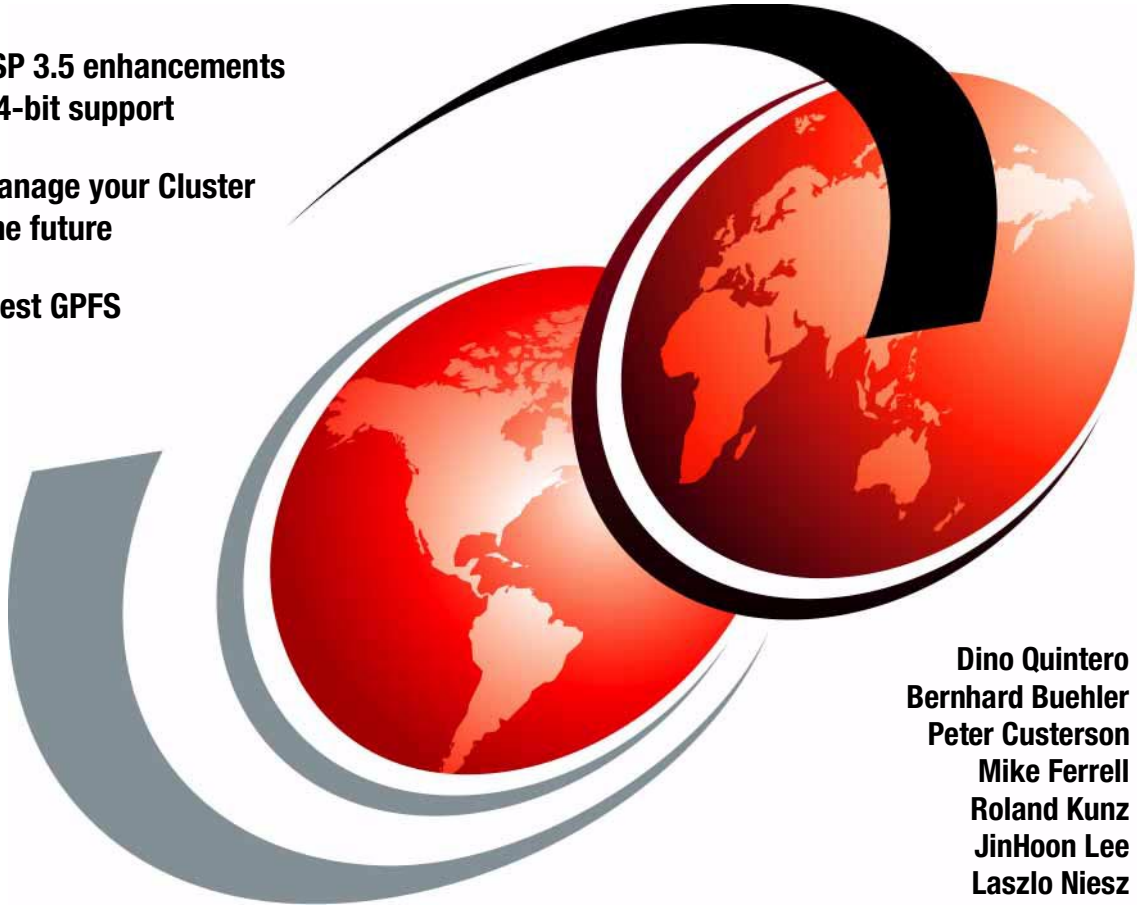
IBM @server Cluster 1600 Managed by PSSP 3.5: What's New

- Explore PSSP 3.5 enhancements including 64-bit support
- Plan and manage your Cluster 1600 into the future
- Tour the latest GPFS features



Dino Quintero
Bernhard Buehler
Peter Custerson
Mike Ferrell
Roland Kunz
JinHoon Lee
Laszlo Niesz

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM @server Cluster 1600 Managed by PSSP 3.5: What's New

December 2002

SG24-6617-00

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

First Edition (December 2002)

This edition applies to Version 3, Release 5, of Parallel System Support Program for use with the AIX operating system, Version 5, Release 1.

Contents

Figures

Notices

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

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

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

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

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

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

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

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

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

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

Trademarks

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

AIX®	IBM @server™	S/370™
AIX 5L™	iSeries™	SAA®
Balance®	LoadLeveler®	Sequent®
DataJoiner®	Micro Channel®	SP™
DB2®	MORE™	Tivoli®
e(logo)™ @	NetView®	TURBOWAYS®
Enterprise Storage Server™	Perform™	Wave®
ESCON®	PowerPC®	xSeries™
GXT1000™	pSeries™	z/VM™
GXT150L™	Redbooks (logo)™ 	zSeries™
GXT150M™	Redbooks™	
IBM®	RS/6000®	

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

Lotus®	Notes®	Word Pro®

The following terms are trademarks of other companies:

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

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

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

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

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

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

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

Preface

This IBM Redbook explores the evolution of the IBM RS/6000 SP system into the IBM @server Cluster 1600 and the impact of pSeries POWER4 LPAR technology in the pSeries clusters. This publication also highlights the new pSeries servers, which can be incorporated into Cluster 1600. This book provides pSeries cluster configuration information, including hardware and software hints and tips, as well as changes in the packaging of the cluster management components: AIX 5L and Parallel System Support Program (PSSP).

An overview of Reliable Scalable Cluster Technology (RSCT) is included to introduce the reader to the latest developments of the RSCT clustering software. The latest enhancements in PSSP 3.5 are included, highlighting in particular the changes made to the switch software and Virtual Shared Disks (VSD). Configuration architectures and examples are included for customers planning to deploy a Cluster 1600 in their computing environment. PSSP 3.5 and General Parallel File System (GPFS) enhancements are explored, including the latest 64-bit support and the latest supported levels of AIX 5L.

This redbook also includes helpful information about software coexistence, migration, and integration in Cluster 1600. Migration scenarios, hints, and tips are provided for customers planning to migrate to the latest software levels in Cluster 1600. Finally, a high-level comparison between PSSP 3.5 and the new IBM @server Cluster 1600 Cluster Systems Management software is provided. Appendices describing Cluster 1600 scalability rules, a sample switch management script, PSSP hints and tips, and AIX device drivers are also included.

The team that wrote this redbook

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

Dino Quintero is a Project Leader at the ITSO Poughkeepsie Center. He currently concentrates on pSeries clustering technologies by writing Redbooks and teaching workshops.

Bernhard Buehler is an instructor, based at the IBM Learning Service Center in Herrenberg, Germany. Before joining the Learning Service Center, he worked as an HACMP specialist at the IBM RS/6000 and AIX Center of Competence, IBM Germany. He has worked at IBM for 21 years and has 12 years of experience in the AIX field. His areas of expertise include AIX, HACMP, RS/6000 SP, and HAGEO. Bernhard is a coauthor of the IBM Redbooks *DataJoiner Implementation and Usage Guide*, *Enterprise-Wide Security Architecture and Solutions Presentation Guide*, and *HACMP Enhanced Scalability Handbook*. He has also contributed to the development of some courses in the AIX curriculum. He is qualified as an IBM Certified Advanced Technical Expert - pSeries and AIX, and he is also Certified in HACMP and SP.

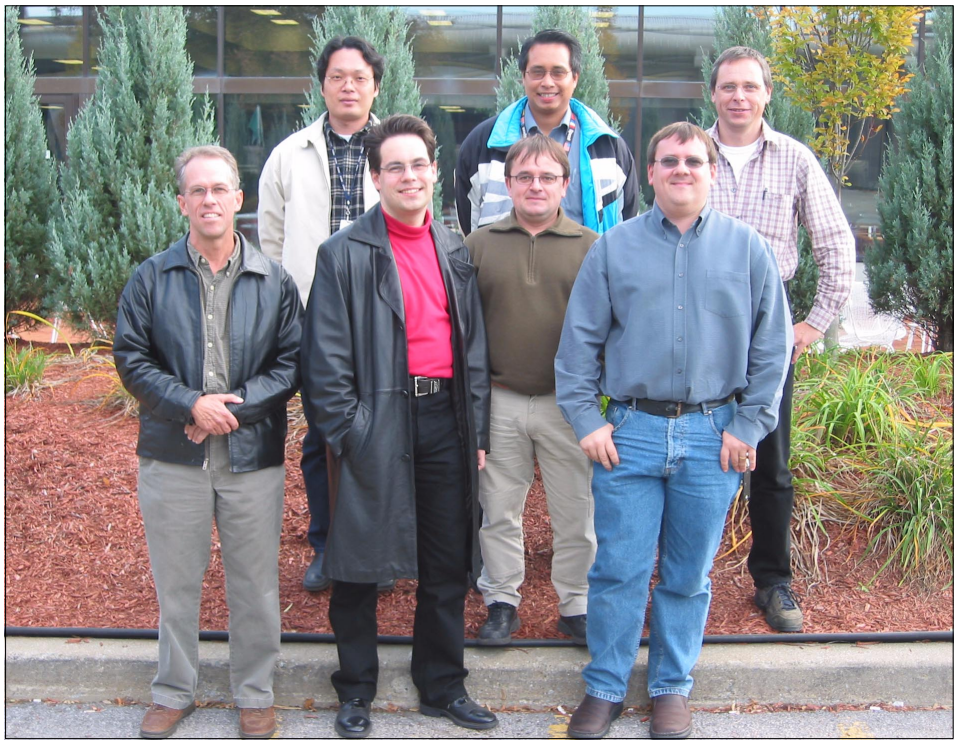
Peter Custerson is a Product Support Specialist based in Weybridge, United Kingdom. He has worked for IBM for six years, four years for the former Sequent organization and two years for the AIX Support line. He currently is the SP Technical Advisor for the UK UNIX Support Centre and concentrates on customer support issues with the PSSP product. He holds a degree in Computer Studies from the University Of Glamorgan in the United Kingdom.

Mike Ferrell is currently a member of the team at the e-tp Design Center for Infrastructure and the Executive Briefing Center in Poughkeepsie, NY. He has 21 years of experience at IBM. He has held various development and marketing positions in IBM mainframe, PC, and RISC-based servers and the associated software. He participated in SP software development from the beginning and is the designer and author of dsh, as well as the designer and team lead on several other PSSP subsystems.

Roland Kunz is a Presales Technical Support Specialist based in Frankfurt, Germany. He has worked with pSeries systems and AIX for five years as a system administrator and software developer and joined IBM in 2001. He holds a degree in physics from the Johann Wolfgang Goethe University of Frankfurt. His areas of expertise include High End pSeries Systems, High Performance Computing, LoadLeveler, and NIM. Roland is qualified as an IBM Certified Advanced Technical Expert - pSeries and AIX.

JinHoon Lee is an AIX system support specialist for IBM Korea. He has worked with the RS/6000 and the pSeries post-sales support team since joining IBM six years ago. His areas of expertise include AIX, system performance tuning, HACMP, High Performance Computing, and GPFS.

Laszlo Niesz is an IBM certified Advanced Technical Expert from IBM Hungary. He is on assignment at IBM e-Business Service Delivery in Germany. He has five years of experience in AIX-based HACMP and PSSP systems. He holds a Computer Programmer degree from University of Szeged, Hungary. His areas of expertise include implementation and management of HA clustering solutions for Oracle and Tivoli. He has written extensively on migration.



Team members (left to right):

Front: Mike Ferrell, Roland Kunz, Peter Custerson
Middle: Laszlo Niesz
Rear: JinHoon Lee, Dino Quintero (project leader), Bernhard Buehler

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

Paul Swiatocha Jr., Brian Croswell, Brian Herr, Waiman Chan, Joan McComb, Dr. Rama Govindaraju, Gordon Mcpheeters, Sarah S. Wong, Mary C. Nisley, Skip Russell, Michael J. Miele, Michael K. Coffey, Octavian Lascu, Deborah Lawrence, Bruno Bonetti, Mike Coffey, Lissa Valletta, Bernard King-Smith
IBM Poughkeepsie

Marge Momberger
Watson Research Center

Einar G. Normann
IBM Austin

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

1

IBM eServer Cluster 1600

This chapter contains a high-level overview of the IBM @server Cluster 1600 and a quick guide to what is new in the Cluster 1600 at the time of this publication.

The focus on this chapter is the Cluster 1600, the cluster components, and the new features of the Cluster 1600. This chapter is primarily conceptual in nature.

This chapter discussed the following topics:

- IBM eServer Cluster 1600 defined
- Components of IBM eServer Cluster 1600
- What's new in Cluster 1600
- PSSP 3.5: Should I upgrade

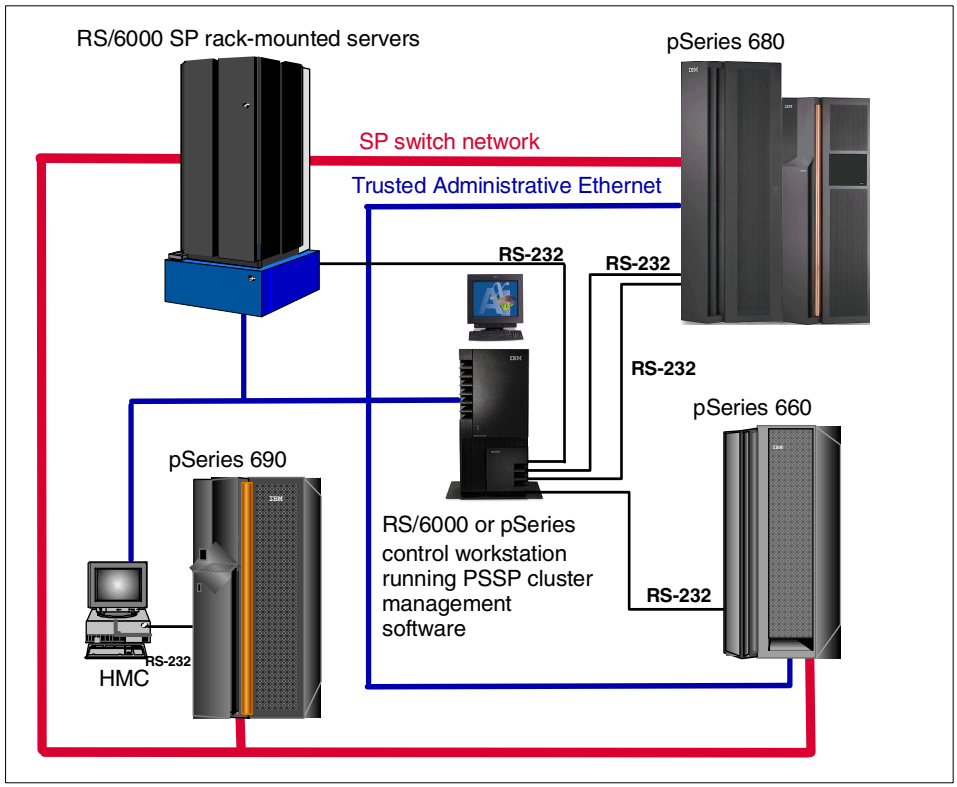


Figure 1-1 Example of an IBM eServer Cluster 1600 managed by PSSP

Figure 1-1 is a fairly elaborate example of a pSeries Cluster 1600. It consists of an RS/6000 SP rack-mounted servers and several large SMP servers, including the IBM @server pSeries p690 Model 681 (p690). In this case, all the systems are attached to a high-speed SP switch network. Also depicted are the Hardware Management Console (HMC), used to control POWER4 hardware, and a central management console. In this example, we are using PSSP as the cluster management software. PSSP calls the central management console a control workstation (CWS). The CWS is serially attached to most of the servers to provide hardware control. The exception is the p690, where the console has an Ethernet connection to the HMC, which in turn, uses a serial connection to control the p690 hardware.

1.1 Clusters defined

A *cluster* is a set of two or more computers with some other unifying characteristics. There are several different types of clusters. Let's distinguish among the types of clusters for the purposes of discussing the Cluster 1600. Each of these clusters is more tightly clustered than the previous:

Islands: This is the cluster in your basement, that pile of PCs without Ethernet cards that you used to play games on. A cluster? No.

Partitions: These can be log*ical* or *physical* partitions, or *virtual machines*, but are all separate computers residing on the same piece of hardware. These computers, all running their own operating system (OS) images and applications, and having their own network addresses, may or may not be part of a cluster, depending on how they are networked and managed.

- An example of physical partitioning is the planned support for growing the xSeries x440 in 4-way stackable boxes from a 4-way all the way to 16-way with a scalable non-uniform memory access (NUMA) interconnect between the chunks. The system can be split up into physical partitions as well, each running its own Linux or Windows OS.
- More flexible is the pSeries POWER4 systems (p690, p670, p655, p650), which can be divided into separate AIX or Linux machines through logical partitioning. A firmware hypervisor slices up the physical resources of the server (I/O, CPU, and memory) so that multiple operating systems can be run simultaneously. In fact, logical partitions can be allocated with fractional central processing unit (CPU) and I/O resources. IBM mainframes have had this kind of fractional logical partitioning for years with the PRSM technology. pSeries currently offers logical partitioning on CPU and I/O slot boundaries, but will offer fractional partitioning in the future.
- IBM invented virtual machines, of course, with the VM hypervisor for the mainframe. Virtual machines are created by a software hypervisor that manages special kernel-involved events, including privileged operation interrupts, I/O interrupts, page faults, timers, and so on. These are dealt with by the hypervisor, but the results are presented to the OS and software on the virtual machines in a way identical to that presented by a real machine. CPU resources are time-sliced by the hypervisor. Thus, the OS and software run as if they had their own hardware. Because only kernel-involved events are simulated, the large bulk of the operation of the virtual machine's software runs directly on the metal during its time-slice, offering good performance. z/VM is still offered on the zSeries and is a popular choice for those wanting to run multiple Linux images on these systems. VMWare ESX Server, in partnership with IBM, offers similar virtualization for the PC server architecture on many xSeries servers.

Connected: Any set of networked computers. A cluster? Maybe, but not in this book.

Distributed manageability: This is a set of computers on a network or set of networks that can be monitored or managed, or both, from a single point of control through specialized software. These computers need not have anything in common, other than running network-enabled agents that communicate with a central management console. Some of these systems may be servers, others desktops. This could be a cluster in the loosest sense, although many would argue that this is not a cluster at all. The IBM NetView product managing far-flung SNMP agents could be considered a cluster in this sense. IBM Director, utilized for managing and controlling hundreds of IBM Intel-based xSeries servers is another example.

High-availability cluster: This consists of two or more computers connected and configured in a way that eliminates single points of failure in a server system. A simple configuration could consist of two similarly-configured servers connected to the same storage device so that if one of the servers stopped providing service, the other could take over, utilizing the same data, IP addresses, and applications that the failed system was using. A sophisticated high-availability clustering software product, such as IBM HACMP/ES, can cluster up to 32 servers in this way.

Manageability cluster: This cluster is more tightly coupled than the distributed manageability cluster, in the sense that the management software has the capability to manage and control computers within it as a group or groups of systems with common characteristics as opposed to a set of individual systems. Manageability clusters often include separate networks for secure management of systems and will typically be used to manage only server computers, often referred to as *nodes*. IBM Cluster Systems Management (CSM) for Linux, for example, can be used quite easily to manage and monitor many xSeries servers running Linux. CSM for Linux is included in the IBM Cluster 1350, a pre-built, rack-mounted Linux cluster utilizing up to 512 1U xSeries servers.

Manageability cluster with hardware control: Modern servers are provided with *service processors*. These are specialized computers that are closely integrated with the server hardware and allow that hardware to be managed, monitored, powered up and down, booted, and so on, from a network connection to the service processor, even when the server is remotely located and powered off. They may offer out-of-band alerts on impending hardware malfunction. They also typically offer an emulation of the server console. This means that an operator need not be near the machine to power it up and boot the OS, and then monitor the boot process, for example.

Service processors vary in function and how they are networked, but it is convenient if cluster management server software understands the protocols of the service processors and allows a single point of control for the cluster hardware integrated with the rest of the server management software. CSM for Linux software understands the service processors in rack-mounted xSeries servers. When coupled with sophisticated rack hardware and remote keyboard, video, mouse (KVM) switches, hundreds of Linux servers can be located within the footprint of several racks and managed from anywhere in the enterprise. The IBM NetBay line of server rack and KVM hardware features hassle-free cabling and power, as well as secure remote console access to the racked servers. IBM Director also has excellent GUI-based hardware management of service processor-equipped IBM Intel servers.

The POWER4-based pSeries servers have a service processor that consists of a PC (Hardware Management Console or HMC) running special software, a serial connection from the HMC to the server, and firmware within the server controlled by the HMC. Cluster 1600 features software to control an HMC, and thus the POWER4 server hardware itself.

Hardware control can be extended beyond the actual server machines. For example, Enhanced Clustered Tools for Linux, available from alphaWorks, takes function from the IBM xCAT Linux cluster configuration software that allows power control of APC switches and terminal servers, as well as the server nodes.

Performance (loosely connected): The clusters described so far consist of machines that may be running software that is completely independent (except in the case of high-availability clusters) of the software applications on the other machines. Many applications require more processing power than is contained in a single machine if they are to complete in a useful amount of time. Multiple nodes can co-operate on a single application using parallel programming techniques. A loosely connected application requires little or no communication between the software on different nodes while it is running. This would be the case, for example, in an application where the data was such that it could be processed in discrete chunks, where boundaries between chunks did not affect processing of other chunks. A parallel application to convert CD tracks to MP3s is an example of such an “embarrassingly parallel” application, because each server could convert part of the CD independently, and all could run concurrently.

Note: Even loosely connected performance clusters can benefit from special software to schedule, distribute, and collect the results of such jobs. IBM LoadLeveler and Parallel Environment applications are well-received examples.

Performance (tightly connected): These are clusters specifically utilized for high-performance computing. They are equipped with special interconnects that provide low latency and high bandwidth communication between nodes, as well as libraries and device drivers that provide message-passing parallel programs with an efficient way to communicate during execution across many nodes. The paradigmatic example is the Cluster 1600 managed by PSSP supercomputer, with nodes connected through an SP switch.

This type of cluster requires a considerable amount of software to support applications. In addition to the software needed to manage conveniently the large number of nodes that can be involved, device drivers for the interconnect, software to manage the interconnect topology, and kernel-level code for the fastest possible data transfer across the interconnect within the message-passing libraries are needed. These kinds of software are exemplified by IBM PSSP, KLAPI, VSD, and Parallel ESSL software utilized in the RS/6000 SP.

Important: We would like to stress that PSSP and CSM could be used to construct any of these types of clusters.

SMP: Symmetric multiprocessor computer, not really a cluster. IBM @server offers up to 32 processors in the pSeries p690, up to 32 on the iSeries i890, up to 16 on the zSeries z900, and up to 16 processors with the xSeries x440.

Multiple CPU microprocessor chip: This is just an opportunity to mention the state-of-the art POWER4 microprocessor utilized in the newest pSeries servers. Each microprocessor chip has two 64-bit 1 GHz CPUs on it. The ultimate 2-way cluster.

Important: Do not confuse this with multi-chip module (MCM).

1.2 IBM eServer Cluster 1600 defined

IBM @server Cluster 1600 is an umbrella name for clusters managed by IBM pSeries servers and associated software. Hardware and software can be chosen and combined based on customer needs. Cluster 1600 provides the building blocks to build the world's most capable UNIX/Linux clusters.

In terms of the cluster typology previously described, Cluster 1600 building blocks can be put together to provide high-availability clusters, manageability clusters, manageability clusters with hardware control, and loosely and tightly connected performance clusters. As described, these types are not necessarily

mutually exclusive. There is no reason why servers in a manageability cluster cannot also be clustered for high availability, for example.

1.2.1 Components of IBM eServer Cluster 1600

The components available for building an IBM @server Cluster 1600 are as follows.

pSeries servers

IBM @server pSeries servers are the world's most reliable and fastest UNIX servers, featuring the most advanced microprocessors and mainframe-inspired reliability, availability, and serviceability (RAS).

AIX 5L

AIX 5L is the IBM enterprise-class UNIX operating system. It now includes Reliable Scalable Cluster Technology (RSCT), software specifically designed to allow AIX servers to be built into a Cluster 1600.

Cluster management software

PSSP 3.5 on AIX 5L Version 5.1 and 5.2 in 2003 for existing or new PSSP, High Performance Computing (HPC), and commercial customers, and Cluster Systems Management (CSM) on AIX 5L Version 5.2 and AIX 5L Version 5.1 (Maintenance Level 3 or later) for new customers are offered as cluster management software components. Both share the same heritage, the IBM RS/6000 SP supercomputer, and both allow clusters of up to 128 servers (or compute nodes), or larger by special bid, to be controlled.

High availability cluster software

High-Availability Cluster Multiprocessing (HACMP) 4.5, the leading UNIX high-availability cluster software, allows robust multiserver configurations to be built, protecting against a wide variety of network, software, and hardware failures. HAGEO/GeoRM 2.4 makes it possible to duplicate data between distant sites so that either automated (HAGEO) or manual (GeoRM) disaster recovery can be implemented.

High performance computing cluster software

Specialized software is required for running, managing, and scheduling parallel supercomputing code and jobs. Cluster 1600 offers Parallel Environment 3.1 for writing and controlling parallel codes, LoadLeveler 3.1 for scheduling jobs, and ESSL 3.3 and Parallel ESSL 2.3 libraries for scientific and engineering calculations. All these are supported by the PSSP 3.5 management functions.

High performance cluster interconnects

Cluster 1600 supports two generations of high-speed interconnect switch hardware to connect cluster nodes together: the SP Switch and the SP Switch2. Although primarily aimed at the HPC customer, these interconnects support IP and thus can be used as a high-bandwidth LAN for purposes of backups, database connectivity, or other bandwidth-hungry applications. In addition, some databases, such as the parallel implementation of DB2, are designed to run a query across multiple nodes in a cluster simultaneously, with each node accessing a subset of the relational tables required for the query. These databases can make use of the switch interconnect for fast communication of query results back to the coordinating node. Data warehouses with dozens of terabytes of online data have been implemented in this fashion.

Cluster file system

GPFS for AIX 2.1 allows multiple systems to concurrently access a large file system over multiple I/O paths, while preserving standard UNIX file system semantics.

1.3 What's new in Cluster 1600

Here, we provide a brief description of what is new in Cluster 1600. This is intended for those already familiar with Cluster 1600.

1.3.1 New hardware support

Cluster 1600 can utilize the latest in pSeries POWER4 technology as follows:

- The pSeries p670, p655, and p630 are now supported as cluster-attached servers, including switch attachment.
- The p630 can be used as a central point of control.
- A new remote input/output (RIO) drawer featuring peripheral component interconnect extended (PCI-X) expansion slots and hot-swappable disk is available as a feature for a node.

For more information about the newest components of Cluster 1600, refer to Chapter 2, "New hardware" on page 13, and the *Read This First* document.

1.3.2 AIX 5L

AIX 5L is not officially part of Cluster 1600, but is integral to it. The enhancements in AIX 5L Version 5.2, in particular Dynamic Logical Partitioning support, are too numerous to go into here. For more information, see *AIX 5L Differences Guide Version 5.2 Edition*, SG24-5765. One feature that deserves mention from a Cluster 1600 perspective is the Reliable Scalable Cluster Technology (RSCT) that is now included in AIX 5L. This technology, originally from PSSP, enables a reliable and scalable cluster-wide view of events, cluster membership, and hardware and software states. It also offers cluster-wide control of hardware and software resources by authenticated and authorized users and processes. This technology will be exploited by both Cluster 1600 and AIX software components going forward, and provides an unequalled framework into which new cluster technology can be plugged. For more detailed information about RSCT, refer to *IBM RSCT for AIX: Guide and Reference*, SA22-7889.

1.3.3 Parallel System Support Program 3.5 on AIX 5L Version 5.1

Parallel System Support Program (PSSP) 3.5 offers the following enhancements and changes from previous releases:

- The 64-bit kernel version of AIX 5L Version 5.1, including any 64-bit kernel extensions and device drivers, can be used within a PSSP 3.5 cluster. The entire software stack supported by PSSP 3.5, GPFS, MPI, LoadLeveler, Parallel ESSL, LAPI, KLAPI, and VSD, works with AIX 5L Version 5.1 64-bit kernels, device drivers, and kernel extensions.
- Virtual Shared Disk (VSD) has been enhanced for performance in IP-based mode and adds improved diagnostics.
- Switch support has been enhanced to allow an administrator to specify sets of nodes to be excluded from serving as switch primary or primary backup.
- IBM intends to support PSSP 3.5 on AIX 5L Version 5.2 in 2003. (AIX 5L Version 5.2 support will not be offered in PSSP 3.4.) Also, PSSP 3.5 does not support AIX 4.3.3.
- PSSP 3.5 (and PSSP 3.4) will be able to utilize redundant HMCs attached to a logically partitionable server, providing better availability.

Attention: PSSP 3.5 will be the last release of PSSP. For more information about the PSSP and CSM plans, contact your IBM technical representative, or see:

`http://www.ibm.com/servers/eserver/clusters/software/`

For more information about the latest features of PSSP 3.5, refer to Chapter 4, “Parallel System Support Program 3.5 enhancements” on page 69.

1.3.4 Cluster Systems Management 1.3 for AIX 5L 5.2

Cluster Systems Management (CSM) 1.3 is available at this time for Cluster 1600 customers that are more interested in *manageability* clusters than *performance* clusters. Customers that do not have a cluster managed by PSSP or need a switch interconnect, and yet desire the cluster system management capability that PSSP customers have known, can choose CSM. In addition, CSM clusters can include Intel-based Linux nodes as of December 2002. CSM requires AIX 5L Version 5.2 for the management server and AIX 5L Version 5.1 with Maintenance Level 3 or later for the managed nodes. For details about CSM 1.3 for AIX 5L 5.2, see *An introduction to CSM 1.3 for AIX 5L*, SG24-6859 or Chapter 7, “Cluster 1600 management: PSSP and CSM” on page 169.

1.3.5 General Parallel File System for AIX Version 2.1

The following features are available in General Parallel File System (GPFS) Version 2.1:

- The AIX 64-bit kernel is supported and exploited.
- Customers using GPFS 2.1 and PSSP 3.4 or 3.5 and AIX 5L Version 5.1 no longer need to purchase and install HACMP as a prerequisite. GPFS 2.1 can utilize the RSCT technology in AIX 5L Version 5.1 instead of HACMP.

For more information about GPFS Version 2.1, see Chapter 5, “General Parallel File System 2.1” on page 99.

1.3.6 High Availability Geographic Cluster and GeoRM 2.4

The following features are new with High Availability Geographic Cluster (HAGEO) and Geographic Remote Mirror (GeoRM) Version 2.4:

- TCP/IP is added to UDP/IP as a data replication transport mechanism between clustered sites. This improves performance when there is more than one LAN segment between sites, for example, if there are bridges or routers between sites.
- Volume group write order can be user-specified with the TCP protocol option.
- HAGEO configuration is simplified by allowing the use of existing HACMP cluster site definitions when defining HAGEO clusters.
- 64-bit kernel support is added for the TCP transport option.

For more information about the features of HAGEO and GeoRM, refer to the redbook *Configuring Highly Available Clusters Using HACMP 4.5*, SG24-6845.

1.4 PSSP 3.5: Should I upgrade

PSSP 3.5 offers functional enhancements over previous releases of PSSP, but customers may have important reasons to upgrade. Such customers may include:

- Customers who want to run in 64-bit kernel environment running 64-bit applications that can exploit existing or utilize 64-bit kernel extensions and device drivers in either AIX or PSSP, or both.
- Customers that use the switch and want to exclude particular nodes from allocation as switch primary or primary backup nodes.
- Customers that want to use AIX 5L Version 5.2 features, such as Dynamic Logical Partitioning. PSSP 3.5 is the only version of PSSP that will be available to support AIX beyond AIX 5L Version 5.1. IBM intends to support PSSP 3.5 with AIX 5L Version 5.2 in 2003.

2

New hardware

This chapter provides information about the latest hardware additions as part of the Cluster 1600 managed by PSSP announcement. For more information, refer to the IBM Parallel System Support Programs V3.5: New function and hardware support software announcement letters 202-263, 202-264, and 202-265 from October 8, 2002.

This chapter also includes descriptions of the following new pSeries servers:

- The p630 server (type 7026-6C4)
- The p650 server (type 7038-6M2)
- The p655 server (type 7039-651)
- The p670 server (type 7040-671)

This chapter describes the new options in Cluster 1600:

- A new SP Switch2 PCI-X Attachment Adapter (FC 8398)
- A new 19-inch switch frame 9076-558 for integrating up to two SP Switch2 switches in a single rack
- A new 24-inch 7040-W42 frame for integrating the p655 and 7040-61D I/O drawers
- A new processor option for legacy Winterhawk-II SP nodes using 450 MHz POWER3 SMP thin and wide nodes

We also introduce the p650 (type 7038-6M2) to the reader, because IBM intends to support it in Cluster 1600 soon.

All features described in this chapter are already supported by PSSP Versions 3.4 and 3.5 except the p650. The following sections introduce each new model and explain the features and benefits in detail, thus making it easier to find the right cluster component for the appropriate workload.

2.1 The p630 server

This section describes the functionality and the features of the IBM pSeries p630 server (type 7028-6C4). We describe the following:

- An overview is given in 2.1.1, “Introduction to the p630 server” on page 15.
- The layout of the CPU board is described in 2.1.2, “CPU board layout” on page 16.
- The design of the system board is outlined in 2.1.3, “System board design” on page 17.
- Software considerations for this machine are given in 2.1.4, “Software requirements” on page 18.
- Considerations for integrating the p630 into a Cluster 1600 are discussed in 2.1.5, “Cluster considerations” on page 19.

2.1.1 Introduction to the p630 server

The IBM pSeries p630 is a 1-, 2-, or 4-way SMP machine running the IBM POWER4 microprocessor at 1.0 GHz. It is a small, densely-packed rack server, suitable for installation in a T42 rack. The 4U height allows up to 10 machines in a single rack. The p630 is the first pSeries server to include PCI-X buses. PCI-X runs at 133 MHz bus speed with a 64-bit wide bus. It is also the first pSeries server where the implementation of the POWER4 chip is realized as a single-chip module (SCM). Two integrated 10/100 Mbps Ethernet ports are included. Two integrated Ultra 3 SCSI Controller and four bays for hot-swapable disk drives provide internal storage of up to 293.6 GB with 73.8 GB disk drives. When used as a stand-alone SMP server, no HMC needs to be attached, but for use in a Cluster 1600, there must be an HMC connection. With PSSP 3.4 and 3.5, up to 64 nodes of this type are supported in a Cluster 1600. The use of logical partitions (LPARs) within this system is not supported by PSSP. Up to 16 p630s can be controlled from a single HMC. Either integrated Ethernet adapter can be used from PSSP as a management Ethernet, eliminating the need for an additional management Ethernet adapter. For more information about this model, refer to *pSeries 630 Models 6C4 and 6E4 Technical Overview and Introduction*, REDP0193.

Figure 2-1 The p630 6C4

2.1.2 CPU board layout

A single POWER4 microprocessor in this machine is packaged on a CPU card as an SCM. All other pSeries (except the p650) have so called multi-chip modules (MCM), where two POWER4 processors are packed into one module. One single-chip module (SCM) is mounted on a CPU board, where 32 MB of Level 3 cache and up to 16 GB DDR memory are installed. The processor communicates with its local memory through its local Level 3 cache and two memory controllers. Each board can use the memory and Level 3 cache of either board. A board is depicted in Figure 2-2 on page 17. The theoretical bandwidth of the local memory on the card is 6.4 GB/s, using all four channels.

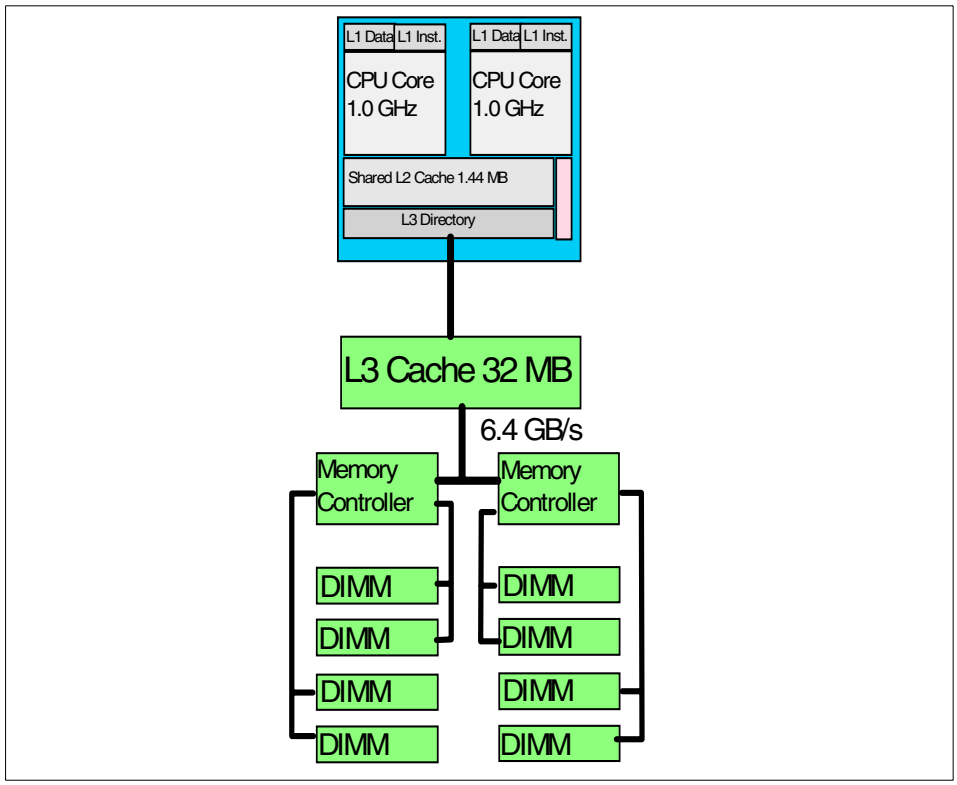


Figure 2-2 Design of the CPU board of the p630 with a POWER4 SCM

2.1.3 System board design

As mentioned in 2.1.1, “Introduction to the p630 server” on page 15, one or two CPU boards can be installed in the machine. The two system boards are connected through a fabric bus providing two 64-bit pathways at 500 MHz, allowing a peak of 8 GB/s shared between the two cards. Both cards can use the memory on either card. The connection to the system board is accomplished using the p690 GX 32-bit wide bus operating at 333.3 MHz. There are two buses, thus allowing a theoretical peak of 2.66 GB/s. The system is connected to the I/O subsystem through a remote I/O bridge. Two different PCI to PCI bridges are included, each of them having two PCI-X slots and an integrated 10/100 Mbps Ethernet controller. All the I/O ports are connected through an Industry Standard Architecture (ISA) bridge. Additionally, an IDE CD-ROM drive can be used. DVD components are small computer system interface-based (SCSI). See Figure 2-3 on page 18.

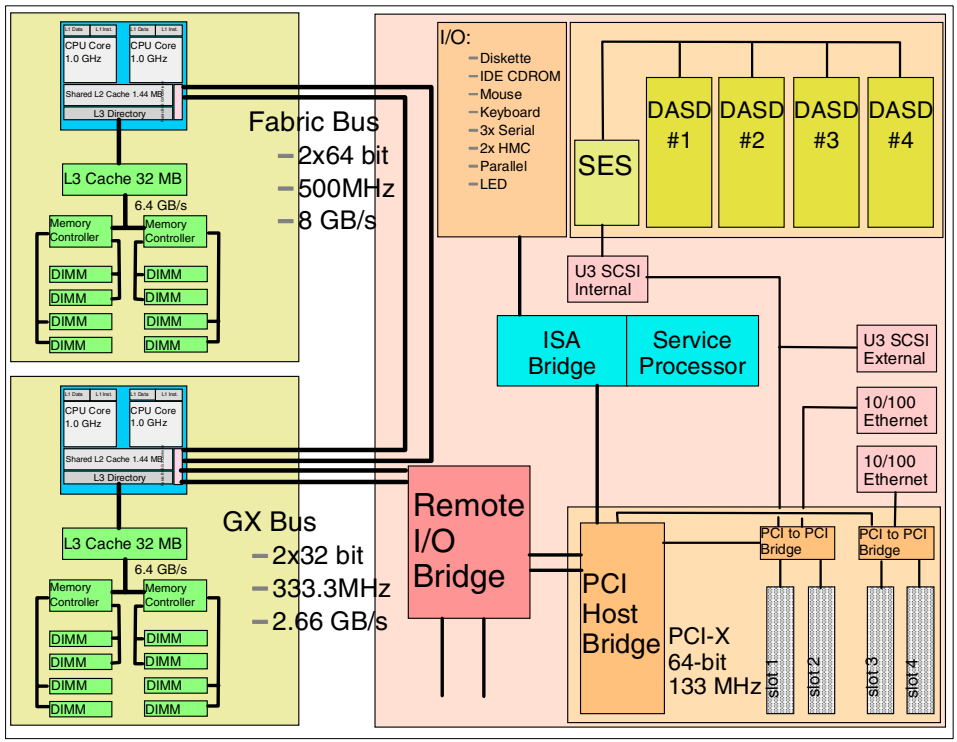


Figure 2-3 Design layout of the p630

2.1.4 Software requirements

To use a 64-bit AIX kernel, AIX 5L Version 5.1 with Maintenance Level 2 is required together with PSSP 3.4 with APAR IY24792 or PSSP 3.5.

Example 2-1 on page 19 shows how a p630 is recognized by PSSP.

Important: Only PSSP 3.5 supports the 64-bit kernel of AIX 5L Version 5.1.

Example 2-1 p630 in a Cluster 1600

```
c166s][/]> splstdata -n 8 1 1
                  List Node Configuration Information

node# frame# slot# slots initial_hostname reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- --------------- --------- --------------- ----------
  113      8     1     1 c166ler01.ppd.pok c166ler01.ppd.pok ""                 9.114.72.125
MP                                2 7028-6C4                1 true            c166ler
```

The speed of its processors and a typical configuration can be determined with the **lsattr** and **lsdev** commands, as shown in Example 2-2.

Example 2-2 CPU speed of the p630 and typical configurations

```
[c166ler01][/]> lsattr -El proc0
state     enable         Processor state False
type      PowerPC_POWER4 Processor type  False
frequency 1000000000     Processor Speed False

[c166ler01][/]> lsdev -Cc adapter
ppa0    Available 01-R1    CHRP IEEE1284 (ECP) Parallel Port Adapter
sa0     Available 01-S1    Standard I/O Serial Port
sa1     Available 01-S2    Standard I/O Serial Port
sa2     Available 01-S3    Standard I/O Serial Port
siokma0 Available 01-K1    Keyboard/Mouse Adapter
fda0    Available 01-D1    Standard I/O Diskette Adapter
ide0    Available 1G-19    ATA/IDE Controller Device
ent0    Available 1L-08    10/100 Mbps Ethernet PCI Adapter II (1410ff01)
scsi0   Available 1S-08    Wide/Ultra-3 SCSI I/O Controller
scsi1   Available 1S-09    Wide/Ultra-3 SCSI I/O Controller
ent1    Available 11-08    10/100 Mbps Ethernet PCI Adapter II (1410ff01)
ent2    Available 14-08    10/100 Mbps Ethernet PCI Adapter II (1410ff01)
ent3    Available 1D-08    10/100 Mbps Ethernet PCI Adapter II (1410ff01)
sioka0  Available 01-K1-00 Keyboard Adapter
sioma0  Available 01-K1-01 Mouse Adapter
css0    Available 1H-08    SP Switch2 Communications Adapter
ent4    Available 1V-08    Gigabit Ethernet-SX PCI-X Adapter (14106802)
```

2.1.5 Cluster considerations

The p630 can be integrated into a Cluster 1600 with the following configurations:

- A switchless cluster with the p630 containing no switch adapters.
- A cluster with a single or a double plane SP Switch2, where the p630 is off the switch.

- A cluster with a single plane SP Switch2, where the p630 is on the switch using one SP Switch2 PCI Attachment Adapter (FC 8397) in one of its PCI-X slots. Use slot 3.

2.2 The p655 server

This section describes the functionality and the features of the IBM pSeries p655 server (7039-651). The following is provided:

- An introduction is given in 2.2.1, “Introduction to the p655 server” on page 20.
- The layout of the CPU board is described in 2.2.2, “CPU board layout” on page 22.
- The design of the system board is outlined in 2.2.3, “System board design” on page 22.
- Software considerations for this machine are given in 2.2.4, “Software requirements” on page 24.
- Considerations for integrating the p655 into a Cluster 1600 are discussed in 2.2.5, “Cluster considerations” on page 26.

2.2.1 Introduction to the p655 server

For no other machine, the statement “The SP frame is dead, long live the SP” is more true. It re-introduces the concept of thin nodes (half rack width) with a height of 4U in a frame. The p655 is a 4- or 8-way SMP machine running the IBM POWER4 microprocessors utilizing the same MCM packaging technology as the p690 and p670. Two versions are available: an 8-way system running at 1.1 GHz and one HPC 4-way system running at 1.3 GHz. Up to 32 p655s are supported in a Cluster 1600 with PSSP 3.4 or PSSP 3.5. This offers over 500 mega floating-point operations per second (MFLOPs) of clustered performance. Although this machine is capable of running AIX 5L Version 5.2, the use of dynamic LPAR as introduced in AIX 5L Version 5.2 is not supported on PSSP. (AIX 5L Version 5.2 will be supported on PSSP 3.5 next year.) Two AIX 5L Version 5.1 LPARs on each p655 are supported, and up to 32 LPARs can be controlled from a single HMC. Because this is a HMC-based protocol system to PSSP, an HMC is required. Two integrated 10/100 Mbps Ethernet connections, which can be both used by PSSP as a management Ethernet, two remote I/O ports, two serial lines (used for HMC connections), and a parallel port are located on the back of the machine. Additionally, three PCI-X slots are integrated together with two Ultra320 SCSI controllers. In front of the machine, two SCSI disks can be mounted. Figure 2-4 on page 21 shows the front view of a p655.

Figure 2-4 The p655 front view without covers and SCSI disks

A minimum of two disks are required, which can be 18.2, 36.4, 73.4 or 146.8 GB each. This machine has no CD-ROM, DVD, tape, or removable disk drive connected to it, so installation can only be done through Network Installation and Maintenance (NIM) or Parallel System Support Program (PSSP). Optionally, a full or half 7040-61D drawer can be attached, providing up to 16 more disks and 20 PCI slots. With the p655, it is also possible to share such a drawer within two nodes, where each node is attached to one side and owns its side exclusively. This machine is very suitable for HPC or business intelligence (BI) applications or for SP customers who want to consolidate their workloads. See also *pSeries p655 Installation Guide*, SA38-0616, for details. A fully-populated rack with 16 p655 weighs up to 1584 kg (3484 lbs). When installing a p655 into the frame, the left side will be filled first, then on the same height position, the right side. An Early Power Off Warning (EPOW) capability is provided to assist an ordered shutdown of the system.

2.2.2 CPU board layout

As in the p670 and the p690, the CPU board design is based on the MCM concept, where four dual-core processors are mounted in one module. In contrast to the p690, where up to four MCMs are screwed on the backplane on the back of the machine, the MCM sits on the system planar. The p655 has a 32 MB cache and memory of up to 64 GB. Two different options exist for this model, a fully populated 8-way system running at 1.1 GHz and an HPC system running at 1.3 GHz, where only four processor cores are enabled. This provides double the size of the L2 and L3 cache and increases the memory bandwidth per processor. Together with the higher frequency, this HPC solution provides power, where bandwidth matters. Figure 2-5 shows the layout of an HPC MCM.

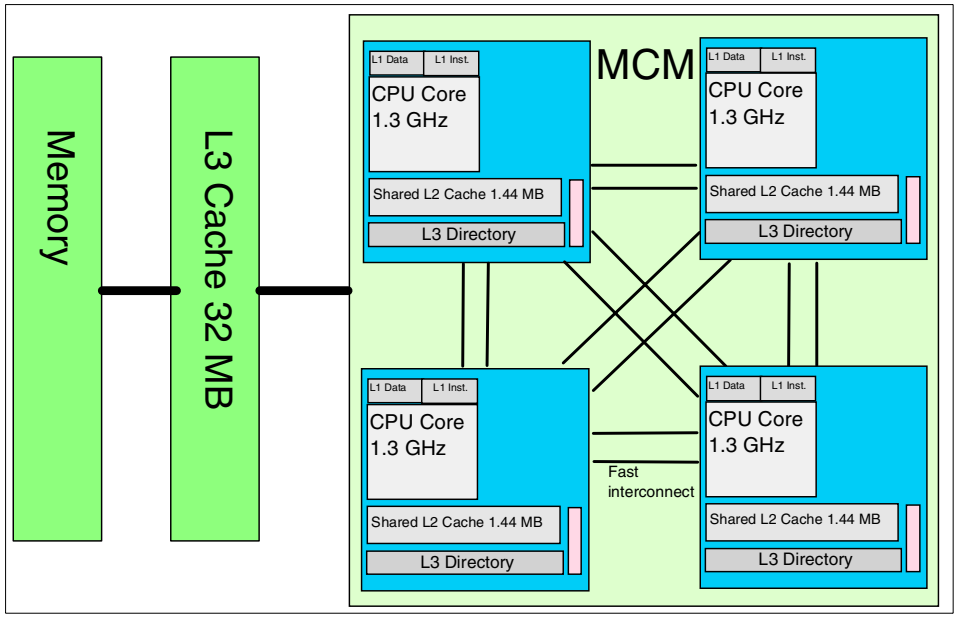


Figure 2-5 Simplified layout of a High Performance Computing MCM

2.2.3 System board design

The p655 has four memory slots that can be populated with 4 GB (FC 4456) or 8 GB (FC 4457) memory modules. Table 2-1 on page 23 gives an overview of the possible configurations. Balanced (equally populated) memory placement is desired, because it can increase throughput for some applications. Each PCI-X slot and the two integrated Ultra320 SCSI controllers can be allocated individually to each LPAR. It is recommended to have at least 4 GB memory per LPAR.

Table 2-1 Memory placement rules for the p655

Total memory	Slot 1	Slot 2	Slot 3	Slot 4
4 GB	4 GB	-	-	-
8 GB	4 GB	-	4 GB	-
16 GB	4 GB	4 GB	4 GB	4 GB
16 GB	8 GB	-	8 GB	-
32 GB	8 GB	8 GB	8 GB	8 GB

Furthermore, two integrated Ultra320 SCSI Controllers, two 10/100 Mbps Ethernet controllers, and a serial and parallel I/O is provided. The MCM is mounted directly on the sysplanar at the bottom of the machine, in contrast to the p670, where the MCMs are mounted and screwed on the back of the machine. Figure 2-6 shows a photo of an open p655.

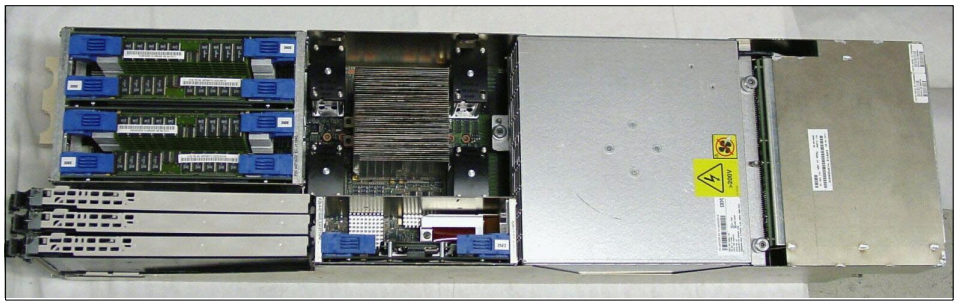


Figure 2-6 Photo of an open p655

Figure 2-7 on page 24 shows a simplified layout of the front, back, and top of the p655.

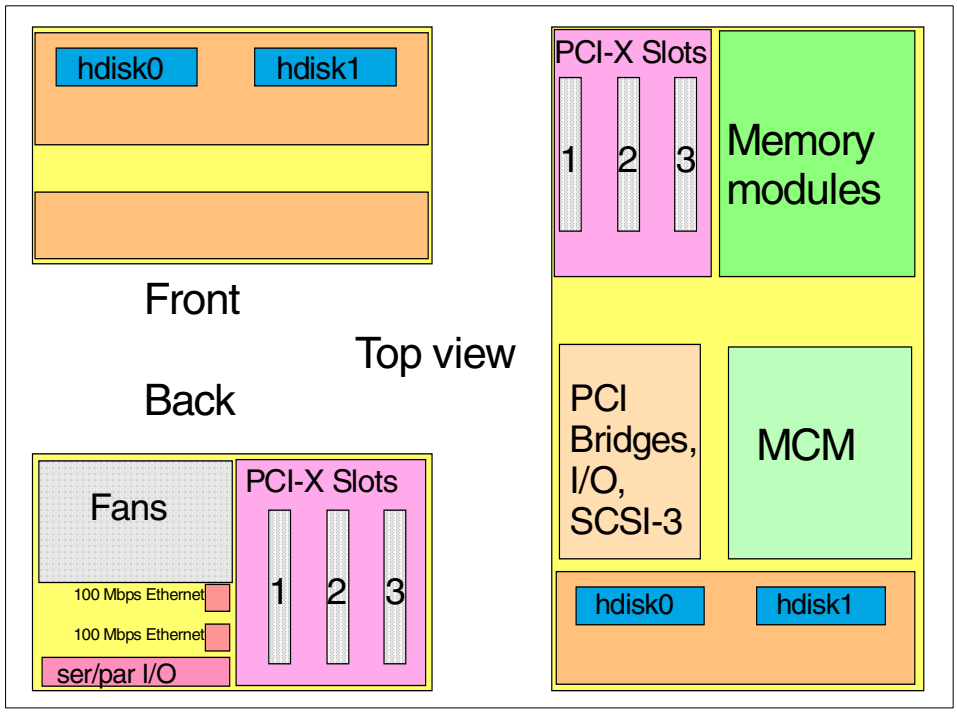


Figure 2-7 Simplified layout of the p655

2.2.4 Software requirements

AIX 5L Version 5.1 with Maintenance Level 3 is needed for installation of the base operating system and the support of PSSP 3.5. Furthermore, APAR IY34495 for PSSP 3.4 or APAR IY34496 for PSSP 3.5 is required for support with PSSP. When running on a 64-bit kernel, only PSSP 3.5 is supported. Example 2-3 shows how PSSP recognizes the p655.

Example 2-3 p655 in a PSSP 3.5 managed cluster

```
[c179s][/]> splstdata -n 11 1 1
                 List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- --------------- --------- --------------- ----------
  161     11     1     1 c59ih04.ppd.pok.i c59ih04.ppd.pok.i ""                9.114.213.125
MP                                8 7039-651               1 true            c59ih04

[c179s][/]> splstdata -b 11 1 1
```

```
                List Node Boot/Install Information

node# hostname          hdw_enet_adr  srvr response     install_disk last_install_image
last_install_time   next_install_image  lppsource_name pssp_ver            selected_vg
----- ----------------- ------------- ---- ------------ ------------ ------ -----------
------------ ------------------ ------------------ -------------- -------------------
  161 c59ih04.ppd.pok.i 00096BE80041     0 disk         hdisk0 bos.obj.node.aix51d
Fri_Sep__6_14:01:04 bos.obj.node.aix51d aix51d0212d    PSSP-3.5            rootvg

c179s][/]> splstdata -g 11 1 1
                List Aggregate IP Database Information

node# adapt  netaddr         netmask         hostname          devicename     update_interval
update_threshold
----- ------ --------------- --------------- ----------------- -------------- -----------------
------------------
  161 ml0    9.114.213.28    255.255.255.192 c179san28.ppd.pok css0,css1      3 10
```

The speed of its processors and a typical configuration can be determined, as shown in Example 2-4.

Example 2-4 Typical output of a p655

```
[c59ih01][/]> lsattr -El proc1
state     enable         Processor state False
type      PowerPC_POWER4 Processor type  False
frequency 1300000000     Processor Speed False
[c59ih01][/]> lscfg
INSTALLED RESOURCE LIST

The following resources are installed on the machine.
+/- = Added or deleted from Resource List.
*   = Diagnostic support not available.

  Model Architecture: chrp
  Model Implementation: Multiple Processor, PCI bus

+ sys0              00-00             System Object
+ sysplanar0        00-00             System Planar
+ mem0              00-00             Memory
+ L2cache0          00-00             L2 Cache
+ proc1             00-01             Processor
+ proc3             00-03             Processor
+ proc5             00-05             Processor
+ proc7             00-07             Processor
* pci2              00-3fffdf08000    PCI Bus
* pci5              10-10             PCI Bus
+ ent0              11-08             10/100 Mbps Ethernet PCI Adapter II
                                      (1410ff01)
```

```
* pci6             10-12            PCI Bus
+ ent1             14-08            10/100 Mbps Ethernet PCI Adapter II
                                    (1410ff01)
* pci7             10-14            PCI Bus
* pci8             10-16            PCI Bus
* pci0             00-3fffdf09000   PCI Bus
* isa0             1G-18            ISA Bus
+ sa0              01-S1            Standard I/O Serial Port
+ tty0             01-S1-00-00      Asynchronous Terminal
* pci1             00-3fffdf0a000   PCI Bus
* pci3             1Y-10            PCI Bus
+ sisscsia0        1Z-08            PCI-X Dual Channel Ultra320 SCSI
                                    Adapter
+ scsi0            1Z-08-00         PCI-X Dual Channel Ultra320 SCSI
                                    Adapter bus
+ hdisk0           1Z-08-00-8,0     16 Bit LVD SCSI Disk Drive (36400
                                    MB)
+ scsi1            1Z-08-01         PCI-X Dual Channel Ultra320 SCSI
                                    Adapter bus
+ hdisk1           1Z-08-01-8,0     16 Bit LVD SCSI Disk Drive (36400
                                    MB)
* pci4             1Y-16            PCI Bus
* css0             1n-08            SP Switch2 Communications Adapter
```

Important: Only PSSP 3.5 supports the 64-bit kernel of AIX 5L Version 5.1.

2.2.5 Cluster considerations

The p655 can be integrated into a Cluster 1600 with the following configurations:

- A switchless cluster with the entire p655 as a single node.
- A switchless cluster with the p655 with two LPARs, each configured as a node.
- A cluster with a single or a double plane SP Switch2, where the p655 is off the switch and the p655 is a single node.
- A cluster with a single or a double plane SP Switch2, where the p655 is off the switch with both LPARs.
- A cluster with a single plane SP Switch2, where the p655 is on the switch using one SP Switch2 PCI-X Attachment Adapter (FC 8398) in one of its PCI-X slots and has no LPARs. Use slot 1.
- A cluster with a single plane SP Switch2, where the p655 is on the switch using one SP Switch2 PCI-X Attachment Adapter (FC 8398) in one of its PCI-X slots in one LPAR, while the other LPAR is off the switch. Use slot 1.

- A cluster with a single plane SP Switch2, where the p655 is on the switch using two SP Switch2 PCI-X Attachment Adapters (FC 8398), one in each LPAR.
- A cluster with a dual plane SP Switch2, where the p655 is on the switch using two SP Switch2 PCI-X Attachment Adapters (FC 8398). Use slots 1 and 3.
- A cluster with a dual plane SP Switch2, where the p655 is on the switch using two SP Switch2 PCI-X Attachment Adapters (FC 8398) in both PCI-X slots for one LPAR; the other LPAR is off the switch. Use slots 1 and 3.

Restriction: No SP Switch2 or SP Switch adapter is supported in the additional 7040-61D drawer.

Restriction: Attachment to an SP Switch is not supported.

2.3 The p670 server

This section describes the functionality and the features of the IBM pSeries p670 server (type 7040-671). We discuss the following topics:

- An overview is given in 2.3.1, “Introduction to the p670 server” on page 27.
- The layout of the CPU board is described in 2.3.2, “CPU board layout” on page 28.
- The design of the system board is outlined in 2.3.3, “System board design” on page 29.
- Software considerations to obtain the full benefit of this machine are given in 2.3.4, “Software requirements” on page 30.
- Considerations for integrating the p670 into a Cluster 1600 are discussed in 2.3.5, “Cluster considerations” on page 31.

2.3.1 Introduction to the p670 server

The p670 (type 7040-671) is a 4-, 8-, or 16-way SMP machine running the IBM POWER4 microprocessor at 1.1 GHz. The 4-way system has only one processor core enabled per chip, while the 8- and 16-way systems have a dual-core processor. The machine is packed in the same 24-inch rack as the p690 and has the same power supply and uses the same I/O drawer. The central electronic complex (CEC) consists of the MCMs, L3 cache, memory books, and I/O books. Beneath the CEC, an 1U height drawer is located, holding the front panel and a diskette drive. In addition, four media bays, two of them accessible from the back, are integrated for holding CD-ROM, DVD, or tape drives. One additional I/O

drawer with 20 hot plug-enabled PCI slots and up to 16 disks connected to 4 integrated Ultra 2 SCSI controllers is included. Up to two additional I/O drawers can be connected, as well as a battery backup feature. When running in a cluster with an SP Switch or an SP Switch2, 32 servers and 4 LPARs per cluster are supported. 16 LPARs are supported in each p670 when using industry standard interconnect, or if you have 12 LPARs off the switch in an SP Switch2 environment. See Appendix A, “Cluster 1600 scalability rules” on page 177 for more scaling rules. Up to eight p670s can be managed with one HMC. Because this is a HMC protocol system for PSSP, a HMC is required even when using the full machine as an SMP server. A picture of the p670 is shown in Figure 2-8.

Figure 2-8 The p670 with three I/O expansion drawers

2.3.2 CPU board layout

As in the p690, the CPU board design is based on the MCM concept, where four dual-core processors are mounted on one module. Each MCM has a 32 MB L3 cache. Two different options exist for this model, a fully populated 8-way system running at 1.1 GHz and a 4-way system running at 1.1 GHz, where only four processor cores are enabled. This provides double the size of the L2 and L3 cache per processor and increases the throughput per processor. The 16-way

p670 uses two MCMs of eight processor cores each. Figure 2-9 provides a high-level overview of the p670 MCM. All processors in the MCM have access to the Level 3 cache and the memory through a fast interconnect switch technology.

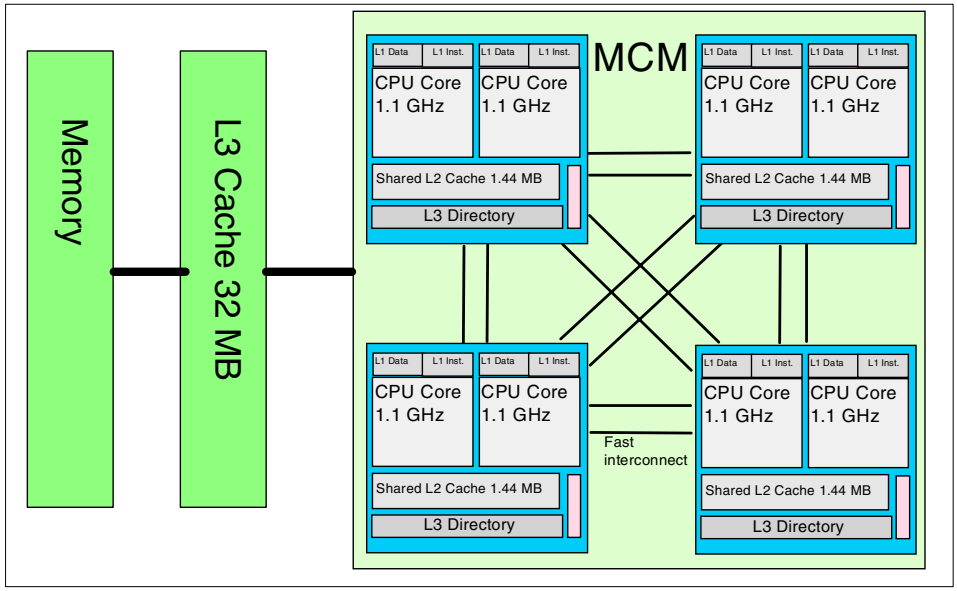


Figure 2-9 Simplified layout of an MCM as installed in a p670 (8-way)

2.3.3 System board design

One or two MCMs, as described in 2.1.2, “CPU board layout” on page 16, can be used in the p670. They are screwed on a backplane on the back of the p670. Both MCMs are connected through a fast switch fabric, allowing each MCM the use of the L3 cache and the memory of the other. Similar to the fast interconnect between the MCMs and their processors, both MCMs connect to the GX bus through the GX slots. On the other side of the bus, up to two I/O books are connected, providing access to the I/O subcomponents. I/O book #1 is mandatory. It includes the serial, HMC and diskette drive connections, as well as the service processor and four remote I/O (RIO) ports. These RIO ports connect to the remote I/O drawers, of which three can exist in a p670. Each drawer needs two connections to an I/O book for redundancy. This drawer is the same drawer utilized for the p690 and the p655. For more information, refer to the *IBM eServer pSeries 690 System Handbook*, SG24-7040. A simplified diagram of this interconnect is shown in Figure 2-10 on page 30.

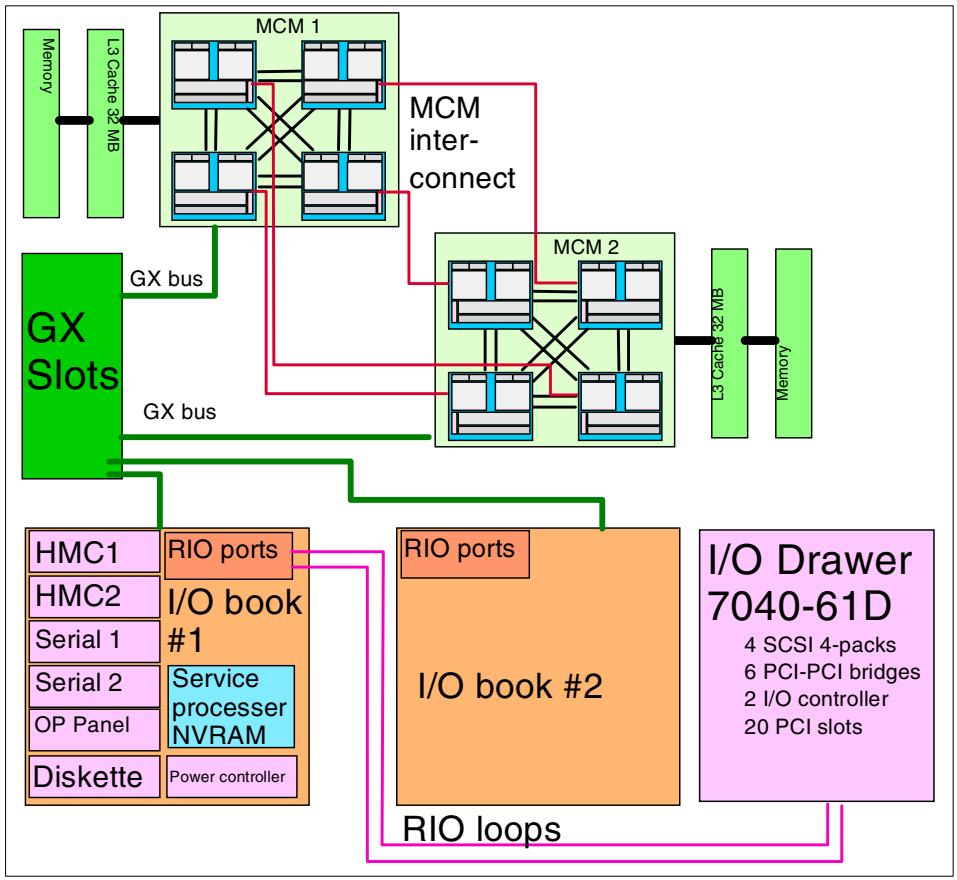


Figure 2-10 Simplified diagram of the p670 interconnection

2.3.4 Software requirements

AIX 5L Version 5.1 with Maintenance Level 2 is required to use all the features and functionality of this server. When using PSSP 3.4, APAR IY30345 for switchless environments, APAR IY29560 for environments with an SP Switch, APAR IY30343 for environments with a single plane SP Switch2, or APAR IY30344 for environments with a dual plane, SP Switch2 is required. If you use PSSP 3.5, no additional APAR is required.

Important: Only PSSP 3.5 supports the 64-bit kernel of AIX 5L Version 5.1.

2.3.5 Cluster considerations

The p670 can be integrated into a Cluster 1600 with the following configurations:

- A switchless cluster with the p670 as a single system.
- A switchless cluster with the p670 with 2-16 LPARs.
- A cluster with an SP Switch, where the p670 is on the switch using one SP Switch Attachment Adapter (FC 8396).
- A cluster with an SP Switch, where the p670 is on the switch using an SP Switch Attachment Adapter (FC 8396) for each partition. Two LPARs per RIO drawer are possible.
- A cluster with a single or double plane SP Switch2, where the p670 is off the switch with the p670 as a single system.
- A cluster with a single or a double plane SP Switch2, where the p670 is off the switch with 1-16 LPARs.
- A cluster with a single plane SP Switch2, where the p670 is on the switch using one SP Switch2 PCI Attachment Adapter (FC 8397) in slot U1.9-P1-I3, U1.9-P1-5, U1.9-P2-I3, or U1.9-P2-I5[1] and has no LPARs.
- A cluster with a single plane SP Switch2, where the p670 is on the switch using one SP Switch2 PCI Attachment Adapter (FC 8397) in one of its PCI slots in all connected LPARs on the switch, while some other LPARs are off the switch.
- A cluster with a single plane SP Switch2, where the p670 is on the switch using one SP Switch2 PCI Attachment Adapter (FC 8397) in one of its PCI slots, one for each LPAR, maximum of 4 LPARs per RIO.
- A cluster with a dual plane SP Switch2, where the p670 is on the switch using two SP Switch2 PCI Attachment Adapters (FC 8397) in its PCI slots as a single system.
- A cluster with a dual plane SP Switch2, where the p670 is on the switch using two SP Switch2 PCI Attachment Adapters (FC 8397) for each LPAR, while some LPARs are off the switch.
- A cluster with a dual plane SP Switch2, where the p670 is on the switch with all LPARs by using two SP Switch2 PCI Attachment Adapters (FC 8397) on each LPAR, allowing two LPARs per RIO drawer.

[1] Assuming only one RIO drawer is installed. Otherwise, the same slots in any other drawer can be utilized.

2.4 The p650 server

This section describes the functionality and the features of the IBM pSeries p650 server (type 7038-6M2). The following topics are discussed in this section:

- An overview is given in 2.4.1, “Introduction to the p650 server” on page 32.
- The layout of the CPU board is described in 2.4.2, “CPU board layout” on page 33.
- The design of the system board is outlined in 2.4.3, “System board design” on page 34.
- Software considerations to obtain the full benefit of this machine are given in 2.4.4, “Software requirements” on page 35.
- Considerations for integrating the p650 into a Cluster 1600 are discussed in 2.4.5, “Cluster considerations” on page 35.

Restriction: Please note, the p650 is not currently supported in a Cluster 1600. IBM intends to support it in the first half of 2003, as described in the “Statement of Direction,” but this is subject to change without notice.

2.4.1 Introduction to the p650 server

The IBM pSeries p650 is a 2-, 4-, 6-, or 8-way SMP machine running the IBM POWER4 microprocessor at 1.2 GHz or 1.45 GHz. It can be equipped with up to 64 GB of memory. The p650 is a rack server, suitable for installation, for example, in a T42 rack. The 8U height allows up to five machines in a single rack. The p650 includes seven PCI-X slots, six running at 133 MHz bus speed with a 64-bit bus design, and the remaining slot is a 5V 32-bit slot at 133 MHz for compatibility. Legacy 32-bit adapters are also supported. This is the first pSeries server where the implementation of the POWER4-II chip is manufactured with higher density and more registers. One integrated 10/100 Mbps Ethernet port and four serial ports, as well as two HMC connections are already included. For enhanced RAS features, a rack status beacon port is integrated. An integrated Ultra 320 SCSI Controller and four bays for hot-swapable disk drives are included, allowing internal storage of up to 587.2 GB. Furthermore, two auto-docking media bays (DVD-ROM, DVD-RAM, CD-ROM, or tape drives) can be included and are attached to the same SCSI bus. For usage as a full SMP server, no HMC needs to be attached. If an HMC is connected, up to eight LPARs are possible with a fully populated machine, four with a minimumly configured system.

Attention: The number of LPARs you can create depends on the number of different boot devices, CPUs, and memory.

For further expansion up to eight 4U height 7311-D10 I/O drawers with six additional hot plug PCI-X bus slots can be added to increase the number of PCI-X slots to 55.

2.4.2 CPU board layout

The POWER4 microprocessor in this machine is packed on CPU cards in a SCM. All other pSeries (except the p630) have so called multi-chip modules (MCM), where four POWER4 processors are packed on one module. One SCM is mounted on a CPU board, in addition to 8 MB of Level 3 cache with the 1.2 GHz processor or 32 MB of Level 3 cache with the 1.45 GHz processor. Up to 16 GB double data rate (DDR) memory can be installed per card. The theoretical bandwidth to the local memory on the card is 6.4 GB/s, using four channels. Four processor cards can be included in the p650, where all cards must have the same frequency and the same amount of memory for optimal parallel work balancing. A simplified illustration of one processor card is shown in Figure 2-11.

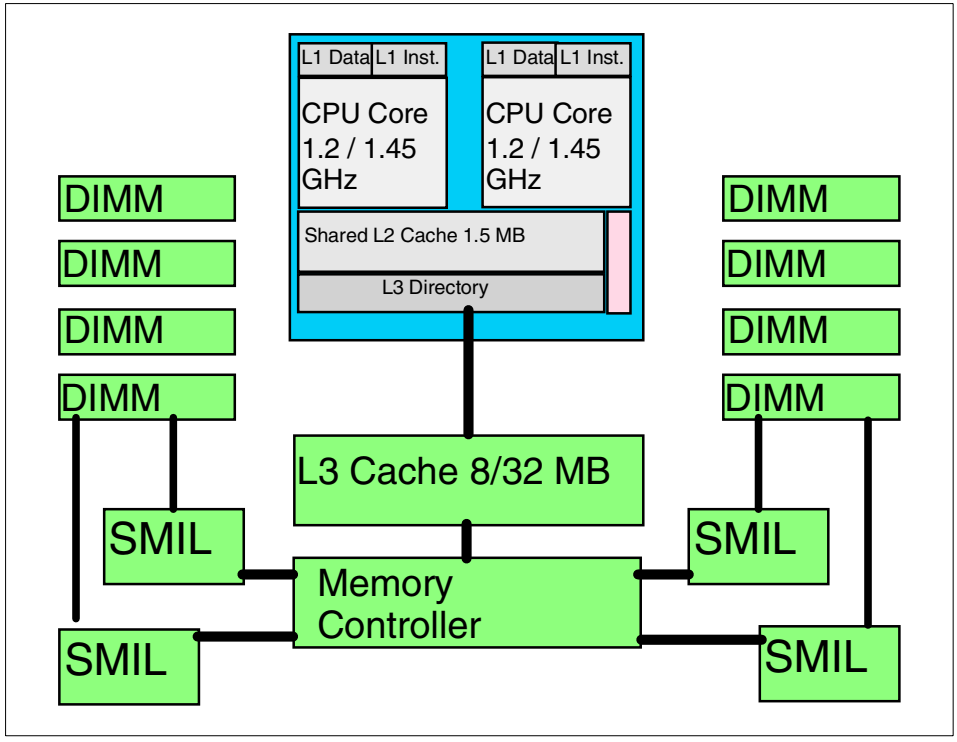


Figure 2-11 Simplified layout of a p650 processor card

2.4.3 System board design

As described in 2.4.1, “Introduction to the p650 server” on page 32, one to four CPU boards can be installed in the machine. The four CPU boards are connected through a fabric interconnect providing two 64-bit pathways at 500 MHz, allowing a peak of 8 GB/s shared between the cards. All cards can use the memory on its own card and on the other. The connection to the system board is accomplished using the GX bus operating at 32 bit and at 333.3 MHz. The first processor cards connect to the mandatory first I/O hub, a second I/O hub for more attached subsystems is optional. Through a remote I/O bridge, the system is connected to the I/O subsystem. Different PCI-X to PCI-X bridges are included, each of them having PCI-X slots, and one has the integrated 10/100 Mbps Ethernet controller, as well as the Ultra 320 SCSI controller. All of the internal disk drives and media devices share the same SCSI bus and thus must be assigned to the same partition. Through a PCI bridge and an ISA bridge, all the I/O ports, such as keyboard and mouse, are connected. With an external RIO connector, up to eight 7133-D10 drawers can be attached. The placement of different PCI and PCI-X cards for optimal performance is complex, see *RS/6000 and eServer pSeries: PCI Adapter Placement Reference*, SA38-0538 for details. Figure 2-12 on page 35 shows a diagram of the data flow within a p650.

Important: When running LPAR with AIX 5L Version 5.1, fully populated processor cards cannot mix memory DIMMs with different capacities on the same card.

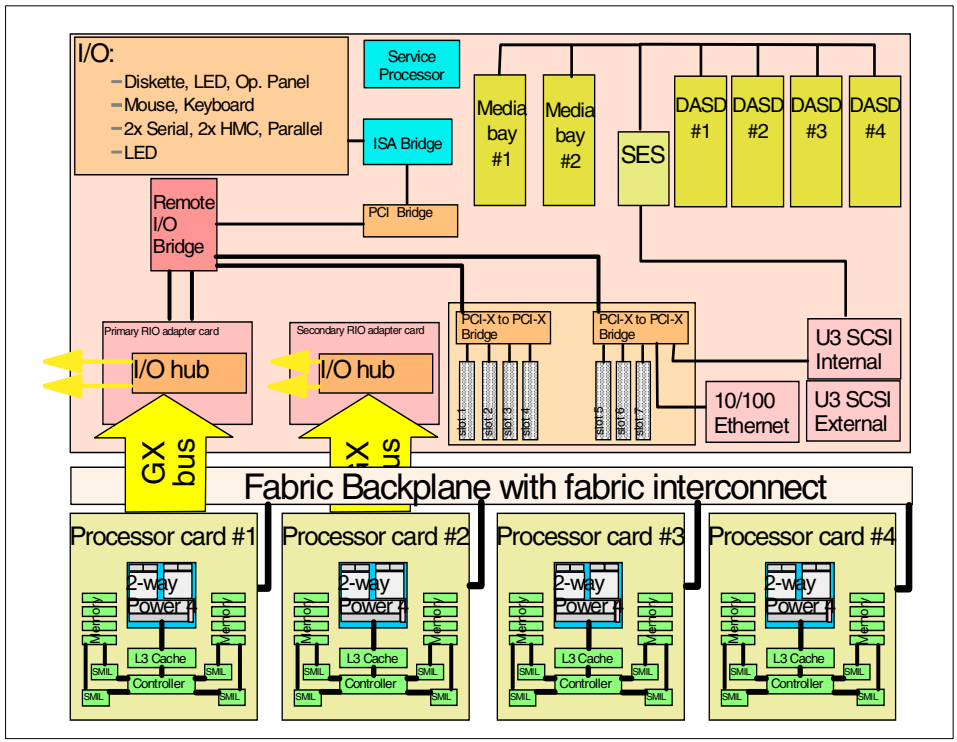


Figure 2-12 Data flow chart of the p650

2.4.4 Software requirements

AIX 5L Version 5.1 with Maintenance Level 3 is needed for the installation of the base operating system.

2.4.5 Cluster considerations

At the time of the publication of this redbook, integration of the p650 into a Cluster 1600 is not supported. IBM intends to provide integration of the p650 into the Cluster 1600 in 2003.

2.5 450 MHz POWER3 SMP thin and wide nodes

This section describes the functionality and the features of the new processor feature for the Winterhawk-II thin and wide nodes. The following concepts are discussed:

- An overview is given in 2.5.1, “Introduction to the 450 MHz SP nodes” on page 36.
- The layout of the CPU board is described in 2.5.2, “CPU board layout” on page 36.
- The design of the system board is outlined in 2.5.3, “System board design” on page 37.
- Software considerations to obtain the full benefit of this machine are given in 2.5.4, “Software requirements” on page 38.
- Considerations for integrating the 450 MHz POWER3 SMP thin and wide nodes into a Cluster 1600 are discussed in 2.5.5, “Cluster considerations” on page 39.

2.5.1 Introduction to the 450 MHz SP nodes

The new processor option for the SP thin and wide nodes allow customers either to upgrade their existing 375 MHz nodes to more processor speed, or to integrate more powerful nodes in their existing SP frames. The new option is supported with PSSP 3.2, 3.4, and 3.5, allowing a wide range of installed software to be run on this node. It contains the latest POWER3-II processor technology with the benefit of the huge 8 MB L2 cache and 20% more CPU speed than before. The Winterhawk-II thin node has an MX slot for attachment to a SP Switch or SP Switch2 fabric, as well as an integrated 10/100 Mbps Ethernet adapter. Two additional PCI slots and two internal disk drives are possible. The wide node extends the use of PCI slots to eight more slots running at 64 bit instead of 32 bit and two more disk drives.

2.5.2 CPU board layout

Two different boards can be plugged into this node, allowing combinations of two or four processors. Each board has up to two POWER3-II processors running on 450 MHz that are connected through the processor-specific 6xx system bus. Each processor has 32 KB of data and 64 KB of instruction cache and its own Level 2 cache with 8 MB. An outline of the CPU board is illustrated in Figure 2-13 on page 37.

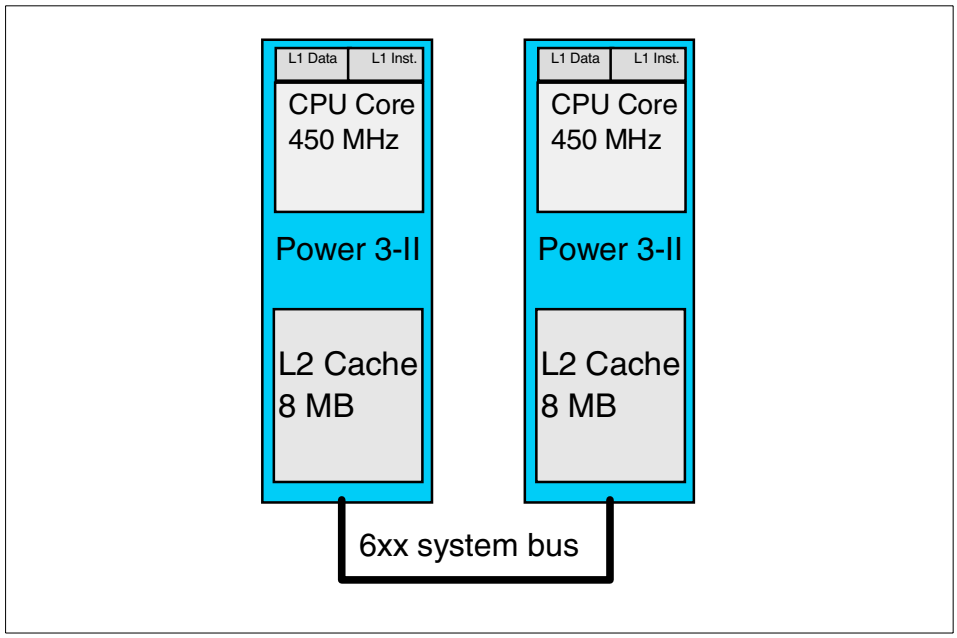


Figure 2-13 A two processor card in a Winterhawk-II SP node

2.5.3 System board design

Up to two processor cards are connected through the 6xx system bus to a combined I/O and memory controller on the system board. The maximum amount that can be installed on the system planar is 16 GB RAM. Through the I/O controller, all other components are attached using the fast 6xx MX bus. Directly attached to this bus is the interconnect to the SP Switch MX port, also an internal peripheral component interconnect (PCI) controller running a bus system at 33 MHz with a 32-bit bus system attached. To this controller, an ISA-bridge for I/O functionality and an U-SCSI controller are also connected. This controller allows two direct access storage devices (DASDs) to be included in the system. A wide node has two additional PCI controllers, each controlling four PCI slots running 64-bit slots at 33 MHz. A simplified outline is given in Figure 2-14 on page 38.

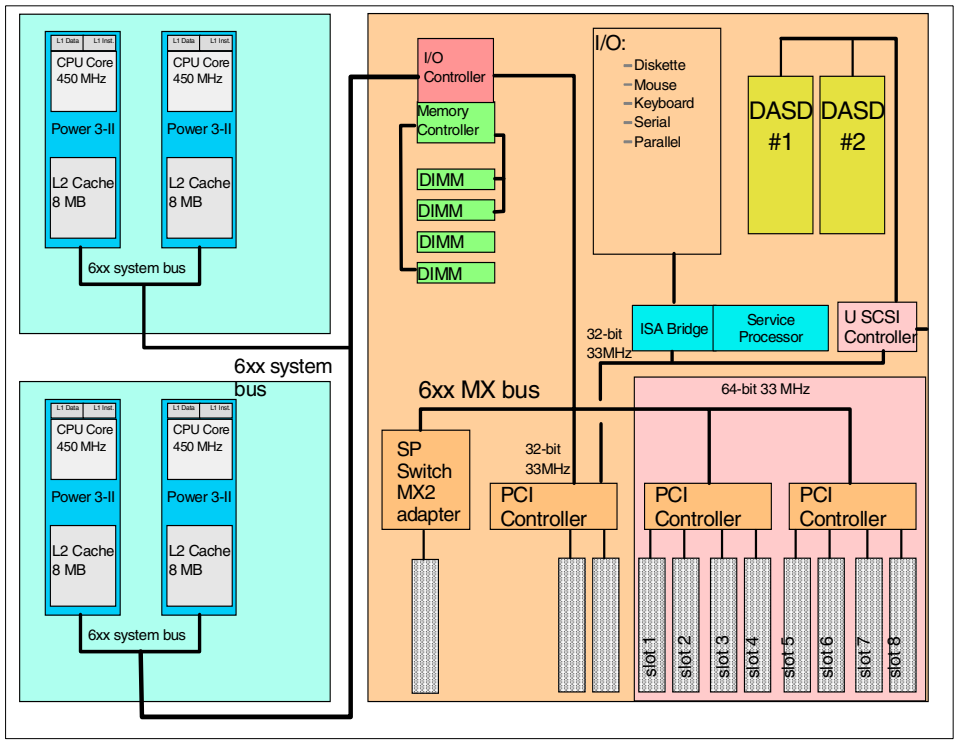


Figure 2-14 System layout of an SP Winterhawk-II wide node

2.5.4 Software requirements

To support this type of node, AIX 4.3.3 or AIX 5L Version 5.1 is required. For PSSP, Version 3.5, 3.4 with APAR IY31115, or PSSP 3.2 with APAR IY28091 are required. In PSSP, the node will be recognized as a 375/450 MHz system. Example 2-5 on page 39 shows the output of the **`splstdata`** command.

Important: Only PSSP 3.5 supports the 64-bit kernel of AIX 5L Version 5.1.

Example 2-5 450 MHz node as seen by PSSP 3.5

```
[c166s][/]> splstdata -n -l 48
                  List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- --------------- --------- --------------- ----------
   48      3    16     1 c177n16.ppd.pok.i c177n16.ppd.pok.i ""                9.114.72.125
MP                                2 375/450_MHz_POW         1 true            ""
```

2.5.5 Cluster considerations

This node can be integrated into a Cluster 1600 with the following options:

- A switchless cluster with the node
- A cluster with a single or a double plane SP Switch2, where the node is off the switch
- A cluster with a single plane SP Switch, where the node is on the switch using one SP Switch MX Attachment Adapter (FC 4023) in the MX slot
- A cluster with a single plane SP Switch2, where the node is on the switch using one SP Switch2 MX Attachment Adapter (FC 4026) in the MX slot

2.6 Overview of new pSeries servers

This section gives a brief overview of the main features of the new pSeries servers introduced in this chapter. Table 2-2 is a quick comparison chart for the new server models.

Table 2-2 Comparison chart of all introduced systems

	Winterhawk II	**p630**	**p655**	**p670**	**p650**
Processor type	POWER 3-II	POWER 4	POWER 4	POWER 4	POWER 4-II
Processor speed	450 MHz	1.0 GHz	1.1/1.3 GHz	1.1 GHz	1.2/1.45 GHz
No. of processors	2, 4	1, 2, 4	4, 8[a]	4, 8, 16	2, 4, 6, 8
Caches L1 (D/I) L2 L3	32 k, 64 k 8 MB N/A	32 k, 64 k 0.72-1.44 MB 16-32 MB	32 k, 64 k 0.72-1.44 MB 16-32 MB	32 k, 64 k 0.72-1.44 MB 16-32 MB	32 k, 64 k 0.72-1.44 MB 4-16 MB

	Winterhawk II	p630	p655	p670	p650
Maximum processors per frame	32-64	40	64-128	16	48
Maximum memory	16 GB	32 GB	32 GB	128 GB	64 GB
Integrated Ethernet	1x 10/100	2x 10/100	2x 10/100	No	1x 10/100
Can use any enX as management Ethernet	No	Yes	Yes	Yes	N/A
Adapters integrated per node	2, 10 PCI	4 PCI-X	3 PCI-X	20-80 PCI	7 PCI-X
Needs HMC for cluster integration	No	Yes	Yes	Yes	N/A
Manageable with PSSP	Yes, PSSP 3.2 or later	Yes, PSSP 3.4 or later	Yes, PSSP 3.4 or later	Yes, PSSP 3.4 or later	Statement of direction for PSSP 3.5
PSSP APARs for 3.4	IY31115	IY24792	IY34495	IY29560 IY30343 IY39344 IY30345	N/A
PSSP APARs for 3.5	N/A	N/A	IY34496	N/A	N/A
Required OS	AIX 4.3.3	AIX 5L V5.1	AIX 5L V5.1	AIX 5L V5.1	AIX 5L V5.1
Required maintenance level	None[b]	2	2	2	3
Support for SP Switch	Yes, MX Adapter FC4023	No	No	Yes, PCI Adapter FC8396	N/A
Support for SP Switch2	Yes, MX Adapter FC4026	Yes, PCI Adapter FC8397	Yes, PCI-X Adapter FC8398	Yes, PCI Adapter FC8397	Statement of direction for PSSP 3.5

	Winterhawk II	p630	p655	p670	p650
Support for SP Switch2 dual plane	No	No	Yes, PCI-X Adapter FC8398[c]	Yes, PCI Adapter FC8397	N/A

a. Four if running on 1.3 GHz, eight if running on 1.1 GHz.
b. We recommend the use of ML10 on AIX 4.3.3 and at least ML02 on AIX 5L Version 5.1.
c. If running two LPARs, only one can be on the switch.

2.7 SP Switch2 PCI-X Attachment Adapter (FC 8398)

To take advantage of the speed of the PCI-X bus design of the newer pSeries server, the SP Switch2 PCI Attachment Adapter (FC 8397) was modified to benefit from the new design. The new SP Switch2 PCI-X Attachment Adapter (FC 8398) complies to the PCI-X standard. For a brief comparison of today's bus speed, Table 2-3 shows the maximum, not to exceed the theoretical bandwidth of some PCI bus designs.

Table 2-3 PCI technology overview: Peak performance

Bus type	Theoretical bandwidth	pSeries model
PCI 32 bit at 33 MHz	133 MB/s	7046-B50
PCI 32 bit at 66 MHz	266 MB/s	N/A
PCI 64 bit at 3 3 MHz	266 MB/s	7026-H10
PCI 64 bit at 50 MHz	403 MB/s	7026-H70
PCI 64 bit at 66 MHz	532 MB/s	7040-671 (p670)
PCI-X 64 bit at 133 MHz	1.06 GB/s	7026-6C4 (p630)

This adapter is currently supported for installation in the new p655 server only. It is not supported in the p630 (the p630 uses SP Switch2 PCI Attachment Adapter), although this machine has a PCI-X bus system. This adapter removes the requirement for an empty adapter slot to its right side that the SP Switch2 PCI Attachment Adapter has, because the PCI-X adapter has smaller heatsinks and an optimized layout of its sandwich card. APAR IY34151 for PSSP 3.4 needs to be installed to support this new adapter. In PSSP 3.5, the support is already included. Figure 2-15 on page 42 shows a diagram of the new adapter.

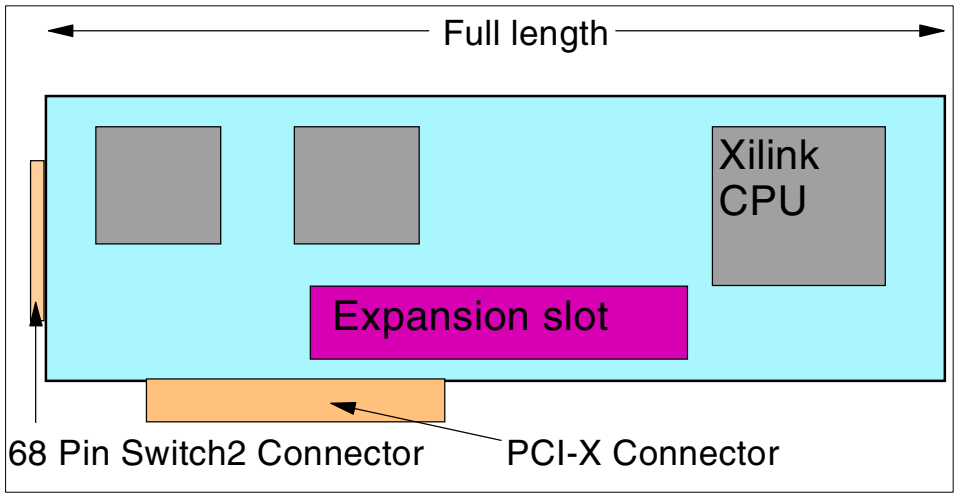


Figure 2-15 Layout of the SP Switch2 PCI-X Attachment Adapter (FC 8398)

2.8 19-inch switch frame 9076-558

The 19-inch frame provides storage for up to two SP Switch2 switches and the corresponding power supplies. N+1 redundancy can be achieved for the power control. The switches are mounted vertically. Customers can have DASD drawers or whatever else is supported in T00 or T42 frames (servers supported in the 7014 rack are allowed). The package consists of a separately ordered IBM 7014-T00 or 7014-T42 frame, one Scalable Electrical Power Base Unit (SEPBU), dual power cord and all the mounting parts, and one or two separately ordered SP Switch2 switches. Figure 2-16 on page 43 shows the layout of this solution.

Important: This model, 9076-558, is just the SEPBU, brackets, power PDU cables, and one or two switches. The frame is not part of the 9076-558. In fact, 9076-558 can be added to an existing 19-inch frame although it does have to be added to the bottom of the frame.

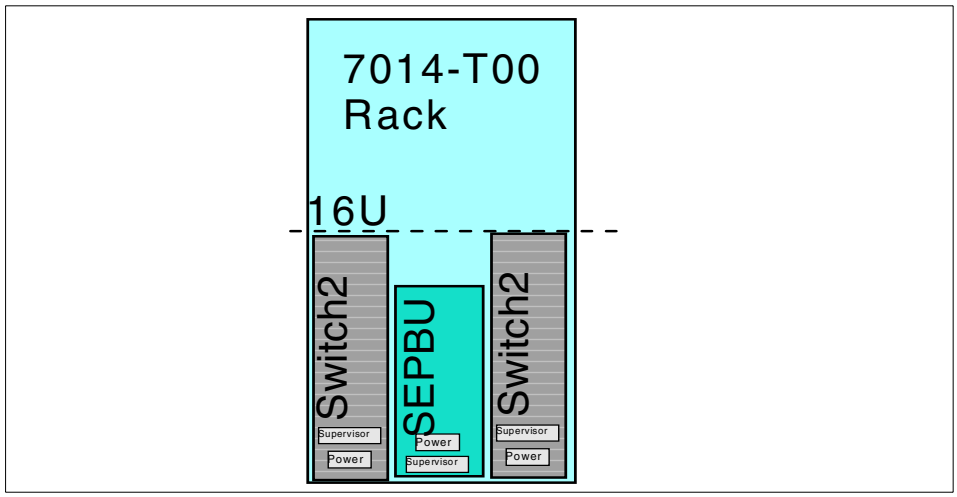


Figure 2-16 The 9078-558 in a 7014-T00 rack

2.9 24-inch 7040-W42 frame

This new frame is a 24-inch wide 42U height frame, designed to hold up to 16 p655 thin nodes or additional 7040-61D remote I/O drawers. It includes a power supply located in the top of the frame using 8U. It uses a 60A or 100A line cord. Internally, a 350V DC bulk power is provided. This power supply is the same as the one used in the p670 and p690 systems. An optional one or two battery packs can be included. The size of the frame is 78.5 cm x 144 cm x 202.5 cm (width, depth, height). A fully populated frame can weigh up to 1584 kg (3484 lbs). For up to nine p655 and 2 7040-61D, redundant bulk power is provided. Otherwise, the bulk power assemblies are tied together allowing N+1 redundancy. It is not supported to install the SP Switch or the SP Switch2 into the same rack. Either a sculptured black front door with copper accent, together with a slim-line rear door, or a sculptured black front acoustic door, together with the acoustic rear door, is available. Figure 2-17 on page 44 shows the frame.

Figure 2-17 The 7040-W42 frame

Tip: When enough space is available, we recommend the use of the acoustic front door with the acoustic rear door, because it provides better air cooling.

The rack provides two special RS-422 ports per base power controller dedicated for HMC attachment.

Note: IBM recommends the use of the integrated battery backup features or an UPS to get benefits from the EPOW capability of the p655.

2.10 New Hardware Management Console

The new IBM Hardware Management Console (HMC) for pSeries (type 7315-C01) provides a set of functions that are necessary to manage the pSeries p655 and LPAR configurations. Up to 16 p655s and 32 LPARs can be controlled. The new HMC supports redundant HMC functionality only on a manual basis, where two HMCs can be connected to each server. PSSP only communicates to one HMC at a time. However, if this HMC fails, you can switch the communication to the second one manually. The new HMC has larger memory and storage capacity than its predecessors. Additionally, a new recovery software through DVD is available that allows continued operation and management of a pSeries system in case the HMC system software requires replacement or recovery. The HMC now comes with the third release of HMC software.

2.11 New control workstation

The p630 (type 7028-6C4) and p630 (type 7028-6E4) are now supported as control workstations (CWS). The 6E4 is a stand-alone tower model of the 6C4. The description in 2.1, “The p630 server” on page 15 concerning internal structures is valid for this machine, too. For a current list of supported control workstations, refer to:

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

2.12 7311 Model D10 I/O drawer

For customers who want to expand the capabilities of the p650 described in 2.4, “The p650 server” on page 32, up to eight 7311 Model D10 I/O 4U height half-width drawers can be attached using RIO connections. Two drawers fit side by side in one position in a 19-inch rack, giving them both a 9.5 inch width. A 4U enclosure carries one or two of them. Each drawer extends the p650 with five additional hot plug PCI-X slots with 3.3V and one additional hot plug 64-bit PCI 5V slot for an older adapter. No disks can be included in this drawer. Redundant hot plug for power and cooling is provided. The outline of this drawer is depicted in Figure 2-18.

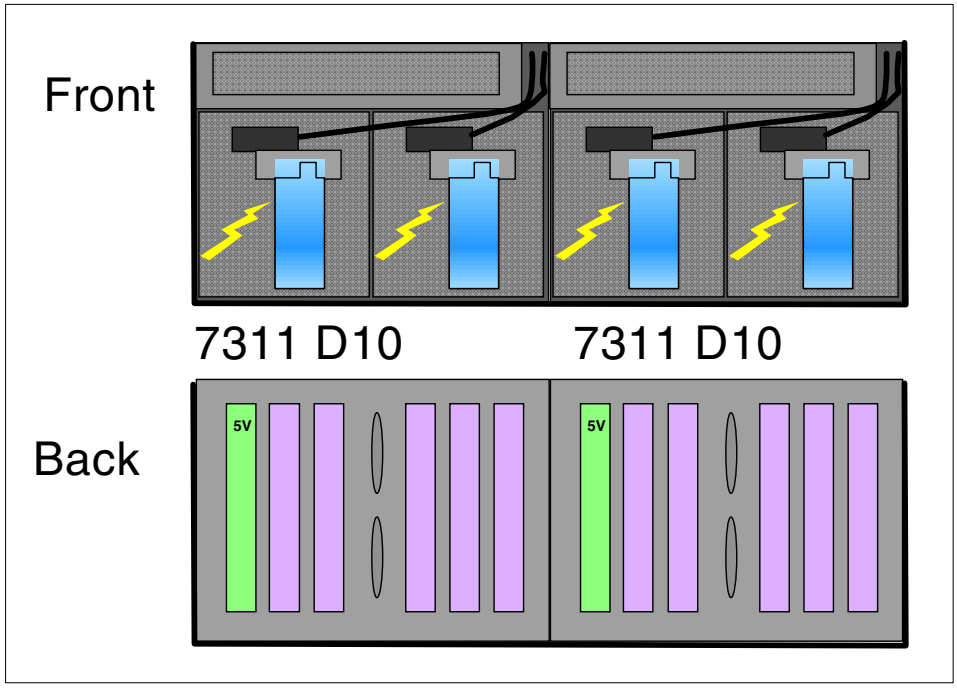


Figure 2-18 Two 7311 D10 I/O drawers side by side

2.13 7311 Model D20 I/O drawer

For customers who want to expand the capabilities of the p630 described in 2.1, “The p630 server” on page 15, up to two 7311 Model D20 I/O 19-inch 4U height drawers can be attached using RIO connections. Each drawer includes seven PCI-X hot pluggable slots and two independent six packs for hot swap SCSI disks. Air cooling and power is redundant. All components can be accessed without tools. A picture of this drawer is shown in Figure 2-19. The p630 FC9575 or FC6675 must be installed, and the December 2002 Update CD of AIX 5L Version 5.1 or later is required.

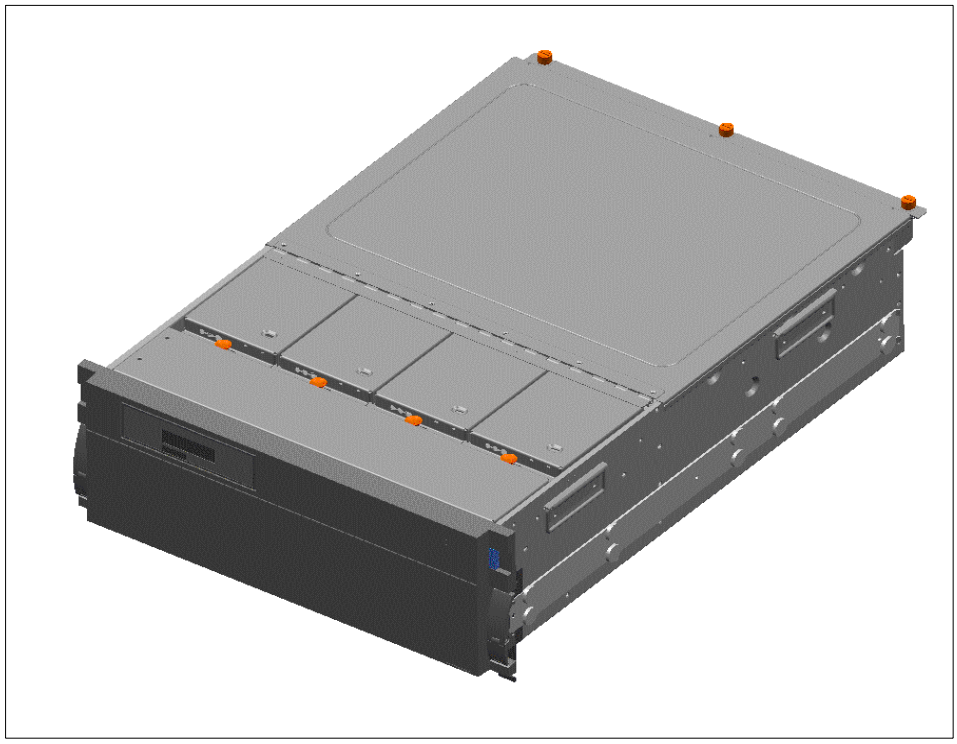


Figure 2-19 7133 Model D20 I/O drawer

Restriction: No SP Switch2 PCI Attachment Adapter (FC 8397) or SP Switch2 PCI-X Attachment Adapter (FC 8398) is supported in the I/O drawer.

3

Reliable Scalable Cluster Technology overview

This chapter provides an overview of Reliable Scalable Cluster Technology (RSCT), its components, and the communication path between these components. We also discuss where it is used and describe a RSCT peer domain.

There are already a number of good IBM manuals, Redbooks, white papers, and Redpapers about RSCT. This chapter focuses on the components found in RSCT and the new components that are incorporated into the RSCT peer domain.

This chapter contains the following sections:

- What is Reliable Scalable Cluster Technology
- Reliable Scalable Cluster Technology components
- Usage of Reliable Scalable Cluster Technology
- RSCT peer domain (RPD)

3.1 What is Reliable Scalable Cluster Technology

Reliable Scalable Cluster Technology (RSCT) is a set of software components that together provide a comprehensive clustering environment for AIX and Linux. RSCT is the infrastructure used by a variety of IBM products to provide clusters with improved system availability, scalability, and ease of use.

3.2 Reliable Scalable Cluster Technology components

This section describes the RSCT components and how they communicate with each other. We focus here on the new design. When RSCT was announced the first time, the structure was slightly different. An overview of these differences is provided in 3.2.2, “Communication between RSCT components” on page 51.

3.2.1 Reliable Scalable Cluster Technology components overview

The main components are as follows. For a more detailed description of the RSCT components, see *IBM RSCT for AIX: Guide and Reference*, SA22-7889.

- Topology Services

  This can be used to provide node and network failure detection.

- Group Services

  This can be used to provide cross-node and process coordination. For a detailed description about how Group Services work and how you can add modifications, see *IBM RSCT: Group Services Programming Guide and Reference*, SA22-7888.

- RSCT cluster security services

  This provides the security infrastructure that enables RSCT components to authenticate the identity of other parties.

- Resource Monitoring and Control (RMC) subsystem

  This is the scalable and reliable backbone of RSCT. It runs on a single machine or on each node (operating system image) of a cluster and provides a common abstraction for the resources of the individual system or the cluster of nodes. In other words, this daemon does run on all nodes of a cluster and all systems using AIX 5L Version 5.1 or later.

You can use RMC for single system monitoring, or for monitoring nodes in a cluster. In a cluster, however, RMC provides global access to subsystems and resources throughout the cluster, thus providing a single monitoring and management infrastructure for clusters. For more information, refer to the redbook *A Practical Guide for Resource Monitoring and Control (RCM)*, SG24-6615.

- RSCT core resource manager

  The resource manager is a software layer between a resource (a hardware or software entity that provides services to some other component) and RMC. A resource manager maps programmatic abstractions in RMC into the actual calls and commands of a resource (status of events and control commands).

3.2.2 Communication between RSCT components

In this section, we give a brief overview of the new and old components and how they communicate with each other. This is due to the fact that both can coexist on the same system. How such an environment may look is described in “Using the old and the new RSCT design on one system” on page 55.

New RSCT design

The Resource Monitoring and Control (RMC) subsystem and RSCT core resource managers (RM) are today the only ones which use the RSCT cluster security services (CtSec). These three components are new. Topology Services and Group Services are still using Remote Procedure Call-based (RPC) communication. This will change so that all components will use CtSec for exchanging data. An advantage of the new design is that it can handle more than one application at a time. For more information about the differences between the new and old designs, see “Comparison of RSCT designs” on page 53.

Group Services is a client of Topology Services, and RMC is a client of Group Services. The RMC application programming interface (API) is the only interface that can be used by applications to exchange data with the RSCT components. RMC manages the RMs and receives data from them. This is done by utilizing the new security service. Figure 3-1 on page 52 shows a brief overview of the RSCT components.

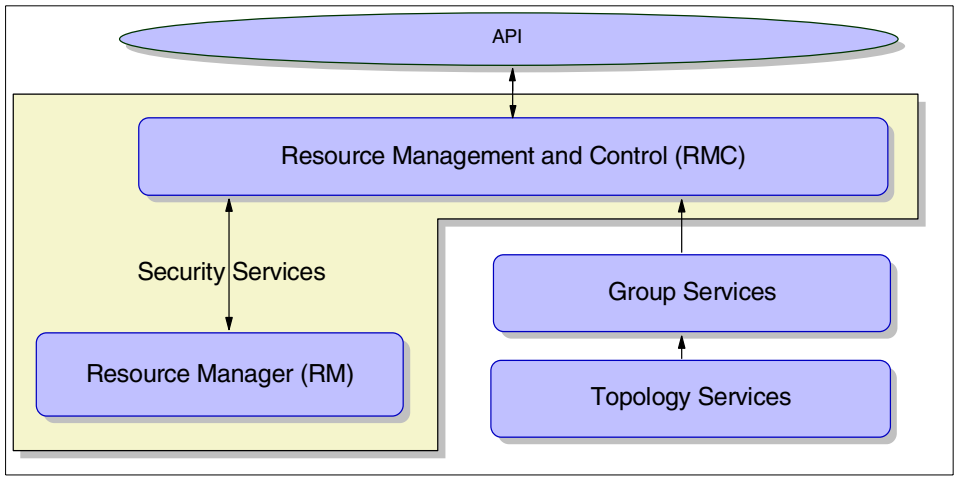


Figure 3-1 Reliable Scalable Cluster Technology components

The RMC can manage more than one RM. It is also able to exchange data with other RMCs using CtSec. An example of possible communications between RMCs, RMs, and applications is shown in Figure 3-2.

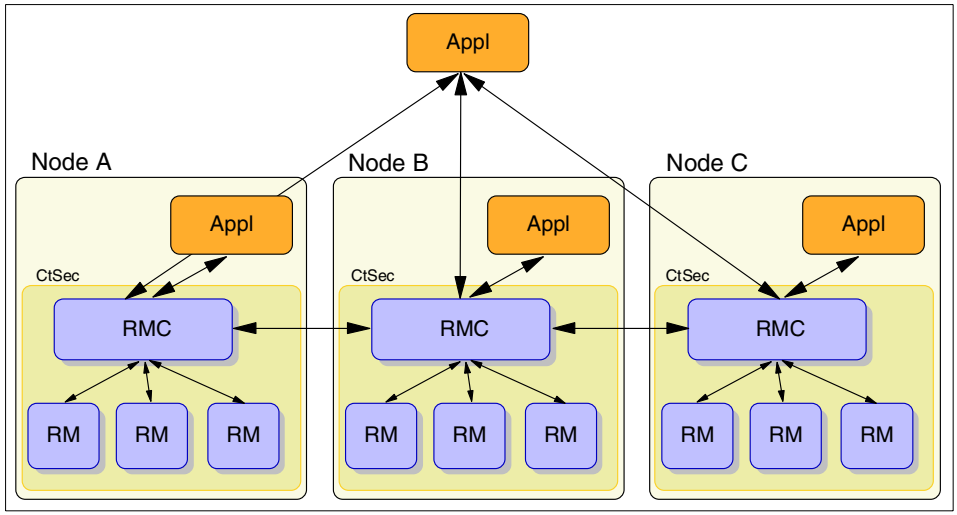


Figure 3-2 Resource Monitoring and Control communication

This new design is used by CSM (see *An Introduction to CSM 1.3 for AIX 5L*, SG24-6859), or GPFS based on RSCT peer domain (see 3.3.3, “General Parallel File System” on page 64 and 5.6, “General Parallel File System on RSCT peer domain” on page 114).

Old RSCT design

When we talk about the old RSCT design, we are referring to RSCT prior to AIX 5L Version 5.1. This is packaged with the products.

In the old design, we had Topology Services, Group Services, Event Management, and Resource Monitors. Each application runs a separate instance of the RSCT daemons. This is illustrated in Figure 3-3.

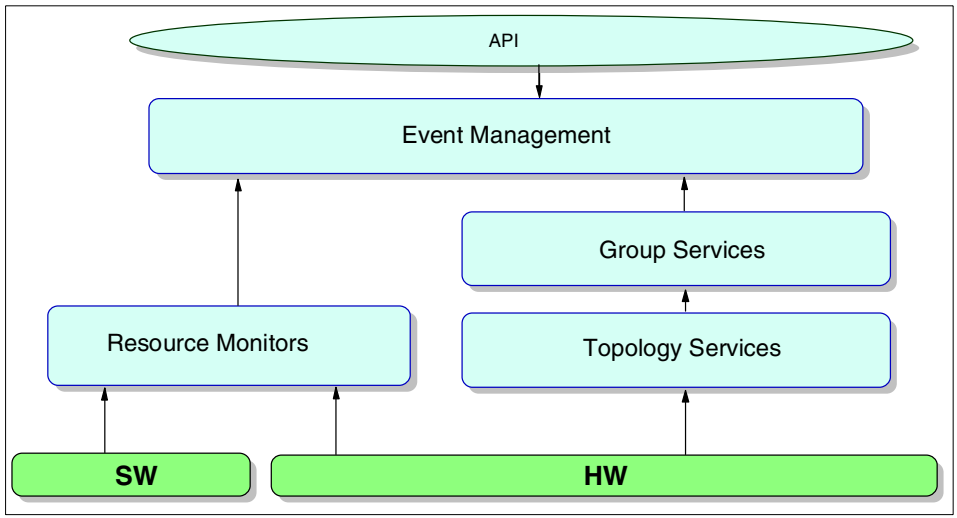


Figure 3-3 Reliable Scalable Cluster Technology components (old)

The RSCT components are as follows:

- Topology Services is basically still the same as today.
- Group Services is also basically still the same as today.
- Event Management is replaced by the Resource Monitoring and Control (RMC) subsystem. The RMC has much more functionality. Resource control is missing in Event Management.
- The Resource Monitors are replaced by RSCT core resource managers (RM). As for RMC, these RMs have much more functionality.

The old design is used by PSSP and HACMP/ES. An example of these two products is in 3.3, “Usage of Reliable Scalable Cluster Technology” on page 61.

Comparison of RSCT designs

As previously discussed, the RSCT design has changed since it was announced the first time. The new design increases the flexibility, scalability, reliability, and functionality. We explain some of the differences in more detail here.

As we have already mentioned, Event Management is replaced by the Resource Monitoring and Control (RMC) subsystem, and the Resource Monitors are replaced by resource managers (RM). The communication between RMC and RMs, and between multiple RMs, runs over a secure layer. As the name indicates, these two daemons do not just detect and communicate events, they can also be used to control resources.

The communication and the security are transparent to the user. Figure 3-4 shows the relationship in the old design and the new design between the daemons and between the nodes.

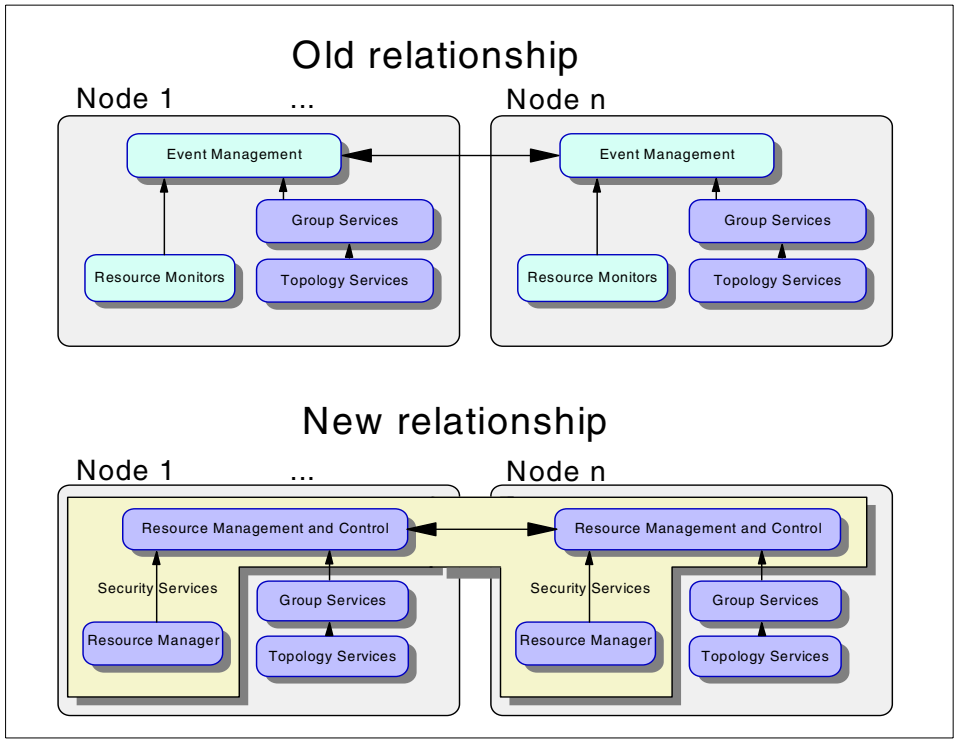


Figure 3-4 Reliable Scalable Cluster Technology communication

In the old design, for each application that is using RSCT, there is a separate set of daemons. In the new design, there is just one daemon of each type for all running applications. The number of daemons that are running depends on your application needs.

If you just use AIX, there is just the ctrmc daemon. If you configure RM, you will find some RM-based daemons as well. Other applications and CSM will start more RM daemons. Topology Services and Group Services daemons are configured and started only in the case of an RPD cluster.

Figure 3-5 compares these two designs from a daemon point of view. Using **`lssrc -a`**, it is possible to check the configured daemons on a machine. Refer to “Using the old and the new RSCT design on one system” on page 55 for information about the additional daemons.

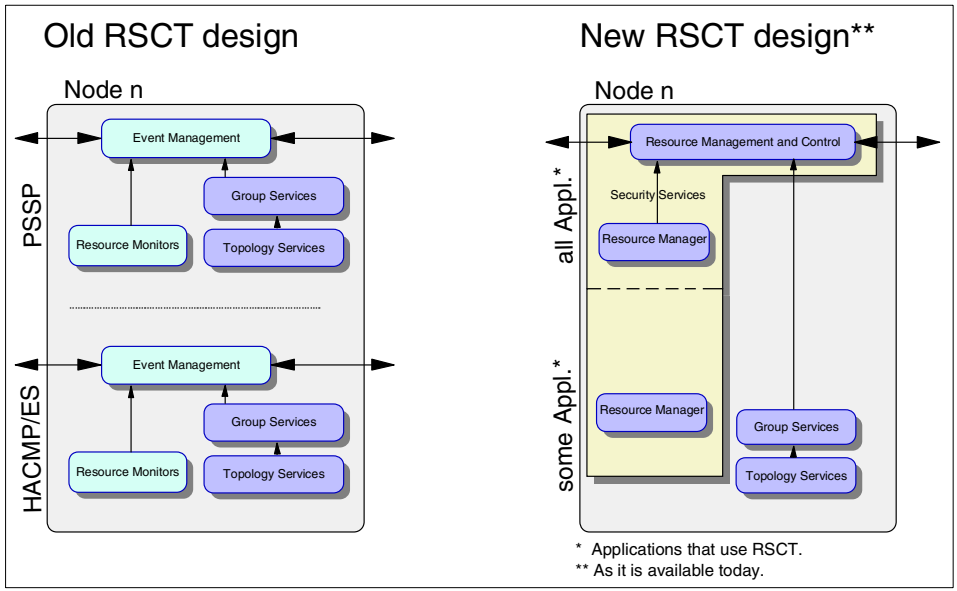


Figure 3-5 Reliable Scalable Cluster Technology daemons

Using the old and the new RSCT design on one system

Today, we might find a combination of the old and the new design running on one system. So far, PSSP and HACMP/ES use the old design. Figure 3-6 on page 56 shows an example of how it might look if you have both designs in use on one system. In this example, the assumption is that you have an application that is using the new design (such as GPFS on RPD), and HACMP/ES or PSSP, or both, on the same system.

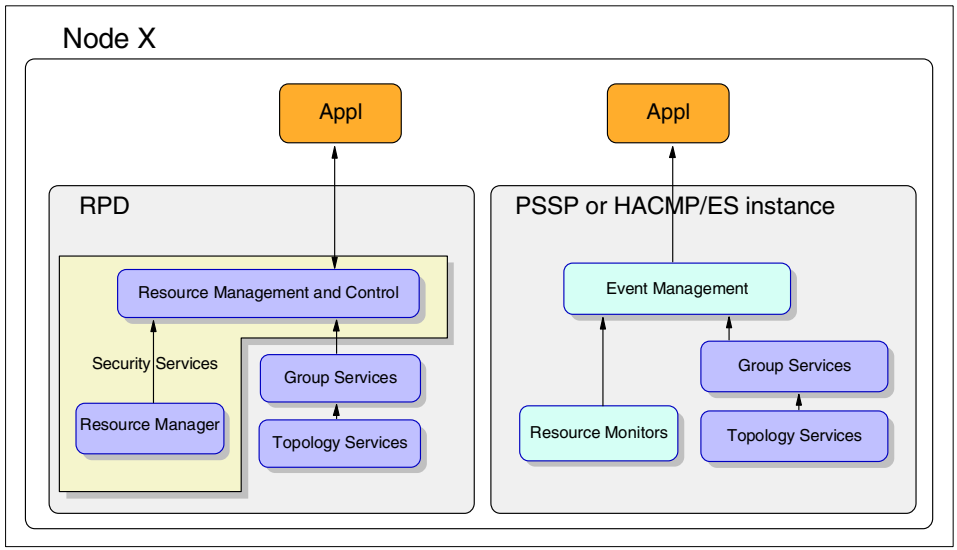


Figure 3-6 Using the old and the new RSCT designs in one system

When you use the `lssrc` command in such an environment, you find the RSCT daemons shown in Table 3-1.

Table 3-1 Reliable Scalable Cluster Technology daemons

Function	PSSP daemon name	HACMP/ES daemon name	RSCT daemon name
Topology Services	hats	topsvcs	cthats[a]
Group Services	hags	grpsvcs	cthags[a]
Group Services Globalized Switch Membership	hagsglsm	grpglsm	cthagsglsm[b]
Event Management	haem	emsvcs	N/A
Event Management AIX Operation System Resource Monitor	haemaixos	emaixos	N/A
Resource Monitoring and Control Subsystem	N/A	N/A	ctrmc
Security Service	N/A	N/A	ctcas

a. For RSCT peer domains (RPD) only.
b. In the RPD environment, this subsystem is not generally used.

Example 3-1 shows what you can get if all the three programs are in use.

Example 3-1 List of Reliable Scalable Cluster Technology daemons

```
# lssrc -a | egrep "rsct |ha|svcs"
 ctrmc           rsct         12126     active
 ctcas           rsct         13418     active
 cthats          cthats       12072     active
 cthags          cthags       14994     active
 topsvcs         topsvcs      15106     active
 grpsvcs         grpsvcs      14058     active
 grpglsm         grpsvcs      16770     active
 emsvcs          emsvcs       17066     active
 emaixos         emsvcs       17562     active
 hats            hats         17852     active
 hags            hags         17996     active
 hagsglsm        hags         18603     active
 haem            haem         18931     active
 haemaixos       haem         19586     active
```

When you use **ps** in such an environment, and you look for instances of hags as a process, you get an output similar to the content of Example 3-2.

Example 3-2 List of Group Services processes

```
# ps -ef | grep hags
   root 14058   3142  0   Oct 08      -  8:52 hagsd grpsvcs
   root 16770   3142  0   Oct 08      -  0:48 hagsglsmd grpglsm
   root 14994   3142  0 14:12:33      -  0:00 hagsd cthags
   root 17996   3142  0   Oct 08      -  0:06 hagsd hags
   root 18603   3142  0   Oct 08      -  0:00 hagsglsmd hagsglsm
   root 28772   8342  0 16:30:33  pts/0  0:00 grep hags
```

3.2.3 Reliable Scalable Cluster Technology relationships

There are three types of RSCT relationships:

- Stand-alone
- Management domain
- Peer domain

The following sections describe these relationships and the combination of a managed domain and a peer domain.

Stand-alone

The following lists what is installed with the base AIX 5L:

- Resource Monitoring and Control (RMC) subsystem
- RSCT core resource managers (RM)
- RSCT cluster security services (CtSec).

The RMs are only started when needed. Figure 3-7 shows what you can have in such an environment.

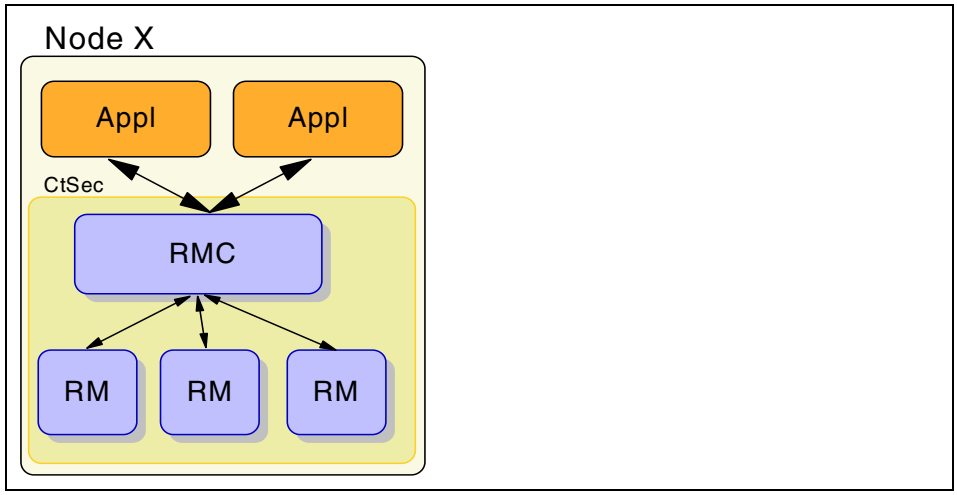


Figure 3-7 Resource Monitoring and Control stand-alone

There are many things we can monitor with a stand-alone RSCT. One example is to monitor the file system usage on a stand-alone IBM @server pSeries or RS/6000 system (running AIX 5L Version 5.1 with Maintenance Level 3).

Management domain

The management domain has at least one management server and one or more managed nodes. Figure 3-8 on page 59 shows an example of what this environment can look like.

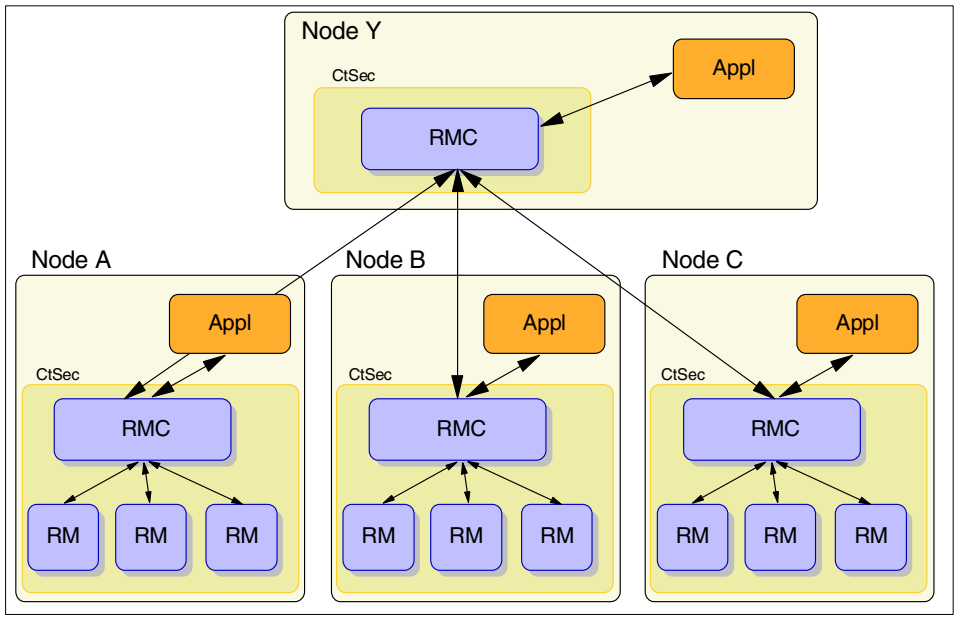


Figure 3-8 Reliable Scalable Cluster Technology management domain

From a design point of view, in such an environment, there is no need to have the same operating system on all the managed nodes. Cluster Systems Management (CSM) is one implementation of this architecture. Each of the managed nodes is autonomous and only the management server knows about the existence of all of the other nodes.

Important: To implement the RSCT managed domain, you need AIX 5L Version 5.1 with Maintenance Level 3, AIX 5L Version 5.2, or Linux installed.

Peer domain

In a RSCT peer domain (RPD), you do not have a management server or a managed node. All nodes are equal. You can use any node to manage the whole domain (one at a time). Figure 3-9 on page 60 shows what such an environment can look like. For more details about RPD, see 3.4, “RSCT peer domain (RPD)” on page 66.

This implementation is used by GPFS. For more information about GPFS on a RSCT peer domain (RPD), see 5.6, “General Parallel File System on RSCT peer domain” on page 114.

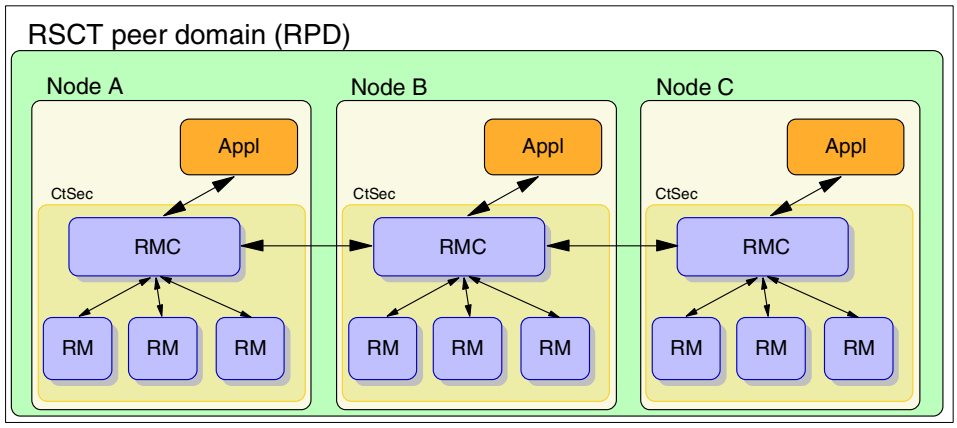


Figure 3-9 Reliable Scalable Cluster Technology peer domain (RPD)

This implementation also makes use of Topology Services and Group Services. This means that you have all nodes and all the RSCT components in use, as shown in Figure 3-1 on page 52.

3.2.4 Combination of Reliable Scalable Cluster Technology domains

You can have a combination of all three types of domains (stand-alone, management domain, and RPD) together with the old RSCT structure as used by PSSP and HACMP/ES.

In this section, we focus on one implementation only: the combination of a RSCT peer domain and a RSCT managed domain. An example of such a combination might be using CSM with RPD-based GPFS on some nodes.

Node Y is a RSCT management server. You have three nodes as managed nodes (Node A, Node B, and Node C). Node B and Node C are also using GPFS on RPD. Therefore, you are required to configure a peer domain for these two nodes, as shown in Figure 3-10 on page 61.

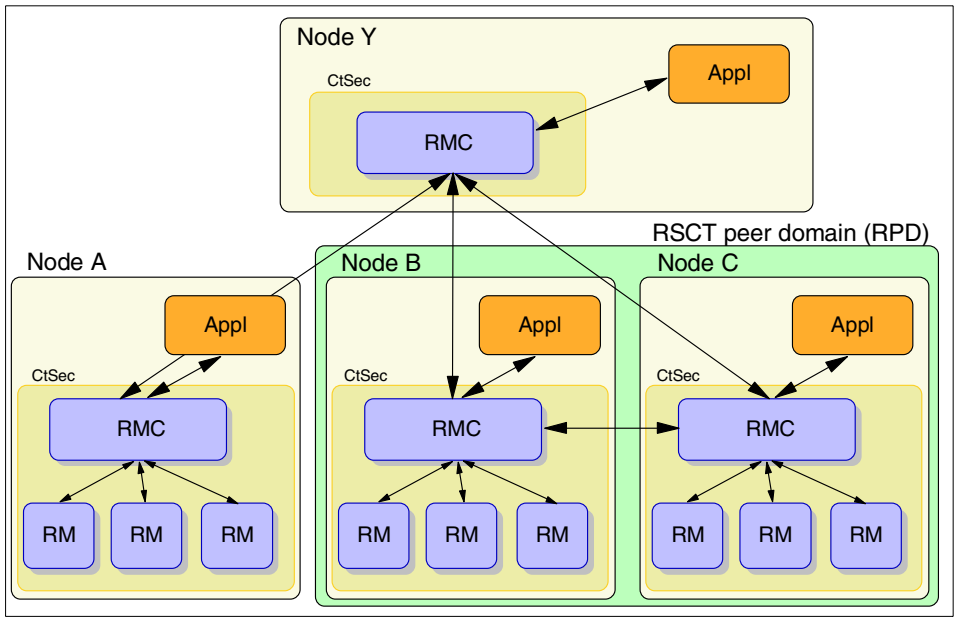


Figure 3-10 Combination of a peer domain and a management domain

3.3 Usage of Reliable Scalable Cluster Technology

Here, we focus on three products for AIX that use RSCT:

- PSSP
- HACMP/ES
- GPFS

RSCT is also available for Linux, and it is used by some other AIX-based products, such as Workload Manager. The number of these products will increase in the future.

3.3.1 Parallel System Support Program

Parallel System Support Program (PSSP) Version 2.2 was the first product that used the High Availability Infrastructure (HAI). HAI was renamed RSCT and packaged differently in PSSP Version 3.1 and later.

In PSSP, there is an Event Manager API for RSCT, so tools, such as Perspectives and PMAN, can utilize its services. Figure 3-11 on page 62 shows the communication in RSCT with these tools.

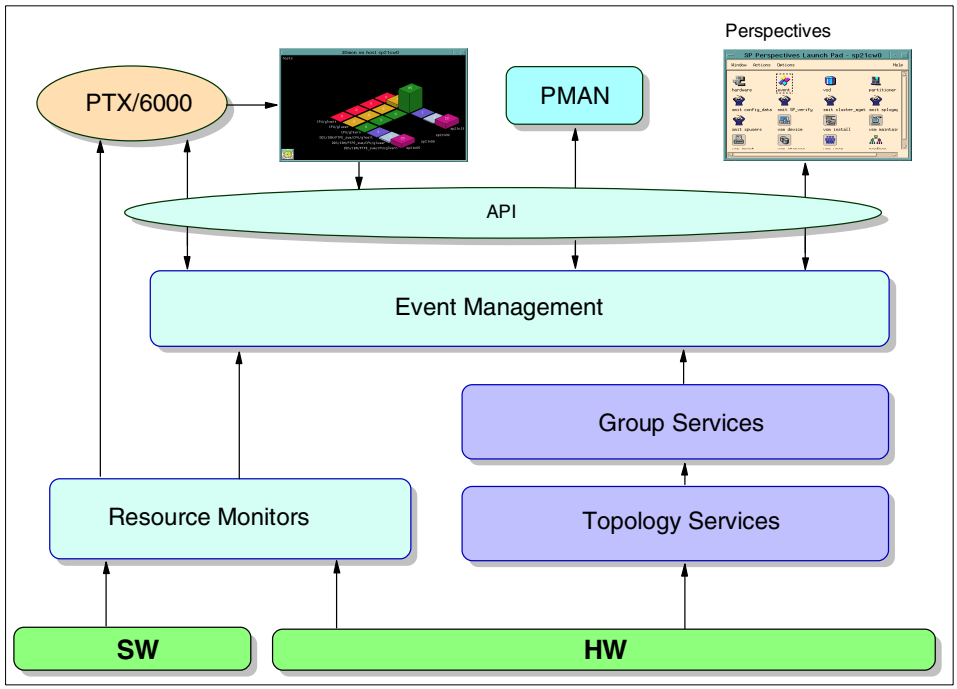


Figure 3-11 Reliable Scalable Cluster Technology and PSSP

As described in 3.2.2, “Communication between RSCT components” on page 51, PSSP is one of the products using the old RSCT design. When you use the `lssrc` command, you find the RSCT daemons for PSSP shown in Table 3-2.

Table 3-2 RSCT daemons in PSSP

Function	PSSP daemon name
Topology Services	hats
Group Services	hags
Group Services Globalized Switch Membership	hagsglsm
Resource Management	haem
Event Management AIX Operation System Resource Monitor	haemaixos

3.3.2 High Availability Cluster Multiprocessing/Enhanced Scalability

High Availability Cluster Multiprocessing/Enhanced Scalability (HACMP/ES) uses the old RSCT infrastructure. RSCT Group Services is used by default. RSCT Event Management can also be configured if necessary. For example, the process application monitoring function of HACMP/ES uses it. Figure 3-12 shows the communication in RSCT to HACMP/ES. This is how it works today (HACMP/ES Version 4.5).

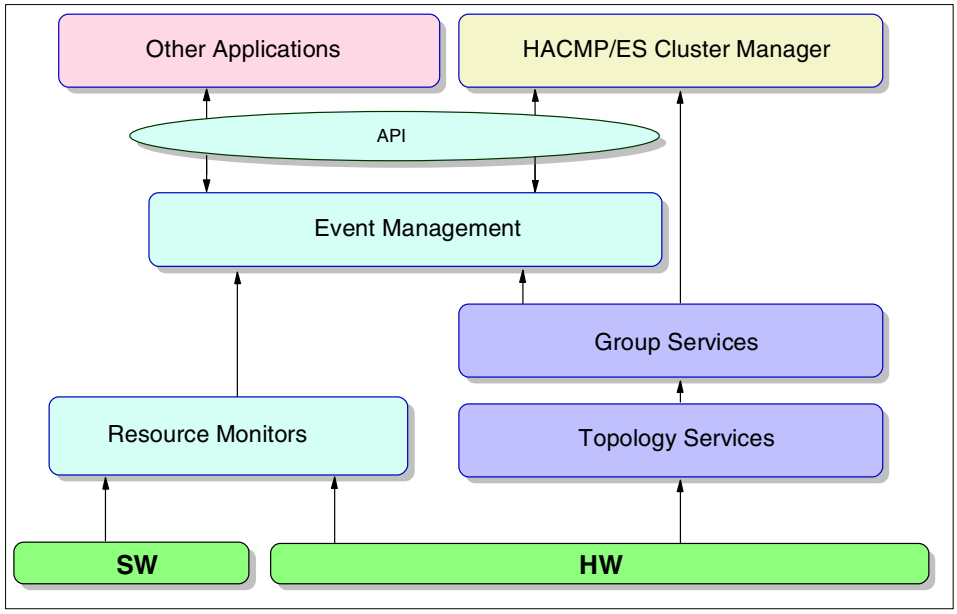


Figure 3-12 Reliable Scalable Cluster Technology and HACMP/ES

As described in 3.2.2, “Communication between RSCT components” on page 51, HACMP/ES is one of the products using the old RSCT design. When you use the `lssrc` command, you find the RSCT daemons for HACMP/ES shown in Table 3-3.

Table 3-3 RSCT daemons in HACMP/ES

Function	HACMP/ES daemon name
Topology Services	topsvcs
Group Services	grpsvcs
Group Services Globalized Switch Membership	grpglsm
Resource Management	emsvcs

Function	HACMP/ES daemon name
Event Management AIX Operation System Resource Monitor	emaixos

3.3.3 General Parallel File System

General Parallel File System (GPFS) is a clustered file system defined over a number of nodes. The overall set of nodes over which GPFS is defined is known as a GPFS cluster. Depending on the operating environment, GPFS can be used in several cluster types. The type names listed here are equal to the value you use in the **mmcrcluster** command. For more information about GPFS, see Chapter 5, "General Parallel File System 2.1" on page 99 or the GPFS manuals *General Parallel File System for AIX 5L: AIX Clusters Concepts, Planning, and Installation Guide*, GA22-7895, and *General Parallel File System for AIX 5L: PSSP Clusters Concepts, Planning, and Installation Guide*, GA22-7899.

The cluster types are as follows:

- sp

  This PSSP cluster environment is based on the shared disk concept of the IBM Virtual Shared Disk (VSD). This is described in more detail in 5.3, "General Parallel File System on Virtual Shared Disk" on page 104.

- lc

  This is based on a Linux operating system. This is described in 5.5, "General Parallel File System on Linux" on page 112.

- hacmp

  This is based on an HACMP/ES cluster. This is described in 5.4, "General Parallel File System on HACMP" on page 108.

- rpd

  This is based on a RSCT peer domain created by the RSCT subsystem of AIX 5L. See 5.6, "General Parallel File System on RSCT peer domain" on page 114 for more details.

GPFS based on RPD is one of the first applications to use the new RSCT design. Figure 3-13 on page 65 shows the communication in GPFS on RPD.

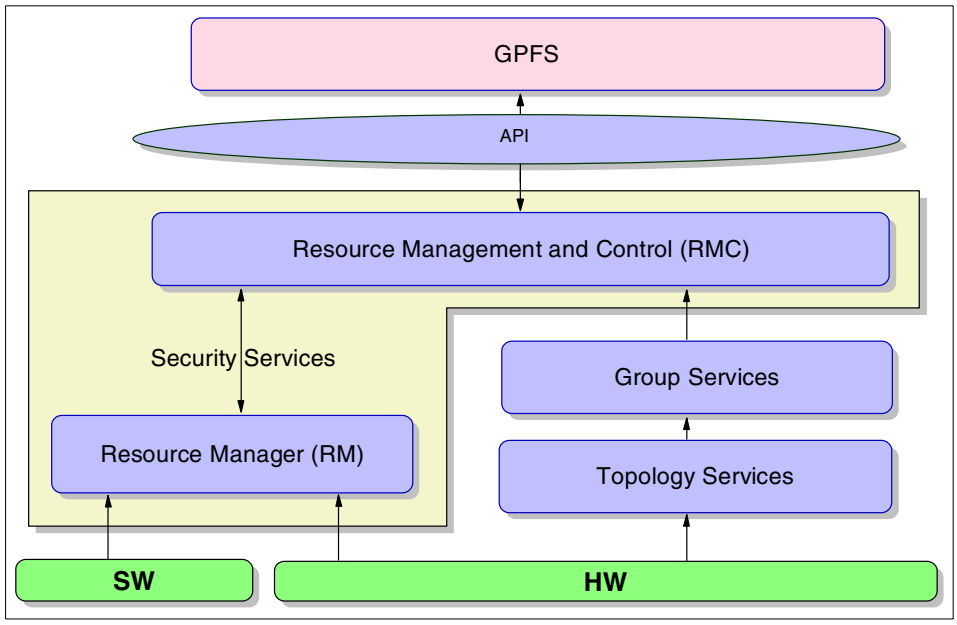


Figure 3-13 Reliable Scalable Cluster Technology and GPFS (using RPD)

As described in 3.2.2, “Communication between RSCT components” on page 51, GPFS based on RPD is one of the products using the new RSCT design. When you use the **`lssrc`** command, you find the RSCT daemons for RPD shown in Table 3-4.

Table 3-4 RSCT daemons in GPFS (using RPD)

Function	RPD daemon name
Topology Services	cthats
Group Services	cthags
Group Services Globalized Switch Membership	cthagsglsm[a]
Resource Management	ctrmc
Security Service	ctcas

a. In the RPD environment, this subsystem is not generally used.

3.4 RSCT peer domain (RPD)

This section briefly describes the new RSCT functionality to build an RSCT peer domain (RPD). For more information, refer to *IBM RSCT for AIX: Guide and Reference*, SA22-7889.

3.4.1 What is RSCT peer domain

When you configure a set of nodes for high availability using the RSCT configuration resource manager, the set of nodes configured is called an RPD. An RPD consists of a number of nodes with no distinguished or master node. All nodes are aware of all other nodes, and administration commands can be issued from any node in the domain.

The Resource Monitoring and Control (RMC) subsystem and RSCT core resource managers are used to manage cluster resources. RMC is a framework for managing, monitoring, and manipulating the physical or logical system entities. RMC runs as a daemon process on individual machines. You can use it to manage and monitor the resources of a single machine or the resources of a RPD. The RMC daemons on the various nodes work together to enable you to manage and monitor the resources of the domain.

The RSCT core resource manager is a daemon process that provides the interface between RMC and the actual physical or logical entities. Therefore, while RMC provides the basic abstractions of resources, a resource manager maps actual entities to the RMC abstractions. RMC and a resource manager provide the administrative and monitoring capabilities of RSCT.

Security services

RSCT cluster security services are used to authenticate and authorize the identity of other parties.

Authentication is the process of ensuring that another party is who it claims to be. Using cluster security services, various cluster applications can check that other parties are genuine and not attempting to gain unwarranted access to the system. Only UNIX host-based authentication is supported, but other security mechanisms may be supported in the future. Authentication is provided by the ctcas daemon, which is started by the RMC service whenever the service starts.

Authorization is the process by which a cluster software component grants or denies resources based on certain criteria. The RSCT component that implements authorization is RMC. It uses access control list (ACL) files in order to control user access to resources. The RMC component subsystem uses cluster security services to map the operating system user identifiers, specified in the ACL file, with network security identifiers to determine if the user has the

correct permissions. This is performed by the identity mapping service, which uses information stored in the identity mapping files ctsec_map.global and ctsec_map.local.

Topology Services subsystem

The Topology Services subsystem is called cthats in an RSCT peer domain. The Topology Services subsystem is used within the RSCT peer domain to provide other RSCT applications and subsystems with network adapter status, node connectivity information, and a reliable messaging service. The Topology Services daemon is contained in the executable file /usr/sbin/rsct/bin/hatsd. This daemon runs on each node in the RSCT peer domain. When each daemon starts, it first reads its configuration from a file, given by the startup command **`cthats`**. This file is called the machines list file (/var/ct/cluster_name/run/cthats/machines.lst), because it has all the machines listed and the IP addresses in the RSCT peer domain. From this file, the hatsd daemon knows the IP address and node number of all the potential heartbeat ring members. That is, the **`cthats`** command obtains the necessary configuration information from the cluster data server, and prepares the environment for the Topology Services daemon in the RSCT peer domain.

In a RSCT peer domain, the configuration resource manager (ConfigRM) controls the Topology Services subsystem. Topology Services is started automatically by the configuration resource manager when you issue the **`startrpdomain`** or **`mkcomg`** command.

Group Services subsystem

The Group Services subsystem is called cthags in an RSCT peer domain. It is used within the RSCT peer domain to provide other RSCT applications and subsystems a distributed coordination and synchronization service. The Group Services subsystem is also started by the configuration resource manager (ConfigRM) when it brings a RSCT peer domain online. It gets the number of the node on which it is running from the local peer domain configuration and tries to connect to the Topology Services subsystem. The

Group Services subsystem monitors the status of all the processes that are joined to a cluster and depend upon Group Services. If either the process or the node on which a process is executing fails, Group Services initiates a failure protocol that informs the remaining nodes in the cluster that one or more nodes have been lost.

3.4.2 Files and directories in a RPD cluster

With the new RPD functionality, we get some new log files and directories. These are for the following:

- The Topology Services subsystem uses the following directories:
  - /var/ct/<cluster_name>/log/cthats, for log files
  - /var/ct/<cluster_name>/run/cthats, for the Topology Services daemon current working directory
  - /var/ct/<cluster_name>/soc/cthats, for the UNIX domain socket files
- The Group Services subsystem uses the following directories:
  - /var/ct/<cluster_name>/lck, for lock files
  - /var/ct/<cluster_name>/log, for log files
  - /var/ct/<cluster_name>/run, for the Group Services daemon current working directories
  - /var/ct/<cluster_name>/soc, for socket files
- The core dumps are located in:
  - /var/ct/<cluster_name>/run/cthags/core*
  - For a HACMP node, the core dumps are located in:
    - /var/ha/run/grpsvcs. cluster/core*
    - /var/ha/run/grpglsm. cluster/core*

4

Parallel System Support Program 3.5 enhancements

This chapter reviews the enhancements made to Parallel System Support Program (PSSP) 3.5, in particular, 64-bit compatibility, switch software command modifications (**`Eprimary`**), and supper user password management. Virtual Shared Disks (VSD), GPFS, and the HPC software improvements are briefly described.

The following topics are discussed:

- 64-bit compatibility
- New software packaging
- Eprimary modifications
- Supper user (supman) password management
- HMC-attached performance improvements
- Virtual Shared Disk and Recoverable Virtual Shared Disk 3.5
- Low-Level Application Programming Interface changes
- General Parallel File System 2.1
- High Performance Computing software stack
- New hardware

4.1 64-bit compatibility

The AIX 64-bit kernel is capable of supporting a larger number of processors, I/O devices, and physical memory (the 32-bit kernel is limited to 96 GB of physical memory). As the internal data structures of the kernel are now extended from 32 to 64 bits, the kernel is also able to support more resources that are used by application programs, such as processes, threads, open files, and shared memory segments. Greater accuracy can also be gained from calculations now performing arithmetic in 64 bits. Since AIX 5L Version 5.1, the 32-bit and 64-bit kernels have the same minimum hardware system requirements. These are 64 MB of physical memory, 128 MB of initial paging space, and 536 MB of disk space to hold the AIX operating system.

Until now, PSSP and all dependent software only operated when used with a 32-bit AIX kernel. PSSP 3.5 has been introduced, and this release has been enhanced to also operate on a 64-bit kernel. The 64-bit kernel support only exists on control workstations and 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 the AIX 64-bit kernel on supported 64-bit hardware

Except for Virtual Shared Disk (VSD) and Kernel Low-Level Application Programming Interface (KLAPI), the kernel extensions and device drivers have been straight ported to the 64-bit environment, meaning that the new 64-bit versions return the same results to the userspace code as the 32-bit versions.

During the design of the 64-bit implementation, it was noted that VSD and KLAPI have additional requirements, because they both make use of kernel addresses across nodes for direct memory access (DMA) buffers. Therefore, the 64-bit versions of these must be able to understand 32-bit addresses, and vice versa. These changes are internal and are transparent to the userspace code.

AIX 5L Version 5.1 installs both the 64- and 32-bit kernels by default. You can switch between 32 bit and 64 bit on nodes running PSSP 3.5. Switching between the kernels is simple and consists of changing a few links to point to the kernel image of choice, rebuilding the boot image, and rebooting the system. The procedure to change from 32 to 64 bit is detailed in Example 4-1 on page 71.

Example 4-1 Switching from 32-bit to 64-bit AIX kernel

```
root $ bootinfo -K
32
root $ ln -sf /usr/lib/boot/unix_64 /unix
root $ ln -sf /usr/lib/boot/unix_64 /usr/lib/boot/unix
root $ bosboot -ad /dev/ipldevice
bosboot: Boot image is 13389 512 byte blocks.
root $ shutdown -Fr

*** System Reboot ***

root $ bootinfo -K
64
```

Note: Some adapters might not be supported under the 64-bit kernel. This means that you will have missing devices when you switch over to 64-bit mode. For compatibility tables, see Appendix D, “AIX device drivers reference” on page 199 and the following Web site:

http://www.ibm.com/servers/aix/os/adapters/51.html

The procedure to change from 64- to 32-bit kernels is detailed in Example 4-2.

Example 4-2 Switching from 64-bit to 32-bit AIX kernel

```
root $ bootinfo -K
64
root $ ln -sf /usr/lib/boot/unix_mp /unix
root $ ln -sf /usr/lib/boot/unix_mp /usr/lib/boot/unix
root $ bosboot -ad /dev/ipldevice
bosboot: Boot image is 13389 512 byte blocks.
root $ shutdown -Fr

*** System Reboot ***

root $ bootinfo -K
32
```

Note: If you switch a node back to 32-bit mode after it is installed with the 64-bit defaults, you will have JFS2 file systems running under a 32-bit kernel. Although this works, it is not recommended.

The commands, shown in Example 4-2, (omitting the **shutdown** command) can be placed in the script.cust script to run and change the kernel type before the first reboot at node installation time. The script will then alter any customized node from that point onward. However, another reboot is needed.

To add support for 64 bit to the High Performance Computing (HPC) set of products, see 4.9, “High Performance Computing software stack” on page 91.

Note: PSSP 3.5 cannot run on any operating system other than AIX 5L Version 5.1. IBM intends to support PSSP 3.5 with AIX 5L Version 5.2 in 2003.

4.2 New software packaging

There have been a couple of small changes to how PSSP software is shipped. The following sections describe the changes in more detail.

4.2.1 Two install images

There are now two mksysb images shipped with PSSP 3.5. The first image is the same as the one that was shipped previously (updated with fixes for 32-bit kernel and JFS root file system). The second is a mksysb image of a system with a 64-bit AIX kernel and JFS2 file systems. The new image is contained on CD 2 of the PSSP set and can be used to install any node you want with a 64-bit environment. The new image is called:

spimg.510_64 3.5.0.0 COMMITTED Minimal AIX 510 64-bit mksysb

When installed, this fileset adds a new mksysb image to your install/images directory. Example 4-3 shows the new mksysb image file, kernel type, and file system sizes of this new install image. All file systems are JFS2.

Example 4-3 New 64-bit mksysb image details

```
-rw-r--r--   1 bin      bin       240445440 Oct 16 11:03 bos.obj.ssp.510
-rw-r--r--   1 bin      bin       239892480 Oct 16 11:04 bos.obj.ssp.510_64
# bootinfo -K
64
# df -k
Filesystem    1024-blocks      Free %Used    Iused %Iused Mounted on
/dev/hd4            65536     57904   12%     1135     8% /
/dev/hd2           458752     84344   82%    12184    38% /usr
/dev/hd9var         65536     60396    8%      483     4% /var
/dev/hd3            65536     64968    1%       23     1% /tmp
/dev/hd1            65536     65160    1%        7     1% /home
/proc                   -         -    -         -     -  /proc
/dev/hd10opt        65536     65164    1%       16     1% /opt
```

Alternatively, you can switch the symbolic links to enable a 64-bit kernel on any of your 64-bit capable nodes after installation.

Note: To find out the file system space requirements for all the PSSP 3.5 filesets and install images, see Chapter 3 in the *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.

Note: It is very important to read the *Read This First* document before doing anything with this new PSSP version. The latest version of this document can be found on the following Web site:

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

Click the **Read This First** link.

4.2.2 Reliable Scalable Cluster Technology

Reliable Scalable Cluster Technology (RSCT) is no longer shipped with the PSSP product set. It is now integrated into AIX 5L Version 5.1 and shipped with those CDs. AIX 5L Version 5.1 installs RSCT by default. Because PSSP 3.5 only runs on AIX 5L Version 5.1, there was no need to package RSCT with PSSP. During the installation of PSSP, rsct.basic.sp is installed to customize RSCT so that it works with PSSP.

For more information about RSCT, see the *IBM RSCT for AIX: Guide and Reference*, SA22-7889.

4.3 Eprimary modifications

Switch error handling is done through the switch itself. Detected faults are communicated to an active node on the switch called the primary node. Another node is also selected to become a backup to this primary node should it fail. For full details of how the primary and backup nodes are selected and how failover is handled, see Chapter 14 of the *PSSP for AIX: Administration Guide*, SA22-7348.

By default, all nodes are enabled to be configured as primary or backup nodes. Until now, it has been impossible to exclude a node from becoming the primary or the backup. PSSP 3.5 includes a new function to allow you to do this. The flags **-d** and **-e** have been added to the **Eprimary** command to allow you to disable and enable any nodes from becoming the primary or backup node, as shown in Example 4-4 on page 74.

Example 4-4 New Eprimary flags

Old syntax:

```
Eprimary [-h] [-p { 0 | 1 | all }] [-init] [node_identifier] [-backup
bnode_identifier]
```

New syntax:

```
Eprimary [-h] [-p { 0 | 1 | all }] [-init] [node_identifier] [-backup
bnode_identifier] [-e { node_id,..,node_id | all }]
[-d { node_id,..,node_id | all }]
```

You can specify a list of nodes to the **-d** and **-e** flags or specify **all**. The **all** function has the effect of setting all nodes to disable or enable. A primary disabled node will not be selected as a primary or backup node provided another primary enabled node is available. If this is not the case, the system selects a disabled node. The selected node has the new **primary_enabled** attribute set to **forced_true** in the SDR node class, and an error is reported to the error log.

In Example 4-5, **Eprimary** automatically selects the node with the lowest IP address for the primary node. The node with the IP address furthest away from the primary node is selected as the backup node. In this case, node1 is the primary and node13 is the backup. We make a decision that only nodes 1, 9, and 13 should be able to become switch primary nodes. To do this, we first disable all nodes from becoming the primary with **Eprimary -d** and then enable node 9. This is a quicker method than disabling each node one by one. Node 1 and 13 cannot be disabled because they are already primary and backup primary nodes, respectively.

Example 4-5 New Eprimary functionality

```
---------------------------------- Frame 1 ----------------------------------
                         Host     Switch   Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- -------- ------- ----- ---------------- -------
  1     1  wide   on     yes      yes     N/A     no  LCDs are blank      no
  3     3  wide   on     yes      yes     N/A     no  LCDs are blank      no
  5     5  wide   on     yes      yes     N/A     no  LCDs are blank      no
  7     7  wide   on     yes      yes     N/A     no  LCDs are blank      no
  9     9  thin   on     yes      yes     N/A     no  LCDs are blank      no
 10    10  thin   on     yes      yes     N/A     no  LCDs are blank      no
 11    11  thin   on     yes      yes    normal   no  LEDs are blank      no
 12    12  thin   on     yes      yes    normal   no  LEDs are blank      no
 13    13  thin   on     yes      yes    normal   no  LEDs are blank      no

root $ Eprimary
1       - primary
```

```
1       - oncoming primary
13      - primary backup
13      - oncoming primary backup
1       - autounfence

root $ Eprimary -d all
All nodes, except primaries, successfully primary disabled
root $ Eprimary -d
Primary disabled nodes
           3
           5
           7
           9
          11
          12
          14
          10

root $ Eprimary -e sp6n09e0
Node sp6n09e0 successfully primary enabled
root $ Eprimary -e
Primary enabled nodes
           1
           9
          13
No primary enabled nodes with a value of forced_true
```

Example 4-6 shows the primary allocation when the current primary node fails. The primary allocation is then moved to the primary backup node. The new primary backup node is then selected from the enabled nodes on the **Eprimary -e** list. In this case, node 13 becomes the primary and node 9 becomes the primary backup.

Example 4-6 Eprimary node selection

```
root $ Eprimary
1       - primary
1       - oncoming primary
13      - primary backup
13      - oncoming primary backup
1       - autounfence
root $ Eprimary -e
Primary enabled nodes
           1
           9
          13
No primary enabled nodes with a value of forced_true
```

```
root $ #Shutdown of the current primary node
root $ spmon -power off node1

---------------------------------- Frame 1 ----------------------------------
                        Host     Switch   Key      Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- -------- ------- ----- ---------------- -------
  1     1  wide  off     no     autojn    N/A     no  OK                  no
                                                     LCD2 is blank
  3     3  wide   on     yes      yes     N/A     no  LCDs are blank      no
  5     5  wide   on     yes      yes     N/A     no  LCDs are blank      no
  7     7  wide   on     yes      yes     N/A     no  LCDs are blank      no
  9     9  thin   on     yes      yes     N/A     no  LCDs are blank      no
 10    10  thin   on     yes      yes     N/A     no  LCDs are blank      no
 11    11  thin   on     yes      yes    normal   no  LEDs are blank      no
 12    12  thin   on     yes      yes    normal   no  LEDs are blank      no
 13    13  thin   on     yes      yes    normal   no  LEDs are blank      no

root $ Eprimary
13      - primary
1       - oncoming primary
9       - primary backup
13      - oncoming primary backup
1       - autounfence
```

Finally, Example 4-7 on page 77 shows that when no node on the enabled list is free, one of the disabled nodes is selected and its state is set to **forced_true**. Nodes 1, 9, and 13 are on the enabled list. Nodes 1 and 13 are lost, leaving only node 9 and no other server on the list to select as a primary backup. **Eprimary** selected node 3 and set its state to **forced_true**.

Attention: In Example 4-7, what happens when node 13 come back? Does node 3 get reset? Or does node 3 have to go away before it is reset?

Nothing is reset automatically. First, you must issue the **Estart** command to make 13 the oncoming primary backup. Then issue the **Eprimary -d 3** command to disable node 3 from becoming a primary or primary backup node. Finally, issue the **Estart** command again to make node 13 the primary backup.

We recommend issuing the **Estart** command when applications are not utilizing the switch.

Example 4-7 Eprimary forced true example

```
root $ spmon -p off node13
root $ Eprimary
9       - primary
1       - oncoming primary
3       - primary backup
13      - oncoming primary backup
1       - autounfence
root $ Eprimary -e
Primary enabled nodes
           1
           9
          13
Primary enabled nodes with a value of forced_true
           3
```

4.4 Supper user (supman) password management

PSSP uses the supman user ID to distribute file collections through **supper**. The user.admin collection contains sensitive data, such as the /etc/security/passwd file. To reduce that security risk, the supman password should be managed like any other password in your system. Management from the control workstation has been enabled by the addition of several commands to allow management of the password. Previously, passwords on each node had to be managed manually.

The commands are as follows:

setsuppwd Sets the password for the supman user. It must be run by the root user on the control workstation. The password is stored in the /spdata/sys1/sup/sysman.key file. A checksum is also created in the same directory so that nodes can check when they receive the key if it was transmitted correctly. The file is ASCII text, but it is root read/write only and protected in a root owned read/write only directory.

usesuppwd This command tells the control workstation to use the password set with the **setsuppwd** command. It also sets the **supman_passwd_enabled** attribute to true in the SP class within the SDR. This enables additional password checking in the file management system. Again, this command must be run by root on the control workstation.

updsuppwd This command only runs on the nodes in your SP system. It collects the /spdata/sys1/sup/sysman.key generated with the **setsuppwd** command. After the password is collected, the /etc/security/passwd file is updated with the new password setting for supman.

If for any reason one of the commands fails, the error will be output to stderr and stored in the /var/adm/SPlogs/filec/suppwd.log log file.

Attention: This feature has been fitted to all supported PSSP releases.

A procedure for updating the supman password and updating all the nodes is in Chapter 7 of the *PSSP for AIX: Administration Guide*, SA22-7348. An example of setting the supper password is shown in Figure 4-1 and Example 4-8 on page 79.

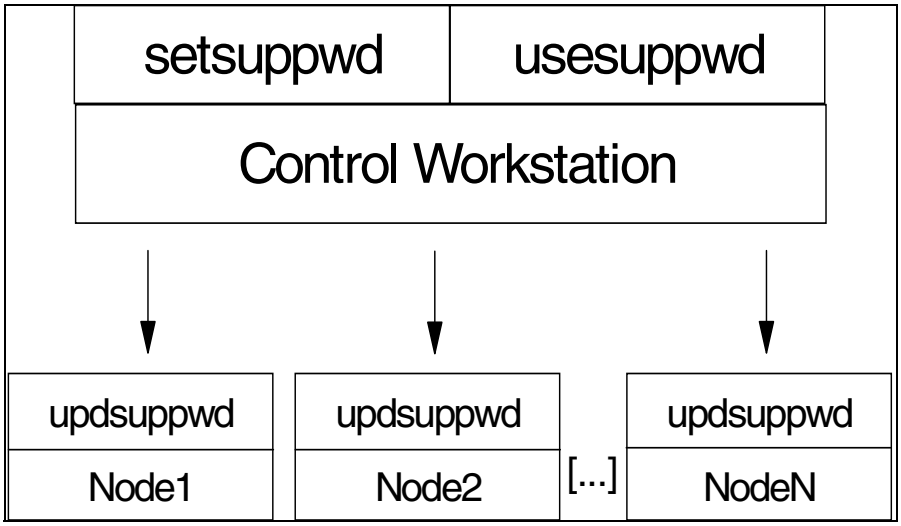


Figure 4-1 Setting the supper password chart

Example 4-8 Setting the supper password

```
root $ setsuppwd
Enter new password for supman:
Enter new password for supman again:
New password saved for supman id.
setsuppwd ended with return code 0.
sp4en0:/
root $ usesuppwd
Password enabled for supman id.
usesuppwd ended with return code 0.
sp4en0:/
root $ dsh -a /usr/lpp/ssp/bin/updsuppwd
sp4n17e0: updsuppwd ended with return code 0.
sp4n17e0:
sp4n01e0: updsuppwd ended with return code 0.
sp4n01e0:
sp4n33e0: updsuppwd ended with return code 0.
```

4.5 HMC-attached performance improvements

With the announcement of PSSP 3.5, a new HMC was also announced that improves performance by using a higher processor clock frequency. This allows the HMC to manage more LPARs than the older ones. For more information, refer to the “IBM 7315-C01 Hardware Management Console Announcement Brief,” announcement date October 8, 2002.

4.6 Virtual Shared Disk and Recoverable Virtual Shared Disk 3.5

IBM Virtual Shared Disk (VSD) and IBM Recoverable Virtual Shared Disk (RVSD) are additional components of the PSSP product that you can optionally install to let multiple nodes share the information they hold. For more details about VSD and RVSD, see *PSSP for AIX: Managing Shared Disks*, SA22-7349.

PSSP 3.5 VSD communication to nodes running PSSP 3.2 and 3.4 can only happen over IP and requires all nodes to run with a 32-bit kernel. VSD communication between 32- and 64-bit kernels is supported, provided all nodes participating in VSD are at PSSP 3.5. LAPI/KLAPI support is only for nodes running PSSP 3.5. Figure 4-2 on page 80 shows how VSD can interoperate with other levels of PSSP and different kernel types.

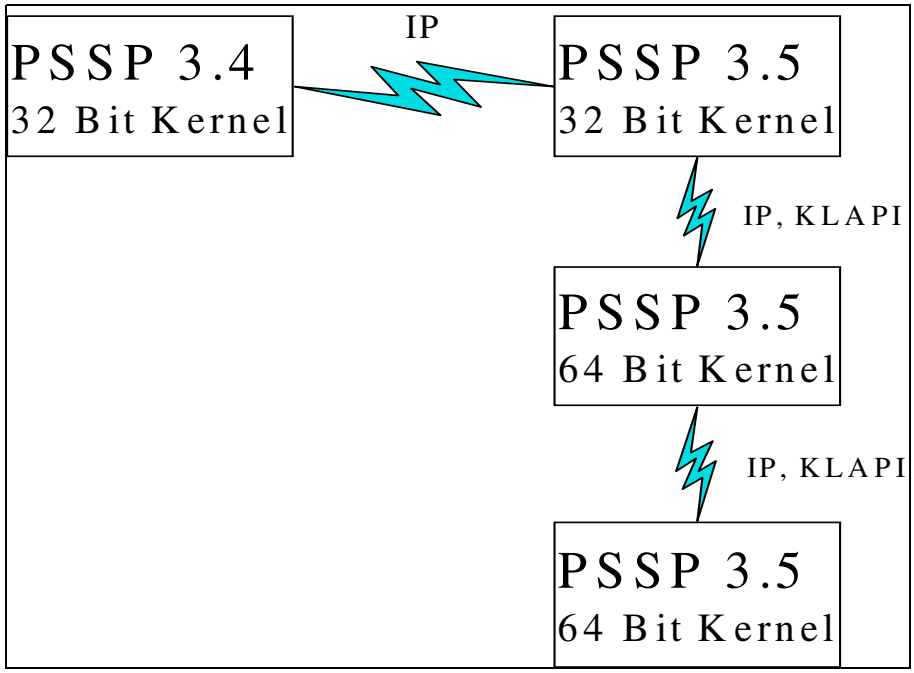


Figure 4-2 Virtual Shared Disk communication

To get the latest level of recovery function with RVSD, the control workstation and each node in the system partition that will use Virtual Shared Disks must have AIX 5L Version 5.1 Maintenance Level 3 or later and PSSP 3.5 with the VSD and RVSD components installed.

RVSD requires the RSCT Group Services and Topology Services utilities to operate. These components must be installed from the AIX installation media.

Note: If you plan to migrate disk servers to AIX 5L Version 5.1 or later, all the servers that share volume groups should be migrated at the same time.

The following sections highlight the new features for VSD and RSVD.

4.6.1 64-bit compatibility

VSD v3.5 will automatically load a 64- or 32-bit version depending on what kernel the AIX system is running at initialization time. 32- and 64-bit coexistence requires that all VSD nodes must be running PSSP 3.5.

4.6.2 Recoverable Virtual Shared Disk integration

IBM Recoverable Virtual Shared Disk (RVSD) is still a separate fileset, but it is no longer a separately Licensed Program Product (LPP). It is now an integrated component of the VSD package.

4.6.3 Expanded Concurrent Virtual Shared Disk support

With previous versions of Concurrent Virtual Shared Disk (CVSD), support was only for SSA attached disks. Version 3.5 has added support for ESS disks. For more information about these subsystems, see:

http://www.storage.ibm.com/hardsoft/products/ess/index.html

4.6.4 New command: updatevsdvg

The **updatevsdvg** command changes VSD global volume group characteristics. This command enables you to change global volume groups from Concurrent Virtual Shared Disk volume groups to Virtual Shared Disk volume groups, and vice versa. It can be used whenever server node numbers change, such as replacing or recabling servers where the new server numbers are different, or when you need to delete a server.

Syntax:

```
updatevsdvg -g global_volgrp {-k VSD -p primary_node -b secondary_node |
-k CVSD -l server_list [-c cluster_name]}
```

Note: This command can be run while the RVSD subsystem is active. No application can use the VSD that is part of the volume group that **updatevsdvg** is working on.

4.6.5 Large and dynamic buddy buffer enhancement

The buddy buffer is pinned kernel memory used to temporarily store data for I/O operations from a client node. The stored data is flushed from the buffer immediately after the clients I/O operation completes.

Important: The term *buddy buffer* is used to mean both the total pinned kernel memory that is being managed and also to refer to the individual temporary I/O buffers that make up this space.

Up until this release of VSD, all of the memory for the buddy buffer was pinned at configuration time and could not be reclaimed for other uses. Now only a quarter of the memory requested for the buffer, up to a maximum of 64 MB, is pinned at

device driver configuration time. The device driver attempts to dynamically expand and contract any additional buddy buffer space up to the maximum specified. Therefore, it is generally advisable to configure a large amount of buddy buffer space, because the system only allocates what is needed. The previous limitation of buddy buffer size was 256 MB. An AIX 32-bit kernel limits the theoretical maximum buddy buffer size to 256 MB corresponding to a 256 MB segment size. An AIX 64-bit kernel supports the allocation of large regions of global memory, allowing a larger buddy buffer size to be specified.

Important: In the following sections, MB refers to the total pinned memory size, and KB refers to the individual buffer sizes that add up to the total size.

PSSP for AIX: Managing Shared Disks, SA22-7349, suggests setting 4096 (4 KB) and 262144 (256 KB), respectively, for the minimum and maximum buddy buffer sizes. The suggested starting value for the total number of maximum size buffers is as follows:

- For an AIX 32-bit kernel, 128 256 KB buffers, which results in an initial pinned buddy buffer size of 8 MB that can increase to 32 MB. This is shown in Figure 4-3 on page 83.

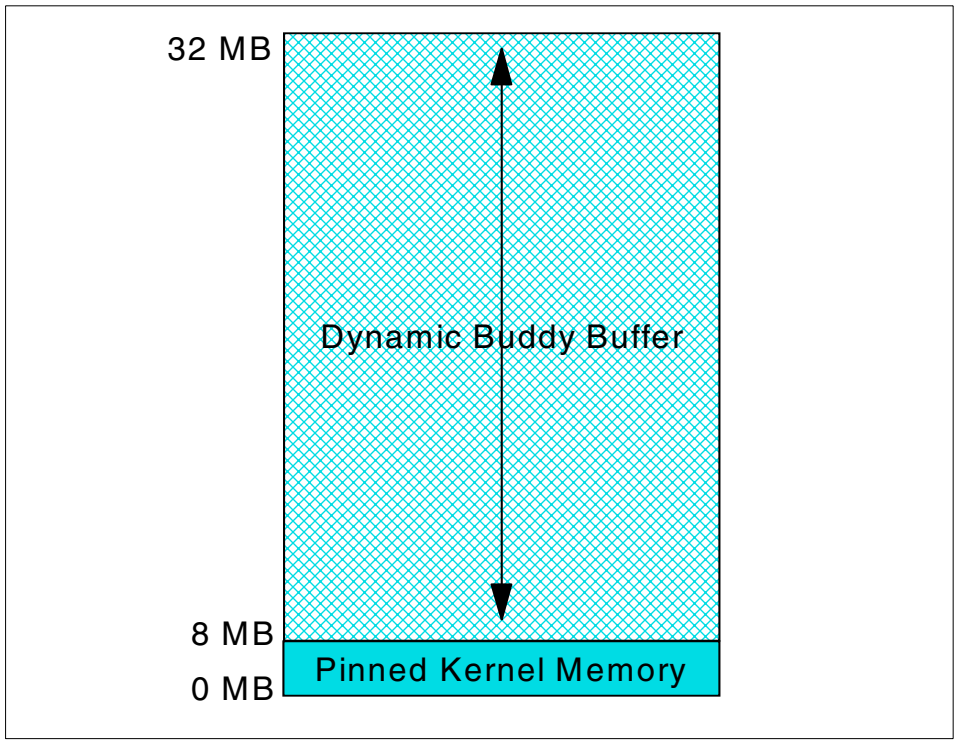


Figure 4-3 32-bit kernel example

- For an AIX 64-bit kernel, 2000 256 KB buffers, which results in an initial pinned buddy buffer size of 64 MB that can increase to 500 MB. The 64-bit kernel is able to allocate a much larger buddy buffer. This is shown in Figure 4-4 on page 84.

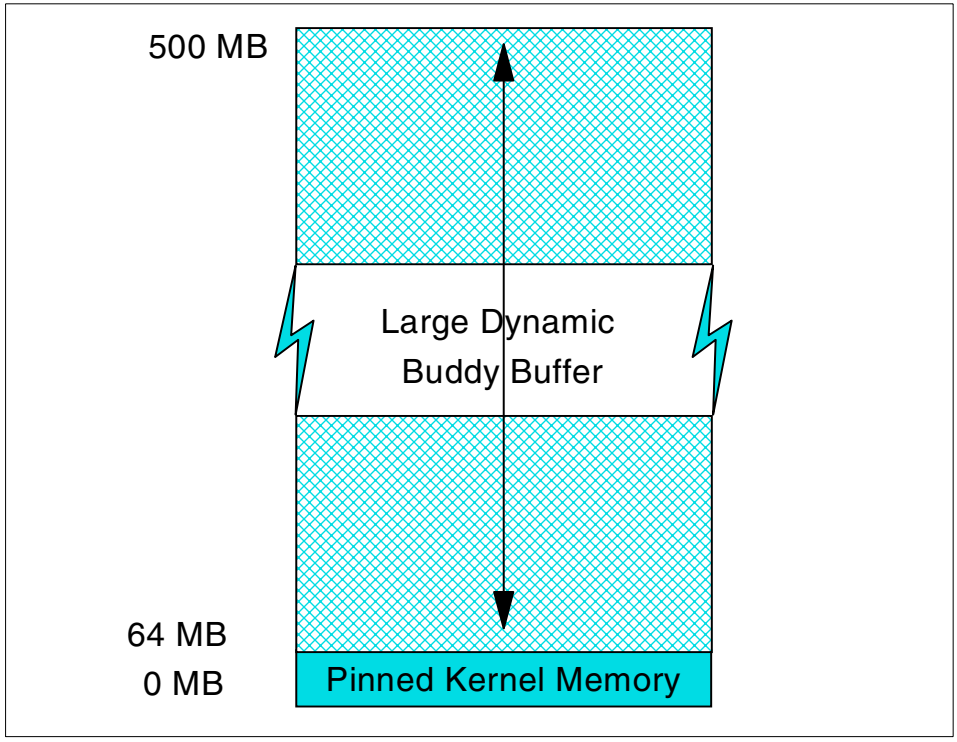


Figure 4-4 Large dynamic buddy buffer

You can display your current buddy buffer settings using the **vsdatalst** command. To set new parameters for buddy buffers, use the VSD perspective graphical user interface or the **vsdnode** command.

Usage:

```
vsdnode node_number  ...  adapter_name  init_cache_buffer_count
max_cache_buffer_count  vsd_request_count  rw_request_count
min_buddy_buffer_size  max_buddy_buffer_size  max_buddy_buffers VSD_maxIPmsgsz
[cluster_name]
```

An example of the **vsdatalst** command output is in Example 4-9 on page 85.

Example 4-9 vsdatalst output

```
root $ vsdatalst -n

       VSD Node Information

                                          Initial Maximum     VSD      rw      Buddy Buffer
  node                 VSD      IP packet   cache   cache request request minimum maximum size: #
number host_name       adapter    size    buffers buffers   count   count    size    size maxbufs
------ --------------- -------- --------- ------- ------- ------- ------- ------- ------- -------
     1 sp6n01e0        css0         61440      64     256     256      48    4096  131072       4
     3 sp6n03e0        css0         61440      64     256     256      48    4096  131072       4
    10 sp6n10e0        css0         61440      64     256     256      48    4096  131072       4
    12 sp6n12e0        css0         61440      64     256     256      48    4096  131072       4
```

There is no user command to display the current size of your buddy buffers, only what the minimum and maximum sizes are set to.

4.6.6 IP flow control

IP flow control has been added to nodes that are running VSD V3.5 or later. The device driver will only implement the flow control if both the source and target support it. This is not user configurable and cannot be switched off.

The client node sends a read request to the server node to access data on one of the Virtual Shared Disks it possesses. The server then does the I/O on the requested data and sends it back to the client. After the data is received by the client, it sends an acknowledgement (ACK) back to the server to acknowledge that the data has been received successfully. The ACK contains the identifiers of all packets successfully received. The server will resend any packets that have not been ACKed after a timeout period is reached. For a typical 256 KB request made up of five packets, there would be one ACK after the five packets have been received. This is shown in Figure 4-5 on page 86.

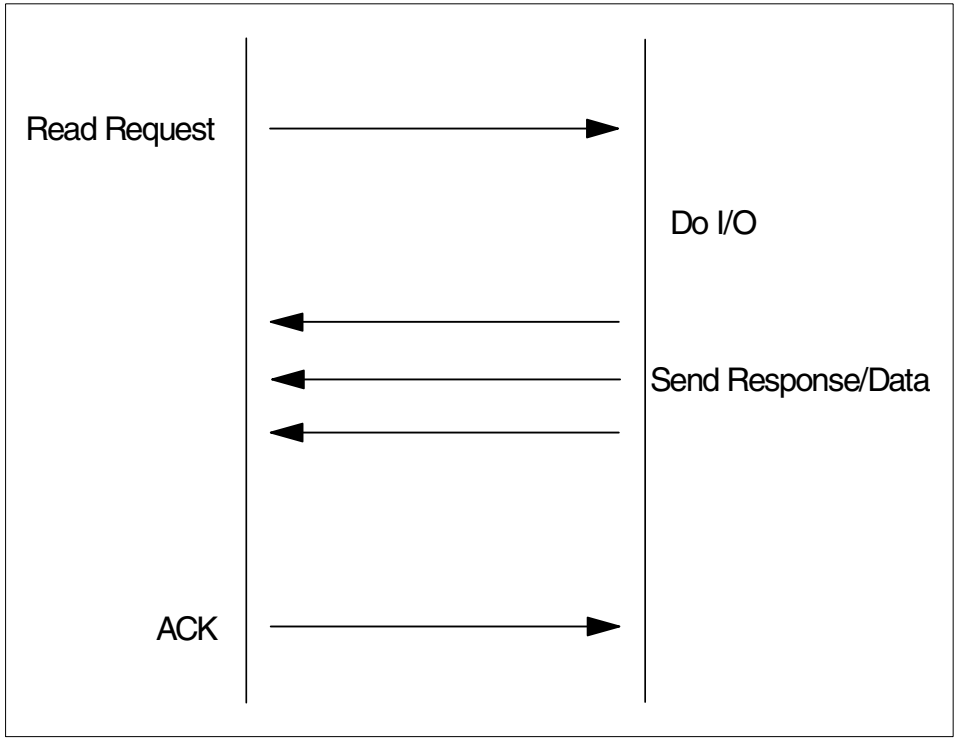


Figure 4-5 IP flow control: Read

A write request is received by the server. It then allocates a buddy buffer to hold the arriving data and responds saying that it is ready for the client to send. On receipt of this, the client sends the data to the server. The server responds with an ACK to the client stating which packets have been received. The client has timers running so that if not all packets are ACKed, and the timeout period is reached, the lost packets are resent from the server again. After all the data is received, the server starts committing the data to the VSD volume. When the data is committed, a reply is sent to the client informing it of the completed I/O. This is shown in Figure 4-6 on page 87.

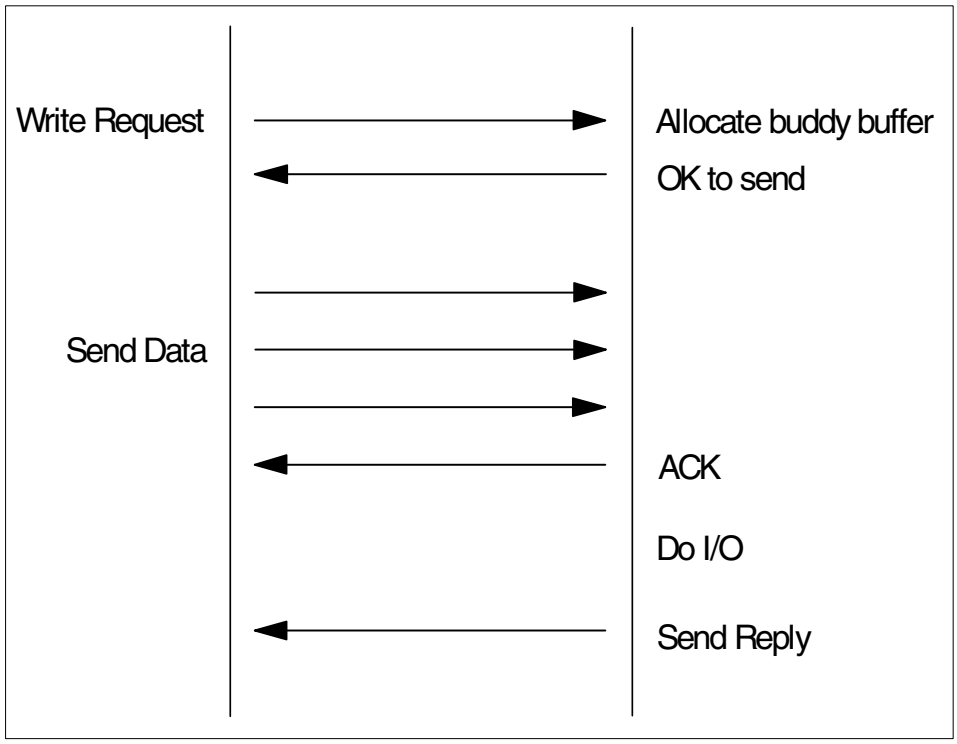


Figure 4-6 IP flow control: Write

Tests have shown that the IP flow control version of the VSD code runs equal to or faster than the previous versions, while offering greater stability because packets are now ACKed. As loads increase, the reads and writes do not show any substantial drop off. This is due to many internal changes to the VSD product. At the time of writing, there are no published performance details we can reference in this redbook.

4.6.7 FAStT support in RVSD

Support has been added to RVSD for the FAStT family of disk subsystems. For more information about these subsystems, see:

http://www.storage.ibm.com/hardsoft/disk/fastt/

4.6.8 AIX trace hooks

Trace hooks have been added to AIX to trap VSD actions should it be required to diagnose system problems. The trace hook is 418. You switch on tracing with the **trace** command and produce a report with the **trcrpt** command. More information about tracing can be found in Chapter 27 of the *AIX General Programming Concepts: Writing and Debugging Programs*:

- For AIX Version 4.3, see:
  http://publib.boulder.ibm.com/doc_link/en_US/a_doc_lib/aixprggd/genprogc/toc.htm
- For AIX 5L Version 5.1, see:
  http://publibn.boulder.ibm.com/doc_link/en_US/a_doc_lib/aixprggd/genprogc/genprogctfrm.htm
- For AIX 5L Version 5.2, see:
  http://publib16.boulder.ibm.com/pseries/en_US/aixprggd/genprogc/genprogc.pdf

Trace hook 418 captures the following VSD actions:

```
  VSD_TRC_CLTBEG 0x01
/* maj|min(of vsd) src|tgt seq# rd|wt count */
  VSD_TRC_SRVBEG 0x02
/* maj|min(of lv) src|tgt seq# rd|wt count */
  VSD_TRC_LCLBEG 0x04
/* maj|min(of lv) src|tgt seq# rd|wt count */
  VSD_TRC_ENDIO 0x08
 /* src|tgt seq# Elapasedtime.sec Elapsedtime.nsec */
  VSD_TRC_ENDRDWT 0x10
/* src|tgt seq# Elapasedtime.sec Elapsedtime.nsec */
```

An example of capturing a VSD trace of a file being copied to a GPFS volume is shown in Example 4-10.

Example 4-10 Using VSD trace hook 418

```
root $ trace -d -j 418  -m "Tracing VSD activity"
-> trcon
-> !cp /etc/hosts /gpfs0fs/redbook.txt
-> trcoff
-> quit
sp6n01e0:/usr/lpp/csd/bin
root $ trcrpt -O "exec=off,pid=off,2line=off,timestamp=3"
>vsd-cp.trace-output.txt
sp6n01e0:/usr/lpp/csd/bin
root $ cat vsd-cp.trace-output.txt

Thu Oct 10 12:25:14 2002
System: AIX sp6n01e0 Node: 5
Machine: 000132374C00
Internet Address: C0A80601 192.168.6.1
The system contains 4 cpus, of which 4 were traced.
```

```
Buffering: Kernel Heap
This is from a 32-bit kernel.
Tracing only these hooks, 418

trace -d -j 418 -m Tracing VSD activity


ID  APPL    SYSCALL KERNEL  INTERRUPT

001                 TRACE ON channel 0
                    Thu Oct 10 12:25:18 2002
418                 vsd_lclbeg: lv  maj/min:0x290001 src/tgt: 0x10001 seqnum:0x013E
                                rd|wt:0x0000 count:0x0200
418                 vsd_endio:                       src/tgt: 0x10001 seqnum:0x013E
                                ElapsedTime:0 . 13263732
418                 vsd_endrdwt:                     src/tgt: 0x10001 seqnum:0x013E
                                ElapsedTime:0 . 13367249
418                 vsd_cltbeg: vsd maj/min:0x260003 src/tgt: 0x10003 seqnum:0x0012
                                rd|wt:0x0001 count:0x2000
418                 vsd_endrdwt:                     src/tgt: 0x10003 seqnum:0x0012
                                ElapsedTime:0 . 17959352
418                 vsd_lclbeg: lv  maj/min:0x270001 src/tgt: 0x10001 seqnum:0x013F
                                rd|wt:0x0001 count:0x0200
418                 vsd_endio:                       src/tgt: 0x10001 seqnum:0x013F
                                ElapsedTime:0 . 16998633
418                 vsd_endrdwt:                     src/tgt: 0x10001 seqnum:0x013F
                                ElapsedTime:0 . 17063862
418                 vsd_lclbeg: lv  maj/min:0x270001 src/tgt: 0x10001 seqnum:0x0140
                                rd|wt:0x0001 count:0x0200
418                 vsd_endio:                       src/tgt: 0x10001 seqnum:0x0140
                                ElapsedTime:0 . 7960933
418                 vsd_endrdwt:                     src/tgt: 0x10001 seqnum:0x0140
                                ElapsedTime:0 . 8012909
418                 vsd_lclbeg: lv  maj/min:0x290001 src/tgt: 0x10001 seqnum:0x0141
                                rd|wt:0x0001 count:0x0200
418                 vsd_endio:                       src/tgt: 0x10001 seqnum:0x0141
                                ElapsedTime:0 . 2535918
418                 vsd_endrdwt:                     src/tgt: 0x10001 seqnum:0x0141
                                ElapsedTime:0 . 2582448
418                 vsd_lclbeg: lv  maj/min:0x290001 src/tgt: 0x10001 seqnum:0x0142
                                rd|wt:0x0001 count:0x0200
418                 vsd_endio:                       src/tgt: 0x10001 seqnum:0x0142
                                ElapsedTime:0 . 10038492
418                 vsd_endrdwt:                     src/tgt: 0x10001 seqnum:0x0142
                                ElapsedTime:0 . 10086491
418                 vsd_lclbeg: lv  maj/min:0x270001 src/tgt: 0x10001 seqnum:0x0143
                                rd|wt:0x0001 count:0x0200
418                 vsd_endio:                       src/tgt: 0x10001 seqnum:0x0143
                                ElapsedTime:0 . 10162564
418                 vsd_endrdwt:                     src/tgt: 0x10001 seqnum:0x0143
                                ElapsedTime:0 . 10208996
002                 TRACE OFF channel 0000 Thu Oct 10 12:25:32 2002
```

4.7 Low-Level Application Programming Interface changes

To facilitate 32-bit and 64-bit communication, a new function, LAPI_Xfer, has been added as a replacement for the following old communication functions:

- LAPI_Amsend, LAPI_Amsendv
- LAPI_Put, LAPI_Putv
- LAPI_Get, LAPI_Getv
- LAPI_Rmw

This routine provides a superset of the functionality of the old routines. However, the original routines still remain and are supported. The new routine provides two important capabilities not present in the original:

- Remote address fields are expanded to 64 bits in length. This allows a 32-bit process to send data to a 64-bit address.
- The new interface allows the origin counter to be replaced with a send completion callback.

Example 4-11 shows the changes to /usr/include/lapi.h to add the new routine.

Example 4-11 LAPI_Xfer from lapi.h

```
LAPI_Xfer(lapi_hndlr_t, lapi_xfer_t *);

typedef enum { LAPI_GET_XFER, LAPI_AM_XFER, LAPI_PUT_XFER,
               LAPI_GETV_XFER, LAPI_PUTV_XFER, LAPI_AMV_XFER,
               LAPI_RMW_XFER, LAPI_LAST_XFER
             } lapi_xfer_type_t;

typedef union {
        lapi_xfer_type_t  Xfer_type;
        lapi_get_t        Get;
        lapi_am_t         Am;
        lapi_rmw_t        Rmw;
#ifndef _KERNEL_LAPI
        lapi_put_t        Put;
        lapi_getv_t       Getv;
        lapi_putv_t       Putv;
        lapi_amv_t        Amv;
#endif /* _KERNEL_LAPI */
} lapi_xfer_t;

typedef struct {
        lapi_xfer_type_t  Xfer_type; /* must be  LAPI_AM_XFER     */
        int               flags;     /* use zero copy for example */
```

```
        lapi_long_t       hdr_hdl;   /* Am  header handler        */
        uint              tgt;       /* target task               */
        uint              uhdr_len;  /* user header length        */
        void              *uhdr;     /* user header data          */
        void              *udata;    /* user data to be xfered    */
        ulong             len;       /* transfer length           */
        scompl_hndlr_t    *shdlr;    /* send completion handler   */
        void              *sinfo;    /* send completion data      */
        lapi_long_t       tgt_cntr;  /* target counter            */
        lapi_cntr_t       *org_cntr; /* origin counter            */
        lapi_cntr_t       *cmpl_cntr;/* origin counter            */
    } lapi_am_t;
```

A second function, LAPI_Address_init64, has been added to allow 32- and 64-bit tasks to exchange addresses. The interface is shown in Example 4-12.

Example 4-12 LAPI_Address_init64 from lapi.h

```
LAPI_Address_init64( lapi_handle_t hndl,
                     void *my_addr,
                     lapi_long_t  *add_tab[]);
```

For more information about these changes, see *PSSP for AIX: Command and Technical Reference, Volume 2*, SA22-7351.

4.8 General Parallel File System 2.1

There are many changes included in the General Parallel File System (GPFS) Version 2.1 release packaged with PSSP 3.5. This topic is discussed in detail in Chapter 5, “General Parallel File System 2.1” on page 99.

4.9 High Performance Computing software stack

PSSP supports a lot of software and the HPC stack, mainly intended for High Performance Computing (HPC). Although the last full releases of the components were with PSSP 3.4 in December 2001, improvements to all the products were achieved through program temporary fixes (PTFs). This section summarizes enhancements to the following:

- LoadLeveler Version 3.1
- Parallel Environment (PE) Version 3.2
- Engineering and Scientific Subroutine Library (ESSL) Version 3.3

- Parallel Engineering and Scientific Subroutine Library (Parallel ESSL) Version 2.3

4.9.1 LoadLeveler

Since its release in December 2001, LoadLeveler has had the following enhancements in its function:

- With APAR IY33664 and IY34168, the software now supports the use of the 64-bit AIX 5L Version 5.1 kernel.

Restriction: Checkpoint/restart support for 64-bit kernel is not available at this time.

- With APAR IY29622, LoadLeveler supports technical large pages as introduced in AIX 5L Version 5.1 Maintenance Level 2. This involves the selective use of large virtual and physical memory pages to back private data segments of a process. When specified, the user process heap, the main program BSS, and the main program data areas are backed by large pages. It is not supported to run the LoadLeveler daemons themselves with large pages. The functionality is exploited using the new LoadLeveler job command file keyword:

  ```
  large_page = <Y | M | N>
  ```

  Here, **Y** means use large page memory if available; otherwise use regular memory. The default option **M** means that it is mandatory to use large page memory, and **N** means not to use large page memory. Example 4-13 shows the new keyword in a LoadLeveler script.

Example 4-13 New LoadLeveler keyword: large_page

```
#!/bin/ksh
# @ output = myfile.out
# @ error = mytest.err
# @ notification = complete
# @ notify_user = lissy@roland.net
# @ requirements = ( Machine == "sp4n01e0" && LargePageMemory > 1000)
# @ preferences = ( TotalMemory > 2500 )
# @ large_page = Y
# @ initialdir = /home/lissy
# @ queue
LDR_CNTRL=LARGE_PAGE_DATA=Y ./myimportantprogramm
```

This example also shows the two new LoadLeveler variables LargePageMemory and TotalMemory. The first defines the amount of large

page memory the user wants, the second defines the total amount of memory. Both values are 64-bit integers, specifying the size in megabytes.

Tip: We recommend setting the VM_IMAGE_ALGORITHM to FREE_PAGING_SPACE_PLUS_FREE_MEMORY in your LoadL.config file. This allows the central manager to consider both the free physical and the free large page memory when deciding if a machine in the cluster has enough virtual memory to run a job step.

- The **llq** command is enhanced by giving additional information about the use of large pages, as shown in Example 4-14.

Example 4-14 llq command enhancements

```
lissy@mikesch: llq -l mikesch.4711.0
==================== Job Step mikesch.4711.0 ================================
Job Step Id: mikesch.4711.0
           Job Name: myjob
          Step Name: step5
  Structure Version: 10
              Owner: lissy
         Queue Date: Fri Oct  4 14:27:02 MET 2002
             Status: Running
   Execution Factor: 1
      Dispatch Time: Tue Oct  8 11:47:44 MET 2002
    Completion Date:
    Completion Code:
      User Priority: 50
       user_sysprio: 9999
      class_sysprio: 0
      group_sysprio: 0
    System Priority: 9860788
          q_sysprio: 9860788
      Notifications: Always
 Virtual Image Size: 1 kb
         Large Page: N
     Checkpointable: no
...
```

- The **llstatus** command has been enhanced to display information associated with the **large_page** keyword, as shown in Example 4-15 on page 94. Now, the total and free memory for both large page memory and regular memory are shown.

Example 4-15 llstatus command enhancements

```
lissy@mikesch: llstatus -l
Name                = mikesch
Machine             = mikesch
Arch                = R6000
OpSys               = AIX51
SYSPRIO             = (((0 -  (QDate /  10)) +  (ClassSysprio *  100)) +
(UserSysprio *  1000))
MACHPRIO            = ((0 -  LoadAvg) -  (10 *  Cpus))
VirtualMemory       = 16771848 kb
Disk                = 125864 kb
KeyboardIdle        = 2950
Tmp                 = 1012956 kb
LoadAvg             = 1.036880
ConfiguredClasses   =
AvailableClasses    =
DrainingClasses     =
DrainedClasses      =
Pool                =
FabricConnectivity  =
Adapter             =
Feature             =
Max_Starters        = 0
Total Memory        = 6399 mb
Memory              = 6400 mb
FreeRealMemory      = 5493 mb
LargePageSize       = 16.000 mb
LargePageMemory     = 0 kb
FreeLargePageMemory = 0 kb
PagesFreed          = 0
```

- The **llsummary** command is enhanced by now giving additional information about the large pages used.
- The **ll_get_data()** function, as defined in llapi.h of the LoadLeveler API, is enhanced so that the large page information of machines can be accessed by the following specifications:
  - LL_MachineLargePageSize64
  - LL_MachineLargePageCount64
  - LL_MachineLargePageFree64
  - LL_StepLargePage
- With APAR IY24116 and IY24117, the checkpoint/restart function is now supported. For more information about this, refer to *Workload Management with LoadLeveler*, SG24-6038.

- With APAR IY25275, a new configuration file keyword is introduced:

  ```
  NEGOTIATOR_CYCLE_TIME_LIMIT = number
  ```

  Where *number* specifies the maximum amount of time in seconds that the negotiator cycle will be allowed to continue. After the specified number of seconds, the negotiator cycle ends, even if there are more jobs to be considered for dispatch. The jobs do not get lost, instead they will be considered in the next subsequent negotiator cycle. The number specified must be a positive integer value or zero. If set to zero, the negotiator behaves as if the command was not set. This means, the negotiator will always consider all jobs for dispatch in one cycle.

Restriction: This keyword applies to the BACKFILL and GANG scheduler only.

- Another keyword was introduced in LoadLeveler with APAR IY25829. This keyword allows the specification of an alternative local directory where LoadLeveler keeps the special files used for UNIX domain sockets for communicating among LoadLeveler daemons running on the same machine. The keyword is:

  ```
  COMM = directory
  ```

  Where directory is the name of an existing directory. The default is /tmp. This keyword allows the administrator to choose a different file system than /tmp for these important files.

Important: If you change the COMM option, you must stop and restart LoadLeveler using **`llctl`**.

- To give administrators a finer granularity of integrating WLM policies into LoadLeveler, APAR IY32415 introduces a new keyword that can be locally different on each machine by integrating it in the LoadL.config.local file. The syntax of this keyword is:

  ```
  ENFORCE_RESOURCE_POLICY = hard | soft | shares
  ```

  Where **`hard`** indicates that Workload Manager (WLM) classes will be created with hard limits representing the percentage of step requested resources per total machine resources. **`Soft`** indicates that WLM classes will be created with soft limits representing the percentage of step requested resources per total machine resources. **`Shares`** indicates that WLM classes will be created with a resource share representing the step requested resources.

Note: The keyword is ignored if the ENFORCE_RESOURCE_USAGE keyword is not set.

4.9.2 Parallel Environment

Parallel Environment (PE) has been enhanced with the introduction of three APARs. For a detailed description of these features, refer to *IBM @server Cluster 1600 and PSSP 3.4 Cluster Enhancements*, SG24-6604. The three APARs delivered that extend the capability of this product are as follows:

- With APAR IY34726, PE is able to exploit the 64-bit kernel of AIX 5L Version 5.1.
- APAR IY32331 delivers enhanced support for technical large pages introduced in AIX 5L Version 5.1 Maintenance Level 2.
- The communication over a two-plane SP Switch2 environment is supported with APAR IY30344.

Note:

- MPI jobs can span PSSP 3.4 and PSSP 3.5.
- MPI jobs can run over mixed 32- and 64-bit AIX kernels. These nodes need to be at PSSP 3.5.
- Stand-alone support for 64-bit systems is provided with APAR IY34726.

4.9.3 Engineering and Scientific Subroutine Library and Parallel ESSL

Since its introduction in December 2001, Engineering and Scientific Subroutine Library (ESSL) and Parallel ESSL have undergone further improvements to exploit the new architecture of the pSeries p690/p670. Table 4-1 lists some of the improvements to ESSL together with the required APARs.

Table 4-1 Improvements to ESSL functions

Functional enhancement	APAR number
Improved DGEMM performance for small matrix sizes on p690, p670, and p655	PQ57448
Improved selected Level 1 BLAS performance on p690	PQ57481
Improved performance for DGEMV and DGER for power of 2 LDA on p690	PQ57570

Functional enhancement	APAR number
Improved DTRMM performance for special shaped problems	PQ57865
Improved performance on p690	PQ59873
Improved DAXPY full cache performance on POWER4	PQ63403
Improved performance on POWER4	PQ63390
Improved FFT performance for small lengths	PQ63401
Improved performance of Rank-K update subroutines on POWER4	PQ67105
Improved performance of short precision matrix add and subtract subroutines on POWER4	PQ67112
Improved performance of CGEMM and ZGEMM on POWER4	PQ67114

Table 4-2 lists an improvement to Parallel ESSL together with the required APARs.

Table 4-2 Improvement to Parallel ESSL

Functional enhancement	APAR number
Improved SMP performance for PDCFT3 and PSCFT3	PQ59854

Note:

- IBM recommends use of the latest available levels for ESSL and Parallel ESSL to fully exploit the speed and functionality of those libraries.
- For ESSL and Parallel ESSL, no explicit APAR is necessary for the support of the 64-bit kernel.
- ESSL is now supported with AIX 5L Version 5.2.

4.10 New hardware

The following additional hardware is supported in PSSP 3.5 and also supported in PSSP 3.4:

- p655
- p630
- p670
- Winterhawk 450 MHz
- SP Switch2 PCI-X Attachment Adapter
- A 19-inch frame for one SP Switch in a single rack
- A 19-inch frame for integrating up to two SP Switch2s in a single rack
- A 24-inch frame for the p655

These hardware additions are also be available for PSSP 3.4 through PTF patches to the product.

The new hardware details are discussed in Chapter 2, “New hardware” on page 13.

5

General Parallel File System 2.1

This chapter discussed the new features in General Parallel File System (GPFS). GPFS 2.1 now has 64-bit kernel exploitation, GPFS on RPD, and other new features. This chapter also contains the different GPFS implementations supported in the Cluster 1600, including the hardware supported by GPFS. Sample implementations are included.

This chapter discusses the following GPFS 2.1 feature:

- 64-bit kernel extensions

The following GPFS implementations are also discussed in this chapter:

- General Parallel File System on Virtual Shared Disk
- General Parallel File System on HACMP
- General Parallel File System on Linux
- General Parallel File System on RSCT peer domain

5.1 Introduction to General Parallel File System

The IBM General Parallel File System (GPFS) enables users shared access to files that can span multiple disk drives on multiple nodes. GPFS offers many of the standard UNIX file system interfaces, allowing most applications to execute without modification or recompiling. It also supports the UNIX file system utilities, so users can use the UNIX commands for ordinary file operations.

GPFS provides file system services to parallel and serial applications. GPFS allows parallel applications to share the same data, or different data spanning multiple disk drives attached to any nodes in the GPFS *nodeset*.

GPFS is a clustered file system defined over multiple nodes. The overall set of nodes over which GPFS is defined is known as a GPFS *cluster*. Within a GPFS cluster, the nodes are divided into one or more GPFS nodesets. A nodeset is a group of nodes that all run the same level of GPFS and operate on the same file system. It is possible for different GPFS versions to coexist in the same cluster.

Note: Three types of AIX-based GPFS documents are available now:

- GPFS on VSD:
  General Parallel File System for AIX 5L: PSSP Clusters Concepts, Planning, and Installation Guide, GA22-7899
- GPFS on RPD and GPFS on HACMP:
  General Parallel File System for AIX 5L: AIX Clusters Concepts, Planning, and Installation Guide, GA22-7895
- Prior to GPFS 2.1:
  IBM General Parallel File System for AIX: Concepts, Planning, and Installation, GA22-7453

5.1.1 What's new in General Parallel File System 2.1

The new features of GPFS 2.1 are as follows:

- Support for AIX 5L Version 5.1 with APAR IY33002 (GPFS on VSD), IY30258 (GPFS in an AIX-related environment), and PSSP 3.5.
- 64-bit kernel exploitation.
- Direct I/O capability for selected files:
  - The direct I/O caching policy bypasses the file cache and transfers data directly from disk into the user space buffer. Applications with poor cache hit rates or very large I/Os may benefit from the use of direct I/O.
  - The **mmchattr** command has been updated with the **-D** option for this support.

- The default changed to *use designation*.
  - The default use designation has been changed from *manager* to *client*.

Note: These changes were included in the **mmconfig** and **mmchconfig** commands.

  - You can list which node is currently assigned as the file system manager by issuing the **mmlsmgr** command. The **mmchmgr** command allows you to change the node that has been assigned as the file system manager.
- The terms to *install/uninstall GPFS quotas* have been replaced by the terms *enable/disable GPFS quota management*.
- For **atime** and **mtime** values, as reported by the **stat**, **fstat**, **gpfs_stat**, and **gpfs_fstat** calls, you can:
  - Suppress updating the value of **atime**.

    When suppressing periodic update, these calls will report the time the file was last accessed when the file system was mounted with the **-S no** option, or for a new file, the time the file system was created.
  - Display the exact value for **mtime**.

    The default is to periodically update the **mtime** value for a file system. If it is more desirable to display the exact modification times for a file system, specify the **-E yes** option.

Note: These changes were included in the **mmcrfs**, **mmchfs**, and **mmlsfs** commands.

- The GPFS documentation is no longer shipped on the product CD-ROM.
  - For the GPFS documents, refer to the following Web site:

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

    or

    http://www.ibm.com/shop/publications/order
  - Two types of the AIX-based GPFS documents are available: one is for GPFS on VSD and the other is for GPFS on RPD or on HACMP. For example, prior to GPFS 2.1, you referred to *IBM General Parallel File System for AIX: Concepts, Planning, and Installation*, GA22-7453. Now, for GPFS 2.1 on VSD, refer to the *General Parallel File System for AIX 5L: PSSP Clusters Concepts, Planning, and Installation Guide*, GA22-7899, and for GPFS 2.1 on RPD or on HACMP, refer to *General Parallel File*

System for AIX 5L: AIX Clusters Concepts, Planning, and Installation Guide, GA22-7895.

– The software can be installed in either an AIX cluster environment or a PSSP cluster environment. Therefore, two sets of man pages are shipped with the product and your MANPATH environment variable should point to the appropriate directory.

Important: GPFS includes new file system functions that are not usable in existing file systems until you authorize these changes by issuing **mmchfs -V**.

5.1.2 General Parallel File System cluster types

GPFS defines several cluster types, depending on the operating environment:

VSD environment The VSD/SP environment is based on the IBM Parallel System Support Programs (PSSP) product and the IBM Virtual Shared Disk (VSD) product. The boundaries of the GPFS cluster in the VSD environment depend on the switch type being used. For more information, refer to 5.3, "General Parallel File System on Virtual Shared Disk" on page 104.

HACMP environment The HACMP environment is created by the High Availability Cluster Multi-Processing for AIX/Enhanced Scalability (HACMP/ES). The boundaries of the GPFS cluster are maintained with the **mmcrcluster**, **mmaddcluster**, and **mmdelcluster** commands. 5.4, "General Parallel File System on HACMP" on page 108, specifies the HACMP environment.

Linux environment The Linux environment is based on the Linux operating system. The boundaries of the GPFS cluster in the Linux environment are maintained with the **mmcrcluster**, **mmaddcluster**, and **mmdelcluster** commands. For more information, refer to 5.5, "General Parallel File System on Linux" on page 112.

RPD environment A Reliable Scalable Cluster Technology (RSCT) peer domain was created by the RSCT subsystem of AIX 5L. In the RPD environment, the boundaries of the GPFS cluster are maintained with the **mmcrcluster**, **mmaddcluster**, and **mmdelcluster** commands. For more information, refer to 5.6, "General Parallel File System on RSCT peer domain" on page 114.

5.1.3 Advantages

GPFS is designed to provide a common file system for data shared among the nodes of the cluster. The characteristics of GPFS are as follows:

- It provides access to all GPFS data from all nodes of the cluster. It provides improved system performance not only by balancing the load across all disks to maximize their combined throughput, but also by increasing aggregate bandwidth of your file system, because multiple servers can access the file system with their own I/O path.
- It uses a token management system to provide data consistency, while allowing multiple independent paths to the same file by the same name from anywhere in the system. Even when nodes are down or hardware resource demands are high, it can find an available path to file system data.
- It maintains replicated data of metadata allowing continued operation when the paths to a disk, or the disk itself, is broken. This feature enables the fast recovery and the restoration of data consistency.
- While the GPFS file system is mounted, you can add or delete disks. You can also add new nodes without having to stop and restart the GPFS daemon except when using LAPI as the communication protocol.

5.2 64-bit kernel extensions

In GPFS 2.1, GPFS kernel extensions exist in both 32-bit and 64-bit forms. GPFS 2.1 also supports interoperability between 32-bit and 64-bit GPFS kernel extensions within a nodeset.

Note: 32-bit and 64-bit kernel extensions can coexist within a nodeset.

If you want to use 64-bit versions of the GPFS programming interfaces, you must recompile your code using the appropriate 64-bit options for your compiler. For more information, refer to 4.1, “64-bit compatibility” on page 70 and *General Parallel File System for AIX 5L: AIX Clusters Concepts, Planning, and Installation Guide*, GA22-7895 or *General Parallel File System for AIX 5L: PSSP Clusters Concepts, Planning, and Installation Guide*, GA22-7899.

In a 32-bit VSD environment, GPFS uses the security configured for PSSP. If this has not been properly configured, you may get GPFS errors and should turn security off. You must also turn off PSSP security prior to starting GPFS if any node in a nodeset is running the 64-bit kernel. That is, PSSP security is not supported in a 64-bit kernel environment. Example 5-1 on page 104 shows how to turn off PSSP security by using the **`mmchconfig`** command.

Example 5-1 Turning off PSSP security

```
root $ mmchconfig useSPSecurity=no -C sp6ns
mmchconfig: Command successfully completed
mmchconfig: 6027-1371 Propagating the changes to all affected nodes.
This is an asynchronous process.
```

5.3 General Parallel File System on Virtual Shared Disk

The implementation of GPFS in the VSD environment relies on VSD and RVSD. This section discusses how to run GPFS using VSD (RVSD). This GPFS environment is only supported on a Cluster 1600 where you have an SP Switch or SP Switch2 available. The only supported adapters are css0 or ml0.

Restriction: When running in a PSSP environment, GPFS use of the Low-Level Application Programming Interface (LAPI) in an SP Switch2 two-plane environment is unavailable. IBM has determined that further testing is necessary before making this support available in a production environment. When available, support will be provided with APAR# IY36170.

Figure 5-1 on page 105 shows the structure of such an environment. In this example, some disks are twin-tailed. The solid lines are the primary connections to the disks, and the dotted line is the backup connection.

The left and the middle nodes are using Concurrent Logical Volume Manager (CLVM) to access the same disk. In this case, when one of these systems fail, the recovery is much faster.

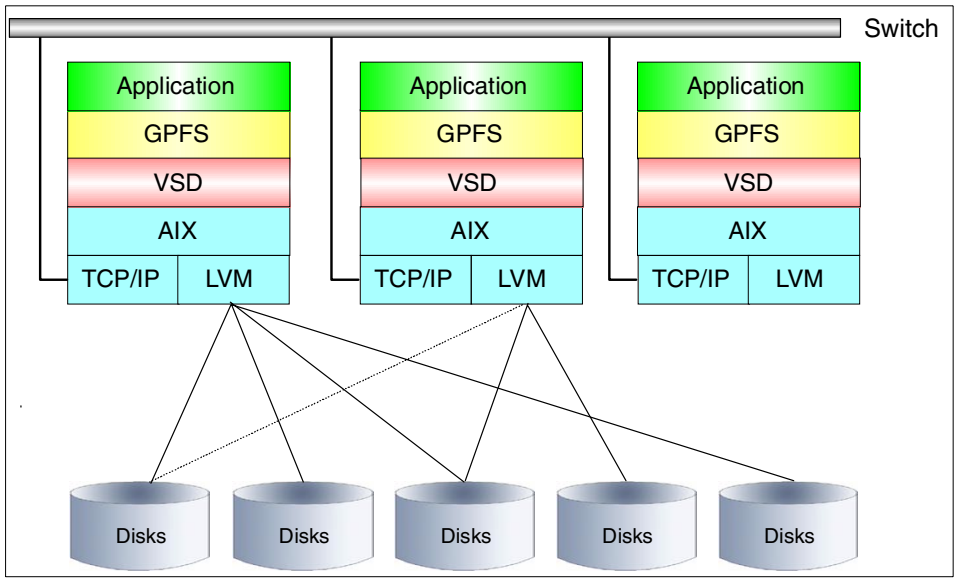


Figure 5-1 General Parallel File System on Virtual Shared Disk

5.3.1 Prerequisites

The GPFS version using VSD that you are able to use is based on the PSSP version you have installed on your nodes. All nodes for such a GPFS cluster must be on the same level. Table 5-1 shows the requirements for GPFS 1.4, GPFS 1.5, and GPFS 2.1.

Table 5-1 GPFS on VSD prerequisites

GPFS 1.4	**GPFS 1.5**	**GPFS 2.1**
AIX 4.3.3 (with APAR IY12051)	AIX 4.3.3 (with APAR IY12051) or AIX 5L Version 5.1	AIX 5L Version 5.1 (with APAR IY33002)
PSSP 3.2	PSSP 3.4	PSSP 3.5
Min. nodes 3	Min. nodes 3	Min. nodes 3

5.3.2 Configuration

Before you can setup your GPFS cluster, you must have all necessary filesets installed, and VSD must be configured.

We list only the main configuration steps here. For a detailed description, see *GPFS on AIX Clusters: High Performance File System Administration Simplified*, SG24-6035 or the PSSP manuals *PSSP for AIX: Managing Shared Disks*, SA22-7349 or *General Parallel File System for AIX 5L: PSSP Clusters Concepts, Planning, and Installation Guide*, GA22-7899.

VSD (RVSD) setup

To configure RVSD, complete the following steps:

1. Check if all necessary VSD and GPFS filesets are installed on your control workstation (CWS) and on your nodes.

   On the CWS:

   – VSD:

   - ssp.basic
   - ssp.css
   - ssp.sysctl
   - vsd.cmi
   - vsd.sysctl
   - vsd.vsdd

   Note: The preceding VSD filesets must be installed on the CWS. When installing these filesets, the following filesets are needed as their prerequisite filesets:

   - vsd.hsd
   - vsd.rvsd.rvsdd
   - vsd.rvsd.scripts

   – GPFS:

   - mmfs.gpfs.rte

   Note: The GPFS daemon will not be available on the CWS when you only install mmfs.gpfs.rte, but you can manage GPFS by using all the GPFS commands and SMIT menus.

On each node in the GPFS nodeset and the VSD server nodes:

- VSD:
  - ssp.basic
  - ssp.css
  - ssp.sysctl
  - vsd.cmi
  - vsd.sysctl
  - vsd.vsdd
  - vsd.rvsd.hc
  - vsd.rvsd.rvsdd
  - vsd.rvsd.scripts

  Note: The preceding VSD filesets must be installed on all nodes in your GPFS nodeset and on any node serving as a VSD server. When installing these filesets, the vsd.hsd fileset is needed as their prerequisite fileset.

- GPFS:
  - mmfs.base.cmds
  - mmfs.base.rte
  - mmfs.gpfs.rte
  - mmfs.msg.en_US
  - mmfs.gpfsdocs.data

  Note: You do not need to install mmfs.gpfsdocs.data on all nodes if the manual pages are not desired.

2. Add your kerberos principal (root.admin) to the /etc/sysctl.acl, /etc/sysctl.vsd.acl (VSD), and /etc/sysctl.mmcmd.acl (GPFS) files on the CWS and copy them to all nodes.
3. Create a dummy VSD for all the nodes in a GPFS cluster. An LV with 1 PP is sufficient. It can be deleted when at least one GPFS file system is created and running.
4. Start the dummy VSD.
5. Start the RVSD daemon, first on CWS, and then on each node in the GPFS cluster.

GPFS setup

To set up GPFS, complete the following steps:

1. Make sure that the RVSD daemon is running on your CWS and the nodes before you continue configuring GPFS.
2. Create a GPFS nodeset.

Note: The **mmcrcluster**, **mmaddcluster**, **mmdelcluster** commands are not supported in the SP (VSD) environment. Therefore, you cannot use these commands to make a GPFS cluster.

3. Start the GPFS daemon.
4. Create the GPFS volume groups and VSDs.
5. Create the GPFS file systems.
6. Mount the GPFS file systems.

Attention: All the nodes in the GPFS cluster must have the same buddy buffer size. Otherwise, the GPFS file systems cannot be mounted.

Note: If you use LAPI as the communication protocol, you must stop the GPFS daemon on all nodes in the nodeset before adding a node or deleting a node.

5.4 General Parallel File System on HACMP

Since the availability of GPFS 1.4, there is no SP Switch or VSD requirement for GPFS. In such an environment, we can use HACMP/ES. It is a requirement except for GPFS Version 2.1. For information about how to implement a GPFS environment without HACMP/ES, see 5.6, “General Parallel File System on RSCT peer domain” on page 114.

Figure 5-2 on page 109 shows the structure of GPFS on an HACMP environment. As shown in this example, all disks must be visible to all nodes in the cluster (nodeset). All nodes use Concurrent LVM (CLVM) to access the same disk or disks. However, you do not have to have HACMP/ES CLVM installed. GPFS just makes use of the AIX CLVM capability.

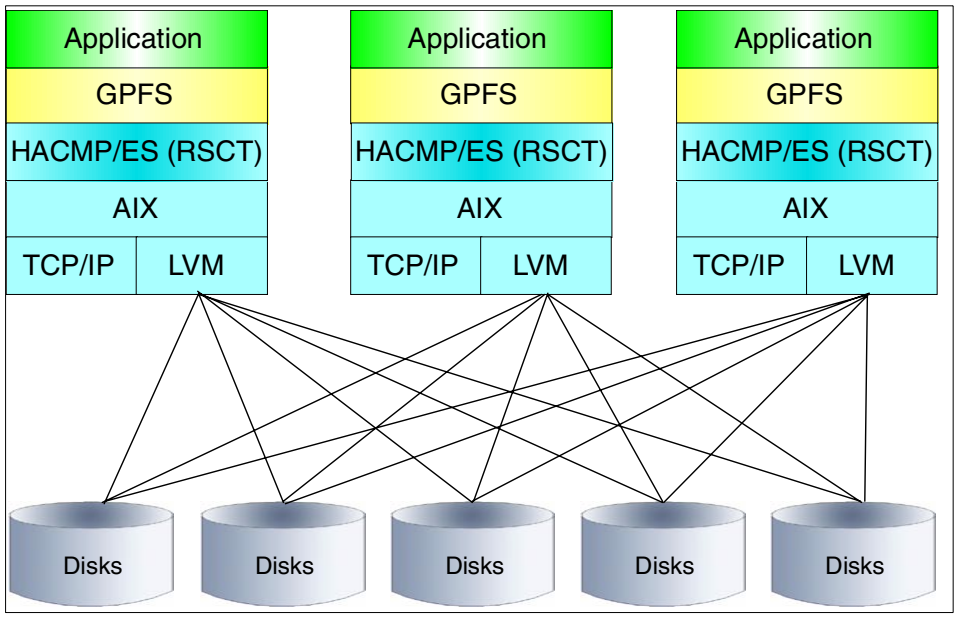


Figure 5-2 General Parallel File System on HACMP

Figure 5-3 on page 110 shows the relationship between the subsystems of the HACMP/ES cluster group and the GPFS subsystem. The implementation of GPFS in the HACMP environment does not make use of the capabilities of HACMP/ES for high availability. HACMP/ES provides the operation environment that GPFS requires for the subsystems of RSCT. GPFS, event management, and the HACMP/ES cluster manager are clients of Group Services. There is no interaction between the subsystems of the HACMP/ES cluster group and the GPFS subsystem.

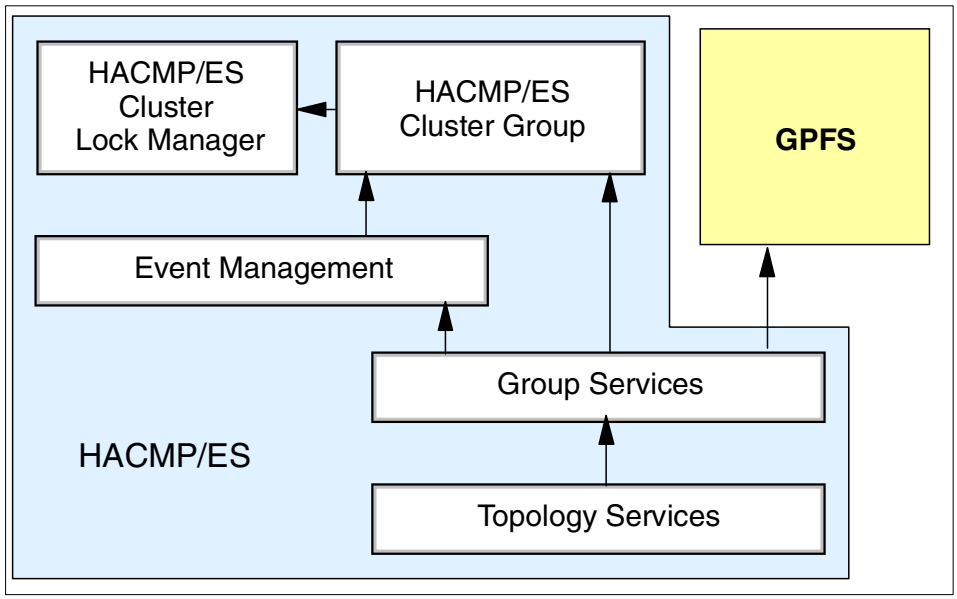


Figure 5-3 Relationship between HACMP and GPFS

5.4.1 Prerequisites

GPFS based on HACMP can be used for GPFS Versions 1.4, 1.5, and 2.1. The requirements for these versions are listed in Table 5-2.

Table 5-2 GPFS on HACMP prerequisites

GPFS 1.4	GPFS 1.5	GPFS 2.1
AIX 4.3.3 (with APAR IY12051)	AIX 4.3.3 (with APAR IY22024) or AIX 5.1 (with APAR IY21957)	AIX 5L Version 5.1 (with APAR IY30258)
HACMP/ES 4.4 or later	HACMP/ES 4.4.1 or later	HACMP/ES 4.4.1 or later
SSA disks	SSA disks or Fibre Channel disks	SSA disks or Fibre Channel disks
Min. 100 MB network or better	Min. 100 MB network or better	Min. 100 MB network or better
Min. nodes 2 with SSA disks	Min. nodes 2 with SSA disks Min. nodes 3 with Fibre Channel disks	min. Nodes 2 with SSA disks min. Nodes 3 with Fibre Channel disks

Note: Be sure to obtain the latest service level for all required software at the following URL:

http://techsupport.services.ibm.com/server/fixes

5.4.2 Configuration

A design where you use both HACMP functionality and GPFS can become very complex. We focus here on what must be done to get GPFS running.

HACMP setup

Before you can set up GPFS, you must configure HACMP. We list the main steps here. For a detailed description, see *General Parallel File System for AIX 5L: AIX Clusters Concepts, Planning, and Installation Guide*, GA22-7895 or *GPFS on AIX Clusters: High Performance File System Administration Simplified*, SG24-6035.

To set up HACMP, complete the following steps.

1. Set up your SSA or Fibre Channel cabling to your disks.
2. Configure the cluster topology.

   For GPFS, you have to define a network with just one (service) adapter per node to HACMP (no standby and no boot).
3. Start HACMP.

GPFS setup

To set up GPFS, complete the following steps:

1. Make sure that all appropriated HACMP subsystems (on all nodes) are running before you continue configuring GPFS.
2. Create the GPFS cluster.
3. Create a nodeset in your GPFS cluster.
4. Start the GPFS daemon.
5. Create the SSA- or Fibre Channel-based volume groups and logical volumes.

   This can become a time consuming step because all volume groups and logical volumes must be known by your nodeset.
6. Create the GPFS file systems.
7. Mount the GPFS file systems.

Note: In an HACMP environment, you cannot protect your file system against disk failure by mirroring data at the LVM level. You must use GPFS replication or RAID devices to protect your data.

5.5 General Parallel File System on Linux

There are two types of disk connectivity you can use when running GPFS in a Linux environment:

- Directly Attached Model
- Network Shared Disk Model

For a detailed description of GPFS on Linux, see *IBM General Parallel File System for Linux: Concepts, Planning, and Installation*, GA22-7844, or see *Linux Clustering with CSM and GPFS*, SG24-6601.

Directly Attached Model

The notion of direct attach is for the disk connection to the nodes. We say a configuration uses direct attachment only when all the nodes from a nodeset have direct access to the disks, as with Fibre Channel disks. A sample of such a configuration is shown in Figure 5-4 on page 113.

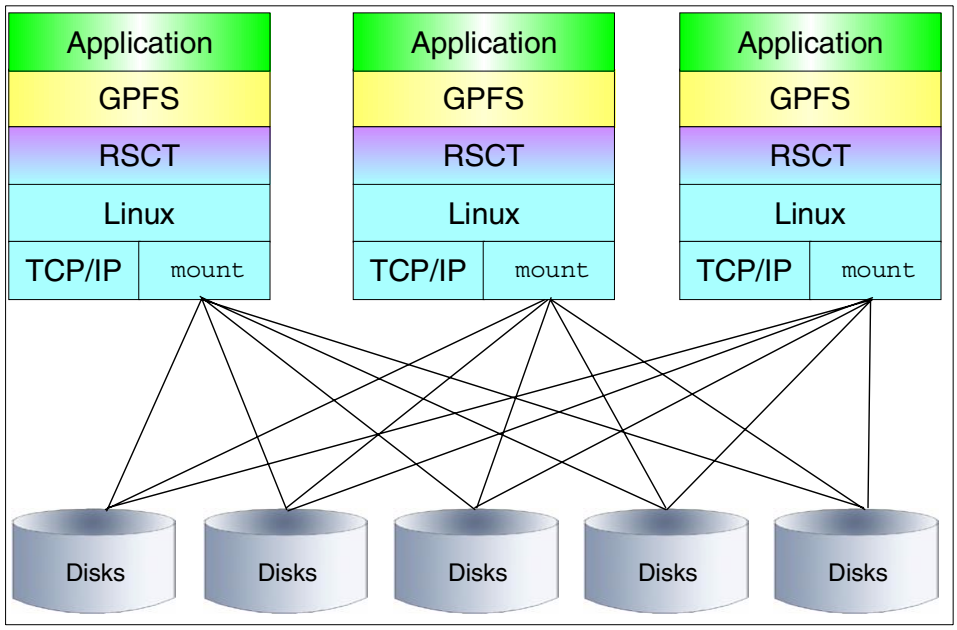


Figure 5-4 General Shared File System on Linux (directly attached)

Network Shared Disk model

We say we have a Network Shared Disk (NSD) configuration when only one or two nodes are directly connected to some disks if we implement redundancy. The other nodes use a communication network to connect to this first node to get access to the disks. This is similar to VSD in AIX. A sample of such a configuration is shown in Figure 5-5 on page 114.

Because this solution is network intensive, it is recommended to use networks, such as Gigabit Ethernet or Myrinet, that are capable of supporting high amounts of traffic.

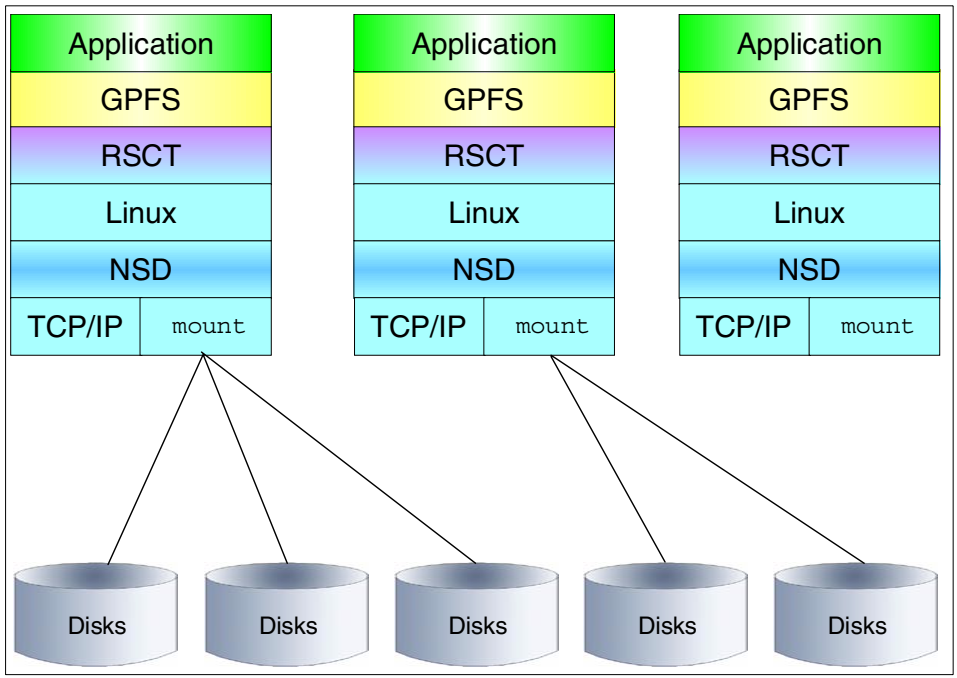


Figure 5-5 General Shared File System on Linux (NSD)

5.6 General Parallel File System on RSCT peer domain

The RSCT subsystem provided by AIX 5L enables you to configure a GPFS cluster in an RPD environment. That is, the RSCT subsystem of AIX 5L removes the existing HACMP prerequisites in the non-SP Switch environment. The RDP-based GPFS is available on GPFS Version 2.1. For more detailed information about RSCT, refer to Chapter 3, “Reliable Scalable Cluster Technology overview” on page 49.

GPFS using the RSCT subsystem of AIX 5L is shown in Figure 5-6 on page 115. For such an environment, you have the same disk requirements as for GPFS on HACMP. This means all nodes must have concurrent access to your disks. The Concurrent LVM (CLVM) functionality of AIX is only required here. You do *not* have to use CLVM.

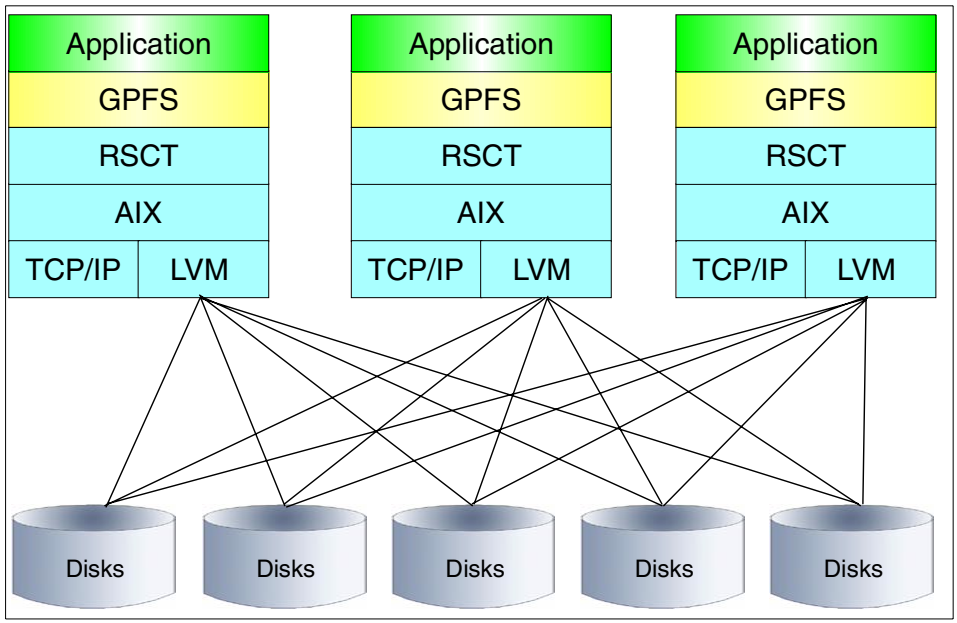


Figure 5-6 General Shared File System on RPD

Figure 5-7 on page 116 shows the relationship between the RPD subsystems and the GPFS subsystem. The implementation of GPFS in the RPD environment does not make use of the capabilities of RPD for high availability. The peer domain provides the operation environment that GPFS requires for the subsystems of RSCT. There is no interaction between the RMC subsystems and the GPFS subsystem. GPFS is just a client of Group Services.

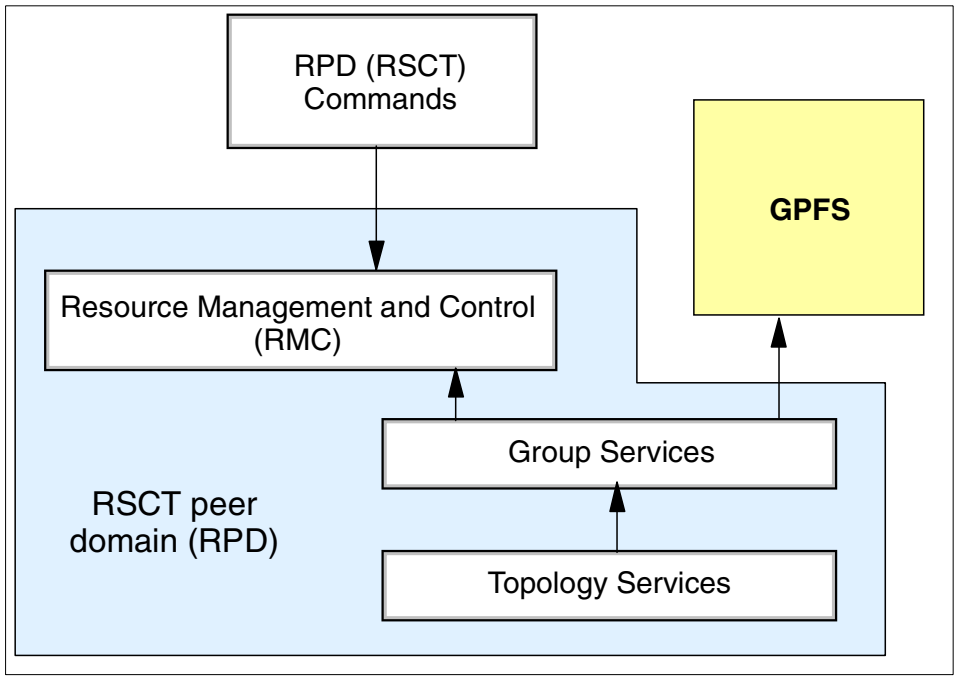


Figure 5-7 Relationship between RPD and GPFS

5.6.1 Prerequisites

If you are planning to use GPFS 2.1 based on RPD, the following prerequisites are needed:

- AIX 5L Version 5.1 with APAR IY32508
- SSA disks or Fibre Channel disks
- A 100 Mbps Ethernet network or faster networks
- Shared disks
  - A minimum of two nodes for SSA disks.
  - A minimum of three nodes for Fibre Channel disks

Note: An RSCT peer domain can have a maximum of 32 nodes. The GPFS cluster size on RPD depends on the disk technology, as GPFS on HACMP does. For example, a maximum of 8 nodes are available with SSA disks and a maximum of 32 nodes with Fibre Channel disks.

Before your can configure your GPFS cluster, you must have all the necessary filesets for GPFS and RSCT installed. The following lists all the RSCT filesets you must have installed on all your GPFS nodes:

- rsct.basic.rte
- rsct.compat.basic.rte
- rsct.compat.clients.rte
- rsct.core.auditrm
- rsct.core.errm
- rsct.core.fsrm
- rsct.core.hostrm
- rsct.core.rmc
- rsct.core.sec
- rsct.core.sr
- rsct.core.utils

5.6.2 Configuring General Parallel File System on RSCT peer domain

In this section, we briefly describe how to configure a new GPFS cluster. If you want to know how to add a node, see 5.6.3, “Adding a node” on page 120, or how to delete a node, see 5.6.4, “Deleting a node” on page 121 in you GPFS cluster.

Before you start to configure a GPFS cluster, you must first configure an RSCT peer domain. For more information about configuring the RSCT peer domain, refer to the *IBM RSCT for AIX: Guide and Reference*, SA22-7889.

To configure a new GPFS cluster in the RPD environment, complete the following steps, which include the configuration of a RSCT peer domain:

1. Establish the initial trust between each node that will be in the peer domain. The node from which you will issue the **mkrpdomain** command is called the originator node.

   ```
   preprpnode originator_node
   ```

   or

   ```
   preprpnode -f node.list
   ```

2. Create a new peer domain definition that consists of a peer domain name, the list of nodes, and the UDP port numbers for the Topology Services and the Group Services daemons. You can issue the command on the originator node:

   ```
   mkrpdomain domain_name node_name [node_name ...]
   ```

   or

   ```
   mkrpdomain -f node.list domain_name
   ```

3. Bring the peer domain online:

   ```
   startrpdomain domain_name
   ```

Attention: The RSCT peer domain must be operational before configuring GPFS.

Example 5-2 shows the status of the peer domain.

Example 5-2 Peer domain

```
root $ lsrpdomain
Name        OpState        RSCTActiveVersion MixedVersions TSPort GSPort
sp4rpdomain Pending online 2.2.1.20          No            12347  12348

root $ lsrpdomain
Name        OpState RSCTActiveVersion MixedVersions TSPort GSPort
sp4rpdomain Online  2.2.1.20          No            12347  12348

root $ lsrpnode
Name     OpState RSCTVersion
sp4n33e0 Online  2.2.1.20
sp4n01e0 Online  2.2.1.20
sp4n17e0 Online  2.2.1.20
```

Now, your RPD configuration is done. Next, you configure GPFS.

4. Create and edit a GPFS node file. You have to specify the node name of the adapter you want to use as a communication device. Example 5-3 shows the contents of a GPFS node file using each Ethernet adapter as their communication devices.

Example 5-3 Contents of the GPFS node file

```
root $ cat /tmp/gpfs.allnodes
sp4n01e0
sp4n17e0
sp4n33e0
```

5. Create a new GPFS cluster. You must specify **rpd** as the type of cluster. Example 5-4 shows how to use the **mmcrcluster** command to create a GPFS cluster on RPD. We defined the primary GPFS configuration data server name only.

Example 5-4 Creating a GPFS cluster

```
root $ mmcrcluster -t rpd -n /tmp/gpfs.allnodes -p sp4n01e0

root $ mmconfig -n /tmp/gpfs.allnodes -A -C sp4ns -M 65000 -p 512M
```

6. Create a new GPFS nodeset. The **mmconfig** command is shown in Example 5-5. You can specify maxFilesToCache and pagepool. In a two-node nodeset, we recommend enabling single-node quorum at this step. The specification of single-node quorum allows the remaining node in a two-node nodeset to continue functioning in the event of the failure of the peer node. Example 5-5 shows how to specify single-node quorum by using the **-U** option.

Example 5-5 Configuring a GPFS nodeset

```
root $ mmconfig -n /etc/gpfs.allnodes -A -C p690ns -M 65000 -p 512M -U yes
mmconfig: Command successfully completed
mmconfig: 6027-1371 Propagating the changes to all affected nodes.
This is an asynchronous process.
```

7. Start the subsystem for GPFS:

   ```
   mmstartup -C nodesetid
   ```

8. Create VGs and LVs for GPFS using the **mmcrlv** command:

   ```
   mmcrlv -F disk_descriptor_file
   ```

   Example 5-6 shows a sample of the disk descriptor file.

Example 5-6 Disk descriptor file

```
root $ cat /etc/sp4dd
hdisk2:sp4n01e0:sp4n33e0:dataAndMetadata:
hdisk3:sp4n33e0:sp4n17e0:dataAndMetadata:
hdisk4:sp4n17e0:sp4n01e0:dataAndMetadata:
hdisk5:sp4n01e0:sp4n17e0:dataAndMetadata:
hdisk6:sp4n33e0:sp4n01e0:dataAndMetadata:
```

9. Make a file system for GPFS. Example 5-7 shows how to make a file system in an GPFS nodeset.

Example 5-7 Making a file system

```
root $ mmcrfs /gpfs0fs gpfs0fs -F /etc/sp4dd -A yes -v no -M 2 -R 2 -C sp4ns

GPFS: 6027-531 The following disks of gpfs0fs will be formatted on node
sp4n01e0:
    gpfs0lv: size 8880128 KB
    gpfs1lv: size 8880128 KB
    gpfs2lv: size 8880128 KB
    gpfs3lv: size 8880128 KB
    gpfs4lv: size 8880128 KB
GPFS: 6027-540 Formatting file system ...
Creating Inode File
Creating Allocation Maps
```

```
Clearing Inode Allocation Map
Clearing Block Allocation Map
Flushing Allocation Maps
GPFS: 6027-572 Completed creation of file system /dev/gpfs0fs.
mmcrfs: 6027-1371 Propagating the changes to all affected nodes.
This is an asynchronous process.
```

10. Mount the GPFS file system.

5.6.3 Adding a node

To add a node to the RSCT peer domain and the GPFS cluster, complete the following steps:

1. Prepare the security environment on the node you want to add to the peer domain. The node from which you issue the **addrpnode** command is called the originator node.

   ```
   preprpnode originator_node
   ```

2. Add the node to the peer domain on the originator node:

   ```
   addrpnode node_name
   ```

   Example 5-8 shows the **lsrpnode** command output after adding the node.

Example 5-8 lsrpnode command output

```
root $ lsrpnode
Name     OpState RSCTVersion
sp4n33e0 Online  2.2.1.20
sp4n01e0 Online  2.2.1.20
sp4n05e0 Offline 2.2.1.20
sp4n17e0 Online  2.2.1.20
```

Note: If the configuration resource manager (ConfigRM) is not running correctly on each node of the peer domain, you cannot add the node to the peer domain.

3. Bring the offline node online from any node in the current peer domain:

   ```
   startrpnode node_name
   ```

   or

   ```
   startrpdomain domain_name
   ```

4. Add the node to the GPFS cluster. The **mmaddcluster** command allows the node to import the existing VGs information at this time.

   ```
   mmaddcluster node_name
   ```

5. Add the node to the GPFS nodeset. The **mmaddnode** command delivers the existing file systems information to the node.

   ```
   mmaddnode -C nodesetid node_name
   ```

6. Issue the **mmstartup** command to start GPFS on the new node.
7. Mount the file systems on the new node.

5.6.4 Deleting a node

To delete a node from the GPFS cluster and the RSCT peer domain, complete the following steps:

1. Issue the **mmshutdown** command on the node to be deleted.
2. Delete the node from the nodeset:

   ```
   mmdelnode -C nodesetid node_name
   ```

3. Delete the node from the cluster:

   ```
   mmdelcluster node_name
   ```

4. To remove the node from a peer domain, take it offline from any online node:

   ```
   stoprpnode node_name
   ```

5. Remove the node from a peer domain:

   ```
   rmrpnode node_name
   ```

5.6.5 Deleting the GPFS cluster and the RSCT peer domain

To change the cluster environment, you might need to delete the GPFS cluster and the RSCT peer domain. The following are some reasons why you would delete the cluster, nodeset, or the file system or peer domain, or both:

- The default for clusters and nodesets for an AIX-based GPFS is 32. GPFS on HACMP or GPFS on RPD has a physical limit of 32 nodes when using Fibre Channel disks and 8 nodes when using SSA disks. If you want to use more than 32 nodes or 8 nodes, you must use the VSD/SP environment. GPFS on VSD has cluster limit of 128 nodes.

  If you are already using VSD, but you used the default during setup, you have the same problem as mentioned above.

- Maximum metadata replicas and data replicas cannot be changed after it is set. If you need metadata and data replication, two copies of metadata and data blocks must be specified at file system creation time by using the appropriate options. These values can be overridden by a system call when the file has a length of 0 and can be changed with recreation of the file system.

- The size of data blocks cannot be changed without recreating the file system. It must be specified at file system creation time using the **-B** option.

There are also other minor reasons:

- You cannot change a nodeset identifier once it is set. So, if you want to change the nodeset identifier, you have to delete the GPFS nodeset.
- The device name of the file system cannot be changed at a later time.

For more information, refer to “Planning for General Parallel File System” on page 192.

To delete the GPFS cluster and the RSCT peer domain, complete the following steps:

1. Umount all the file systems in the GPFS cluster.
2. Delete all the file systems in the GPFS cluster:

   `mmdelfs` *`filesystem_name`*
3. Stop the GPFS daemons running on each node in the cluster:

   `mmshutdwon -a`
4. Delete all the nodes in the GPFS cluster:

   `mmdelnode -C` *`nodesetid`* `-a`

Attention: If you are in the VSD/SP environment at this time and want to move to the RPD environment, additional steps are required:

a. Delete all the files except mmfs.log file under the /var/mmfs/gen directory on each node in the GPFS cluster.

b. Delete the SDR information about GPFS:

   `SDRDeleteFile` *`mmsdrfs2`*

5. Delete the GPFS cluster:

   `mmdelcluster -n` *`gpfs_node_list`*
6. Delete the RSCT peer domain:

   `rmrpdomain` *`domain_name`*

6

Coexistence, migration, and integration

This chapter discusses PSSP and related software coexistence in a Cluster 1600 and provides information that must be considered before migration. Migration scenarios are discussed in this chapter as well. Detailed explanations about coexistence and migration activities can be found in the following guides:

- *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281
- *PSSP for AIX: Installation and Migration Guide*, GA22-7347

This chapter contains the following sections:

- Software coexistence
- Considerations for migration
- Migration related information about Cluster 1600: Hardware, 64-bit kernel support, and Parallel System Support Program and General Parallel File System
- Migration scenarios
- Integration of SP-attached servers managed through SAMI, CSP, and HMC protocols
- Migration tips

6.1 Software coexistence

The support of multiple versions or levels of hardware or software building blocks, or both, in a single system is called coexistence. This is an important feature that allows:

- Running part of the system on current software levels, while upgrading and testing new software versions.
- Reduction in the length of the maintenance window, because it is possible to upgrade a component of the system without disturbing the operation of other components.

Coexistence in PSSP enables one-by-one node migration or staged migration from previous PSSP levels to PSSP 3.5, while portions of the cluster continue to operate. In partitionable systems, it is possible to have nodes installed with different PSSP levels in the same system partition. However, some of the PSSP-related licensed programs, such as Parallel Environment (PE), are restricted in a mixed system partition.

Important: The control workstation (CWS) has to be running at the highest PSSP and AIX levels in the system. The boot/install server must be on the highest level of PSSP and AIX that it is to serve.

The coexistence limitations are as follows:

- PSSP 3.5 is supported only on AIX 5L 5.1, Maintenance Level 03.
- Coexistence of PSSP 3.5 on the CWS and PSSP 3.1.1 on nodes is not supported.
- GPFS 2.1 does not interoperate with earlier releases of GPFS in the same nodeset.
- 32-bit and 64-bit applications may coexist within a GPFS nodeset. However, if any node is running the 64-bit kernel, PSSP security may not be used.
- To use PSSP security, all the nodes in a GPFS nodeset must run the 32-bit version of the kernel.
- LAPI can not run in an environment mixed with different PSSP levels.
- KLAPI can not run in an environment mixed with different PSSP levels.
- Different switch types can not coexist in the same Cluster 1600 system.

Important: IBM intends to provide support for AIX 5L Version 5.2 with PSSP 3.5 in a Cluster 1600 in 2003.

Table 6-1 illustrates the possible coexistence of PSSP and related software levels.

Table 6-1 Coexistence levels for PSSP and related software

PSSP	AIX	GPFS	HACMP	RSCT	LL	PE
3.2	4.3.3	1.2/1.3/1.4	4.3, 4.4, 4.4.1	1.2	2.2	3.1
3.4	4.3.3	1.3/1.4/1.5	4.4.1	1.2.1	2.2	3.1
3.4	5.1 (with RSCT)	1.5	4.4.1, 4.5	2.2	3.1	3.2
3.5	5.1 (with RSCT)	1.5/2.1	4.4.1, 4.5	2.2	3.1	3.2

Other coexistence matrixes can be found in *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.

6.2 Considerations for migration

Before planning the migration, you need to understand your present system configuration and what considerations led you to this configuration. Review your overall system goals and plan a migration with them in mind. It is possible that migration of one part of the system may require the upgrade or migration of other parts as well.

Check the PSSP coexistence and migration scenarios in Table 6-2 on page 130 before starting any migration activities.

This chapter does not cover all migration steps. However, it provides some guidance for planning necessary migration activities. For detailed migration instructions, refer to the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

Important: Before applying any new software to your system, check the *Read This First* document. The latest version of PSSP documentation can be found at the following URL:

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

6.2.1 Hardware

In the following section, we provide information to keep in mind before changing the hardware configuration in your Cluster 1600.

Upgrading existing hardware or introducing new hardware into the Cluster 1600 may require software upgrades as well. System firmware and microcode upgrades may be necessary when adding new hardware to the cluster or when installing new software levels.

To get status information about the installed microcode, and to find out whether the system needs an upgrade, enter:

```
spsvrmgr -G -r status all
```

Attention: The old 66 MHz Power2 MCA nodes in one of the Cluster 1600 systems does not give back any information as a result of the `spsvrmgr -G -r status all` command.

For example, to upgrade from the SP Switch to the SP Switch2, we have to change the adapter in the nodes and install at least PSSP 3.2 for SP Switch2 support.

If the control workstation and all of the nodes are running PSSP 3.4 or later, we have optional SP Switch2 connectivity. We can connect a selection of nodes to the SP Switch2 and leave some nodes off the switch.

Note: Notice that nodes prior to the 332 MHz SMP and the SP-attached servers S70 and S7A do not support the SP Switch2.

While changing the switch configuration, the *primary* and *primary backup* node roles may be moved between nodes. Using PSSP 3.5, we can enable or disable nodes from serving as primary or primary backup nodes. For more information, refer to 4.3, “Eprimary modifications” on page 73.

When adding pSeries p670 and p690 servers to the Cluster 1600, new filesets must be installed and copied to the CWS into the appropriate lppsource directory. These files are Java130.xml4j.* and openCIMOM*. For details, refer to Chapter 2 in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

In PSSP 3.5, it is possible to expand a switchless Cluster 1600 system with new attached servers even if you do not have available switch ports in the existing frame. For this, we have to change the force_non_partitionable parameter either in the SP site environment SMIT panel or with the **spsitenv** command. For a detailed explanation about system partitioning, refer to *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281 and Chapter 16 of the *PSSP for AIX: Administration Guide*, SA22-7348. Details about hardware reconfiguration can be found in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

6.2.2 Direct migration

We differentiate between two kinds of PSSP migrations. For example, direct migration is migrating PSSP from 3.4 to 3.5 and AIX from 4.3.3 to 5.1 on a node in one step. When PSSP 3.1.1 is installed on a node, it is not possible to upgrade to PSSP 3.5 in a single step. In this case, an intermediate software level of PSSP must be installed prior to PSSP 3.5. The migration path could be more complicated if we have other applications based on PSSP, such as LoadLeveler or GPFS.

Attention: Any PSSP 3.1.1 node migrations must be done before migrating the CWS to PSSP 3.5 because PSSP 3.5 is not supported in coexistence with PSSP 3.1.1. In 3.5, PSSP 3.1.1 was removed from the SMIT panels.

AIX migration from Version 4.3.3 to 5.1 needs careful planning because LoadLeveler and PE versions cannot coexist and interoperate between AIX levels. For more details, see Figure 6-1 on page 125.

6.2.3 AIX

Let's investigate AIX 64-bit kernel support and JFS2 file systems from a coexistence and migration point of view.

64-bit kernel support

Before trying to run a 64-bit kernel on a node, ensure that both the processor and all the adapters in the node are capable of running in 64-bit mode. Refer to Appendix D, "AIX device drivers reference" on page 199.

Nodes can be running in 32- or 64-bit kernel mode and coexist with PSSP 3.4 nodes using a 32-bit kernel. It is possible to activate the 64-bit kernel on an AIX system after initial installation. For detailed instructions, see Example 4-1 on page 71.

Note: 64-bit applications for AIX 4.3.3 need to be recompiled to run on 64-bit AIX 5.

When they are called from within a 64-bit process, 32-bit functions can fail or cause failures. Because of this, all applications that link to 32-bit libraries must also be linked in 32-bit mode. Refer to the following list for combinations that work:

- 32-bit kernel, 32-bit application, 32-bit libraries
- 64-bit kernel, 32-bit-bit application, 32-bit libraries

- 32-bit kernel (AIX 5L), 64-bit application, 64-bit libraries
- 64-bit kernel, 64-bit application, 64-bit libraries

JFS and JFS2

A migration install from AIX 4.3.3 to AIX 5L Version 5.1 will activate the 64-bit kernel, but all the file systems remain JFS. This does not happen automatically. You must select a 64-bit image to get a 64-bit kernel on migration. Newly created file systems will be JFS2. If we activate the 32-bit kernel on AIX 5L Version 5.1, and then we create a file system, this will be JFS by default. These machines will contain mixed-type file systems. There might by some operations where AIX 5L Version 5.1 running 32-bit kernel and JFS2 file systems can cause a performance slowdown.

6.2.4 Parallel System Support Program

VSD communication between nodes installed with PSSP 3.2, 3.4, and 3.5 will run over IP even if it is set to KLAPI. After migrating all nodes to PSSP 3.5, the communication will resume to KLAPI.

VSD/RVSD supports 32- and 64-bit coexistence between nodes running PSSP 3.5. However, coexistence with previous levels requires that all nodes run in 32-bit mode.

When migrating to PSSP 3.5, and if we want to use VSD, it is necessary to install an additional AIX fileset, bos.clvm.enh, that is not part of the default AIX installation.

To be able to use RVSD in a mixed system partition, stop RVSD on all nodes and use the **`rvsdrestrict`** command to specify the functionality level. Then restart the RVSD instance on the nodes. After the migration of the last node, another restart is necessary in order to pick up the new function. For instructions on stopping RVSD, see Example 6-11 on page 140. The RVSD restart can be done by running the **`ha_vsd reset`** command. For node migration details, refer to Chapter 11 in *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.

Attention:

- IBM no longer supports PSSP 3.1.1.
- IBM will support PSSP 3.2 until the end of 2002.
- IBM will support PSSP 3.4 until AIX 4.3.3 support ends.

6.2.5 General Parallel File System

The CWS must run GPFS 2.1 if any of the nodes will be installed with this version.

The migration of the nodes must be done at the same time for all nodes in the nodeset. To activate the file system with the new function, run the **`mmfsch -V`** command.

GPFS Version 1.5 depends on RVSD 3.4 level of function. This means that if we have PSSP 3.5 installed on the nodeset, we must run RVSD with the previous functionality. The ability to configure this is provided with the **`rvsdrestrict`** command.

Attention: GPFS Versions 1.3 and 1.4 are supported until the end of 2002, and GPFS Version 1.5 is supported until the end of 2003.

6.2.6 LoadLeveler

The LoadLeveler central manager node must be migrated to PSSP 3.5 before any other nodes.

Note: When running on PSSP 3.5, LoadLeveler Version 3.1 must run with APAR IY33664 for AIX 5L Version 5.1 64-bit kernel support.

6.2.7 High-Availability Cluster Multiprocessing

The latest version of HACMP is 4.5. For detailed information about the new features and use of HACMP in a Cluster 1600 environment, refer to the redbook *Configuring Highly Available Clusters Using HACMP 4.5*, SG24-6845.

6.3 Migration scenarios

In this section, we provide details about the migration scenarios we tried in our lab environment. Notice that the migration options chosen were driven by the actual system environment.

We highlight only the steps that caused problems or that provided particularly useful information. We closely followed the instructions in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347. For detailed migration problem determination, refer to the redbook *Universal Clustering Problem Determination Guide*, SG24-6602.

We split our migration activity into five major parts:

1. Apply any prerequisite program temporary fixes (PTFs) to the CWS and the nodes and prepare the system for migration.
2. Migrate the CWS to the highest level of PSSP and AIX of any node it serves.
3. If we have to partition the system because of possible coexistence problems, we should do it at this time.
4. Migrate a test node.
5. Migrate the boot/install servers.
6. Migrate the nodes.

Possible migration paths are shown in Table 6-2.

Table 6-2 Migration paths

From	To
PSSP 3.1.1 and AIX 4.3.3	PSSP 3.2 and AIX 4.3.3
PSSP 3.1.1 and AIX 4.3.3	PSSP 3.4 and AIX 4.3.3
PSSP 3.2 and AIX 4.3.3	PSSP 3.4 and AIX 4.3.3
PSSP 3.2 and AIX 4.3.3	PSSP 3.4 and AIX 5L Version 5.1
PSSP 3.2 and AIX 4.3.3	PSSP 3.5 and AIX 5L Version 5.1
PSSP 3.4 and AIX 4.3.3	PSSP 3.4 and AIX 5L Version 5.1
PSSP 3.4 and AIX 4.3.3	PSSP 3.5 and AIX 5L Version 5.1
PSSP 3.4 and AIX 5L Version 5.1	PSSP 3.5 and AIX 5L Version 5.1

Attention: PSSP 3.1.1 and AIX 4.3.3 to PSSP 3.2 and AIX 4.3.3, and PSSP 3.1.1 and AIX 4.3.3 to PSSP 3.4 and AIX 4.3.3 must be done before the CWS is migrated to PSSP 3.5.

It is good practice to migrate only one node first to see if everything works properly and then migrate the other nodes. It is possible that some applications will not work with mixed levels. When a system has more than one node running the same cluster application, or when some applications on different nodes work together, you should migrate the nodes in groups to keep them in a consistent state.

Important: Until the migration is complete, avoid any configuration changes in the system, such as adding or deleting frames and nodes and changing host names or IP addresses.

Attention: We had many hardware- and network-related problems and, therefore, did some operations that may be unwise for a production system. Continuing a failed migration is not the best practice. However, there may be a situation when a restore of the whole system takes much more time than continuing with the failed migration.

6.3.1 Migrating PSSP 3.2 and AIX 4.3.3 to PSSP 3.5 and AIX 5.1

In this scenario, we have AIX 4.3.3 and PSSP 3.2 installed on the CWS and on all the nodes. VSD/RVSD filesets and GPFS 1.4 are also installed on the CWS and on five of the nodes. See Figure 6-1 illustrates this configuration. This is a heterogeneous cluster with four types of nodes that might represent a customer who has been using PSSP for long time.

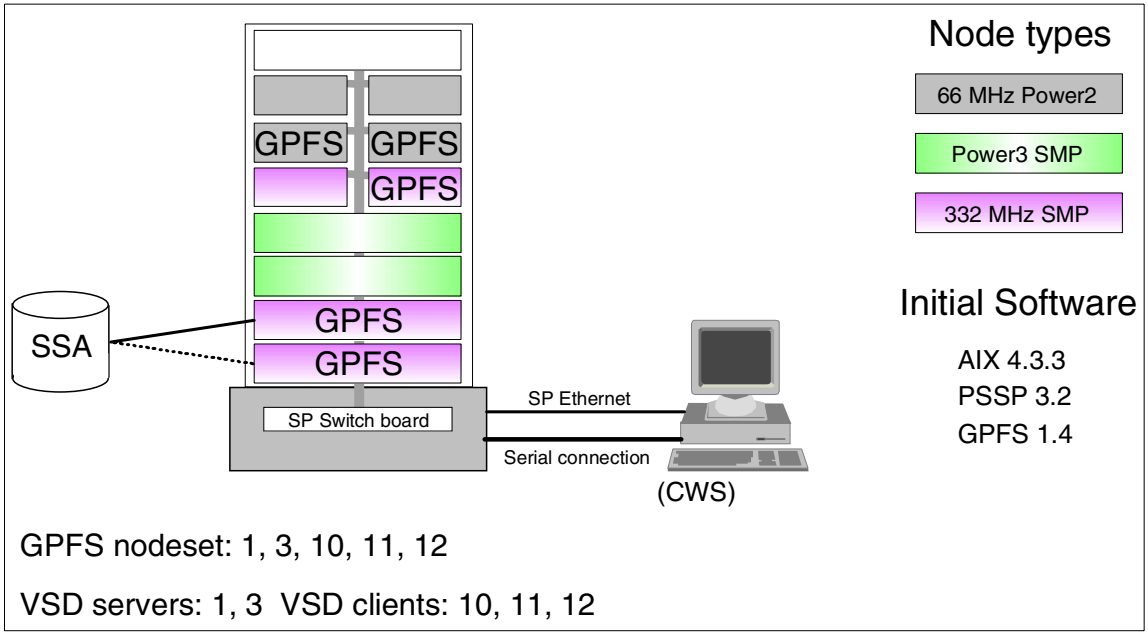


Figure 6-1 Migration from PSSP 3.2 in one step

Example 6-1 on page 132shows the system partition information and security settings.

Example 6-1 System Data Repository (SDR) system partition information

```
sp6cws:/
root $ splstdata -p
List System Partition Information

System Partitions:
------------------
sp6cws

Syspar: sp6cws
---------------------------------------------------------------------------
syspar_name     sp6cws
ip_address      9.12.6.79
install_image   default
syspar_dir      ""
code_version    PSSP-3.2
haem_cdb_version 1033078071,366945734,0
auth_install    k4:std
auth_root_rcmd  k4:std
ts_auth_methods compat
auth_methods    k4:std
```

We have two choices for the CWS migration. We can do it all in one maintenance window or we can split it into three steps. The staged migration for the CWS provides a shorter maintenance window and more possibilities to check the system state and go back if something goes wrong. The cumulative maintenance time, however, will be longer.

The staged migration consists of the following steps:

1. Migrate from PSSP 3.2 to PSSP 3.4, AIX remains on Level 4.3.3.
2. Migrate to AIX 5L Version 5.1 ML3.
3. Migrate to PSSP 3.5.

We migrated the CWS in one maintenance window. After that, all of the nodes are migrated in one maintenance window. Because of this, there were no coexistence issues to be considered.

Thee following list describes the migration steps and the problems we encountered:

1. Migrate the CWS to PSSP 3.5, AIX 5L Version 5.1, and GPFS 2.1:
   - Stop GPFS on nodes, suspend all VSDs, and stop all VSDs.
   - We migrated AIX 5L Version 5.1 without any problems, but after the restart of the CWS, **spmon** did not work. This was because the migration install of the OS overwrites the /tmp directory where the Kerberos ticket cache file

is saved. After reinitializing Kerberos with **k4init**, **spmon** showed host and switch responds for all nodes. This is documented in “Step 6” on page 146 of *PSSP for AIX: Installation and Migration Guide*, GA22-7347. We tried some of the standard test commands, such as **SDR_test**, **spmon_itest**, **spmon_ctest**, **CSS_test**, and **SYSMAN_test**. None of the commands failed, and all PSSP 3.2 subsystems seemed to run. After the migration, the machine started with the 32-bit kernel, because our CWS was a 32-bit machine.

Note: PSSP 3.2 is not supported on AIX 5L Version 5.1.

- “Step 16: Stop the daemons on the control workstation and verify” says to stop all SP daemons on the CWS. RVSD is installed on the CWS, and after rebooting the CWS, it starts automatically. So we have to stop the RVSD subsystem as well.
- After the successful migration of AIX and PSSP, we have to install the new version of VSD and GPFS on the CWS. As a prerequisite for VSD, we have to install the bos.clvm.enh fileset. Because the VSD version on the nodes will be earlier until we migrate them, we have to run the **rvsdrestrict** command and restart the VSD subsystem on the nodes. After this, we tested our defined GPFS file systems. Start GPFS on the nodes with the **mmstartup** command.
- With an AIX migration install, copying the LPPs for AIX and PSSP maintenance window should take about four hours. The necessary time can be reduced by first copying the AIX and PSSP LPPs to the CWS before the maintenance window using the **bffcreate** command. Remember to run **inutoc** after any file copy into the lppsource or pssplpp.

2. Migrate the nodes to AIX 5L Version 5.1 and PSSP 3.5:
   - These steps are defined in Chapter 4 of the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.
   - Stop GPFS and RVSD on all nodes.
   - After running **setup_server** in Step 3, check the log:

     ```
     nim -o showlog spot_aix51
     ```

   - In Step 6, we have to shut down the nodes. For the older MCA nodes that we have in this Cluster 1600 system, the **cshutdown** command returned the error messages shown in Example 6-2 on page 134.

Example 6-2 Cluster shutdown failed on one node

```
sp6cws:/tmp
root $ cshutdown -F -G ALL
Progress recorded in /var/adm/SPlogs/cs/cshut.0930201450.39518.

cshutdown: 0036-163 Problem with Frame Controller Interface (FCI) routine.
                    NodePowerOff() was unsuccessful with return code 1.
cshutdown: 0036-120 Could not switch power off to node 11.
```

- The **spled** command shows **000** for all MCA nodes. We have to run **spmon -p off node11** several times to be able to power off the node. This could be because the SP serial connection was overloaded.
- The migration on node1 was not successful because of a prerequisite failure for the vacpp filesets. However, we received a host response for that node. After some investigation, we found that AIX migrated, but PSSP 3.2 is running on the machine. After checking the log files, we found that this failure came in the pssp_script at node customization. The log file for pssp_script is sp6n01e0:/var/adm/SPlogs/sysman/sp6n01e0.config.log.7230. Before the migration, we installed Visual Age C++, vacpp.cmp Version 6.0 filesets to run some benchmark tests. For this software, there is a prerequisite fileset called xlC.adt.include 6.0.0.0. After installing the software, the installp needs xlC.rte 6.0.0.0 and not the version that we have in our lppsource directory for AIX 5L Version 5.1. To solve this problem, we did the following:

  i. On the CWS, put the xlC Version 6 filesets in the /spdata/sys1/install/name/lppsource/installp/ppc directory, named aix51 in our case.

  ii. On the CWS, run **inutoc** for this directory.

  iii. On the CWS, unallocate and recreate the Shared Product Object Tree (SPOT) using the **unallnimres**, **delnimres**, and **setup_server** commands. For details, refer to “Rebuilding the SPOT” on page 188.

  iv. On the CWS, change the node to customize with the **spbootins -r customize 1 1 1** command.

  v. On the node, ftp the pssp_script from the CWS /usr/lpp/ssp/install/bin directory to the node’s same directory.

  vi. On the node, run pssp_script. We had a problem where the script hangs when started in the background. The LED on the node was c42. To continue, bring the script to the foreground with the **fg** command.

  vii. Restart the node.

– MCA node migration failed because of the lost connection on the SP LAN Ethernet network. The AIX migration was successful, as in the example above, but some PSSP filesets remained at the old version because the pssp_script died. To solve this problem, we repeated the steps iv, v, vi, and vii in the previous section.

– Our SP LAN was very unstable, as seen from these examples. We had an NFS server connection problem as well, but in this case, it was much more dangerous than the earlier problem. This time, monitoring the AIX migration install with the **s1term** returned the message shown in Example 6-3.

Example 6-3 AIX migration error

```
0503-434 installp:  There are incomplete installation operations
        on the following filesets.  Run installp -C to clean up
        the previously failed installations before continuing.

        sysmgt.websm.webaccess
        sysmgt.websm.diag
+-----------------------------------------------------------------------------+
                           RPM  Error Summary:
+-----------------------------------------------------------------------------+
The following RPM packages were requested for installation
but they are already installed or superseded by a package installed
at a higher level:
mtools-3.9.8-1 is already installed.
cdrecord-1.9-4 is already installed.
mkisofs-1.13-4 is already installed.

Basic operating system support could not be installed.
System administrator should see /var/adm/ras/devinst.log for further
information.  Probable cause of failure is insufficient free disk space.
Type '2' to perform system maintenance to correct the problem, then type
'exit' to continue the installation, or restart the installation with
different installation options.
   ID#        OPTION
     1        Continue
     2        Perform System Maintenance and Then Continue
   Enter ID number: 2
```

– By pressing 2, we got a limited AIX shell where we could fix the problem. First, clean up the failed installation with **installp -C**, and then fix the network connection to the NFS server (CWS), and continue the installation by typing **exit** at the prompt. The failed filesets are installed automatically.

– When pssp_script fails, it is possible that we have to run the **installp -C** command on the node to clean up the interrupted installation. Check the log file for the scripts in the /var/adm/SPlogs/sysman directory.

3. After the migration, we re-established the switch communication and ran verification tests. We did not have any problems with these steps.
4. Install new versions of VSD and GPFS and run tests on the file systems. For VSD, we installed the bos.clvm.enh fileset as a prerequisite.

6.3.2 Migrating PSSP 3.1.1 and AIX 4.3.3 to PSSP 3.5 and AIX 5.1

Attention: PSSP 3.1.1 is not supported by IBM. Direct migration from PSSP 3.1.1 to PSSP 3.5 is not supported. We developed this section to show that it is possible to use the latest level of PSSP software with older SP nodes as well.

For this scenario, we chose a system with four 112 MHz SMP high nodes connected by an SP Switch. AIX 4.3.3 with Maintenance Level 10 and PSSP 3.1.1 are installed on the CWS and on the nodes. We have the first node configured as an NFS server with a file system that is mounted on the rest of the nodes.

There is no direct migration path available from the software level we have on this cluster. Therefore, we decided to move to PSSP 3.4 first on both the CWS and the nodes, and then to AIX 5L Version 5.1 ML3 and PSSP 3.5. For the first part, we followed the *PSSP for AIX: Installation and Migration Guide*, GA22-7347 for PSSP 3.4 and then for PSSP 3.5.

We completed the following steps:

1. Migrate the CWS to PSSP 3.4:
   - Before migration, check for non-ASCII data in the SDR by running **`SDRScan`**. Verify system configuration and *connectivity*.

Important: To avoid the following problem in “Step 22: Run SbR and system monitor verification test” in Chapter 4 of the *PSSP for AIX: Installation and Migration Guide*, GA22-7347, reset the Hardware Monitor daemon by running the **`hmreinit`** command. This is correctly documented in step 20 of the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

   - After migration, we found that **`spmon`** and **`spsvrmgr`** do not give back information about the nodes. The following examples contain the related error messages we found. General messages from SPdaemon.log are shown in Example 6-4 on page 137.

Example 6-4 Error message in /var/adm/SPlogs/SPdaemon.log

```
Oct  4 13:41:10 sp3cws hardmon[32052]:
LPP=PSSP,Fn=hm_tty.c,SID=1.21.4.15,L#=264, hardmon: 0026-850 Data length
mismatch in packet from tty /dev/tty0 (Frame 1): calculated = 15, received =
14.
Oct  4 13:42:09 sp3cws sphwlog[32106]:
LPP=PSSP,Fn=splogd.c,SID=1.16.7.3,L#=1537, 0026-107 Failure; Frame 1:0;
frPowerModCbad; Power module - DC power loss.
```

The SDR configuration messages are shown in Example 6-5.

Example 6-5 SDR_init log file: /var/adm/SPlogs/sdr/SDR_config.log

```
SDR_init: SDR_init was invoked at Fri Oct  4 15:36:04 EDT 2002 with flag values
of debug=0, log=1 and verbose=0.
SDR_init: 0016-082 An error has been encountered while internally executing the
command "/usr/lpp/ssp/bin/hmmon -Q -v type -r 1,5,9,13 2>  /dev/null". The
return code from the command was 1. SDR_init is continuing.
SDR_init: 0016-082 An error has been encountered while internally executing the
command "/usr/lpp/ssp/bin/hmmon -Q -v type -r 1,5,9,13 2>  /dev/null". The
return code from the command was 1. SDR_init is continuing.
SDR_init: 0016-705 Problem while attempting to read Hardmon data.
SDR_init: 0016-733 SDR_init completed unsuccessfully with a return code value
of 2.
```

Hardmon-related messages are shown in Example 6-6.

Example 6-6 Hardmon daemon log file /var/adm/SPlogs/spmon/hmlogfile.277

```
hardmon: 0026-801I Hardware Monitor Daemon started at Fri Oct  4 13:20:31 2002
hardmon: 0026-802I Server port number is 8435, poll rate is 5.000000 seconds
hardmon: 0026-805I 1 frames have been configured.
hardmon: 0026-803I Entered main processing loop 0000001c
hardmon: 0026-808I Received command to quit from SIGTERM at sp3cws/0.
hardmon: 0026-801I Hardware Monitor Daemon started at Fri Oct  4 13:27:39 2002
hardmon: 0026-802I Server port number is 8435, poll rate is 5.000000 seconds
hardmon: 0026-805I 1 frames have been configured.
hardmon: 0026-803I Entered main processing loop 00000038
hardmon: 0026-808I Received command to quit from SIGTERM at sp3cws/0.
hardmon: 0026-850 Data length mismatch in packet from tty /dev/tty0 (Frame 1):
calculated = 15, received = 14.
```

– After restarting the hardmon daemon with the **/usr/lpp/ssp/install/bin/hmreinit** command on the CWS, everything worked fine.

2. Migrate PSSP on the nodes:
   - In this step, we installed the new version for PSSP only. AIX remained on Version 4.3.3. PSSP migration needs node customization only. For this, after changing the node information in the SDR, pssp_script must be copied to the nodes and it must be started. We found that one node did not finish the software installation during the first run. After a second start of pssp_script on that node, PSSP 3.4 was installed on every node, and all the verification commands succeeded.
3. Install VSD for PSSP 3.4 and GPFS Version 1.5 in this step for a coexistence and migration test:
   - Every node is designated as a VSD node. We created four VSDs on one node on an integrated disk just for this test. Example 6-7 shows the list of defined VSDs.

Example 6-7 VSD list on node sp3n01e0

```
sp3cws:/
root $ dsh -w sp3n01e0 lsvsd -l
sp3n01e0: minor state server lv_major lv_minor vsd-name option size(MB)
server_list
sp3n01e0:  1      ACT    9       0       0      vsd1n9 nocache 256          9
sp3n01e0:  2      ACT    9       0       0      vsd2n9 nocache 256          9
sp3n01e0:  3      ACT    9       0       0      vsd3n9 nocache 256          9
sp3n01e0:  4      ACT    9       0       0      vsd4n9 nocache 256          9
```

   - The VSD server node number is 9. We changed the VSD configuration, as shown in Example 6-8.

Example 6-8 Output of vsdatalst -n

```
sp3n09e0:/
root $ vsdatalst -n
       VSD Node Information
                                          Initial Maximum     VSD      rw      Buddy Buffer
  node                 VSD      IP packet   cache   cache request request minimum maximum size: #
number host_name       adapter    size    buffers buffers   count   count    size     size maxbufs
------ --------------- -------- --------- ------- ------- ------- ------- ------- ------- -------
     1 sp3n01e0        css0        61440     256     256     256      48    4096  262144       2
     5 sp3n05e0        css0        61440     256     256     256      48    4096  262144       2
     9 sp3n09e0        css0        61440     256     256     256      48    4096  262144      32
    13 sp3n13e0        css0        61440     256     256     256      48    4096  262144       2
```

4. Migrate to AIX 5L Version 5.1 on the CWS:
   - Before BOS migration, stop VSD on the nodes.
   - Quiesce the switch using the **Equiesce** command.

– Keep in mind that as a final step of the AIX migration, the licence agreement must be accepted. After this, the AIX Installation Assistant starts. This means that if the machines are in a separate room after putting in the last disk, we have to go and finish the previous steps to get a running CWS.

– The vacpp.ioc.aix50.rte fileset must be installed, because after the AIX migration, only the vacpp.ioc.aix43.rte fileset will be on the system.

– Check the AIX software level with the **oslevel** command, as shown in Example 6-9.

Example 6-9 Output of the oslevel command

```
sp3cws:/spdata/sys1/install/aix51/lppsource
root $ oslevel -l 5.1.0.0
sp3cws:/spdata/sys1/install/aix51/lppsource
root $ oslevel -r
5100-03
```

– The migration from disk extends the file systems if it is needed, but the amount of added space is only enough for the migration itself. Check the space in the /usr, /var, and /spdata file systems before any further activity. The space requirements for /spdata can be found in the *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.

– Go through the rest of the necessary steps for AIX migration on CWS as they are listed in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347. This also includes the preparation of lppsource for the installation of AIX 5L Version 5.1.

– The AIX migration install removes everything from /tmp, which means that the Kerberos ticket files are removed as well. The result of this can be seen in Example 6-10. Run **k4init** for the users for which you need Kerberos authentication. This is documented in Step 6 on page 146 of the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

Example 6-10 Authorization problems after migration

```
sp3cws:/
root $ spmon -d
1.  Checking server process
    Process 21690 has accumulated 1 minutes and 3 seconds.
    Check successful

2.  Opening connection to server
    Connection opened
    Check successful
```

```
3.  Querying frame(s)
spmon: 0026-064 You do not have authorization to access the Hardware Monitor.
spmon: 0026-059 Could not query frames.
sp3cws:/
root $ k4list
Ticket file:    /tmp/tkt0
k4list: 2504-076 Kerberos V4 ticket file was not found
```

5. Migrate to AIX 5L Version 5.1 on nodes 9 and 13:
   - Enter and verify the node configuration data in the SDR.
   - Stop VSD on nodes 9 and 13. Example 6-11 shows VSD availability after stopping the server node.

Example 6-11 Stopping VSD

```
sp3cws:/
root $ dsh -w sp3n09e0,sp3n13e0 suspendvsd -a
sp3cws:/
root $ dsh -w sp3n09e0,sp3n13e0 stopvsd -a
sp3cws:/
root $ dsh -w sp3n09e0,sp3n13e0 ha.vsd stop
sp3n09e0: 0513-044 The rvsd Subsystem was requested to stop.
sp3n09e0: ha.vsd: Wed Oct  9 11:41:02 EDT 2002 Waiting for 16302 to exit.
sp3n09e0: ha.vsd: Wed Oct  9 11:41:08 EDT 2002 16302 has exited.
sp3n13e0: 0513-044 The rvsd Subsystem was requested to stop.
sp3n13e0: ha.vsd: Wed Oct  9 11:41:02 EDT 2002 Waiting for 15864 to exit.
sp3n13e0: ha.vsd: Wed Oct  9 11:41:08 EDT 2002 15864 has exited.
sp3cws:/
root $ dsh -w sp3n09e0,sp3n13e0 stopsrc -s hc.hc
sp3n09e0: 0513-044 The hc.hc Subsystem was requested to stop.
sp3n13e0: 0513-044 The hc.hc Subsystem was requested to stop.
sp3cws:/
root $ dsh -w sp3n01e0,sp3n05e0 lsvsd -l
sp3n01e0: minor  state server lv_major lv_minor vsd-name option     size(MB)
server_list
sp3n01e0:  1      STP   -1       0       0      vsd1n9 nocache        256
sp3n01e0:  2      STP   -1       0       0      vsd2n9 nocache        256
sp3n01e0:  3      STP   -1       0       0      vsd3n9 nocache        256
sp3n01e0:  4      STP   -1       0       0      vsd4n9 nocache        256
sp3n05e0: minor  state server lv_major lv_minor vsd-name option     size(MB)
server_list
sp3n05e0:  1      STP   -1       0       0      vsd1n9 nocache        256
sp3n05e0:  2      STP   -1       0       0      vsd2n9 nocache        256
sp3n05e0:  3      STP   -1       0       0      vsd3n9 nocache        256
sp3n05e0:  4      STP   -1       0       0      vsd4n9 nocache        256
```

– Go through the other preparation steps and network boot the nodes. Because we are migrating only two nodes of the four we have on the switch, we have to check which node is the primary and primary backup for the switch. It is possible to change them by running the **Eprimary "node number1" -backup "node number2"** and **Estart** commands. Fence the nodes from the switch using the **Efence** command.

– Stop and network boot nodes 9 and 13.

– The AIX migration did not start on node 13, because we lost hdisk1 from rootvg. To continue, we opened a read/write **s1term** and set the migration configuration to use the remaining disk for the AIX install. The SDR has the data about the volume group information in the Volume_Group class. This can be listed with the **SDRGetObjects Volume_Group** or **splstdata -v** commands. In this case, even if the AIX migration used only one disk for rootvg, the SDR contains two disks for this node.

– The other migration on node 9 failed as well. We did not have enough free Logical Partitions (LP) in rootvg to extend the size of the /usr file system. This extension is done automatically by the migration process if you have enough space in rootvg. In this case, the safest way is to restore the backup and start the migration again. However, this is a test environment, and there is no application running on this system. Therefore, we connected to the system with **s1term** and tried to repair the system with the following steps:

  i. Check the level of the AIX filesets with **oslevel**. However, the command did not give back any answer. Fortunately, bos.mp and some main parts of AIX were installed for Version 5.1. The **oslevel -l 5.1.0.0** command showed many filesets with Level 4.3.3.

  ii. In AIX 5L Version 5.1, the **chps** command has a new flag that enables decreasing the size of the migration paging space. The command creates a temporary paging space and activates it. It is possible to deactivate a paging space in AIX 5L as well. So the command deactivates the original, deletes it, and creates a new one using the same name. Then it changes the active paging space to the new. This new feature enabled us to decrease the size of the paging space to win some space for the /usr file system. Example 6-12 on page 142 shows the output of **chps -d 22 hd6**.

Example 6-12 Decreasing the paging space

```
shrinkps: Temporary paging space paging00 created.
shrinkps: Dump device moved to temporary paging space.
shrinkps: New boot image created with temporary paging space.
shrinkps: Paging space hd6 removed.
shrinkps: Paging space hd6 recreated with new size.
shrinkps: New boot image created with resized paging space.
```

iii. The migration failed before changing bootp_response in the SDR class from node to disk. We tried to run **nodecond** to migrate the node again, but the migration did not start.

iv. We mounted the lppsource directory from the CWS and installed the bos.up fileset.

v. We ran **update_all** from SMIT and updated all the filesets from AIX 4.3.3 to AIX 5L Version 5.1. The AIX level on the nodes after this step is shown in Example 6-13.

Example 6-13 Run oslevel on the CWS

```
sp3cws:/
root $ dsh -a oslevel -r
sp3n01e0: 4330-09
sp3n05e0: 4330-09
sp3n09e0: 5100-03
sp3n13e0: 5100-03
```

vi. We customized the node using the standard method.

– At this stage, we have node 1 and 5 installed with AIX 4.3.3 and PSSP 3.4, and node 9 and 13 installed with AIX 5L Version 5.1 and PSSP 3.4. The VSD server is node 9. We tried VSD functionality using IP and KLAPI between the nodes without any problem. The RVSD recovery and quorum function worked fine, too.

6. Migrate the CWS to PSSP 3.5:

– Go through the steps in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347 for CWS migration.

– After stopping the SDR daemon on the CWS, we cannot use the **dsh** command with the **-a** option. However, with the **-w** option, we can specify a node for the remote command.

– In Step 18, we should start the installation of all ssp filesets. As a part of this fileset, ssp.vsdgui would be installed as well, but for this, the other VSD filesets are a prerequisites. Because of this, we first have to install the bos.clvm.enh fileset. Instead of using the command for installing everything under the ssp name, it is better to use SMIT, where we can exclude ssp.hacws and ssp.vsdgui for now.

– Check the /.spgen_klogin file as stated in Step 21. We had to modify the file as it is in the document because our maintenance level was not the latest for PSSP.

– Verify the PSSP 3.5 installation on the CWS.

– Install the latest version of VSD:

```
installp -acgX -d /spdata/sys1/install/pssplpp/PSSP-3.5 vsd ssp.vsdgui
```

– Install GPFS 2.1 on the CWS.

– Because we have nodes installed with an older version of VSD, if we start the RVSD subsystem, we get the error message shown in Example 6-14.

Example 6-14 RVSD start problem

```
sp3cws:/
root $ ha_vsd reset
Stopping subsystem rvsd.sp3cws.
0513-004 The Subsystem or Group, rvsd.sp3cws, is currently inoperative.
0513-083 Subsystem has been Deleted.
Stopping subsystem hc.hc.
0513-004 The Subsystem or Group, hc.hc, is currently inoperative.
0513-083 Subsystem has been Deleted.
There are 4 nodes in partition sp3cws.
Making RVSD subsystem in partition sp3cws.
ha.vsd: Thu Oct 10 17:49:24 EDT 2002 Making SRC object "rvsd.sp3cws".
0513-071 The rvsd.sp3cws Subsystem has been added.
ha.vsd: 2506-112 Thu Oct 10 17:49:25 EDT 2002 RVSD can not start. Backlevel
nodes were detected. See the rvsdrestrict command.
sp3cws:/
root $ lssrc -g rvsd
Subsystem         Group            PID     Status
 rvsd.sp3cws      rvsd                     inoperative
```

– Use the **rvsdrestrict** command to change the working level of RVSD as in Example 6-15 on page 144.

Example 6-15 The rvsdrestrict command

```
sp3cws:/
root $ rvsdrestrict -s RVSD3.4
rvsdrestrict level is RVSD3.4
sp3cws:/spdata/sys1/install/pssplpp/PSSP-3.5
root $ ha_vsd reset
Stopping subsystem rvsd.sp3cws.
0513-004 The Subsystem or Group, rvsd.sp3cws, is currently inoperative.
0513-083 Subsystem has been Deleted.
There are 4 nodes in partition sp3cws.
Making RVSD subsystem in partition sp3cws.
ha.vsd: Thu Oct 10 17:50:55 EDT 2002 Making SRC object "rvsd.sp3cws".
0513-071 The rvsd.sp3cws Subsystem has been added.
ha.vsd: 2506-111 Thu Oct 10 17:50:57 EDT 2002 The rvsdrestrict command forces
RVSD to reduce its function to 3.4.0.0.
0513-059 The rvsd.sp3cws Subsystem has been started. Subsystem PID is 33282.
```

- Copy and install the latest PSSP PTFs onto the CWS and recreate the .toc file as stated in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

7. Install PSSP 3.5 on nodes 9 and 13.
   - Follow the instructions in the book for changing node information and then customizing the nodes. The PSSP installation is done by running the already copied pssp_script on the nodes. Example 6-16 shows a way to find the log file for the script running on the node.

Example 6-16 The pssp_script log file

```
sp3n09e0:/var/adm/SPlogs/sysman
root $ ps -ef|grep pssp
    root 20302 26400   4 14:17:30  pts/0  0:00 grep pssp
    root 25608     1   0 14:04:37      -  0:01 ksh /tmp/pssp_script
sp3n09e0:/var/adm/SPlogs/sysman
root $ ls -l *.25608
-rw-r--r--  1  root   system      35727 Oct 11 14:17 sp3n09e0.config.log.25608
```

   - Install the latest PSSP PTFs from the CWS and restart the nodes.
   - The customization does not install the latest version of VSD and GPFS, so this must be done after the initial PSSP testing.
   - Run system verification tests.

8. Migrate nodes 1 and 5 to AIX 5L Version 5.1 and install PSSP 3.5 in one step.
   - As in the case of any other node migration, the applications running on the nodes and the switch communication must be stopped before starting the migration.

- This step involves changing the rootvg object in the SDR with the **spchvgobj** command, setting the node boot response to migrate with the **spbootins** command, running **setup_server**, and running the **nodecond** command, which changes the bootlist of the node and restarts it. It is possible to run these commands for more than one node. Proper planning is needed for multiple installations at the same time to avoid overloading the network and the CWS or other boot/install servers. For more information, refer to Chapter 6 of the *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.
- PSSP uses the AIX NIM environment in a special way to create the SP-related NIM resources and allocate them for NIM clients. At migration preparation time, the pssp_script script is defined as a NIM script resource for the nodes. After AIX migration, this script is started by NIM to install the new PSSP code and customize the node. For hints on customizing the Cluster 1600 system NIM environment and keeping the settings after running setup_server, refer to “NIM and PSSP coexistence” on page 189. It is possible to follow the AIX migration and the pssp_script script by watching the LED codes on the machines. For a detailed description, refer to *PSSP for AIX: Diagnosis Guide*, GA22-7350.
- Install the latest PSSP PTFs from the CWS and restart the nodes.
- The migration and customization do not install the latest version of VSD and GPFS, so this must be done after the initial PSSP testing.

6.3.3 Migrating PSSP 3.4 and AIX 4.3.3 to PSSP 3.5 and AIX 5.1

In this scenario, the first step is the AIX migration of the CWS.

PSSP 3.5 migration must first be done on the CWS before any nodes are migrated to PSSP 3.5. The steps for the CWS can be done either in one or two maintenance windows. The PSSP migration of the nodes is just a node customization.

We have compiled the possible migration steps in Table 6-3.

Table 6-3 Migration routes

Route 1	Route 2	Route 3
Migrate AIX on CWS	Migrate AIX on CWS	Migrate AIX on CWS
Migrate AIX on nodes	Install PSSP 3.5 on CWS	Install PSSP 3.5 on CWS
Migrate PSSP 3.5 on CWS	Migrate AIX and install PSSP 3.5 on nodes	Migrate AIX on nodes

Route 1	Route 2	Route 3
Migrate PSSP 3.5 on nodes		Migrate PSSP 3.5 on nodes

For more information, refer to 6.3.2, “Migrating PSSP 3.1.1 and AIX 4.3.3 to PSSP 3.5 and AIX 5.1” on page 136, where we migrated the system first to an intermediate level and then to the latest AIX and PSSP levels.

6.3.4 Migrating PSSP 3.4 and AIX 5.1F to PSSP 3.5 and AIX 5.1F

This scenario is described in 6.3.3, “Migrating PSSP 3.4 and AIX 4.3.3 to PSSP 3.5 and AIX 5.1” on page 145.

6.4 Integration of SP-attached servers

In this section, we discuss, step-by-step, the integration of external nodes (SP-attached servers) to our Cluster 1600. The hardmon daemon uses different protocols for each kind of SP-attached server. The names of the protocols are shown in Table 6-4.

Table 6-4 Hardware protocols for available Cluster 1600 nodes

Protocol name	Servers
SP	SP nodes
SAMI	RS/6000 S70, S7A, and S80 or IBM @server pSeries 680 servers
CSP	RS/6000 H80, M80, and IBM @server pSeries 660 servers (6H0, 6H1, and 6M1)
HMC	IBM @server pSeries 630, 670, and 690 servers

Hardware protocol information:

- In the case of an Enterprise Server (S70, S7A, S80, p680) that uses SAMI, the hardmon daemon on the CWS does not have a direct connection to the node and frame supervisor card installed in the external system. The connection is made through another daemon running on the CWS for every attached server.
- The IBM @server pSeries 660 servers have an SP Attachment adapter card that is used to communicate with the CWS over a serial line using the CSP protocol. For these servers, no other daemon is necessary to translate the communication protocol for hardmon.

- There is one additional daemon running for each HMC server on the CWS to provide communication between the hardmon daemon and the HMC server using the HMC protocol.

The protocol name can be found for every frame by running the **splstdata -f** command, as shown in Example 6-17.

Example 6-17 SDR frame information

```
sp4en0:/
root $ splstdata -f
                  List Frame Database Information

frame# tty        s1_tty    frame_type      hardware_protocol  control_ipaddrs domain_name
------ ---------  ------    --------------- ------------------ --------------- -----------
     1 /dev/tty0  ""        switch          SP                 ""              ""
     2 /dev/tty1  ""        noswitch        SP                 ""              ""
     3 /dev/tty2  ""        ""              CSP                ""              ""
     4 ""         ""        ""              HMC                192.168.4.251   Enterprise
     5 /dev/tty3  /dev/tty4 ""              SAMI               ""              ""
```

The data flow of the hardware control subsystem is shown in Figure 6-2.

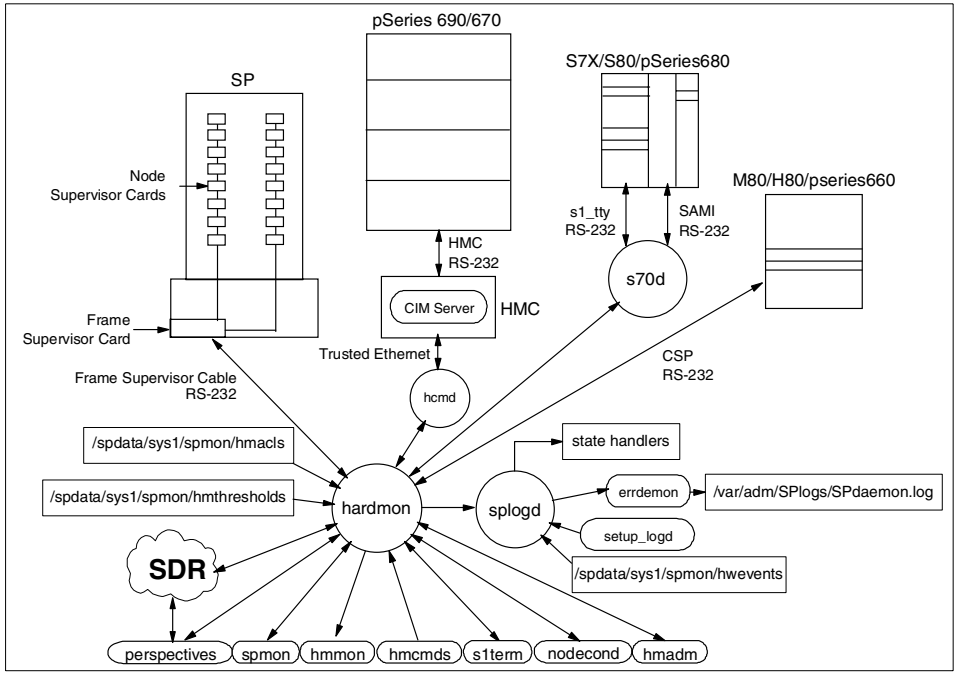


Figure 6-2 Hardware control

For a description of the Cluster 1600 control components, refer to the IBM Redbook *RS/6000 SP Cluster: The Path to Universal Clustering*, SG24-5374.

The server with SAMI and CSP hardware management protocols behaves like a frame with one node installed in it. In the case of pSeries 630, 670, and 690 servers managed by an HMC, if there are LPARs configured, every LPAR will look like a separate node from the CWS. Node information is created in the SDR for every node reachable from the hardmon daemon.

As with the SP frames and nodes, the serial and SP LAN connection must be prepared between the frames, nodes, and the CWS before any change in the SDR. These preparations are different for every kind of hardware management protocol, and we discuss them in the following sections.

6.4.1 pSeries 660, Model 6H1

In this scenario, we have an SP frame with two high nodes and two HMC managed LPARs. We also have an SP Switch2 in this configuration. We add two pSeries 660 servers as external nodes.

The two nodes are as follows:

- sp4n17e0: We keep the existing AIX installed on this server. This will simulate an existing node in our enterprise.
- sp4n33e0: A new installation with AIX and PSSP. The new node is seen as a new server.

Hardware considerations

Before you integrate a pSeries p660 server into a Cluster 1600, a special interface, the SP System Attachment Adapter (FC 3154), must be integrated into the machine. Although not connected to the PCI bus, it occupies one slot. This provides the connection to the serial port of the CWS. Additionally, you have to provide a management Ethernet adapter, which is recommended by IBM to reside in slot 1. This is because of the way the p660 defines its en0 that is assigned to the leftmost Ethernet adapter on the PCI bus, excluding the integrated adapter. However, as long as your Ethernet adapter is in the leftmost slot, this is a supported configuration. If your production environment does not allow integration of an additional adapter in the leftmost slot, see “Identifying Ethernet adapters on the pSeries p660” on page 185.

Tip: We recommend upgrading the firmware of the p660 6H0 and 6H1 to at least CM020807, and to at least MM020807 for the p660 6M1.

Integrating a CSP protocol server

In this section, we highlight the unique steps for these types of external nodes.

Example 6-18 shows the initial configuration for this scenario.

Example 6-18 Initial configuration

```
root $ spmon -d
---------------------------------- Frame 1 ----------------------------------
                      Host     Switch   Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- -------- ------- ----- ---------------- -------
  1     1  high   on     yes      yes     N/A     no  LCDs are blank      no
  5     5  high   on     yes      yes     N/A     no  LCDs are blank      no


---------------------------------- Frame 4 ----------------------------------
                      Host     Switch   Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- -------- ------- ----- ---------------- -------
  1    49  thin   on     yes    noconn    N/A    N/A  LCDs are blank     N/A
  2    50  thin   on     yes      yes     N/A    N/A  LCDs are blank     N/A


root $ splstdata -n
                 List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
------------- -------------------- --------------- --------- --------------- ----------
    1      1     1     4 sp4n01e0          sp4n01e0          ""                192.168.4.250
MP                              16 375_MHz_POWER3_         1 true            ""
    5      1     5     4 sp4n05e0          sp4n05e0          ""                192.168.4.250
MP                              16 375_MHz_POWER3_         1 true            ""
   49      4     1     1 sp4n49e0          sp4n49e0          ""                192.168.4.250
MP                               4 7040-681                0 false           NCC1701-A
   50      4     2     1 sp4n50e0          sp4n50e0          ""                192.168.4.250
MP                               4 7040-681                1 true            NCC1701-B
```

The steps unique to these types of external nodes are as follows:

1. After the preparation of the serial connection and the SP LAN, we can add the two new frames. We have to add a frame for each of the nodes. These frames will contain only one node. The commands to add a CSP type frame are shown in Example 6-19. By running the **splstdata -f** command, we can see that the hardware protocol for the new frames is CSP. The basic node information is added automatically. The node type is extrn in the output of **spmon -d**, and the node reserves one slot if we check by running **splstdata -n**.

Example 6-19 Adding the frames

```
root $ /usr/lpp/ssp/bin/spframe -p CSP -r yes 2 2 /dev/tty1
0025-322  SDRArchive: SDR archive file name is /spdata/sys1/sdr/archives/backup.02282.1311
sp4en0:/
root $ splstdata -f
                  List Frame Database Information

frame# tty        s1_tty         frame_type      hardware_protocol  control_ipaddrs domain_name
------ -------    -------------- --------------- ------------------ --------------- -----------
     1 /dev/tty0  ""             switch          SP                 ""              ""
     2 /dev/tty1  ""             ""              CSP                ""              ""
     3 /dev/tty2  ""             ""              CSP                ""              ""
     4 ""         ""             ""              HMC                192.168.4.251   Enterprise
sp4en0:/

root $ splstdata -n
                  List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ---------------- ---------- ---------------- ---------------
-------------- ------------------ --------------- --------- --------------- -----------
    1      1     1     4 sp4n01e0          sp4n01e0          ""               192.168.4.250
MP                              16 375_MHz_POWER3_         1 true            ""
    5      1     5     4 sp4n05e0          sp4n05e0          ""               192.168.4.250
MP                              16 375_MHz_POWER3_         1 true            ""
   17      2     1     1 ""                ""                ""               ""
MP                               1 ""                      0 false           ""
   33      3     1     1 ""                ""                ""               ""
MP                               1 ""                      0 false           ""
   49      4     1     1 sp4n49e0          sp4n49e0          ""               192.168.4.250
MP                               4 7040-681                0 false           NCC1701-A
   50      4     2     1 sp4n50e0          sp4n50e0          ""               192.168.4.250
MP                               4 7040-681                1 true            NCC1701-B
```

2. Next, we add the SP Ethernet information for node number 33 into the SDR, as shown in Example 6-20 on page 151.

Example 6-20 Adding Ethernet information

```
root $ /usr/lpp/ssp/bin/spadaptrs -e 192.168.4.250 -t tp -d half -f 10 3 1 1  en0 192.168.4.33
255.255.255.0
root $ splstdata -n -l 33
                  List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- --------------- --------- --------------- ----------
   33      3     1     1 sp4n33e0          sp4n33e0           ""                192.168.4.250
MP                                1 ""                        0 false           ""
```

3. To add more PSSP managed adapters into the server, run the **spadaptrs** command for each of them. In this configuration, we have an SP Switch2 installed. In this case, there are no restrictions for the available switch node numbers. If your system contains an SP Switch, check the restrictions in the *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.
4. In the following steps, we add the boot/install information for node number 33. Obtain the hardware address and prepare the boot/install server on the CWS, as shown in Example 6-21. The install image can be any AIX 5L Version 5.1 image made from an operational node or the image that is provided by IBM for PSSP 3.5.
5. Node number 33 will be installed this time, so we use the **sphrdwrad** command without any preparation. Keep in mind that if you have a operational node, and the command can not find a hardware address for that node in the SDR or the /etc/bootptab.info file, the node is rebooted to get the hardware address.

Example 6-21 Prepare the boot/install server

```
root $ splstdata -b -l 33
                  List Node Boot/Install Information

node# hostname hdw_enet_adr srvr response  install_disk last_install_image last_install_time
next_install_image  lppsource_name pssp_ver            selected_vg
----- -------- ------------ ---- --------- ------------ ------------------ -------------------
------------------- -------------- ------------------- -------------------
   33 sp4n33e0         000000000033     0 install     hdisk0                initial
initial             default             default        PSSP-3.5           rootvg
root $ /usr/lpp/ssp/bin/spchvgobj -r rootvg -h hdisk0 -c 1 -n 0 -i mksysb.51f_64 -v aix51 -p
PSSP-3.5 3 1 1
spchvgobj: Successfully changed the Node and Volume_Group objects for node number 33, volume
group rootvg.
spchvgobj: The total number of changes successfully completed is 1.
```

```
spchvgobj: The total number of changes which were not successfully completed is 0.
sp4en0:/
root $ splstdata -b -l 33
                  List Node Boot/Install Information

node# hostname hdw_enet_adr  srvr response install_disk last_install_image  last_install_time
next_install_image  lppsource_name pssp_ver            selected_vg
----- -------- ------------ ---- -------- ------------ ------------------- -------------------
------------------- -------------- ------------------- -------------------
   33 sp4n33e0         000000000033     0 install     hdisk0                 initial
initial             mksysb.51f_64       aix51          PSSP-3.5            rootvg

root $ /usr/lpp/ssp/bin/sphrdwrad 3 1 1
Acquiring hardware Ethernet address for node 33
Hardware ethernet address for node 33 is 000629DC2595
Ping to default_route successful for node 33.
sp4en0:/usr/lpp/ssp/bin
root $ splstdata -b -l 33
                  List Node Boot/Install Information

node# hostname hdw_enet_adr srvr response install_disk last_install_image  last_install_time
next_install_image  lppsource_name pssp_ver            selected_vg
----- -------- ------------ ---- -------- ------------ ------------------- -------------------
------------------- -------------- ------------------- -------------------
   33 sp4n33e0         000629DC2595     0 install     hdisk0                 initial
initial             mksysb.51f_64       aix51          PSSP-3.5            rootvg

root $ setup_server
setup_server: Running services_config script to configure SSP services.This may take a few
minutes...
...
Lines omitted
...
mknimclient: Client node 33 (sp4n33e0) defined as NIM client on
server node (NIM master) 0 (sp4en0).
export_clients: File systems exported to clients from server node 0.
...
Lines omitted
...
allnimres: Node 33 (sp4n33e0) prepared for operation: install.
```

6. Start the network install of the new node with the **nodecond** command.

7. Adding the node where we want to keep the existing installation differs from the above, because we set the node to *customize* instead of install. The command is **spbootins -s no -r customize 2 1 1**. After this the PSSP software installation is done by running **pssp_script** on the node. The following steps provide the high level procedures:
   a. Mount /spdata/sys1/install/pssplpp/PSSP-3.5 from the CWS.
   b. Install ssp.basic on the node.
   c. Copy /etc/SDR_dest_info from the CWS to the node.
   d. Run **/usr/lpp/ssp/install/bin/pssp_script** on the node.
   e. Check the switch communication if you configured a switch adapter. **Eunfence** the node if it is necessary.
8. Use **spmon -d** to check the host response and switch response for the new nodes. Run verification tests as listed in *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

6.4.2 pSeries 690, Model 681

Because the integration of new HMC-based servers to a Cluster 1600 requires special treatment, we discuss the necessary considerations and decisions before we integrate them into our cluster.

Hardware considerations

The new HMC protocol type server does not require a serial attachment to the CWS. Instead, an IP connection from the control workstation to the HMC is used for the protocol flow. This connection can be either on the existing management Ethernet, or through an additional trusted network, containing only the HMC and the CWS. The HMC itself is connected through a serial line and an IP interface to each server it manages. This reduces the amount of serial lines needed to connect to different nodes compared to, for example, a cluster of 6H1s servers. Figure 6-3 on page 154 shows an example of an HMC managing different pSeries servers controlled by the CWS.

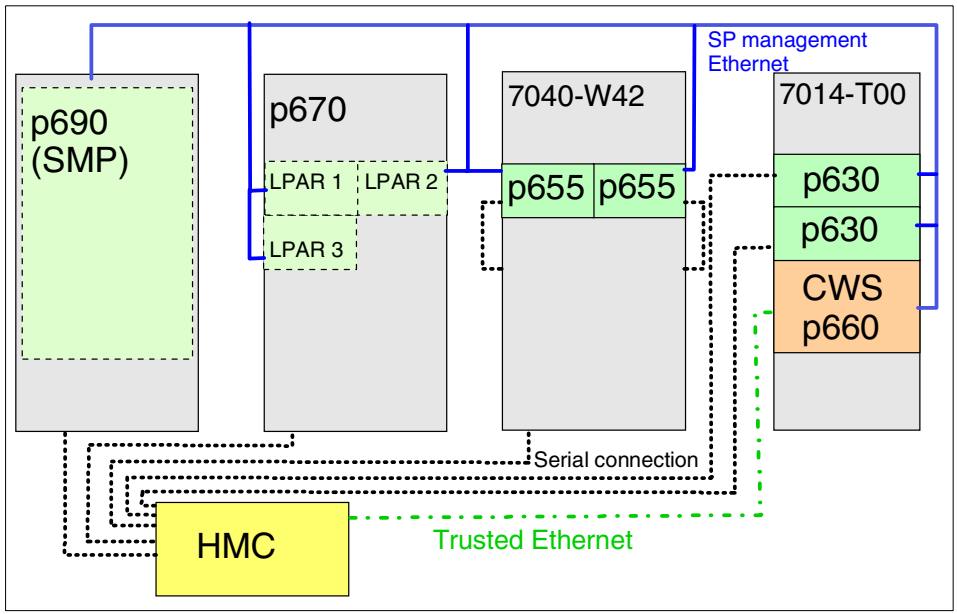


Figure 6-3 Connection between the HMC and the CWS

Tip: Although the performance of the HMC itself is high, the serial connections to the connected servers can be a bottleneck if too many servers are connected to one HMC. If you have a large cluster, we recommend distributing the managed nodes equally if possible.

HMC preparation

In general, be sure to have the latest software level on the HMC. For attaching the p670/p690, at least Version 2, Release 1.1, and for the p655/p630, at least Version 3, Release 1.0, should be installed on the HMC. Be sure to upgrade the HMC software first, before you upgrade the firmware on your pSeries server.

Attention: When applying a software service to an HMC, the associated HMC daemon on the CWS must be stopped while the software service is applied.

Tip: Be aware that PSSP orders the LPARs as thin nodes in the frame and numbers them as they are numbered in the HMC. This is not necessarily the order in which the HMC display shows the LPARs.

If only one pSeries server is connected to the HMC, the first native serial port is used for the RS232 TTY connection. If more than one server is connected to one single HMC, an 8-port or 128-port Async PCI card is needed. The second native serial port is reserved for a modem connection. In an IBM Cluster 1600, the Object Manager Security Mode on the HMC needs to be set to plain socket. This is necessary for the PSSP hardware control and monitor functions. If the mode is set to Secure Sockets Layer (SSL), PSSP will not be able to perform the hardware monitor and control functions. The Object Manager Security menu is located in the System Manager Security folder of the WebSM interface. Figure 6-4 shows how the settings should look.

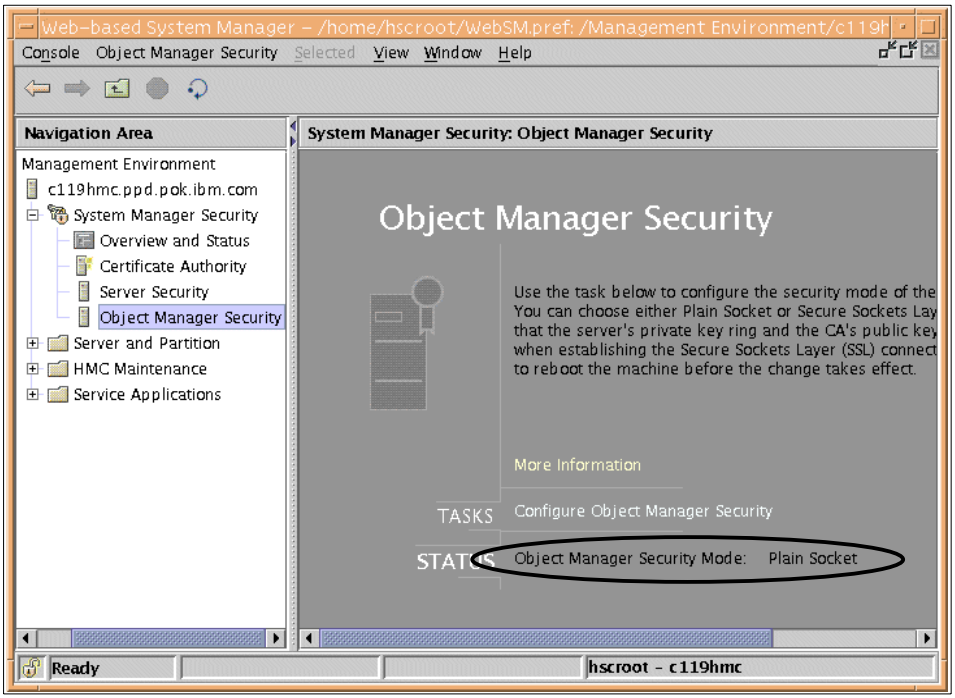


Figure 6-4 Setting the Object Manager Security

pSeries p630, 655, 670, and 690 preparation

Each pSeries server has two dedicated ports for attachment to the HMC. Keep in mind that the cable distance between the HMC and server is at most 15 m. For every pSeries server or LPAR, you need a uniquely dedicated Ethernet. For the p655 and p630, an integrated Ethernet adapter will do, even when running two LPARs on the p655. For the p670 and 690, you have to have an additional FC 4962 Ethernet adapter for each LPAR. Check for the newest microcode of that adapter at:

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

Also consider having a boot device for each LPAR. The firmware for the p670 and p690 must be at least at RH20413, for the p630, RR20927, and for the p655, RJ020829. Example 6-22 shows how to list the firmware level installed in your machine.

Example 6-22 Obtaining the firmware level on a p655

```
[c59ih01][/]> lscfg -vp | grep -p -e Firmware
      Platform Firmware:
        ROM Level.(alterable).......RJ020829
        Version.....................RS6K
        System Info Specific.(YL)...U1.5-P1-X1/Y1
      Physical Location: U1.5-P1-X1/Y1

      System Firmware:
        ROM Level.(alterable).......RG020805_GA3
        Version.....................RS6K
        System Info Specific.(YL)...U1.5-P1-X1/Y2
      Physical Location: U1.5-P1-X1/Y2
```

If you plan to use the SP Switch2 PCI Attachment Adapter (FC 8397) or the SP Switch2 PCI-X Attachment Adapter (FC 8398), new functionality is included in PSSP that allows the update to a newer microcode level. How to determine whether you need to upgrade is shown in Example 6-23.

Example 6-23 Obtaining information about the SP Switch2 Adapter

```
[c59ih01][/]> /usr/lpp/ssp/css/read_regs -l css0 -X | grep 0x00100030
  0x0C000008F9100030   0x00100030          PCI Trace Reg 1
```

The third nibble can have one of three values, where 8 indicates a properly working adapter in 64-bit mode, 4 indicates that the adapter is not properly stated, and 0 means an update is required. Therefore, the /usr/lpp/ssp/css/xilinx_file_core file is shipped with the firmware for the adapter. After applying PTFs for ssp.basic, you should check for a new version of this file. The update is performed by issuing the following command:

```
/usr/lpp/ssp/css/load_xilinx_cor -l css0 -P -f\
/usr/lpp/ssp/css/xilinx_file_core
```

This can take several minutes and can end with three different messages.

- “This card cannot be field updated.” No update is possible.
- “Reflash not needed.” The card is up to date.
- “Programming function complete.”

If you have the SP Switch2 MX2 Adapter (FC 4026), you have to reboot the node. Otherwise, follow these steps:

1. Quiesce all jobs running on the switch that are using this node.
2. Detach the node with **`Eunfence`**.
3. Detach the network with **`/usr/lpp/ssp/css/ifconfig css0 down detach`**.
4. Stop hats and hags with **`stopsrc -s hats && stopsrc -s hags`**.
5. Kill the Worm on the object node with **`/usr/lpp/ssp/css/css_cdn`**.
6. Issue **`/usr/lpp/ssp/css/ucfgcor -l css0`** to unconfigure SP Switch2 MX2 adapter.
7. Kill all processes using **`css0`** by issuing **`fuser /dev/css0`**.
8. Remove power to the slot by issuing **`echo | /usr/sbin/drslot -R -c pci -l css0 -I`**.
9. Reboot the node.

Note: LPAR resources defined to PSSP need to be uniquely tied to a single LPAR. Therefore, the rootvg, SPLAN adapter, and any other adapter defined to PSSP must be defined to only a single LPAR.

CWS preparation

In contrast to the attachment of a CSP or SAMI protocol server, additional software is required on the CWS to communicate with the HMC:

- csm.clients
- openCIMOM-0.61-1.aix4.3.noarch.rpm
- Java130.xml4j.*
- Java130.rte
- devices.chrp_lpar*

Be sure to obtain the latest level and put the filesets in the correct places in your lppsource, which is the installp/ppc/ subdirectory for installp packages and RPMS/ppc/ for the rpm files. After this, you need to update your SPOT.

Adding an HMC managed server

The following steps highlight what is unique to this type of external node:

1. In a HMC-managed environment, the hardmon daemon does not communicate with the server hardware. It connects to the HMC through the daemon named hmcd running on the CWS. To secure the connection, we need a user ID and a password specified for hardmon. This must be done for every HMC we want to add to the Cluster 1600 system, as shown in Example 6-24.

Example 6-24 Setting hardmon authentication for HMC

```
sp4en0:/
root $ sphmcid sp4hmc hscroot
Password:
Verifying, please re-enter Password:
sphmcid: HMC entry updated.
```

2. We have to add the frame information for every p690 into the SDR. The protocol is HMC in this case. The IP address is the HMC server address. The domain name is the p690 server domain name as it is defined in the HMC. As shown in Figure 6-5, the domain name for the server is Enterprise.

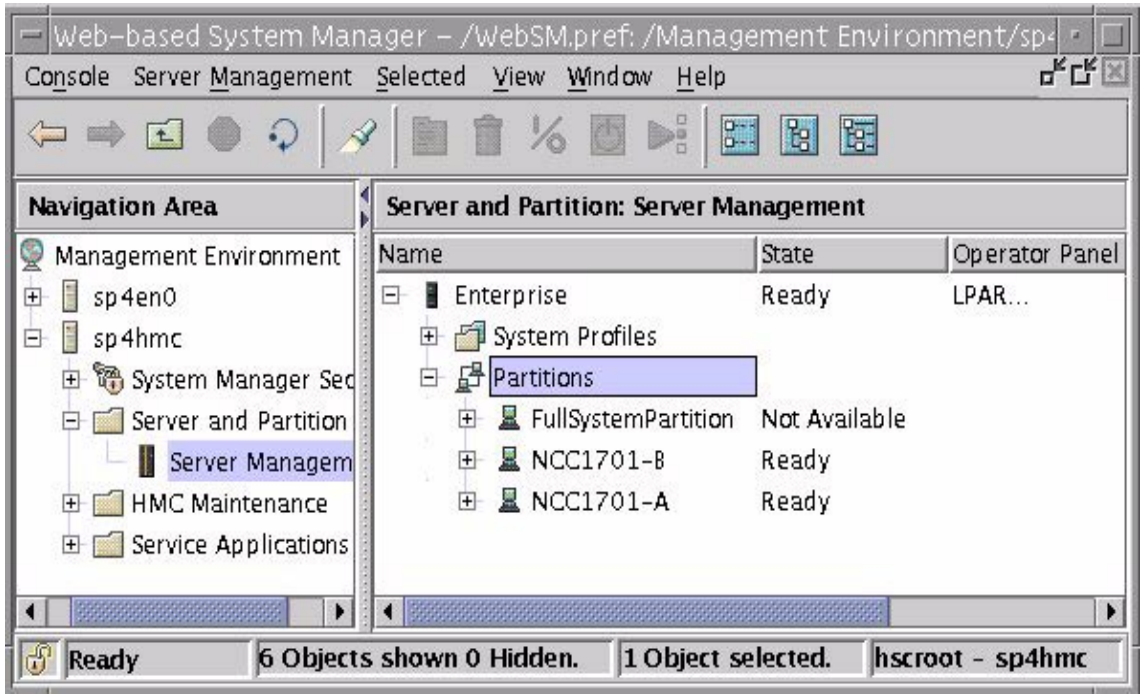


Figure 6-5 Web-based System Manager Console for HMC

Example 6-25 shows the syntax of the **spframe** command. Notice that there is no tty information for HMC frames.

Example 6-25 Adding an HMC-managed frame

```
sp4en0:/
root $ /usr/lpp/ssp/bin/spframe -p HMC -r yes -d Enterprise -i 192.168.4.251 4
0025-322  SDRArchive: SDR archive file name is
/spdata/sys1/sdr/archives/backup.02281.1529
0513-044 The splogd Subsystem was requested to stop.
0513-044 The hardmon Subsystem was requested to stop.
0513-059 The hardmon Subsystem has been started. Subsystem PID is 38238.
0513-059 The splogd Subsystem has been started. Subsystem PID is 40846.
SDR_config: SDR_config completed successfully.
sp4en0:/
root $ splstdata -f
                 List Frame Database Information

frame# tty      s1_tty frame_type hardware_protocol control_ipaddrs domain_name
------ -------- ------ ---------- ----------------- --------------- -----------
     1 /dev/tty0 ""      switch    SP                ""              ""
     2 /dev/tty1 ""      ""       CSP               ""              ""
     3 /dev/tty2 ""      ""       CSP               ""              ""
     4 ""        ""      ""       HMC               192.168.4.251   Enterprise
```

3. The nodes are added automatically and the hmcd daemon started for the frame on the CWS, as show in Example 6-26. At this moment, there is not much information about the nodes. The LPAR name is shown at the end of the line for each node.

Example 6-26 SDR node information and hmcd daemon

```
sp4en0:/
root $ splstdata -l 49,50
                List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- --------------- --------- --------------- ----------
   49      4     1     1 ""                ""                 ""                ""
MP                                 1 ""                     0 false           NCC1701-A
   50      4     2     1 ""                ""                 ""                ""
MP                                 1 ""                     0 false           NCC1701-B
sp4en0:/
root $ spmon -d
...
Lines omitted
...
```

```
---------------------------------- Frame 4 ----------------------------------
                        Host     Switch   Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- -------- ------- ----- ---------------- -------
  1    49  thin  off     no     noconn    N/A    N/A  LCDs are blank     N/A
  2    50  thin  on      no     noconn    N/A    N/A  LCDs are blank     N/A
```

4. The next step is to check the adapter slot information for the SP Ethernet. For this, run the **spadaptr_loc** command. This command obtains the physical location codes for SP-configurable adapters and it also collects the hardware addresses. The nodes will be powered off by the command. No tty connection can be open when running this command. This command is useful when there is a new node added to the system. For an operating node, we should use another method. On the running LPAR, there is a command called **lsslot** to show adapter location codes. The port number has to be added to the slot ID when we configure the adapters into the SDR. If the command gives **U1.9-P2-I3**, and this is a single port Ethernet adapter, we should use **U1.9-P2-I3/E1** as the physical location code. In the case of a four-port adapter, use **E"*port number*"**. Instead of the adapter name, in this case, we have to use the adapter type **en** in the **spadaptrs** command, as shown in Example 6-27.

Example 6-27 The spadaptrs command

```
sp4en0:/
root $ /usr/lpp/ssp/bin/spadaptrs -P U1.9-P2-I3/E1 -e 192.168.4.250 -t tp -d full -f 100 4 2 1
en 192.168.4.50 255.255.255.0
sp4en0:/
root $ splstdata -n 4 2 1
                  List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- --------------- --------- --------------- ----------
   50      4     2     1 sp4n50e0          sp4n50e0           ""                 192.168.4.250
MP                                  1 ""                        0 false           NCC1701-B
sp4en0:/
root $ splstdata -a 4 2 1
                  List LAN Database Information

node# adapt  netaddr         netmask         hostname          type         t/r_rate enet_rate
duplex other_addrs     adapt_cfg_status  physical_location SPLAN
----- ------ --------------- --------------- ----------------- ------------ -------- ----------
------ --------------- ----------------- ----------------- -----
   50 en     192.168.4.50    255.255.255.0   sp4n50e0          tp           NA       100
full   ""              ""                U1.9-P2-I3/E1        1
```

Notice that the HMC console shows the location codes in a different format. See Figure 6-6.

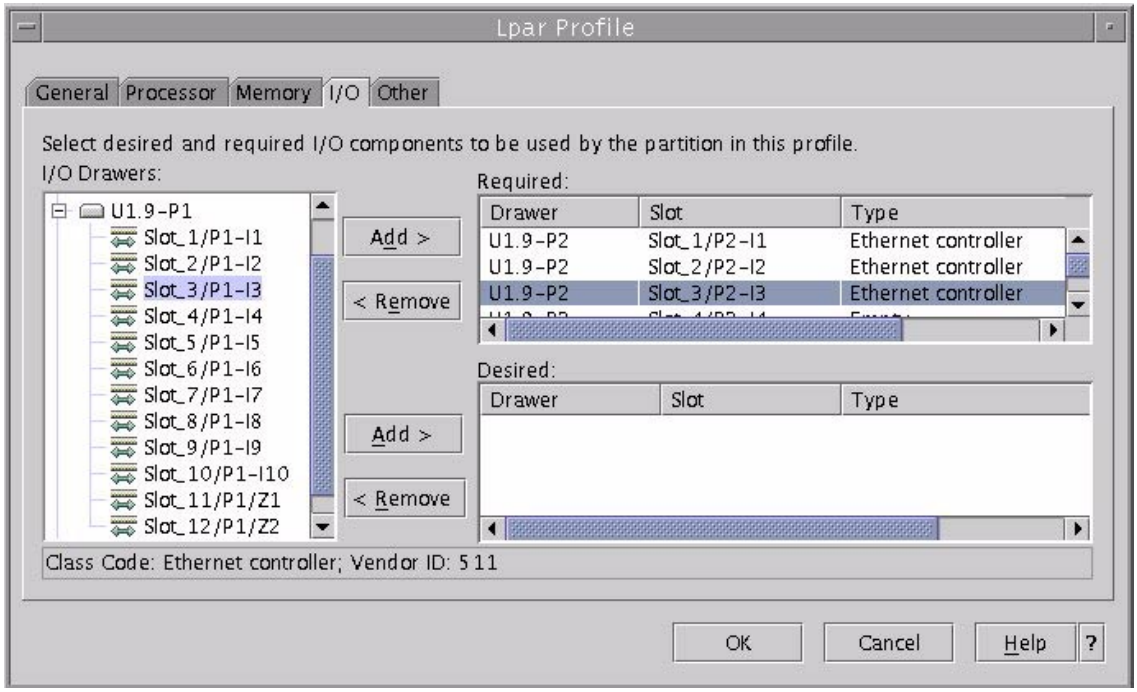


Figure 6-6 Hardware Management Console LPAR I/O profile

5. The hardware addresses can be queried with the **`netstat -I`** command or with the **`lscfg -vl ent2`** command, and they have to be listed in the /etc/bootptab.info file on the CWS for every operational node. We have to use the format as it is listed in the output of the **`lscfg`** command.
6. For the SP Switch definition, we can use the adapter name css0. The **`spadaptrs`** command use is the same as for an SP node.
7. Add the rootvg information to the SDR in the usual way. For an operational LPAR, set the node to customize. For a new LPAR, set it to install. These steps are the same for all Cluster 1600 nodes.
8. Run **`setup_server`**. In Example 6-28 on page 162, we show the output of setup_server. We had several error messages. The reason was that we added node information to only one node (LPAR) from the two that were defined in the HMC for our p690 server. The **`spframe`** command, however, created all the nodes but did not specify any networking attribute. For node 49 there was no *reliable hostname* and *lppsource* information. At this time, setup_server does not provide checking mechanism to exclude the nodes

with missing information. We had to define all the LPARs that were available with a completed attribute list to the SDR and rerun **setup_server**.

Example 6-28 The setup_server output

```
sp4en0:/
root $ setup_server
setup_server: There is no reliable hostname assigned to node 49.
setup_server: No NIM resources will be allocated for node 49.
setup_server: Running services_config script to configure SSP services.This may
take a few minutes...
...
Lines omitted
...
mknimmast: Node 0 (sp4en0) already configured as a NIM master.
create_krb_files: 0016-428 Can not create the client srvtab file for node
number 49. No host name information was found in the SDR.
create_krb_files: tftpaccess.ctl file and client srvtab files created/updated
on server node 0.
...
Lines omitted
...
0042-001 nim: processing error encountered on "master":
   0042-001 m_mk_lpp_source: processing error encountered on "master":
   0042-154 c_stat: the file or directory
"/spdata/sys1/install/default/lppsource" does not exist
mknimres: 0016-375 The creation of the lppsource resource named
lppsource_default
 had a problem with return code 1.
setup_server: 0016-279 Internal call to /usr/lpp/ssp/bin/mknimres was not
successful; rc= 2.
Tickets destroyed.
setup_server: Processing incomplete (rc= 2).
```

9. To finish the operational LPAR integration, run the steps of a normal node conditioning:
   a. Copy or ftp /etc/SDR_dest_info from the CWS to the node.
   b. Mount pssplpp from the CWS.
   c. Install ssp.basic.
   d. Run **pssp_script**.
   e. Reboot the LPAR.
   f. Update all the PSSP filesets with the newest PTFs on the media.
10. For a new LPAR installation, run **nodecond** for the node on the CWS.

11. Check the host responds and switch responds for the new nodes. Run the verification tests listed in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

6.4.3 S70 Enterprise Server

In this scenario, we have an SP frame with two high nodes and an SP Switch2 in Cluster 1600. We add one S70 as a SAMI attached node to Cluster 1600.

Hardware considerations

Before you integrate a pSeries p680 or an Enterprise Server 7017-S70, S80, or S85 into a Cluster 1600, a special SAMI interface (FC 3150, FC 3151) must be integrated into the machine. This provides the connection to the serial port of the CWS for hardware control, and the first integrated serial line is connected to the CWS for use by **`s1term`**. Additionally, you have to provide a management Ethernet adapter, which has to be in slot 1.

Tip: We recommend upgrading the firmware of the S70 and S7A to at least 20020514 (Sys) and 20010824 (SvP), and for the 7017-S80 and S85 and the p680 to at least 20020411 (Sys) and 20020411 (SvP).

Integrating a SAMI protocol server

Example 6-29 shows the initial configuration of our Cluster 1600.

Example 6-29 S70 attach initial configuration

```
root $ spmon -d
---------------------------------- Frame 1 ----------------------------------
                      Host     Switch   Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- -------- ------- ----- ---------------- -------
  1     1  high   on     yes      yes     N/A     no  LCDs are blank      no
  5     5  high   on     yes      yes     N/A     no  LCDs are blank      no

root $ splstdata -n
                  List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- -------------- --------- --------------- ----------
    1      1     1     4 sp4n01e0          sp4n01e0          ""                192.168.4.250
MP                              16 375_MHz_POWER3_         1 true            ""
    5      1     5     4 sp4n05e0          sp4n05e0          ""                192.168.4.250
MP                              16 375_MHz_POWER3_         1 true            ""
```

The following steps highlight what is unique to these external nodes:

1. SAMI attached nodes require an extra serial connection, one to the SAMI card and one to the integrated serial port. The SAMI connection is for talking to the supervisor card and the other for console login. We prepared these connections and the connection to the SP LAN. Each SAMI attached node is seen as an individual frame containing one node. Example 6-30 shows us adding the new frame.

Example 6-30 Adding the CSP frame

```
root $ /usr/lpp/ssp/bin/spframe -s 'tty2' -p 'SAMI' -r 'no' 2 1 tty1
root $ splstdata -f
                 List Frame Database Information

frame# tty              s1_tty           frame_type      hardware_protocol  control_ipaddrs
domain_name
------ ---------------- ---------------- --------------- ------------------ ---------------
-----------
     1 /dev/tty0       ""              switch          SP                 ""             ""
     1 /dev/tty1       /dev/tty2        ""             SAMI               ""             ""

root $ splstdata -n
                 List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- --------------- --------- --------------- ----------
    1      1     1     4 sp4n01e0          sp4n01e0          ""                192.168.4.250
MP                              16 375_MHz_POWER3_         1 true            ""
    5      1     5     4 sp4n05e0          sp4n05e0          ""                192.168.4.250
MP                              16 375_MHz_POWER3_         1 true            ""
   17      2     1     1 ""                ""                ""                ""
MP                               1 ""                      0 false           ""
```

   - **spmon -d** shows the node type as **extern**.
   - **splstdata -f** shows the **hardware_protocol** as **SAMI**.
   - **splstdata -n** lists basic node configuration information.

2. The next step is to configure the SP LAN information. Example 6-31 on page 165 shows this information being added with the **spadaptrs** command.

Example 6-31 Configuring the S70 SP LAN information

```
root $ /usr/lpp/ssp/bin/spadaptrs -e 192.168.4.250 -t tp -d half -f 10 2 1 1  en0 192.168.4.17
255.255.255.0
root $ splstdata -n -l 33
                  List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname      default_route
processor_type processors_installed description     on_switch primary_enabled LPAR_name
----- ------ ----- ----- ----------------- ----------------- ----------------- ---------------
-------------- -------------------- --------------- --------- --------------- ----------
   17      2     1     1 sp4n17e0          sp4n17e0           ""                192.168.4.250
MP                                 1 ""                        0 false           ""
```

3. Add any adapters that you require with the **spadaptrs** command.
4. Next, configure the boot and install information for the new node. We need to set the install disk, install image name, lppsource, PSSP version, and the MAC address of the Ethernet card we are using on the SP LAN. Because this is a new node, we can use the **sphrdwrad** command to probe the machine for the MAC address. This requires the machine to be rebooted. If you have an existing node you want to integrate into the cluster, you should use the /etc/bootptab.info file to set the MAC address so that **phrdwrad** does not actually need to probe the machine.

Example 6-32 Preparing S70 boot and install information

```
root $ splstdata -b -l 17
                  List Node Boot/Install Information

node# hostname          hdw_enet_adr  srvr response     install_disk
last_install_image  last_install_time   next_install_image  lppsource_name pssp_ver
selected_vg
----- ----------------- ------------- ---- ------------ ----------------------
------------------- ------------------- ------------------- -------------- -------------------
-------------------
   17 sp4n17e0          000000000017     0 install      hdisk0                 initial
initial             default             default        PSSP-3.5            rootvg
root $
root $ /usr/lpp/ssp/bin/spchvgobj -r rootvg -h hdisk0 -c 1 -n 0 -i mksysb.51f_64 -v aix51 -p
PSSP-3.5 2 1 1
spchvgobj: Successfully changed the Node and Volume_Group objects for node number 17, volume
group rootvg.
spchvgobj: The total number of changes successfully completed is 1.
spchvgobj: The total number of changes which were not successfully completed is 0.
sp4en0:/
root $ splstdata -b -l 17
                  List Node Boot/Install Information
```

```
node# hostname          hdw_enet_adr  srvr response     install_disk
last_install_image  last_install_time   next_install_image  lppsource_name pssp_ver
selected_vg
----- ---------------- ------------ ---- ------------ ----------------------
------------------ ------------------ ------------------ -------------- ------------------
------------------
   17 sp4n17e0         000000000017     0 install      hdisk0                 initial
initial             mksysb.51f_64       aix51          PSSP-3.5            rootvg

root $ /usr/lpp/ssp/bin/sphrdwrad 3 1 1
Acquiring hardware Ethernet address for node 17
Hardware ethernet address for node 33 is 000629DC5904
Ping to default_route successful for node 17.

root $ splstdata -b -l 17
                  List Node Boot/Install Information

node# hostname          hdw_enet_adr  srvr response     install_disk
last_install_image  last_install_time   next_install_image  lppsource_name pssp_ver
selected_vg
----- ---------------- ------------ ---- ------------ ----------------------
------------------ ------------------ ------------------ -------------- ------------------
------------------
   17 sp4n17e0         000629DC5904     0 install      hdisk0                 initial
initial             mksysb.51f_64       aix51          PSSP-3.5            rootvg
```

5. The last set of PSSP preparation is to run **setup_server** to configure the SP system for the new node.
6. Start the network install of the new node with the **nodecond** command.
7. Adding the node where we want to keep the existing installation differs from the above, because we set the node to *customize* instead of install. The command is **spbootins -s no -r customize 2 1 1**. After this, the PSSP software installation is completed by running the **pssp_script** on the node. The high-level steps for this are as follows:
   a. Mount /spdata/sys1/install/pssplpp/PSSP-3.5 from the CWS.
   b. Install ssp.basic on the node.
   c. Copy /etc/SDR_dest_info from the CWS to the node.
   d. Run **/usr/lpp/ssp/install/bin/pssp_script** on the node.
   e. Check the switch communication if you configured a switch adapter. **Eunfence** the node if necessary.
8. Use **spmon -d** to check the host response and switch response for the new nodes. Run the verification tests listed in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

6.5 Migration tips

The following list contains tips that are well documented in the *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281 and in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347. However, we would like to collect them in one place for your consideration before and during the migration:

- Read the *Read This First* document before doing any migration or integration activity. The latest version of the PSSP documentation can be found at:

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

- Use *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281, for planning the maintenance.
- Create a system backup before any maintenance activity.
- Archive the SDR on the CWS before any PSSP-related maintenance activity.
- Use the *PSSP for AIX: Installation and Migration Guide*, GA22-7347 as a reference for all steps of the migration activity. It contains information about the new steps you have to do because of newly introduced features in the software.
- Check the error logs and PSSP log files both on the CWS and the nodes for any hidden problems. They can cause lots of trouble during migration.
- Ensure that all the PSSP subsystems are running well, and the communication between the nodes and the CWS is running without any problem before the migration.
- Check available disk space and space requirements before installation. The AIX migration can stop and wait for intervention when the space on the disks in rootvg is not enough for the new software level.
- The install process extends the file system with the amount of space just enough for that installation. After migration, check the free space in the /usr and /var file systems again and extend the file systems needed.
- Monitor your running migration using **`tail`** on the appropriate log file. A comprehensive list of log files can be found in the “Using the SP logs” section in Chapter 4 of the *PSSP for AIX: Diagnosis Guide*, GA22-7350.
- Create and use scripts for activities on multiple nodes to avoid typing mistakes. Consult smit.log and smit.script in the users home directory for help.
- Deactivate and export any non-rootvg volume groups on the nodes before migration activities.

- Stop any application on the node before starting the migration process. The node customization reconfigures the network adapters while it is running, and this can cause problems for running applications. Stop HACMP on the nodes as well to prevent any unwanted takeover.
- After the CWS migration, do the migration steps on a test node first, and test the node before changing the other nodes.
- Do one step at a time and do not change anything else in the system configuration until the migration finishes.
- After migrating the CWS to AIX 5L Version 5.1, prepare the AIX lppsource before the start of the maintenance window for the node migration. The PSSP lppsource can be created before the CWS AIX migration if desired.
- The first copy of the mksysb images from the PSSP product CD must be done by installing the files to the CWS. After this, it is possible to copy these files to other machines.
- We had problems running the **pssp_script** in the background. If you have to do this, check the LED for the node, and if the node hangs at code c42, bring it into the foreground.
- The final step in AIX migration from CD is accepting the licence agreement. The system will wait for this after copying the last CD.
- Use the **lsslot** command to collect network adapter location code information for HMC-managed LPARs. Put the ***E"port number"*** after the location code the command returns, where port number is the port that is used on that card for the specific connection. The port number is always 1 for a one-port adapter.
- The hardware addresses for the adapter PSSP uses for node installation can be collected using the **sphrdwrad** command. Running this command can take several minutes, and it needs to restart the node. If the nodes are running before installation, collect the hardware addresses and put them into the /etc/bootptab.info file on the CWS.
- Use **spled** to monitor the migration and customization activities. For code explanation, refer to "SP-specific LED/LCD values" in Chapter 32 of the *PSSP for AIX: Diagnosis Guide*, GA22-7350.
- After copying any PTFs to lppsource, update the SPOT from lppsource.
- The setup_server configures the boot/install server for all defined nodes in the SP complex. If some boot/install information is missing, the script fails. This means that all LPARs from HMC-based frames must be fully configured before setup_server starts.

7

Cluster 1600 management: PSSP and CSM

This chapter discusses concepts relating to managing a Cluster 1600 with Parallel System Support Program (PSSP) and the current features of Cluster Systems Management (CSM). In this chapter, we introduce the CSM components and briefly describe the similarities and differences between CSM and PSSP. The cluster functions provided with AIX 5L Version 5.2 in 2002 are just the beginning. There are many enhancements planned for future releases of the cluster software.

Attention: More information regarding the Cluster 1600 management software will be provided in later documentation releases. For additional documentation about AIX 5L Version 5.2 and Cluster 1600 hardware, software, and peripherals, refer to:

`http://www.ibm.com/redbooks`

We discuss the following topics:

- PSSP and CSM for cluster management
- A brief comparison of PSSP and CSM for AIX
- PSSP and CSM for cluster management
- Cluster 1600 assistance

7.1 PSSP and CSM for cluster management

The PSSP distributed server management technology has been used for commercial and high-performance computing environments for server and workload consolidation through a single-point-of-control for years. As more and more customers take advantage of the clustering systems management capability of PSSP and pSeries servers, it has become clear that IBM cluster technology has a role in any new, medium- or large-sized pSeries installation.

Recognition of this has led to the integration of much of this software into AIX itself. AIX 5L Version 5.1 and 5.2 contain Reliable Scalable Cluster Technology (RSCT), a distributed cluster-enabling layer offering a highly available view and control of resources and events throughout the cluster from any place within the cluster.

IBM is actively developing software to exploit this technology. One example is GPFS on Cluster 1600, which with PSSP 3.5 and AIX 5L Version 5.1, no longer needs HACMP or VSD as a prerequisite.

Another example is CSM for AIX 5L Version 5.2. It utilizes RSCT, a fair amount of PSSP-based technology, and well-proven open source software to provide much of the usefulness of PSSP in software bundled with AIX.

This is an advantage for customers for several reasons:

- The clustering software is shipped directly with AIX.
- The cluster management function is no longer logically tied to a particular hardware model (SP frames, nodes, and so on) and can be used within more generic clusters, including those with Linux OS, whether Intel- or POWER4-based servers.
- The cluster software is introduced with the OS, instead of being a separate product.
- The look and feel of the product is integrated with AIX.

CSM 1.3 for AIX is available and shipping with AIX 5L Version 5.2. Should customers consider using CSM instead of PSSP for building their Cluster 1600? (Note that you cannot use both at the same time.) Although PSSP is in its last release, customers need to understand the differences between the products until CSM provides the same functionality as PSSP.

Attention: CSM will be replacing PSSP as IBM's main clustering software.

PSSP 3.5 is intended for existing SP and Cluster 1600 customers using PSSP, as well as new High Performance Computing (HPC) cluster customers, while CSM is intended for new Cluster 1600 customers in the commercial and HPC space.

PSSP currently supports the software that High Performance Computing (HPC) customers require. This includes GPFS, LoadLeveler, Parallel Environment, Parallel ESSL, and ESSL. CSM for AIX 5L Version 5.2 does not support the HPC software stack at this time, although IBM intends to provide support for the HPC software stack in 2003.

PSSP also contains support for SP switch interconnects, such as the SP Switch and the SP Switch2. Customers using these need PSSP, because CSM does not support the switch technology at this time. The next switch generation will be supported by CSM.

Customers primarily interested in total cost of ownership improvements from Cluster 1600 manageability, and with no need for connection to an SP, SP-attached nodes, or to the switch technology, may be interested in CSM.

Attention: CSM is in its infancy and IBM is planning “final parity” with PSSP in the AIX 5.3 time frame.

CSM is shipped with AIX 5L Version 5.2 and supplied *try-and-buy* for those who are new to clustering and want to try it out. PSSP (3.5 only) will be available on AIX 5L Version 5.2 later in 2003. Because new customers will probably want to make use of the new features of AIX 5L Version 5.2, such as dynamic logical partitioning on the pSeries POWER4 machines, CSM is the better choice if the cluster is needed now.

7.1.1 A brief comparison of PSSP and CSM for AIX

Attention: PSSP and CSM share similar concepts but contain very different interfaces. Additional PSSP features are being considered for future introduction into CSM. However, this does not guarantee future availability.

CSM utilizes much of the well-received technology that has enabled over 10,000 PSSP customers to manage systems from small half-frame size to hundreds of nodes containing thousands of processors.

CSM offers PSSP-like functionality, such as:

- Central Point of Control: A single machine, configured as a CSM management server, can control and manage the entire cluster. This is analogous to the PSSP control workstation. The management GUI offered by PSSP, Perspectives, is not used in CSM. It is replaced by the AIX WebSM plug-in technology, bringing the cluster management more into the mainstream AIX look and feel.
- Manage cluster membership and attributes: Nodes can be added or deleted from the cluster. The CSM management server maintains a database containing information about the attributes of the nodes, and these attributes can be set, displayed, and used by the management server to more efficiently manage the nodes. The AIX RSCT cluster registry is used as the management database.
- Monitor cluster-wide hardware and software state: PSSP includes the event management and problem management subsystems, which together, allow monitoring and automated actions based on many hardware and software attributes across the cluster. Hardware, operating system, and application events and conditions on the nodes throughout the cluster can be configured for monitoring and alerting from the central console. Provided and user-written software can respond to these events in customizable ways, such as sending SNMP traps, running scripts, or issuing pager alerts. CSM has similar functionality, but it is enhanced through the exploitation of the AIX RSCT software.
- Node software installation: Node software, including the OS, can be installed from the management server, possibly across multiple nodes simultaneously. After installation, custom scripts can be run on the nodes, built from the configuration information supplied from the cluster database. PSSP uses AIX Network Installation and Maintenance (NIM) under the covers with scripts such as setup_server. CSM exposes NIM functionality more directly. This eases problem determination if the install should fail and allows for greater use of the rich set of features of NIM.
- Cluster command execution: Tools are provided to run commands concurrently across nodes in the cluster. The PSSP **dsh** command has been made available in AIX, is supplied in CSM, and is enhanced to enable use with **ssh**, as well as the default **rsh**.
- Cluster file management: Typically used with configuration and system files, this feature ensures that changes in the managed files are propagated across the correct set of nodes in a timely fashion, without user intervention. Both PSSP and CSM offer this function. PSSP uses its *file collections* component, whereas CSM utilizes the rdist open source package.

- Node grouping: Sets of nodes can be given names, and these names can be used to refer to and manage the nodes in the set as a unit. This is useful when a subset of nodes is running a particular application, such as HTTP serving, and needs to be managed based on common characteristics. PSSP offered this within Perspectives. CSM offers *dynamic grouping*. A dynamic group creates a group defined by a set of node attributes, instead of a list of node names. When nodes join the CSM cluster, they automatically are added to the proper dynamic group, or when node attributes change, they are automatically removed or added to the appropriate group.
- Cluster diagnostics: Software may be provided to collect logs from groups of nodes to a central location. Cluster-specific software may be instrumented for diagnostic purposes. PSSP and CSM provide this feature.
- Cluster security: Cluster-wide policies on user and process authentication and authorization can be configured and maintained from the management server in a secure way. Security is implemented differently in CSM. Instead of trusted third-party authentication with Kerberos, host-based authentication using public/private keys is used.
- Scalability: Both PSSP and CSM for AIX support 128-way clusters, with larger clusters available through special bids.
- **`dsh`** offers improved performance, DCEM, a user-friendly, WebSM-based, GUI wrapper for **`dsh`**, and the option of using **`ssh`** instead of **`rsh`** as the underlying command transport.

In comparison to PSSP, CSM does not provide the following at this time:

- SP Switch interconnect technology. CSM will provide support for the next generation of IBM switch technology.
- Support of the HPC software stack. CSM will provide support for the HPC software stack in 2003.
- Hardware control on legacy SP node and cluster-attached servers. CSM 1.3 supports hardware control only on HMC-based servers and LPARs at this time. PSSP will continue to support all the servers it does now, which include SP nodes, non-HMC-based cluster-attached servers, as well as HMC-based servers.
- Network Time Protocol (NTP) cluster configuration is not available, although it may be offered in the future.
- Special cluster-wide management for the AIX user IDs and passwords used on the nodes.
- Cluster-wide startup/shutdown commands.
- High-availability management server function, although this may be offered in the future.

- Cluster-wide accounting.
- The pcommands, parallel administrative commands, such as **pfind**, **p_cat**, **pexec**, and so on.

In comparison with PSSP, CSM adds:

- The ability to manage Linux/xSeries nodes from the central cluster console, as well as AIX/pSeries nodes. The management server must run AIX 5L Version 5.2.
- Tighter integration with AIX through RSCT and WebSM.
- Faster cluster boot time, which will improve ultimate scalability over time.

7.2 Decision trees

Here are two decision maps to evaluate which Cluster 1600 management software meets your requirements at this point in time. Figure 7-1 on page 175 provides a decision tree to help you evaluate whether a Cluster 1600 managed by PSSP or a Cluster 1600 managed by CSM is the appropriate solution to your clustering requirements in the 2002 to first half 2003 time frame.

Attention: These decision trees may not request all the necessary information to evaluate which Cluster 1600 management software is required. However, they do provide a set of questions to evaluate and provide initial guidelines about which management software, PSSP or CSM, your Cluster 1600 may require for your individual needs.

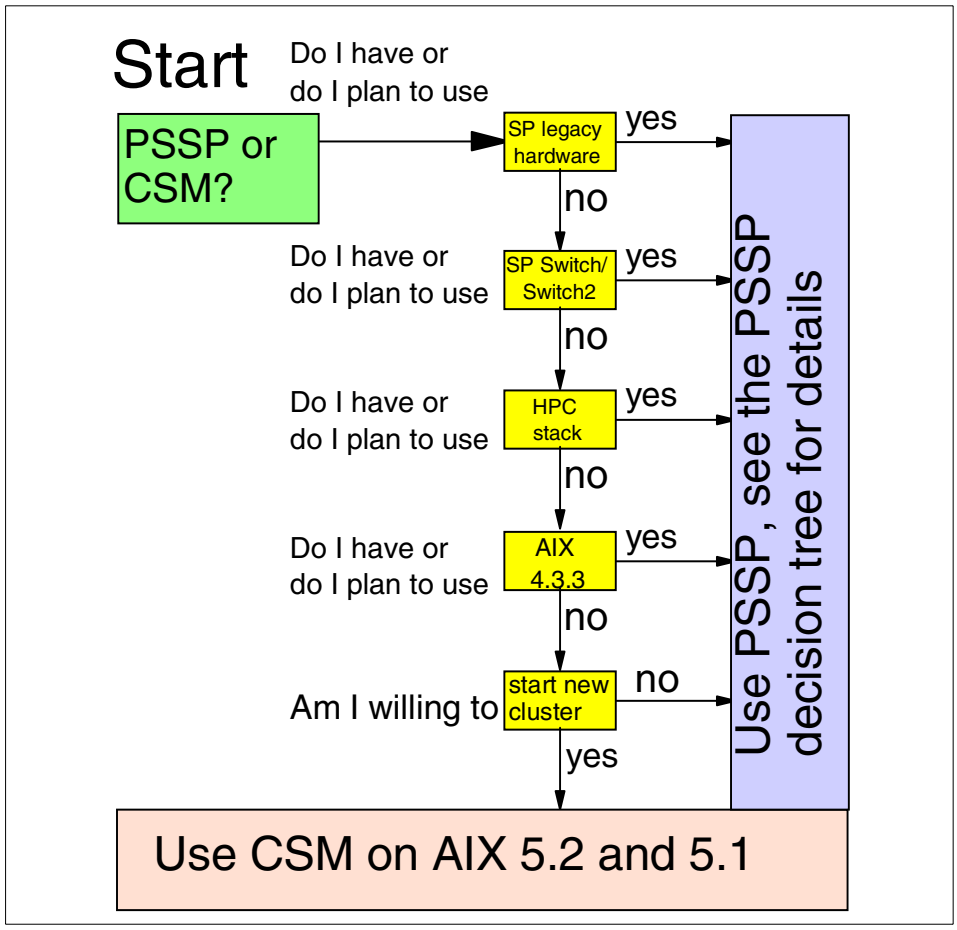


Figure 7-1 Considerations when planning Cluster 1600 in 2002-03 time frame

Figure 7-2 on page 176 helps you decide which release of PSSP supports a given Cluster 1600.

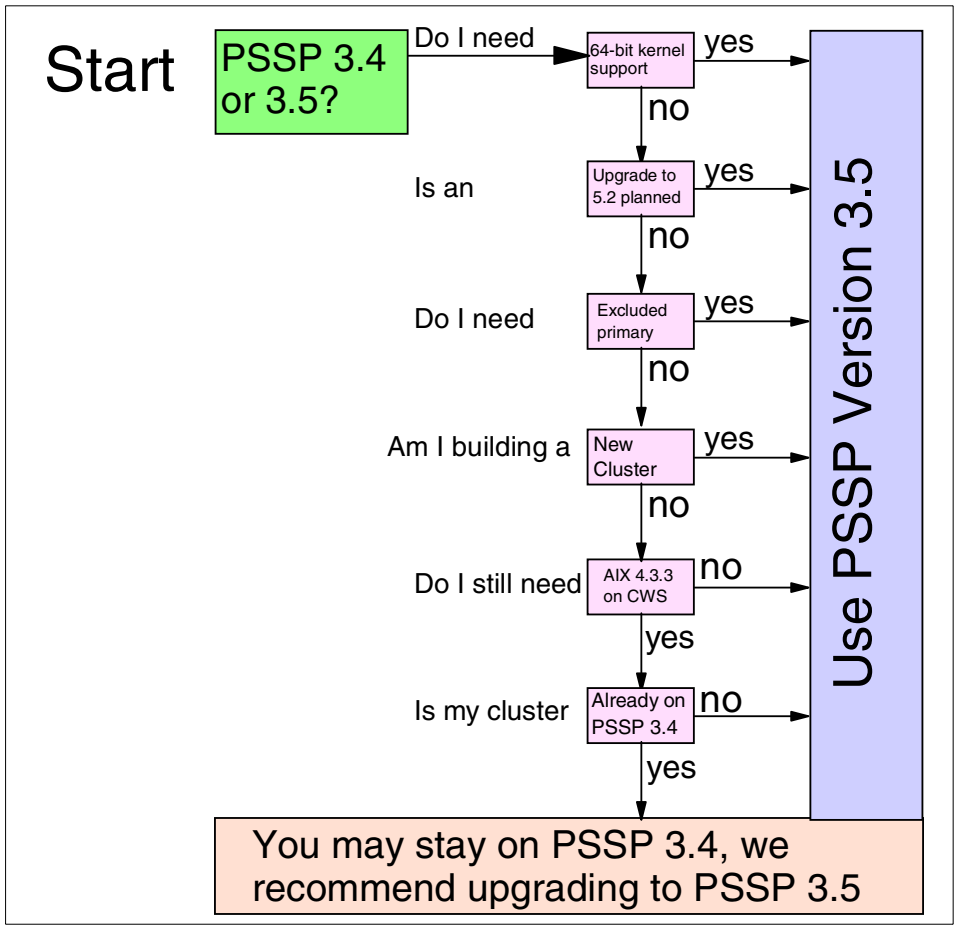


Figure 7-2 Considerations when planning Cluster 1600 managed by PSSP

7.3 Cluster 1600 assistance

IBM provides customers with a variety of documentation, education, and services to evaluate current and future Cluster 1600 management software requirements. Please contact your local IBM contact for additional information about the current plans for the Cluster 1600 management software.

A

Cluster 1600 scalability rules

This appendix describes maximum sizing and scalability rules for the Cluster 1600.

Cluster 1600 scaling

A Cluster 1600 with PSSP can consist of 2 to 128 AIX OS images (the 128 limit is due to availability of systems for IBM labs to test, higher scalability configurations are available by special bid).

An AIX OS image is defined as:

- A 7040 server (p690/p670) running as a full system partition or a 7028 server (p630)
- An LPAR of a 7040 server
- A 7026 server (H80, M80, p660, 6H0/6H1, 6M1)
- A 7017 server (S70, S7A, S80, p680)
- An SP node

The following scalability rules apply:

- No more than 128 AIX OS images from the set (7040, 7028, 7026, 7017, 9076)

- No more than 32 physical servers from the set (7040)
- No more than 64 physical servers from the set (7026, 7028, 7040, 7017)
- No more than 16 physical servers from the set (7017)
- No more than 128 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
- 12 p690s with 4 LPARs each, 16 p680s, 16 p660s, and 48 SP nodes

B

Sample switch management script

This appendix contains an example script shown in Example B-1 on page 180 to check if the oncoming primary and the oncoming primary backup assignments are assigned to nodes that have no switch response. It then changes this by issuing the **Eprimary** command. The syntax is as follows:

```
check_primary.sh [-h][-e][-p oncoming_primary oncoming_primary_backup]
```

The script has the following options:

- -h: Help.
- -e: The **Eprimary** command is not performed, all other checks were made. This is for simulating the behavior of this script.
- -p: Pretend, for testing purposes, that these two node numbers will become primary and primary backup. This changes nodes even if they have no switch response.

Attention: This is not an official IBM script, it is just an example. Although it has been successfully tested in our environment, no guarantee is given or implied.

Example: B-1 The check_primary.sh script

```
#!/bin/sh
#
# check_primary.sh: Script to check if oncoming primary and oncoming primary
backup are
#                   valid and have script response
#
# Version 1.1 10212002 RK
#
# Change history
# Version 1.1: New Flag -e to show command execution instead of doing it, RC
added
# Version 1.0: Initial release
#
# Definition of global variables
# OPRIMARY: Holds the node number of the oncoming primary
# OBACKUP:  Holds the node number of the oncoming primary backup
#
# EPRIMARY: The Eprimary command with path
# SDRGET:   The SDRGetObjects command with path
#
# SWREP:    Switch response of oncoming primary as recorded in the SDR
# SWREPB:   Switch response of oncoming primary backup as recorded in the SDR
# PSSPV:    Version of PSSP runninag on the system
#
# FIELD:    Variable for differentiating a Switch and Switch2 environment, where
the Switch2 env. can have
#           dual plane
# RESPCMD:  SDR attribute containing the switch_response or switch_response0
# RC:       Event counter
# PRETEND:  Flag if command is executed (1) or not (0)
#
# LIMITATIONS: This script applies to PSSP 3.2, 3.4 and 3.5. Only plane 0 in a
dual plane env. is checked!
#              Program does not check if a switch is installed
#
# DISCLAIMER: This script is provided as is. It is public domain and not part of
any IBM product, no software
#             service applies and IBM is not reliable for any damage it may
cause.
#
# Note: The main function is at the end of the script!

# 1. Clear all variables and set the defaults
OPRIMARY=""
OBACKUP=""
EPRIMARY=""
SDRGET=""
SWREP=""
```

```
SWREPB=""
PSSPV=""
FIELD=1
RESPCMD="switch_responds"
RC=0
PRETEND=1

#
# Subroutine func_checkenv checks for a valid system and figures out which
switch is running
#
func_checkenv(){
    # See if we have an Eprimary command and assign it
    if [ ! -f /usr/lpp/ssp/bin/Eprimary ]; then
   echo "0000-01 Eprimary command not found!"
   echo "No PSSP installed?"
   exit 1
    fi
    EPRIMARY="/usr/lpp/ssp/bin/Eprimary"

    # See if we have an SDRGetObject command and assign it
    if [ ! -f /usr/lpp/ssp/bin/SDRGetObjects ]; then
   echo "0000-02 SDRGetObjects not found!"
   echo "No PSSP installed?"
   exit 1
    fi
    SDRGET="/usr/lpp/ssp/bin/SDRGetObjects"

    #See, if we run in a supported env and which switch we have
    PSSPV=`/usr/lpp/ssp/bin/splst_versions -t -n0 | cut -f 2 -d " "`
    STYPE=`$SDRGET -x Switch switch_type`

    #Assign the command to use with different PSSP and SWITCH types
    case $PSSPV in
   "PSSP-3.5")
       echo "PSSP 3.5 with switch type " $STYPE "found"
       if [[ $STYPE -eq 132 ]]; then
      FIELD=4
      RESPCMD="switch_responds0"
       fi
       return
       ;;
   "PSSP-3.4")
       echo "PSSP 3.4  with switch type " $STYPE "found"
       if [[ $STYPE -eq 132 ]]; then
      FIELD=4
      RESPCMD="switch_responds0"
       fi
       return
       ;;
```

```
   "PSSP-3.2")
       echo "PSSP 3.2 found"
       FIELD=1
       RESPCMD="switch_responds"
       return
       ;;
    esac
    echo "000-09 Version " $PSSPV " not supported"
    exit 1
}

#
# func_getcurrent gets the current oncoming primary and the oncoming primary
backup
#
func_getcurrent(){
    OPRIMARY=`$EPRIMARY | grep "oncoming primary" | grep -v backup | cut -f
$FIELD -d " "`
    OBACKUP=`$EPRIMARY | grep "oncoming primary backup" | cut -f $FIELD -d " "`
}

#
# func_checkcurrent check the current oncoming primary and the oncoming primary
backup for switch responds
#
func_checkcurrent(){
    SWREP=`$SDRGET -x -d . switch_responds node_number==$OPRIMARY $RESPCMD`
    SWREPB=`$SDRGET -x -d . switch_responds node_number==$OBACKUP  $RESPCMD`
}

#
# func_changeprimary assigns new primary and primary backups
#
func_changeprimary(){

    # $1 includes the new primary and backup in a format as 1.7
    NEWPRIMARY=`echo $1 | cut -f 1 -d.`
    NEWSECONDARY=`echo $1 | cut -f 2 -d.`

    # Commands run seperately for different rc, maybe SEC does not exist
    if [[ $PRETEND -eq 1 ]]; then
   if [[ -n $NEWPRIMARY ]]; then
       $EPRIMARY -init $NEWPRIMARY
   fi
   if [[ -n $NEWSECONDARY ]]; then
       $EPRIMARY -backup $NEWSECONDARY
   fi
    else
   if [[ -n $NEWPRIMARY ]]; then
       echo "Eprimary -init " $NEWPRIMARY
```

```
   fi
   if [[ -n $NEWSECONDARY ]]; then
       echo "Eprimary -backup "  $NEWSECONDARY
   fi
    fi
}

#
# func_action tries to assign new nodes if neccessary
#
func_action(){
    LOCALRES=""
    let z=0
    if [[ $1 -ne 1 ]]; then
   # No Switchresponse!
      # Actions: Check if some nodes have switchresponds, record them
   echo "0000-05 Warning! Oncoming has no switch responds"
   for i in `$SDRGET -x switch_responds node_number`
     do
     LOCALRES=`$SDRGET -x switch_responds node_number==$i $RESPCMD`
     if [[ $LOCALRES -eq 1 ]]; then
         # Node has switchresponds, is candidate
        NEWNODE=$NEWNODE$i.
        let z=z+1
        echo "0000-03 Found node with switch: "$i
     fi
     if [[ $z -gt 1 ]]; then
         # enough nodes found
        echo "0000-04 Found enough nodes: "$z
        break
     fi
   done
   if [[ $z -eq 0 ]]; then
       echo "0000-06 Alert! No nodes with switch responds found!"
       exit 2
   fi
   if [[ $z -eq 1  ]]; then
       echo "0000-07 Alert! Only one node with switch respond found"
       RC=1
   fi
   func_changeprimary $NEWNODE
    else
   echo "0000-08 Commander: Everythings calm"
    fi
}

func_help(){
    echo "Usage: "$0" [-p oncoming_primary oncoming backup][-h]"
    echo "-p: Pretend oncoming primary and backup nodes for testing"
    echo "-e: Do not execute Eprimary change, just show it"
```

```
    echo "-h: This help"
}

#
# 2. The main function, calls every subroutine
#

echo $0 "Version 1.0 starting"

# Go to check the environment the script runs
func_checkenv

# Check for command line options
if [[ $# -ne 0 ]]; then

    # Three command line options provided, -p for simulation and -h for help
    case $1 in
   "-p")
       if [[ $# -eq 3 ]]; then
      OPRIMARY=$2
      OBACKUP=$3
       else
      func_help
      exit 1
       fi
       ;;
   "-e")
       PRETEND=0
       ;;
   "-h")
       func_help
       exit 0
       ;;
    esac
fi
    # Get the current oncoming primary and oncoming prinmary backup
func_getcurrent

# Check if both have host responds
func_checkcurrent
# Check if action is needed for both primary and backup
func_action $SWREP "-init"
if [[ $RC -eq 0 ]]; then
    func_action $SWREPB "-backup"
fi
exit $RC
```

C

Hints and tips

In this section, we present some hints and tips we found useful in our work. Note that the examples and workarounds may not be officially supported by IBM.

PSSP hints and tips

This section gives some ideas and hints for working with PSSP 3.5.

Identifying Ethernet adapters on the pSeries p660

Although officially only supported on pSeries p690, p670, p630, and p655, the functionality of selecting `smitty` → **RS/6000 System Management** → **RS/6000 Configuration** → **Database Management** → **Enter Database Information** → **Node Database Information** → **Get Adapter Physical Location Information**, which is, in fact, the `/usr/lpp/ssp/bin/spadaptr_loc` command, is also possible with some other pSeries servers, for example, the p660 6H1. Example C-1 on page 186 shows the obtained location codes.

Example: C-1 Location codes obtained for a p660

```
root $ splstdata -f
                  List Frame Database Information

frame# tty              s1_tty           frame_type      hardware_protocol
control_ipaddrs domain_name
------ ---------------- ---------------- --------------- ------------------
--------------- -----------
    1 /dev/tty0       ""             switch         SP              ""
""
    2 /dev/tty1       ""             ""             CSP             ""
""
    3 /dev/tty2       ""             ""             CSP             ""
""
sp4en0:/
root $ /usr/lpp/ssp/bin/spadaptr_loc 2 1 1
Acquiring adapter physical location codes for node 17
node# adapter_type physical_location_code MAC_address
----- ------------ --------------------- ------------
   17 Ethernet     U0.1-P1-I1/E1          000629DC5904
   17 Ethernet     U0.1-P1/E1             0004AC57489A
```

Example C-1 gives two location codes for node 17, the upper one is the Ethernet controller located in drawer U0.1 at position P1 in slot I1. The E1 denotes the first port of this adapter, which is useful if you own a 4-port Ethernet adapter.

Attention: The **sphardwrad** command powers off the system.

The obtained MAC address can be entered in /etc/bootptab.info, which is read by **smitty** → **RS/6000 System Management** → **RS/6000 Configuration** → **Database Management** → **Enter Database Information** → **Node Database Information** → **Get Hardware Ethernet Addresses**, which uses the **sphrdwrad** command, as shown in Example C-2.

Example: C-2 Getting the hardware Ethernet address

```
root $ cat /etc/bootptab.info
17 000629DC5904
root $ sphrdwrad -l 17
Acquiring hardware ethernet address for node 17 from /etc/bootptab.info
```

The location can be entered using **smitty** → **RS/6000 System Management** → **RS/6000 Configuration** → **Database Management** → **Enter Database Information** → **Node Database Information** → **SP Ethernet Information**, as shown in Example C-3. This executes:

```
/usr/lpp/ssp/bin/spadaptrs -l '17' -P 'U0.1-P1-I1/E1' -e '192.168.4.250' -t
'tp' -d 'full' -f '10' en 192.168.4.17 255.255.255.0
```

Example: C-3 SMIT panel for entering SP Ethernet information

```
SP Ethernet Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[MORE...13]                                              [Entry Fields]
  (C) Node List                                         [17]

  * You must fill in only one of the following sets
  (D) or (E) of fields:

  (D) Adapter Name                                      []

  OR

  (E) Physical Location Code                            [U0.1-P1-I1/E1]
  (E) Adapter Type                                      [en]                    +


[MORE...7]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Important: After this, you have to enter the physical location code instead of the adapter name, for example, when setting host names from long to short.

A tip on a Cluster 1600 lpp_source

The choice of lpp_source is very important and should be done very carefully. Often merely copying data from the CDs is not enough. We recommend the following actions:

- Keep your lppsource up to date with the newest AIX, vac, and PSSP PTFs.
- Keep your lppsource as small as possible by removing language filesets you do not need.

- The lppsource has to be at least at the level of the SPOT and the machines it supports.
- Have the newest version of the xlC.rte and vacpp.* in your lppsource and your SPOT. Otherwise, you might encounter migration problems.

Investigating PTFs

Usually, it is recommended to always apply the latest PTFs obtained by IBM. Sometimes in a stable production environment, only a few PTFs are applied after testing. In this case, it can be necessary to not only read the information obtained by the **instfix** command, but also to take a look at the package itself, as shown in Example C-4.

Example: C-4 Listing the contents of a PTF

```
sp3cws:/spdata/sys1/install/pssplpp/PSSP-3.5/PTFs
root $ cat PTF001.ssp.vsdgui.bff | restore -Tq -f -
./
./lpp_name
./usr
./usr/lpp
./usr/lpp/ssp.vsdgui/ssp.vsdgui/3.5.0.1
./usr/lpp/ssp.vsdgui/ssp.vsdgui/3.5.0.1/liblpp.a
./usr/lpp/ssp/perspectives/bin/spvsd
```

Rebuilding the SPOT

When making changes to the lppsource directory, such as adding PTFs, the Shared Product Object Tree (SPOT) must be updated. In order to update SPOT, perform the following steps on the control workstation and all of the boot/install servers:

1. Deallocate the SPOT from all clients using the **unallnimres -l <Node>** command.
2. On the control workstation only, copy all the install images for the PTFs to the lppsource directory that corresponds to the appropriate SPOT.
3. For boot/install server (BIS) nodes, it is necessary to add the BIS host name to the /.rhost file on the control workstation.
4. Issue **inutoc** in the lppsource directory.
5. Issue **nim -o check -F <lppsourcename>**.
6. Issue **smit nim_res_op**.
7. Select the appropriate SPOT.

8. Select **update_all**.
9. Press F4 in the Source of Install Images field, and select the appropriate lppsource.
10. Press Enter twice to initiate the update.
11. After the update completes, run **setup_server** to reallocate the SPOT to the necessary clients.

NIM and PSSP coexistence

NIM is one of the most powerful tools delivered with AIX. Besides the installation of multiple machines concurrently, it also provides a wide variety of management tools.

When installed on a CWS, NIM is completely controlled by PSSP. Any NIM configuration made without the PSSP tools will be deleted after PSSP, and especially **setup_server**, runs. For many reasons, it still can be useful to exploit more of the functionality of NIM than is provided by PSSP. In particular, NIM can be used to manage resources not managed by PSSP. Here, we provide some ideas of how to do this.

Attention: This may not be supported by IBM and should only be done if you are very familiar with AIX, NIM, and PSSP.

You can define any resource you would usually create for NIM with the **nim** command, smit, or WebSM, but because any unknown NIM resource is unconfigured, you should add all of them in a shell script. We have done this in a very simple example. You might want to use more sophisticated scripts or even a database to create the necessary inputs.

Example: C-5 Simple script to add NIM resources

```
#!/bin/ksh
# cr_nimres.sh:
#
# Sample Script to add a node automatically
#

# This is node sp6cws an other CWS which is not part of the Cluster 1600

nim -o define -t standalone\
   -a platform=chrp\
   -a if1="spnet_en1 sp6cws 0"\
   -a cable_type1=tp\
   -a netboot_kernel=mp\
   -a comments="Machine not in SP cluster"\
```

```
   sp6cws

# This defines a bosinst.data ressource, not handled by PSSP
nim -o define -t bosinst_data\
   -a server=master\
   -a location=/spdata/sys1/install/pssp/bosinst_data_sp6\
   -a comments="Other bosinst.data" mybosinst_data

# This defines a network outside of the PSSP network
nim -o define -t ent\
   -a net_addr=192.168.6.0\
   -a snm=255.255.255.0\
   -a comments="Other Network"

#This defines a different lpp_source
nim -o define -t lpp_source\
   -a server=master\
   -a location=/spdata/sys2/lppsource/\
   -a comment="Other LPP Source"
```

Example C-6 shows how **setup_server** affects the NIM definition not controlled by PSSP. The non-PSSP definitions are highlighted.

Example: C-6 NIM resources and setup_server

```
sp4en0:/tmp
root $ lsnim
master              machines        master
boot                resources       boot
nim_script          resources       nim_script
spnet_en0           networks        ent
spnet_en1           networks        ent
psspscript          resources       script
prompt              resources       bosinst_data
noprompt            resources       bosinst_data
migrate             resources       bosinst_data
1_noprompt          resources       bosinst_data
1_migrate           resources       bosinst_data
sp4n01e0            machines        standalone
mksysb_2            resources       mksysb
lppsource_aix51     resources       lpp_source
mksysb_1            resources       mksysb
spot_aix51          resources       spot
sp4n05e0            machines        standalone
sp4n17e0            machines        standalone
sp6cws              machines        standalone
sp4n33e0            machines        standalone
mynetwork           networks        ent
mybosinst_data      resources       bosinst_data
```

```
newlpp              resources       lpp_source
myspot              resources       spot
mynewspot           resources       spot

sp4en0:/tmp
root $ setup_server
...
sp4en0:/tmp
root $ lsnim
master              machines        master
boot                resources       boot
nim_script          resources       nim_script
spnet_en0           networks        ent
spnet_en1           networks        ent
psspscript          resources       script
prompt              resources       bosinst_data
noprompt            resources       bosinst_data
migrate             resources       bosinst_data
1_noprompt          resources       bosinst_data
1_migrate           resources       bosinst_data
sp4n01e0            machines        standalone
mksysb_2            resources       mksysb
lppsource_aix51     resources       lpp_source
mksysb_1            resources       mksysb
spot_aix51          resources       spot
sp4n05e0            machines        standalone
sp4n17e0            machines        standalone
sp4n33e0            machines        standalone
mynetwork           networks        ent
mybosinst_data      resources       bosinst_data
newlpp              resources       lpp_source
mynewspot           resources       spot

sp4en0:/
root $ ./cr_res.sh
sp4en0:/
root $ lsnim
master              machines        master
boot                resources       boot
nim_script          resources       nim_script
spnet_en0           networks        ent
spnet_en1           networks        ent
psspscript          resources       script
prompt              resources       bosinst_data
noprompt            resources       bosinst_data
migrate             resources       bosinst_data
1_noprompt          resources       bosinst_data
1_migrate           resources       bosinst_data
sp4n01e0            machines        standalone
```

```
mksysb_2            resources       mksysb
lppsource_aix51     resources       lpp_source
mksysb_1            resources       mksysb
spot_aix51          resources       spot
sp4n05e0            machines        standalone
sp4n17e0            machines        standalone
sp6cws              machines        standalone
sp4n33e0            machines        standalone
mynetwork           networks        ent
mybosinst_data      resources       bosinst_data
newlpp              resources       lpp_source
mynewspot           resources       spot
```

Tip: We recommend adding NIM objects that can also be part of a PSSP cluster as spot, lppsource, or mksysb within the PSSP directory structure.

Coexistence of s1term and vterm for HMC-based servers

PSSP uses the HMC for the control of the HMC-based servers, such as the p630, p655, p670, and p690. It uses the same method provided by the HMC, the virtual terminal (vterm). Limitations on the HMC allow only one vterm per LPAR. If the HMC already has one vterm open, all s1term-related operations on the CWS will fail. You can, however, either **ssh** to the HMC and get the GUI by issuing **startHSC**, or by using the WebSM client on the CWS and then selecting the partition and closing the terminal. This closes the terminal wherever it is opened.

Tip: It is a good practice to issue all commands, even HMC-related ones, on the CWS to guarantee a single point of control.

Planning for General Parallel File System

This section describes some restrictions that must be taken into account when you configure GPFS using the **mmcrcluster**, **mmconfig**, and **mmcrfs** commands.

GPFS on HACMP/RPD (AIX-related environment)

In the AIX-related environment, a GPFS cluster is created by issuing the **mmcrcluster** command. The GPFS cluster creation options on the **mmcrcluster** command are shown in Table C-1 on page 193. You must make sure the GPFS cluster type cannot be changed later by the **mmchcluster** command.

Table C-1 GPFS cluster creation options in an AIX-related environment

Options	mmcrcluster	mmchcluster	Default value
Nodes in an GPFS cluster	O	You can add or delete nodes by using the **mmaddcluster** or **mmdelcluster** command.	None
Primary server	O	O	None
Secondary server	O	O	None
Cluster type	O	This cannot be changed.	None
Remote shell command	O	O	/usr/bin/rsh
Remote file copy command	O	O	/usr/bin/rcp
Note: O indicates the option is available on the command.			

A GPFS nodeset is created by issuing the **mmconfig** command. Table C-2 on page 194 shows the configuration options specified with the **mmconfig** command. The GPFS nodeset identifier cannot be changed later with the **mmchconfig** command.

Table C-2 GPFS configuration options in an AIX-related environment

Options	mmconfig	mmchconfig	Default value
Nodes in an GPFS nodeset	O	You can add or delete nodes by using the **mmaddnode** or **mmdelnode** command.	All the nodes in the GPFS cluster
Nodeset identifier	O	This cannot be changed.	An integer value beginning with 1 and increasing sequentially
Starting GPFS automatically	O	O	No
Path for the storage of dumps	O	O	/tmp/mmfs
Single-node quorum	O	O	No
pagepool	O	O	20 M
maxFilesToCache	O	O	1000
maxStatCache	Default value initially used	O	4 x maxFilesToCache
maxblocksize	Default value initially used	O	256 K
dmapiEventTimeout		O	86400000
dmapiSessionFailureTimeout		O	0
dmapiMountTimeout		O	60

Notes:
1. O indicates the option is available on the command.
2. An empty cell indicates the option is not available on the command.

A GPFS file system is created by issuing the **mmcrfs** command. Table C-3 on page 195 shows the file system creation options specified with the **mmcrfs** command. The estimated number of nodes, the size of data blocks, maximum metadata replicas, maximum data replicas, and the device name of the file system cannot be changed later with the **mmchfs** command.

Table C-3 GPFS file system creation options in an AIX-related environment

Options	mmcrfs	mmchfs	Default value
Automatic mount	O	O	Yes
Estimated node count	O	This cannot be changed.	32
Block size	O	This cannot be changed	256K
Maximum number of files	O	O	File system size/1 MB
Default metadata replicas	O	O	1
Maximum metadata replicas	O	This cannot be changed.	1
Default data replicas	O	O	1
Maximum data replicas	O	This cannot be changed.	1
Automatic quota activation	O	O	No
Disk verification	O		Yes
Enable DMAPI	O	O	No
Mountpoint directory	O	O	None
Device name of the file system	O	This cannot be changed	None
Disks for the file system	O	You can add or delete disks by using the **mmadddisk** or **mmdeldisk** command.	None
Nodeset	O	O	The nodeset from which the **mmcrfs** command is issued

Notes:
1. O indicates the option is available on the command.
2. An empty cell indicates the option is not available on the command.

GPFS on VSD (PSSP-related environment)

In the VSD environment, you cannot use the **mmcrcluster** command. The default cluster type is sp. You can find the cluster type by issuing the **mmlsconfig** command or by looking at the /var/mmfs/etc/mmfs.cfg file.

A GPFS nodeset is configured by issuing the **mmconfig** command. Table C-4 on page 196 shows the configuration options specified with the **mmconfig** command. The GPFS nodeset identifier cannot be changed later with the **mmchconfig** command.

Table C-4 GPFS configuration options in a VSD environment

Options	mmconfig	mmchconfig	Default value
Nodes in a GPFS nodeset	O	You can add or delete nodes by using the **mmaddnode** or **mmdelnode** command.	All the nodes in the GPFS cluster
Nodeset identifier	O	This cannot be changed.	An integer value beginning with 1 and increasing sequentially
Starting GPFS automatically	O	O	No
Path for the storage of dumps	O	O	/tmp/mmfs
GPFS daemon communication protocol	O	O	TCP/IP
Single-node quorum	O	O	No
pagepool	O	O	20 M
maxFilesToCache	O	O	1000
maxStatCache	Default value initially used	O	4 x maxFilesToCache
maxblocksize	Default value initially used	O	256 K
dmapiEventTimeout		O	86400000
dmapiSessionFailureTimeout		O	0
dmapiMountTimeout		O	60

Notes:
1. O indicates the option is available on the command.
2. An empty cell indicates the option is not available on the command
3. The GPFS daemon communication protocol is a unique option only available in the VSD environment and can be either TCP/IP or LAPI.

A GPFS file system is created by issuing the **mmcrfs** command. Table C-5 on page 197 shows the file system creation options specified with the **mmcrfs** command. The estimated number of nodes, the size of data blocks, maximum metadata replicas, maximum data replicas, and the device name of the file system cannot be changed later with the **mmchfs** command.

Table C-5 GPFS file system creation options in a VSD environment

Options	mmcrfs	mmchfs	Default value
Automatic mount	O	O	Yes
Estimated node count	O	This cannot be changed.	32
Block size	O	This cannot be changed.	256 K
Maximum number of files	O	O	File system size/1 MB
Default metadata replicas	O	O	1
Maximum metadata replicas	O	This cannot be changed.	1
Default data replicas	O	O	1
Maximum data replicas	O	This cannot be changed.	1
Automatic quota activation	O	O	No
Disk verification	O		Yes
Enable DMAPI	O	O	No
Mountpoint directory	O	O	None
Device name of the file system	O	This cannot be changed.	None
Disks for the file system	O	You can add or delete disks by using the **mmadddisk** or **mmdeldisk** command.	None
Nodeset	O	O	The nodeset from which the **mmcrfs** command is issued

Notes:
1. O indicates the option is available on the command.
2. An empty cell indicates the option is not available on the command.

D

AIX device drivers reference

This section matches the AIX device driver names and hardware. The problem is that often a system administrator has either the name of a device or the feature code and now wants to know the AIX fileset representing this device. If you want to keep the NIM environment in your CWS up-to-date, you will find this information useful.

The device drivers for most devices supported by various AIX releases are already included in the base operating system release. For some OEM-supplied hardware, there are additional device drivers that are not included in this document.

Matching AIX device drivers to devices

AIX device drivers are usually on the installation media and are coded as follows:

```
devices.pci.14100401.rte
```

Where the first part denotes that this is a device driver, the second part specifies the bus, the third is the special hardware this driver supports and the last part denotes the purpose. The usual supported busses are pci, isa, and mca. Some of the typical purposes are rte (for real-time environment), com (for common software to more then one device), diag (for diagnostic support), and X11 (for support of the X-Windows system).

The following tables list AIX device drivers. The first column identifies the device driver and the device number, and the second column shows the bus type the hardware is connected to. The next columns list the parts of the driver and the release level of AIX that supports it. Note, that there may be newer versions of the specific software due to PTFs. The seventh column shows the feature code of the hardware, and if available, the label that is printed on the card. The last column gives a brief description of the product.

PCI-attached hardware

PCI bus systems are standard, industry-based bus systems. Various adapters exist that are supported by AIX 5L Version 5.2, 5.1, and 4.3.3, as listed in Table D-1.

Table D-1 PCI device drivers

Device number	Bus	Fileset	AIX Version 4.3.3	5.1.0	5.2.0	Feature Code	Description
00100100	PCI	com rte	77 76	35 25	0 0	#6208 (4-A) #6207 (4-L)	Standard NCR53C810 SCSI software for Common Symbios PCI SCSI I/O Controller for SCSI-2 SE Fast/Wide PCI Adapter, PCI Differential Ultra SCSI Adapter
00100300	PCI	diag rte	77 25	35 0	0 0	#2408 (4-A)[d] #2409 (4-B)[d]	PCI 16-bit SCSI I/O Controller
00100b00	PCI	diag rte	25 25	15 10	0 0	#6205 (4-R)	SYM53C896 PCI Dual-Channel Ultra2 SCSI Adapter
00100c00	PCI	diag rte	2.25 25	15 10	0 0	N/A	SYM53C895 PCI LVD SCSI I/O Controller software

Device number	Bus	Fileset	AIX Version 4.3.3	5.1.0	5.2.0	Feature Code	Description
00100f00	PCI	diag rte	0.25 78	15 26	0 0	#6208 (4-A) #6209 (4-B) #6206 (4-K) #6207 (4-L) #6204 (4-U) #9136	SYM53C8xxA PCI SCSI I/O Controller for SCSI-2 SE Fast/Wide PCI Adapter and PCI Single-Ended Ultra SCSI Adapter, PCI Differential Ultra SCSI Adapter, PCI Universal Differential Ultra SCSI Adapter
00102100	PCI	diag rte	0.0 0.0	15 0	0 0	#6203 (4-Y)	Dual-Channel Ultra3 SCSI Adapter SYM53C1010
0e100091	PCI	X11 diag rte	10 0 0	0 0 0	N/A N/A N/A[a]	#2657 (*)[d]	S15/H10 Graphics Adapter
14100401	PCI	diag rte	51 79	26 27	0 0	#2969 (9-U) #1117 (SP) #2975 (A-A)	Gigabit Ethernet-SX PCI Adapter, 10/100/1000 Base-T Ethernet
14101800	PCI	diag rte	50 77	25 0	0 0	#2979 (8-T)	Auto LANstreamer Token-Ring PCI Adapter
14101b00	PCI	X11 diag rte	75 0 0	10 0 0	N/A N/A N/A	#2648 (*)	GXT150P Graphics Adapter
14101b02	PCI	X11 diag rte	4 1 4	25 25 N/A	0 0 N/A	#2843 (1-Z)	GXT6500P Graphics Adapter software
14101c00	PCI	rte	0	N/A	N/A	N/A	Power Management Controller software
14101c02	PCI	X11 diag rte	4 1 4	25 0 26	0 0 0	#2842 (1-Y)	GXT4500P Graphics Adapter
14102000	PCI	X11 diag rte ucode	0 0 11 0	0 0 0 N/A	N/A N/A N/A N/A	#2856 (1-H)	GXT1000 PCI Graphics Adapter

Device number	Bus	Fileset	AIX Version 4.3.3	5.1.0	5.2.0	Feature Code	Description
14102e00	PCI	diag rte vsmit	78 27 0	26 2 N/A[b]	0 0 N/A	#2493 (4-H) #2494 (4-T) #2498 (4-X)	SCSI-2 Fast/Wide IBM PCI SCSI RAID Adapter, PCI 3-Channel Ultra2 SCSI RAID, 4-Channel Ultra3 SCSI RAID Adapter
14103c00	PCI	X11 com diag rte	75 25 0 2.0	10 0 0 0	0 0 0 0	#2851 (1-M)[d] #2852 (1-N)	GXT250P/GXT255P Graphics Adapter
14103e00	PCI	diag rte	75 75	26 26	0 0	#2920 (9-O) #4959 (9-Y)	IBM PCI Token-Ring Adapter 16 Mbps, 100Mbps Token-Ring PCI Adapter
14103302	PCI	X11 diag rte	6 0 2	26 25 26	0 0 0	#2848 (1-X)	GXT135P Graphics Adapter
14104000	PCI	X11 rte	2.1.1 1.0	N/A N/A	N/A N/A	N/A	GXT 5000P Graphics Adapter
14104500	PCI	diag rte	75 1.2	0 15	0 0	#6218 (4-J)[e] #6215 (4-N)[d] #6225 (4-P) #6230 (4-P)	PCI SSA 4-Port RAID Adapter, PCI SSA Multi-Initiator/RAID EL Adapter and SSA Fast-Write Cache Option Card, PCI SSA Advanced SerialRAID, PCI SSA Advanced SerialRAID Plus Adapters
14104e00	PCI	diag rte	0 50	0 0	0 0	#2963 (9-J)	TURBOWAYS 155 PCI UTP ATM Adapter
14104f00	PCI	diag rte	0 2.50	0 0	0 0	N/A	PCI ATM Adapter
14105000	PCI	diag rte	0 2.50	0 0	0 0	#2988 (9-F)	TURBOWAYS 155 PCI MMF ATM Adapter
14105300	PCI	diag rte	50 25	0 0	0 0	#2998 (*)[d]	TURBOWAYS 25 ATM PCI Adapter

Device number	Bus	Fileset	AIX Version 4.3.3	5.1.0	5.2.0	Feature Code	Description
14105400	PCI	X11 diag rte	25 10 25	0 0 0	0 0 0	#2854 (1-I) #2855 (1-J)	GXT500P/GXT550P Graphics Adapter
14105e00	PCI	X11 diag rte	25 10 25	0 0 0	0 0 0	#2853 (1-K) #2859 (1-L)	GXT800P Graphics Adapter
14105e01	PCI	com diag rte	80 77 75	26 25 15	0 0 0	#2946 (A-B)	TURBOWAYS 622 PCI MMF ATM
14106001	PCI	diag rte	0 75	0 10	0 0	#4957	64-bit/66 MHz PCI ATM 155 MMF Adapter software
14106602	PCI	diag rte	N/A N/A[c]	35 35	0 0	N/A	PCI-X Ultra 320 SCSI Adapter (Dual Channel)
14106802	PCI	diag rte	N/A N/A	35 35	0 0	#5700	Gigabit Ethernet SX PCI-X Adapter
14106902	PCI	diag rte	N/A N/A	35 35	0 0	#5701	Gigabit Ethernet Base TX PCI-X Adapter
14106e01	PCI	X11 diag rte	78 0 77	25 0 26	0 0 0	#2826 (1-V)	GXT4000P PCI Graphics Adapter
14107001	PCI	X11 diag rte	78 0 77	25 0 26	0 0 0	#2827 (1-W)	GXT6000P PCI Graphics Adapter
14107c00	PCI	com diag rte	76 0 76	25 0 15	0 0 0	#2988 (9-F) #2963 (9-J)	TURBOWAYS 155 MMF ATM Adapter and Turboways 155 UTP ATM Adapter
14107d01	PCI	X11 diag rte	78 0 75	25 0 10	0 0 0	#2841 (1-U)	GXT300P 2D Graphics Adapter
14108c00	PCI	rte	52	25	0	#2947 (9-R)[d]	ARTIC960Hx 4-Port Selectable PCI Adapter

Device number	Bus	Fileset	AIX Version 4.3.3	5.1.0	5.2.0	Feature Code	Description
14108e00	PCI	X11 diag rte	78 10 75	25 0 15	0 0 0	#2825 (1-R)	GXT3000P Graphics Adapter
14109100	PCI	diag rte	50 26	0 15	0 0	#6225 (4-P) #6230 (4-P)	PCI SSA Advanced SerialRAID, PCI SSA Advanced SerialRAID Plus Adapters
14109f00	PCI	diag rte	2 1	25 26	0 0	#4958 (6-H)[d]	Crypto Accelerator Adapter software.
1410b800	PCI	X11 diag rte	78 0 75	25 0 15	0 0 0	#2823 (1-S)	GXT2000P Graphics Adapter
1410c101	PCI	rte	75	15	0	#4953	64-bit/66MHz PCI ATM 155 UTP Adapter software
1410e601	PCI	diag rte	75 75	25 26	0 0	#4960	IBM e-business Crypto Accelerator Adapter software
1410ff01	PCI	diag rte	3 4	26 26	0 0	#4962 (A-F)	10/100 Mbps Ethernet PCI Adapter II
1c104ac2	PCI	X11 rte	2.1.0 2.1.1	N/A	N/A	(*) (*)	G10 Graphics Adapter
22100020	PCI	diag rte	50 25	0 0	0 0	#2985 (8-Y)[d] #2987 (8-Z)[d]	IBM PCI T2 Ethernet Adapter, IBM PCI T5 Ethernet Adapter
23100020	PCI	diag rte	50 80	26 28	0 0	#2968 (9-P)[d] #4951 (9-Z) #4961 (A-E)	10/100 Ethernet Tx PCI Adapter, 4-Port 10/100 Base-TX Ethernet, 4-Port 10/100 Ethernet Base-TX
2b101a05	PCI	X11 diag rte	79 10 25	26 0 0	0 0 0	#2838 (1-P)	GXT120P Graphics Adapter

Device number	Bus	Fileset	AIX Version 4.3.3	5.1.0	5.2.0	Feature Code	Description
2b102005	PCI	X11 diag rte	0 11 75	0 1 0	0 0 0	#2830 (1-T)	GXT130P Graphics Adapter
331121b9	PCI	com diag rte	25 51 51	0 25 25	0 0 0	#2962 (9-L)[d] #2962 (9-V)[d]	2-Port Multiprotocol PCI Adapter
31114571	PCI	diag rte	0 2.2.1	N/A	N/A	#2638 (7-9)	Ultimedia Video Capture Adapter
31114671	PCI	diag rte	0 2.1	N/A	N/A	#2639	Ultimedia Video Capture Adapter
33531188	PCI	X11 diag rte	0 0 2.0	0 0 0	N/A	#2839 (*)[d] #2837 (*)[d]	GXT110P Graphics Adapter, MVP POWER Multi-Monitor Adapter
33531288	PCI	X11 rte	2.1.1 2.1.1	N/A	N/A	#2837 (*)	MVP POWER Multi-Monitor Adapter
3353b088	PCI	X11 rte	1.5.0 1.5.0	N/A	N/A	N/A	Unknown adapter
3353c088 3353c188 3353c288 3353c388	PCI	X11 com rte	75 25 2.0	10 0 0	N/A	N/A	E15 Graphics Adapter family
48110040	PCI	diagsa diag rte	75 4.0.1.0 1.0.0.13	15 4.0.1.0 1.0.0.13	0	#2741 (*)[d] #2742 (*)[d] #2743 (*)[d]	PCI FDDI Adapter Models 7024-E20/E30 7025-F30 7248-100/120/132
4f111100	PCI	asw com diag rte	11 79 1 2	N/A 27 25 25	0 0 0 0	#2943 (3-B)	PCI 8-Port Asynchronous Adapter
4f111b00	PCI	asw diag rte	11 1 1	0 25 0	0 0 0	#2944 (3-C)	PCI 128-Port Asynchronous Adapter
86808404	PCI	com rte	2.1.0 0	25 0	N/A	N/A	ISA Bus software

Device number	Bus	Fileset	AIX Version 4.3.3	5.1.0	5.2.0	Feature Code	Description
8d100100	PCI	N/A	N/A	N/A	N/A	#8246 (*)	Olicom Token-Ring Adapter
ad100501	PCI	com rte	N/A 4	N/A 25	0 0	N/A	IDE Adapter Driver for Winbond 553F Chip software
b7105059	PCI	N/A	N/A	N/A	N/A	#8242 (*)	3Com Ethernet 3C590/595 10/100 Mbps Adapter
b7105090	PCI	N/A	N/A	N/A	N/A	#2986 (*)[e]	3Com 3C905 Fast EtherLink XL PCI 10/100 Ethernet Adapter
c1110358	PCI	diag rte	N/A N/A	0 25	0 0	N/A	USB Open Host Controller Adapter software
df1000f7	PCI	com diag rte	83 78 76	28 25 16	0 0 0	#6227 (4-S)	Gigabit Fibre Channel Adapter
df1000f9	PCI	diag rte	75 76	15 15	0 0	#6228 (4-W)	Gigabit Fibre Channel Adapter 64 bit
ed101073	PCI	rte	2.1.0	N/A	N/A	N/A	Unknown
artic960	PCI	rte	1.2.0.0	1.4.4.0		#2947 (9-R)[d] #2948 (9-S) #2949 (9-7)	ARTIC960Hx 4-Port Selectable PCI Adapter, ARTIC960Hx 4-Port T1/E1 PCI Adapter, ARTIC960Hx DSP Resource
esconCH	PCI	rte	2.1.3.0			#2751 (5-5)[d]	Escon Channel Emulator ESCON Channel PCI Adapter
esconCU	PCI	diag rte	2.1.2.0 2.1.3.0			N/A	Escon Control Unit Connectivity

a. This adapter is no longer supported in AIX 5L Version 5.2.
b. AIX 5L Version 5.1 and 5.2 do not support vsmit.
c. This adapter is supported only in AIX 5L.
d. This adapter is only supported for the 32-bit kernel of AIX 5L Version 5.1.
e. This adapter is not supported in AIX 5L.

MCA-attached hardware

Microchannel bus systems were included in the very first RS/6000 servers and are not longer included in any pSeries systems. Nevertheless, support by AIX is included up to AIX 5L Version 5.1. Table D-2 show the available device drivers.

Restriction: None of the device drivers or machines support the use of the AIX 5L 64-bit kernel.

Table D-2 Microchannel device drivers

Device number	Bus type	Fileset	AIX Version 4.3.3	AIX Version 5.1.0	Feature Code	Description
0072	MCA	N/A	N/A	N/A	#2402 (8-5)	Network Terminal Accelerator 256 Adapter
0095	MCA	N/A	N/A	N/A	#2403 (8-6)	Network Terminal Accelerator 2048 Adapter
0200	MCA	rte diag	0.0 0.0	0 0	N/A	Wide SCSI Adapter
0210	MCA	rte diag	0 1.0	0 0	N/A	Turboways 25 MCA ATM Adapter
61fd	MCA	rte diag	50 0.0	15 0	#6400 (3-6)	64-Port Asynchronous Controller
8787	MCA	diag rte ucode	0 0 0.0	0 10 N/A	#2801 #2802 (6-2)	5080/85/86/88 Attachment Adapter
8d77	MCA	diag rte ucode	50 0 0.0	35 35 N/A	#2410 (4-4) #2831 (4-4) #2420 (4-2) #2835 #2828 #2929 (4-1)	SCSI-2 8-bit Single-Ended High-Performance, SCSI-1 Single-Ended Int/Ext I/O Controller (4-1), Internal/External I/O Controller, SCSI-2 Differential 8-bit External I/O Controller
8ee3	MCA	X11 diag rte ucode	10 0 10 0	0 0 0 N/A	#2795 #2790 #2796 #2791 #2711 #2712 #2713 (1-5) #2777 (1-6) #2776 (1-8) #2768 (1-9)	Gt4/Gt4x/Gt4xi/Gt3/gt4e/Gt3i Graphics Adapter Software

Device number	Bus type	Fileset	AIX Version 4.3.3	AIX Version 5.1.0	Feature Code	Description
8ee4	MCA	X11 diag rte	10 0 0	0 0 0	#2770 (1-1)	AIXwindows Color Graphics Display Adapter
8ee5	MCA	X11 diag rte	0 0.0 0	0 0 0	#2760 (1-2)	Grayscale Graphics Display Adapter
8ee6	MCA	N/A	N/A	N/A	#2780 (1-3) #2781 (1-3)	8-bit 3D Color Graphics Processor, 24-bit 3D Color Graphics Processor
8ef2	MCA	com diag rte	0 0 1.0	15 0 0	(*)	Integrated SCSI 7006/7008/7009/ 7011-250/7012 and SCSI-1 7011-220/230
8ef3	MCA	diag rte	0 1.0	15 0	#9000 #4221 (2-8) #9001 #4222 (2-9)	Integrated Ethernet for Ethernet Riser Cards Thick/Thin and Integrated Etherne Riser Cards Twisted-Pair
8ef4	MCA	diag rte ucode	75 75 0.25	10 10 N/A	#2720 (2-6) #2722 (2-7) #2724 (2-R) #2725 (2-S) #2726 (2-U) (2-5)	FDDI Single Ring Adapter, FDDI Dual Ring Upgrade Kit Adapter, FDDI Fiber Single Ring Adapter, FDDI Fiber Dual Ring Upgrade Kit, FDDI STP Single Ring Adapter, FDDI STP Dual Ring Upgrade Kit Adapter
8ef5	MCA	diag rte	0 0	15 0	#2964 (9-Q) #2980 (2-1)	10/100 Mbps Ethernet UNI only, Ethernet High-Performance LAN Adapter
8efc	MCA	com diag rte	0 0 10	35 15 0	#2412 (4-C) #2413 #2416 #9217 (4-6) #2414 #2415 #9216 (4-7)	Enhanced SCSI-2 Differential Fast/Wide Adapter, SCSI-2 Differential Fast/Wide Adapter, SCSI-2 Single-Ended Fast/Wide Adapter #2414/2415/9216 (4-7)

Device number	Bus type	Fileset	AIX Version 4.3.3	5.1.0	Feature Code	Description
8f61	MCA	X11 diag rte	25 0 11	0 0 0	#2850 (1-Q)	GXT800M Graphics Adapter
8f62	MCA	diag rte	50 25	0 0	#2964 #2994 (9-K)	10/100 Mbps Ethernet SMP and UNI
8f64	MCA	diag rte	0 0	0 0		155 Mbps MCA ATM Adapter
8f66	MCA	diag rte ucode	0 0 1.0	0 0 N/A	#2999 (9-E)	155 Mbps ATM MPEG Adapter
8f67	MCA	com diag rte ucode	51 0 0 1.0	25 15 0 0	#2989 (9-9)	Turboways 155 ATM Adapter
8f70	MCA	diag rte mpqp	0 0 25	10 0 10	#2700 (2-3)	Integrated SCSI-1 7012/7013 / 7015, 4-Port Multiprotocol Communications Controller
8f78	MCA	diag rte ucode	25 0 0	15 0 N/A	(4-3) (4-5) (4-8)	Serial Linked Disk Adapter/Controller/DASD High-Performance Disk Drive Subsystem
8f7f	MCA	diag rte ucode	0.0 0.0 1.0	0 0 N/A	#2998 #2984(8-W)	Turboways 100 ATM Adapter
8f95	MCA	diag rte	50 50	0 0	#2992 (8-U) #2993 (8-V)	10 Mbps Ethernet High-Performance LAN Adapter
8f96	MCA	rte	2.0	N/A	#2404 #2405 (7-5)	Ultimedia Video Adapter
8f97	MCA	com diag rte	0 75 0	0 0 15	#6214 (4-D) #6216 (4-G) #6217 (4-I) #6219 (4-M)	SSA 4-Port MCA Adapter, SSA Enhanced Adapter MCA, SSA 4-Port RAID MCA Adapter, SSA Multi-Initiator/RAID EL MCA Adapter
8f98	MCA	diag rte	0 1.0	15 0	(*)	10 Mbps Integrated Ethernet

Device number	Bus type	Fileset	AIX Version 4.3.3	AIX Version 5.1.0	Feature Code	Description
8f9a	MCA	X11 diag rte ucode	50 0 0 0.0	0 0 0 N/A	#2650 (1-D) #2650 (1-E) #2767 #9651 #2665	GXT150M Graphics Adapter, GXT150 for 7011-250, GXT150L GXT155L
8f9d	MCA	diag rte	0 1	0 35	N/A	LAN SCSI Adapter
8fa2	MCA	diag rte	0 10	0 0	#2972 (8-S)	Auto Token-Ring LANstreamer MC 32 Adapter
8fba	MCA	com diag rte	0 0 0.0	N/A 0 35	N/A	Common NCR53C7xx software
8fbc	MCA	X11 diag rte ucode	50 0 0 0	0 0 0 N/A	#2820 (1-A)	GXT1000 Graphics Adapter
8fc3	MCA	diag rte	0 1.1.0.10	N/A	#2756 (5-3) #2754 (5-3)	Escon Channel Adapter, Escon Channel Emulator Adapter
8fc8	MCA	diag rte ucode	0 75 0.0	15 10 N/A	#2970 (2-2)	Token-Ring High-Performance Network Adapter
8fe2	MCA	diag rte	2.1.0 2.0.1	N/A	#1904 #1902 (9-A)	Fibre Channel 1063 Adapter Short Wave
8fe5	MCA	N/A	N/A	N/A	#2735 (8-A) (8-B)	High-Performance Parallel Interface (HIPPI) Adapter
8ff4	MCA	diag rte	0.0 0	0 35	N/A	Standard NCR 53C700 software
8ffd	MCA	rte	2.1.1	N/A	#4350	Graphics Subsystem Adapter and GTO 7235-001/002 Parts #4350 (1-4)
dee6	MCA	rte	0	0	N/A	Standard I/O Adapter
deff	MCA	diag rte sdlc	0 0 0	0 N/A 10	#2959 (2-P)	Multiprotocol Adapter

Device number	Bus type	Fileset	AIX Version 4.3.3	5.1.0	Feature Code	Description
df5f	MCA	com diag rte	N/A 0 0.0	0 0 0	N/A	Standard I/O Adapter
df9f	MCA	diag rte	25 0	15 0	(*)	Direct Attached Disk
dfe5	MCA	rte	0.1	N/	#6302 (7-6)	Ultimedia Audio Adapter
e555	MCA	rte	2.1.0.23 5	N/A	#9291 #9295 (6-5)	Voice Server Attachment Adapter (VSAA/VSCA)
e556	MCA	rte ucode	2.2.2.31 09 2.2.2.30 00	N/A	#9291 #9295 (6-6)	Voice Server Dual Attachment Adapter (VSDA)
e1ff	MCA	diag rte	0 0	0 10	#2990 (5-1)	3270 Connection Adapter
edd0	MCA	com diag rte	50 0.0 0.0	0 0 0	#2930 (3-1) #2940 (3-2) #2950 (3-3) #2955 (3-4) #2957 (3-5) #2755 (5-2	Common Async Adapter Support, 8-Port Async Adapter EIA-232, 8-Port Async Adapter EIA-422A, 16-Port Async Adapter EIA-232, 16-Port Async Adapter EIA-422.
edd1	MCA	diag rte	0.0 0.0	0 0	#7002 #7004 #7028 (2-G)	8-Port EIA-422A Multiport/2 Adapter
edd2	MCA	diag rte	0.0 0.0	0 0	N/A	8-Port Asynchronous Adapter MIL-STD 188
edd3	MCA	diag rte	0.0 1.0	0 0	N/A	16-Port Asynchronous Adapter EIA-422
edd5	MCA	X11 com diag rte	0 0 0 0	N/A N/A 0 0	#2810 (6-1)	Graphics Input Device Adapter
edd6	MCA	diag rte	0.0 1.0	0 0	N/A	16-Port Asynchronous Adapter
efbd	N/A	diag rte ucode	0.0 0.0 0.0	0 0 N/A	#2840 (6-8)	5080 CoaxAttachment Adapter 7011-200 Series and 7006 41T/41W

Device number	Bus type	Fileset	AIX Version 4.3.3	AIX Version 5.1.0	Feature Code	Description
eff0	MCA	diag rte	0 0	0 0	#2960 (2-4)	X.25 Interface CO-Processor/2
fe92	MCA	diag rte	0 2.0.7	N/A	#2755 (5-2)	IBM S/370 Block Multiplexer Channel Adapter
fed9	MCA	rte	0.0	0	N/A	Standard I/O Adapter
f6f4	MCA	diag rte	1.0 0	0 0	N/A	MCA Keyboard and Mouse Adapter
f6f8	MCA	diag rte	1.0 0	0 0	N/A	MCA Keyboard and Mouse Adapter
f6fe	MCA	rte	0	0	N/A	Standard I/O Adapter
ffe1	MCA	diag rte ucode	0 79 10	0 35 0	#8128 (3-7)	128-Port Asynchronous Adapter

SP Switch Attachment Adapters

The high-performance IBM Switch Systems, introduced in the SP and now an important part of the Cluster 1600, provide a low-latency and high-bandwidth interconnect. They are managed by PSSP, and therefore, the device driver is integrated into PSSP. A single LPP ssp.css is responsible for all the different switch adapters. Table D-3 gives an overview of all existing adapters.

Table D-3 SP Switch Attachment Adapters

Bus type	Fileset	PSSP Version 3.50	PSSP Version 3.4.0	Feature Code	Description
MCA	css	0	10	#4018	High Performance Switch Adapter Card
MCA	css	0	10	#4020 (6-9)	Scalable PowerParallel Switch Adapter
MX	css	0	10	#4022 (6-A)	SP Switch MX Adapter
MX2	css	0	10	#4023 (6-C)	SP Switch MX2 Adapter
MX2	css	0	10	#4025 (6-D)	SP Switch2 Communications Adapter
MX2	css	0	10	#4026 (6-M)	SP Switch2 MX2 Attachment Adapter
PCI	css	0	10	#8396 (6-F)	SP Switch PCI Attachment Adapter
PCI	css	0	10	#8397 (6-L)	SP Switch2 PCI Attachment Adapter

Bus type	Fileset	PSSP Version 3.50	3.4.0	Feature Code	Description
PCI-X	css	0	10	#8398	SP Switch2 PCI-X Attachment Adapter

Other attached hardware

Additional bus systems for special purposes were integrated into some older models. Table D-4 lists the drivers for those adapters.

Restriction: These drivers do not support the 64-bit kernel and are not available beyond AIX 5L Version 5.1.

Table D-4 Other attached hardware

Device number	Bus type	Fileset	AIX Version 4.3.3	5.1.0	Feature Code	Description
00004001	BUC	X11 com rte	75 0 2.0	10	#2766	GXT100 Graphics Adapter (7011-250)
00004002	BUC	X11 diag rte	25 0 10	0 0 0	#2643 #2645	GXT500 Graphics Adapter, GXT500D Graphics Adapter
00004005	BUC	X11 diag rte	0 0 2.0	0 0 0	N/A	GXT150 Graphics Adapter
00004006	BUC	X11 diag rte	0 4.3.2. 0	0 0 0	N/A	GXT150L Graphics Adapter
00004007	BUC	X11 diag rte	0 0 2.0	0 0 0	N/A	GXT155L Graphics Adapter
c1x	ISA	com diag rte	N/A 0 0	15 10 N/A	#2961 (*)	X.25 Interface Co-Processor ISA Adapter
cxia	ISA	com diag rte ucode	79 0 0 10	35 N/A N/A N/A	#2931 (3-8)	8-Port Asynchronous EIA-232 ISA Adapter

Device number	Bus type	Fileset	AIX Version 4.3.3	AIX Version 5.1.0	Feature Code	Description
cxia128	ISA	diag rte ucode	0 0 10	N/A 0 N/A	#2933 (3-9)	128-Port Asynchronous Controller ISA
mm2	ISA	diag rte mpqp	0 10 75	N/A N/A 15	#2701 (*)	Co-Processor Multiport Adapter, Model 2
pc8s	ISA	diag rte	0 0	0 0	#2932 (3-A)	8-Port Asynchronous EIA-232E/RS-422A ISA Adapter
PNP80CC	ISA	rte	0	0	#2971 (*)	16/4 Token-Ring ISA Adapter
IBM0010	ISA	rte	10	0	#2981 (*)	ISA Ethernet Adapter for 7020/7248

Miscellaneous hardware

Some older adapters are not supported in AIX 4 and 5, some others have filesets that do not follow the previously described conventions, and for some adapters, the information about the drivers are unclear. They are listed in the following sections.

Not supported on AIX 4 and AIX 5L

The following adapter is not supported on AIX 4 and AIX 5L:

- Fibre Channel/266 Adapter #1906 (8-X) is not supported beyond AIX 3.2.5.

Artic device family

The following adapters are part of the Artic device family:

- PCI ARTIC960RxD Quad Digital Trunk #6310 (6-E) (only supports the 32-bit kernel of AIX 5L Version 5.1).
- PCI ARTIC960RxF Digital Trunk Resource #6311 (6-G) (only supports the 32-bit kernel of AIX 5L Version 5.1).
- PCI Digital Trunk Quad Adapter #6309 (6-B) (not supported in AIX 5L).
- ARTIC960 Adapter - #2921, 2924, 2928 (9-1) separate LPP sx25.* (MCA cards).

The common driver is devices.artic960.rte 1.4.4.0 (AIX 4 and 5L).

Drivers with other naming conventions

For PCI cards, the following adapters do not have drivers following the naming conventions:

- PCI Digital Trunk Quad Adapter #6309 (6-B)
- Eicon ISDN DIVA Pro 2.0 S/T 2708 #2708 (9-N) (driver on diskette, not supported on AIX 5L)
- IBM Short-wave Serial HIPPI PCI Adapter (#2732, only 32-bit mode on AIX 5L)
- IBM Long-wave Serial HIPPI PCI Adapter (#2733, only 32-bit mode on AIX 5L)

The following are MCA adapters:

- X.21 Communications Controller #2938 (9-2) seperate LPP sx25.*.
- EIA-232E Communications Controller #2929 (9-3) seperate LPP sx25.*.
- SAMI SP Attach Clustered Server Control Panel to CWS #3154 (6-K) is included in the server base microcode.

List of common devices

This section lists common device drivers that support a special class (for example, SCSI) or common functionality of multiple devices. They are alphabetically ordered for the particular class. For example, devices.common.base.rte.4.3.3.75 is in the common subsection.

BASE: (devices.base)

- diag
  - Common diagnostics
  - 5.1.0.25, 4.3.3.75
- rte
  - Real-time environment
  - 5.1.0.10, 4.3.3.50

CHRP: (devices.chrp)

- base
  - RISC PC Base System Device Software (CHRP)
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.79

  - diag: 5.2.2.0, 5.1.0.26, 4.3.3.78
  - ServiceRM: 1.2.0.0, 1.1.0.30
- pci
  - PCI Bus Software (CHRP)
  - rte: 5.2.0.0, 5.1.0.25, 4.3.3.75

CHRP_LPAR (devices.chrp_lpar)
Only AIX V5

- base
  - RISC PC Base System Device Software for lpar (CHRP)
  - ras: 5.2.0.0, 5.1.0.15
  - rte: 5.2.0.0, 5.1.0.35

COMMON: (devices.common)

- base
  - Common Base System Diagnostics
  - diag: 5.2.0.0, 5.1.0.25, 4.3.3.75
- rspcbase
  - rte: 5.2.0.0, 4.3.3.0
- IBM.async
  - Asynchronous Software
  - rte: 5.2.0.0, 5.1.0.0, 4.3.3.25
- IBM.atm
  - ATM Software
  - rte: 5.2.0.0, 5.1.0.25, 4.3.3.81
- IBM.bbl
  - Graphics Adapter Diagnostics
  - 4.3.3.0
- IBM.cx
  - CX Adapter Software
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.77
- IBM.disk
  - Common Disk Software

  - rte: 5.2.0.0, 5.1.0.25, 4.3.3.0
- IBM.ethernet
  - Common Ethernet Software
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.80
- IBM.esconCU.mpc
  - Multipath Channel Driver
  - rte: 2.1.4.1
- IBM.fc
  - Common Fibre Channel Software
  - rte: 5.2.0.0, 5.1.0.10, 4.3.3.75
  - hba-api: 5.2.0.0
- IBM.fda
  - Common Diskette Support
  - rte: 5.2.0.0, 5.1.0.25, 4.3.3.1
  - diag: 5.2.0.0, 5.1.0.25, 4.3.3.51
- IBM.fddi
  - Common FDDI Software
  - rte: 5.2.0.0, 5.1.0.0, 4.3.3.50
- IBM.hdlc
  - Common HDLC Software
  - rte: 5.2.0.0, 5.1.0.15, 4.3.3.51
  - sdlc: 5.2.0.0, 5.1.0.35, 4.3.3.25
- IBM.iscsi
  - Common ISCSI Software (5.2 only)
  - rte: 5.2.0.0
- IBM.ktm_std
  - Common Keyboard, Mouse, and Tablet Software
  - diag: 5.2.0.0, 4.3.3.0
  - rte: 5.2.0.0, 4.3.3.0
- IBM.modemcfg
  - Modem configuration
  - data: 5.2.0.0, 4.3.1.0

- IBM.mpio
  - MPIO Disk Path Control (5.2 only)
  - data: 5.2.0.0, 4.3.1.0
- IBM.pmmd_chrp
  - Power Management Software
  - rte: 4.3.3.0
- IBM.ppa
  - Parallel Printer Adapter
  - diag: 5.2.0.0, 4.3.3.25
  - rte: 5.2.0.0, 4.3.3.0
- IBM.rby
  - GXT1000 Graphics Adapter Diagnostics
  - 4.3.3.0
- IBM.scsi
  - SCSI I/O Controller Software
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.75
- IBM.son
  - GXT Common Graphics Adapter Software
  - rte: 5.1.0.15, 4.3.3.75
  - diag: 5.2.0.0, 5.1.0.25, 4.3.3.76
- IBM.ssa
  - SSA Adapter Software
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.76
  - diag: 5.2.0.0, 5.1.0.15, 4.3.3.76
- IBM.tokenring
  - Common Token-Ring Software
  - rte: 5.2.0.0, 5.1.0.10, 4.3.3.50
- IBM.usb
  - Common USB Software (V5 only)
  - rte: 5.2.0.0, 5.1.0.35
  - diag: 5.2.0.0, 5.1.0.25
  - IBM.ARTIC

- Common Artic Software
- diag: 4.3.3.0, 5.1.0.0

FCP: (devices.fcp)

- disk.array
  - Fibre Channel SCSI RAID Software
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.81
  - diag: 5.2.0.0, 4.3.3.50
- disk
  - Fibre Channel SCSI CD-ROM, Disk Device Software
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.76
- tape
  - Fibre Channel SCSI Tape Device Software
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.77

GRAPHICS: (devices.graphics)

- com
  - Graphics Adapter Common
  - 5.2.0.0, 5.1.0.35, 4.3.3.75
- voo
  - Stereo and VOO Software
  - 5.2.0.0, 4.3.3.0

IDE: (devices.ide)

- cdrom
  - IDE CD-ROM support
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.76
  - diag: 5.2.0.0, 5.1.0.35, 4.3.3.1
- disk
  - rte: 5.2.0.0, 5.1.0.0, 4.3.3.0

ISA_SIO: (devices.isa_sio)

- chrp.ecp
  - CHRP IEEE1284 Parallel Port Adapter
  - rte: 5.2.0.0, 5.1.0.23, 4.3.3.76

  - diag: 5.2.0.0, 4.3.1.25
- chrp.8042
  - ISA Keyboard and Mouse Software
  - diag: 5.2.0.0, 5.1.0.0, 4.3.3.0
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.51
- km
  - ISA Keyboard and Mouse Software
  - diag: 5.1.0.0, 4.3.3.50
  - rte: 5.1.0.35, 4.3.3.51
- baud
  - Audio Device Software RISC Ultimedia
  - rte: 5.1.0.25, 4.3.2.1
- pnpPNP.400
  - Standard Parallel Adapter Software
  - diag: 5.2.0.0, 4.3.1.0
  - rte: 5.2.0.0, 4.3.0.0
- pnpPNP.501
  - CHRP Serial Adapter Software
  - diag: 5.2.0.0, 4.3.0.0
  - rte: 5.2.0.0, 4.3.3.0
- pnpPNP.700
  - CHRP Diskette Adapter
  - diag: 5.2.0.0, 4.3.0.0
  - rte: 5.2.0.0, 4.3.3.0
- IBM0005.IBM8301
  - ISA Power Management Controller
  - rte: 4.3.3.0
- IBM000E
  - Ultimedia RISC Audio Device
  - rte: 4.3.2.0
- IBM0012
  - Empty (for Gameport)

  - N/A: 4.3.3.0
- IBM0013
  - Empty (MIDI Support)
  - N/A: 4.3.3.0
- IBM0017
  - Audio Device RISC PC
  - rte: 5.2.0.0, 4.3.2.0, 5.1.0.15
  - diag:5.2.0.0, 4.3.3.76, 5.1.0.25
- IBM0019
  - ISA Tablet Software
  - diag: 5.2.0.0, 5.1.0.0, 4.3.3.0
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.25
- IBM001C
  - EPP Parallel Port Adapter Software
  - rte: 4.3.3.0
- IMB001D
  - Empty (Yamaha Audio Support)
  - N/A: 4.3.3.0
- IBM001E
  - Service Processor Software
  - diag: 4.3.1.0
  - rte: 4.3.3.0
- IBM001F
  - Ring Indicate Power-On
  - diag: 4.3.3.0
  - rte: 4.3.3.0
- PNP0303
  - ISA Keyboard Software
  - diag: 4.3.3.51, 5.1.0.35
  - rte: 4.3.3.0
- PNP0400
  - RISC PC Standard Parallel Port Adapter

  - rte: 5.1.0.10, 4.3.3.75
  - diag: 4.3.1.0
- PNP0401
  - RISC PC ECP Parallel Port Adapter
  - rte: 5.1.0.0, 4.3.3.0
  - diag: 4.3.1.0
- PNP0501
  - RISC PC Standard Serial Adapter
  - rte: 5.1.0.35, 4.3.3.51
  - diag: 4.3.3.0
- PNP0600
  - IDE Adapter Device
  - com: 5.1.0.35, 4.3.1.1
  - rte: 5.1.0.25, 4.3.3.76
- PNP0700
  - Diskette Adapter Software
  - diag: 4.3.3.0
  - rte: 4.3.3.25
- PNP0E00
  - PCMCIA Bus Software
  - rte: 4.3.3.0
- PNP0F03
  - ISA Mouse Software
  - diag: 4.3.3.25
  - rte: 4.3.3.0

ISCSI: (devices.pci) (5.2 only)

- disk
  - iSCSI Disk Software
  - rte: 5.2.0.0
- tape
  - iSCSI Tape Software

- rte: 5.2.0.0

PCI: (devices.pci)

- ibmccm
  - Common Character Mode Graphics Adapter (V5 only)
  - rte: 5.2.0.0, 5.1.0.35
- isa
  - ISA Bus Bridge (CHRP)
  - rte: 5.2.0.0, 4.3.3.0
- pci
  - PCI Bus Bridge
  - rte: 5.2.0.0, 4.3.3.0
- PNP0A00
  - ISA Bus Bridge (V5 only)
  - 5.1.0.0
- PNP0A03
  - PCI Bus Bridge (V5 only)
  - 5.1.0.0

PCMCIA: (devices.pcmcia)

- ethernet
  - PCMCIA Ethernet
  - 4.3.3.0
- serial
  - PCMCIA Serial Port
  - com: 4.3.3.1
- tokenring
  - PCMCIA Token-Ring
  - com: 4.3.3.0

RS6KSMP: (devices.rs6ksmp)

- base
  - Multiprocessor Base System
  - rte: 4.3.3.50, 5.1.0.0

RSPC: (devices.rspc)

- base
  - RISC PC Base System
  - diag: 4.3.3.51, 5.1.0.25
  - rte: 4.3.3.76, 5.1.0.35

SCSI: (devices.scsi)

- disk
  - SCSI CD-ROM, Disk
  - diag.com: 5.2.0.0, 4.3.3.77, 5.1.0.35
  - diag.rte: 5.2.0.0, 4.3.3.75, 5.1.0.35
  - rspc: 5.2.0.0, 4.3.3.25, 5.1.0.0
  - rte: 5.2.0.0, 4.3.3.76, 5.1.0.35
- safte
  - SCSI Accessed Fault-Tolerant Enclosure
  - rte: 5.2.0.0, 5.1.0.0
  - diag: 5.2.0.0
- scarray
  - 7135 RAIDiant Array
  - diag: 5.2.0.0, 4.3.3.50, 5.1.0.0
  - rte: 5.2.0.0, 4.3.3.50, 5.1.0.0
- ses
  - SCSI Enclosure Services
  - diag: 5.2.0.0, 4.3.3.77, 5.1.0.25
  - rte: 5.2.0.0, 4.3.3.75, 5.1.0.35
- tape
  - SCSI Tape Device
  - diag: 5.2.0.0, 4.3.3.77, 5.1.0.35
  - rspc: 5.2.0.0, 4.3.3.25, 5.1.0.0
  - rte: 5.2.0.0, 4.3.3.12, 5.1.0.35
- tm
  - SCSI Target Mode

- rte: 5.2.0.0, 4.3.3.75, 5.1.0.35

SERIAL: (devices.serial)

- gio
  - General IO for serial graphics input adapter
  - rte: 5.2.0.0, 4.3.3.50, 5.1.0.0
  - diag: 5.2.0.0, 5.1.0.0, 4.3.3.0
  - X11: 5.2.0.0, 5.1.0.0
- sb1
  - Spaceball 3-D input device V5 only
  - X11: 5.2.0.0, 5.1.0.0
- tablet1.X11
  - AIXwindows Serial Tablet Input Device
  - X11: 5.2.0.0, 5.1.0.0

SIO (devices.sio)

- fda
  - Diskette Drive Adapter
  - diag: 4.3.3.0
- ktma
  - Keyboard, Tablet, and Mouse
  - diag: 5.1.0.0, 4.3.3.0
  - rte: 5.1.0.35, 4.3.3.1
- ppa
  - Parallel Printer Adapter
  - rte: 4.3.3.75, 5.1.0.10
  - diag: 5.1.0.0, 4.3.1.0
- sa
  - Built-in Serial Adapter
  - diag: 5.1.0.0
  - rte: 5.1.0.0, 4.3.2.0

SSA: (devices.ssa)

- disk
  - SSA DASD Software
  - rte: 5.2.0.0, 4.3.3.76, 5.1.0.25
- IBM_raid
  - SSA Raid Manager
  - rte: 5.2.0.0, 4.3.3.50, 5.1.0.0
- tm
  - rte: 5.2.0.0, 4.3.3.26, 5.1.0.35
- network_agent
  - rte: 4.3.3.0

SYS: (devices.sys)

- PNP0A03
  - V5 only
- mca
  - Microchannel Bus
  - rte: 5.1.0.15, 4.3.3.1
- pci
  - PCI Bus
  - rte: 4.3.3.75, 5.1.0.35
- sga
  - Graphics Slot for Gt1 Graphics Adapter 7011-220 (V4 only)
  - X11
  - diag
  - rte
- sgabus
  - Special Graphics Slot
  - rte: 5.1.0.0, 4.3.3.0
- wga
  - Graphics Slot for Gt1x Graphics Adapter for 7011-220/230
  - X11

  - diag
  - rte
- slc
  - Serial Optical Link
  - diag: 5.1.0.0, 4.3.3.0
  - rte: 5.1.0.0, 4.3.3.0

TTY: (devices.tty)

- rte
  - Device Driver Support Software
  - rte: 5.2.0.0, 5.1.0.35, 4.3.3.0

USBIF: (devices.usbif)

- 030101.rte
  - USB Keyboard Client Driver
  - rte: 5.2.0.0, 5.1.0.35
- 030102.rte
  - USB Mouse Client Driver
  - rte: 5.2.0.0, 5.1.0.35

Abbreviations and acronyms

ACK	acknowledgement
ACL	access control list
AIX	Advanced Interactive Executive
APAR	authorized program analysis report
APC	Automated Power Control
API	application programming interface
ASCII	American Standard Code for Information Interchange
BI	business intelligence
BOS	basic operating system
CD-ROM	compact disk, read only mode
CEC	central electronic complex
CLVM	Concurrent Logical Volume Manager
ConfigRM	configuration resource manager
CPU	central processing unit
CSM	Cluster Systems Management
CSP	Converged Support Processor
CSS	Communication Subsystem
CtSec	RS/6000 RSCT Cluster Security Service
CVSD	Concurrent Virtual Shared Disk
CWS	control workstation
DASD	direct access storage device
DCEM	Distributed Cluster Execution Manager
DDR	double data rate
DMA	direct memory access
DVD	digital versatile disk
EPOW	Early Power Off Warning
ESS	Enterprise Storage Server
ESSL	Engineering and Scientific Subroutine Library
FC	Feature Code
GB	gigabyte (10^9 byte)
GeoRM	Geographic Remote Mirror
GHz	gigahertz (10^9 Hz)
GPFS	General Parallel File System
GUI	graphical user interface
HA	high availability
HACMP	High-Availability Cluster Multiprocessing
HACMP/ES	High-Availability Cluster Multiprocessing/Enhanced Scalability
HAI	High Availability Infrastructure
HAGEO	High Availability Geographic Cluster
HMC	Hardware Management Console
HPC	High Performance Computing
HTTP	Hypertext Transfer Protocol
I/O	input/output
IBM	International Business Machines Corporation
IDE	integrated drive electronics
IP	Internet Protocol
ISA	Industry Standard Architecture
ITSO	International Technical Support Organization
JFS	Journaled File System

KLAPI	Kernel Low-Level Application Programming Interface
KVM	keyboard, video, mouse
LAN	local area network
LAPI	Low-Level Application Programming Interface
LED	light-emitting diode
LL	LoadLeveler
LPAR	logical partition
LPP	Licensed Program Product
LV	Logical Volume
LVM	Logical Volume Manager
MAC	medium access control
MACN	Management and Control Node or control workstation
MB	megabyte (10^6 byte)
MCA	Micro Channel Architecture
MCM	multi-chip module
MFLOP	mega floating point operations per second
MHz	megahertz (10^6Hz)
MP3	Motion Picture Group Encoding Standard
MPI	message passing interface
MX	Memory Expansion
NFS	Network File System
NUMA	non-uniform memory access
NIM	Network Installation and Maintenance
NSD	Network Shared Disk
OS	operating system
PAM	pluggable authentication module
PCI	peripheral component interconnect
PCI-X	peripheral component interconnect extended
PE	Parallel Environment
Parallel ESSL	Parallel Engineering and Scientific Subroutine Library
PSSP	Parallel System Support Program
PTF	program temporary fix
RAS	reliability, availability, and serviceability
RIO	remote input/output
RM	resource manager
RMC	Resource Monitoring and Control subsystem
RPC	remote procedure call
RPD	RSCT peer domain
RSCT	Reliable Scalable Cluster Technology
RVSD	Recoverable Virtual Shared Disk
SAMI	Service and Manufacturing Interface
SCSI	small computer system interface
SDR	System Data Repository
SEPBU	Scalable Electrical Power Base Unit
SMIT	System Management Interface Toolkit
SMP	symmetric multiprocessing
SNMP	Simple Network Management Protocol
SP	Scalable Parallel
SPOT	Shared Product Object Tree
SSL	Secure Sockets Layer
TCP	Transmission Control Protocol
TTY	Teletype
UDP	User Datagram Protocol
UPS	uninterruptible power supply
URL	Uniform Resource Locator
VG	Volume Group

VM	virtual machine
VSD	Virtual Shared Disk
WebSM	Web-Based System Manager
WLM	Workload Manager

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 235.

- *A Practical Guide for Resource Monitoring and Control (RCM)*, SG24-6615
- *AIX 5L Differences Guide Version 5.2 Edition*, SG24-5765
- *An Introduction to CSM 1.3 for AIX 5L*, SG24-6859
- *Configuring Highly Available Clusters Using HACMP 4.5*, SG24-6845
- *GPFS on AIX Clusters: High Performance File System Administration Simplified*, SG24-6035
- *IBM eServer Cluster 1600 and PSSP 3.4 Cluster Enhancements*, SG24-6604
- *IBM eServer pSeries 690 System Handbook*, SG24-7040
- *Linux Clustering with CSM and GPFS,* SG24-6601
- *pSeries 630 Models 6C4 and 6E4 Technical Overview and Introduction*, REDP0193
- *RS/6000 SP Cluster: The Path to Universal Clustering*, SG24-5374
- *Universal Clustering Problem Determination Guide*, SG24-6602
- *Workload Management with LoadLeveler*, SG24-6038

Other resources

These publications are also relevant as further information sources:

- *AIX General Programming Concepts: Writing and Debugging Programs*

  For AIX Version 4.3, see:

  http://publib.boulder.ibm.com/doc_link/en_US/a_doc_lib/aixprggd/genprogc/toc.htm

For AIX 5L Version 5.1, see:

http://publibn.boulder.ibm.com/doc_link/en_US/a_doc_lib/aixprggd/genprogc/genprogctfrm.htm

For AIX 5L Version 5.2, see:

http://publib16.boulder.ibm.com/pseries/en_US/aixprggd/genprogc/genprogc.pdf

- *General Parallel File System for AIX 5L: AIX Clusters Concepts, Planning, and Installation Guide*, GA22-7895
- *General Parallel File System for AIX 5L: PSSP Clusters Concepts, Planning, and Installation Guide*, GA22-7899
- *IBM General Parallel File System for AIX: Concepts, Planning, and Installation*, GA22-7453
- *IBM General Parallel File System for Linux: Concepts, Planning, and Installation*, GA22-7844
- *IBM RSCT for AIX: Guide and Reference*, SA22-7889
- *IBM RSCT: Group Services Programming Guide and Reference*, SA22-7888
- *pSeries p655 Installation Guide*, SA38-0616
- *PSSP for AIX: Administration Guide*, SA22-7348
- *PSSP for AIX: Command and Technical Reference, Volume 2*, SA22-7351
- *PSSP for AIX: Diagnosis Guide*, GA22-7350
- *PSSP for AIX: Installation and Migration Guide*, GA22-7347
- *PSSP for AIX: Managing Shared Disks*, SA22-7349
- *RS/6000 and eServer pSeries: PCI Adapter Placement Reference*, SA38-0538
- *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281

Referenced Web sites

These Web sites are also relevant as further information sources:

- For GPFS documentation, refer to the following Web sites:

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

  Or

  http://www.ibm.com/shop/publications/order

- To obtain the latest service level for all required software, refer to the following Web site:

  http://techsupport.services.ibm.com/server/fixes

- For information about FAStT disk subsystems, see:

  http://www.storage.ibm.com/hardsoft/disk/fastt/

- For information about Enterprise Storage Server (ESS) disk subsystems, see:

  http://www.storage.ibm.com/hardsoft/products/ess/index.html

- For information about AIX 64-bit compatibility with adapters, see:

  http://www.ibm.com/servers/aix/os/adapters/51.html

- For the latest PSSP documentation and a current list of supported control workstations, see:

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

How to get IBM Redbooks

You can order hardcopy Redbooks, as well as view, download, or search for Redbooks at the following Web site:

ibm.com/redbooks

You can also download additional materials (code samples or diskette/CD-ROM images) from that site.

IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Index

D

E

F

G

H

I

J

K

L

M

N

P

Q

R

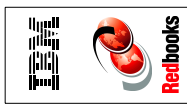
S

T

V

W

X



IBM @server Cluster 1600 Managed by PSSP 3.5: What's New

IBM @server Cluster 1600 Managed by PSSP 3.5: What's New

Explore PSSP 3.5 enhancements including 64-bit support

Plan and manage your Cluster 1600 into the future

Tour the latest GPFS features

This IBM Redbook explores the evolution of the IBM RS/6000 SP system into the IBM @server Cluster 1600 and the impact of pSeries POWER4 LPAR technology in the pSeries clusters. This publication also highlights the new pSeries servers, which can be incorporated into Cluster 1600. This book provides pSeries cluster configuration information, including hardware and software hints and tips, as well as changes in the packaging of the cluster management components: AIX 5L and Parallel System Support Program (PSSP).

An overview of Reliable Scalable Cluster Technology (RSCT) is included to introduce the reader to the latest developments of the RSCT clustering software. The latest enhancements in PSSP 3.5 are included, highlighting in particular the changes made to the switch software and Virtual Shared Disks (VSD). Configuration architectures and examples are included for customers planning to deploy a Cluster 1600 in their computing environment. PSSP 3.5 and General Parallel File System (GPFS) enhancements are explored, including the latest 64-bit support and the latest supported levels of AIX 5L. This redbook also includes helpful information about software coexistence, migration, and integration in Cluster 1600. Finally, a high-level comparison between PSSP 3.5 and the new IBM @server Cluster 1600 Cluster Systems Management software is provided.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-6617-00 ISBN 0738426652