.

The Resource Center provides access to a variety of information including publications, READMEs, Redbooks, white papers, product information, as well as up-to-date service information.

- ▶ PSSP supports a multi-threaded, standards-compliant Message Passing Interface (MPI) through an IBM Parallel Environment for AIX (PE), as well as maintaining its single-threaded MPI support.

In addition, PSSP includes a Low-level Application Programming Interface (LAPI) with a flexible, active-message style, communications programming model on the SP Switch.

- ▶ PSSP offers perl programming language for developing system-wide shell scripts.
- ▶ PSSP comes with Tool command language (Tcl) for controlling and extending applications.

4.2.9 New in PSSP 3.5

PSSP 3.5 was announced on the October 8, 2002. It offers functional enhancements to the previous release, including the following:

Use of 64-bit kernel

- ▶ PSSP 3.5 provide support for use of the 64-bit kernel version of AIX 5L V5.1 or later. This includes the software stack supported by PSSP 3.5 (GPFS, MPI, Parallel ESSL, LAPI, KLAPI and IBM VSD). This also extends the capability to address physical memory beyond the 96 GB limit on the 32-bit kernel.

Important: The 64-bit kernel support only exists on the CWS and the managed nodes that meet the following requirements:

- ▶ Running PSSP 3.5 or later
- ▶ Running AIX 5L Version 5.1 Maintenance Level 3 (IY32749) or later
- ▶ Running AIX 64-bit kernel on supported 64-bit hardware

Hardware support

- ▶ PSSP 3.5 software supports logical partitions (LPARs) of pSeries as a PSSP node in a Cluster 1600 or SP system. These nodes require AIX 5L V5.1.
- ▶ The pSeries servers are supported in an SP-attached configuration with the SP Switch2, the SP Switch, or no switch, or in a clustered server configuration.

Switch support

- ▶ Switch support now allows the administrator to specify which set of nodes to exclude from serving as a switch primary or primary backup, and which nodes are available for that purpose.

Note: IBM does not intend to support the next generation IBM eServer High Performance Switch on PSSP-managed clusters.

IBM Virtual Shared Disk

- ▶ Virtual Shared Disk (VSD) has been enhanced for performance using IP-based mode, and adds improved diagnostics.

For details about the enhancements on PSSP 3.5, see Chapter 4, *IBM @server Cluster 1600 Managed by PSSP 3.5: What's New*, SG24-6617.

4.2.10 Software requirements

Table 4-2 lists the minimum software requirements for PSSP 3.5.

Table 4-2 PSSP 3.5 software requirements

AIX version	Other software
5L V5.1 with AIX service at 5100-03	C for AIX, V6.0 or later. VisualAge® C++ Professional for AIX V6.0 or later.

Important: PSSP 3.5 is the last release of PSSP. For more information about PSSP and CSM plans, contact your IBM technical representative or see:

<http://www.ibm.com/servers/eserver/clusters/software>

4.2.11 Software compatibility matrix

Table 4-3 shows the software compatibility matrix for PSSP 3.5.

Table 4-3 PSSP 3.5 software compatibility matrix

Software stack products	Supported version
General Parallel File System (5765-F64)	Version 2 Release 1
LoadLeveler (5765-E69)	Version 3 Release 2
Parallel Environment (5765-D93)	Version 4 Release 1
ESSL for AIX (5765-C42)	No dependencies on PSSP. ESSL Version 4 Release 1 and Version 3 Release 3 will be supported on the Cluster 1600 with the appropriate level of the AIX operating system.
Parallel ESSL for AIX (5765-C41)	Version 3 Release 1
HACMP (5765-E54)	Version 5 Release 1

4.2.12 Documentation references - PSSP

You may wish to refer to the following documentation for more information about PSSP:

- ▶ *IBM @serverCluster 1600 Managed by PSSP 3.5: What's New*, SG24-6617
- ▶ *Parallel System Support Programs for AIX, Administration Guide*, SA22-7348

- ▶ *Parallel System Support Programs for AIX, Managing Shared Disks, SA22-7349*
- ▶ *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment, GA22-7281*

4.3 Cluster Systems Management (CSM)

Cluster Systems Management (CSM) for AIX and Linux is designed for simple, low-cost management of distributed and clustered IBM pSeries and xSeries servers in technical and commercial computing environments. CSM, included with the IBM Cluster 1600 and Cluster 1350, dramatically simplifies administration of a cluster by providing management from a single point of control. CSM is available for managing homogeneous clusters of xSeries servers running Linux or pSeries servers running AIX, or heterogeneous clusters which include both. CSM is offered as program number 5765-F67.

CSM offers the following advantages:

- ▶ It provides the basis for consistent cluster management for both AIX and Linux, which will enable a tighter integration of these solutions over time.
- ▶ It allows server management of a hybrid collection of AIX or Linux servers (or nodes) from a single point of control.
- ▶ It is comprised of a modular architecture, where both IBM software and certain open source software can be integrated into a complete system management solution.
- ▶ It leverages the rich heritage and proven technology of the SP by utilizing and deriving software from PSSP.
- ▶ It is built on Reliable Scalable Cluster Technology (RSCT), which is an AIX-based clustering technology. It can provide tighter integration with AIX, because CSM is now part of the AIX 5L distribution.
- ▶ It is designed to handle large scaling and has the ability to handle distributed operations in parallel.
- ▶ It is designed to be independent of any switch topology or other types of domains, such as those used in the SP.
- ▶ It can manage multiple High Availability clusters inside a given CSM domain.
- ▶ It exploits advance hardware management features of the servers in the Cluster 1600 and Cluster 1350.

The discussion of CSM in this section mainly focuses on managing clusters of pSeries running the AIX operating system.

4.3.1 Administration and operation

CSM has been designed to facilitate easy and efficient systems management of a cluster of servers with an integrated collection of tools, utilities, and functions. Conceptually, CSM-managed clusters are similar to the ones managed by PSSP.

From a high level perspective, the entire cluster consists of a management server and one or more managed servers, as shown in Figure 4-4. The CSM software is packaged into a “server package” that runs on the management server and a “client package” that runs on each managed server.

The size of the CSM client packages are small, thus putting a minimum of overhead on the nodes in the cluster.

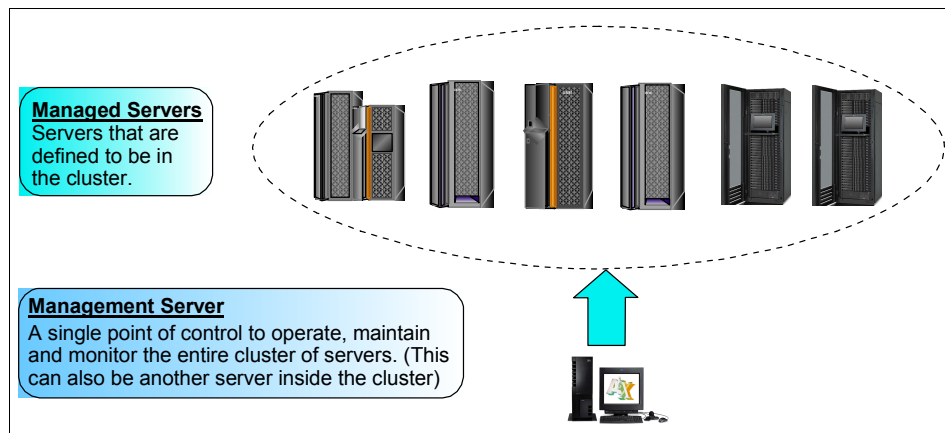


Figure 4-4 An overview of CSM-managed cluster

Tip: While a CSM-managed cluster can consist of a set of AIX-based pSeries, Linux-based pSeries, Linux-based xSeries, or a combination of the three, a CSM-managed Cluster 1600 system is uniquely identified by a pSeries management server running the AIX 5L operating system.

Cluster management from a single point of control

CSM allows the system administrator/operator to perform all system administrative functions from the management server. This makes the management server a single point of control.

The cluster management capability allows an administrator to:

- Add nodes to the cluster (installation of the operating system and/or CSM can be done at that time), and remove nodes from the cluster.

- View and/or change information for one or more nodes in the cluster.

Concept of node groups

Node groups can be formed inside the CSM cluster and managed and monitored as distinct entities, as shown in Figure 4-5.

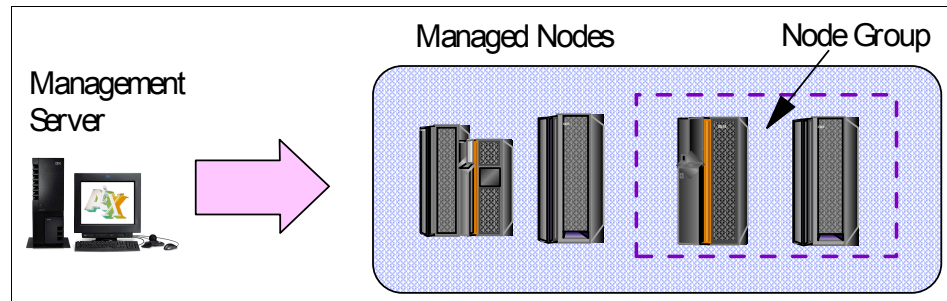


Figure 4-5 Forming a node group within a cluster

Node groups can be defined to be a static list of nodes, or they can be defined to be dynamic and have nodes inside them that correspond to one or more characteristics. This is illustrated in the following examples:

1. We predefined a dynamic node group for Linux nodes, and a dynamic node group for AIX nodes. As nodes are added to the cluster, they will automatically become part of one of their corresponding node groups.
2. An administrator can create a node group for a particular type of hardware and monitor that hardware. If new machines of that hardware type are added into the cluster, monitoring will automatically start for them.

Installation and configuration

Software installation and configuration can be performed centrally on all cluster nodes using the Network Installation Manager (NIM). NIM facilitates the remote installation of AIX on the cluster nodes.

Figure 4-6 on page 183 illustrates the following flow:

- The administrator installs the management server and defines the nodes to be in the cluster; then CSM can remotely do a parallel network operating system install of the nodes.
- CSM updates the software on the nodes, such as new CSM versions and new open source updates. If a node is down during an update, CSM automatically performs the update when the node comes back up.

- CSM automatically sets up the security configuration for the underlying cluster infrastructure. It can also do the necessary setup for rsh or ssh (exchange of ssh keys).

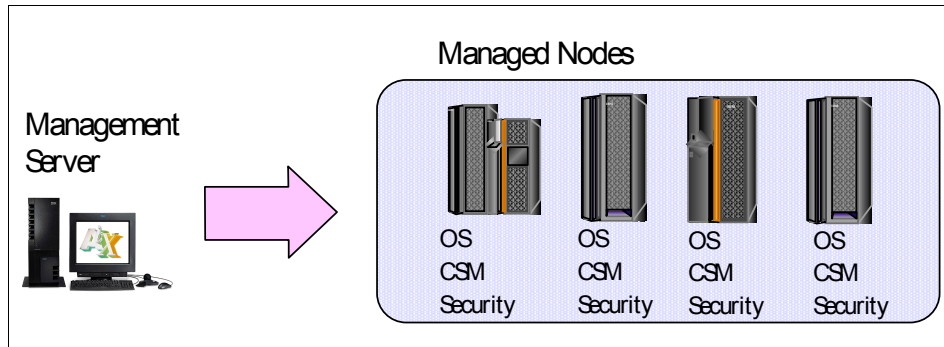


Figure 4-6 CSM installation and setup

For details of AIX installation on cluster nodes, see Chapter 6 of *IBM Cluster Systems Management for AIX 5L, Planning and Installation Guide*, SA22-7919, or Chapter 5 of *An Introduction to CSM 1.3 for AIX 5L*.

Tip: CSM can also allow you to tailor and run user customization scripts during installation and update.

Remote and distributed commands

UNIX basic remote shell (**rsh**), secure shell (**ssh**), and distributed shell (**dsh**) are available to systems administrators to help manage the nodes efficiently.

- Administrators can run commands in parallel across nodes or node groups in the cluster, and gather the output using the **dsh** command.
- Administrators can choose to use either the **rsh** or **ssh** command within **dsh**.
- The **dshbak** command can format the output returned from **dsh** if desired. For example, identical output from more than one node can be collapsed so that it is displayed only once.

Configuration File Management (CFM)

CFM is the file distribution technology used by CSM. It allows common configuration files and group-specific configuration files to be maintained once in a CSM server directory (**/cfmroot**), and then be automatically distributed to all the nodes or node group in the cluster, so administrators do not have to copy files manually across the nodes in the cluster. This file maintenance and propagation is illustrated in Figure 4-7.

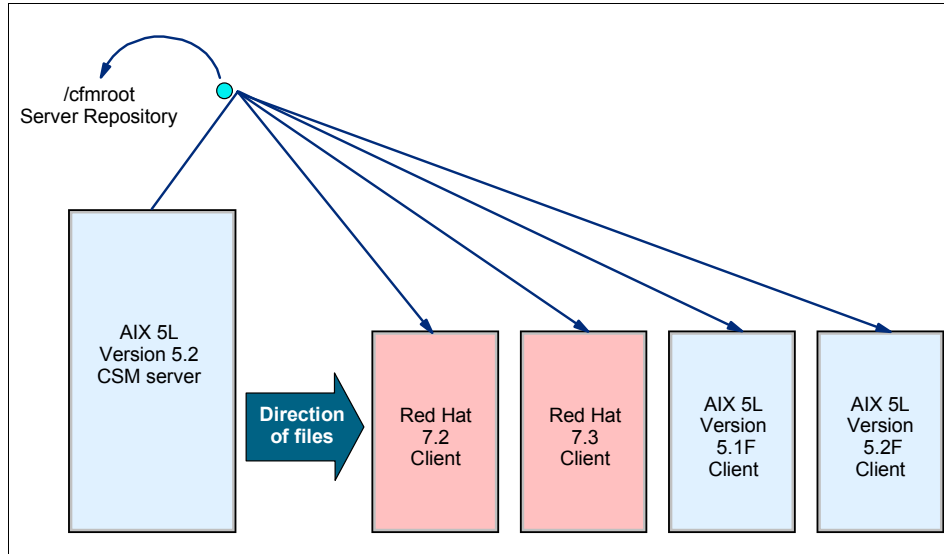


Figure 4-7 Configuration file maintenance and propagation using CFM

CFM makes use of the **rdist** command for file transfer. The **rdist** command can use **rsh** or **ssh**, so administrators can decide which one to use.

Tip: CFM can use make use of variables for IP address or hostname substitution in files being transferred. Preprocessing and postprocessing scripts can be run before and after a file is copied, such as to stop and start daemons.

For details of CFM, refer to Chapter 1 in *IBM Cluster Systems Management for AIX 5L Administration Guide*, SA22-7918.

Management interfaces

CSM functions can be administered using the following management interfaces:

- ▶ Command Line Interface
- ▶ Distributed Command Execution Manager GUI
- ▶ AIX Systems Management Interface Tool (SMIT)
- ▶ Web-based System Manager

CSM uses a set of Web-based System Manager plug-ins to provide a common look and feel in a Web-based System Manager environment. This provides an interface for monitoring and managing a CSM cluster. For an overview of the plug-ins, see Chapter 6 of *An Introduction to CSM 1.3 for AIX*

5L, SG24-6859. This redbook also shows screenshots of Web-based System Manager interfaces.

Hardware control capability

CSM provides hardware control capability to power on and power off, reboot, bring up a remote hardware console through a tty, and query nodes in the cluster. This can be done remotely without being physically present at the server or hardware console. This capability is currently available for xSeries machines, BladeCenter, and HMC-managed pSeries servers, such as those shown in Figure 4-8.

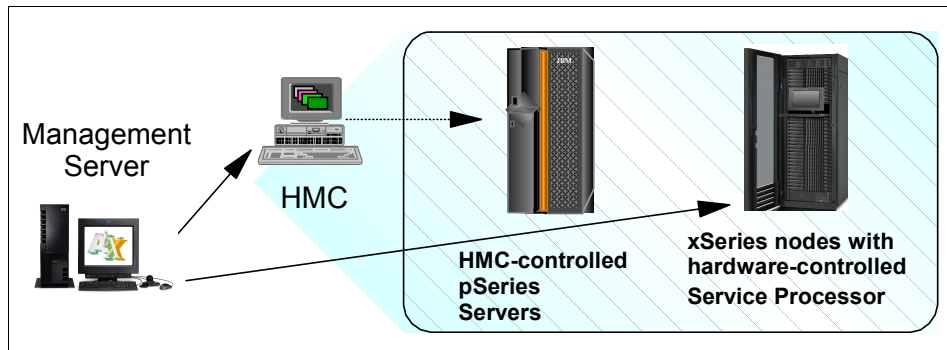


Figure 4-8 Hardware control capability in CSM

The hardware control capability is designed with a layer of code (which performs hardware functions) that references a library based on the hardware type of the node; this allows you to “plug in” a library for the specific hardware to be supported. This capability was specifically designed so that you can easily add additional hardware support.

Cluster management across geographies

CSM remote hardware support on HMC-managed servers extends the AIX cluster management domain beyond the data center. This is illustrated in the example in Figure 4-9 on page 186, which shows an organization with data center operations in four different geographies.

All the HMC-managed servers can be managed remotely from a single CSM management server. While the hardware function is managed through the HMC located at the respective data center, software management is performed through network connectivity from the management server to the clusters.

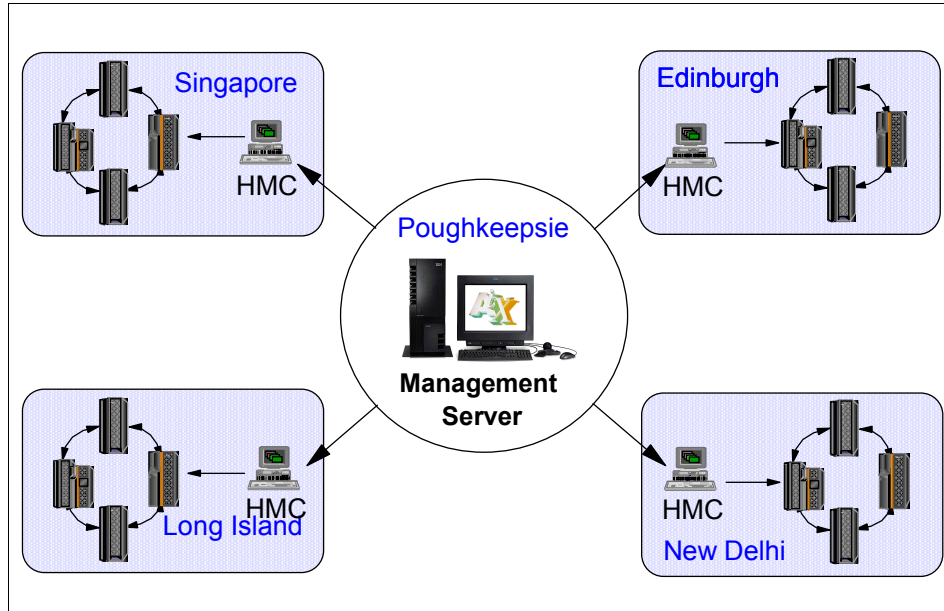


Figure 4-9 Extending cluster management across geographies

System event monitoring

An administrator can set up monitoring for various conditions across nodes or node groups in the cluster and have actions run in response to events that occur in the cluster. This is illustrated in Figure 4-10.

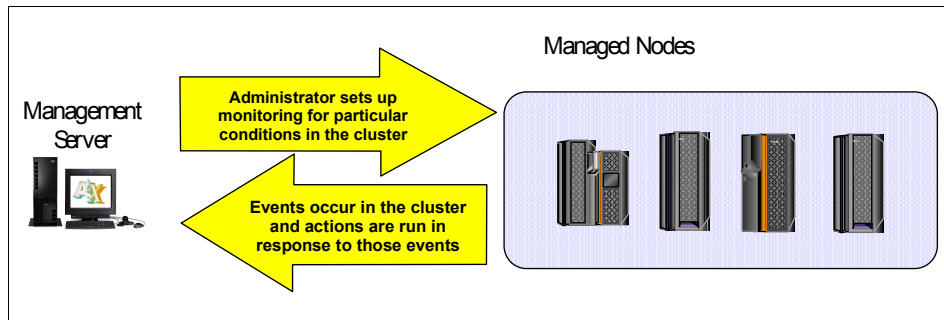


Figure 4-10 Event monitoring and automated response in a CSM cluster

Conditions that can be monitored

Conditions that can be monitored include network accessibility, power status, application or daemon status and utilization level of CPU, memory and file system, among others.

Predefined “Conditions” are shipped with CSM for every type of information that can be monitored, so that monitoring can be started “out of the box”.

Actions in response to occurring conditions

You can run the following actions in response to events that occur in the cluster:

- ▶ Commands can be run on the management server or on any node of the cluster.
- ▶ Notification actions such as logging, e-mailing or paging can be run.
- ▶ SNMP traps can be generated in response to events in the cluster.
- ▶ Predefined “Responses” for e-mail notification, SNMP traps, logging and displaying a message to a console are provided.

Users can quickly associate a condition with a response and start monitoring. And administrators can easily customize such conditions and responses to fit their own needs.

In addition to notification actions, administrator-defined recovery actions can also be run. Such recovery actions can include cleaning up file systems that are filling up, or taking actions to help restart a critical application that went down.

Resources and events in a CSM-managed cluster are monitored through the Resource Monitoring and Control (RMC) infrastructure. RMC is described in 4.3.2, “Reliable Scalable Cluster Technology (RSCT)” on page 188. You can also find useful information in *A Practical Guide for Resource Monitoring and Control (RMC)*, SG24-6615.

Diagnostic probes

Administrators can run diagnostic probes provided by CSM to automatically perform “health checks” of the particular software function if a problem is suspected. These probes are small standalone programs that can be run on a specific part of a system or subsystem to automate problem determination and isolation. These probes can also be run periodically or automatically as a response to a condition occurring in the system.

Current probes shipped with CSM include probes to diagnose network connectivity, NFS health, and the status of daemons that CSM runs. Administrators can also write their own probes to add into the existing CSM probe infrastructure.

For a complete list of probes in CSM 1.3, refer to *IBM Cluster Systems Management for AIX 5L, Administration Guide*, SA22-7918.

4.3.2 Reliable Scalable Cluster Technology (RSCT)

RSCT, a set of software components, provides a comprehensive clustering technology on AIX and Linux. It brings an infrastructure that helps achieve improved system availability, scalability, and ease of use for several IBM products. CSM utilizes RSCT as its clustering infrastructure. A diagrammatic representation of the RSCT components on AIX is shown in Figure 4-2 on page 172, and the main components are discussed in the next sections.

Resource Monitoring and Control (RMC)

RMC is a distributed system monitoring application provided by RSCT. It allows you to define conditions on the system to be monitored, and can automatically respond to the system events that occur when those predefined conditions are met.

For example, you can configure RMC to monitor the usage of file systems and automatically expand if a predefined threshold is exceeded. You can also set different predefined thresholds for file systems in different node groups.

AIX 5L Version 5.2 includes more than 80 predefined conditions that can be used to configure automatic actions in response to a system event. You can find details of RMC in Chapter 3, *IBM Cluster Systems Management for AIX 5L, Administration Guide*, SA22-7918 and in *A Practical Guide for Resource Monitoring and Control (RMC)*, SG24-6615.

Resource managers (RMs)

Resource managers are daemon processes that provide the interface between RMC and actual physical or logical resources. RSCT provides a core set of resource managers for managing base resources on single systems and across clusters. Additional resource managers are provided by cluster licensed program products (such as CSM, which contains the Domain Management resource manager).

The RSCT core resource managers are:

- ▶ Audit Log resource manager (AuditLogRM)
- ▶ Event Response resource manager (ERRM)
- ▶ File System resource manager (FileSystemRM)
- ▶ Host resource manager (HostRM)
- ▶ Sensor resource manager (SensorRM)

For details on RMC and the core resource managers, see Chapter 3 of *IBM Reliable Scalable Cluster Technology for AIX 5L, Administration Guide*, SA22-7889.

Cluster security services (CtSec)

CtSec is used by RMC to determine and authenticate a node within the cluster.

Note: This is not to be confused with authorization (granting or denying access to resources), which is handled by RMC.

CtSec uses credential-based authentication that enables:

- ▶ A client process to present information to a server in a way that cannot be imitated
- ▶ A server process to clearly identify the client and validity of the information

Credential-based authentication uses a third party that both the client and server trust. In this release of CSM, only UNIX host-based authentication is supported.

Group Services and Topology Services

Although included in RSCT, these are not used in the management domain structure of CSM. Group Services and Topology Services are used in peer domains for applications, such as HACMP/ES and GPFS. These provide node/process coordination and node/network failure detection. These services are often referred to as hags, or high availability groups services, and hats, or high availability topology services. Their corresponding daemons are known as hagsd and hatsd.

For more information about RSCT, refer to *IBM RSCT for AIX: Guide and Reference*, SA22-7889.

4.3.3 New in CSM 1.3.2 for AIX

CSM 1.3.2, announced in September, 2003, includes the following enhancements:

- ▶ Back up and restore of scripts for CSM.
- ▶ Install customization scripts.
- ▶ Install and hardware control probes for enhanced problem diagnosis.
- ▶ CSM installation and configuration usability enhancements.
- ▶ NIM enhancement to support secondary adapter configuration. This will allow additional adapters, such as Ethernet or IBM high performance switch adapters, to be configured at install time.
- ▶ Support of a new command **cmstat**, which provides a snapshot of a cluster similar to the **spmon -d** command in PSSP.

Other enhancements include:

- ▶ Linux on pSeries support
 - SuSE Linux Enterprise Server (SLES) 8 on pSeries p630, p630+, p650, p655, p615
 - Support for mixed AIX and Linux partitions
- ▶ 1024-way scaling (Linux on xSeries)
- ▶ RedHat 9 and SuSE 8.2 (Linux on xSeries)
- ▶ Replace use of System Installation Suite (SIS) with native SuSE Installer, AutoYaST for SuSE and SLES installs
- ▶ AMD Opteron support

4.3.4 Supported platform

CSM version 1.3.2 is supported on three platforms:

- ▶ AIX 5L on pSeries
- ▶ Linux on pSeries
- ▶ Linux on xSeries

Table 4-4 lists the platform requirements for CSM 1.3.2.

Table 4-4 CSM 1.3.2 software and hardware requirements

Operating systems	Hardware requirement
AIX 5.1 or 5.2 SLES 8	p615, p630, p650, p655, p670, p690
Red Hat 7.2, 7.3, or 8.0 Red Hat AS 2.1 SuSE 8.0 or 8.1 SLES 7 or 8	x330, x335, x342, x345, x360, x440 BladeCenter IntelliStation Model 6221

Note: AIX 5L Version 5.2 or higher is required for the management server.

Switch support

IBM intends to support the IBM @server High Performance Switch with CSM.

Note: Since the current SP Switch technology is dependent on PSSP, it is not supported by CSM on AIX.

4.3.5 PSSP-to-CSM transition

CSM is designed to be the next generation cluster systems management software. It encompasses and exploits the best of breed advantage of the highly successful, cluster-proven PSSP software and AIX operating systems to meet the demands of present and future cluster computing needs. Figure 4-11 shows a diagrammatic representation of the transition of the PSSP software, AIX operating systems, and CSM. The RSCT and VSD components, which used to be in PSSP, have now been developed and packaged in AIX.

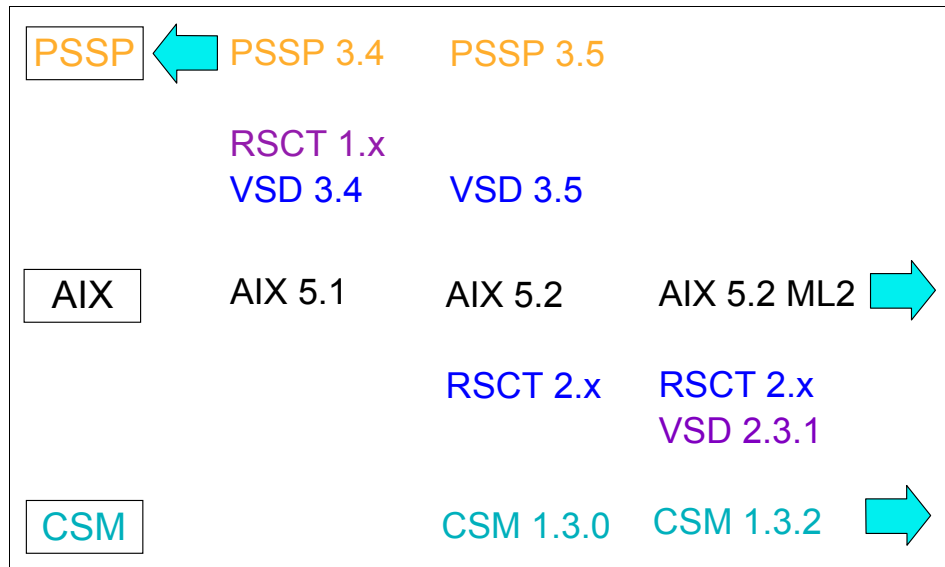


Figure 4-11 PSSP, AIX and CSM software components transition

4.3.6 Documentation references - CSM

The following are suggested documentation for further reading and references for CSM:

- ▶ *An Introduction to CSM 1.3 for AIX 5L*, SG24-6859
- ▶ *A Practical Guide for Resource Monitoring and Control (RMC)*, SG24-6615
- ▶ *IBM Cluster Systems Management for AIX 5L, Administration Guide*, SA22-7918
- ▶ *IBM Cluster Systems Management for AIX 5L, Planning and Installation Guide*, SA22-7919
- ▶ *IBM Reliable Scalable Cluster Technology for AIX 5L, Administration Guide*, SA22-7889

4.4 General Parallel File System (GPFS)

General Parallel File System (GPFS), the IBM high performance file system, provides a single, global file system for multiple nodes within a cluster. It allows parallel and serial applications shared access to files with high performance, availability, and scalability characteristics that surpass those of traditional distributed file system.

GPFS is used extensively on the IBM RS/6000 SP system to scale file system I/O in order to meet demanding requirements of applications in areas such as scientific and technical computing, business intelligence, and shared data access in server consolidation.

GPFS is offered on the IBM Cluster 1600, and on clusters of IBM pSeries nodes as GPFS for AIX and IBM Cluster 1350, and on clusters of selected IBM xSeries nodes as GPFS for Linux. In this section, the discussion will focus on GPFS for AIX.

GPFS for AIX 5L Version 2 is offered as program number 5765 -F64. You can also find information on GPFS for Linux in the white paper “An Introduction to GPFS v1.3 for Linux, June 2003”, available at:

http://www.ibm.com/servers/eserver/clusters/whitepapers/gpfs_linux_intro.html

Advantages of GPFS:

- ▶ GPFS is designed to provide high performance by striping I/O across multiple disks or storage subsystems (on multiple servers).
- ▶ The file system performance can scale with additional server nodes and disks storage added to the cluster.
- ▶ It can be used as a general purpose file system, as it is suitable for many kinds of workloads, from technical to commercial computing.
- ▶ GPFS provides global access (or uniform access) to files.
- ▶ GPFS offers many standard UNIX file system interfaces that cater to application portability.
- ▶ GPFS offers data consistency through a token-management system.
- ▶ GPFS is designed for high availability through file system logging, replication, and server and disk failover.

4.4.1 Architecture

A simple conceptual diagram of the GPFS environment is shown in Figure 4-12 on page 193. It shows storage devices with data access paths to a set of

application nodes within a GPFS nodeset. These nodes are capable of sharing the file system where the files are striped across the disks.

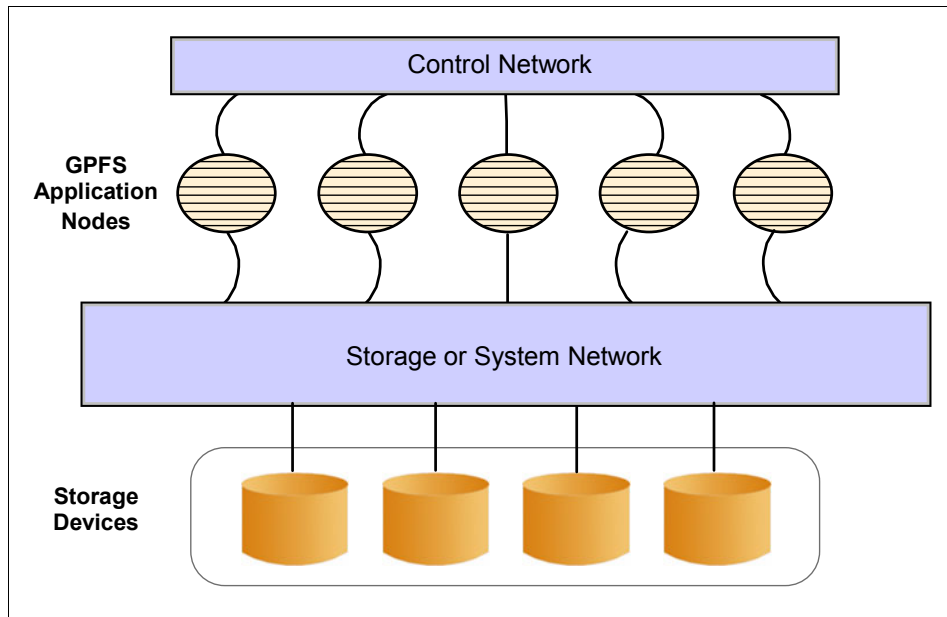


Figure 4-12 GPFS environment

There are two types of data access mechanisms and connectivity, and they are shown in the examples in Figure 4-13 on page 194 and Figure 4-14 on page 195.

Direct storage attachment to nodes in GPFS nodeset

Figure 4-13 on page 194 shows the direct disk attachment of the disk storage subsystem to each node of the GPFS nodeset. The connectivity may be Serial Storage Architecture (SSA) or *switched* fibre channel (FC).

Important: The use of FC loops without a switch was not fully tested at the time of writing. Physical connections other than SSA and FC are also theoretically possible, but untested. We recommend that you contact your IBM representative before trying GPFS in these environments.

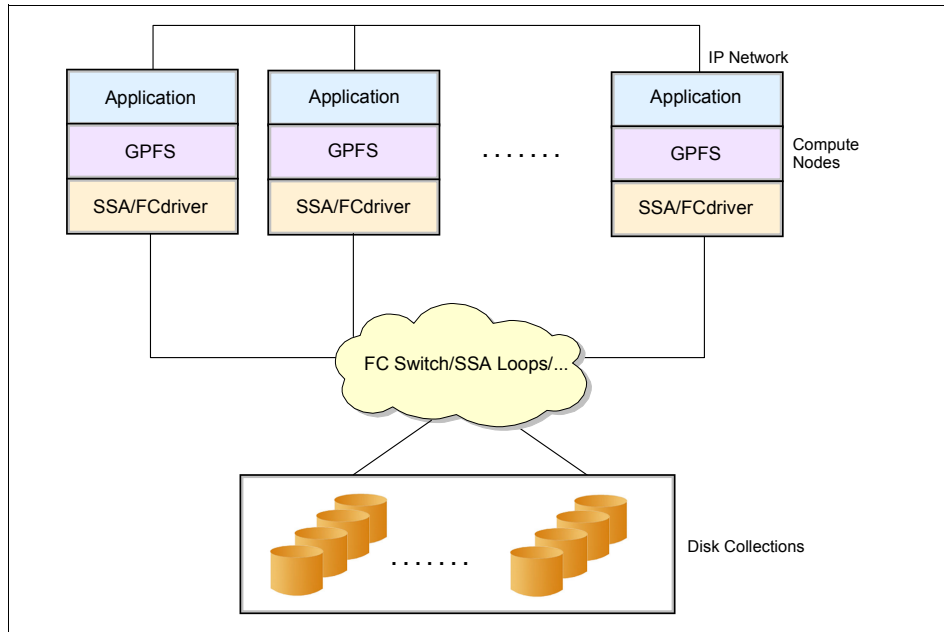


Figure 4-13 GPFS in a direct-attached SAN environment

Virtual Shared Disk (VSD)

Figure 4-14 on page 195 shows a mechanism involving the use of the Virtual Shared Disk (VSD). For this example, only a subset of all nodes, called the “I/O nodes” or “VSD server nodes”, have direct physical connectivity to the disks subsystem.

VSD is a software simulation of an SAN that runs across a high-speed interconnect (such as the IBM SP Switch2) which interconnects the nodes in the GPFS nodeset. The VSD servers act as “gateway”, routing disk operations from the disk subsystem to the application nodes.

Note: The use of the SP Switch will require the implementation of PSSP and its software prerequisites. You can choose to use CSM when using the IBM eServer High Performance switch for interconnecting the nodes.

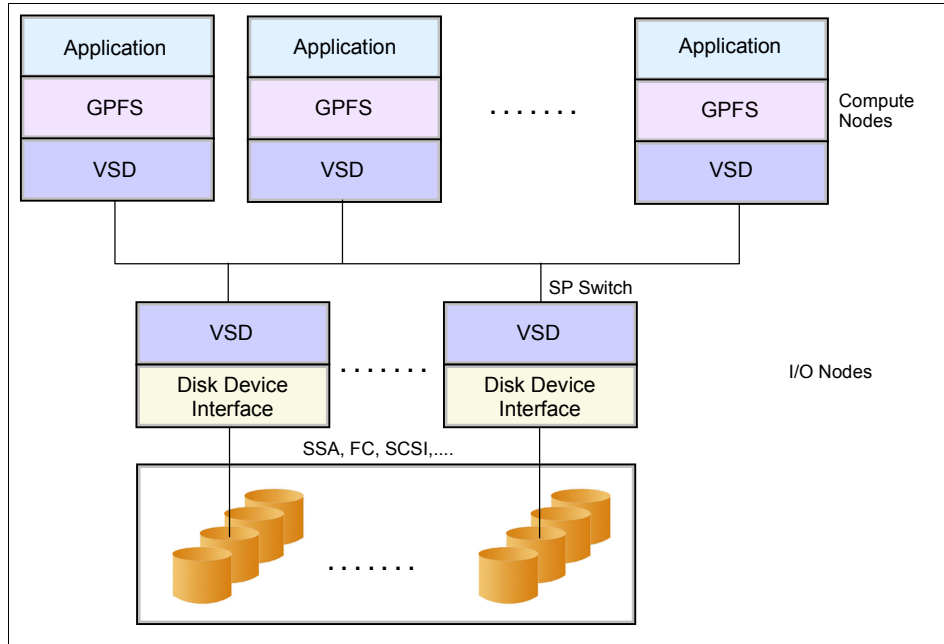


Figure 4-14 GPFS using VSD over a high performance interconnect

The IBM Recoverable Virtual Shared Disk (RVSD) component allows for failure recovery of the VSD. It automatically manages the VSDs by detecting error conditions, such as node failures, adapter failures, and disk failures, and then switches access to the disk from the primary node to the secondary node so that your application can continue to operate normally.

Tip: We recommend that GPFS installations twin-tail their storage to multiple RVSD servers if using RVSD. A simplified diagram of this setup is shown in Figure 4-15 on page 196.

For details of the GPFS architecture and operations, refer to the white paper *An Introduction to GPFS v2.1 for AIX*, available at:

http://www.ibm.com/servers/eserver/pseries/software/whitepapers/gpfs_intro.html

You can also refer to the white paper *GPFS Primer for AIX Clusters*, available at:

http://www.ibm.com/servers/eserver/pseries/software/whitepapers/gpfs_primer.html

For more useful information, see *General Parallel File System for AIX 5L in an RSCT Peer Domain, Concepts, Planning, and Installation Guide*, GA22-7940.

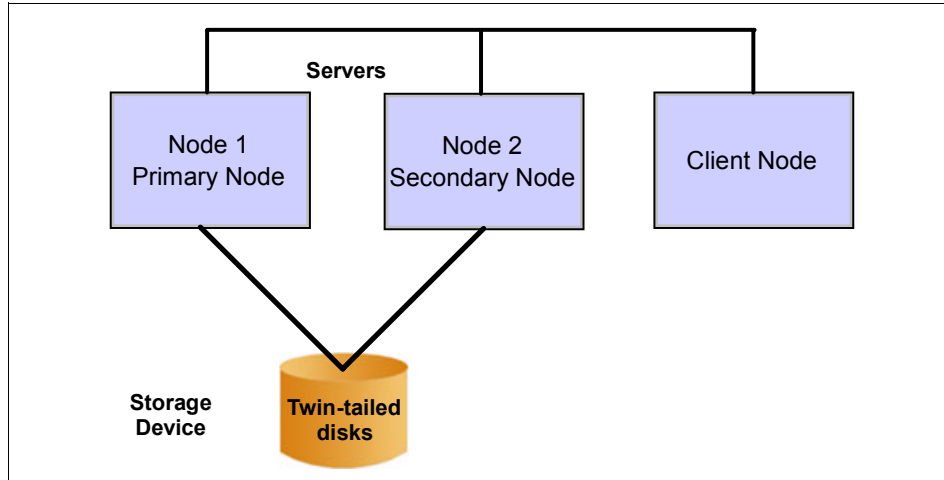


Figure 4-15 Simplified view of a twin-tailed disk

4.4.2 Administration and operation

GPFS provides functions that simplify multinode administration. It allows the system administrator to issue commands from any node/server in the cluster. These functions are based on, and are in addition to, the AIX administrative commands. A single GPFS multinode command can perform a file system function across the entire GPFS cluster. In addition, most existing UNIX utilities will also run unchanged. All of these capabilities allow GPFS to be used where parallel optimization is desired.

GPFS, like all file systems, provides a number of data management services, such as the ability to duplicate files, remove files, rename files, and so forth. GPFS supports the file system standards of X/Open 4.0, with minor exceptions. As a result, most AIX and UNIX applications can use GPFS data without modification and most UNIX utilities will run unchanged.

4.4.3 Higher performance/scalability

GPFS is designed to provide superior performance; it has the following features:

- ▶ It allows parallel applications simultaneous access to the same file or different files, from any node in the GPFS nodeset.
- ▶ It supports very large file systems, up to 100 TB per file system.
- ▶ It allows data striping across multiple storage devices or storage subsystems, to achieve increased aggregate bandwidth of your file system.

- ▶ It provides load balancing across all disks, to maximize their combined throughput.
- ▶ It allows concurrent read and write from multiple nodes.
- ▶ It provides enhanced parallel programming capabilities in conjunction with the implementation of the MPI-IO industry standard.

GPFS scales beyond single-server (node) performance limits by delivering file performance across multiple nodes and disks. In a parallel environment, GPFS can easily outperform other file systems such as Network File System (NFS), Journalled File System (JFS), and Distributed File System (DFS). Unlike NFS and JFS, GPFS file performance scales with additional file server nodes and disks added to the cluster.

Incremental improvements may be made to the file system by adding additional hardware of the same, or even lesser, capability. You can even add or delete disks from a mounted GPFS and mount a GPFS file system from every node/server in the cluster.

4.4.4 Recoverability

GPFS can survive many system and I/O failures. It is designed to transparently fail over locked servers and other GPFS central services. GPFS can be configured to automatically recover from node, disk connection, disk adapter, and communication network failures.

- ▶ In a PSSP cluster environment, availability is achieved through the use of the clustering technology capability of PSSP and PSSP Recoverable Virtual Shared Disk (RVSD) functions, or disk-specific recovery capabilities.
- ▶ In an AIX cluster environment, availability is achieved through the use of the cluster technology capabilities of either an RSCT peer domain or an HACMP cluster, in combination with the Logical Volume Manager (LVM) component or disk-specific recovery capabilities.

GPFS supports data and metadata replication to further reduce the chances of losing data if storage media fail. GPFS is a logging file system that allows the recreation of consistent structures for quicker recovery after node failures. GPFS also provides the capability to mount multiple file systems, each of which can have its own recovery scope in the event of component failures.

You can eliminate single points of failure in a GPFS solution by organizing the hardware into a number of failure groups. Establishing a redundant access path to the data will allow GPFS find an available path to the data at all times.

4.4.5 Migration

Upgrading to a new release of GPFS can be tested on a system currently running GPFS. This eases migration by allowing testing of a new level of code without inhibiting the production GPFS application.

4.4.6 New in GPFS 2.1 for AIX 5L

GPFS 2.1 for AIX 5L includes the following enhancements:

- ▶ It supports the 64-bit kernel in AIX 5L.
- ▶ High Availability Cluster Multi-Processing (HACMP) is no longer required to run GPFS in a non-SP or non-Parallel System Support Programs (non-PSSP) environment.
- ▶ It is supported on AIX 5L only.

4.4.7 Software requirements

Table 4-5 lists the software requirements for GPFS in a PSSP-managed cluster.

Table 4-5 Software requirements for GPFS in a PSSP environment

GPFS version	AIX version	PSSP version	Other software
2.1	5.1	3.5	IBM VSD and RVSD of PSSP
1.5	5.1 (32-bit kernel) 4.3.3	3.4	IBM VSD and RVSD of PSSP

Important: Note the following:

- ▶ The SP Switch or the SP Switch2 is required for node interconnect in a PSSP environment.
- ▶ VSD servers are required for disks connectivity.
- ▶ IBM intends to support AIX 5L Version 5.2 in December 2003 with PSSP 3.5 with a new release of the GPFS.

Table 4-6 lists the software requirements for GPFS in an RSCT Peer Domain (rpd)-managed cluster.

Table 4-6 Software requirements for GPFS in an rpd environment

GPFS version	AIX version	CSM version	Other software
2.1	5.2 5.1	1.3.2	CSM is not required for GPFS in an RPD cluster. The component that is required to create the peer domain is RSCT, which is delivered with AIX 5L.

Table 4-7 lists the software requirements for GPFS in an HACMP environment.

Table 4-7 Software requirements for GPFS in an HACMP environment

GPFS version	AIX version	HACMP version	CSM version	Other software
2.1	5.2	4.5	1.3.2	CSM is not required for GPFS in an HACMP cluster. The component that is required to create the peer domain is RSCT, which is delivered with HACMP 4.5.
2.1	5.1	4.5	1.3.2	CSM is not required for GPFS in an HACMP cluster. The component that is required to create the peer domain is RSCT, which is delivered with HACMP 4.5.

4.4.8 Documentation references

The following are suggested documents for further reading and references for GPFS:

- *An Introduction to GPFS v2.1 for AIX*, white paper, June 2003

http://www.ibm.com/servers/eserver/pseries/software/whitepapers/gpfs_intro.html

- *An Introduction to GPFS v1.3 for Linux*, white paper, June 2003

http://www.ibm.com/servers/eserver/clusters/whitepapers/gpfs_linux_intro.html

- *GPFS Primer for AIX Clusters*, white paper, March 24, 2003

http://www.ibm.com/servers/eserver/pseries/software/whitepapers/gpfs_primer.html

- *GPFS Primer for Linux Clusters*, white paper, March 24, 2003

http://www.ibm.com/servers/eserver/clusters/whitepapers/gpfs_linux_primer.html

General Parallel File System Questions and Answers (a useful resource for answers to frequently asked questions about GPFS)

http://www.ibm.com/servers/eserver/pseries/software/sp/gpfs_faq.pdf

- *General Parallel File System for AIX 5L, PSSP Clusters Concepts, Planning, and Installation Guide*, GA22-7899
- *General Parallel File System for AIX 5L, AIX Clusters Concepts, Planning, and Installation Guide*, GA22-7895
- *General Parallel File System for AIX 5L in an RSCT Peer Domain, Concepts, Planning, and Installation Guide*, GA22-7940

4.5 LoadLeveler

LoadLeveler for AIX 5L is an integral part of the total systems management solution for the Cluster 1600, RS/6000 SP nodes, or clusters of pSeries or RS/6000 servers. It is a network job management and scheduling system that allows users to run more jobs in less time by matching the jobs' processing needs with the available resources. Figure 4-16 shows an example of a LoadLeveler configuration.

LoadLeveler (LL) supports load balancing for interactive Parallel Operating Environment sessions, and focuses on parallel job scheduling and scalability. It can also schedule jobs written for Network Queuing System (NQS) to run on machines outside the LoadLeveler cluster.

LoadLeveler for AIX 5L, Version 3 is available as program number 5765-E69.

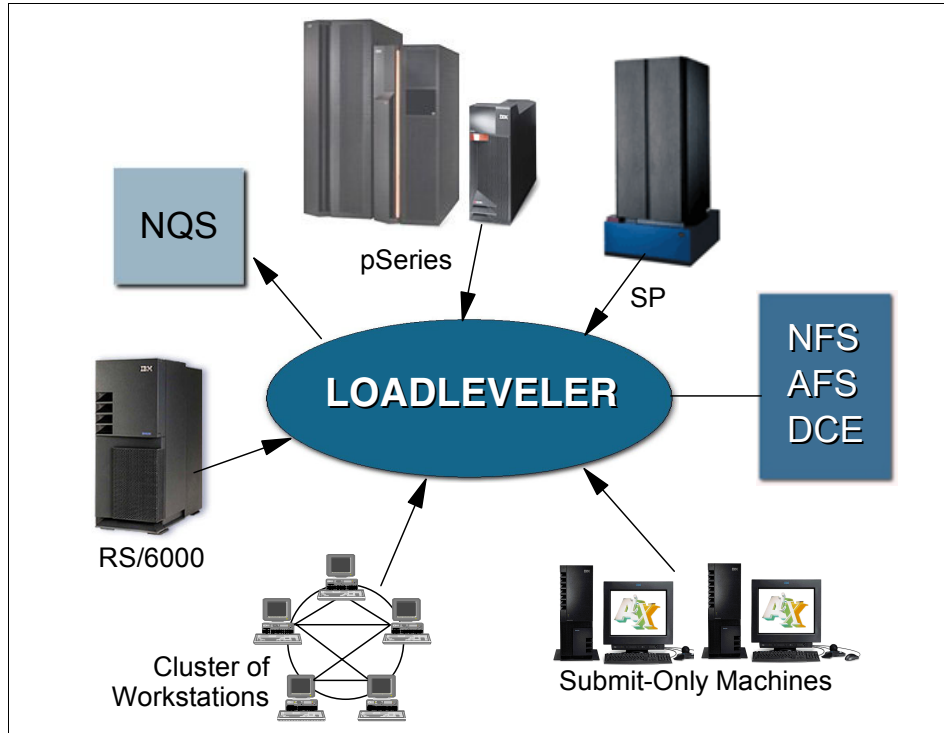


Figure 4-16 Example of a LoadLeveler configuration

LoadLeveler offers the following advantages:

- ▶ The efficient scheduling algorithm helps improve overall system utilization and responsiveness.
- ▶ It offers a single point of control for administration, job management and workload scheduling.
- ▶ LoadLeveler provides users full scalability across processors and jobs, and supports jobs scaling across hundreds of cluster nodes.
- ▶ It provides checkpoint/restart enhancements to facilitate job recovery.
- ▶ It exploits many features of AIX and integrates with the AIX 5L Workload Manager (WLM) for resource enforcement and monitoring. For details, refer to *Workload Management with LoadLeveler*, SG24-6038.

4.5.1 Administration and operations

LoadLeveler schedules jobs, and provides functions for building, submitting, and processing jobs quickly and efficiently in one or more cluster nodes under its

control. The batch jobs can be submitted, either in serial or parallel, to run in the background without input from the user. LoadLeveler accepts the job and reviews its requirements; it then determines, from the availability of consumable resources, on which cluster nodes the job will run. (Consumable resources include items like the number of CPUs and the amount of memory.)

LoadLeveler interfaces

LoadLeveler uses three types of interfaces. They allow users to build, create, submit, and delete jobs, and allow system administrators to configure LoadLeveler, control running jobs, and do accounting.

- ▶ **Common line interface**

This interface provides basic and administrative functions.

- ▶ **Application programming interface (API)**

This interface allows application programs written by users to interact with the LoadLeveler environment.

- ▶ **Graphical User Interface (GUI)**

This interface is similar to the Command Line Interface and is designed for ease of use. (However, experienced users and administrators may find the command line interface more efficient than the GUI.)

LoadLeveler cluster

LoadLeveler can submit jobs to a machine if that machine is configured as a member of a LoadLeveler cluster. The machines in a LoadLeveler cluster can have one or more roles in job scheduling:

- ▶ **Scheduling node**

This is used to manage jobs from submission to completion.

- ▶ **Central manager**

This is a central resource manager and workload balancer. One or more alternate central managers can be set up to prevent a single point of failure.

- ▶ **Execute node**

Jobs dispatched by the central manager are run here.

- ▶ **Submit node**

This is used to submit jobs to LoadLeveler from outside the cluster, and runs no daemon. (For an explanation of daemons, refer to “LoadLeveler daemons” on page 203.)

An example of job scheduling in LoadLeveler is shown in Figure 4-17 on page 203.

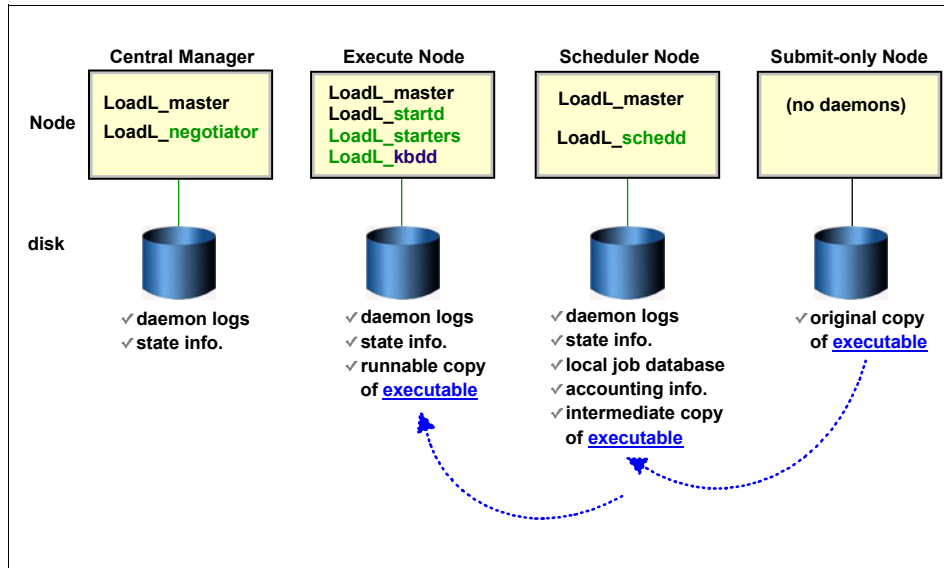


Figure 4-17 An example of job scheduling in a LoadLeveler cluster

LoadLeveler daemons

LoadLeveler operates through a set of daemons. These daemons control the processes that move the jobs through the LoadLeveler cluster. These process control daemons are managed by a master daemon. You can find detailed descriptions of LoadLeveler daemons in Chapter 1 of *IBM LoadLeveler for AIX 5L: Using and Administering*, SA22-7881; other chapters contain complete information on configuring, administering, and using LoadLeveler.

Job definition

Each step of a job is defined in a job command file. LoadLeveler uses the information specified in the job command file to identify the job and to execute the job steps. A detailed description of submitting and managing jobs can be found in Chapter 5 of *IBM LoadLeveler for AIX 5L: Using and Administering*, SA22-7881.

4.5.2 Capabilities

LoadLeveler has the following capabilities, including parallel processing, individual control, central control, and more.

Parallel processing

LoadLeveler interfaces with the Parallel Environment (PE) software to support batch jobs in a parallel operating environment (POE). Parallel Environment is described in “Parallel Environment (PE)” on page 216.

Individual control

Owners of machines in the LoadLeveler cluster can restrict the use of their resources. They can specify the availability of their resources and how they are to be used. For example, some users might allow their workstations to accept any job during the night—but only certain jobs during the day, when they need their resources most. Others might allow LoadLeveler to monitor their keyboard activity and make their workstations available whenever it has been idle for a period of time.

Central control

LoadLeveler gives administrators a complete view of the status of all jobs on the system and the resources available to those jobs. It allows administrators to change the availability of these resources to best meet their computing needs.

Scalability

Nodes can be added to the LoadLeveler cluster to increase the consumable resources available in LoadLeveler.

Automated job tracking

Process tracking allows LoadLeveler to remove any processes (throughout the entire cluster) that are left behind when a job terminates. This ensures that resources held by the orphaned processes are freed for future use. You can also automate the process of removing old checkpoint files that are no longer needed on the system.

Multiple user space tasks

LoadLeveler supports multiple user space tasks on both SP Switch and SP Switch2. Parallel Applications that use LAPI user space API or Parallel Environment MPI and LoadLeveler can use up to 4096 user space tasks.

Additional scheduling algorithms

LoadLeveler supports the Gang Scheduler and Backfill scheduler, in addition to standard reservation scheduling, for more efficient use of system resources.

4.5.3 New in LoadLeveler 3.2

LoadLeveler 3.2, introduced in September, 2003, includes the following enhancements:

- ▶ AIX Cluster Security Services

LoadLeveler now includes an additional security mechanism, AIX Cluster Security Services (CtSec), which allows for identity authentication of LoadLeveler daemons and users. These identities are then used to determine authorization, thus preventing unauthorized users and programs from misusing resources or disrupting services. Cluster Security Services is a security mechanism similar to DCE, but tailored for a clustered environment.

- ▶ Multi-adapter connection support

LoadLeveler now supports multi-adapter connection support. This means you can now use multiple adapters for job processing, and allocate separate memory support for each. You can also request multiple windows per task per protocol, and have the adapters automatically share the workload for the number of requested windows. To facilitate this, LoadLeveler now refers to an adapter by its connection to the network, not by its name.

- ▶ High performance switch support

LoadLeveler now supports the IBM eServer pSeries High Performance Switch (HPS). The HPS provides significantly greater communication bandwidth, along with reduced latency and improved fault tolerance. It is able to support multiple adapters per network per node and multiple windows on the same adapter or across multiple adapters

- Enhancements for running in an RSCT peer domain

A new command, **llextrPD**, is added to help you set up your administration files. It extracts machine and adapter data from the RSCT peer domain and formats it into stanzas for use in building administration files.

A new keyword is added to specify whether GSmonitor uses RMC API access or PSSP SDR access routines:

- The keyword is CSMONITOR_DOMAIN=PEER for RMC API access.
- The keyword is CSMONITOR_DOMAIN=PSSP For PSSP SDR access routines.

- Dynamic adapter configuration support

LoadLeveler now offers the option to dynamically detect and handle adapters and adapter changes for machines in the cluster. This option is available only on RSCT peer domain clusters.

- ▶ Automatic resume for drained startd
Improved switch management and recovery now allows LoadLeveler to automatically resume a startd that was drained due to switch table unload errors.
- ▶ Support for POE Wait option for interactive jobs
Interactive jobs can now be queued to wait for resources to become available.
- ▶ Enhancements to the **llmodify** command
Attributes such as job class, account number for idle job steps, and job step wall clock limits can be modified with the **llmodify** command.
- ▶ Enhancements to Data Access API
Job class information can be queried through the LoadLeveler Data Access API.

4.5.4 New in LoadLeveler 3.1

LoadLeveler 3.1, announced on November 13, 2001, includes the following enhancements:

- ▶ Striped Communication

Striping increases performance, because it allows the job to use multiple communication paths to the node.

- ▶ AIX Workload Manager Integration

LoadLeveler can be configured to use WLM to enforce usage of consumable CPU and memory.

- ▶ Gang Scheduling

This can be used to improve the overall system utilization and responsiveness to parallel workloads.

- ▶ Checkpoint/Restart

This function is enhanced so that it allows you to periodically save the state of jobs automatically.

- ▶ 64-bit support

Support for 64-bit applications for interactive and batch jobs that run on POWER nodes and AIX 5L is introduced.

- ▶ Multiple User Space Tasks per adapter

This is supported on the SP Switch and SP Switch2.

4.5.5 Software requirement

Table 4-8 lists the software requirements for LoadLeveler on PSSP.

Table 4-8 Software requirements for LoadLeveler 3.1 and 3.2 on PSSP

LL version	AIX version	Domain	Other software
3.2	5.2 or higher	RSCT 2.3.1 and PSSP 3.5	<ul style="list-style-type: none">▶ PE for AIX Version 4.1, if you plan to run POE jobs in user space mode.▶ Java Runtime Environment filesets (Java131.rte.bin, Java131.rte.lib 1.3.1), if you plan to use the graphical user interface or configuration tasks wizard.▶ TaskGuide Runtime Environment fileset (sysmgt.sguide.rte 5.2), if you plan to use the configuration tasks wizard.▶ RSCT 2.3.1.0, if you plan to run in an RSCT peer domain or configure LoadLeveler to support RSCT security services.
3.1	5.1 or higher	3.4 or 3.5	<ul style="list-style-type: none">▶ PE for AIX Version 3.2, if you plan to run POE jobs in user space mode.▶ If you plan to configure LoadLeveler to support DCE security services, then fileset ssp.clients 3.2 is required.▶ Java Runtime Environment filesets (Java130.rte.bin, Java130.rte.lib 1.3.0), if you plan to use the graphical user interface or configuration tasks wizard.▶ TaskGuide Runtime Environment fileset (sysmgt.sguide.rte 5.1), if you plan to use the configuration tasks wizard.

4.5.6 LoadLeveler configuration suggestions

The following are some suggestions for LoadLeveler configurations:

- ▶ Place LoadLeveler executables in the shared file system.
- ▶ Place logs, execute, spool in the local file system (/var).
- ▶ Place configuration files in the shared file system.
- ▶ Place local configuration files in a subdirectory.
- ▶ Soft-link worker local configuration files.

4.5.7 Documentation references - LoadLeveler

The following are suggested documents for further reading and reference on LoadLeveler:

- ▶ *Workload Management with LoadLeveler*, SG24-6038
- ▶ *IBM LoadLeveler for AIX 5L: Using and Administering*, SA22-7881

4.6 Scientific subroutine libraries

In this section, we describe the scientific subroutine libraries and their features.

4.6.1 Engineering and Scientific Subroutines Library (ESSL) family of products

The Engineering and Scientific Subroutine Library (ESSL) family of products consists of:

- ▶ Parallel Engineering and Scientific Subroutine Library (Parallel ESSL) for Advanced Interactive Executive (AIX), program number 5765-F84
- ▶ Engineering and Scientific Subroutine Library (ESSL) for AIX, program number 5765-F82

These products are state-of-the-art collections of mathematical subroutines that provide a wide range of mathematical functions for many different scientific and engineering applications.

You can use these subroutine libraries to develop and enable many different types of scientific and engineering applications. New applications can be designed and developed to take full advantage of all the capabilities of ESSL product family. Existing applications can be enabled by replacing comparable subroutines and in-line code with calls to ESSL subroutines.

Some of the types of applications that can take advantage of the ESSL capabilities are:

- ▶ Structural analysis time series analysis
- ▶ Computational chemistry computational techniques
- ▶ Fluid dynamics analysis mathematical analysis
- ▶ Seismic analysis dynamic systems simulation
- ▶ Reservoir modeling nuclear engineering
- ▶ Quantitative analysis electronic circuit design
- ▶ Time series analysis
- ▶ Computational techniques
- ▶ Mathematical analysis
- ▶ Dynamic systems simulation
- ▶ Nuclear engineering
- ▶ Electronic circuit design

Advantages of ESSL and Parallel ESSL

These subroutines can be called from application programs written in Fortran, C, and C++. They have been designed with an easy-to-use call interface and error-handling ability. They are compatible with public domain subroutine libraries such as Basic Linear Algebra Subprograms (BLAS), Scalable Linear Algebra Package (ScaLAPACK), and Parallel Basic Linear Algebra Subprograms (PBLAS), thereby making it easy to migrate from these libraries.

You can obtain high performance on SMP processors without requiring extensive knowledge of parallelization techniques. The subroutines support a 64-bit environment.

4.6.2 Operations

ESSL and Parallel ESSL offer mathematical subroutines available in several computational areas, and these are summarized in Table 4-9.

Table 4-9 Wide range of mathematical functions for different applications

	ESSL	Parallel ESSL
Computational areas	<ul style="list-style-type: none"> ▶ Linear Algebra Subprograms ▶ Matrix Operations ▶ Linear Algebraic Equations ▶ Eigensystem Analysis ▶ Fourier Transforms, Convolutions and Correlations, and Related Computations ▶ Sorting and Searching ▶ Interpolation ▶ Numerical Quadrature ▶ Random Number Generation ▶ Utilities 	<ul style="list-style-type: none"> ▶ Level 2 PBLAS ▶ Level 3 PBLAS ▶ Linear Algebraic Equations ▶ Eigensystem Analysis and Singular Value Analysis ▶ Fourier Transforms ▶ Random Number Generation ▶ Utilities

	ESSL	Parallel ESSL
Supported platform	<ul style="list-style-type: none"> ▶ POWER, POWER2, POWER3, POWER3-II, POWER4, POWER4+, PowerPC®, and Symmetric Multi-Processing (SMP) PowerPC processors. ▶ IBM RS/6000 SP 	<ul style="list-style-type: none"> ▶ IBM RS/6000 SP ▶ Clusters of pSeries and RS/6000 workstations

For details of the computational areas, refer to *IBM ESSL Products General Information*, GA22-7903.

4.6.3 New in ESSL 4.1

The ESSL libraries are tuned for the POWER4 and POWER4+ processors. ESSL now supports the AIX 5L V5.2 32-bit and 64-bit kernels.

The Dense Linear Algebraic Equation Subroutines has been enhanced to include new LAPACK subroutines as follows:

- ▶ General Matrix Factorization and Multiple Right-Hand Side Solve
- ▶ Positive Definite Real Symmetric Matrix Factorization and Multiple Right-Hand Side Solve
- ▶ Positive Definite Real Symmetric or Complex Hermitian Matrix Factorization and Multiple Right-Hand Side Solve
- ▶ Positive Definite Real Symmetric Matrix Factorization
- ▶ Positive Definite Real Symmetric Matrix Multiple Right-Hand Side Solve
- ▶ General Matrix Inverse
- ▶ Positive Definitive Complex Hermitian Matrix Inverse
- ▶ Triangular Matrix Inverse

SMP support has been added to the SCFT and FDCFT subroutines when computing a single, large transform.

4.6.4 New in Parallel ESSL 3.1

The Parallel ESSL Libraries are tuned for the POWER4 and POWER4+ processors.

The Dense Linear Algebraic Equations Subroutines now include:

- ▶ Positive Definite Real Symmetric and Complex Hermitian Matrix Inverse

- ▶ Estimation of the Reciprocal of the Condition Number of a Positive Definite Real Symmetric or Complex Hermitian Matrix Inverse

The Eigensystems Analysis Subroutines now include:

- ▶ Reduce a General Matrix to Bidiagonal Form
- ▶ Singular Value Decomposition of a General Matrix

The Utilities now include:

- ▶ Initialize a Type-1 Array Descriptor
- ▶ Initialize a Type-1 Array Descriptor with Error Checking
- ▶ Compute the Ceiling of the Division of Two Integers
- ▶ Compute the Least Common Multiple of Two Positive Integers
- ▶ Compute the Local Row or Column Index of a Global Element of a Block-Cyclically Distributed Matrix
- ▶ Compute the Process Row or Column Index of a Global Element of a Block-Cyclically Distributed Matrix
- ▶ Compute the Global Row or Column Index of a Local Element of a Block-Cyclically Distributed Matrix
- ▶ Compute the Starting Local Row or Column Index and Process Row or Column Index of a Global Element of a Block-Cyclically Distributed Matrix
- ▶ Compute the Starting Local Row and Column Indices and the Process Row and Column Indices of a Global Element of a Block-Cyclically Distributed Matrix
- ▶ Real Symmetric, Complex Symmetric or Complex Hermitian Matrix Norm

4.6.5 Software requirements

The software requirements for ESSL version 4.1 and 3.3 and Parallel ESSL version 3.1 and 2.3 are tabulated in this section.

Table 4-10 lists the software requirements for ESSL.

Table 4-10 Requirements for ESSL 4.1 and 3.3

ESSL version	AIX version	Other software
4.1	5.2 with 5200-01	<p>For compiling (one of these):</p> <ul style="list-style-type: none"> ▶ XL Fortran for AIX, Version 8.1 or later (5765-F70) ▶ IBM VisualAge C++ Professional for AIX Version 6.0 (5765-F56) ▶ C for AIX, Version 6.0 (5765-F57) <p>For linking, loading or running:</p> <ul style="list-style-type: none"> ▶ XL Fortran Run-Time Environment for AIX, Version 8.1, or later (5765-F71) ▶ C libraries (included in the AIX 5L V5 Application Development Toolkit)
3.3	5.1 or higher	<p>For compiling (one of these):</p> <ul style="list-style-type: none"> ▶ XL Fortran for AIX, Version 7.1.1 (5765-E02) ▶ IBM VisualAge C++ Professional for AIX Version 5.0.2 (5765-E26) ▶ C for AIX, Version 5.0.2 (5765-E32) <p>For linking, loading, or running:</p> <ul style="list-style-type: none"> ▶ XL Fortran Run-Time Environment for AIX, Version 7.1.1 (5765-E03) ▶ C libraries (included in the AIX 5L Version 5 Application Development Toolkit)

The software requirements for Parallel ESSL are listed in Table 4-11.

Table 4-11 Requirements for Parallel ESSL 3.1 and 2.3

PESSL version	AIX version	PSSP version	Other software
3.1	5.2	3.5 or higher	<p>For compiling (one of these):</p> <ul style="list-style-type: none"> ▶ XL Fortran for AIX, Version ▶ IBM VisualAge C++ Professional ▶ C for AIX, Version <p>For linking, loading, or running:</p> <ul style="list-style-type: none"> ▶ XL Fortran Run-Time Environment for AIX, Version ▶ ESSL V4.1 ▶ Parallel Environment for AIX Version 4.1 ▶ C libraries (included in the AIX 5L Version 5 Application Development Toolkit)
2.3	AIX 5L Version 5.1 or higher	3.4 or higher	<p>For compiling (one of these):</p> <ul style="list-style-type: none"> ▶ XL Fortran for AIX, Version 7.1.1 (5765-E02) ▶ IBM VisualAge C++ Professional ▶ C for AIX, Version 5.0.2 (5765-E32) <p>For linking, loading, or running:</p> <ul style="list-style-type: none"> ▶ XL Fortran Run-Time Environment for AIX, Version 7.1.1 (5765-E03) ▶ ESSL V3.3 (5765-C42) ▶ Parallel Environment for AIX Version 3.2 (5765-D93) ▶ C libraries (included in the AIX 5L Version 5 Application Development Toolkit)

4.6.6 Documentation references - ESSL and PESSL

The following are suggested documents for further reading and reference on ESSL and PESSL:

- ▶ *ESSL Products General Information*, GA22-7903
- ▶ *ESSL Guide and Reference*, SA22-7904
- ▶ *ESSL Installation Guide*, GA22-7886
- ▶ *Parallel ESSL Guide and Reference*, SA22-7906

4.6.7 Mathematical Acceleration Subsystem (MASS)

Another Scientific Subroutine Library that you can use with the pSeries server or Cluster 1600 is the Mathematical Acceleration Subsystem (MASS). MASS is not an orderable product from IBM; it is provided for use as a download from the MASS support Web site:

<http://techsupport.services.ibm.com/server/mass>

The download is available at no additional charge to users on the World Wide Web, subject to a set of terms and conditions.

Note: MASS is provided *as is*. IBM makes no warranties, expressed or implied, including the implied warranties of merchantability and fitness for a particular purpose. IBM has no obligation to defend or indemnify against any claim of infringement, including, but not limited to, patents, copyright, trade secret, or intellectual property rights of any kind.

MASS consists of libraries of tuned mathematical intrinsic functions. Each new MASS version includes the material from previous versions that has not been changed, except that Version 3.0 no longer includes the library that runs only on POWER2 processors. The POWER2 library remains available in Version 2.7 of MASS. For details of the versions of MASS, refer to the MASS support Web site.

Scalar library

The MASS scalar library, libmass.a, contains an accelerated set of frequently used math intrinsic functions in the AIX system library libm.a (now called libxlf90.a in the IBM XL Fortran publication).

sqrt, rsqrt, exp, log, sin, cos, tan, atan, atan2, sinh, cosh, tanh, dnint, x**y (FORTRAN), pow (C)

The libmass.a library can be used with either FORTRAN or C applications and will run under AIX on all of the IBM pSeries and RS/6000 processors. Because MASS does not check its environment, it must be called with the IEEE rounding mode set to round-to-nearest and with exceptions masked off (the default XLF environment). MASS may not work properly with other settings. In some cases MASS is not as accurate as the system library, and it may handle edge cases differently from libm.a (sqrt(Inf), for example). The trigonometric functions (sin, cos, tan) return NaN (Not-a-Number) for large arguments ($\text{abs}(x) > 2^{50} \pi$).

These functions accept double-precision arguments and return a double-precision result. You can refer to the MASS support Web site for details of FORTRAN and C declarations for the functions.

Vector libraries

The general vector library, `libmassv.a`, contains vector functions that will run on all computers in the IBM pSeries and RS/6000 families. The library `libmassvp3.a` contains some functions that have been tuned for the POWER3 architecture, while the remaining functions are identical to those in `libmass.a`.

Similarly, the library `libmassvp4.a` contains some functions that have been tuned for the POWER4 architecture, while the remaining functions are identical to those in `libmass.a`.

The vector libraries `libmassv.a`, `libmassvp3.a`, and `libmassvp4.a` can be used with either FORTRAN or C applications. When calling the library functions from C, only call by reference is supported, even for scalar arguments. As with the scalar functions, the vector functions must be called with the IEEE rounding mode set to round-to-nearest and with exceptions masked off. The accuracy of the vector functions is comparable to that of the corresponding scalar functions in `libmass.a`, though results may not be bit-wise identical.

► Double-precision functions:

`vrec`, `vdiv`, `vsqrt`, `vsqrt`, `vexp`, `vlog`, `vsin`, `vcos`, `vtan`, `vasin`, `vacos`, `vatan2`,
`vsincos`, `vcosisin`, `vdint`, `vdnint`

► Single-precision functions:

`vsrec`, `vsdiv`, `vssqrt`, `vsqrt`, `vsexp`, `vslog`, `vssin`, `vscos`, `vstan`, `vsasin`, `vsacos`,
`vsatan2`, `vssincos`, `vscosisin`, `vsdint`, `vsdnint`

These functions accept double-precision (or single-precision) vector input and output arguments, and an integer vector-length parameter. You can refer to the MASS support Web site for details of FORTRAN and C declarations for the functions.

The MASS vector FORTRAN source library enables application developers to write portable vector codes. The source library, `libmassv.f`, includes FORTRAN versions of all the vector functions in the MASS vector libraries.

Further information on MASS

Information about the installation, use, performance, and accuracy of the MASS libraries is beyond the scope of the redbook and is not discussed here. You can obtain this information from the MASS support Web site:

<http://techsupport.services.ibm.com/server/mass>

For questions, you may write to:

MASS Support
IBM Toronto Laboratory
Mail Stop D2-515
8200 Warden Avenue
Markham, Ontario, L6G 1C7
Canada

Or send e-mail to:

`masslib@ca.ibm.com`

4.7 Parallel Environment (PE)

IBM Parallel Environment (PE) for AIX is a high-function environment for the development and execution of parallel applications on the Cluster 1600 system. It allows organizations to develop, debug, analyze, tune, and execute parallel C/C++ and FORTRAN applications that use the industry-standard message passing interface. PE is available as program number 5765-D93.

The advantages of Parallel Environment include the following:

- ▶ It facilitates parallel application development and portability to clusters of pSeries or RS/6000 systems.
- ▶ It exploits MPI message passing on all nodes, including symmetric multiprocessors (SMPs).
- ▶ It supports Low-level Application Programming Interface (LAPI) programs.
- ▶ It provides enhanced tools for program debugging and application performance analysis.
- ▶ LoadLeveler can be used for POE batch jobs.
- ▶ It allows you to checkpoint and restart POE jobs.

4.7.1 Parallel Programming support

PE supports two basic parallel programming models:

- ▶ Single Program Multiple Data (SPMD)
In this model, the programs running the parallel tasks are identical, but the tasks work on different sets of data.
- ▶ Multiple Program Multiple Data (MPMD)
In this model, different programs may be running on each node.

- ▶ PE is a distributed memory message-passing system. Two types of subroutine calls are available to the developer to parallelize the application:
- ▶ Message Passing Interface (MPI)
- ▶ Low-level Application Programming Interface (LAPI)

The processor nodes that handle the parallel tasks communicate using the following protocols:

- ▶ Internet Protocol (IP) Communication Subsystem
- ▶ User Space (US) Communication Subsystem

Note: Although LAPI is used for data communication in conjunction with PE, it is available as a component of PSSP or shipped as part of RSCT in AIX 5.2.

4.7.2 Operation

PE consists of the following:

Message passing and collective communication API subroutine libraries

Developers can use these subroutines for code parallelization. For details of the message passing subroutine calls, refer to *IBM Parallel Environment for AIX: MPI Subroutine Reference*, SA22-7423, and *IBM Parallel Environment for AIX: Hitchhiker's Guide*, SA22-7424.

Parallel Operating Environment (POE)

POE is an execution environment that facilitates a smooth transition from serial to parallel processing. It allows you to invoke your parallel program from a home node and have the partition manager start the parallel tasks on a number of remote nodes that you specify in a host list file. A number of parallel compiler scripts and POE environment variables are available to enhance the operation of POE.

Debugging and profiling tools

The following tools are available in PE to facilitate debugging and profiling:

- ▶ Parallel Debugging Facility - the pdbx facility is a line-oriented parallel debugger based on the dbx debugger.
- ▶ Parallel Profiling Capability - once the parallel program is debugged, you can profile the program using the AIX Xprofiler graphical performance tool and the AIX commands `prof` and `gprof`.

Performance analysis tools

The PE Benchmarker tools consist of a suite of applications and utilities:

- ▶ Performance collection tools
This set of tools is built using the Dynamic Probe Class Library (DPCL) where probes are placed in running executables to collect the required information.
- ▶ Unified Trace Environment (UTE) utilities
This is a set of enhanced trace utilities that allows you to obtain an MPI trace of all or part of a parallel application.
- ▶ Performance visualization tools
This set of tools allows you to view system profiles from data collected from the performance collection tools.

Note: DPCL is no longer part of the PE licensed program, but is shipped with PE for convenience. It is available as an open source offering that supports PE.

For details of the operations and use of PE, refer to *IBM Parallel Environment for AIX: Operation and Use, Volume 1*, SA22-7948, and *Volume 2*, SA22-7949.

4.7.3 New in PE 4.1

PE version 4.1 includes the following functional enhancements:

- ▶ MPI is enhanced to use LAPI as the common transport protocol, and provides collective communications performance enhancements.
- ▶ PE now supports the new generation of pSeries switches, in addition to the pSeries (RS/6000 SP) switches.
- ▶ Threaded library support only is provided, with binary compatibility for signal (non-threaded) library applications.
- ▶ Electronic Licensing is provided.
- ▶ Cluster-based security support is provided.
- ▶ Program Marker Array is removed.
- ▶ MPL is no longer supported.
- ▶ Xprofiler is now part of AIX.
- ▶ The parallel utility subroutine MP_QUERYINTRDELAY, mpc_queryintrdelay is no longer supported. When invoked, it returns a value of zero.

- ▶ There is a new library to be used when converting MPI trace files to the slog2 file format used by the latest version of Jumpshot (an MPI trace viewer program available from Argonne National Laboratory).
- ▶ User-written SIGIO handlers are no longer invoked when a packet arrives.

4.7.4 Software requirements

Table 4-12 lists the software requirements of PE.

Table 4-12 *Software requirements for PE 4.1 and 3.2*

PE version	AIX version	PSSP version	Other software
4.1	5.2 or higher	3.5	<ul style="list-style-type: none"> ▶ XL Fortran V or later, if FORTRAN programs are to be compiled and run. FORTRAN Run-Time Environment for AIX V or later, if FORTRAN programs are to be run. ▶ If User Space batch or interactive jobs are to be submitted under PE, then LoadLeveler V3.1 (5765-E69) is required. ▶ At least one concurrent use license of C for AIX compiler or C++ compiler installed on the SP/server complex that includes the control workstation. ▶ PE Benchmark requires Java Runtime Environment V1.3
3.2	5.1	3.4 or higher. PSSP is not required when Parallel Environment is used on a standalone or independently managed cluster of RS/6000 servers.	<ul style="list-style-type: none"> ▶ XL Fortran, V7.1.0.2 or later, if FORTRAN programs are to be compiled and run. XL Fortran Run-time Environment for AIX Version 7.1.1 or later, if FORTRAN programs are to be run. ▶ If User Space batch or interactive jobs are to be submitted under PE, then LoadLeveler V3.1 (5765-E69) is required. ▶ At least one concurrent use license of C for AIX compiler or C++ compiler installed on the SP/server complex that includes the control workstation. ▶ PE Benchmark requires Java Runtime Environment V1.3.

4.7.5 Documentation references - Parallel Environment (PE)

The following resources are suggested references for further reading and references:

- ▶ *PE for AIX 5L Installation*, GA22-7418
- ▶ *PE for AIX 5L Hitchhiker's Guide*, SA22-7424
- ▶ *PE for AIX 5L Operation and Use, Volume 1*, SA22-7948
- ▶ *PE for AIX 5L Operation and Use, Volume 2*, SA22-7949
- ▶ *PE for AIX 5L MPI Programming Guide*, SA22-7422
- ▶ *PE for AIX 5L MPI Subroutine Reference*, SA22-7423

4.8 IBM High Availability Cluster Multi-Processing for AIX (HACMP)

HACMP for AIX is a proven high availability solution for business-critical environments. For over 10 years, HACMP has been providing reliable high-availability services, monitoring capabilities, and dependable detection of application failures.

HACMP manages the failover of business application environments to backup servers. With the introduction of the new optional package, HACMP/XD (Extended Distance), HACMP will also manage failover to backup servers at remote sites.

HACMP/XD provides long distance remote failover for ESS/PPRC peers, and unlimited distance failover for IP-connected peers using proven IBM High Availability Geographic Cluster (HAGEO) technology. Now there is a single, world-class source of protection for mission-critical applications. HACMP is offered as program number 5765-F62.

It is important to understand that a solution with HACMP is not a fault-tolerant solution—but when properly designed and configured, it helps to minimize unplanned outage. The term “high availability” refers to a computing configuration that has the ability to detect and recover from failures and provide a better level of protection against system downtime than the standard hardware and software alone.

An HACMP solution provides these means to detect and recover from any unplanned server hardware and application failures. It also gives you the means to take down an individual server (node) for planned maintenance and upgrades, without having to take down the entire cluster.

HACMP offers the following advantages:

- ▶ It minimizes expensive downtime for both planned and unplanned outages by quickly restoring essential services, as follows:
 - HACMP makes use of redundant hardware configured in a cluster to keep an application running, restarting it on a backup server if necessary.
 - HACMP can also detect software problems that are not severe enough to interrupt proper operation of the system, such as process failure or exhaustion of system resources.
 - HACMP can monitor, detect and react to failure events, allowing the system to stay available during the occurrence of random and unexpected problems.
 - HACMP can be configured to react to hundreds of system events.
- ▶ It provides the flexibility to accommodate changing business needs.
- ▶ It provides horizontal growth with rock-solid reliability, where up to 32 servers can participate in an HACMP cluster.
- ▶ HACMP exploits the systems and network management capabilities of AIX 5L and RSCT.

4.8.1 HACMP operations

The primary goal of high availability clustering software is to minimize or ideally, eliminate, the need to take your resources out of service during maintenance and reconfiguration activities.

HACMP is a software product that executes on each node in a loosely coupled cluster. It provides application availability by detecting and reacting to failures of systems, processors, adapters, networks, disks, or applications. When these failures occur, HACMP makes use of redundant hardware in the cluster to keep the application running. In the event of a complete node failure, HACMP restarts the application on a backup node.

HACMP responds to failure and recovery events based on policies specified when the cluster was defined. It also allows the systems administrator to customize extra operations or accommodate additional resource types.

Components of an HACMP cluster

A HACMP cluster consists broadly of the following components:

- ▶ Physical resources
These are the hardware nodes, network interfaces and volume groups.

► Logical resources

These are the applications, service IP addresses and mounted file systems, that can be instantiated on any one of an equivalent set of physical resources. For example:

- An application can run on any of a set of nodes.
- A service IP address can be made active on any of a set of network interfaces.
- A volume group can be varied on any of a set of nodes.

► Resource groups

Resource groups allow you to combine related resources into a single logical entity for easier management. They are moved by HACMP from one node to another when the conditions in the cluster change.

► Policies

Policies determine which physical resource will hold a logical resource, when there are multiple choices available.

HACMP will then, in accordance with the specified policies, move logical resources around so as to keep applications running despite hardware and software failures.

The takeover relationships among cluster nodes determine which cluster nodes control a resource group and which cluster nodes take over control of the resource group when the original node relinquishes control. Takeover relationships are defined by assigning one of the following HACMP takeover relationships:

Cascading	The resource is always owned by the active node having the highest priority for that resource.
Rotating	The resource rotates among all the nodes defined in the chain. The node with the highest priority in the resource chain for the resource takes over for a failed node, but does not immediately return the resources when the failed node returns.
Concurrent	The resource is owned and accessed simultaneously by all owning nodes.
Custom	The administrator has the choice and control over the takeover policy.

4.8.2 Administration and operation

HACMP management and administration facilities include the following:

- ▶ The Cluster Manager in HACMP is responsible for monitoring the status of cluster resources, reacting to failures, and responding to administrative requests. It uses the RSCT services in AIX.
- ▶ HAView allows you to monitor HACMP clusters through the NetView® network management platform.
- ▶ HATivoli allows you to monitor the state of an HACMP cluster and its components through your Tivoli Framework enterprise management system.
- ▶ Version Compatibility allows nodes running earlier versions of HACMP to interoperate with those running HACMP V5.1. A customer can upgrade an existing cluster running HACMP Version V4.5 or Version 4.4, without taking the entire cluster offline.
- ▶ The Cluster Snapshot captures a cluster configuration, creating text files that contain all the information necessary to configure a similar cluster.
- ▶ Cluster Single Point of Control (C-SPOC) enables the user to perform certain common administrative operations across the cluster from a single SMIT session.
- ▶ Dynamic Reconfiguration allows the user to change the configuration of a running cluster. The changes take effect immediately, without having to stop and restart the HACMP daemons, and without having to disrupt the applications running on the cluster.
- ▶ Resource Group Management (clRGmove) enables customers to use the SMIT interface to move resource groups between nodes and to bring them online or offline. Other options allow groups to be “stuck” to a particular node or to temporarily suspend application monitoring.

The Application Availability Analysis Tool provides a tool for measuring application availability. A log file is maintained to capture application and node startup and outages. The analysis tool reads the log and generates a report including availability metrics.

For details of HACMP systems administration, see *High Availability Cluster Multi-Processing for AIX, Administration and Troubleshooting Guide*, SC23-4862.

4.8.3 New in HACMP 5.1

HACMP 5.1, announced in July 2003, includes the following enhancements:

- ▶ Consolidation of all previous forms of HACMP (HAS, CRM, ES, ESCRM) into a single HACMP offering.
- ▶ Reduced failover time using fast disk takeover (which happens within 10 seconds).
- ▶ A streamlined configuration interface that requires only six user inputs to build a simple HA cluster.
- ▶ Non-IP heartbeating protection over disks where no additional hardware is required.
- ▶ Enhanced security mechanism, removing the need for `/.rhosts`.
- ▶ Increased administration productivity through faster cluster verification and synchronization.
- ▶ Greater control over resources owning application startup and failover behavior.
- ▶ More cluster status information readily available in the cluster monitor.
- ▶ The addition of multiple disaster recovery technologies to keep the system accessible if disaster strikes.

4.8.4 Software requirements

Table 4-13 lists the software requirements for HACMP 5.1 on PSSP- and CSM-managed clusters.

Table 4-13 Software requirements for HACMP 5.1

AIX version	PSSP version	CSM version	Other software
5.1 with 5100-03 5.2 with 5200-01	3.5	1.3.0 or later	▶ Neither PSSP nor CSM is required to configure HACMP 5.1.

4.8.5 Documentation references - HACMP

The following are suggested documents for further reading and reference on HACMP:

- ▶ *HACMP for AIX: Concepts and Facilities*, SC23-4864
- ▶ *HACMP for AIX: Planning and Installation Guide*, SC23-4861
- ▶ *HACMP for AIX: Administration and Troubleshooting Guide*, SC23-4862
- ▶ *HACMP for AIX: Programming Client Applications*, SC23-4865
- ▶ *HACMP for AIX: Programming Locking Applications*, SC23-4866

- ▶ *HACMP for AIX: Glossary, SC23-4867*
- ▶ *HACMP Remote Copy: ESS PPRC Guide, SC23-4863*

4.9 Performance Toolbox (PTX) and Performance AIDE (PAIDE)

The AIX Performance Toolbox (PTX) and Performance AIDE (PAIDE) are licensed programs designed to support pSeries and RS/6000 systems. It provides a comprehensive tool for monitoring and tuning system performance in distributed environments. PTX is offered as program number 5765-E74, and PAIDE is offered as program number 5765-E68.

PTX and PAIDE offer the following advantages:

- ▶ They provide a quick and simple solutions for obtaining and analyzing detailed system information.
- ▶ They allow customization of views for monitoring.
- ▶ They can generate reports on system activity over hours, days, or weeks.
- ▶ They support distributed performance monitoring.
- ▶ They provide access to thousands of system metrics.

4.9.1 Administration and operation

PTX

PTX provides an easy method of monitoring a single server or managing a distributed environment in the following ways:

- ▶ It can provide a high level view of system performance.
- ▶ It generates reports over specified time periods, or “drills down” on a specific system, to collect data or analyze recorded data. These reports can be customized to show hourly, weekly, or monthly trends.
- ▶ It provides access to thousands of performance metrics.
- ▶ It provides customizable graphical interfaces to view system status and tune parameters.
- ▶ It provides a customizable menu interface.
- ▶ Built-in analysis tools combine with 2-dimensional and 3-dimensional graphing capabilities to pinpoint immediate or long-term bottlenecks.
- ▶ It provides additional utilities to convert recorded data into formats for import by third-party spreadsheets.

PAIDE

PAIDE can be used in conjunction with PTX to allow the user to visualize live performance on multiple systems concurrently.

PAIDE offers the following functions:

- ▶ It provides agents for creating 24/7 recordings of large sets of performance metrics.
- ▶ It provides data filtering based on user-customized criteria.

The filtering criteria can be used for event generation to a monitoring console, or to execute administration scripts.

A new Java-based interface, Jtopas, focuses on pre-filtering the set of performance metrics available on large-scale systems to provide simplified snapshot views of overall system activity.

4.9.2 Platform requirements

Table 4-14 lists the platform requirements for PTX and PAIDE for POWER Version 3.

Table 4-14 Platform requirements for PTX and PAIDE for POWER version 3

Hardware	AIX
<ul style="list-style-type: none">▶ RS/6000 or pSeries system with a minimum of 256 MB memory▶ Graphics adapter for Performance Toolbox user interface operation▶ Network adapter needed for Performance Toolbox remote monitoring	V5.1 with 5100-01

4.10 Software ordering and configuration

You can order the Cluster 1600 software from your IBM sales representative or IBM Business Partners. A technical review of your requirements with an IBM technical representative may be helpful.

Note for IBM sales representative and IBM Business Partners:

The following resources can assist you in configuring and ordering Cluster 1600 software:

- ▶ IBM Universal Sales Manual
- ▶ *IBM eServer pSeries & RS/6000 Cluster Software Ordering Guide*, available from the IBM Systems Sales Web site:
<http://www.ibm.com/servers/eserver/pseries>



Solutions and offerings “best practices”

This chapter presents scenarios and solutions that came to our attention while writing this redbook. To the degree possible, the team has verified each of the solutions and best practices. Also, where possible, we have solicited feedback from subject matter experts.

We cover the following scenarios and solutions:

- ▶ High Performance Computing (HPC) - view of a hypothetical HPC environment
- ▶ Replacing SP nodes with LPAR technology
- ▶ Virtual serial port implications with LPAR technology
- ▶ Hardware Management Console and Web-based System Manager considerations
- ▶ HMC and Web-based System Manager coexisting with firewall

5.1 High Performance Computing (HPC) environment

In this section, we discuss the Cluster 1600 in a hypothetical high performance computing environment. The discussion includes applications on different hardware that can be clustered and incorporated as part of the Cluster 1600.

5.1.1 A hypothetical solution

The example in this section is based on common HPC requirements in the field.

Note: The example shown here does not refer to or imply any particular installation, nor has this example been specifically configured and tested. It is used here solely to apply the concepts and details described in earlier chapters.

Background requirement

This hypothetical solution assumes that this HPC customer technical requirement includes workload ranging from 32-bit and 64-bit applications running both serial and parallel applications. Some applications also require large shared memory.

Some applications use Message Passing Interface (MPI) and require high bandwidth, low latency, node-to-node communications, and the compute platform must be able to scale as the workload requirement increases.

For performance reasons, the ability to share files globally in the same application domain is strongly preferred. Some applications (such as databases) are mission critical, and require high availability. Job scheduling is required to ensure optimized job throughput and resource usage.

5.1.2 Solution architecture

Figure 5-1 on page 231 shows a hypothetical solution of a Cluster 1600 solution that consists of pSeries servers running AIX and Linux, as well as xSeries servers running Linux to meet the requirements stated in “Background requirement”.

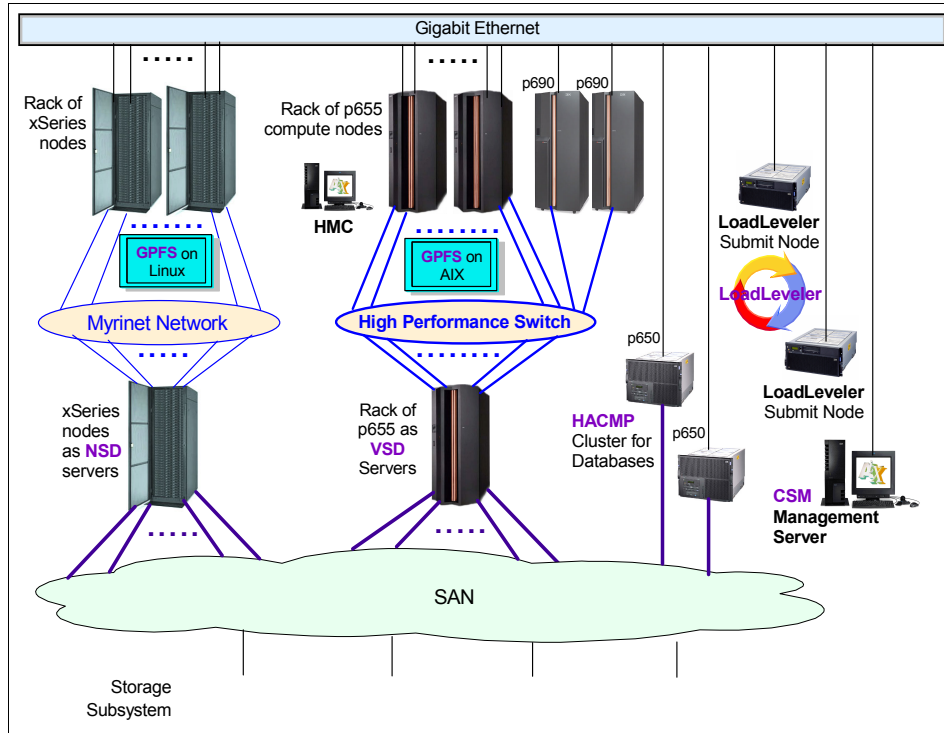


Figure 5-1 A hypothetical HPC environment

This environment contains several components, which are described in the following sections.

xSeries servers

The rack of two of xSeries nodes are running the Linux operating system. Some of these nodes are used as Network Shared Disk (NSD) to manage GPFS I/O, while the rest are used fully for computational purposes. The NSD nodes have data connectivity to the Storage Area Network (SAN). All the nodes are interconnected via a high speed interconnect, which can be Gigabit Ethernet or Myrinet. Detailed discussion of the Cluster 1350 is beyond the scope of this redbook; for more information on that topic, refer to “Documentation and suggested reading” on page 237.

pSeries servers

The pSeries servers in the cluster run the AIX operating system. They consist of the following:

- ▶ p655s used for computation for jobs requirement that are no larger than a 4-way scaling.
- ▶ p690s with up to 32 processor and up to 512 GB of memory are used for applications that require large shared memory and high memory bandwidth.
- ▶ A set of servers, which can be p655 or LPAR p690, are used as Virtual Shared Disk (VSD) servers with data connectivity to the SAN.
- ▶ A pair of p650s or p670s configured for high availability for mission-critical databases.
- ▶ The compute p655 and p690 are interconnected via a high bandwidth, low latency High Performance Switch (HPS).
- ▶ A pair of p615s or p630s that are used for submitting jobs to the pSeries clusters.
- ▶ A p630 that is used as the CSM management server.

You can find the details of the pSeries servers and the HPS in Chapter 2, “Cluster 1600 hardware” on page 11.

Storage Area Network (SAN)

The SAN consists of storage subsystems such as the Enterprise Storage Server® (ESS) and FAStT Storage Server (FAStT), Tape Storage and SAN switches. Details of the SAN are beyond the scope of this redbook; for further information on that subject, refer to *Introduction to Storage Area Networks*, SG24-5470.

Network

The communication network consists of the following:

- ▶ A private compute network, such as the myrinet network for xSeries and HPS network for pSeries (p655 and p690).
- ▶ A common Ethernet backbone, such as the Gigabit Ethernet (1000Mbps) or Fast Ethernet (100 Mbps), can be used for CSM installs and client access.

5.1.3 Solution discussion

In this section, we provide a discussion of the various components and considerations for designing clusters.

pSeries compute cluster

The pSeries compute cluster is discussed in terms of the p655 and p690 platforms and the software covered in Chapter 2, “Cluster 1600 hardware” on page 11 and Chapter 4, “Software support” on page 169.

p690 for shared memory applications

The p690 is an ideal platform for workloads requiring large shared memory. It can also be clustered together to create compute clusters to run large parallel workloads with bigger compute requirements per node than the p655.

The p690 is also an ideal platform to run multiple logical partitions (LPAR) to allow consolidation of application servers on the same hardware. The same is true when LPARs are created separately for a development environment.

The LPARs can also be dynamically reconfigured to allow flexibility and adaptability to changing workload requirements. This is accomplished with the dynamic logical partitions (DLPAR) feature which allows processor, memory, and I/O slot resources to be added to or deleted from running partitions, or moved between running partitions without requiring any AIX instance to be rebooted. You can find details of using LPAR in *pSeries Systems Handbook 2003 Edition*, SG24-5120 and *The Complete Partitioning Guide for IBM eServer pSeries Servers*, SG24-7039.

Important: DLPAR is supported on AIX 5L 5.2 or later.

You can consider creating an on-demand environment using the IBM DLPAR tool set for pSeries. This application consists of a set of tools that enhance the usability of the DLPAR feature in AIX 5.2. In this tool set, both time-based and load-based scenarios for moving processor resources and memory resources among partitions are explored. It includes sample scripts for monitoring LPARs and automating DLPAR operations, which you can use or customize to fit your requirements.

Tip: The IBM DLPAR tool set for pSeries is currently available at the alphaWorks Web site. For details and download, go to:

<http://www.alphaworks.ibm.com/tech/dlpar>

IBM intends to include the features of the DLPAR Tool Set for pSeries in the next release of AIX 5L 5.3.

p655 compute cluster

The p655 compute clusters provide a good compute platform for both 32-bit and 64-bit applications. Besides running serial applications on the individual nodes,

the p655 is ideally suited for running parallel applications across the cluster. The p655 uses the same MCM technology as the p690 and p670. It provides an advantage to applications requiring high memory bandwidth.

Parallel computing

Parallel jobs using MPI on the cluster of p655 and p690 can involve message passing via the HPS. It can also be executed in the multiprocessor environment of the p690. The Parallel Environment (PE) can be set up for running parallel jobs using the Parallel Operating Environment (POE). The scientific subroutines, such as the Parallel ESSL, can be used for parallel computing, while ESSL or MASS can be used for serial or shared-memory applications. For discussions of PE, ESSL, PESSL and MASS, see 4.7, “Parallel Environment (PE)” on page 216 and 4.6, “Scientific subroutine libraries” on page 208.

Job scheduling on the pSeries

The compute jobs on these nodes can be managed with job scheduling. LoadLeveler can be configured in this cluster to schedule the jobs. Two submit nodes allow users to submit their jobs for queuing and execution on the cluster. These nodes can also function as the *central manager* and the *scheduler node*; refer to 4.5, “LoadLeveler” on page 200 for further information.

Parallel file system

Files can be shared and accessed in parallel using General Parallel File System (GPFS). This is set up using Virtual Shared Disk (VSD) servers. The mechanism involves software simulation of a SAN which runs across the HPS interconnecting the nodes in the GPFS node set.

The node set in this case refers to the compute nodes participating in the sharing of the GPFS. GPFS gives you the flexibility to group and manage the cluster nodes into one or more nodeset. The VSD servers act as “gateway”, routing disk operations from the disk subsystem to the compute cluster node; refer to 4.4, “General Parallel File System (GPFS)” on page 192 for more information on GPFS.

High availability solution

A pair of p650s, as shown in Figure 5-2, are used to serve the purpose of a database server. The database server is clustered using IBM HACMP software. Both servers have direct fiber channel connectivity to the SAN and use “heartbeat” over IP and “heartbeat” over shared disks.

In this scenario, we are assuming that there is only a single database instance running on the active server. In the case of a failover situation, the database instance is started on the standby server.

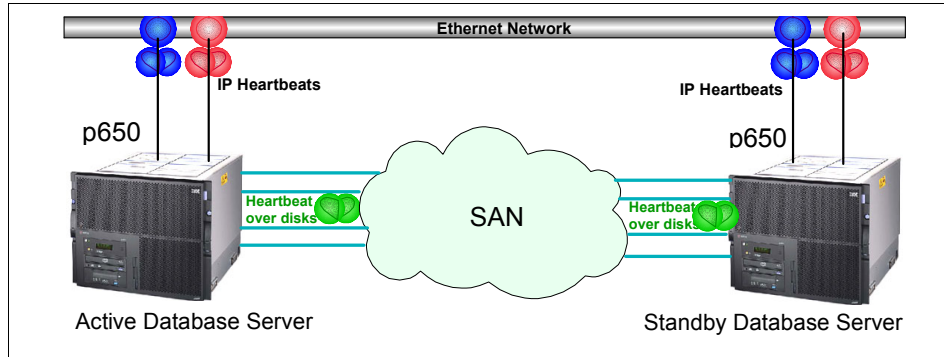


Figure 5-2 A basic two-server HACMP cluster configuration

There is flexibility in designing high availability clusters using HACMP and LPAR. If there is more than one instance of the databases, you can keep them isolated in different LPARs and run them on the same server—or on different servers.

As an example, if we have two instances of databases, the design of the high availability cluster can look like the logical diagram shown in Figure 5-3 on page 235. Distributing the active database server across the physical hardware ensures that system resources are better utilized and that a failure of the hardware only affects the databases running on that hardware. The individual LPAR can be managed as a single server within the CSM cluster.

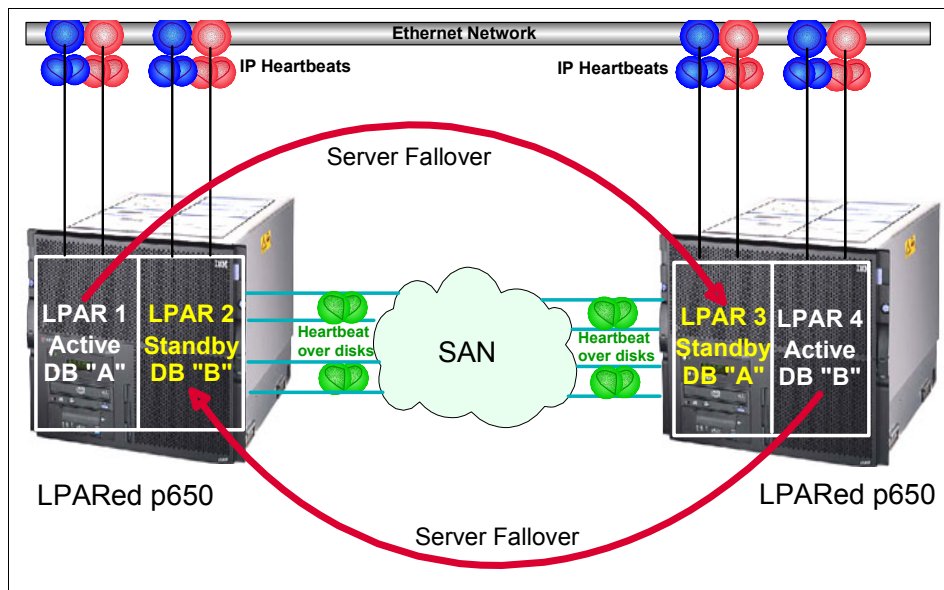


Figure 5-3 HACMP configuration in an LPAR environment

xSeries compute cluster

The xSeries compute cluster provides similar functions of parallel computing to the p655 clusters. The xSeries is a 32-bit platform and may provide better price/performance than the 64-bit p655 platform for certain applications. It is ideally suited for certain 32-bit Linux applications.

The NSD works similarly to the traditional VSD on the pSeries in terms of functionality, and the GPFS on the Cluster 1350 allows high performance file sharing on the Cluster 1350. Jobs on the xSeries cluster servers can be scheduled via some third-party applications, such as Load Sharing Facility offered by Platform Computing, or some open source applications such as Portable Batch Scheduler (PBS). More information can be found at the following sites:

<http://platform.com>

<http://www.openpbs.org>

5.1.4 Cluster management considerations

In this section, we discuss cluster management of a heterogeneous solution consisting of AIX on pSeries and Linux on xSeries.

Cluster Systems Management (CSM)

This heterogeneous platform can be centrally managed via a CSM management server, making the management server the single point of control. CSM provides low cost, consistent management of distributed and clustered IBM pSeries and xSeries servers. Refer to 4.3, “Cluster Systems Management (CSM)” on page 180 for more information on CSM.

Tip: In addition to AIX on pSeries and Linux on xSeries servers, selected pSeries running the Linux operating system can also be managed via CSM in the same cluster. This provides the benefit of a broader range of Linux source codes that can be compiled on Linux on pSeries. You can find details of Linux on pSeries at the following Web site:

<http://www.ibm.com/servers/eserver/pseries/linux>

Compilers such as IBM XL Fortran for Linux on pSeries and IBM VisualAge C++ for Linux on pSeries are also available. These compilers are optimized for the pSeries platform. Evaluation copies are currently available at:

<http://www.ibm.com/developerworks/offers/linux-speed-start/download-p.html>

Managing clusters across geographies

In “Cluster management across geographies” on page 185, we discuss cluster management across geography using CSM. This concept may be applied in some scenarios where the clusters are housed separately in separate data centers.

For example, the xSeries clusters may be owned by Department A and located physically in Building A, while the pSeries cluster may be owned by Department B and housed physically in Building B. The system administrator operates out of the operations center in Building C. The management server can be located at the operations center in Building C, and yet be able to provide a single point of management of the clusters at the different locations.

Documentation and suggested reading

The following documents are suggestions for further reading and references:

- ▶ *Linux Cluster with CSM and GPFS*, SG24-6601
- ▶ *IBM @server xSeries Clusters Planning Guide*, SG24-5845
- ▶ *Linux Handbook: A Guide to IBM Linux Solutions and Resources*, SG24-7000
- ▶ *Introduction to Storage Area Networks*, SG24-5470
- ▶ *pSeries Systems Handbook 2003 Edition*, SG24-5120
- ▶ *The Complete Partitioning Guide for IBM @server pSeries Servers*, SG24-7039

5.2 Transition from SP nodes to LPARs

Many IBM pSeries install bases are already utilizing POWER3-based servers, and are in-plan to refresh and/or replace their production environments with pSeries servers. In this section we present one of the scenarios that will be most common in the realm of Cluster 1600: it addresses install bases trying to migrate off a “node” in an SP frame into partitions in LPAR-enabled systems such as p690, p670, p655, p650, or p630.

Note: In your case, you can use different methods to perform this migration, or devise creative solutions or scripts on your own. However, in this redbook we only present a solution that has been implemented in our own sandbox environment and that has been shown to be effective in this type of migration.

Moving to a POWER4 hardware ensures that you can benefit from the enhanced features of the improved hardware architecture. For more details on POWER4, refer to Chapter 2, “Cluster 1600 hardware” on page 11.

As you know, SP nodes are POWER3 architecture and all of the new LPAR-enabled servers are POWER4 or POWER4+ architectures. Therefore, for simplistic reference, we refer to the SP nodes and LPAR servers as POWER3 and POWER4+, respectively.

5.2.1 System migration by utilizing alternate disk migration

Alternate disk migration is one of the methods available to migrate a server from an older version of AIX (for example, AIX 4.3.x) to a newer version of AIX (for example, AIX 5L) without disrupting the production environment during the course of the server migration from POWER3 to POWER4+. Using this method, you need a spare disk that is equal or greater in size in rootvg to use for the migration.

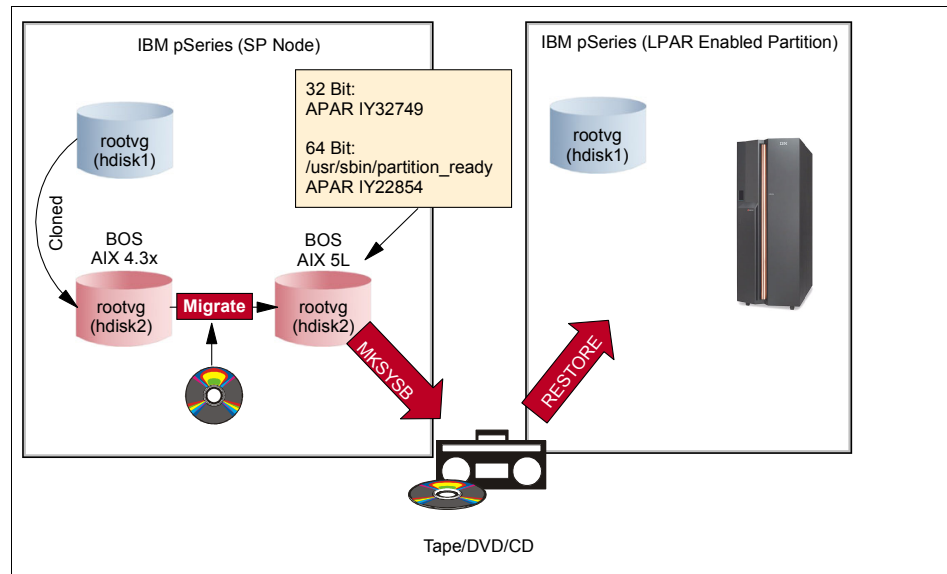


Figure 5-4 alt_diskinstall process

The advantage to using this method is that you are able to clone your rootvg and then migrate it from AIX 4.3.x to AIX 5L while your server's production rootvg is kept safely if rollback is needed. By utilizing this method, you are in fact creating two bootable images, as shown in Figure 5-4:

- ▶ Your original rootvg with the AIX 4.x bootable image intact (hdisk1)
- ▶ The cloned rootvg to which the migration from AIX 4.3.x to AIX 5L can be applied (hdisk2)

Once the process of migration has been completed and verified, the downtime associated with normal BOS updates and migrations are minimized because you are able to migrate the rootvg in the background. All you need is a reboot, with the primary boot device pointing to the *migrated* (AIX 5L rootvg). This methodology offers a useful rollback or contingency plan, if there is a need for one.

Steps for alt_diskinstall migration

To migrate from POWER3 to POWER4++™ by utilizing the alt_diskinstall method, follow these steps:

1. Enter: smitty alt_clone
 - a. Select -> **Software Installation and Maintenance -> Alternate Disk Installation.**
 - b. Clone rootvg to an alternate disk.
2. Modify bootlist to boot from the cloned physical device (hdisk-x).
3. Reboot.

Once booted with the cloned rootvg, verify that APAR IY32749 has been properly applied

1. If your hardware is 64-bit architecture, verify that these filesets are installed:
 - devices.chrp_lpar.base.ras
 - devices.chrp_lpar.base.rte
 - devices.chrp.base.rte
 - devices.chrp.base.ServiceRM
 - bos.64bit
 - bos.mp64
 - APAR IY22854
2. Execute: /usr/sbin/partition_ready.
3. Create a mksysb of the system onto either a DVD or bootable tape media.

If your rootvg spans one physical disk (hdisk0 - rootvg), the results of alt_diskinstall create a second rootvg on a pdisk that you specify (that is, hdisk2 - altinst_rootvg). If you issue the command **lspv**, you will see that hidsk0 is still “active”. At this point, you need to modify your bootlist to boot from hdisk2.

Upon reboot, issue an **lspv** command once again to verify that the hdisk0’s label has changed to “old_rootvg”, and that hdisk2’s label has indeed changed to “rootvg” and is set to “active”. This will ensure that the 5L migration you are about to execute is applied to the cloned alternate disk.

Note: More detailed guidance about this migration can be found at:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/howto/HT_insgdrf_altldiskinstall_clone.htm#insgdrf_altldiskinstall_clone

Important: AIX 5L version 5.1, recommended maintenance level 5100-03, is the minimum AIX install CD set that supports 32-bit-to-64-bit install migration. This version ensures that the operating system image is bootable on a partitioned POWER4 system.

Using an earlier package (for example, AIX 5L version 5.1, recommended maintenance level 5100-01) causes the booting process to halt on LED 0c43 when you boot after restoring the backup to an LPAR.

Tasks to be performed on the POWER4++ partition

On the POWER4++ partition, perform these tasks:

1. Allocate a bootable tape drive or DVD (in a media format that is compatible with the POWER3 system).
2. Load the media.
3. Boot the LPAR to an SMS state, and select tape/ DVD device as the boot device.
4. Issue the **boot** command.
5. Configure the booted images as needed

Your BOS migration from POWER3 to POWER4+ should now be completed and ready for non-system data to be migrated to the LPAR system.

5.3 Virtual Serial port implications with LPARS

In various discussion forums for IBM pSeries, the following question is often asked.

Question: Can the integrated serial port on a LPAR enabled pSeries be used for an HACMP Heartbeat?

Answer: Yes, but only in SMP (non-LPAR) mode. If the pSeries is configured in LPAR mode, it requires an 8-port Async adapter. You will need to find other means. Also note that the p655 Model 651 has no built-in native serial port. Refer to “Moving the media drawer between active LPARs” on page 243 for more detail.

5.3.1 Console device

The HMC provides one virtual tty console access for each partition (LPAR), which removes most of the need for partition access to native serial ports. The two native serial ports on the p670/p690 are under the control of the service processor and are not available for functions such as HACMP heartbeat cabling

or UPS control, since they may be dynamically allocated to a single LPAR at any time.

For HACMP heartbeat applications on LPAR-enabled servers, you may choose to utilize “heartbeat on disk”, which requires no other hardware. In AIX 5L, RSCT utilizes its Topology Services layer to provide a heartbeat mechanism to ensure application takeover can occur at any time.

Tip: The HMC provides two integrated serial ports: the first native serial port is required for server connectivity, and the second native serial port is reserved for a modem if “Service Agent call home” is implemented. Therefore, if there is more than one managed partition and the call home function is implemented, you have to assign an 8-port (F/C 2943) or 128-port (F/C 2944) serial adapter.

Note that 128-Port asynchronous adapters are not supported on the HMC (F/C 7316).

Each system that supports a Hardware Management Console (HMC) has two serial port connections, so that you may optionally attach a second HMC to the same system. The benefits of using two HMCs are as follows:

- ▶ This ensures that access to the HMC management function capabilities are not interrupted.
- ▶ It ensures access if the network is down.

There is no physical console on the partition unless you assign it explicitly (by assigning the serial port in the media drawer). Two built-in native serial (RS-232) ports on the p690 can only be assigned to one partition (LPAR) at the same time.

If an explicit system console is needed on more than one partition at the time, your only option to facilitate this would be to implement a “graphical console” environment for the partitions. A graphical console is available on a partition by configuring a graphical adapter (F/C 2943) with a graphical display attached and a USB keyboard and mouse adapter per partition needing explicit system console. As with standalone servers, only one graphical console is configurable per each partition; the number of graphical console can equal the number of LPARs.

Important: For installation or service processor supported functions, you must use the Virtual Terminal Window (VTW) function of the HMC.

5.3.2 Serial port implication in an LPAR/SP environment

There is no serial connection to any LPAR or p690 server. Connectivity from the control workstation (CWS) to the p690, p670, p655, p650, and p630 (the LPAR enabled systems) is achieved through the Hardware Management Console (HMC) via a network connection to the SP LAN.

The LPAR-enabled systems are supported as SP-attached servers, or in a Cluster 1600 configuration with the SP Switch2, the SP Switch, or no switch.

The first step in installing the SP system is to prepare the CWS. This involves connecting your SP frames and attached servers to the serial ports of your CWS, configuring the Ethernet connections on your CWS, and verifying name resolution. You must then install the PSSP code on the CWS and apply the necessary PTFs.

The control workstation uses one serial port per SP frame for hardware monitoring and control.

For attached servers, the number of serial ports you must allocate depends on the servers:

- ▶ SAMI protocol servers use two tty ports for communicating with the CWS.
- ▶ CSP protocol servers use one tty port.
- ▶ For pSeries 690 servers, you do not need to reserve a tty port, but the Hardware Management Console (HMC) must be connected to the SP Ethernet administrative LAN.

5.3.3 Virtual terminal window

In an LPAR environment there are no physical consoles, unless you assign it explicitly (with the limitations previously mentioned). To facilitate console outputs and to provide the means for diagnostics, firmware implementations and virtual tty device, a Virtual Terminal Window (VTW) is offered. AIX sees this VTW as a standard tty device. The output of the VTW is streamed to the HMC, and acts as the console for the LPAR.

Note: VTW is designed for limited purposes. It is used during AIX installation and diagnostics services. VTW does *not* support:

- ▶ Printing to a virtual terminal
- ▶ Transparent print service
- ▶ Modem connections
- ▶ Real time application usage

VTW emulates a VT320. If a VTW is used in a full partition mode, the I/O of the serial-1 port is redirected to the VTW window, and communication to and from the serial-1 port is interrupted. Normal operation of the serial-1 port is resumed as soon as the VTW window is closed.

Tip: For AIX configurations and systems management, we recommend that you use the network or the serial ports (using the serial adapter assigned to the partition).

In a “No Power” state of the system, you can still access the service processor via the VTW.

If you are booting an AIX partition that did not have the native serial adapters assigned as one of its resources at the time of BOS installation, the device driver for the serial adapter is not available in that partition's BOS.

So in order to install device drives, you may have to dynamically move the media drawer to each of the partitions temporarily to define, configure, and load device drives, as explained in the following section.

Moving the media drawer between active LPARs

You can use the Hardware Management Console (HMC) to move the managed system's media drawer from one active logical partition to another active logical partition without rebooting the partition's operating system.

To use the HMC to reassign the managed system's media drawer between active partitions, follow these steps:

1. Log in to the HMC using either the System Administrator or Advanced Operator roles.
 - Click the console's icon to expand the tree.
 - Click the **Server and Partition** folder.
 - Click the **Server Management** icon.
 - Click the managed system's icon to expand the tree.
 - Select the partition from which you want to move the CD-ROM.
 - Select **Dynamic Logical Partitioning**.
 - Select **Adapters**.
2. The Dynamic Reconfiguration window opens. Click **Move resource to a partition**.
3. Select the CD-ROM. The CD-ROM is listed as group U1.18-P1-I10 in the HMC interface.
4. Select the name of the partition to which you want to move the CD-ROM.

5. In the Task time-out field, select the number of minutes you want the system to wait before it stops the task.
6. In the Details field, select the level of feedback you would like to see while the HMC performs the task. Details shown include the operating system's standard output and standard error information.

When you are finished selecting the correct information, click **OK**.

5.4 HMC considerations

When implementing a Cluster 1600 environment, one of the key components is the HMC. The HMC provides a standard user interface for configuring and managing partitions or SMP systems. HMC also facilitates features to monitor system and hardware problems.

5.4.1 Redundant HMC

PSSP now supports automatic failover of hardware control and monitoring (hardmon) communications to a redundant HMC. The pSeries Hardware Management Console allows you to configure two HMCs to control a single HMC-controlled server. The PSSP **spframe** command allows you to enter a list of IP addresses to be used for the hardware control point for the logical frame being defined for the HMC-controlled server.

PSSP initially tries to use the first IP address when attempting to establish a communication session with an HMC for the HMC-controlled server. If communication cannot be established to the first HMC in the list of IP addresses, the hardmon subsystem attempts to establish a communication session with the other IP address defined for that logical frame.

In redundant HMC configurations, both HMCs are fully active and accessible at all times, enabling you to perform management tasks from either HMC at any time. There is no primary or backup designation. To avoid conflicts, mechanisms in the communication interface between HMCs and the managed systems allow an HMC to temporarily take exclusive control of the interface, effectively locking out the other HMC. Usually this locking is done only for the duration of time it takes to complete an operation, after which the lock is released for further operations.

HMCs are also automatically refreshed of any changes that occur in the managed systems, so the results of commands issued by one HMC are visible to the other. For example, if you select to activate a partition from one HMC, you will observe the partition going to the Starting and Running states on both HMCs.

The locking between HMCs does not prevent users from running commands that might seem to be in conflict with each other. For example, if the user on one HMC selects to activate a partition, and a short time later, a user on the other HMC selects to power off the system, the system will power off. Effectively, any sequence of commands that you can do from a single HMC is also permitted when your environment contains redundant HMCs.

For this reason, it is important to carefully consider how you want to use this redundant capability to avoid such conflicts. You might choose to use them in a primary and backup role, even though the HMCs are not restricted in that way.

The interface locking between two HMCs is automatic and is usually of short duration. Most console operations wait for the lock to release without requiring user intervention. However, if one HMC experiences a problem while in the middle of an operation, it may be necessary to manually release the lock.

Because authorized users can be defined independently for each HMC, you should determine whether the users of one HMC should be authorized on the other. If so, the user authorization must be set up separately on each HMC.

Tip: Because both HMCs provide Service Focal Point and Service Agent functions, connect a modem and phone line to only one of the HMCs, and enable its Service Agent. To prevent redundant service calls, enable Service Agent on only one HMC.

Perform HMC software maintenance separately on each HMC, at separate times, so that there is no interruption in accessing HMC function. This situation allows one HMC to run at the new fix level, while the other HMC can continue to run at the previous fix level. Make sure that both HMCs are moved to the same fix level as soon as possible.

5.5 Web-based System Manager client solutions

Web-based System Manager is a tool that provides a complete set of Web-based system management interfaces for the entire pSeries domain. It enables administration of single/multiple AIX machines/partitions, AIX clusters, and SP nodes from any client platform, including the browser on the PC at the administrator's home.

HMC can be managed through Web-based System Manager interfaces. However, in an enterprise computing environment, you have to work through the firewall to enable the functionality of the Web-based System Manager and the

HMC. In this section, we present a solution to facilitate communication between the HMC and a Web-based System Manager client through a firewall.

5.5.1 Web-based System Manager functionality through the firewall

In a corporate environment, firewalls are common, if not mandatory. It is also common that systems administration tasks are often performed outside of the physical boundaries of a data center; often systems administrators are across town or across the country from the data center.

With the HMC, you are able to manage multiples of servers from various corporate locations, enabling the systems administrators to be freed from the distance limitations of RS-232 connections, and to still be able to manage hardware and configuration issues on their servers.

However, systems administrators are still limited by the security measures (firewall) put in place between the intranet and Internet. If you want to put the HMC's management capabilities outside the compounds of trusted networks, you are confronted with the limitation of not knowing how the HMC and the Web-based System Manager client communicate through the firewall. Capturing the port numbers utilized by the HMC and Web-based System Manager in a single session and opening those specific ports is rendered useless on subsequent communications between the HMC and the Web-based System Manager.

For that reason, we offer the following solution to address the firewall limitation imposed on HMC and Web-based System Manager communications.

Note: This solution is not documented and may not be supported by IBM at this time.

We verified this solution on the following:

- ▶ IBM p690, LPARed with AIX5.1 and AIX5.2
- ▶ Checkpoint Firewall 4.1 running on Sun Ultra60 - Solaris 7
- ▶ HMC 7315-C02, with software version 1.3
- ▶ Web-based System Manager HMC-Client on Windows® XP (SP1®) workstation
- ▶ DSL (static IP) Internet connectivity

Intranet HMC-to-Web-based System Manager communication

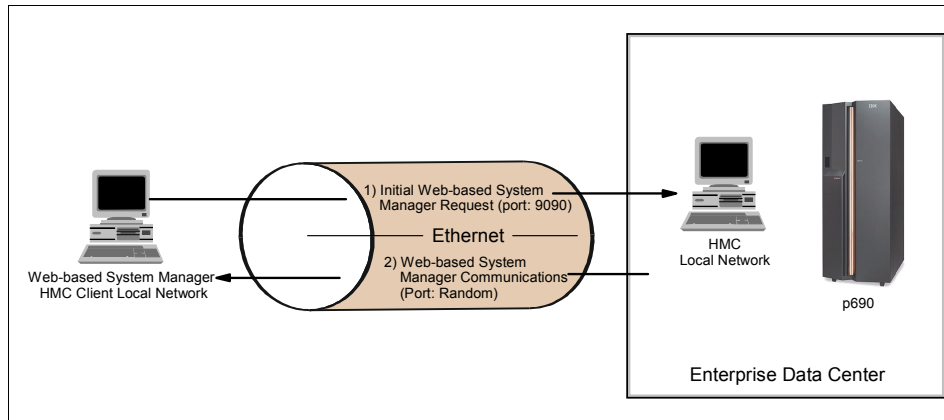


Figure 5-5 Intranet communication between HMC/Web-based System Manager

Figure 5-5 depicts a sequence of initial requests from the HMC and Web-based System Manager and their subsequent communication. The HMC and Web-based System Manager initialize communication on Port 9090 (this is the factory default, which may be changed as described in “Changing the default client communication port” on page 247).

Once the initial request is received by the HMC, it returns a handshake to Web-based System Manager, using any one of the random port numbers between 1024 to 65535. (It is more than likely that HMC will not be using a commonly used port number, for example, 8080.) At this point, communication is established between the HMC and the Web-based System Manager HMC client.

Changing the default client communication port

If your environment requires that you change the Web-based System Manager’s initialization port from 9090 to some other port, issue the following command:

```
/usr/websm/bin/wmsmserver -enable -listenport <desired port number>
```

You must enable or open these ports on your firewall to facilitate the communication.

Implications of a firewall

After we set up the LPAR-enabled pSeries, along with the HMC, within the enterprise data center, we tested all HMC functionality and verified that it worked within the intranet. However, soon as the Web-based System Manager client was removed from the bounds of the trusted network, as shown in Figure 5-6, communication between the HMC and Web-based System Manager stopped.

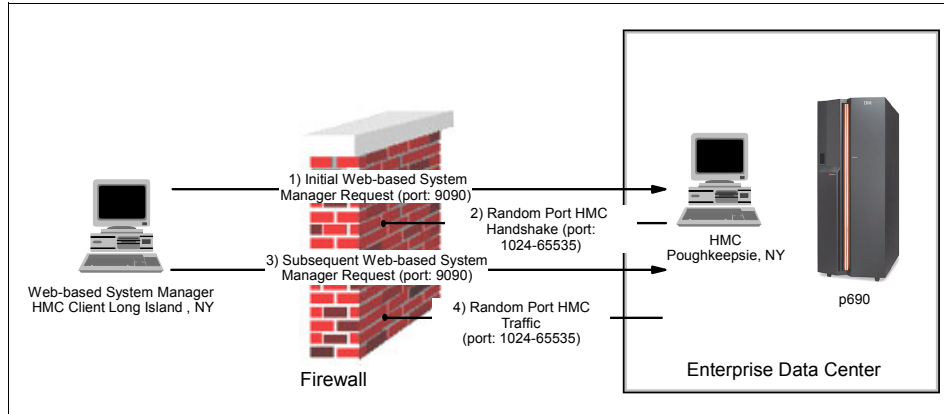


Figure 5-6 Implications of the firewall

The initial Web-based System Manager communication coming through port 9090 (assuming that this port is enabled in the firewall) reached the HMC; however, the handshake and all subsequent requests and communications were held at the firewall, because the HMC's follow-on communication is via a random port numbers.

An easy way to enable communications between the HMC and the Web-based System Manager client is to enable all ports on your firewall. However, enabling all ports between port number 1024 and 65535 defeats the purpose of having a firewall in the first place.

Important: We noticed that the handshake coming from the HMC to the Web-based System Manager server utilizes a different, random port number each session. Therefore, capturing the initial handshake port number and opening that specific port on your firewall will not resolve this issue.

Solution

Solving this problem is rather simple if you control the range of ports which the HMC uses to communicate with the Web-based System Manager client. Then the number of ports to be enabled on the firewall is manageable, as shown in "Changing the default client communication port" on page 247.

Enabling ranges of ports on your firewall to facilitate the Web-based System Manager communications to the HMC presents a level of security risk. However, you can minimize the risk by implementing security schemes (for example, a time-based connection).

Defining HMC-to-Web-based System Manager communication ports

To shrink the range of ports that the HMC uses to communicate through port numbers 10010 to 10020, issue the following command:

```
/usr/websm/bin/wsmserver -enable -portstart <start port>10010 -portend  
<end port>10020
```

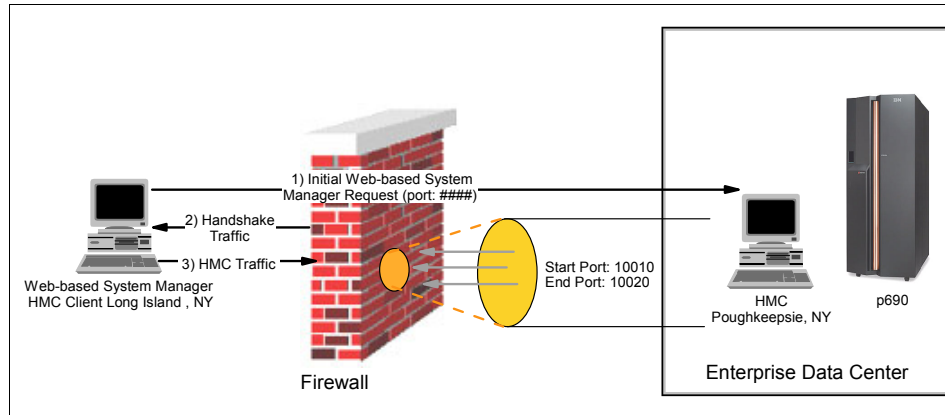


Figure 5-7 Limiting the HMC's communication ports

In summary, to ensure reliable communications, reduce the number of ports that the HMC utilizes for Web-based System Manager client communications from 64511 possibilities to about 5 to 10 possibilities (as shown in Figure 5-7), and open those ports on the firewall.



Performance

The communication performance seen by applications running in AIX® cluster nodes and communicating through the SP Switch or SP Switch2 is comprised of a number of elements. The SP Switch and SP Switch2 provide the base communications performance capability.

Important: All performance data is subject to change. For the latest performance information, refer to the following Web site:

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

The performance characteristics of the SP Switch and SP Switch2 adapters in the node, and the time the node takes to process the communication protocol stack, determine how much of the SP Switch and SP Switch2 performance capability can be sustained during application-to-application communication. The time for processing the communication protocol stack is, in turn, determined by the software path length and the performance characteristics of the processor.

In this appendix, we discuss the performance capabilities of both the SP Switch and SP Switch2 fabrics and their adapters. We also present actual application performance measurements. A discussion of the software path length in processing communication protocol stacks is beyond the scope of this redbook.

We use the term “switch” in this appendix to refer to any switch type, and we use the term “adapter” to refer to any adapter type, unless a specific switch or adapter is named.

A.1 Switch performance

The SP Switch2 was first introduced in 2000 for the interconnection of SP nodes via internal MX slots. As the next step in the evolution of the SP interconnection fabric, it offered significant improvements in bandwidth, latency, and Reliability, Availability, and Serviceability, or RAS, over the previous generation SP Switch. The bandwidth gains resulted from design improvements in the adapters and switch hardware.

In 2001, the SP Switch2 was enhanced to include the SP Switch2 PCI Attachment Adapter, allowing interconnection of any PCI-based server supported in an IBM Cluster 1600. It was also enhanced to include dual plane support, which is the ability to support up to two adapters per node (or logical partition), thereby increasing bandwidth and allowing for redundant adapters and higher degrees of RAS. These enhancements enabled switch communication performance to keep pace with the increased performance of the SP nodes and Cluster 1600 systems.

With the announcement of the pSeries™ 655 in December of 2002, a new SP Switch2 adapter, the SP Switch2 PCI-X adapter, was introduced. Initially, this adapter is intended for attaching pSeries 655 servers to the SP Switch2 by way of one of the available PCI-X slots.

Because the SP Switch2 design improvements are evolutionary steps based on the SP Switch and previous high performance switches, the SP Switch2 is designed to be fully compatible with applications written for the older SP switches. As with these prior switches, the SP Switch2 supports the industry-standard Message Passing Interface (MPI) and Internet Protocol (IP), as well as the IBM Low-level Application Programming Interface (LAPI).

The raw peak performance of the SP Switch and new SP Switch2 are listed in Table A-1 on page 253.

Table A-1 Switch peak performance

Number of cluster nodes	Switch type	Latency (µsec)	Bandwidth (MB/sec)	
			Uni-directional	Bi-directional
Up to 16	SP Switch2	1.0	500	1000
17 to 80		1.5		
81 to 512		2.5		
Up to 16	SP Switch	1.3	150	300
17 to 80		1.9		
81 to 512		3.3		

This peak (for example, not-to-exceed) performance cannot be achieved by an application.

A.2 Adapter performance

The raw peak performance of the SP Switch and SP Switch2 adapters are listed in Table A-2.

Table A-2 Switch adapter peak performance

Adapter	Bandwidth (MB/sec)	
	Uni-directional	Bi-directional
SP Switch2 adapter	500	1000
SP Switch2 PCI of PCI-X adapters	500	1000
SP Switch MX2 adapter	150	300
SP System Attachment adapter	150	150

Greater detail about these adapters is given in the following sections.

SP Switch2 adapters

- ▶ The SP Switch2 adapter (SP Switch2 adapter, SP F/C 4025) is available for the 222 MHz and 375 MHz POWER3™ high nodes.
- ▶ The SP Switch2 MX2 adapter (SP Switch2 MX2 adapter, SP FC 4026) is available for the 375 and 450 MHz POWER3 thin and wide nodes.
- ▶ The SP Switch2 PCI-X adapter (SP Switch2 PCI-X Attachment Adapter, F/C 8398) is available for pSeries 655 nodes.
- ▶ The SP Switch2 PCI adapter (SP Switch2 PCI Attachment Adapter, F/C 8397) adapter is available to attach selected RS/6000® and pSeries servers to the SP Switch2 to PCI slots in the server or I/O drawer (as in the case of the pSeries 670 or 690 servers).

SP Switch adapters

- ▶ The SP Switch MX2 adapter (SP Switch MX2 adapter, SP F/C 4023) is available for all MX slot-enabled SP nodes
- ▶ The SP System Attachment Adapter (SP System Attachment Adapter, server F/C 8396) is available to attach various servers to the SP Switch using a PCI slot in the node or server.

These peak performance numbers represent the maximum rate at which data can be given to or taken from the switch by the node and, effectively, become the base communications performance capability of the switch subsystems. Similar to the switch peak performance, the switch adapter peak (that is, not-to-exceed) performance cannot be achieved by an application.

When the communication occurs between nodes with different adapters supported on the same switch type, the effective peak performance of the switch subsystem is that of the slower adapter.

A.3 Node I/O slot performance

Since the performance of the SP Switch and SP Switch2 fabrics is faster than some I/O slots available in various pSeries servers, there are several configurations where the same adapter, in different nodes or servers, will perform at different rates. This is due to differences in the I/O subsystems of different RS/6000 or pSeries servers.

In the case of the SP Switch2, with a bandwidth of 500 MB/s full duplex, this is much faster than the individual PCI or PCI-X slots that it plugs into. Different I/O drawers or internal slots on various models of the pSeries servers have different maximum bandwidths per PCI slot. This shows up as slightly different performance when identical switch adapters are used in different server models.

Therefore, depending on the I/O subsystem used, the same SP Switch2 PCI or PCI-X adapters will perform differently in various server models.

When the SP Switch2 adapter for the 6xx direct memory slot is used in the SP POWER3 SMP High Nodes, the much faster memory slot allows full utilization of the bandwidth in the SP Switch2. Over time, as advances are made in the I/O slots used in pSeries servers, the performance of the I/O buses should catch up to the performance of the SP Switch adapters. This is why initially adapters for the SP Switch and SP Switch2 used memory bus attachment and later used standard I/O slots.

Both adapters are now available in both memory buses and PCI bus versions. Initially, the next generation fabric adapters will plug into the direct memory bus. After the standard buses catch up in performance, the adapters will plug into a standard bus.

A.4 Application performance

High performance internode communication is a key component of the overall performance of many user applications. The most basic measurements to characterize the performance of the communication subsystem are latency and bandwidth.

- ▶ *Latency* is the overhead associated with sending data between two processors, and is usually quantified in microseconds (μsec).
- ▶ *Bandwidth* is the rate at which data can be transmitted between two processors, and is typically measured in megabytes per second (MB/s). In this document, for bandwidth, when using MPI, a megabyte is defined as 10^6 bytes. When using IP, a megabyte is defined as 2^{20} bytes.

Historically, these have been the definitions used when measuring these two protocols.

We will characterize internode communication performance over the switch for two communication protocols: the so-called “user space” protocol (used to support the industry standard MPI), and the industry standard IP family of communication protocols (which include TCP/IP and UDP/IP). The user space protocol is sometimes referred to as a “lightweight” protocol because it requires fewer processor cycles to transmit a given amount of data compared to heavier protocols like TCP/IP.

User space is most commonly used for scientific and technical computing applications via a message-passing interface. MPI was used to measure user space performance. TCP/IP is utilized in socket interface communication for

many commercial applications, and is the basis for popular network protocols such as Network File System (NFS) and File Transfer Protocol (FTP). Performance measurements for these two protocols are presented in the next two sections.

Several factors contribute to the communication performance that is obtained by a user application. Internode communication performance depends on the processor, the memory subsystem, the switch adapter, and the switch fabric. Therefore, when considering communication performance measurements, it is extremely important to understand the exact configuration of the system to which the data applies. In addition to hardware considerations, the system software contributes to the overhead involved in sending data between processors.

Note the following points:

- ▶ The latency and bandwidth measurements represent performance as seen by an application.
- ▶ MPI measured bandwidth increases asymptotically as message size grows very large. Each bandwidth measurement presented in these tables represents the asymptotic values for very large messages.
- ▶ The measurements for a given node were made using the latest release of software generally available at the time of the announcement of that node.
- ▶ The latencies and bandwidths for older nodes are included in these tables for reference.
- ▶ Latencies and bandwidths are measured on two nodes connected to the same switch chip.
- ▶ On the 375 MHz POWER3 High Node, MPI tasks were bound to CPUs in order to reduce performance fluctuations

The performance data shown in the following tables was measured under ideal conditions. In all cases except the pSeries 680 server, we configured the maximum number of CPUs and memory per node or server. For the pSeries 680 measurements, we only configured 12 CPUs and 16 Gigabytes of memory. However, this less than full configuration did not impact the performance of the adapters.

For each measurement, we configured up to two adapters using customer configurable guidelines. In most cases, both adapters were configured in the same I/O drawer as would be expected in customer configurations.

The best possible tuning was done on the systems for these measurements, as well as the latest level of AIX and Parallel System Support Programs for AIX (PSSP) available at the time of measurement.

A.5 MPI/user space

On a distributed-memory system, parallel applications perform interprocessor communication via some form of message passing. IBM fully supports MPI as an industry standard. This standardized interface for message passing greatly improves the portability of parallel application codes among different parallel systems. We will discuss the performance of the SP Switch for parallel applications in terms of what can be measured using MPI.

Table A-3 through Table A-6 on page 259 show interprocessor communication performance measurements from a FORTRAN program with MPI calls, using the user space protocol. Latency is a measure of the time it takes to send a zero byte message between two processors using `mpi_send` and `mpi_recv` from the MPI library. It is calculated as half the time needed for a round trip between processors for that zero byte message. Latency represents the time used to set up a single message for transfer at the level of an application, and may be seen as the initialization overhead for transferring information between applications.

Table A-3 through Table A-6 also contain data for bandwidth measurements using MPI over the user space interface. The uni-directional bandwidth, sometimes called “point-to-point” bandwidth, was measured for messages of several megabytes in size. The bi-directional bandwidth, sometimes called “exchange” bandwidth, implies simultaneous sending and receiving of messages between processors, thereby achieving a slightly higher data rate. The bi-directional data rate is the sum of the simultaneous data rates in both directions.

In Table A-3 through Table A-6, the SP Switch adapter has the lowest latency of all adapters. Table A-3 shows SP Switch2 MPI user space performance with a single MPI per node connected to the same switch.

Table A-3 SP Switch2 MPI user space performance

Node type	Switch adapter	SP Switch2 latency (µsec)	SP Switch2 bandwidth (MB/sec)	
			Uni-directional	Bi-directional
375 MHz POWER3 high node	SP Switch2	17.5	350	350
450 MHz POWER3 SMP node	SP Switch2 MX2	19.4	168	197
pSeries 680	SP Switch2	25	111	122

pSeries 660-6h1	SP Switch2 PCI	21.1	161	167
pSeries 660-6M1	SP Switch2 PCI	20.6	162	162
pSeries 690	SP Switch2 PCI	18.5	179	225
pSeries 655	SP Switch2 PCI-X	18	239	245

All but one of the results in Table A-4 were obtained using the threaded MPI library with MP_SINGLE_THREAD=yes. (The pSeries 680/SP Switch Attachment adapter results were obtained using the non-threaded MPI library.) Table A-4 shows the SP Switch MPI user performance with a single MPI task per node connected to the same switch chip.

Table A-4 SP Switch MPI user space performance

Node type	Switch adapter	Latency (μsec)	Bandwidth (MB/sec)	
			Uni-directional	Bi-directional
375 MHz POWER3 thin/wide node	SPSMX2	19.7	140	192
pSeries 680	SPSAA	36.6	71	87

Measurements for the SP Switch2 Attachment adapter for the pSeries 660 (Models 6H0, 6H1 and 6M1) and the pSeries 680 are included for completeness, even though applications deployed on attached servers will generally use IP rather than MPI communication. These servers will generally be used for commercial computing applications, while MPI is generally used by scientific and technical computing applications. Table A-5 shows the MPI user space performance on nodes with multiple MPI tasks per node for the SP Switch2.

Table A-5 MPI user space performance with multiple MPI tasks

Node type	Switch adapter	Number of MPI tasks per node	Bandwidth (MB/sec)	
			Uni-directional	Bi-directional
pSeries 655	SP Switch2 PCI-X	1 or more	235	254

pSeries690	SP Switch2 PCI	1 or more	179	225
pSeries 6M1	SP Switch2 PCI	1 2 or more	162 163	162 213
pSeries 660-6h1	SP Switch2 PCI	1 2 or more	161 161	167 212
450 MHz POWER3 SMP node	SP Switch2 MX2	1 or more	168	197
pSeries 680	SP Switch2	1 or more	111	122
375 MHz POWER3 high node	SP Switch2 PCI-X	1 2 4 or more	350 445 445	350 680 720

Table A-6 shows MPI user space performance on nodes with multiple MPI tasks per node for the Dual fabric SP Switch2.

Table A-6 MPI user space performance

Node type	Switch adapter	Number of MPI tasks per node	Bandwidth (MB/sec)	
			Uni-directional	Bi-directional
pSeries 655	SP Switch2 PCI-X	1 2 or more	412 460	440 460
pSeries690	SP Switch2 PCI	1 2 or more	337 337	375 385
pSeries 6M1	SP Switch2 PCI	1 2 4 or more	188 309 310	188 347 366
pSeries 660-6h1	SP Switch2 PCI	1 2 4 or more	168 245 295	168 261 324
pSeries 680	SP Switch2	1 2 or more	149 164	151 193
375 MHz POWER3 high node	SP Switch2	1 2 4 8 or more	350 664 880 880	350 675 1150 1270

A.6 TCP/IP

TCP/IP is a more common industry standard communication protocol used to transfer information between two systems running the IP family of protocols. It is a robust protocol that supports multiple users and reliable transport of data.

However, since it supports networking function not currently used by MPI (such as multiplexing and guaranteed delivery), it requires higher processor overhead compared to the user space protocol using MPI. This increases the CPU overhead for the same bandwidth when compared to MPI.

The performance of the TCP/IP socket protocol on various nodes was measured using a modified version of the Netperf public-domain benchmark, and the results are listed in Table A-7 through Table A-12 on page 263.

All Netperf measurements were memory-to-memory to eliminate slower devices (such as disks) from impacting the performance. As with the user space measurements, the same hardware configuration was used for the TCP/IP measurements as for the MPI measurement listed in A.5, "MPI/user space" on page 257.

Table A-7 TCP/IP SP Switch performance

Number of TCP/IP sessions	Bandwidth (MB/sec), Uni-/bi-directional			
	375 MHz POWER3 thin/wide node		pSeries 680 node	
	Uni-	Bi-	Uni-	Bi-
1 or more	135.4	174.6	74.2	90.7
Switch adapter	SPSMX2		SPSAA	

On the SP Switch, the 375 MHz POWER3 SMP Node with the SPSMX2 adapter was the fastest performing adapter and node combination; see Table A-7.

In both the p680 and 375 MHz SMP Node, the adapter could be saturated by a single TCP/IP session.

Table A-8 SP Switch2 TCP/IP SP Switch performance

Number of TCP/IP sessions	Bandwidth (MB/sec), Uni-/Bi-directional single fabric					
	375 MHz POWER3 high node		pSeries 690		450 MHz POWER3 SMP node	
	Uni-	Bi-	Uni-	Bi-	Uni-	Bi-
1	132	253	168	230	136	210
2	252	472	167	230	156	210
4 or more	440	650	167	230	157	210
Switch adapter	SP Switch 2		SP Switch2 PCI		SP Switch2 MX2	

As shown in Table A-8, the 375 MHz POWER3 SMP high node delivers the highest bandwidth of all node and adapter combinations for the SP Switch or SP Switch2. This performance can primarily be attributed to the use of the memory bus to attach the adapters. All other nodes or servers are limited by the throughput of the bus into which the adapter is plugged.

When using the SP Switch2 adapter and switch, you get higher bandwidth when using more than one CPU on the 375 MHz POWER3 high node; see Table A-9.

Table A-9 SP Switch2 TCP/IP SP Switch performance

Number of TCP/IP sessions	Bandwidth (MB/sec), Uni-/Bi-directional single fabric					
	pSeries 660-6M1		pSeries 660-6H1		pSeries 680	
	Uni-	Bi-	Uni-	Bi-	Uni-	Bi-
1 or more	176	223	178	132	124	130
Switch adapter	SP Switch 2 PCI		SP Switch2 PCI		SP Switch2 PCI	

When using dual switch fabrics and using the global interface (ml0), the performance of TCP/IP increases significantly; see Table A-10 on page 262.

We do not get perfect scaling to two adapters because of the limitations on aggregate bandwidth to a single I/O drawer. In addition, the global interface introduces some overhead managing both interfaces.

Table A-10 SP Switch2 TCP/IP SP Switch performance

Number of TCP/IP sessions	Bandwidth (MB/sec), Uni-/bi-directional			
	pSeries 630		pSeries 655 node	
	Uni-	Bi-	Uni-	Bi-
1 or more	169	231	238	259
Switch adapter	SP Switch 2 PCI		SP Switch 2 PCI-X	

The bandwidth measured is the maximum obtainable both uni-directional and bi-directional over TCP between two identical applications running on two identical nodes; see Table A-11.

All Netperf measurements were memory-to-memory to eliminate slower devices such as disks from impacting the performance.

Table A-11 SP Switch2 dual fabric TCP/IP SP Switch performance

Number of TCP/IP sessions	Bandwidth (MB/sec), Uni-/bi-directional dual fabric (mI0 interface)			
	pSeries 690		pSeries 655	
	Uni-	Bi-	Uni-	Bi-
1	317	385	449	495
2 or more	308	397	435	491
Switch adapter	SP Switch 2 PCI		SP Switch 2 PCI-X	

The nodes can take advantage of multiple processors if there are multiple IP connections running at the same time. If only one TCP/IP socket is used, the maximum throughput will be similar to the single-processor throughput no matter how many processors are configured in the node.

A single TCP/IP socket currently cannot take advantage of multiple processors, due to the single-threaded nature of memory-to-memory copies and the TCP/IP stack; see Table A-12 on page 263.

Table A-12 SP Switch2 dual fabric TCP/IP SP Switch performance

Number of TCP/IP sessions	Bandwidth (MB/sec), Uni-/Bi-directional single fabric					
	pSeries 660-6M1		pSeries 660-6H1		pSeries 680	
	Uni-	Bi-	Uni-	Bi-	Uni-	Bi-
1	197	353	199	323	175	200
2 or more	287	366	288	324	174	199
Switch adapter	SP Switch 2 PCI		SP Switch2 PCI		SP Switch2 PCI	

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 267.

Note: Note that some of the documents referenced here may be available in softcopy only.

- ▶ *An Introduction to CSM 1.3 for AIX 5L*, SG24-6859
- ▶ *Linux Clustering with CSM and GPFS*, SG24-6601
- ▶ *RS/6000 SP and Clustered IBM @server pSeries System Handbook*, SG24-5596
- ▶ *IBM @server pSeries 690 and pSeries 670 System Handbook*, SG24-7040
- ▶ *Exploiting RS/6000 SP Security: Keeping It Safe*, SG24-5521
- ▶ *pSeries Systems Handbook 2003 Edition*, SG24-5120
- ▶ *IBM @server Certification Study Guide: Cluster 1600 Managed by PSSP*, SG24-7013
- ▶ *Effective System Management Using the IBM Hardware Management Console for pSeries*, SG24-7038
- ▶ *The Complete Partitioning Guide for IBM pSeries Servers*, SG24-7039
- ▶ *pSeries 650 Model 6M2 Technical Overview and Introduction*, REDP0194
- ▶ *Configuring a p690 in an IBM Cluster 1600*, REDP0187

Other publications

These publications are also relevant as further information sources:

- ▶ *IBM @server Cluster 1600 Planning, Installation and Service Guide*, GA22-7863

- ▶ *IBM @server Cluster 1600 Planning Volume 2, Control Workstation and Software Environment*, GA22-7281
- ▶ *PSSP 3.5 Administration Guide*, SA22-7348
- ▶ *IBM @server Cluster 1600 Planning Volume 1, Hardware and Physical Environment*, GA22-7280
- ▶ *RS/6000 and eServer pSeries: PCI Adapter Placement Reference*, SA38-0538
- ▶ *Implementing a Firewalled RS/6000 SP System*, GA22-7874
- ▶ *HMC Operations Guide*, SA38-0590
- ▶ *PSSP Administration Guide*, SA22-7348
- ▶ *PSSP Installation and Migration Guide*, SA22-7347
- ▶ *IBM CSM for AIX 5L: Administration Guide*, SA22-7918
- ▶ *SP Switch Router Adapter Guide*, GA22-7310
- ▶ *pSeries High Performance Switch Planning, Installation, and Service Manual*, GA22-7951

Online resources

These Web sites and URLs are also relevant as further information sources:

- ▶ White paper on POWER4 System Micro architecture
<http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/power4.html>
- ▶ Microcode Discovery Service Web site
<http://techsupport.services.ibm.com/server/aix.invscountMDS>
- ▶ pSeries and RS/6000 Microcode Updates Web site
<http://techsupport.services.ibm.com/server/mdownload>
- ▶ PSSP V3.5 documentation at the Web site
http://www.rs6000.ibm.com/resource/aix_resource/sp_books/
- ▶ More information on IBM eServer xSeries servers
<http://www.ibm.com/servers/eserver/education/xseries/xref.html>

- ▶ Detailed guide on alt_diskinstall
http://publib16.boulder.ibm.com/pseries/en_US/infocenter/howto/HT_insgdrf_alt_diskinstall_clone.htm#insgdrf_alt_diskinstall_clone
- ▶ Detailed Base Operating System Migration Guide
http://publib16.boulder.ibm.com/pseries/en_US/infocenter/howto/HT_insgdrf_migrationinstall.htm#insgdrf_migrationinstall

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

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Abbreviations and acronyms

AC	Alternating Current	CWS	Control Work Station
AIX	Advanced Interactive Executive	C-SPOC	Cluster Single Point of Control
AFS	Andrew File System	DASD	Direct Access Storage Device
APAR	Authorized Program Analysis Report	DC	Direct Current
API	Application Programming Interface	DCE	Distributed Computing Environment
ARP	Address Resolution Protocol	DDR	Double Data Rate
ATM	Asynchronous Transfer Mode	DFS	Distributed File System
ATM	Asynchronous Transfer Mode	DHCP	Dynamic Host Control Protocol
BIST	Built In Self Tests	DIMM	Dual Inline Memory Module
BLAS	Basic Linear Algebra Subprograms	DVD-RAM	Digital Versatile Disk - Random Access Memory
BNC	Barrel Nut Connector	DVD-ROM	Digital Versatile Disk - Read Only Memory
BPC	Bulk Power Controller	EIA	Electronic Industries Alliance
BTU	British Thermal Unit	ESSL	Engineering and Scientific Subroutine Library
CCGA	Ceramic Column Grid Array	F/C	Feature Code
CD	Compact Disk	FCS	Fibre Channel Standard
CD-ROM	Compact Disk Read Only Memory	FDDI	Fiber Distributed Data Interface
CE	Customer Engineer	FFDC	First Failure Data Capture
CEC	Central Electronics Complex	FRU	Field Replaceable Unit
CES	Clustered Enterprise Server	FTP	File Transfer Protocol
CFM	Configuration File Management	Gb	Giga bits
CIU	Core Interface Unit	GB	Gigabyte (10 ⁹ byte)
CPU	Central Processing Unit	Gb/s	Giga bits per second
CSM	Cluster Systems Management	GB/s	Giga Bytes per second
CSS	Communication Subsystem	GHz	Gigahertz (10 ⁹ hertz)
CtSec	Cluster Security Services	GPFS	General Parallel File System
CUoD	Capacity Upgrade on Demand	GUI	Graphical User Interface
		HACMP	IBM High Availability Cluster Multi-Processing for AIX

HACMP/XD	HACMP Extended Distance	MASS	Mathematical Acceleration Subsystem
HACWS	High Availability Control Work Station	Mb	Mega bits
HIPPI	High Performance Parallel Interface	MB	Mega Bytes
HMC	Hardware Management Console	Mb/s	Mega bits per second
HPC	High Performance Computing	MCA	Micro Channel Architecture
HPS	High Performance Switch	MCM	Multi Chip Module
Hz	Hertz (Cycles per Second)	MES	Miscellaneous Equipment Specification
I/O	Input Output	MHz	megahertz (10 ⁶ hertz)
IBM	International Business Machines Corporation	ML	Maintenance Level
IP	Internet Protocol	mm	Millimeter
IPv4	Internet Protocol Version 4	MPI	Message Passing Interface
IPv6	Internet Protocol Version 6	MS	Management Server
ISB	Intermediate Switch Board	NC	Non-Cachable
ITSO	International Technical Support Organization	NFS	Network File System
JFS	Journaled File System	NIM	Network Installation and Maintenance
KB	Kilo Bytes	NQS	Network Queuing System
kg	Kilo-Gram	NTP	Network Time Protocol
KLAPI	Kernel Low level Application Programming Interface	OS	Operating System
kVA	KiloVolt-Ampere	PAIDE	Performance AIDE
L1	Level 1	PC	Personal Computer
L2	Level 2	PCI	Peripheral Component Interconnect
L3	Level 3	PCI-X	Peripheral Component Interconnect Extended
LAN	Local Area Network	PE	Parallel Environment
LAPI	Low Level Application Programming Interface	PDU	Power Distribution Unit
lb	pounds (weight)	POE	Parallel Operating Environment
LL	LoadLeveler	POR	Power-on Reset
LPAR	Logical Partition	PSSP	Parallel System Support Programs
LPP	Licensed Programme Product	PTX	Performance Toolbox
LVM	Logical Volume Manager	RAN	Remote Access Node
M/T	Machine Type	RAS	Reliability, Availability, Scalability

RIO	Remote Input/Output	SVC	Switch Virtual Circuit
RM	Resource Managers	TCP	Transmission Control Protocol
RMC	Resource Monitoring and Control Subsystem	TP	Twisted Pair
RPQ	Request for Price Quotation	TTY	Terminal Type (UNIX terminal interface)
RSA	Remote Server Access	UDP	User Datagram Protocol
RSCT	Reliable Scalable Cluster Technology	US	User Space
SAE	Service Action Event	USB	Universal Serial Bus
SAS	Service Agent gateway Server	UTP	Unshielded Twisted Pair
ScaLAPACK	Scalable Linear Algebra Package	V	Volts
SCM	Single Chip Module	VLAN	Virtual Area Network
SCSI	Small Computer System Interface	VSD	Virtual Shared Disks
SDR	System Data Repository	WAN	Wide Area Network
SDRAM	Synchronous Dynamic Random Access Memory	WLM	Workload Manager
SES	SCSI Enclosure Services		
SFP	Service Focal Point		
SIS	System Installation Suite		
SLES	SuSE Linux Enterprise Server		
SMIT	System Management Interface Tool		
SMP	Symmetric Multiprocessors		
SNI	System Network Interface		
SNM	Switch Network Manager		
SNMP	System Network Management Protocol		
SOI	Silicon On Insulator		
SP	Scalable Parallel		
SP	Service Processor		
SSA	Serial Storage Architecture		
SSB	Server Switch Board		
SSH	Secure SHell		
SSL	Secure Sockets Layer		
std	standard		

Index

A

- accounting 173
- AIX Performance Toolbox (PTX) 225
- Alternate disk migration 238
- APAR
 - IY21957 96
 - IY22854 239
 - IY28102 87
 - IY32749 239
 - IY36001 81
 - IY36002 81
 - IY37884 80–81
 - IY37885 80–81
 - IY39794 57, 63, 73, 82, 91, 99, 119
 - IY39795 57, 63, 73, 82, 91, 99, 119
 - IY41696 81
 - IY42352 72, 81
 - IY42353 57, 63, 73, 82, 91, 99, 119
 - IY42356 57, 63, 73, 82, 91, 99, 119
 - IY42359 72
 - IY42369 57, 63, 72, 82, 90, 99, 119
 - IY42377 57, 63, 73, 82, 91, 99, 119
 - IY42379 57, 63, 73, 82, 91, 99, 119
 - IY42782 57, 63, 73, 82, 91, 99, 119
 - IY42783 57, 63, 73, 82, 91, 99, 108, 119
 - IY42847 57, 63, 73, 82, 91, 99, 119
 - Y42358 72
- asynchronous card 27
- Audit Log resource manager (AuditLogRM) 188
- automounter 5

B

- Built-in Self Test (BIST) 34
- bulk power controller (BPC) 106

C

- Capacity on Demand (CuOD) 69
- Capacity Upgrade on Demand (CUoD) 17
- Central Electronics Complex (CEC) 114
- central manager 234
- Ceramic Column Grid Array (CCGA) 37
- Cluster 1600 2, 13, 230, 237

- components 13
 - hardware components 52
- Cluster Enterprise Servers (CES) 12
- cluster management 237
- Cluster Systems Management (CSM) xi, 11, 151, 170, 282
 - advantages 180
 - modular architecture 5
 - nodes concept 2
- cluster VLAN 154
- command
 - cmstat 189
 - dsh 4–5, 7, 154, 183
 - dshbak 8, 183
 - llextrPD 205
 - llmodify 206
 - lsmcode 47
 - lsmcode -A 47
 - lspv 239
 - rconsole 153
 - rpower 153
 - rsh 8, 183
 - spframe 148, 244
 - spmon -d 189
 - ssh 154, 183
- command line 4
- Communication Subsystems Support 176
- Configuration File Manager (CFM) 8
- control workstation (CWS) 16, 242
- Core Interface Unit (CIU) 33
- CSP 242

D

- daemon
 - hagsd 189
 - hatsd 189
- directory
 - /cfmroot 183
 - /spdata/sys1/install/pssplpp/PSSP-3.4 81
 - /spdata/sys1/install/pssplpp/PSSP-3.5 81
 - /tmp 4
 - /var 4
- Distributed Command Execution Manager (DCEM)

6
 Distributed Computing Environment (DCE) 175
 dynamic logical partitions (DLPAR) 233
 Dynamic Probe Class Library (DPCL) 218

E

electronic service agent 48
 Engineering and Scientific Subroutine Libraries (ES-
 SL) 170
 Enterprise Storage Server (ESS) 232
 ESSL (Scientific Subroutine Library) 169
 Event Response resource manager (ERRM) 188

F

F 70, 107
 FASTT Storage Server (FASTT) 232
 feature code
 F/C 0008 90, 99
 F/C 0009 90, 99
 F/C 0010 82
 F/C 0011 82
 F/C 0012 56, 62
 F/C 0013 62
 F/C 0014 72
 F/C 0015 72
 F/C 1500 15
 F/C 2031 15
 F/C 2032 15
 F/C 2050 17
 F/C 2051 17
 F/C 2052 17
 F/C 2053 17
 F/C 2054 18
 F/C 2056 18
 F/C 2057 18
 F/C 2058 18
 F/C 2934 43
 F/C 2943 41, 43, 241
 F/C 2944 41, 43, 79, 107, 241
 F/C 2968 118, 123
 F/C 2969 123
 F/C 2975 123
 F/C 2985 123
 F/C 2987 122–123
 F/C 3124 43
 F/C 3125 43
 F/C 3151 118
 F/C 3154 118

F/C 3166 107
 F/C 3167 107
 F/C 3256 107
 F/C 3257 107
 F/C 3628 44
 F/C 3636 44
 F/C 3756 106
 F/C 4011 16, 101
 F/C 4012 16
 F/C 4020 101
 F/C 4022 101
 F/C 4023 101
 F/C 4598 90, 99
 F/C 4962 43, 89, 98, 118, 122–123
 F/C 5122 67
 F/C 5126 61
 F/C 5131 61
 F/C 5133 54, 61
 F/C 5208 67
 F/C 550 15
 F/C 5511 77
 F/C 5513 77
 F/C 5515 77
 F/C 5518 77
 F/C 6234 106
 F/C 6266 55
 F/C 6273 61
 F/C 6404 97
 F/C 6418 97
 F/C 6420 81, 111
 F/C 6432 99, 110–111
 F/C 6433 106
 F/C 6434 99, 108, 110–111
 F/C 6435 106
 F/C 6436 106
 F/C 6437 106
 F/C 6563 77, 86, 89–90, 95, 98
 F/C 6571 77, 86, 89–90, 95, 98
 F/C 6578 70
 F/C 6579 70
 F/C 7039-6420 106
 F/C 7040-6432 106
 F/C 7040-6434 106
 F/C 7176 69
 F/C 7177 69
 F/C 7178 69
 F/C 7315 38
 F/C 7316 38, 241
 F/C 8120 43, 56, 62, 72, 81, 90, 98

F/C 8121 43, 56, 62, 72, 81, 90, 98
 F/C 8122 43, 81, 106, 161
 F/C 8123 43, 81, 106, 161
 F/C 8131 43
 F/C 8132 43
 F/C 8133 43
 F/C 8136 43
 F/C 8137 43
 F/C 8396 89–90, 98–99, 108–110, 117–118
 F/C 8397 62, 89–90, 98–99, 108–110, 117–118
 F/C 8398 62, 71–72, 81, 89–90, 98–99,
 108–110
 F/C 8691 107
 F/C 9047 106
 F/C 9049 106
 F/C 9123 118
 F/C 9125 118
 F/C 9172 68
 F/C 9176 69
 F/C 9177 69
 F/C 9178 69
 F/C 9222 123
 F/C 9223 123
 F/C 9302 108
 F/C 9305 108
 F/C 9310 108
 F/C 9315 108
 F/C 9320 108
 F/C 9581 61
 Fiber Distributed Data Interface (FDDI) 137
 Field Replaceable Unit (FRU) 160
 Field Replacement Unit (FRU) 50
 file
 /etc/switch.info 149
 File System resource manager (FileSystemRM)
 188
 File Transfer Protocol (FTP) 256
 First Failure Data Capture (FFDC) 34
 frames 15

G

General Parallel File System (GPFS) 170, 192, 234
 Gigabit Ethernet 231
 GPFS (General Parallel File System) 169
 Group Services 189

H

HACMP (High Availability Cluster Multi-Processing)

169
 HACMP/XD (Extended Distance) 220
 hags 189
 Hardware Management Console (HMC) 17, 38
 redundant 40
 hats 189
 heartbeat 234
 High Availability Cluster Multiprocessing (HACMP)
 170
 High Availability Control Workstation (HACWS) 20,
 22
 High Availability Geographic Cluster (HAGEO) 220
 Host resource manager (HostRM) 188

I

IBM Recoverable Virtual Shared Disk 175
 IBM Virtual Shared Disk 175
 inter switch boards (ISB) 105
 intermediate switch board (ISB) 157
 Internal Ethernet 122
 Inventory Scout 46

K

Kerberos 51

L

LAPI 177
 legacy hardware 13
 LoadLeveler 234
 LoadLeveler (LL) 170, 200
 logical partition (LPAR) 27, 233
 Logical Volume Manager (LVM) 197
 Low-level Application Programming Interface 177
 LPAR concept diagram 29

M

machine type
 7039 41
 M/T 14
 M/T 7028 41, 163
 M/T 7039 41
 M/T 7040 41, 161
 M/T 7040-61R 161–162
 M/T 7040-W42 162
 M/T 7045-SW4 104
 managed nodes 169
 managed servers 169

management server 16, 26, 166
management VLAN 153
Message Passing Interface 177
microcode discovery service 46
microcode level 47
microcode update application 45
microcode update files and discovery tool 47
MPI 177
Myrinet 231

N

Network File System (NFS) 256
Network Shared Disk (NSD) 231
Network Time Protocol (NTP) 5
node 169, 237

P

p630 32
P630 Internal Structure 54, 59, 65
p650 32
p670 32
p690 32
Parallel Environment (PE) 170, 234
Parallel ESSL 170, 234
Parallel Operating Environment (POE) 234
Parallel System Support Program (PSSP) xi
 advantages 172
 frame concept 2
Parallel System Support Programs (PSSP) 11, 25,
170, 282
partitions 237
PE (Parallel Environment) 169
Performance AIDE (PAIDE) 169, 225
Performance Toolbox (PTX) 169
perl 177
Placement 124
Power Distribution Unit (PDU) 68
POWER3 237
POWER4 237
pSeries and RS/6000 microcode updates 46
pSeries High Performance Switch (HPS) 151
pSeries servers 32
 p655 32
PSSP (Parallel System Support Programs) 169
public VLAN 154

R

Recoverable Virtual Shared Disk (RVSD) 197
redbook xi
Redbooks Web site 267
 Contact us xiv
redundant HMC 244
Reliability, Availability, Serviceability (RAS) 13
Reliable Scalable Cluster Technology (RSCT) 4,
180
Remote I/O (RIO) 114
request for price quotation (RPQ) 30
Resource Monitoring and Control (RMC) 139
restricted root access 51
roll-back 239
RS/6000 SP 2

S

SAMI 242
scheduler node 234
Scientific Subroutines Libraries (SSL) 170
SCSI Enclosure Services (SES) 76
SDR 173
Sensor resource manager (SensorRM) 188
server switch board (SSB) 105, 157
Service Action Event (SAE) 49
Service Agent 245
service director 48
Service Focal Point (SFP) 48, 245
Service Processor (SP) 34
single administrative domain 2
Single Chip Modules (SCM) 58
Software Update Protocol (SUP) 4
SP frame 237
SP Switch 16, 100, 125
SP Switch2 16, 100, 125–126
SP switches 16
Storage Area Network (SAN) 231
subnets 138
switch network interface (SNI) 105
Switch Network Manager (SNM) 105, 158
symmetric multiprocessor (SMP) 53
System Data Repository (SDR) 25, 173
System Management Interface Tool (SMIT) 151

T

Tcl 177
time synchronization 176
Tool command language 177

Topology Services 189
trusted Ethernet connection 166
trusted network 139

U

Unified Trace Environment (UTE) 218

V

Virtual Shared Disk (VSD) 194, 232
Virtual Terminal Window (VTW) 242
VLAN 26

W

Workload Manager (WLM) 201



IBM *@server* pSeries Cluster Systems Handbook

(0.5" spine)
0.475" <-> 0.875"
250 <-> 459 pages



IBM *@*server pSeries Cluster Systems Handbook



Overview of Cluster 1600 software components

Cluster 1600 machine types, models, and feature codes

Solution hints and tips

The IBM *@*server Cluster 1600 server, which was introduced to meet the rigorous demands of mission-critical enterprise applications, continues to offer outstanding performance, scalability, reliability, availability, serviceability, and management capabilities. In this IBM Redbook, we highlight the benefits of using a Cluster 1600, and describe which hardware components can be managed by either Parallel System Support Programs (PSSP) Version 3, Release 5, or Cluster Systems Management (CSM) Version 1, Release 3, Modification 2.

This publication contains the following information on the Cluster 1600:

- Cluster 1600 hardware components
- Networking components and considerations
- Cluster 1600 software components
- Scalability of the Cluster 1600
- Solutions and offerings scenarios

The Cluster 1600 helps to reduce the complexities and costs of system management, thus lowering the total cost of ownership and allowing simplification of application service level management. It also provides the infrastructure that supports availability, data sharing, and response time. This redbook will be useful for IT professionals seeking to implement Cluster 1600 mission-critical solutions to address business intelligence applications, server consolidation, and collaborative computing.

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