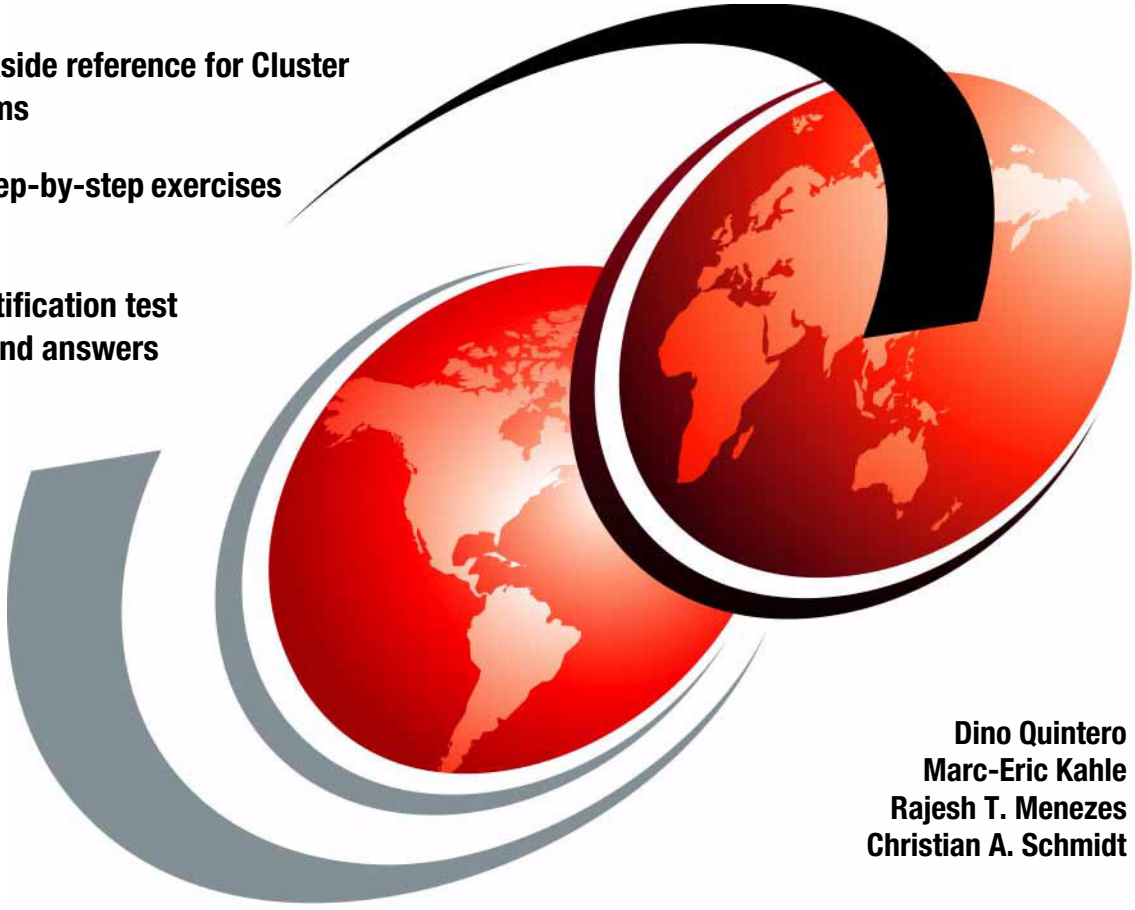
IBM @server

IBM

IBM @server Certification Study Guide: Cluster 1600 Managed by PSSP

- Handy deskside reference for Cluster 1600 systems
- Detailed, step-by-step exercises
- Sample certification test questions and answers



Dino Quintero
Marc-Eric Kahle
Rajesh T. Menezes
Christian A. Schmidt

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM @server Certification Study Guide: Cluster 1600 Managed by PSSP

December 2003

SG24-7013-00

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

First Edition (December 2003)

This edition applies to PSSP Version 3, Release 5, and AIX Version 5, Release 2.

Contents

Notices

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:
This information contains sample application programs in source language, which 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:

AFS®	IBM®	Redbooks™
AIX 5L™	LoadLeveler®	Redbooks (logo) ™
AIX®	Magstar®	Requisite®
Domino®	Micro Channel®	RS/6000®
DFS™	PowerPC®	SP2®
Enterprise Storage Server®	POWER3™	TotalStorage®
ESCON®	POWER4™	TURBOWAYS®
@server ™	pSeries®	Versatile Storage Server™
FICON™	PTX®	zSeries®

The following terms are trademarks of other companies:

Intel, Intel Inside (logos), and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

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

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

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

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

Preface

This IBM Redbook is designed as a study guide for professionals wishing to prepare for the certification exam to achieve IBM Certified Specialist - Cluster 1600 managed by PSSP.

The Cluster 1600 managed by PSSP certification validates the skills required to install and configure PSSP system software and to perform the administrative and diagnostic activities needed to support multiple users in an SP environment. The certification is applicable to specialists who implement and/or support Cluster 1600 managed by PSSP systems.

This redbook helps Cluster 1600 specialists seeking a comprehensive and task-oriented guide for developing the knowledge and skills required for certification. It is designed to provide a combination of theory and practical experience needed for a general understanding of the subject matter. It also provides sample questions that will help in the evaluation of personal progress and provides familiarity with the types of questions that will be encountered in the exam.

This redbook does not replace the practical experience you should have. Instead, it is an effective tool that, when combined with education activities and experience, should prove to be a very useful preparation guide for the exam. Due to the practical nature of the certification content, this publication can also be used as a desk-side reference. So, whether you are planning to take the RS/6000 SP and PSSP exam, or whether you just want to validate your RS/6000 SP skills, this book is for you.

The AIX® and RS/6000® Certifications offered through the Professional Certification Program from IBM® are designed to validate the skills required of technical professionals who work in the powerful and often complex environments of AIX and RS/6000. A complete set of professional certifications is available. They include:

- IBM Certified AIX User
- IBM Certified Specialist - RS/6000 Solution Sales
- IBM Certified Specialist - AIX 5L™ System Administration
- IBM Certified Specialist - AIX System Support
- IBM Certified Specialist - Cluster 1600 Managed by PSSP
- Cluster 1600 Sales Qualification
- IBM Certified Specialist - AIX HACMP
- IBM Certified Specialist - Domino® for RS/6000
- IBM Certified Specialist - Web Server for RS/6000

- IBM Certified Specialist - Business Intelligence for RS/6000
- IBM Certified Advanced Technical Expert - RS/6000 AIX

Each certification is developed by following a thorough and rigorous process to ensure that the exam is applicable to the job role and is a meaningful and appropriate assessment of skill. Subject matter experts who successfully perform the job participate throughout the entire development process. These job incumbents bring a wealth of experience into the development process, thus making the exams more meaningful than the typical test that only captures classroom knowledge. These experienced subject matter experts ensure that the exams are relevant to the *real world* and that the test content is both useful and valid. The result is a certification of value, which appropriately measures the skill required to perform the job role.

For additional information about certification and instructions on how to register for an exam, call IBM at 1-800-426-8322 or visit the IBM Certification Web site at:

http://www.ibm.com/certify

The team that wrote this redbook

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

Dino Quintero is an IBM Certified Senior IT Specialist at the International Technical Support Organization, Poughkeepsie Center. He spends time assessing, designing, and implementing pSeries/AIX technical solutions for various customer sets, including those in clustered environments, writing Redbooks™, and teaching workshops.

Marc-Eric Kahle is a pSeries® Hardware Support specialist at the IBM ITS Central Region Back Office in Ehningen, Germany. He has experience in the RS/6000, pSeries, and AIX fields since 1993. He has worked at IBM Germany for 16 years. His areas of expertise include RS/6000 and pSeries hardware, including the SP, and he is also an AIX certified specialist. He has participated in the development of three other redbooks.

Rajesh T. Menezes is a pSeries Senior Specialist in IBM Global Services India. He has 11 years of experience in the IT support and services field. He has worked at IBM for four years. His areas of expertise include RS/6000, SP, pSeries, AIX, PSSP, HACMP, CLuster 1600, and storage. He has also conducted customer training for AIX 5L (Basic and Advanced) for customers and IBM internal (country level) and BP, training for SP administration. He is also certified on pSeries AIX5L Version 5.1.

Christian A. Schmidt is an Advisory IT Specialist working for the Strategic Outsourcing Division of IBM Global Services in Denmark. He has worked for IBM for 10 years. During the last five years he has been specializing in Cluster 1600, providing support and 2nd level support for AIX and Cluster 1600 configurations for the IBM Software Delivery and Fulfillment in Copenhagen. His area of expertise includes designing and implementing highly available Cluster 1600 solutions, AIX, PSSP, GPFS, CSM, security, system tuning and performance. Christian is an IBM Certified Specialist in SP and System Support. He is also the co-author of *Managing IBM (e)server Cluster 1600 - Power Recipes for PSSP 3.4,* SG24-6603.

Team members from left to right: Christian A. Schmidt, Dino E. Quintero (project leader), Rajesh T. Menezes, Marc-Eric Kahle

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

Scott Vetter
International Technical Support Organization, Austin Center

Hans Mozes
IBM Germany

Christopher V Derobertis, David Wong, Paul J Swiatocha, Lissa Valletta, ShujunZhou, Ed Biro, Dave Delia
pSeries clusters development lab, Poughkeepsie

Alfred Schwab, editor
International Technical Support Organization, Poughkeepsie

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

Introduction

This guide is not a replacement for the SP product documentation or existing ITSO redbooks, or for the real experience of installing and configuring SP environments.

SP knowledge alone is not sufficient to pass the exam. *Basic AIX* and *AIX admin* skills are also required.

You are supposed to be fluent with all topics addressed in this redbook before taking the exam. If you do not feel confident with your skills in one of these topics, you should go to the documentation referenced in each chapter.

The SP Certification exam is divided into two sections:

- *Section One* - Is a series of general SP- and PSSP-related questions.
- *Section Two* - Is based on a scenario in a customer environment that begins with a basic SP configuration. In this scenario, as a customer's requirements evolve, so does the SP configuration. As the scenario develops, additional partitions, nodes, frames, and system upgrades are required.

In order to prepare you for both sections, we have included a section in each chapter that lists the key concepts that should be understood before taking the exam. This scenario is described in 1.2, "The test scenario" on page 3.

1.1 Book organization

This guide presents you with all domains in the scope of the IBM eServer Cluster 1600 managed by PSSP certification exam. The structure of the book follows the normal flow that a standard Cluster 1600 installation takes.

Part 1, “System planning” on page 5, contains chapters dedicated to the initial planning and setup of a standard Cluster 1600 managed by PSSP. It also includes concepts and examples about PSSP security and user management.

Part 2, “Installation and configuration” on page 273, contains chapters describing the actual implementation of the various steps for installing and configuring the control workstation, nodes, and switches. It also includes a chapter for system verification as a post-installation activity.

Part 3, “Application enablement” on page 393, contains chapters for the planning and configuration of additional products that are present in most SP installations, such as the IBM Virtual Shared Disk and the IBM Recoverable Virtual Shared Disk, as well as GPFS. There is also a section dedicated to problem management tools available in PSSP.

Part 4, “On-going support” on page 417, contains chapters dedicated to software maintenance, system reconfiguration including migration, and problem determination procedures and checklists.

Each chapter is organized as follows:

- *Introduction* - This contains a brief overview and set of goals for the chapter.
- *Key concepts you should study* - This section provides a list of concepts that need to be understood before taking the exam.
- *Main section* - This contains the body of the chapter.
- *Related documentation* - Contains a comprehensive list of references to SP manuals and redbooks with specific pointers to the chapters and sections covering the concepts in the chapter.
- *Sample questions* - A set of questions that serve two purposes: To check your progress with the topics covered in the chapter, and to become familiar with the type of questions you may encounter in the exam.
- Exercises - The purpose of the exercise questions is to further explore and develop areas covered in the chapter.

There are many ways to perform the same action in an SP environment: Command line, SMIT or SMITTY, spmon -g (PSSP 2.4 or earlier), IBM SP Perspectives, and so on. The certification exam is not restricted to one of these

methods. You are expected to know each one, in particular the syntax of the most useful commands.

1.2 The test scenario

To present a situation similar to the one you may encounter in the SP Certification exam, we have included a test scenario that we use in all sections of this study guide. The scenario is depicted in Figure 1-1.

We start with the first frame (Frame 1) and 11 nodes, and then add a second frame (Frame 2) later when we discuss reconfiguration in Part 3.

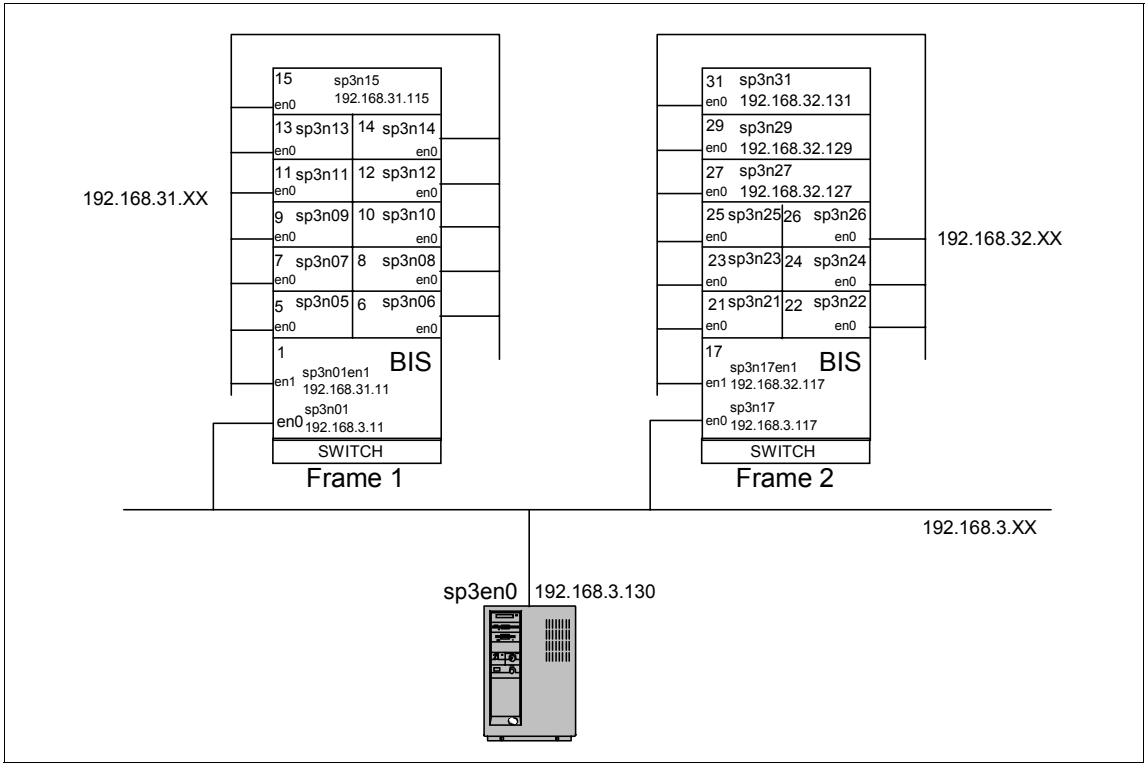


Figure 1-1 Study guide test environment

The environment is fairly complex in the sense that we have defined two Ethernet segments and a boot/install server (BIS) to better support our future expansion to a second frame, where we will add a third Ethernet segment and an additional boot/install server for the second frame.

Although, strictly speaking, we should not need multiple Ethernet segments for our scenario, we decided to include multiple segments in order to introduce an environment where networking, and especially routing, has to be considered. Details about networking can be found in Chapter 3, “Cluster 1600 networking” on page 101.

The boot/install servers were selected following the default options offered by PSSP. The first node in each frame is designated as the boot/install server for the rest of the nodes in that frame.

The frame numbering was selected to be consecutive because each frame has thin nodes in it; hence, it cannot have expansion frames. Therefore, there is no need skipping frame numbers for future expansion frames.

Part 1

System planning

This part contains chapters dedicated to the initial planning and setup of a standard Cluster 1600 managed by PSSP. It also includes concepts and examples about PSSP security and user management.

2

Validate hardware and software configuration

In this chapter the hardware components of the SP and IBM @server Cluster 1600 are discussed, such as node types, control workstations, Hardware Management Console, Logical Partitioning, frames, and switches. It also provides additional information on disk, memory, and software requirements

2.1 Key concepts you should study

The topics covered in this section provide a good preparation toward the RS/6000 SP and IBM eServer Cluster 1600 certification exam. Before taking the exam, you should understand the following key concepts:

- The hardware components that comprise an SP system.
- The types and models of nodes, frames, and switches.
- Hardware and software requirements for the control workstation.
- Levels of PSSP and AIX supported by nodes and control workstations (especially in mixed environments).

2.2 Hardware

The basic components of the IBM eServer Cluster 1600 and RS/6000 SP are:

- The frame with its integral power subsystems
- External and internal processor nodes
- Optional dependent nodes that serve a specific function, such as high-speed network connections
- Optional SP Switch, Switch-8, and SP Switch2 to expand your system
- A control workstation (a high-availability option is also available) and Hardware Management Console (HMC)
- Network connectivity adapters and peripheral devices, such as tape and disk drives

These components connect to your existing computer network through a local area network (LAN), thus making the RS/6000 SP system accessible from any network-attached workstation.

Figure 2-1 on page 9 shows a sample of IBM eServer Cluster 1600 components.

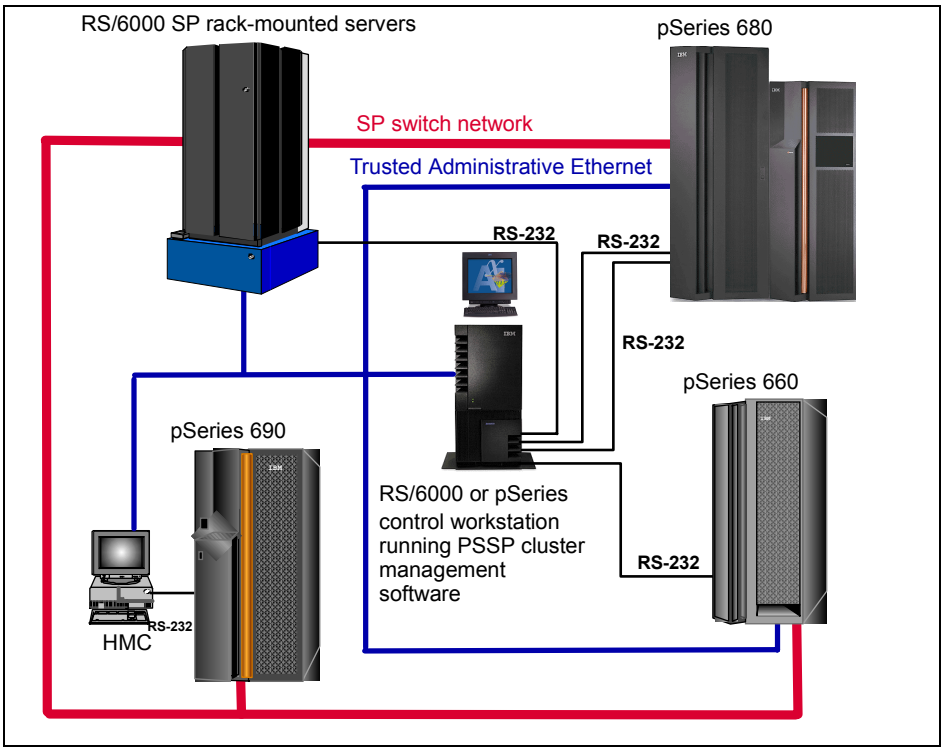


Figure 2-1 Sample of Cluster 1600 managed with PSSP

The building block of RS/6000 SP is the *frame*. There are two sizes: The tall frame (1.93 meters high) and the short frame (1.25 meters high). RS/6000 SP internal nodes are mounted in either a tall or short frame. A tall frame has eight drawers, while a short frame has four drawers. Each drawer is further divided into two slots. A thin node occupies one slot; a wide node occupies one drawer (two slots), and a high node occupies two drawers (four slots). An internal power supply is included with each frame. Frames get equipped with optional processor nodes and switches.

There are five current types of frames:

- The tall model frame
- The short model frame
- The tall expansion frame
- The short expansion frame
- The SP Switch frame

2.2.1 Overview of the available frames

For the Cluster 1600, RS/6000 SP internal nodes and SP Switch/SP Switch2 several frames are available to fit the needs of the desired configuration. A quick summary shows the available frames and their description:

- 7040-T00 19-inch system rack containing p630 (7028-6C4) or p660 (7026-6xx) servers or I/O drawers
- 7040-T42 19-inch system rack containing p630 (#7028) or p660 (#7026) servers or I/O drawers
- 7040-W42 24-inch system frame containing p655 servers (#7039-651) or I/O drawers
- 9076-550 SP frame containing nodes or nodes and switch, including expansion frame (F/C 1550)
- 9076-55H SP frame containing nodes or nodes and switch (administrative field conversion of legacy 79-inch model frame), including expansion frame (F/C 1550)
- 9076-555 SP Cluster Switch Frame, which can contain up to one SP Switch
- 9076-556 SP Switch2 Frame, which can contain up to eight SP Switch2s
- 9076-557 SP Switch Model for 19-inch rack, which can contain up to one SP Switch
- 9076-558 SP Switch2 Model for 19-inch rack, which can contain up to two SP Switch2s

Model 550 - a 1.93m frame

- With eight empty node drawers for either eight wide nodes, 16 thin nodes, or four high nodes
- A 10.5 kW, three-phase SEPBU power subsystem

Model 555 - a 1.93m SP Switch frame

This frame provides switch ports for Cluster 1600 systems of two to 16 logical nodes. For configurations of 17 to 32 nodes, you need to add an SP expansion frame (F/C 1555) containing one SP Switch.

- With one SP Switch node switch board, F/C 4011
- A 10.5 kW, three-phase SEPBU power subsystem

Model 556 - a 2.01m SP Switch2 frame

The Model 556 and F/C 2034 share the 79 in./2.01m frame and covers of the pSeries 690.This frame provides node switch board (NSB) ports for Cluster 1600 and SP processor nodes for both single and two-plane switch configurations

within scaling limits. Installation of more than four SP Switch2s requires the addition of a frame extender for cable management (F/C 9941).

- It has an integral SEPBU power supply
- One to eight SP Switch2 node switch boards F/C 4012, but no processor nodes.

Model 557 - SP Switch package

The Model 557 consists of one SP Switch (F/C 4011) packaged with an SEPBU power supply having redundant power input cables. This package is installed at the customer site. It occupies 16 EIA positions in a separately-ordered, IBM 19" rack (M/T 7014 model T00). The two Model 557 power input cables connect to one receptacle in each of two separate Power Distribution Bus (PDB) outlets in the rack.The Model 557 provides switch ports for Cluster 1600 systems. See Figure 2-2 for a physical overview of the Model 557 package.

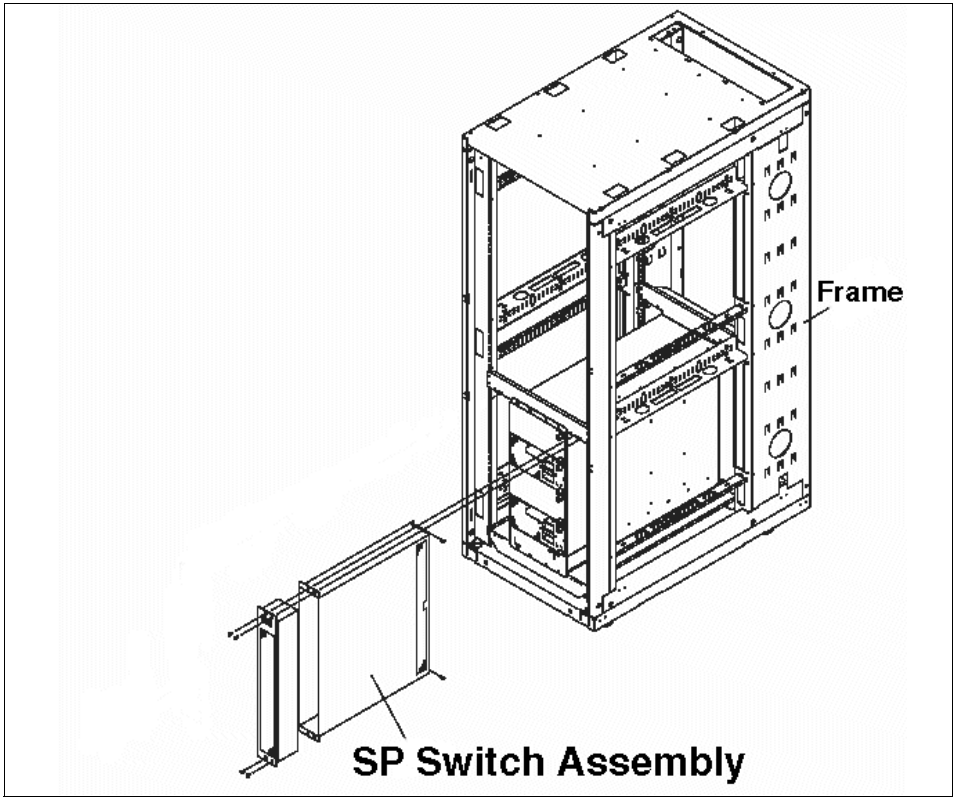


Figure 2-2 Model 557 SP Switch package overview

Model 558 - SP Switch2 package

The Model 558 consists of one to two SP Switch2s (F/C 4012) packaged with an SEPBU power supply having redundant power input cables. This package is installed at the customer site. It occupies 16 EIA positions in a separately-ordered, IBM 19-inch rack (M/T 7014 Model T00 or M/T 7014 Model T42). The Model 558 power input cables connect to one receptacle in each of two separate PDB outlets in the rack.The Model 558 provides node switch board (NSB) ports for Cluster 1600 systems. See Figure 2-3 for a physical overview of the Model 558 SP Switch2 package.

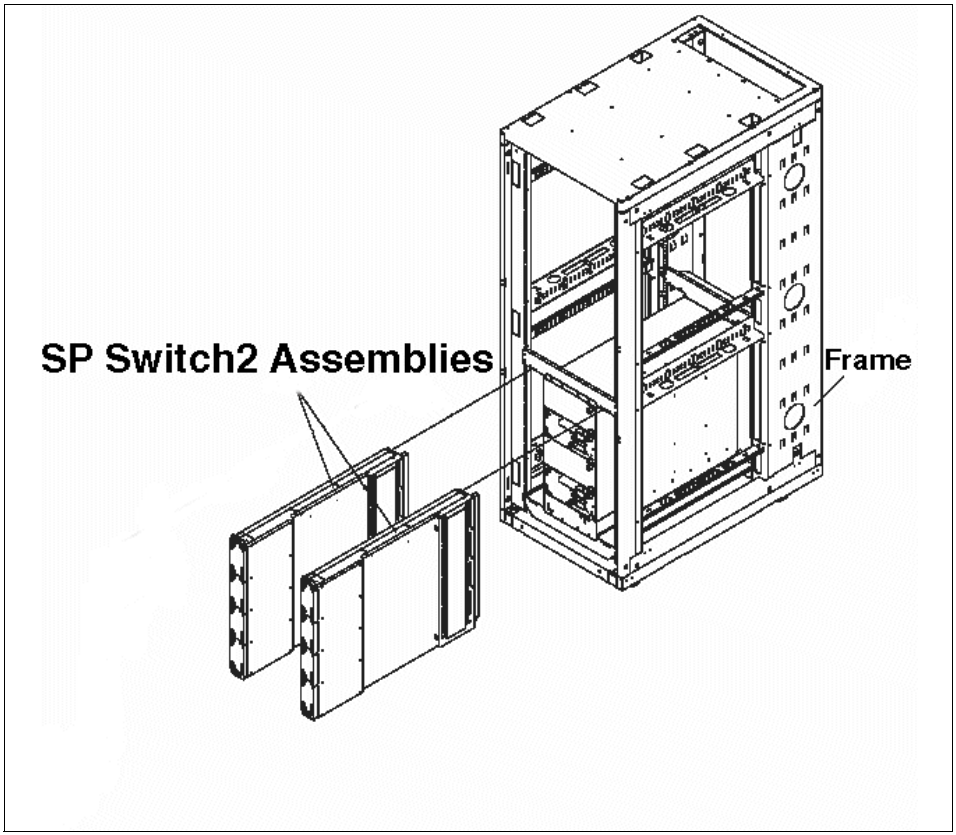


Figure 2-3 Model 558 SP Switch2 package overview

Expansion frame F/C 1550

This expansion frame is a 1.93 m (75.8 inch) tall frame with eight empty node drawers and a 10.5 kW three-phase SEPBU power supply, and is used in Model 550, 3BX, 20X, 30X, 40X, and 55H systems. For these frames, you order the processor nodes and optional switches separately. A switch and up to sixteen thin nodes, eight wide nodes, or four high nodes can be installed in each frame

as permitted by system configuration rules. There are also two possible expansion frame configurations:

- Non-switched expansion frame - only processor nodes can be installed. The unused switch ports of another frame can be used.
- Switched expansion frame - both processor nodes and switch board are installed in the frame.

SP Switch frame F/C 2031

An SP Switch Frame (F/C 2031) is a base (empty), tall frame with integral SEPBU power supply, equipped with four SP Switch intermediate switch boards (ISB) but no processor nodes. The SP Switch Frame is required for systems using more than five SP Switches; it interconnects all the switches in the system. An SP Switch Frame supports systems with from 65 to 128 nodes; however, it can also be configured into systems with fewer than 65 nodes to greatly simplify future expansion as more switches are added.

SP Switch2 frame F/C 2032

An SP Switch2 Frame is a base, tall frame with integral SEPBU power supply, equipped with four SP Switch2 intermediate switch boards (ISB), including their interposers, but no processor nodes. The SP Switch2 frame is required for systems having more than five switches; it interconnects all the switches in the system.For two-plane applications (F/C 9977), four additional SP Switch2 ISBs (F/C 2033) are installed in the SP Switch2 Frame. An SP Switch2 frame supports systems with from 65 to 128 logical nodes; however, it can also be configured into systems with fewer than 65 nodes to greatly simplify future expansion as more switches are added.

SP Switch2 expansion frame F/C 2034

An SP Switch2 Expansion Frame shares the 79 in./2.01m frame and covers of the pSeries 690. It has an integral SEPBU power supply, equipped with four SP Switch2 intermediate switch boards (ISB), including their interposers, but no processor nodes. This frame is required for systems configured with the Model 556 having more than five node switch boards in one switch plane; it interconnects all the node switch boards in the system.

The model frame is always the first frame in an SP system. It designates the type or *model class* of your SP system. The optional model types are either a tall frame system or a short frame system. Other frames that you connect to the model frame are known as expansion frames. The SP Switch frame is used to host switches or Intermediate Switch Boards (ISB), which are described later in this chapter. This special type of frame can host up to eight switch boards.

Since the original RS/6000 SP product was made available in 1993, there have been a number of model and frame configurations. The frame and the first node

in the frame were tied together forming a model. Each configuration was based on the frame type and the kind of node installed in the first slot. This led to an increasing number of possible prepackaged configurations as more nodes became available.

The introduction of a new tall frame in 1998 is the first attempt to simplify the way frames and the nodes inside are configured. This new frame replaces the old frames. The most noticeable difference between the new and old frame is the power supply size. Also, the new tall frame is shorter and deeper than the old tall frame. With the new offering, IBM simplified the SP frame options by telecopying the imbedded node from the frame offering. Therefore, when you order a frame, all you receive is a frame with the power supply unit(s) and a power cord. All nodes, switches, and other auxiliary equipment are ordered separately.

All new designs are completely compatible with all valid SP configurations using older equipment. Also, all new nodes can be installed in any existing SP frame provided that the required power supply upgrades have been implemented in that frame.

Note: Tall frames and short frames cannot be mixed in an SP system.

2.2.2 Tall frames

The tall model frame (model 55x) and the tall expansion frame (F/C 1550) each have eight drawers, which hold internal nodes and an optional switch board. Depending on the type of node selected, an SP tall frame can contain up to a maximum of 16 thin nodes, eight wide nodes, or four high nodes. Node types may be mixed in a system and scaled up to 128 nodes (512 by special request). There is also the Tall Frame model 555 (feature 1555) that requires one SP Switch and a minimum of two clustered RS/6000 or pSeries servers controlled by a CWS. The other new model frame 556 requires an SP Switch2 and also a minimum of two clustered RS/6000 or pSeries servers controlled by a CWS.

2.2.3 Short frames

The short model frame (model 500) and the short expansion frame (F/C 1500) each have four drawers, which hold internal nodes, and an optional switch board. Depending on the type of node selected, an SP short frame can contain up to a maximum of eight thin nodes, four wide nodes, or two high nodes. Also, node types can be mixed and scaled up to only eight nodes. Therefore, for a large configuration or high scalability, tall frames are recommended.

Only the short model frame can be equipped with a switch board. The short expansion frame cannot hold a switch board, but nodes in the expansion frame can share unused switch ports in the model frame.

Figure 2-4 illustrates short frame components from the front view.

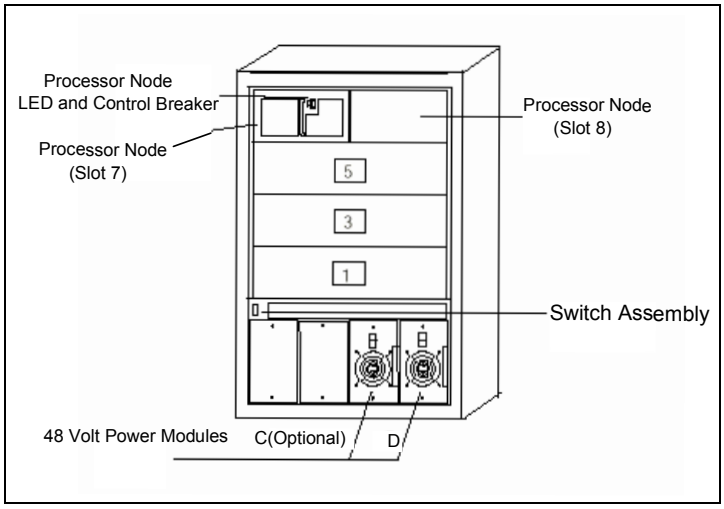


Figure 2-4 Front view of short frame components

2.2.4 SP Switch frames

The SP Switch frame is defined as a base offering tall frame equipped with either four or eight Intermediate Switch Boards (ISB). This frame does not contain processor nodes. It is used to connect model frames and switched expansion frames that have maximized the capacity of their integral switch boards. Switch frames can only be connected to data within the local SP system.

The base level SP Switch frame (F/C 2031) contains four ISBs. An SP Switch frame with four ISBs supports up to 128 nodes. The base level SP Switch frame can also be configured into systems with fewer than 65 nodes. In this environment, the SP Switch frame will greatly simplify future system growth. Figure 2-5 on page 16 shows an SP Switch frame with eight ISBs.

Note: The SP Switch frame is required when the sixth SP frame with an SP Switch board is added to the system and is a mandatory prerequisite for all large scale systems.

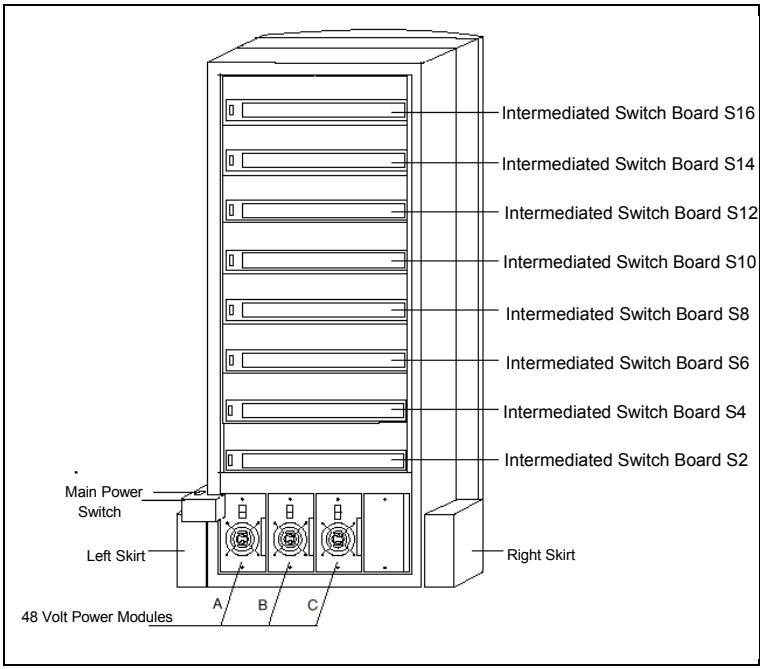


Figure 2-5 SP Switch frame with eight Intermediate Switch Boards (ISB)

2.2.5 Power supplies

Tall frames come equipped with redundant (N+1) power supplies; if one power supply fails, another takes over. Redundant power is an option with the short frames (F/C 1213). These power supplies are self-regulating units. Power units with the N+1 feature are designed for concurrent maintenance; if a power unit fails, it can be removed and repaired without interrupting the running processes on the nodes.

A tall frame has four power supplies. In a fully populated frame, the frame can operate with only three power supplies (N+1). Short frames come with one power supply, and a second, optional one, can be purchased for N+1 support.

Figure 2-6 on page 17 illustrates tall frame components from front and rear views.

The power consumption depends on the number of nodes installed in the frame. For details, refer to *RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment,* GA22-7280.

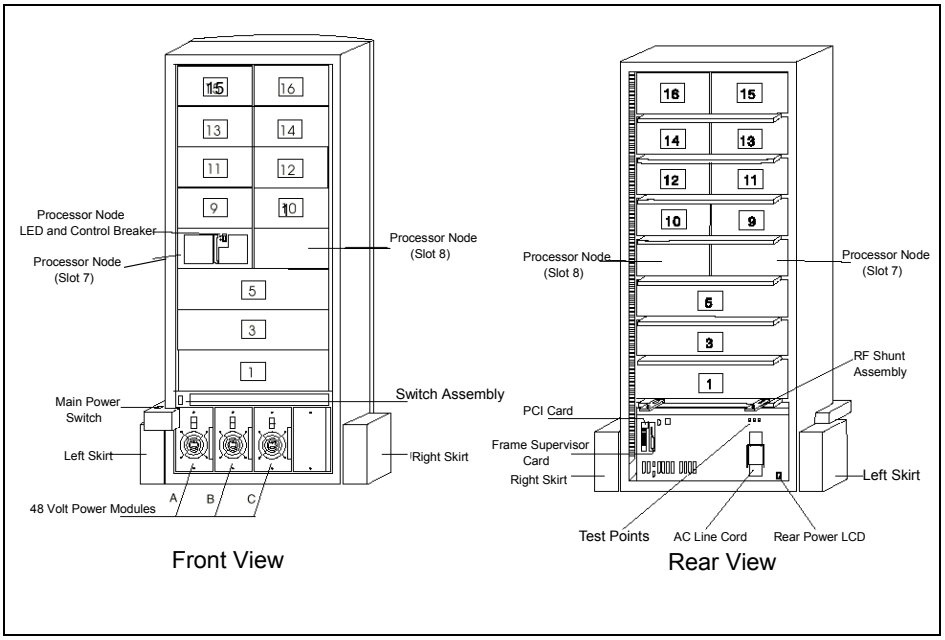


Figure 2-6 Front and rear views of tall frame components

2.2.6 Hardware control and supervision

Each frame (tall and short) has a supervisor card. This supervisor card connects to the control workstation through a serial link, as shown in Figure 2-7 on page 18.

The supervisor subsystem consists of the following components:

- Node supervisor card (one per processor node)
- Switch supervisor card (one per switch assembly)
- Internal cable (one per thin processor node or switch assembly)
- Supervisor bus card (one per thin processor node or switch assembly)
- Frame supervisor card
- Serial cable (RS-232)
- Service and Manufacturing Interface (SAMI) cable

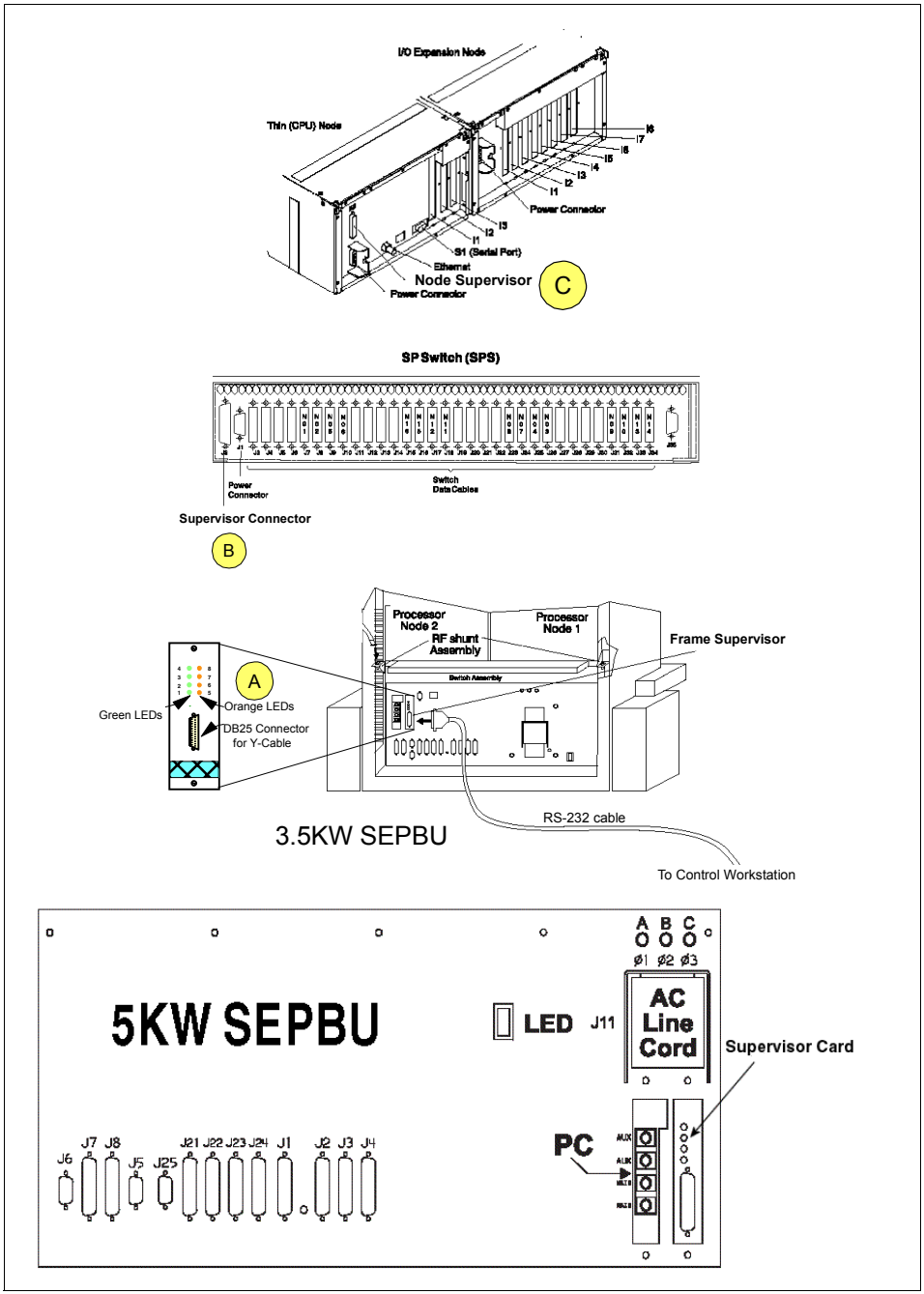


Figure 2-7 Frame supervisor attachment

A cable connects from the frame supervisor card (position A) to the switch supervisor card (position B) on the SP Switch or the SP-Switch-8 boards and to the node supervisor card (position C) of every node in the frame. Therefore, the control workstation can manage and monitor frames, switches, and all in-frame nodes.

2.3 Cluster 1600 nodes

The Cluster 1600 includes both the basic RS/6000 SP building block with standard nodes, and internal nodes and the external nodes that are known as SP-attached servers. Each node is a complete server system comprising of processors, memory, internal disk drive, expansion slots, and its own copy of the AIX operating system. The basic technology is shared with standard RS/6000 and pSeries workstations and servers, but differences exist that allow nodes to be centrally managed. There is no special version of AIX for each node. The same version runs on all RS/6000 and pSeries systems.

Standard nodes can be classified as those that are inside the RS/6000 SP frame and those that are not.

2.3.1 Internal nodes

Internal nodes can be classified, based on their physical size, as thin, wide, and high nodes. Thin nodes occupy one slot of an SP frame, while wide nodes occupy one full drawer of an SP frame. A high node occupies two full drawers (four slots).

Since 1993, when IBM announced the RS/6000 SP, there have been 17 internal node types, excluding some special on-request node types. The three most current nodes are: 375/450 MHz POWER3™ SMP thin nodes, 375/450MHz POWER3 SMP wide nodes and the 375 MHz POWER3 SMP high nodes.

Note:

With PSSP 3.5 and AIX5L, the following internal nodes are still supported:

- The 332 MHz SMP thin and wide nodes
- The 200 MHz POWER3 SMP thin and wide nodes
- The 222 MHz POWER3 SMP high node

With AIX 5.2, no more microchannel RS/6000 type nodes are supported.

375/450 MHz POWER3 SMP thin nodes

The 375 MHz POWER3-II SMP thin node can have two or four 64-bit 375 MHz processors, 256 MB to 16 GB of memory (in two memory cards), two 32-bit PCI slots, a slot for either the SP Switch MX2 adapter or the SP Switch2 MX adapter, integrated Ethernet (10/100 Mbit), 4.5 GB to 18.2 GB of mirrored internal DASD, integrated Ultra SCSI, and an external RS-232 connection with active heartbeat used only by the HACMP application. The faster 450 MHz POWER3-II processor is also available. It contains the latest POWER3-II processor technology with the benefit of the huge 8 MB L2 cache and 20% more CPU speed. The thin node can also be upgraded to a wide node.

375/450 MHz POWER3 SMP wide nodes

The 375 MHz POWER3 SMP wide node can have up to two 64-bit 375 MHz processor cards with two CPUs on each card, 256 MB to 16 GB of memory (in two memory cards), two 32-bit PCI slots and eight 64-bit PCI slots, a slot for either the SP Switch MX2 adapter or the SP Switch2 MX adapter, integrated Ethernet (10/100 Mbit), four internal disk bays 4.5 GB to 111.6 GB of mirrored internal DASD, integrated Ultra SCSI, and an external RS-232 connection with active heartbeat used only by the HACMP application.The 375/450 MHz thin and wide nodes are equivalent to the IBM RS/6000 7044-270 workstation.The faster 450 MHz POWER3-II processor is also available. It contains the latest POWER3-II processor technology with the benefit of the huge 8 MB L2 cache and 20% more CPU speed.

375 MHz POWER3 SMP high nodes

The POWER3 SMP high node consists of 4, 8, 12, or 16 CPUs (in 4 processor cards) at 375 MHz. Each node has 1 GB to 64 GB of memory (in 4 card slots), an integrated 10/100 Mbps , an integrated Ultra SCSI bus, 2 disk storage bays, 5 PCI slots (one 32-bit slot, four 64-bit slots), and 6 expansion I/O unit connections that support three SP Expansion I/O Unit loops.

Table 2-1 POWER3 nodes overview

Resource	375/450MHz thin/wide	375MHz SMP high node
Processor	2 or 4 way POWER3-II at 375 MHz or 450 MHz	4, 8, 12, 16 way at POWER3-II 375 MHz
Memory	256 MB to 16 GB	1 GB to 64 GB
Cache	8 MB L2	
PCI slots	2 (thin)/10 (wide)	5 (one 32-bit, four 64-bit)
I/O drawers/additional PCI slots	N/A	1 to 6 (each drawer has eight 64-bit slots) / 48

2.3.2 External nodes

An external node is a kind of processor node that cannot be housed in the frame due to its large size. These nodes, such as the RS/6000 7017-S70, 7026-H80, and 7040-690 are also known as SP-attached servers. Table 2-2 shows supported SP-attached server configurations.

Table 2-2 Supported SP-attached servers

SP-attached Servers	HMC-attached	Standard RS232 connection to CWS	Custom RS232 connection to CWS
M/T 7040			
IBM @server pSeries 670	Yes	Yes	No
IBM @server pSeries 690	Yes	Yes	No
M/T 7039[a]			
IBM @server pSeries 655	Yes	Yes	No
M/T 7038			
IBM @server pSeries 650	Yes	Yes	No
M/T 7028			
IBM @server pSeries 630	Yes	Yes	No
M/T 7026[b]			
IBM @server pSeries 660 models 6H1, 6H0 and 6M1	No	No	Yes
RS/6000 models M80 and H80	No	No	Yes
M/T 7017			
IBM @server pSeries 680	No	No	Yes
RS/6000 model S80, S7A and S70	No	No	Yes

a. M/T 7039 requires RS-422 connections between the HMC and the Bulk Power Controllers on the M/T 7040-W42 frame used with the 7039 server.
b. For M/T 7026 servers only, an SP Internal Attachment Adapter.

Overall, an Ethernet connection to the SP LAN may require an SP-supported card and a customer-supplied cable.

SP-attached servers

The RS/6000 SP performance can be enhanced by using SP-attached servers. These external nodes operate as nodes within the SP system and provide scalability. They excel in capacity and scalability in On-line Transaction Processing (OLTP), Server Consolidation, Supply Chain Management, and Enterprise Resource Planning (ERP), such as SAP, where single large database servers are required. The attachment to the SP system is done in several ways. For more information, refer to *IBM (e)server Cluster 1600: Planning, Installation, and Service,* GA22-7863.

Note: With PSSP 3.5 and AIX5L, the following external nodes are still supported:

- The RS/6000 Enterprise Server H80, M80, and S80
- The RS/6000 Enterprise Server S70 or S7A

2.3.3 POWER4™ technology

The POWER4 processor is a high-performance microprocessor and storage subsystem utilizing IBM's most advanced semiconductor and packaging technology. A POWER4 system logically consists of multiple POWER4 microprocessors and a POWER4 storage subsystem, interconnected to form an SMP system. Physically, there are three key components:

- The POWER4 processor chip contains two 64-bit microprocessors, a microprocessor interface controller unit, a 1.41 MB (1440 KB) level-2 (L2) cache, a level-3 (L3) cache directory, a fabric controller responsible for controlling the flow of data and controls on and off the chip, and chip/system pervasive functions.
- The L3 merged logic DRAM (MLD) chip, which contains 32 MB of L3 cache. An eight-way POWER4 SMP module will share 128 MB of L3 cache consisting of four modules each of which contains two 16 MB merged logic DRAM chips.
- The memory controller chip features one or two memory data ports, each 16 bytes wide, and connects to the L3 MLD chip on one side and to the Synchronous Memory Interface (SMI) chips on the other.

The POWER4 chip

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

Four POWER4 chips can be packaged on a single module to form an 8-way SMP. Four such modules can be interconnected to form a 32-way SMP. To accomplish this, each chip has five primary interfaces. To communicate to other POWER4 chips on the same module, there are logically four 16-byte buses. Physically, these four buses are implemented with six buses, three on and three off.

To communicate to POWER4 chips on other modules, there are two 8-byte buses, one on and one off. Each chip has its own interface to the off chip L3 across two 16-byte wide buses, one on and one off, operating at one third processor frequency. To communicate with I/O devices and other compute nodes, two 4-byte wide GX buses, one on and one off, operating at one third processor frequency, are used. Finally, each chip has its own JTAG interface to the system service processor.

The POWER4+ chip

POWER4+ is IBM's newest 64-bit microprocessor, which takes advantage of the most advanced 0.13 micron fabrication process and contains over 180 million transistors. The POWER4+ chip is available in the 1.2, 1.45, 1.5 and 1.7 GHz versions.

POWER4+ is based on POWER4 and also contains two processors, a high-bandwidth system switch, a large memory cache and I/O interface. L1, L2 caches and L2, L3 directories on the POWER4+ chip are manufactured with spare bits in their arrays that can be accessed via programmable steering logic to replace faulty bits in the respective arrays. This is analogous to the redundant bit steering employed in main store as a mechanism to avoid physical repair that is also implemented in POWER4+ systems. The steering logic is activated during processor initialization and is initiated by the Built-in System Test (BIST) at Power On time.

L3 cache redundancy is implemented at the cache line granularity level. Exceeding correctable error thresholds while running causes invocation of a dynamic L3 cache line delete function, capable of up to two deletes per cache. In the rare event of solid bit errors exceeding this quantity, the cache continues to run, but a message calling for deferred repair is issued. If the system is rebooted without such repair, the L3 cache is placed in bypass mode and the system comes up with this cache deconfigured.

M/T 7017 server overview

M/T 7017 servers

The RS/6000 7017 Enterprise Server Model S70, Model S7A, Model S80 and pSeries 680 Model S85 are packaged in two side-by-side units. The first unit is the Central Electronics Complex (CEC). The second unit is a standard 19-inch I/O tower. Up to three more I/O towers can be added to a system. Figure 2-8 on page 25 shows the RS/6000 7017 Enterprise Server scalability.

Table 2-3 M/T 7017 overview

Resource	S70	S7A	S80	S85
Processors	4, 8, 12 way 64-bit PowerPC® RS64-I at 125 MHz	4, 8, 12 way 64-bit PowerPC RS64-II at 262 MHz	6, 12, 18, 24 way 64-bit PowerPC RS64-III at 450 MHz	6, 12, 18, 24 way 64-bit PowerPC RS64-III at 450 MHz or 64-bit RS64-IV at 600 MHz
Memory	0.5 GB to 32 GB	1 GB to 32 GB	2 GB to 96 GB	2 GB to 96 GB
Cache per processor	64 KB Data instruction L1 4 MB L2	64 KB Data instruction L1 8 MB L2	128 KB Data instruction L1 8 MB L2	128 KB Data instruction L1 16 MB L2
I/O drawer	minimum 1 maximum 4			
PCI slots 32-bit/64-bit at 33 MHz bus speed	14 to 56			

Each I/O rack accommodates up to two I/O drawers (maximum four drawers per system) with additional space for storage and communication subsystems. The base I/O drawer contains:

- A high-performance 4.5 GB GB Ultra SCSI disk drive
- A CD-ROM
- A 1.44 MB 3.5-inch diskette drive
- A service processor
- Eleven available PCI slots
- Two available media bays
- Eleven available hot-swappable disk drive bays

Each additional I/O drawer contains:

- Fourteen available PCI slots (nine 32-bit and five 64-bit) providing an aggregate data throughput of 500 MB per second to the I/O hub

- Three available media bays
- Twelve available hot-swappable disk drive bays

When all four I/O drawers are installed, the 7017 contains twelve media bays, forty-eight hot-swappable disk drive bays, and fifty-six PCI slots per system.

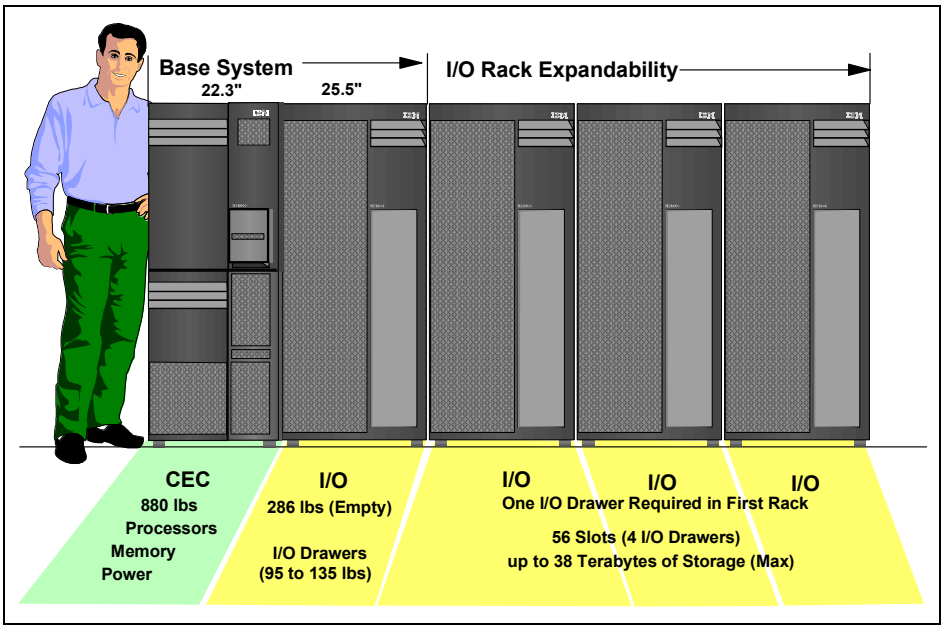


Figure 2-8 RS/6000 7017 Enterprise Server S70/S7A/S80 system scalability

M/T 7026 pSeries 660 overview

M/T 7026 pSeries 660 servers

The 7026 servers offer 64-bit performance with either RS64-III or RS64-IV processors installed. There are 1, 2, 4, 6, and 8-way configurations available with possible 450 MHz, 500 MHz, 600 MHz, 668 MHz and 750 MHz processor speeds, depending on the model. Reliability features such as chipkill, power supply redundancy, hot plug, and more, are already implemented. Memory is also expandable from 256 MB up to 64 GB. The pSeries 660 offers an innovative design that makes it ideal for mission-critical ERP/SCM, Data Warehouse/Data Mart, OLTP, and e-business applications. It blends copper silicon-on-insulator technology and larger memory capacity with a proven system architecture to give you reliability you can count on. Refer to Table 2-4 on page 26.

Table 2-4 M/T 7026 overview

Resource	7026-H80[a]	7026-M80[b]	p660 6H0[c]	p660 6H1[d]	p660 6M1[e]
Processor	1, 2, 4 or 6-way at 450 MHz or 500 MHz	2, 4, 6 or 8-way at 500 MHz	1, 2 or 4-way at 450 MHz, 600 MHz or 750 MHz	1, 2, 4-way at 450 MHz, 600 MHz or 750 MHz, 6-way at 668 MHz or 750 MHz	2, 4-way at 500 MHz or 750 MHz or 6, 8 way at 750 MHz
Memory	256 MB to 16 GB	1GB to 32GB	256 MB to 32 GB	256 MB to 32 GB	2 to 64 GB
Cache	2 MB L2 (1-way) or 4 MB L2	4MB L2	2MB L2 (1-way) or 4MB L2 (450 MHz/600 MHz) or 8 MB L2 (750 MHz)	2 MB L2 (1-way 450/600 MHz) 4MB (450/600 MHz) 8 MB (668/750 MHz)	4MB L2 (500 MHz) or 8 MB L2 (750 MHz)
PCI slots	14 to 28 (4 32-bit slots at 33 MHz bus speed and 10 64-bit slots at 66 MHz bus speed per drawer)	14 to 56 (4 32-bit slots at 33 MHz bus speed and 10 64-bit slots at 66 MHz bus speed per drawer)	14 to 28 (4 32-bit slots at 33 MHz bus speed and 10 64-bit slots at 66 MHz bus speed per drawer)	14 to 28 (4 32-bit slots at 33 MHz bus speed and 10 64-bit slots at 66 MHz bus speed per drawer)	14 to 56 (4 32-bit slots at 33 MHz bus speed and 10 64-bit slots at 66 MHz bus speed per drawer)
I/O Drawer	1 to 2	1 to 4	1 to 2	1 to 2	1 to 4
Internal disk storage	0 to 36.4 GB	0 to 36.4 GB	0 to 72.8 GB	0 to 72.8 GB	0 to 72.8 GB
Internal media	Diskette drive CD-ROM Tape drive	Diskette drive CD-ROM Tape drive	Diskette drive CD-ROM Tape drive	Diskette drive CD-ROM Tape drive	Diskette drive CD-ROM Tape drive

a. With RS64-III processor
b. With RS64-III processor
c. With 450 MHz RS64-III processor or 600/750 MHz RS64-IV processor
d. With 450 MHz RS64-III processor or 600/668/750 MHz RS64-IV processor
e. With 500 MHz RS64-III processor or 750 MHz RS64-IV processor

M/T 7028 server overview

M/T 7028 pSeries 630 server

The 7028 IBM eServer pSeries 630 Model 6C4 is a rack-mount server. The Model 6C4 provides the power, capacity, and expandability required for e-business computing. It offers 64-bit scalability via the 64-bit POWER4 or POWER4+ processor packaged as 1-way and 2-way cards. With its two-processor positions, the Model 6C4 can be configured into 1-, 2- or 4-way

configurations. The processor cards operate at 1.0 GHz with 32 MB of L3 cache per processor card, or 1.2 and 1.45 GHz with 8 MB of L3 cache per processor. Memory DIMMs are mounted on the CPU card and can contain up to 32 GB of memory.

Table 2-5 M/T 7028 p630 overview

Resource	7028 p630 6C4
Processor	1, 2 or 4 way at 1 GHz POWER4 or 1.2 GHz, 1.45 GHz POWER4+
Memory	1 to 32 GB
Cache	32 KB - 64 KB Data - instruction L1 cache 1.5 MB L2 cache 8 MB L3 cache
Maximum logical partitions (LPARs)	4
Maximum 64-bit PCI-X slots (at 133MHz bus speed)	4 with 1GHz POWER4 processor 6 with 1.2, 1.45 GHz POWER4+
Internal disk storage	18.2 GB to 587.2 GB
7311-D20 I/O drawer/additional PCI-X slots/hot swap media bays	0 to 2/7 to 14/12-24
Internal media	CD-ROM, DVD-RAM, DVD-ROM, diskette drive

M/T 7039 server overview

M/T 7039 pSeries 655 server

Using the POWER4 and POWER4+ 64-bit processors with different processor speeds, the goal was to make as many as 128 processors per frame available for High Performance computing. In a 24-inch rack, 16 nodes (each node is a thin node) can fit, with up to 8 processors in each node. Advanced Multichip Module (MCM) packaging, similar to that used in IBM zSeries®™ servers, places either four 1.3 GHz or eight 1.1 GHz POWER4, or four 1.7 GHz or eight 1.5 GHz POWER4+ processors into a package that can fit in the palm of your hand.

To further enhance performance, 128 MB of Level 3 (L3) cache are packaged with each MCM. L3 cache helps stage information more efficiently from system memory to application programs.

The additional I/O drawers are 7040-61D drawers that are connected either with RIO (500 MHz speed) or RIO-2 (1 GHz speed) loops. The maximum bandwidth can be achieved when both RIO loops are connected to one CPU drawer (CEC).

Refer to Table 2-6 for an overview of the system specifications. For a system configuration overview refer to Figure 2-9 on page 29.

Table 2-6 M/T 7039 p655 overview

Resource	7039 p655 model 651
Processor	4-way 1.7 GHz, high memory bandwidth processors 8-way 1.5 GHz, dual core implementation for greater density 4-way 1.3 GHz, high memory bandwidth processors 8-way 1.1 GHz, dual core implementation for greater density
Memory	4 GB to 32 GB[a]
Cache	32 KB/64 KB Data - instruction L1 5.6MB L2 128MB L3
Maximum logical partitions (LPARs)	4
Internal 64-bit PCI-X slots	3
I/O drawer	One 7040-61D I/O Drawer with 20 hot swap PCI slots and additional 16 Ultra-3 SCSI hot swap disk slots. The I/O contains two I/O planars with 10 PCI slots each and each I/O planar can attach to one single CEC.
Internal disk storage	18.2 GB to 146.8 GB (mirrored)

a. 64 GB of memory is available by special order on POWER4+ systems.

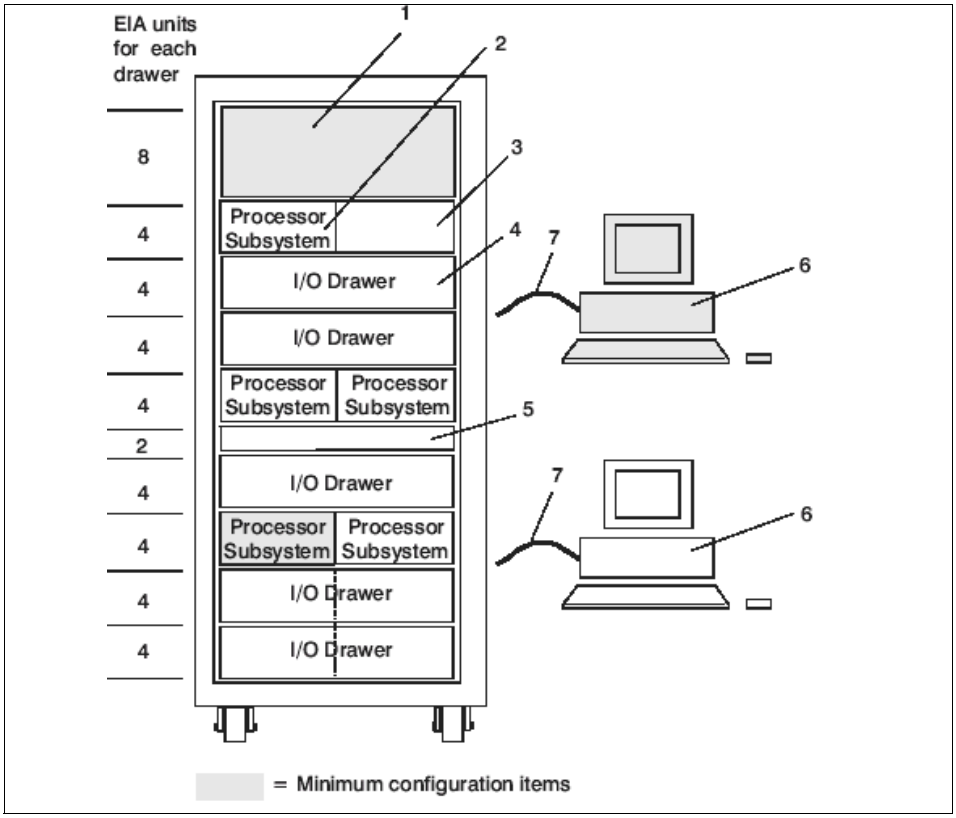


Figure 2-9 M/T 7039 system maximum configuration overview

Table 2-7 shows detailed descriptions for the components of the 7039 maximum configuration shown in Figure 2-9.

Table 2-7 Description for p655 maximum configuration

Item	What?	Description
1	7040 Model W42 Bulk Power Subsystem	N/A
2	pSeries 655 Processor Subsystem	Five or six processor subsystems maximum with five I/O drawers
3	May contain sixth pSeries 655 Processor Subsystem or may be empty	N/A

Item	What?	Description
4	7040 Model 61D I/O Subsystem	At least four I/O subsystems must be daisy-chained to processor subsystems to achieve the 5-drawer maximum.
5	Empty in this maximum I/O configuration	No IBF installed
6	Hardware Management Console	One HMC standard, one optional
7	Cables	Five RS 232 cables (one to each processor subsystem for the 5-processor configuration) Six RS 232 cables for the 6-processor configuration Two RS 422 cables from each HMC attach to each BPC. A maximum of 4 RS 422 cables per rack.

M/T 7040 servers overview

M/T 7040 pSeries 670 and 690 servers

The eServer pSeries 670 and 690 offer reliability features such as chipkill memory, power supply and fan redundancy, hot plug, dynamic CPU decollation—to name just a few. The scalability ranges from four up to 32 processors, 8 GB to 512 GB of memory, and a maximum of 160 PCI slots available. Dynamic logical partitioning (DPLAR) makes it possible to add or remove resources without interrupting the running application. Memory, processors, and adapters can be added or removed. For more information, refer to *IBM pSeries 670 and pSeries 690 System Handbook,* SG24-7040.

Table 2-8 M/T 7040 overview

Resource	pSeries 670	pSeries 690 / 690+
Processors	4, 8 or 16 64-bit POWER4 / POWER4+ at 1.1 GHz / 1.5 GHz	8,16, 24 or 32 POWER4 / POWER4+ at 1.1 GHz, 1.3 GHz / 1.5 GHz, 1.7 GHz
Memory	8 GB to 256 GB	8 GB to 512 GB
Cache	32 KB-64 KB Data instruction L1 5.7 MB/6.0 MB L2 128 MB L3	
Maximum logical partitions (LPARs)	16	32

Resource	pSeries 670	pSeries 690 / 690+
Maximum 64-bit PCI slots (133 MHz bus speed)	60	160
Internal disk storage	minimum 36.4 GB maximum 7.0 TB	minimum 36.4 GB maximum 18.7 TB
I/O drawer	minimum 1 maximum 3	minimum 1 maximum 8
Internal media	diskette drive DVD-RAM and/or CD ROM 20/40 GB 4mm tape drive	

The I/O drawers have 20 PCI/PCI-X slots available and offer a hot-plug function to add/remove/replace adapters from the drawer without powering down the machine.

pSeries firmware verification and update

pSeries firmware verification and update

All the Cluster 1600-supported pSeries servers have firmware installed that resides in the machine itself and controls some of the hardware functions. It is responsible for failure detection and much more.

You can check your firmware level on your pSeries system with the **lsmcode** command shown in Example 2-1.

Example 2-1 Check firmware level with the lsmcode command

```
root@c3pocws:/home/root> lsmcode -c
System Firmware level is SPH02066
Service Processor level is sh020307
```

The **lsmcode** command has several flags available for all kinds of reports to look at. You can use **lsmcode -A** to display all microcode levels and firmware levels of your machine. The flags are shown in Table 2-9.

Table 2-9 lsmcode flags

Flag	Description
-A	Displays microcode level information for all supported devices. Using this flag assumes the -r flag.
-c	Displays the microcode/firmware levels without using menus.
-d Name	Displays microcode level information for the named device.

Flag	Description
-r	Displays the microcode/firmware levels in a tabular format. The microcode level is preceded by a Type if supported or required.

For all systems you can go to the IBM homepage and check the latest firmware level available for your machine. It is important for some machines to be on a certain level of firmware, since enhancements such as LPAR enablement and so on depend on the installed level. Refer to this homepage for the latest available levels:

http://techsupport.services.ibm.com/server/mdownload2/download.html

The Microcode Discovery Service is the latest tool and is available on a CD. You can also download the CD ISO image and produce your own CD. This tool helps you to verify and check your systems for the latest microcode and firmware versions. To use it, you need to install the Inventory Scout first, which is automatically installed on all HMCs for pSeries.

Microcode Discovery Service gives you two ways to generate a real-time comparison report showing subsystems that may need to be updated. One possible way is to use a secure Internet connection, the other is a report generated on a system without Internet access which can be sent from another system with secure Internet access. For more details, refer to:

https://techsupport.services.ibm.com/server/aix.invscoutMDS?filename=cd.html

Cluster 1600 server attachment

The SP attached servers are not physically mounted in the existing SP frames themselves, so they are in their own racks and need be connected to the CWS. Due to the fact that the maximum allowed distance between the CWS and a attached server is 15 m, planning is very important. Depending on the SP attached node type, several connections via RS-232 and Ethernet cables have to be made. The SP system must view the SP-attached server as a frame. Therefore, the SP system views the SP-attached server as an object with both frame and node characteristics. For more detailed information, refer to Chapter 5, “Cluster 1600 installation and administration” on page 159. For an overview of a Cluster 1600, refer to Figure 2-10 on page 33.

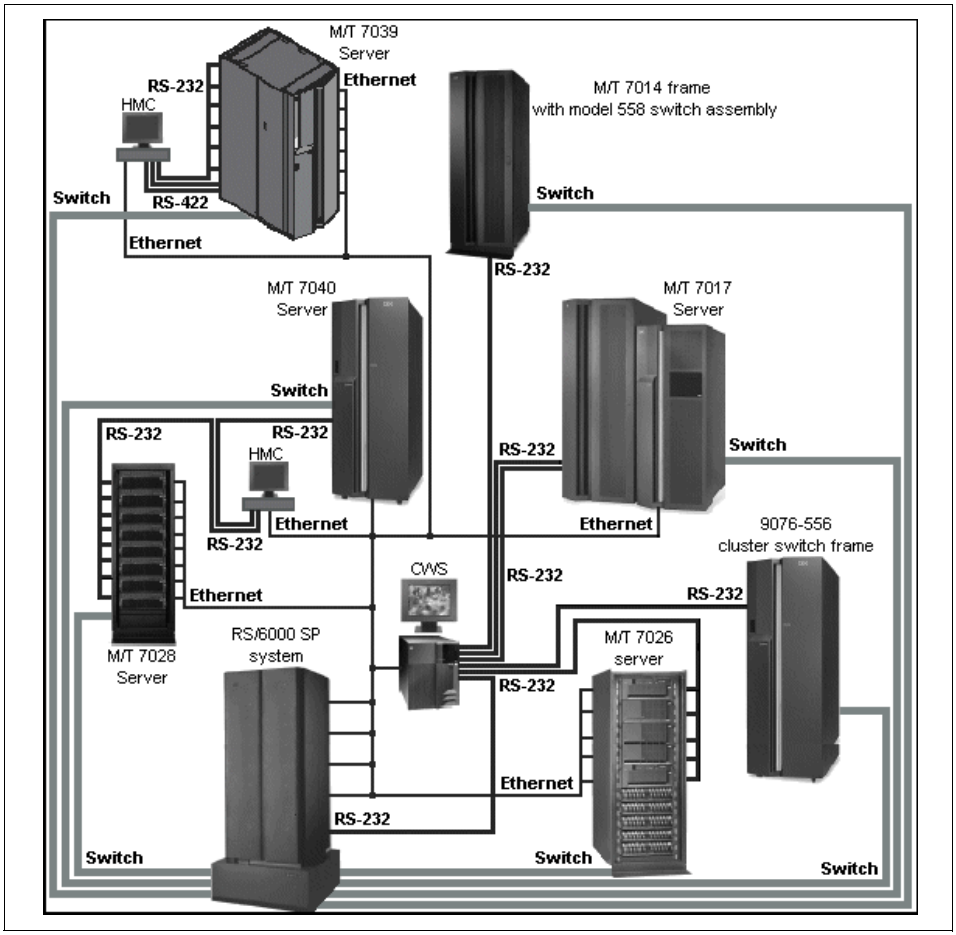


Figure 2-10 Cluster 1600 overview

2.4 Dependent nodes

Dependent nodes are non-standard nodes that extend the SP system's capabilities but cannot be used in all of the same ways as standard SP processor nodes. A dependent node depends on SP nodes for certain functions but implements much of the switch-related protocol that standard nodes use on the SP Switch. Typically, dependent nodes consist of four major components, as follows:

- A physical dependent node - The hardware device requiring SP processor node support.

- A dependent node adapter - A communication card mounted in the physical dependent node. This card provides a mechanical interface for the cable connecting the physical dependent node to the SP system.
- A logical dependent node - Made up of a valid, unused node slot and the corresponding unused SP Switch port. The physical dependent node logically occupies the empty node slot by using the corresponding SP Switch port. The switch port provides a mechanical interface for the cable connecting the SP system to the physical dependent node.
- A cable - To connect the dependent node adapter with the logical dependent node. It connects the extension node to the SP system.

2.4.1 SP Switch Router

A specific type of dependent node is the IBM 9077 SP Switch Router. The 9077 is a licensed version of the Ascend GRF (Goes Real Fast) switched IP router that has been enhanced for direct connection to the SP Switch. The SP Switch Router was known as the High Performance Gateway Node (HPGN) during the development of the adapter. These optional external devices can be used for high-speed network connections or system scaling using High Performance Parallel Interface (HIPPI) backbones or other communication subsystems, such as ATM or 10/100 Ethernet (see Figure 2-11 on page 35).

Note: The SP Switch Router (M/T9077) is only supported on an SP Switch and will only work at PSSP 3.5 if the SP Switch is used. It won't work with SP Switch2. There is also no plan to implement SP Switch2 for the GRF, since it is withdrawn from marketing already.

An SP Switch Router may have multiple logical dependent nodes, one for each dependent node adapter it contains. If an SP Switch Router contains more than one dependent node adapter, it can route data between SP systems or system partitions. For an SP Switch Router, this card is called a Switch Router Adapter (F/C 4021). Data transmission is accomplished by linking the dependent node adapters in the switch router with the logical dependent nodes located in different SP systems or system partitions.

In addition to the four major dependent node components, the SP Switch Router has a fifth optional category of components. These components are networking cards that fit into slots in the SP Switch Router. In the same way that the SP Switch Router Adapter connects the SP Switch Router directly to the SP Switch, these networking cards enable the SP Switch Router to be directly connected to an external network. The following networks can be connected to the SP Switch Router using available media cards:

- Ethernet 10/100 Base-T

- FDDI
- ATM OC-3c (single or multimode fiber)
- SONET OC-3c (single or multimode fiber)
- ATM OC-12c (single or multimode fiber)
- HIPPI
- HSSI (High Speed Serial Interface)

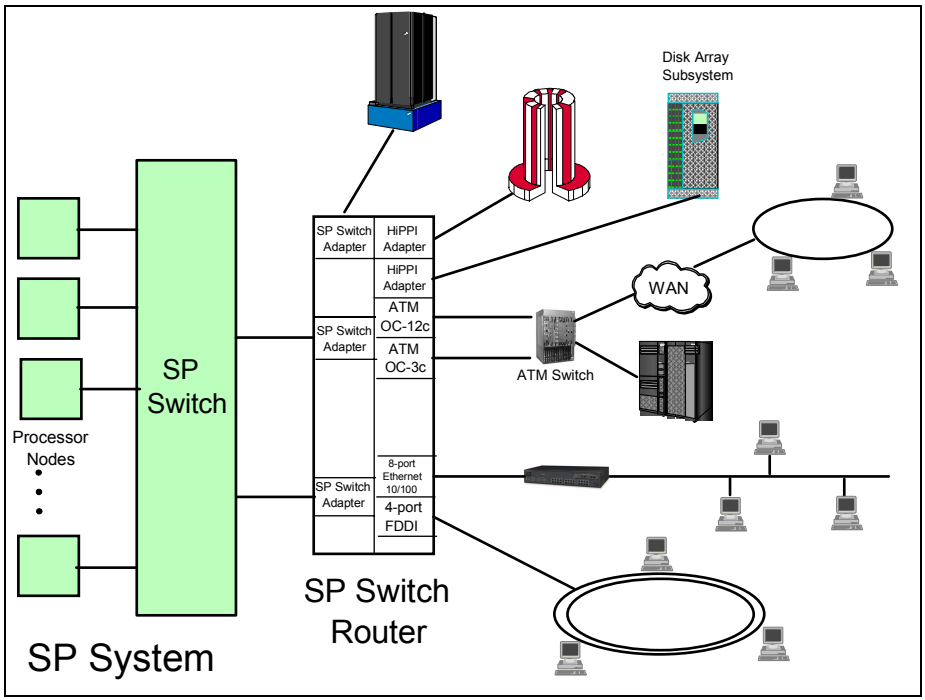


Figure 2-11 SP Switch router

Although you can equip an SP node with a variety of network adapters and use the node to make your network connections, the SP Switch Router with the Switch Router Adapter and optional network media cards offers many advantages when connecting the SP to external networks:

- Each media card contains its own IP routing engine with separate memory containing a full route table of up to 150,000 routes. Direct access provides much faster lookup times compared to software driven lookups.
- Media cards route IP packets independently at rates of 60,000 to 130,000 IP packets per second. With independent routing available from each media card, the SP Switch Router gives your SP system excellent scalability characteristics.

- The SP Switch Router has a dynamic network configuration to bypass failed network paths using standard IP protocols.
- Using multiple Switch Router Adapters in the same SP Switch Router, you can provide high performance connections between system partitions in a single SP system or between multiple SP systems.
- A single SP system can also have more than one SP Switch Router attached to it, further insuring network availability.
- Media cards are hot swappable for uninterrupted SP Switch Router operations.
- Each SP Switch Router has redundant (N+1) hot swappable power supplies.

Two versions of the RS/6000 SP Switch Router can be used with the SP Switch. The Model 04S (GRF 400) offers four media card slots, and the Model 16S (GRF 1600) offers sixteen media card slots. Except for the additional traffic capacity of the Model 16S, both units offer similar performance and network availability as shown in Figure 2-12.

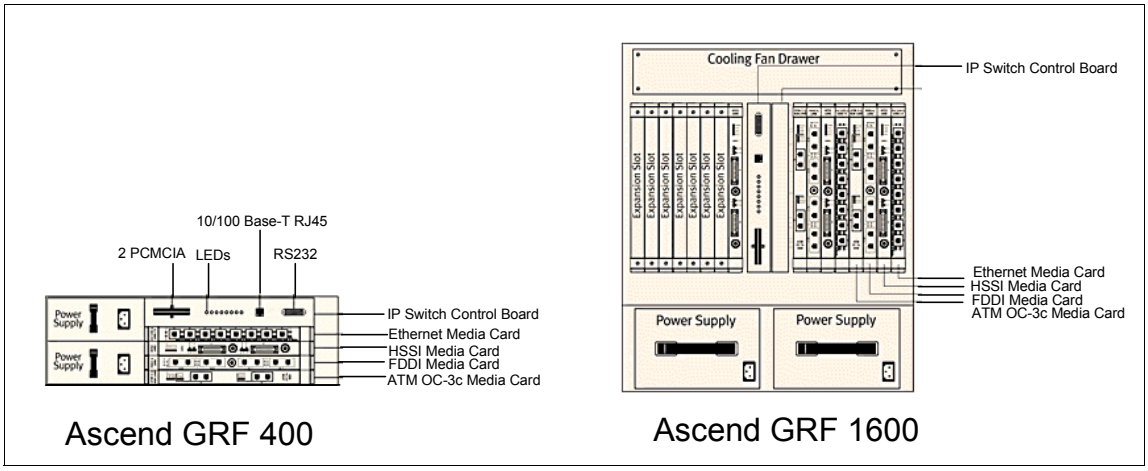


Figure 2-12 GRF models 400 and 1600

2.4.2 SP Switch Router attachment

The SP Switch Router requires a minimum of three connections with your SP system in order to establish a functional and safe network. These connections are:

1. A network connection with the control workstation - The SP Switch Router must be connected to the control workstation for system administration purposes. This connection may be either:

- A direct Ethernet connection between the SP Switch Router and the control workstation.
- An Ethernet connection from the SP Switch Router to an external network, which then connects to the control workstation.

2. A connection between an SP Switch Router Adapter and the SP Switch - The SP Switch Router transfers information into and out of the processor nodes of your SP system. The link between the SP Switch Router and the SP processor nodes is implemented by:
   - An SP Switch Router adapter
   - A switch cable connecting the SP Switch Router adapter to a valid switch port on the SP Switch
3. A frame-to-frame electrical ground - The SP Switch Router frame must be connected to the SP frame with a grounding cable. This frame-to-frame ground is required in addition to the SP Switch Router electrical ground. The purpose of the frame-to-frame ground is to maintain the SP and SP Switch Router systems at the same electrical potential.

For more detailed information, refer to *IBM 9077 SP Switch Router: Get Connected to the SP Switch*, SG24-5157.

2.5 Control workstation

The RS/6000 SP system requires an RS/6000 workstation. The control workstation serves as a central point of control with the PSSP and other optional software for managing and maintaining the RS/6000 SP frames and individual processor nodes. It connects to each frame through an RS232 line to provide hardware control functions. The control workstation connects to each external node or SP-Attached server with two custom RS232 cables, but hardware control is minimal because SP-Attached servers do not have an SP frame or SP node supervisor. A system administrator can log in to the control workstation from any other workstation on the network to perform system management, monitoring, and control tasks.

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

An optional High Availability Control Workstation (HACWS) allows a backup control workstation to be connected to an SP system. The second control

workstation provides backup when the primary workstation requires update service or fails. Planning and using the HACWS will be simpler if you configure your backup control workstation identical to the primary control workstation. Some components must be identical, others can be similar.

2.5.1 Supported control workstations

There are two basic types of control workstations:

- MCA-based control workstations
- PCI-based control workstations

Both types of control workstations must be connected to each frame through an RS-232 cable and the SP Ethernet BNC cable. These 15 m cables are supplied with each frame. Thus, the CWS must be no more than 12 m apart, leaving 3 m of cable for the vertical portion of the cable runs. If you need longer vertical runs, or if there are under-floor obstructions, you must place the CWS closer to the frame. Refer to Table 2-10 for the supported control workstations.

Supported control workstations

Table 2-10 Supported control workstations

Machine Type	Model
Currently available	
7044	170
7025	6F1
7026	6H1
7028	6C1, 6E1, 6C4, 6E4
No longer available (not supported for HMC controlled servers)	
7012 (MCA[a])	37T, 370, 375, 380, 39H, 390, 397, G30, G40
7013 (MCA)	570, 58H, 580, 59H, 590, 591, 595, J30, J40, J50
7015 (MCA)	97B, 970, 98B, 980, 990, R30, R40, R50
7024	E20, E30
7025	F30, F40, F50, F80
7026	H10, H50, H80
7030 (MCA)	3AT, 3BT, 3CT
7043	140, 240

a. MCA=Microchannel

Note: The supported machines (7043 and 7044) should only be used for regular SP systems with up to four frames. They should not be used for Cluster 1600 and 7017 machines attached to SP systems.

Refer to Figure 2-13 for a quick overview of how the attachment of the CWS is done, and some of its physical restrictions.

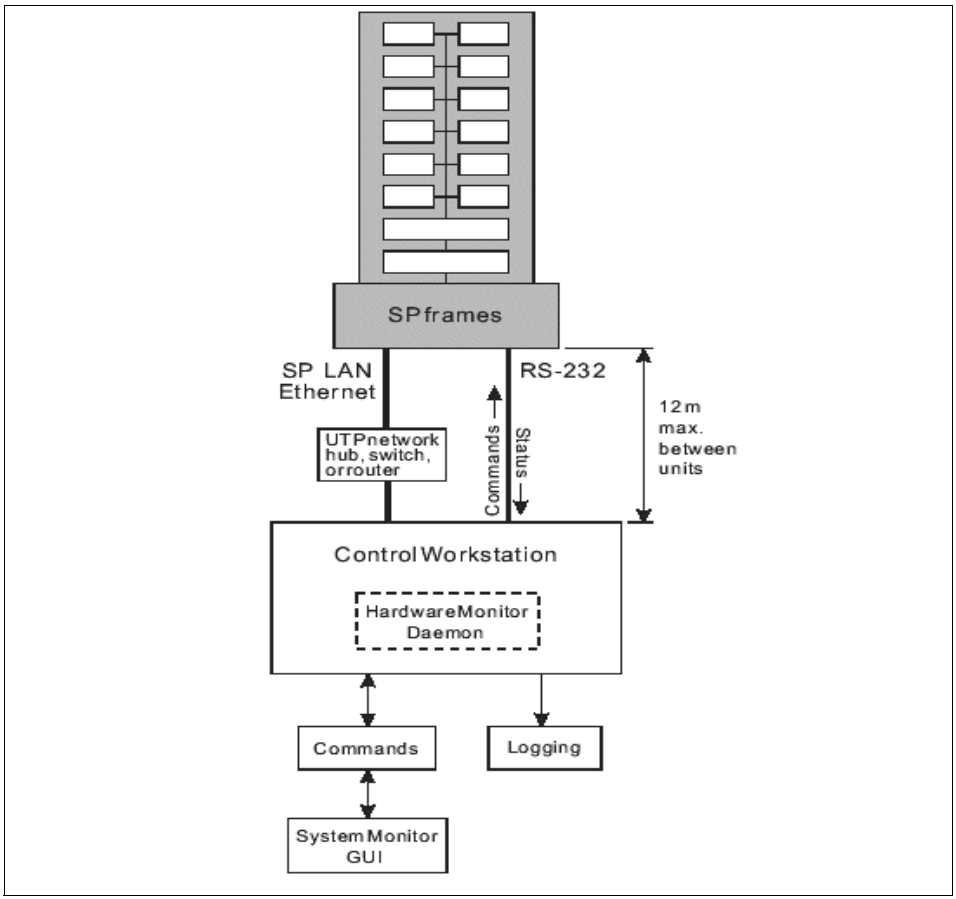


Figure 2-13 Overview of CWS attachment to SP Frame and physical restrictions

2.5.2 Control workstation minimum hardware requirements

The minimum hardware requirements for the control workstation are:

- At least 128 MB of main memory. An extra 64 MB of memory should be added for each additional system partition. For SP systems with more than 80

nodes, 256 MB is required, 512 MB of memory is suggested. For systems containing HMC-controlled servers, a minimum of 2 GB is suggested.

- At least 9 GB of disk storage. If the SP is going to use an HACWS configuration, you can configure 9 GB of disk storage in the rootvg volume group and 9 GB for the /spdata in an external volume group.
- Physically installed to within 12 meters of an RS-232 cable to each SP frame or eServer pSeries or RS/6000 server.
- With the following I/O devices and adapters:
  - A 3.5 inch diskette drive
  - A four or eight millimeter (or equivalent) tape drive
  - A SCSI CD-ROM drive
  - One RS232 port for each SP frame
  - Keyboard and mouse
  - A color graphics adapter and color monitor. An X-station model 150 and display are required if an RS/6000 that does not support a color graphics adapter is used.
  - An appropriate network adapter for your external communication network. The adapter does not have to be on the control workstation. If it is not on the control workstation, the SP Ethernet must extend to another host that is not part of the SP system. A backup control workstation does not satisfy this requirement. This additional connection is used to access the control workstation from the network when the SP nodes are down. SP Ethernet adapters are used for connection to the SP Ethernet (see 3.3.1, “The SP Ethernet admin LAN” on page 108 for details).

2.5.3 High Availability Control Workstation

The design of the SP High Availability Control Workstation (HACWS) is modeled on the High Availability Cluster Multi-Processing for RS/6000 (HACMP) licensed program product. HACWS utilizes HACMP running on two RS/6000 control workstations in a two-node rotating configuration. HACWS utilizes an external disk that is accessed non-concurrently between the two control workstations for storage of SP-related data. There is also a Y cable connected from the SP frame supervisor card to each control workstation. This HACWS configuration provides automated detection, notification, and recovery of control workstation failures. Figure 2-14 on page 41 shows a logical view of the HACWS attachment.

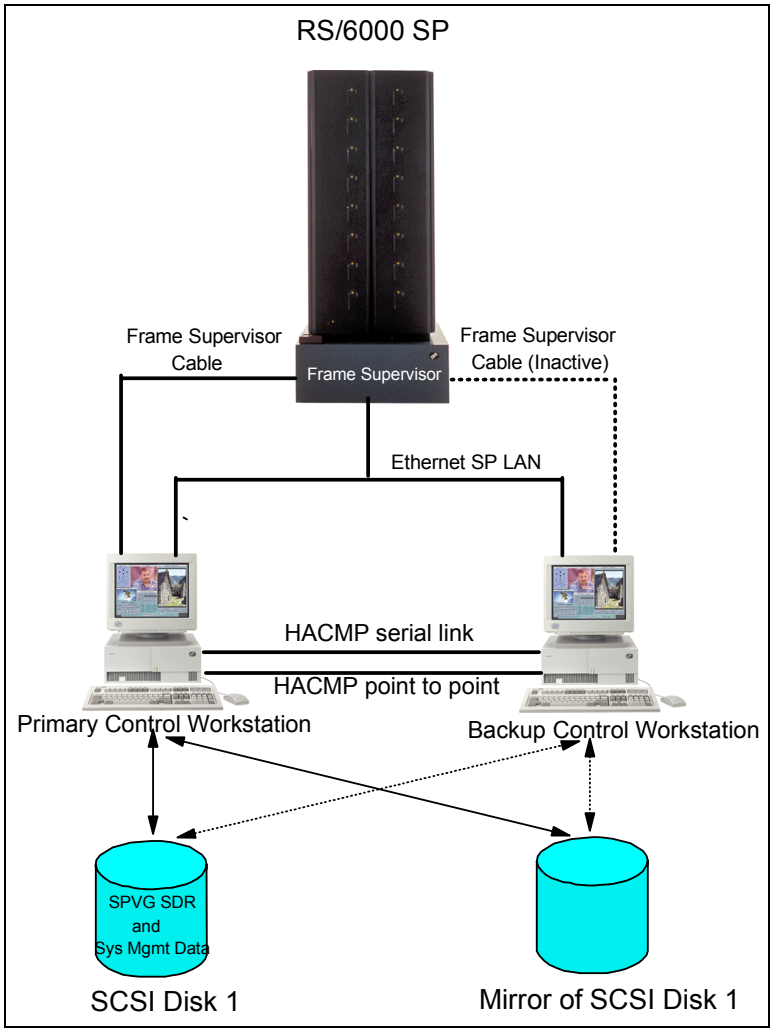


Figure 2-14 High Availability Control Workstation (HACWS) attachment

The primary and backup control workstations are also connected on a private point-to-point network and a serial TTY link or target mode SCSI. The backup control workstation assumes the IP address, IP aliases, and hardware address of the primary control workstation. This lets client applications run without changes. The client application, however, must initiate reconnects when a network connection fails.

The HACWS has the following limitations and restrictions:

- You cannot split the load across a primary and backup control workstation. Either the primary or the backup provides all the functions at one time.
- The primary and backup control workstations must each be a RS/6000. You cannot use a node in your SP as a backup control workstation.
- The backup control workstation cannot be used as the control workstation for another SP system.
- The backup control workstation cannot be a shared backup of two primary control workstations.
- There is a one-to-one relationship of primary to backup control workstations; a single primary and backup control workstation combination can be used to control only one SP system.
- If a primary control workstation is an SP authentication server, the backup control workstation must be a secondary authentication server.
- For SP-attached servers that are directly attached to the control workstation through one or two RS232 serial connections (see Table 2-2 on page 21), there is no dual RS232 hardware support as there is for SP frames. These servers can only be attached to one control workstation at a time. Therefore, when a control workstation fails, or scheduled downtime occurs, and the backup control workstation becomes active, you will lose hardware monitoring and control and serial terminal support for your SP-attached servers.

  For SP-attached servers controlled by an HMC, there is no direct serial connection between the server and the CWS. The SP-attached servers will have the SP Ethernet connection from the backup control workstation; so, PSSP components requiring this connection will still work correctly. This includes components such as the availability subsystems, user management, logging, authentication, the SDR, file collections, accounting, and others.

For more detailed information, refer to *RS/6000: Planning Volume 2*, GA22-7281.

Enablement of HACWS on HMC-controlled servers using PSSP 3.5

For cluster customers using PSSP, the existing feature called HACWS was not supported on HMC-controlled servers before PSSP 3.5. This application previously supported RS/6000 SP nodes in the cluster. It did not fully support "attached servers." HACWS is now enhanced to fully support HMC-controlled servers in an IBM Cluster 1600 configuration. For more information about HACWS, refer to *Configuring Highly Available Clusters Using HACMP 4.5,* SG24-6845.

2.6 Hardware Management Console (HMC)

When the machine types M/T 7040, M/T 7039, M/T 7038 and M/T 7028 are attached to a Cluster 1600, a Hardware Management Console (HMC) is required for the control of these machines. In a Cluster 1600, the HMC is attached to the CWS on the administrative LAN.

What is the HMC

The IBM Hardware Management Console for pSeries (HMC) provides a standard user interface for configuring and operating partitioned and SMP systems. The HMC supports the system with features that allow a system administrator to manage configuration and operation of partitions in a system, as well as to monitor the system for hardware problems. It consists of a 32-bit Intel-based desktop PC with a DVD-RAM drive.

What is the HMC doing?

- Creating and maintaining a multiple-partitioned environment
- Displaying a virtual operating system session terminal for each partition
- Displaying virtual operator panel values for each partition
- Detecting, reporting, and storing changes in hardware conditions
- Powering managed systems on and off
- Acting as a service focal point for service representatives to determine an appropriate service strategy and enable the Service Agent Call-Home capability
- Activating additional resources on demand

There is no serial RS-232 connection between the CWS and the HMC-controlled servers or between the HMC and the CWS. Only the HMC has a serial connection to the HMC ports of the servers. For M/T 7039 additional RS-422 connections are needed between the Bulk Power Controllers (BPC) and the HMC.

Since the HMC has only two integrated RS-232 ports, additional multiport adapters need to be installed for these connections. An 8-port adapter can be used. This adapter supports both RS-232 and RS-422 connections. See Figure 2-15 on page 44 for a overview of a pSeries p655 attached to a HMC.

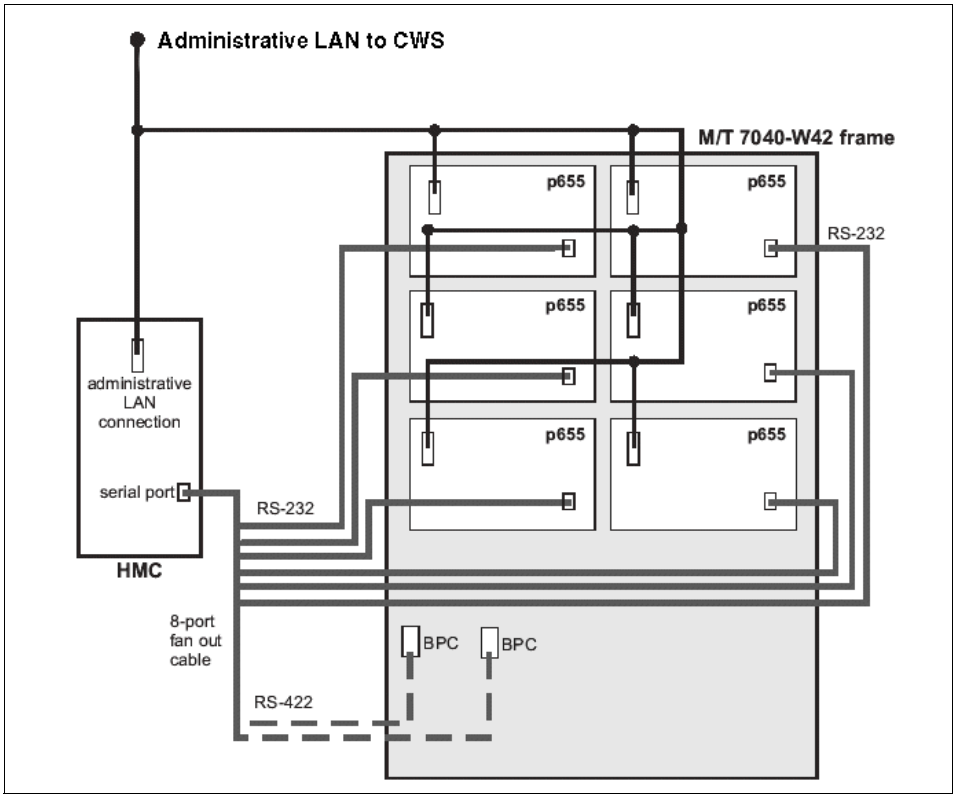


Figure 2-15 HMC attachment for p655 servers

Note: If more than six p655 servers are used, a 128-Port adapter is needed for the RS-232 communication. The 8-Port adapter will then be used for the RS-422 communication only to the BPCs in the frames.

Supported HMC models

Supported HMC models

The following models are the currently supported HMCs:

- M/T 7315-C01 (withdrawn from marketing 07/2003)
  - Has a 2.0 GHz Intel Pentium 4 processor, 1 GB memory, 40 GB hard drive, 2 integrated serial ports, one graphics port, one integrated Ethernet port, DVD-RAM drive, 4 USB ports, and 3 PCI slots.

- M/T 7315-C02
  - Has a 2.4 GHz Intel Pentium 4 processor, 1 GB memory, 40 GB hard drive, 2 integrated serial ports, one graphics port, one integrated Ethernet port, DVD-RAM drive, 6 USB ports, and 3 PCI slots.

Supported Multiport adapters

Supported Multiport adapters

These adapters are supported for the HMCs to connect to the attached servers:

- 8-port asynchronous adapter PCI BUS EIA-232/RS-422 F/C 2943
- 128-port asynchronous controller PCI bus F/C 2944
  - 2.4 MB/sec enhanced remote asynchronous node (RAN) 16-port EIA-232 F/C 8137
  - 128-port asynchronous controller cable, 4.5 m (1.2 MB/sec transfers) F/C 8131
  - 128-port asynchronous controller cable, 23 cm (1.2 MB/sec transfers) F/C 8132
  - RJ-45 to DB-25 converter cable F/C 8133
  - 1.2 MB/sec rack-mountable remote asynchronous node (RAN) 16-port EIA-232 F/C 8136
  - Asynchronous terminal/printer cable, EIA-232 (2.4 MB/sec transfers) F/C 2934
  - Serial port to serial port cable for drawer-to-drawer connections (2.4 MB/sec transfers) F/C 3124
  - F/C 3125 Serial port to serial port cable for rack-to-rack connections (2.4 MB/sec transfers) F/C 3125

HMC redundancy

The 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.

HMC functionality and software

The HMC controls the attached servers. The hardware control known in the SP through the supervisor bus is not available on the pSeries HMC-controlled servers. Therefore, the service processor is used to control, manage and collect data of the machine. The service processor stores error codes, configuration, and much more. The HMC uses the RS-232 interface to communicate with the Service Processor firmware and the hypervisor. The hypervisor firmware

provides major additions to firmware functionality. If an operating system instance in a partition requires access to hardware, it first invokes the hypervisor using hypervisor calls. The hypervisor allows privileged access to an operating system instance for dedicated hardware facilities and includes protection for those facilities in the processor. Figure 2-16 shows an overview of how the HMC communicates with its so-called managed server.

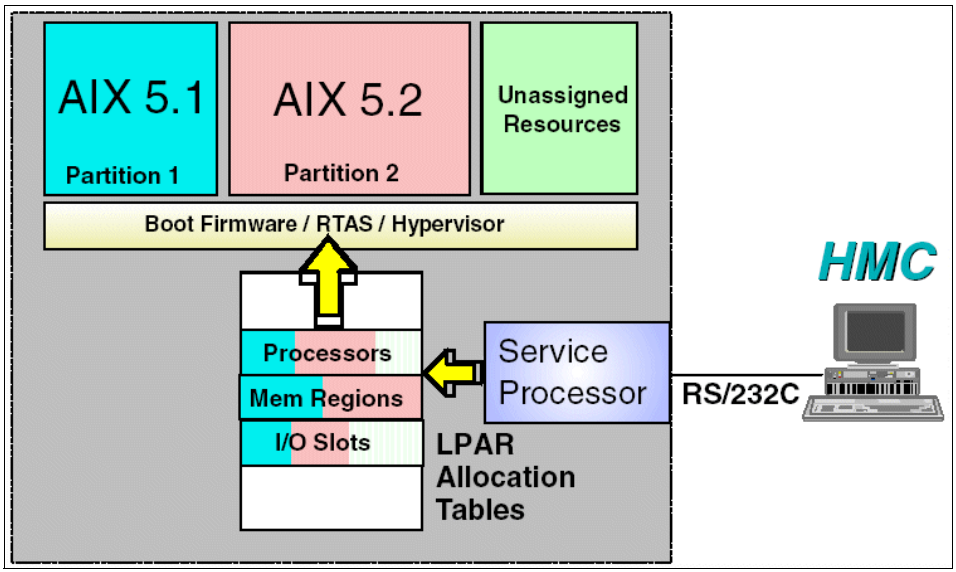


Figure 2-16 Overview of HMC control

Since it is a closed system, only IBM-approved software is allowed to run on the HMC, which runs a modified Linux operating system and system management software. The software also has integrated service features such as:

- Service Agent

  Service Agent (SA) accepts information from the Service Focal Point (SFP). It reports serviceable events and associated data collected by SFP to IBM for service automatically. The Service Agent Gateway HMC maintains the database for all the Service Agent data and events sent to IBM, including any Service Agent data from other Client HMCs. The Service Agent uses a modem connection. The modem is supplied with the machine, the phone line must be supplied by the customer.

- Service Focal Point

  Service representatives use the Service Focal Point application to start and end their service calls and provide them with event and diagnostic information. The HMC can also automatically notify service representatives of hardware failures by using a Service Agent. You can configure the HMC to

use Service Agent's call-home feature to send event information to your service representative.

- Inventory Scout

  Inventory Scout Is a tool that surveys managed systems for hardware and software information. Inventory Scout provides an automatic configuration mechanism and eliminates the need for you to manually reconfigure Inventory Scout Services. The Inventory Scout collects VPD and microcode data of the system.

For more information about the HMC, refer to *Hardware Management Console for pSeries Installation and Operations Guide,* SA38-0590.

The HMC is preloaded with the HMC code. The service representative will initially install the HMC and enable, if agreed to by the customer, the Service Agent and Service Focal Point functions. Everything else will be done by the customer. The software versions are changing from time to time and need to be updated. Every new release has improvements to the overall functionality and, of course, enhancements. See the following link for the latest software updates:

`http://techsupport.services.ibm.com/server/hmc/corrsrv.html`

The latest HMC code can either be downloaded from that homepage or can be ordered from your AIX support center on CD.

HMC vterm and s1term considerations

HMC vterm and s1term considerations

PSSP uses the HMC for 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 good practice to issue all commands, even HMC-related ones, on the CWS to guarantee a single point of control.

2.7 Cluster 1600 feature codes

The Cluster 1600 gathers all the internal and external nodes that are shown in 2.3, "Cluster 1600 nodes" on page 19. Therefore, a special machine type is available now to show that a pSeries machine is a Cluster 1600 node. This is

important for both the handling with service contracts and the support structures that need to handle the questions and possible problems that occur in a Cluster 1600 system. It makes a big difference whether a 7017-S80 is standalone or an SP-attached server in a Cluster 1600 environment. The new machine type that exists now is 9078-160. Refer to Table 2-11 for an overview of all the available feature codes (F/C).

Table 2-11 Cluster 1600 feature codes

9078 Model and feature	Description	Minimum system requirement	Maximum system requirement
9078 Model 160	IBM @server Cluster 1600	1	1
F/C 0001	M/T 7017 servers	0	16
F/C 0002	M/T 7026 servers	0	64
F/C 0003	SP Switch connections Model 555/557	0	Subject to scaling limits
F/C 0004	SP Switch2 connections Model 556/558	0	Subject to scaling limits
F/C 0005	9076 SP models	0	1
F/C 0006	9076 SP expansion frames	0	33
F/C 0007	Control Workstation	1	2 (SP HACWS only)
F/C 0008	M/T 7040 servers	0	16
F/C 0009	M/T 7040 LPARs	0	48
F/C 0010	M/T 7039 servers	0	32
F/C 0011	M/T 7039 switched LPARs	0	64
F/C 0012	M/T 7028 servers	0	64

Note: One server F/C must be ordered for each server in the cluster. These F/Cs only provide identification for the server as part of a Cluster 1600. All cluster hardware must be ordered separately under the appropriate machine type and model.

2.8 Boot/install server requirements

By default, the control workstation is the boot/install server. It is responsible for AIX and PSSP software installations to the nodes. You can also define other nodes to be a boot/install server. If you have multiple frames, the first node in each frame is selected by default as the boot/install server for all the nodes in its frame.

When you select a node to be a boot/install server, the setup_server script will copy all the necessary files to this node, and it will configure this node to be a NIM master. All mksysbs and PSSP levels served by this boot/install server node will be copied from the control workstation the first time setup_server is run against this node. The only NIM resource that is not maintained locally in this node is the lppsource. The lppsource always resides on the control workstation; so, when the lppsource NIM resource is created on boot/install servers, it only contains a pointer to the control workstation. The Sequence Power Off Timer (SPOT) is created off the lppsource contents, but it is maintained locally on every boot/install server.

Generally, you can have a boot/install server for every eight nodes. Also, you may want to consider having a boot/install server for each version of AIX and PSSP (although this is not required).

The following requirements exist for all configurations:

- Each boot/install server's en0 Ethernet adapter must be directly connected to each of the control workstation's Ethernet adapters.
- The Ethernet adapter configured as en0 must always be in the SP node's lowest hardware slot of all Ethernets.
- The NIM clients that are served by boot/install servers must be on the same subnet as the boot/install server's Ethernet adapter.
- NIM clients must have a route to the control workstation over the SP Ethernet.
- The control workstation must have a route to the NIM clients over the SP Ethernet.

Figure 2-17 on page 50 shows an example of a single frame with a boot/install server configured on node 1.

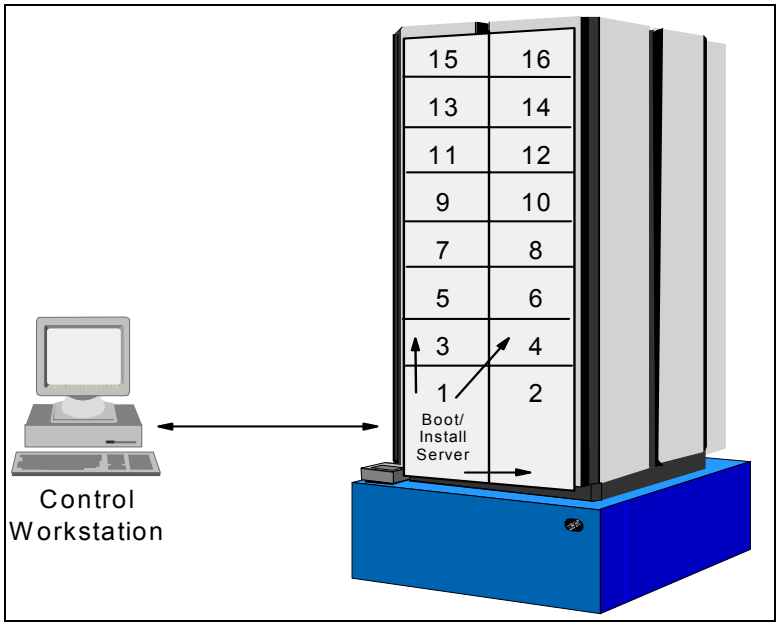


Figure 2-17 Boot/Install servers

2.9 SP Switch and SP Switch2 communication network

During the initial development of the SP system, a high-speed interconnection network was required to enable communication between the nodes that made up the SP complex. The initial requirement was to support the demands of parallel applications that utilize the distributed memory MIMD programming model.

There are two current switch types available: The older SP Switch and the newer SP Switch2, which has an even higher bandwidth and performance. The SP Switch2 also allows the use of a two-plane configuration. Therefore, a second SP Switch2 adapter needs to be installed in a node and also connected to the SP Switch2 board.

SP Switch

SP Switch

More recently, the SP Switch network has been extended to a variety of purposes:

- Primary network access for users external to the SP complex (when used with SP Switch Router)
- Used by ADSM for node backup and recovery

- Used for high-speed internal communications between various components of third-party application software (for example, SAP's R/3 suite of applications)

All of these applications are able to take advantage of the sustained and scalable performance provided by the SP Switch. The SP Switch provides the message passing network that connects all of the processors together in a way that allows them to send and receive messages simultaneously.

There are two networking topologies that can be used to connect parallel machines: Direct and indirect.

In direct networks, each switching element connects directly to a processor node. Each communication hop carries information from the switch of one processor node to another.

Indirect networks, on the other hand, are constructed such that some intermediate switch elements connect only to other switch elements. Messages sent between processor nodes traverse one or more of these intermediate switch elements to reach their destination. The advantages of the SP Switch network are:

- Bisectional bandwidth scales linearly with the number of processor nodes in the system.

  Bisectional bandwidth is the most common measure of total bandwidth for parallel machines. Consider all possible planes that divide a network into two sets with an equal number of nodes in each. Consider the peak bandwidth available for message traffic across each of these planes. The bisectional bandwidth of the network is defined as the minimum of these bandwidths.
- The network can support an arbitrarily large interconnection network while maintaining a fixed number of ports per switch.
- There are typically at least four shortest-path routes between any two processor nodes. Therefore, deadlock will not occur as long as the packet travels along any shortest-path route.
- The network allows packets that are associated with different messages to be spread across multiple paths, thus, reducing the occurrence of hot spots.

The hardware component that supports this communication network consists of two basic components: The SP Switch Adapter and the SP Switch Board. There is one SP Switch Adapter per processor node and, generally, one SP Switch Board per frame. This setup provides connections to other processor nodes. Also, the SP system allows switch boards-only frames that provide switch-to-switch connections and greatly increase scalability.

SP Switch2

SP Switch2

SP Switch2 was first introduced in 2000 for the interconnection of SP nodes via internal MX slots. As the next step in the evolution of the SP interconnection fabric, it offered significant improvements in bandwidth, latency, and RAS (Reliability, Availability, and Serviceability) over the previous generation SP Switch.

SP Switch2 is designed to be fully compatible with applications written for the older SP switches. It provides a low-latency, high-bandwidth, message-passing network that interconnects the nodes in your SP system. The N+1 feature of these switches allows for concurrent replacement of any failed power supplies or cooling fans. The supervisor can also be replaced while the switch is operating. PSSP level 3.2 software was the minimum requirement for using these switches. PSSP 3.5 still supports SP Switch2. For the different node types that can be attached to the SP Switch2 network, refer to Table 2-12 for the SP Switch2 adapter types.

Restriction:

- SP Switch2 and its adapters are not compatible with the SP Switch or the High Performance Switch series or their adapters; they cannot coexist in the same SP system.
- SP Switch2 cannot be connected to an SP Switch Router.
- SP Switch2 is not supported in 2.01 m frames.

Table 2-12 SP Switch2 adapter types

Node type	SP Switch2 adapter feature code
POWER3 thin/wide	SP Switch2 F/C 4026
POWER3 high node	SP Switch2 F/C 4025
SP-attached server with standard PCI	SP Switch2 PCI F/C 8397
SP-attached server with PCI-X	SP Switch2 PCI-X F/C 8398

2.9.1 Adapter placements for SP Switch and SP Switch2 adapters

For machines attached to SP Switch or SP Switch2 that have no dedicated switch adapter slot, such as the MX slot in the POWER3 nodes, special placement rules apply. In general, the basic requirements for an SP Switch or SP Switch2 adapter in an SP-attached server are:

- One valid, unused switch port on the switch corresponding to a legitimate node slot in your SP configuration

- Three media card slots (M/T 7017) or two media card slots (M/T 7026 and 7040) in an I/O drawer for each adapter
- Two server card slots (M/T 7028) for each adapter
- One server card slot (M/T 7039) for each adapter

The specific restrictions for the various machine types and SP Switch adapters are shown in Table 2-13.

Table 2-13 Placement restrictions for SP Switch and SP Switch2 adapters

Machine type	SP Switch	SP Switch2
M/T 7040	For SP System Attachment Adapter (F/C 8396) – Install in I/O subsystem slot 8 only (one adapter per LPAR).	For SP Switch2 Attachment Adapter (F/C 8397) **Single-plane** – Install in I/O subsystem slot 3 or 5, or both if on separate LPARs (one adapter per LPAR).
		Two-plane – Install in I/O subsystem slot 3 for css0 and slot 5 for css1 (one adapter per LPAR).
M/T 7039	N/A	For SP Switch2 PCI-X Attachment Adapter (F/C 8398) **Single-plane** – Install in server slot 1 or 3 or both if on separate LPARs (one adapter per LPAR, 2 max. per server).
		Two-plane – Install in server slot 1 for css0 and server slot 3 for css1.
M/T 7028	N/A	The adapter must be installed in slot 1 of the server. Slot 2 must remain empty.
M/T 7026	The SP System Attachment Adapter (F/C 8396) must be installed in slot 6 of the primary I/O drawer only.	**Single-plane** - The SP Switch2 Attachment Adapter (F/C 8397) must be installed in slot 5 of the server primary I/O drawer. Slot 6 must remain empty.
		Two-plane - The adapter must be installed in slot 3 of the server primary I/O drawer. Slot 4 must remain empty.

Machine type	SP Switch	SP Switch2
M/T 7017	SP System Attachment Adapter (F/C 8396) must be installed in slot 10 or the primary I/O drawer. Slots 9 and 11 must remain empty.	**Single-plane** - SP Switch2 Attachment Adapter (F/C 8397) must be installed in slot 10 of the primary I/O drawer. Slots 9 and 11 must remain empty.
		Two-plane - The adapter must be installed in slot 10 of the secondary I/O drawer. Slots 9 and 11 must remain empty.

Restriction: The following restrictions apply to M/T 7039 servers:

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

2.9.2 SP Switch hardware components

This section discusses the hardware components that make up the SP Switch network: The Switch Link, the Switch Port, the Switch Chip, the Switch Adapter, and the Switch Board. The Switch Link itself is the physical cable connecting two Switch Ports. The Switch Ports are hardware subcomponents that can reside on a Switch Adapter that is installed in a node or on a Switch Chip that is part of a Switch Board.

SP Switch Board

An SP Switch Board contains eight SP Switch Chips that provide connection points for each of the nodes to the SP Switch network as well as for each of the SP Switch Boards to the other SP Switch Boards. The SP Switch Chips each have a total of eight Switch Ports that are used for data transmission. The Switch Ports are connected to other Switch Ports through a physical Switch Link.

In summary, there are 32 external SP Switch Ports in total. Of these, 16 are available for connection to nodes, and the other 16 to other SP Switch Boards.

The SP Switch Board is mounted in the base of the SP Frame above the power supplies.

A schematic diagram of the SP Switch Board is shown in Figure 2-18.

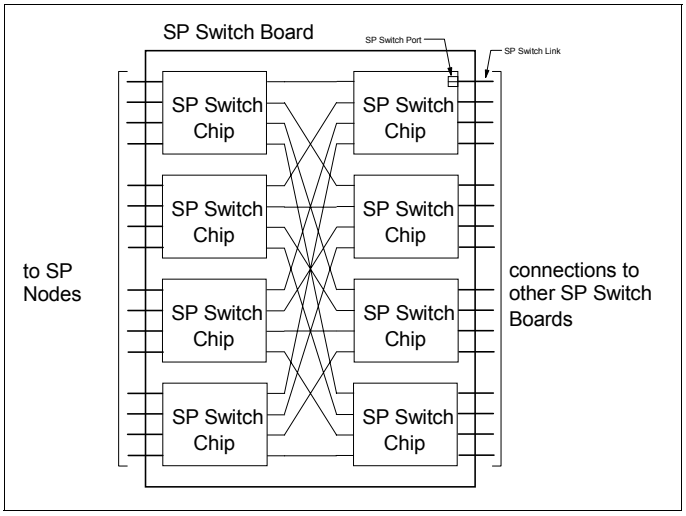


Figure 2-18 SP Switch board

SP Switch Link

An SP Switch Link connects two switch network devices. It contains two channels carrying packets in opposite directions. Each channel includes:

- Data (8 bits)
- Data Valid (1 bit)
- Token signal (1 bit)

The first two elements here are driven by the transmitting element of the link, while the last element is driven by the receiving element of the link.

SP Switch Port

An SP Switch Port is part of a network device (either the SP Adapter or SP Switch Chip) and is connected to other SP Switch Ports through the SP Switch Link. The SP Switch Port includes two ports (input and output) for full duplex communication.For SP Switch2 especially, each occupied switch port in SP Switch2 contains an interposer card. Interposer cards can be changed or added while the switch is operating. Any unused switch ports must have blank interposer cards installed. These prevent contamination of the connector and ensure proper cooling air flow.

The relationship between the SP Switch Chip Link and the SP Switch Chip Port is shown in Figure 2-19.

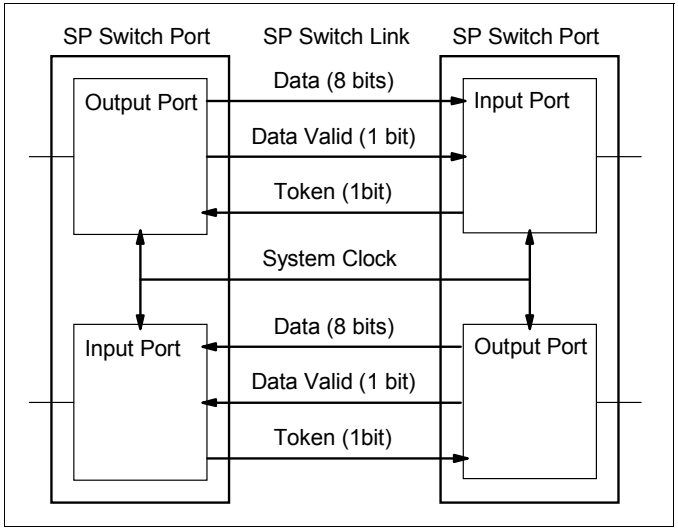


Figure 2-19 Relationship between switch chip link and switch chip port

SP Switch Chip

An SP Switch Chip contains eight SP Switch Ports, a central queue, and an unbuffered crossbar that allows packets to pass directly from receiving ports to transmitting ports. These crossbar paths allow packets to pass through the SP Switch (directly from the receivers to the transmitters) with low latency whenever there is no contention for the output port. As soon as a receiver decodes the routing information carried by an incoming packet, it asserts a crossbar request to the appropriate transmitter. If the crossbar request is not granted, the crossbar request is dropped (and, hence, the packet will go to the central queue). Each transmitter arbitrates crossbar requests on a least recently served basis. A transmitter will honor no crossbar request if it is already transmitting a packet or if it has packet chunks stored in the central queue. Minimum latency is achieved for packets that use the crossbar.

A schematic diagram of the SP Switch Chip is shown in Figure 2-20 on page 57.

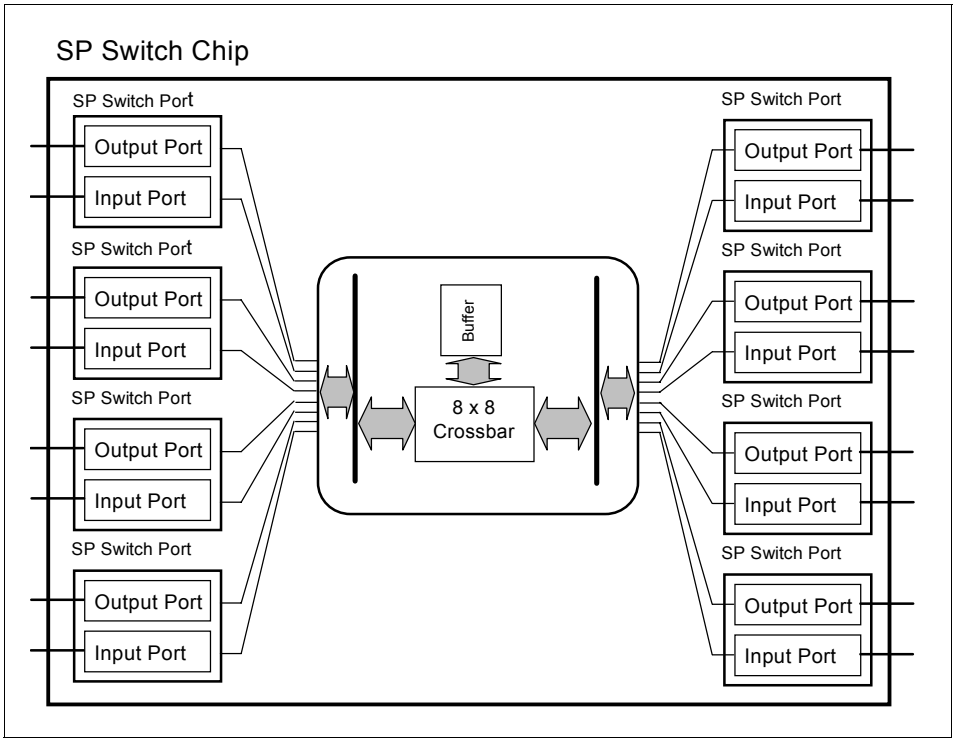


Figure 2-20 SP Switch Chip diagram

SP Switch/ SP Switch2 Adapter

Another network device that uses an SP Switch Port is the SP Switch Adapter. An SP Switch Adapter includes one SP Switch Port that is connected to an SP Switch Board, and is installed in an SP node.

Nodes based on PCI bus architecture (older 332 MHz SMP thin and wide nodes, the 375/450 MHz POWER3 SMP thin and wide nodes) must use the MX-based switch adapters (#4022 and #4023, respectively) since the switch adapters are installed on the MX bus in the node. The older SP Switch MX adapter used in the 332 MHz Nodes is withdrawn but can still be used in an SP Switch environment. The so-called mezzanine, or MX bus, allows the SP Switch Adapter to be connected directly to the processor bus providing faster performance than adapters installed on the I/O bus. The SP Switch MX2 adapter is used in the POWER3 nodes and provides better performance since the MX2 bus is faster than the older MX bus.

External nodes, such as the M/T 7017, M/T 7026, and M/T 7038, are based on standard PCI bus architecture. If these nodes are to be included as part of an SP

Switch network, then the switch adapter installed in these nodes is a PCI-based SP Switch adapter with F/C 8396. The equivalent SP Switch2 adapter is F/C 8397. For the latest PCI-based machines, such as pSeries 655, pSeries 650, pSeries 630, and pSeries 670/690, have the enhanced PCI-X adapter slots. For this PCI-X the SP Switch2 PCI-X adapter with F/C 8398 is needed.

Figure 2-21 shows a schematic diagram of the SP Switch Adapter.

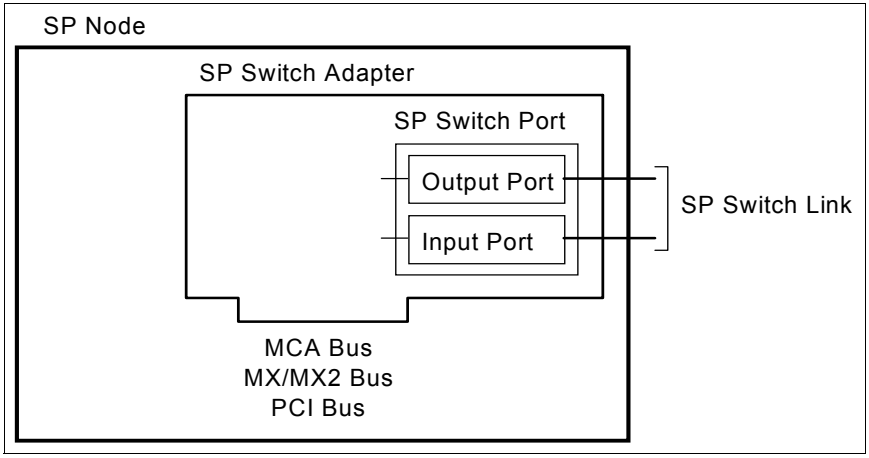


Figure 2-21 SP Switch adapter

SP Switch system

The SP Switch system in a single frame of an SP is illustrated in Figure 2-22 on page 59. In one SP frame, there are 16 nodes (maximum) equipped with SP Switch Adapters and one SP Switch Board. Sixteen node SP Switch Adapters are connected to 16 of 32 SP Switch Ports in the SP Switch Board. The remaining 16 SP Switch Ports are available for connection to other SP Switch Boards.

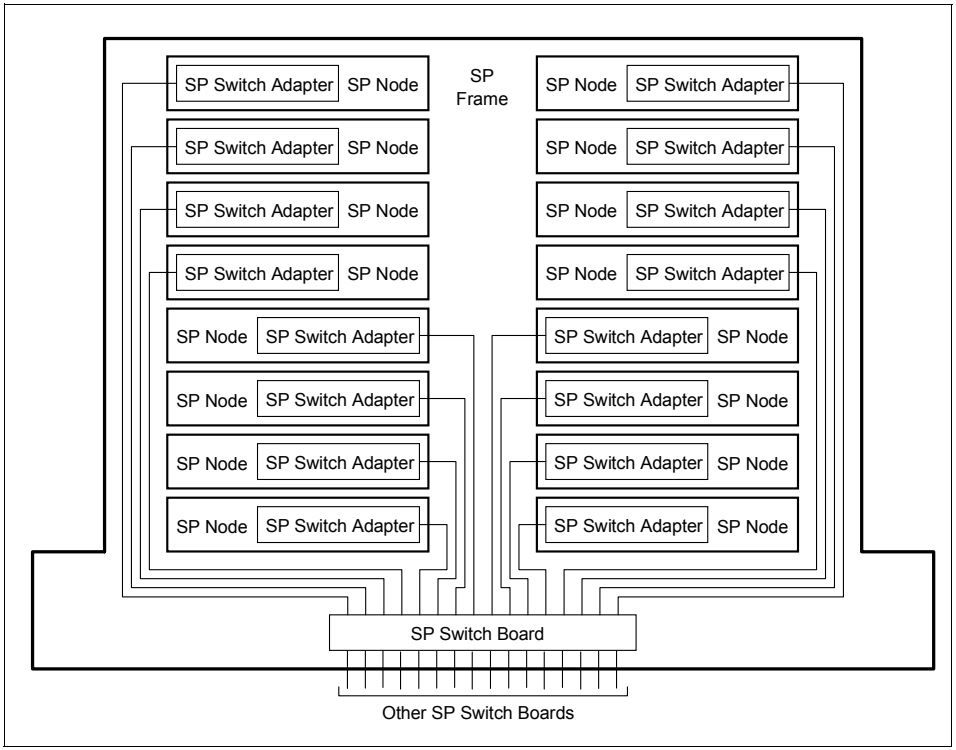


Figure 2-22 SP Switch system

2.9.3 SP Switch networking fundamentals

When considering the network topology of the SP Switch network, nodes are logically ordered into groups of 16 that are connected to one side of the SP Switch Boards. A 16-node SP system containing one SP Switch Board is schematically presented in Figure 2-23 on page 60. This SP Switch Board that connects nodes is called a Node Switch Board (NSB). Figure 2-23 on page 60 also illustrates the possible shortest-path routes for packets sent from node A to two destinations. Node A can communicate with node B by traversing a single SP Switch chip and with node C by traversing three SP Switch chips.

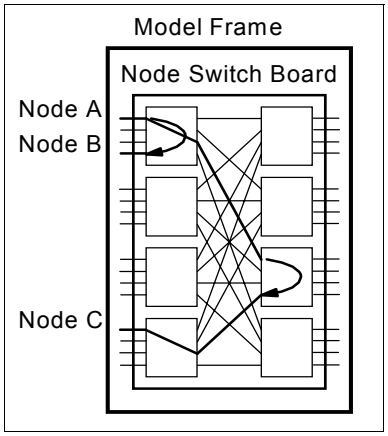


Figure 2-23 16-node SP system

The 16 unused SP Switch ports on the right side of the Node Switch Board are used for creating larger networks. There are two ways to do this:

- For an SP system containing up to 80 nodes, these SP Switch ports connect directly to the SP Switch ports on the right side of other node switch boards.
- For an SP system containing more than 80 nodes, these SP Switch ports connect to additional stages of switch boards. These additional SP Switch Boards are known as Intermediate Switch Boards (ISBs).

The strategy for building an SP system of up to 80 nodes is shown in Figure 2-24 on page 61. The direct connection (made up of 16 links) between two NSBs forms a 32-node SP system. Example routes from node A to node B, C, and D are shown. Just as for a 16-node SP system, packets traverse one or three SP Switch Chips when the source and destination pair are attached to the same Node Switch Board. When the source and destination pair are attached to different Node Switch Boards, the shortest path routes contain four SP Switch Chips. For any pair of nodes connected to separate SP Switch Boards in a 32-node SP system, there are four potential paths providing a high level of redundancy.

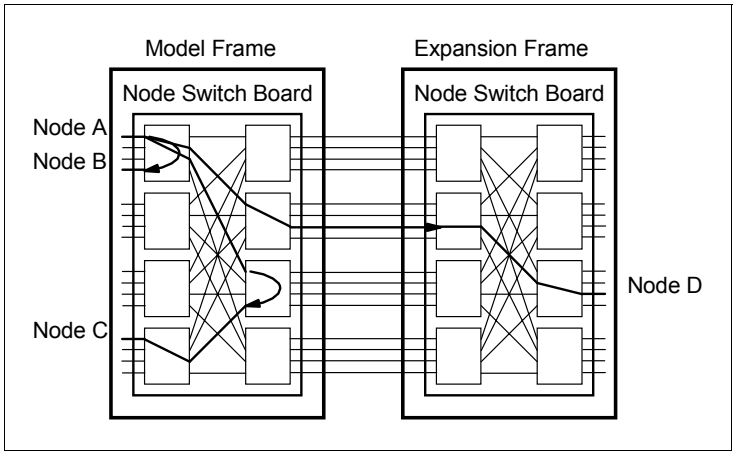


Figure 2-24 32-node SP system

If we now consider an SP system made up of three frames of thin nodes (48 nodes in total, see Figure 2-25), we observe that the number of direct connections between frames has now decreased to eight. (Note that for the sake of clarity, not all the individual connections between Switch ports of the NSBs have been shown; instead, a single point-to-point line in the diagram has been used to represent the eight real connections between frames. This simplifying representation will be used in the next few diagrams.) Even so, there are still four potential paths between any pair of nodes that are connected to separate NSBs.

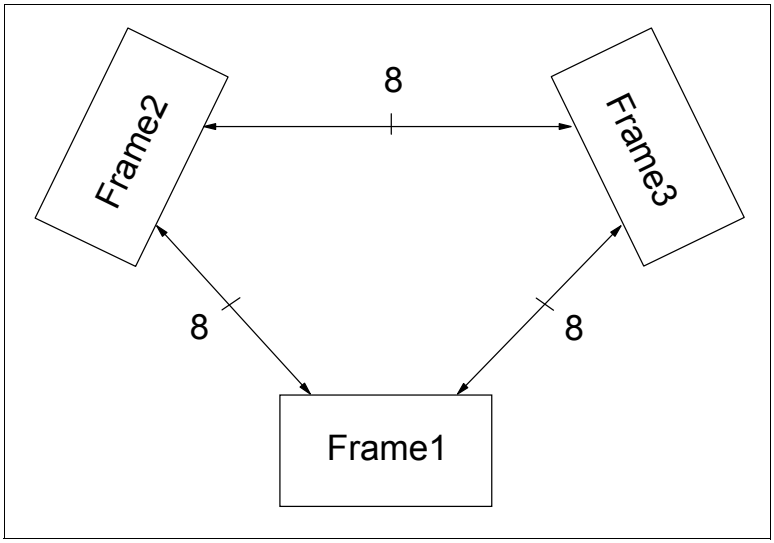


Figure 2-25 SP 48-Way system interconnection

Adding another frame to this existing SP complex further reduces the number of direct connections between frames. The 4-frame, 64-way schematic diagram is shown in Figure 2-26. Here, there are at least five connections between each frame, and note that there are six connections between Frames 1 and 2 and between Frames 3 and 4. Again, there are still four potential paths between any pair of nodes that are connected to separate NSBs.

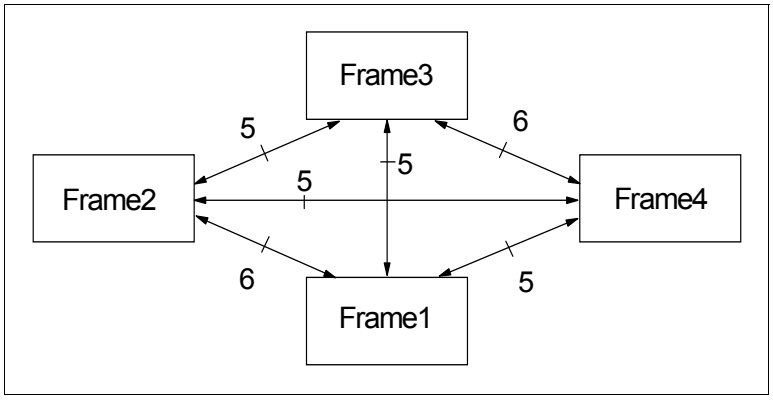


Figure 2-26 64-Way system interconnection

If we extend this 4-frame SP complex by adding another frame, the connections between frames are reduced again (see Figure 2-27 on page 63); at this level of frame expansion, there are only four connections between each pair of frames. However, there are still four potential paths between any pair of nodes that are connected to separate NSBs.

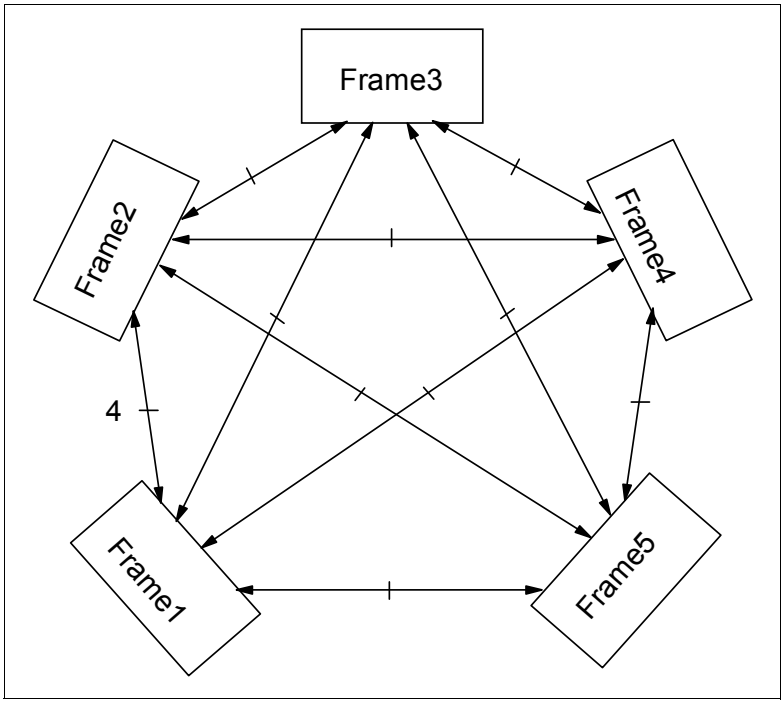


Figure 2-27 SP 80-Way system interconnection

The addition of a sixth frame to this configuration would reduce the number of direct connections between each pair of frames to below four. In this hypothetical case, each frame would have three connections to four other frames and four connections to the fifth frame for a total of 16 connections per frame. This configuration, however, would result in increased latency and reduced switch network bandwidth. Therefore, when more than 80 nodes are required for a configuration, an (ISB) frame is used to provide 16 paths between any pair of frames.

The correct representation of an SP complex made up of six frames with 96 thin nodes is shown in Figure 2-28 on page 64. Here, we see that all interframe cabling is between each frame's NSB and the switches within the ISB. This cabling arrangement provides for 16 paths between any pair of frames, thus increasing network redundancy and allowing the network bandwidth to scale linearly.

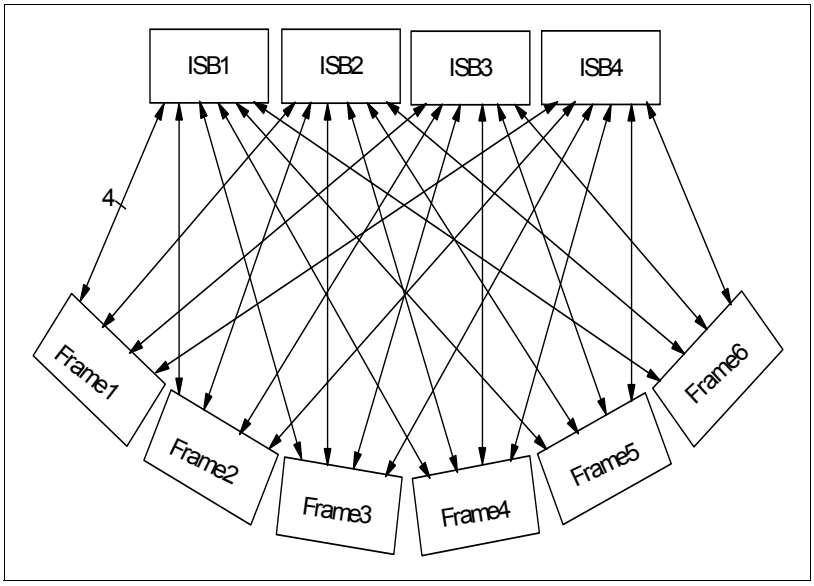


Figure 2-28 SP 96-way system interconnection

2.9.4 SP Switch network products

Since the original SP product was made available in 1993, there have been three evolutionary cycles in switch technology. The latest available switch is called the SP Switch2. SP Switch and SP Switch2 provide the base communications performance capability.

SP Switch

The operation of the SP Switch (F/C 4011) has been described in the preceding discussion. When configured in an SP order, internal cables are provided to support expansion to 16 nodes within a single frame. In multiswitch configurations, switch-to-switch cables are provided to enable the physical connectivity between separate SP Switch Boards. The required SP Switch Adapter connects each SP node to the SP Switch Board.

SP Switch-8

To meet some customer requirements, eight port switches provide a low-cost alternative to the full-size 16-port switches. The 8-port SP Switch-8 (SPS-8, F/C 4008) provides switch functions for an 8-node SP system. It is compatible with high nodes. SP Switch-8 is the only low-cost switch available for new systems.

SP Switch-8 has two active switch chip entry points. Therefore, the ability to configure system partitions is restricted with this switch. With the maximum eight nodes attached to the switch, there are two possible system configurations:

- A single partition containing all eight nodes
- Two system partitions containing four nodes each

SP Switch2

The SP Switch2 (F/C 4012) offers better performance and is described in “SP Switch2” on page 52. When configured in an SP order or Cluster 1600 configuration, internal cables are provided to support expansion to 16 nodes within a single frame. In multiswitch configurations, switch-to-switch cables are provided to enable the physical connectivity between separate SP Switch2 Boards. Each occupied switch port in SP Switch2 contains an interposer card (RS/6000 SP F/C 4032). Interposer cards can be changed or added while the switch is operating. Any unused switch ports must have blank interposer cards (RS/6000 SP F/C 9883) installed. These prevent contamination of the connector and ensure proper cooling air flow.

If a switch is configured in an SP system, an appropriate switch adapter is required to connect each SP node to the switch subsystem .Table 2-14 summarizes the switch adapter requirements for particular node types. We have also included here the switch adapter that would be installed in the SP Switch router. An overview of this dependent node, along with installation and configuration information, can be found in *IIBM 9077 SP Switch Router: Get Connected to the SP Switch,* SG24-5157.

Table 2-14 Supported switch adapters

SP Node type	Supported Switch adapter
SP Switch	
332 MHz SMP thin/wide node	F/C 4022 SP Switch MX adapter
POWER3 thin/wide & 332 MHz thin/wide node and POWER3 high node	F/C 4023 SP Switch MX2 adapter
SP Switch router (M/T 9077)	F/C 4021
SP attached server	F/C 8396 SP Switch PCI
SP Switch2	
POWER3 thin/wide	F/C 4026 SP Switch2
POWER3 high node	F/C 4025 SP Switch2
SP attached Server with standard PCI	F/C 8397 SP Switch2 PCI

SP Node type	Supported Switch adapter
SP attached Server with PCI-X[a]	F/C 8398 SP Switch2 PCI-X

a. PCI-X support on M/T 7039, M/T 7028 and M/T 7040 with RIO-2 PCI-X Backplane

Note: M/T 7028 pSeries 630 has the following PSSP requirements for SP Switch2 PCI-X:

- IY42359 PSSP V3.4 support for p630 with SP Switch2 PCI-X in RIO-2 mode IY42358 PSSP V3.5 support for p630 with SP Switch2 PCI-X in RIO-2 mode.
- The 7038-6M2 with FC 8398 requires PSSP V3.5.
- IY42352 PSSP V3.5 support for p650 with SP Switch2 PCI-X in RIO mode The 7038-6M2 with FC 8398 requires PSSP V3.4 or PSSP V3.5.
- IY42359 PSSP V3.4 support for p650 with SP Switch2 PCI-X in RIO-2 mode.
- IY42358 PSSP V3.5 support for p650 with SP Switch2 PCI-X in RIO-2 mode.

The 332 MHz and 200 MHz SMP PCI-based nodes listed here have a unique internal bus architecture that allows the SP Switch Adapters installed in these nodes to have increased performance compared with previous node types. A conceptual diagram illustrating this internal bus architecture is shown in Figure 2-29 on page 67.

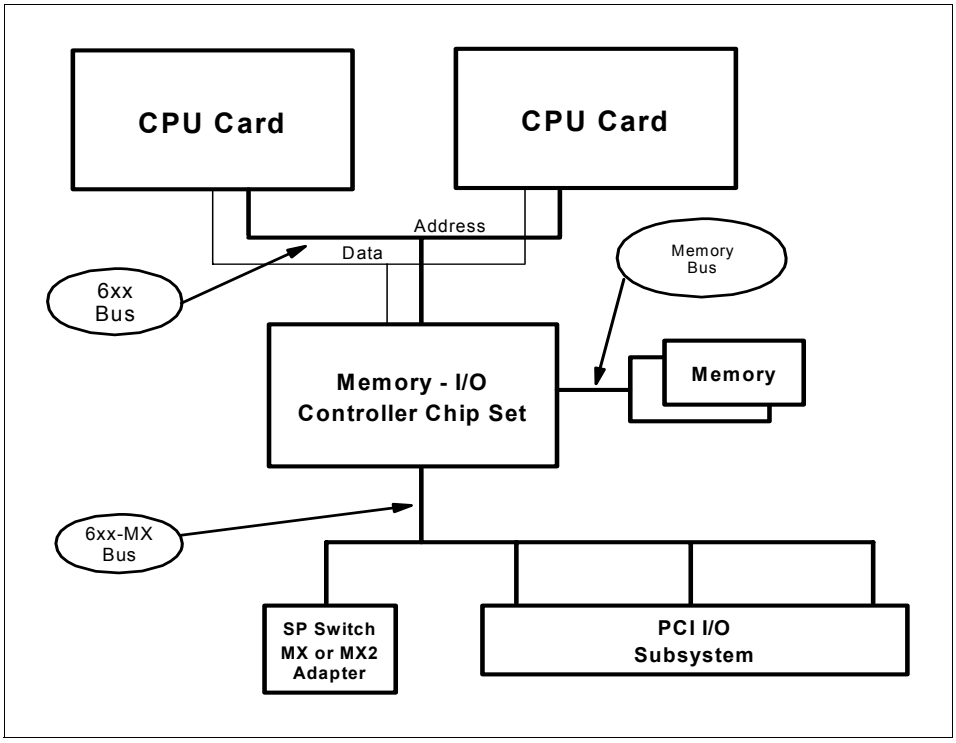


Figure 2-29 Internal Bus Architecture for PCI-based SMP nodes

These nodes implement the PowerPC MP System Bus (6xx bus). In addition, the memory-I/O controller chip set includes an independent separately clocked mezzanine bus (6xx-MX) to which 3 PCI bridge chips and the SP Switch MX or MX2 Adapter are attached. The major difference between these node types is the clocking rates for the internal buses. The SP Switch Adapters in the these nodes plug directly into the MX bus - they do not use a PCI slot. The PCI slots in these nodes are clocked at 33 MHz. In contrast, the MX bus is clocked at 50 MHz in the 332 MHz SMP nodes and at 60 MHz in the 200 MHz POWER3 SMP nodes. Thus, substantial improvements in the performance of applications using the switch can be achieved.

2.10 Peripheral devices

The attachment of peripheral devices, such as disk subsystems, tape drives, CD-ROMs, and printers, is very straightforward on the SP. There are no SP-specific peripherals; since the SP uses mainstream pSeries node technology, it simply inherits the array of peripheral devices available to the

pSeries family. The SPs shared-nothing architecture gives rise to two key concept when attaching peripherals:

1. Each node has I/O slots. Think of each node as a stand-alone machine when attaching peripherals. It can attach virtually any peripheral available to the RS/6000 family, such as SCSI and SSA disk subsystems, Magstar® tape drives, and so on. The peripheral device attachment is very flexible, as each node can have its own mix of peripherals or none at all.
2. From an overall system viewpoint, as nodes are added, I/O slots are added. Thus, the scalability of I/O device attachment is tremendous. A 512-node high node system would have several thousand I/O slots.

When you attach a disk subsystem to one node, it is not automatically visible to all the other nodes. The SP provides a number of techniques and products to allow access to a disk subsystem from other nodes.

There are some general considerations for peripheral device attachment:

- Devices, such as CD-ROMs and bootable tape drives, may be attached directly to SP nodes. Nodes must be network-installed by the CWS or a boot/install server.
- Many node types do not have serial ports. High nodes have two serial ports for general use.
- Graphics adapters for attachment of displays are not supported.

2.11 Network connectivity adapters

The required SP Ethernet LAN that connects all nodes to the control workstation is needed for system administration and should be used exclusively for that purpose. Further network connectivity is supplied by various adapters, some optional, that can provide connection to I/O devices, networks of workstations, and mainframe network. Ethernet, FDDI, Token-Ring, HIPPI, SCSI, FCS, and ATM are examples of adapters that can be used as part of an SP system.

Administrative LAN and Ethernet adapter rules

SP administrative LAN Ethernet adapter rules

For the SP internal nodes that reside in the SP frame the SP Ethernet uses the integrated Ethernet ports. It has to be en0 for the SP LAN. For some older nodes the adapter has to be in slot 1. In SP-attached servers that have LPAR capabilities the Ethernet adapter for the SP administrative LAN does not have to be en0; it can occupy any available slot.

The supported Ethernet adapters are:

- 10/100 Mbps Ethernet PCI Adapter II (F/C 4962) – Required for 7026 servers
- 10/100 Ethernet 10BASE-TX adapter (F/C 2968) (withdrawn 12/01)
- 10 MB AUI/RJ-45 Ethernet adapter (F/C 2987) (withdrawn 7/01)
- 10 MB BNC/RJ-45 Ethernet adapter (F/C 2985)

Table 2-15 shows the SP administrative LAN requirements for the Ethernet adapters.

Table 2-15 Administrative Ethernet adapter locations

SP-attached server	Location for Ethernet adapter
M/T 7017	Must be installed in slot 5 of the primary I/O drawer (en0).
M/T 7026	Must be installed in slot 1 of the primary I/O drawer (en0).
M/T 7028[a]	Native, integrated Ethernet port used. Connected to the administrative Ethernet adapter in the HMC, and also connects to the Ethernet adapter in the CWS.
M/T 7039[a]	Native, integrated Ethernet port used. Connected to the administrative Ethernet adapter in the HMC, and also connects to the Ethernet adapter in the CWS.
M/T 7040	Generally one Ethernet adapter in each LPAR and one administrative Ethernet adapter in the HMC.
M/T 7040 LPAR with SP Switch2	The LAN adapters must be placed in I/O subsystem slot 8 using the same respective LPAR as the switch adapter. Place a second LAN adapter in slot 9.
M/T 7040 LPAR with SP Switch	The LAN adapter must be placed in the same respective LPAR as the switch adapter, but does not need to be in the same I/O subsystem.

a. If the administrative LAN connections are made to the native Ethernet ports on the HMC and the p655 or p630 servers, then additional adapters are not required for those components.

For the M/T 7038 and M/T 7039 servers, the Ethernet adapters are restricted to 10/100 Mbit for the administrative Ethernet. The use of Gigabit Ethernet adapters is allowed for other external network connections. See Figure 2-30 on page 70 for an overview of the administrative LAN connections.

Restriction: If you plan to attach a pre-existing M/T 7017 or 7026 model server to your system, you must place an administrative LAN Ethernet adapter in the en0 position inside the server. This is slot 5 on 7017 models and slot 1 on 7026 models. Due to the fact that the Ethernet adapter in this slot must be configured for PSSP communications, any nonsupported Ethernet adapter that is in the en0 slot must be removed.

Additionally, if a pre-existing Ethernet adapter in slot en0 is either of F/C 2968, 2985 or 2987, that adapter must be deconfigured and then reconfigured as an administrative LAN Ethernet adapter.

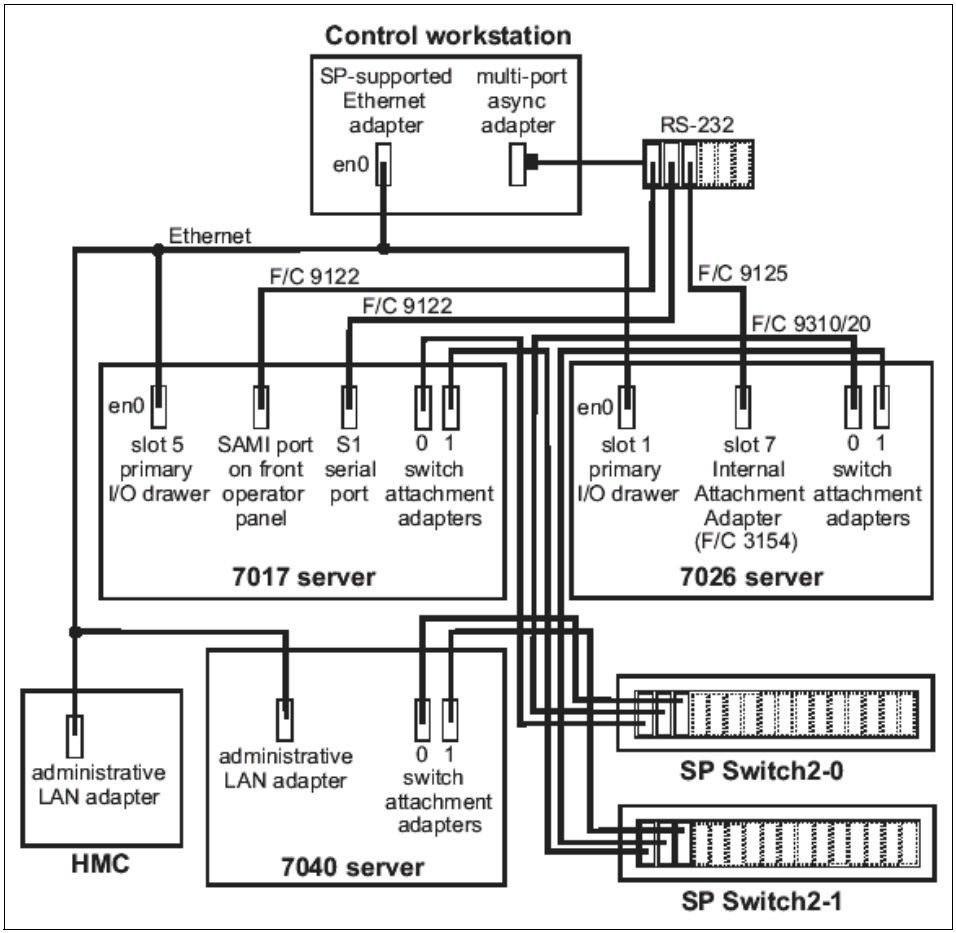


Figure 2-30 Overview of administrative LAN connections

2.12 Space requirements

You must sum the estimated sizes of all the products you plan to run. These include:

- An image comprised of the minimum AIX filesets
- Images comprised of required PSSP components
- Images of PSSP optional components and the graphical user interface (in this case, the Resource Center, PTPE, and IBM Virtual Shared Disk)

You can find more information on space requirements in 7.6.1, “AIX Automounter” on page 267.

2.13 Software requirements

The SP system software infrastructure includes:

- AIX, the base operating system
- Parallel System Support Program (PSSP)
- Other IBM system and application software products
- Independent software vendor products

With PSSP 3.5, the coexistence of several PSSP and AIX versions is supported when several specifications are met. Basically PSSP 3.5 is supported only on AIX 5L 5.1 and AIX 5L 5.2.

PSSP 3.5 supports multiple levels of AIX and PSSP in the same system partition. Only certain combinations of PSSP and AIX are supported to coexist in a system partition. Some licensed programs state that multiple levels can coexist but not interoperate. When coexistence does not include interoperability, it is explicitly stated where applicable in the subsections that follow. Refer to Table 2-16 on page 72 for supported AIX and PSSP levels in a mixed system partition.

Important: An unpartitioned system is actually a single default system partition. Coexistence is supported in the same system partition or a single default system partition (the entire Cluster 1600 system managed by PSSP).

Table 2-16 Supported AIX and PSSP levels in a mixed system partition

	PSSP 3.2	PSSP 3.4	PSSP 3.5
AIX 5L 5.2	No	No	Yes
AIX 5L 5.1	No	Yes	Yes
AIX 4.3.3	Yes	Yes	No

The nodes can run any combination of AIX and PSSP if the CWS is running PSSP 3.5 with AIX 5L 51. or AIX 5L 5.2:

- PSSP 3.5 and AIX 5L 5.2
- PSSP 3.5 and AIX 5L 5.1
- PSSP 3.4 and AIX 5L 5.1
- PSSP 3.4 and AIX 4.3.3
- PSSP 3.2 and AIX 4.3.3

The application the customer is using may require specific versions of AIX. Not all the versions of AIX run on all the nodes; so, this too must be considered when nodes are being chosen.

PSSP provides the functions required to manage an SP system as a full-function parallel system. PSSP provides a single point of control for administrative tasks and helps increase productivity by letting administrators view, monitor, and control system operations.

With PSSP 3.4 and higher the support for the SP-attached servers is enhanced and includes all the machines shown in Table 2-2 on page 21. PSSP 3.5 enables the 64-Bit PSSP use and other enhancements. You can refer to:

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

for the latest README file.

The software requirements for the internal and external nodes are very important. For the proper operation of a Cluster 1600 environment, the requirements are:

- Software requirements for M/T 7040, 7026, 7017, and 9076
  - AIX 5L 5.1 and PSSP 3.4 or AIX 5L 5.2 PSSP 3.5
  - AIX 4.3.3 and PSSP 3.4
- Software requirements for M/T 7039
  - AIX 5L for POWER V5.1 with the 5100-03 recommended maintenance package and PSSP 3.4 or PSSP 3.5

- Software requirements for M/T 7028

The p630 servers, the HMC, and their associated control workstations, require one of the following software levels when used as part of a Cluster 1600:

  - AIX 5L 5.1 and PSSP 3.4 or PSSP 3.5
  - AIX 5L 5.2 and PSSP 3.5

Note: Each Cluster 1600 system server requires its own PSSP license. PSSP is available as a CD.

2.14 System partitioning with the SP Switch

In a switched SP, the switch chip is the basic building block of a system partition. If a switch chip is placed in the system partition, then any nodes connected to that chip's node switch ports are members of that partition. Any system partition in a switched SP is comprised physically of the switch chip, any nodes attached to ports on those chips, and links that join those nodes and chips.

A system partition can be no smaller than a switch chip and the nodes attached to it, and those nodes would occupy some number of slots in the frame. The location of the nodes in the frame and their connection to the chips is a major consideration if you are planning on implementing system partitioning.

Note: Systems with SP Switch2 or clustered servers cannot be partitioned, only SP Switch allows partitioning.

Switch chips connect alternating pairs of slots in the frame. Switch boundaries are:

- Nodes 1, 2, 5, 6
- Nodes 3, 4, 7, 8
- Nodes 9, 10, 13, 14
- Nodes 11, 12, 15, 16

For a single frame system with 16 slots, the possible systems partitioning the number of slots per partition are:

- One system partition: 16
- Two system partitions: 12-4 or 8-8

- Three system partitions: 4-4-8
- Four system partitions: 4-4-4-4

System partitioning is shown in Figure 2-31.

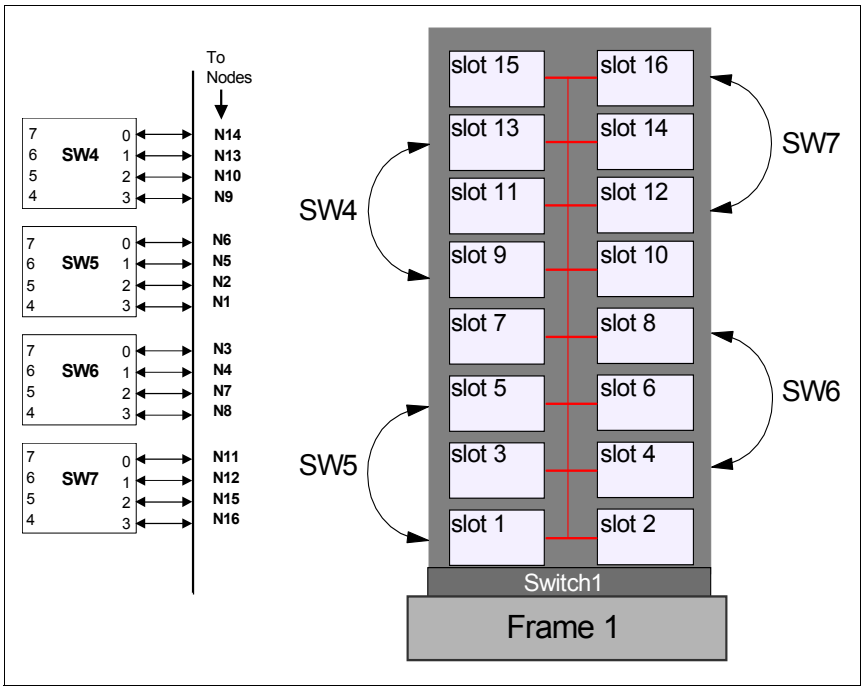


Figure 2-31 System partitioning

2.15 Cluster 1600 configuration rules

The eServer Cluster 1600 system has extremely wide scalability. For standard configuration, the Cluster 1600 system can consist of up to 128 AIX operating system images or logical nodes. This section provides you with information on how you can expand your system and what kind of configuration fits your requirement. We also provide a set of rules and sample configurations to introduce you to the design of more complex Cluster 1600 configurations. You may use these configuration rules as a checklist when you configure your system.

There are several variables that determine the total number of logical nodes in a cluster managed by PSSP. These are:

- The type of servers installed

- Whether or not the system contains a switch
- The type of switch used
- Whether or not a server is divided into LPARs

Internal Nodes

SP internal nodes are:

- 332 MHz SMP thin node (F/C 2050)
- 332 MHz SMP wide node (F/C 2051)
- 200 MHzPOWER3 SMP thin node (F/C 2052)
- 200 MHzPOWER3 SMP wide node (F/C 2053)
- 375/450 MHz POWER3 SMP thin node (F/C 2056)
- 375/450 MHz POWER3 SMP wide node (F/C 2057)
- 222 MHz POWER3 SMP high node (F/C 2054)
- 375 POWER3 SMP high node (F/C 2058)

External nodes (switched system feature codes)

The switched system feature codes for external nodes are:

- SP-attached node M/T 7017, two RS-232 cables (F/C 9122).
- SP attached node M/T 7026 CSP RS-232 cable (F/C 9125 and F/C 3154)
- Specify Code – M/T 7040 switched LPAR, M/T 7039 switched LPAR, M/T 7039 server, or M/T 7028 server (F/C 9126)

External nodes (switchless system feature codes)

The switchless system feature codes for external nodes are:

- One RS-232 cable to S1 port M/T 7017 only (F/C 3150)
- CSP/SAMI RS-232 cable (F/C 3151)
- Internal Attachment Adapter (F/C 3154)

Frames

The available frames are:

- Short model frame (model 500)
- Tall model frame (model 550)
- Short expansion frame (F/C 1500)
- Tall expansion frame (F/C 1550)

- SP Switch frame (F/C 2031)
- SP-attached server frame (F/C 9123)

Switches

The available switches are:

- SP Switch-8 (8-port switch, F/C 4008)
- SP Switch (16-port switch, F/C 4011)
- SP Switch2 (16-port switch, F/C 4012)

Switch adapter

The available switch adapters for external and internal nodes are:

- SP Switch Adapter (F/C 4020)
- SP Switch MX adapter (F/C 4022)
- SP Switch MX2 adapter (F/C 4023)
- SP System attachment adapter (F/C 8396)
- SP Switch2 Adapter POWER3 high Node (F/C 4025)
- SP Switch2 Adapter thin/wide nodes (F/C 4026)
- SP Switch2 PCI adapter (F/C 8397)
- SP Switch2 PCI-X adapter (M/T 7039 only, F/C 8398)

The SP configurations are very flexible. Several types of processor nodes can be intermixed within a frame. However, there are some basic configuration rules that come into place.

Configuration Rule 1: The tall frames and short frames cannot be mixed within an SP system.

All frames in an SP configuration must either be tall frames or short frames but not a mixture. An SP Switch frame is classified as a tall frame. You can use an SP Switch frame with tall frame configurations.

Configuration Rule 2: If there is a single PCI thin node in a drawer, it must be installed in the odd slot position (left side of the drawer).

With the announcement of the POWER3 SMP nodes in 1999, a single PCI thin node is allowed to be mounted in a drawer. In this circumstance, it must be installed in the odd slot position (left side). This is because the lower slot number

is what counts when a drawer is not fully populated. Moreover, different PCI thin nodes are allowed to be mounted in the same drawer, such as you can install a POWER3 SMP thin node in the left side of a drawer and a 332 MHz thin node in the right side of the same drawer.

Based on the configuration rule 1, the rest of this section is separated into two major parts. The first part provides the configuration rule for using short frames, and the second part provides the rules for using tall frames.

2.15.1 Short frame configurations

Short frames can be developed into two kinds of configurations: non-switched and switched. The supported switch for short frame configurations is SP Switch-8. Only one to eight internal nodes can be mounted in short frame configurations. The SP-attached servers are not supported in short frame configurations. Add the following to configuration rule 2: A single PCI thin node must be the last node in a short frame.

Configuration Rule 3: A short model frame must be completely full before a short expansion frame can mount nodes. You are not allowed any imbedded empty drawers.

Non-switched short frame configurations

This configuration does not have a switch and mounts one to eight nodes. A minimum configuration is formed by one short model frame and one PCI thin node, or one wide node, or one high node, or one pair of MCA thin nodes, as shown in Figure 2-32.

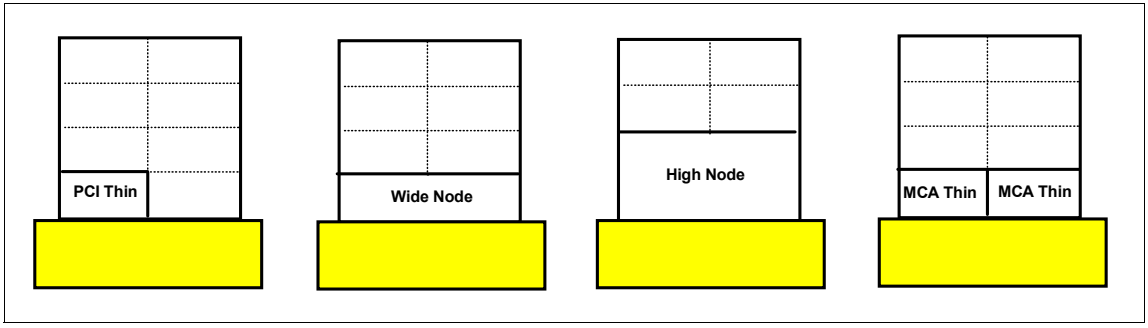


Figure 2-32 Minimum non-switched short frame configurations

The short model frame must be completely full before the short expansion frame can mount nodes, as shown in Figure 2-33.

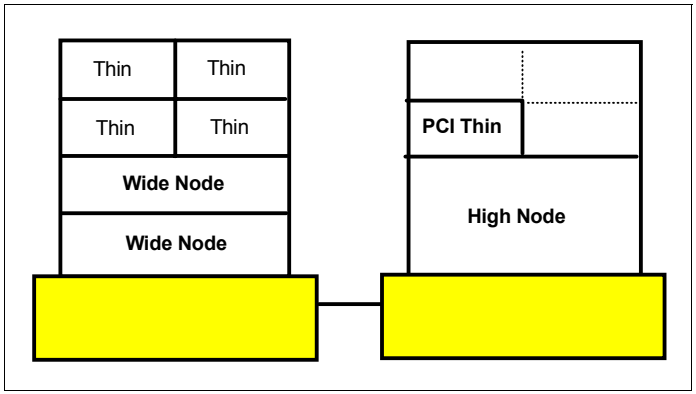


Figure 2-33 Example of non-switched short frame configuration

SP Switch-8 short frame configurations

This configuration mounts one to eight nodes and connects through a single SP Switch-8. These nodes are mounted in one required short model frame containing SP Switch-8 and additional non-switched short expansion frames. Each node requires supported SP Switch adapters. Nodes in the non-switched short expansion frames share unused switch ports in the short model frame. Figure 2-34 on page 79 shows an example of a maximum SP Switch-8 short frame configuration.

Configuration Rule 4: A short frame supports only a single SP Switch-8 board.

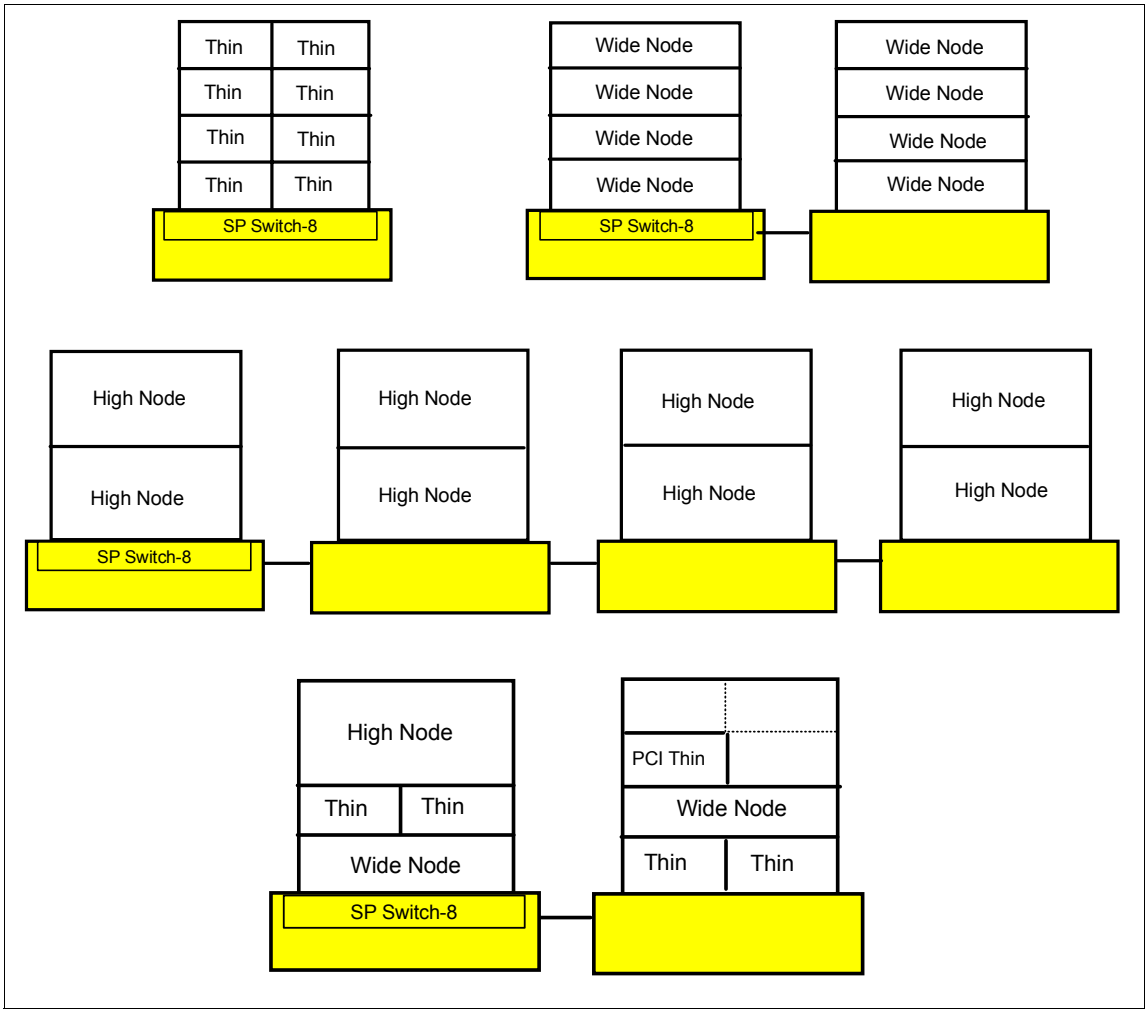


Figure 2-34 Maximum SP Switch-8 short frame configurations

2.15.2 Tall frame configurations

The tall frame offers several configurations, and it is more flexible than the short frame. The SP-attached servers are supported in tall frame configurations. There are four kinds of tall frame configurations, based on the switch type:

1. Non-switched
2. SP Switch-8
3. Single stage SP Switch
4. Two-stage SP Switch

Configuration Rule 5: Tall frames support SP-attached servers.

Non-switched tall frame configuration

This configuration does not have a switch. A minimum configuration is formed by one tall model frame and a single PCI thin node, or one wide node. In contrast to the short frame configuration, the tall expansion frame can mount nodes even when the model frame has some empty drawers. It provides more flexibility in adding more nodes in the future.

SP Switch-8 tall frame configurations

This configuration mounts one to eight nodes and connects through a single SP Switch-8. A minimum configuration is formed by one tall model frame equipped with an SP-Switch-8 and single PCI thin node, or one wide node, or one high node, or one pair of MCA thin nodes. Each node requires a supported SP Switch adapter. A non-switched tall expansion frame may be added, and nodes in a expansion frame share unused switch ports in the model frame. You are not allowed any imbedded empty drawers. Again, if there is a single PCI thin node in a drawer, it must be placed at the last node in a frame. Figure 2-35 shows example of SP Switch-8 tall frame configurations.

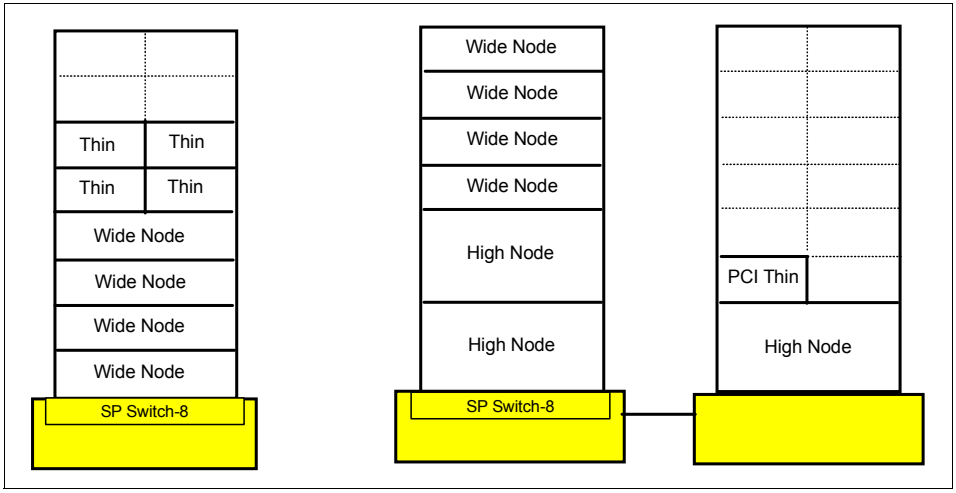


Figure 2-35 Example of SP Switch-8 tall frame configurations

Single-stage SP Switch configuration

This is probably the most common SP configuration. It provides both scalability and flexibility. This configuration can mount one to eighty processor nodes in one required tall model frame with an SP Switch and additional switched and/or non-switched expansion frames. A minimum configuration is formed by one tall

model frame equipped with an SP Switch and single PCI thin node, or one wide node, or one high node, or one pair of MCA thin nodes. Each node requires a supported SP Switch adapter. Empty drawers are allowed in this configuration.

Single-stage SP Switch with single SP Switch configuration

If your SP system has no more than 16 nodes, a single SP Switch is enough. In this circumstance, non-switched expansion may be added depending on the number of nodes and node locations (see 2.16.4, “The switch port numbering rule” on page 92 and Figure 2-43 on page 94).

Figure 2-36 on page 82 shows an example of a single-stage SP Switch configuration with no more than 16 nodes. In configuration (1), four wide nodes and eight thin nodes are mounted in a tall model frame equipped with an SP Switch. There are four available switch ports that you can use to attach SP-Attached servers or SP Switch routers. Expansion frames are not supported in this configuration because there are thin nodes on the right side of the model frame.

Configuration Rule 6: If a model frame on a switched expansion frame has thin nodes on the right side, it cannot support non-switched expansion frames.

In configuration (2), six wide nodes and two PCI thin nodes are mounted in a tall model frame equipped with an SP Switch. There are also a high node, two wide nodes, and four PCI thin nodes mounted in a non-switched expansion frame. Note that all PCI thin nodes on the model frame must be placed on the left side to comply with configuration rule 6. All thin nodes on an expansion frame are also placed on the left side to comply with the switch port numbering rule. There is one available switch port that you can use to attach SP-attached servers or SP Switch routers.

In configuration (3), there are eight wide nodes mounted in a tall model frame equipped with an SP Switch and four high nodes mounted in a non-switched expansion frame (frame 2). The second non-switched expansion frame (frame 3) is housed in a high node, two wide nodes, and one PCI thin node. This configuration occupies all 16 switch ports in the model fame. Note that wide nodes and PCI thin nodes in frame 3 have to be placed on high node locations.

Now let’s try to describe configuration (4). If you want to add two POWER3 thin nodes, what would be the locations?

A maximum of three non-switched expansion frames can be attached to each model frame and switched expansion frame.

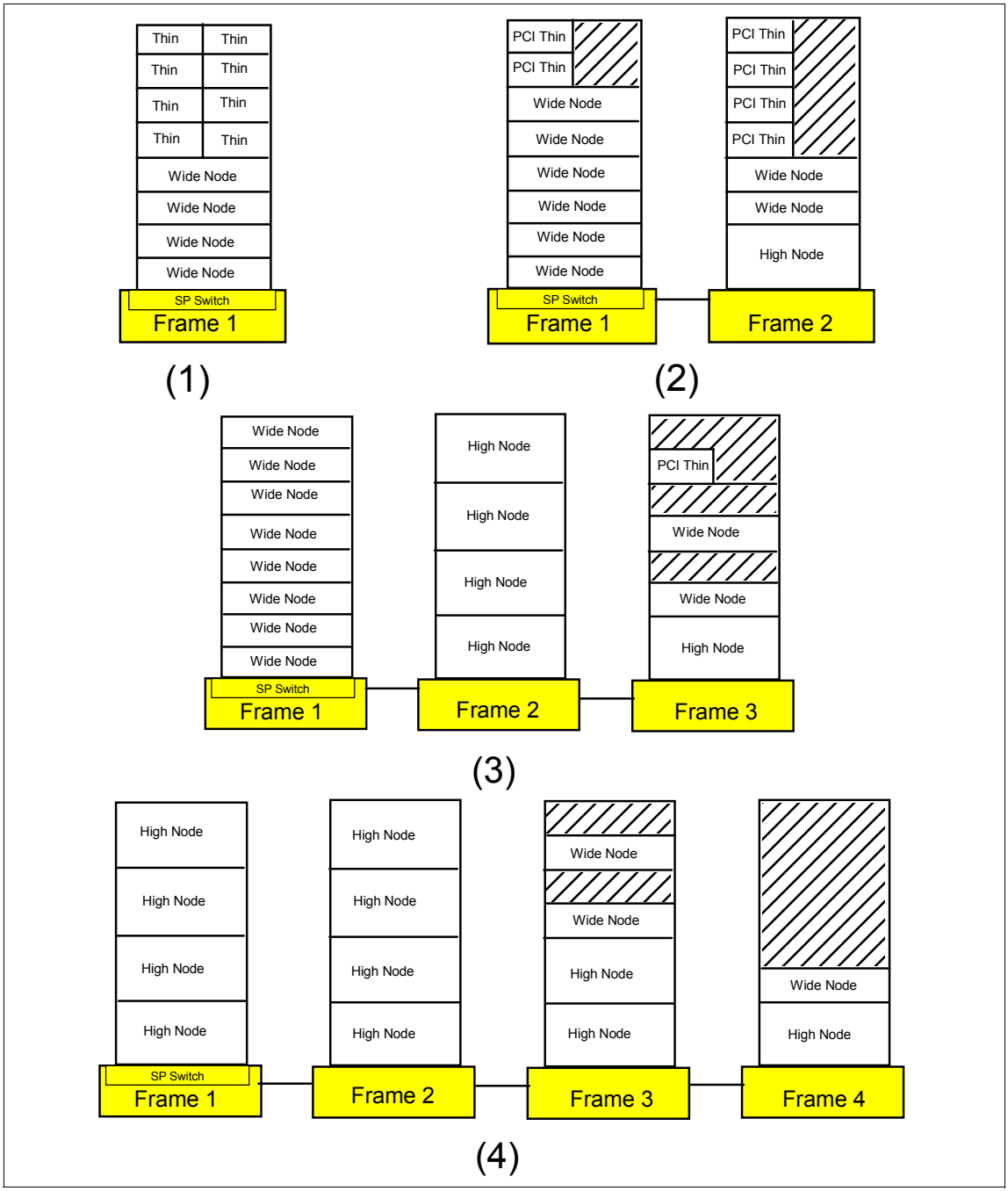


Figure 2-36 Example of single SP Switch configurations

Configuration Rule 7: SP Switch and SP Switch2 cannot coexist in the same Cluster 1600 system.

Single-stage with multiple SP Switch configurations

If your SP system has 17 to 80 nodes, switched expansion frames are required. You can add switched expansion frames and non-switched expansion frames. Nodes in the non-switched expansion frame share unused switch ports that may exist in the model frame and in the switched expansion frames. Figure 2-37 shows an example of a Single Stage SP Switch with both switched and non-switched expansion frame configurations. There are four SP Switches; each can support up to 16 processor nodes. Therefore, this example configuration can mount a maximum of 64 nodes.

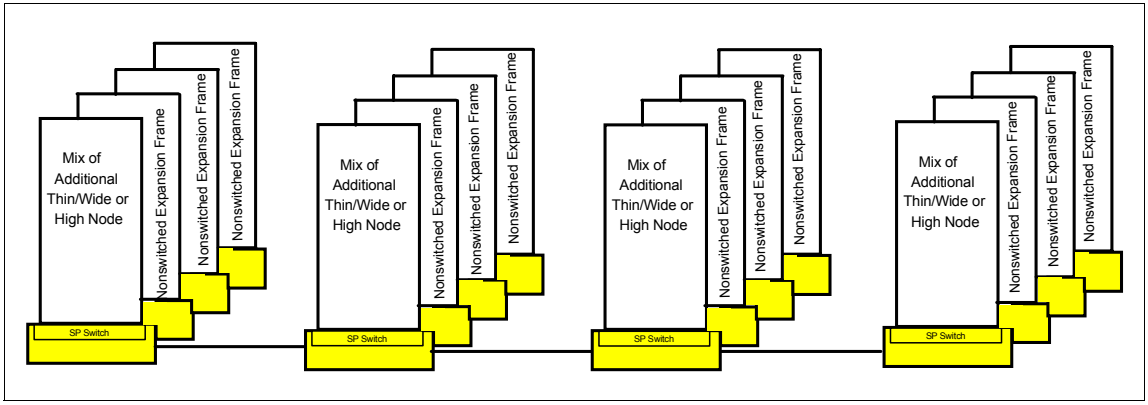


Figure 2-37 Example of a multiple SP Switch configuration

Two-stage SP Switch configurations

This configuration (Figure 2-38 on page 84) requires an SP Switch frame that forms the second switching layer. A minimum of 24 processor nodes is required to make this configuration work. It supports up to 128 nodes. Each node requires a supported SP Switch adapter. These nodes are mounted in one required tall model frame equipped with an SP Switch and at least one switched expansion frame. The SP Switch in these frames forms the first switching layer. The SP Switch frame is also required if you want more than 80 nodes or more than four switched expansion frames. This configuration can utilize both switched and non-switched expansion frames as well. Nodes in the non-switched expansion frame share unused switch ports that may exist in the model frame.

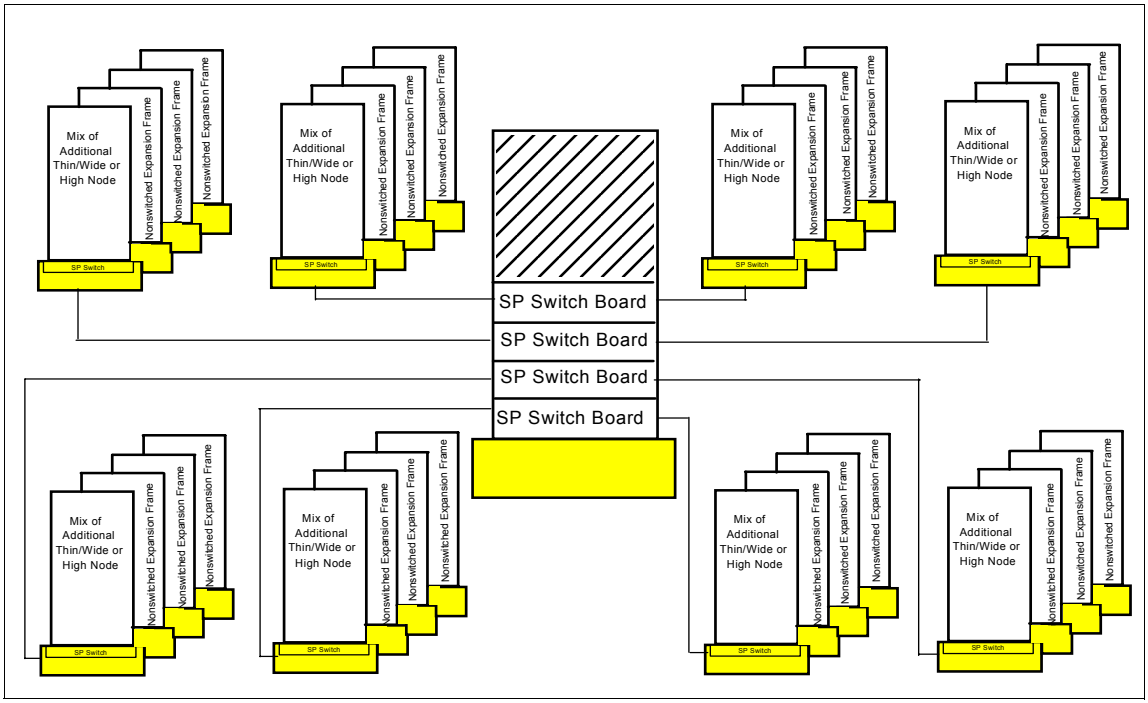


Figure 2-38 Example of two-stage SP Switch configuration

2.16 Numbering rules

Thin, wide, and high nodes can coexist in the same frame and in the same SP system partition. Whether or not you use nodes with varied physical node sizes, you must carefully consider the set of supported frame configurations. Extension nodes (like an SP Switch Router), SP-attached servers, and SP Expansion I/O Units must also be accommodated. For the Cluster 1600 numbering rules we give information on all the various configurations and numberings.

In order to place nodes in an SP system, you need to know the following numbering rules:

- The frame numbering rule
- The slot numbering rule
- The node numbering rule
- The SP Switch port numbering rule

2.16.1 The frame numbering rule

The administrator establishes the frame numbers when the system is installed. Each frame is referenced by the *tty* port to which the frame supervisor is attached and is assigned a numeric identifier. The order in which the frames are numbered determines the sequence in which they are examined during the configuration process. This order is used to assign global identifiers to the switch ports, nodes, and SP Expansion I/O Units. This is also the order used to determine which frames share a switch.To allow for growth you can skip frame numbers, but the highest number you can use for frames with nodes is 128.

If you have an SP Switch frame, you must configure it as the last frame in your SP system. Assign a high frame number to an SP Switch frame to allow for future expansion.

Frames can be the following types:

- A frame that has nodes and a switch is a *switched* frame.
- A frame with nodes and no switch is a *non-switched* expansion frame.
- A frame that has only switches for switch-to-switch connections is a *switch-only* frame. This is also called an intermediate switch board (ISB) frame.
- A frame that has only SP Expansion I/O Units, no nodes and no switch, is a *non-node* frame.

Note: The highest frame number that can be used for nodes is 128 and the highest node number is 2047. Frame numbers from 129 through 250 can be used for frames that do not have nodes.

2.16.2 The slot numbering rule

A tall frame contains eight drawers that have two slots each for a total of 16 slots. A short frame has only four drawers and eight slots. When viewing a tall frame from the front, the 16 slots are numbered sequentially from bottom left to top right.

The position of a node in an SP system is sensed by the hardware. That position is the slot to which it is wired and is the slot number of the node.

- A thin node occupies a single slot in a drawer, and its slot number is the corresponding slot.
- A wide node occupies two slots, and its slot number is the odd-numbered slot.
- A high node occupies four consecutive slots in a frame. Its slot number is the first (lowest number) of these slots.

Figure 2-39 shows slot numbering for tall frames and short frames.

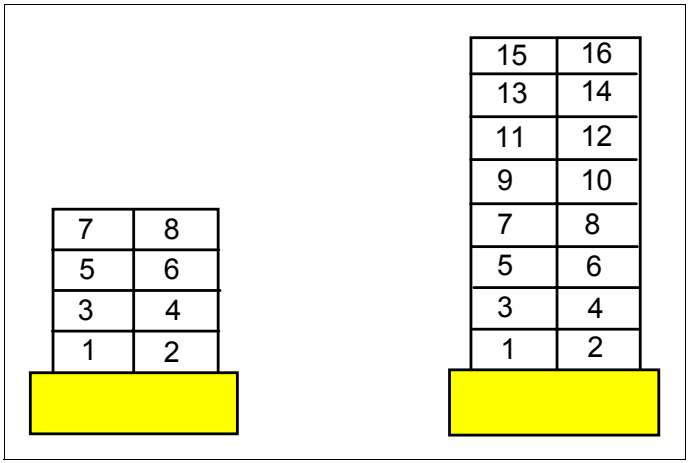


Figure 2-39 Slot numbering for short frames and tall frames (front view)

Rules for M/T7017-S70, S7A, S80, p680, and 7026-p660

The rules for these attached servers are:

- This SP-attached server type is viewed as a single frame containing a single node.
- This SP-attached server occupies the slot one position.
- Each SP-attached server installed in the SP system subtracts one node from the total node count allowed in the system. However, as the server has frame-like features, it reserves sixteen node numbers that are used in determining the node number of nodes placed after the attached server. The algorithm for calculating the node_number is demonstrated in Figure 2-40 on page 87.

```
node_number = (frame_number -1) * 16 + slot_number
```

Rules for HMC-controlled servers p690, p670, p655, p650, and p630

Note: An unpartitioned server has one LPAR and is seen by PSSP as one node. A partitioned server is seen by PSSP as one frame with as many nodes as there are LPARs. The number of these servers counts toward the total number of servers in one system.

- Each of these servers is shown as a single frame.
- When no partitioning (LPAR) is configured, one node is shown for that frame.

- When partitioning (LPAR) is used, each LPAR is shown as a node in that frame. An LPAR functions like an SP node. But since originally only 16 nodes per frame are allowed in an SP frame, you can only configure up to 16 LPARs in p690 and p690+ systems. No more than 16 LPARs will show up in a frame.There is no support from IBM for LPARs 17 to 32 configured on a system. Refer to “Cluster limits for supported pSeries server” on page 164.

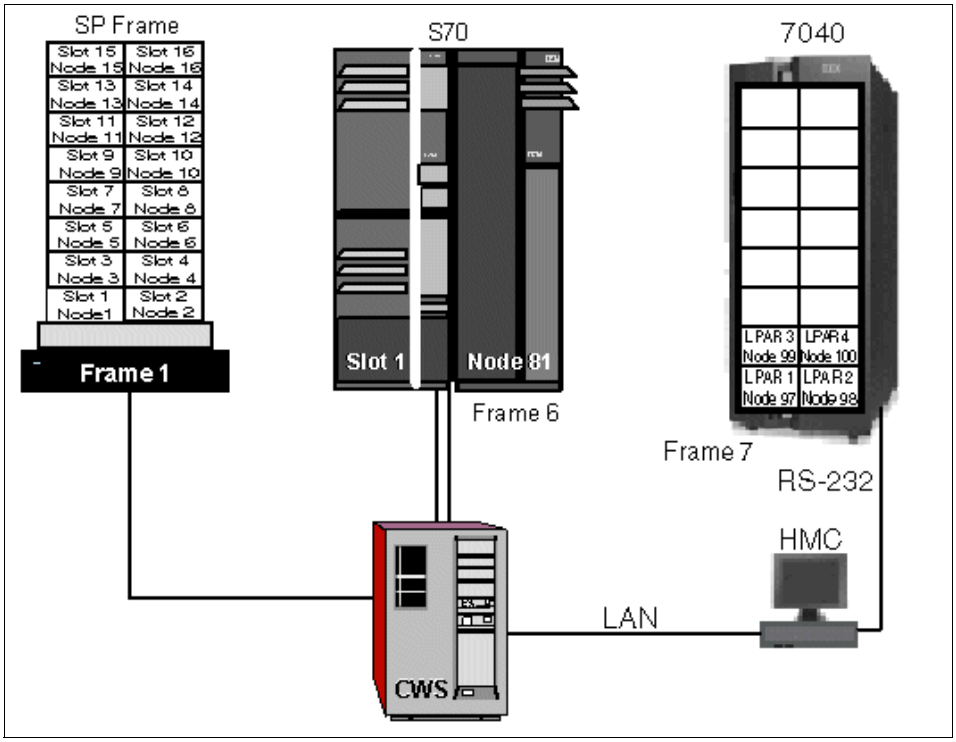


Figure 2-40 Node numbering

Important information on LPARs

Based on the preceding information, you could configure a p690+ server to have 17 LPARs with LPAR numbers 1- 17. Realizing that the p690+ server can only be attached to the Cluster 1600 system managed by PSSP if, and only if, it has no more than 16 LPARs on it, you may think that you can delete any LPAR to meet the restriction. However, this is not the case. If you deleted the LPAR with LPAR number 5, the p690+ server would only have 16 LPARs on it, but they would have LPAR numbers 1 - 4 and 6 - 17. The LPAR with LPAR number 17 violates one of the conditions for p690+ attachment. Therefore, instead of deleting the LPAR with LPAR number 5, you must delete the LPAR with LPAR

number 17 in order to properly configure the p690+ server for attachment to the cluster.

An SP-attached server is managed by the PSSP components because it is in a frame of its own. However, it does not enter into the determination of the frame and switch configuration of your SP system. It has the following additional characteristics:

- It cannot be the first frame.
- It connects to a switch port of a model frame or a switched expansion frame.
- It cannot be inserted between a switched frame and any non-switched expansion frame using that switch.

2.16.3 The node numbering rule

A node number is a global ID assigned to a node. It is the primary means by which an administrator can reference a specific node in the system. Node numbers are assigned for all nodes including SP-attached servers regardless of node or frame type. Replace node number with expansion number for the global ID of an SP Expansion I/O Unit. Global IDs are assigned using the following formula:

```
node_number = ((frame_number - 1) x 16) + slot_number
```

where *slot_number* is the lowest slot number occupied by the node or unit. Each type (size) of node occupies one slot or a consecutive sequence of slots. For each node, there is an integer n such that a thin node or expansion unit occupies slot *n*, a wide node occupies slots *n, n+1*, and a high node occupies *n, n+1, n+2, n+3*. An SP-attached server is considered to be one node in one frame. For single thin nodes (not in a pair), wide nodes, and high nodes, n must be odd. For an SP-attached server, n is 1. Use *n* in place of *slot_number* in the formula.

Node numbers are assigned independent of whether the frame is fully populated. Figure 2-41 on page 89 demonstrates node numbering. Frame 4 represents an SP-Attached server in a position where it does not interrupt the switched frame and companion non-switched expansion frame configuration. It can use a switch port on frame 2, which is left available by the high nodes in frame 3. Its node number is determined by using the previous formula.

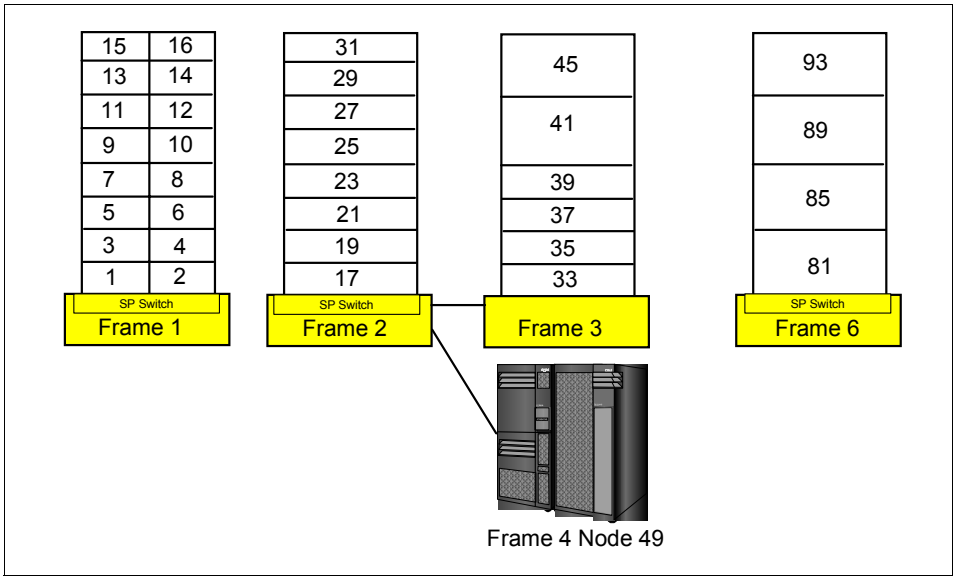


Figure 2-41 Node numbering for an SP system

An SP Switch2 system is more flexible than an SP Switch system concerning node placement. Actually, SP Switch2 has no restrictions on node placement. A node can be placed anywhere it is allowed by the physical constraints.

The only SP Switch2 rules have to do with the system configuration and with nodes connecting to the SP Switch2, which are as follows:

- You can connect a maximum of sixteen nodes to one switch.
- In a two-plane configuration, each node that is to use the switch must have two adapters with the first adapter (css0) connected to the first switch (plane 0) and the second (css1) connected to the second switch (plane 1). Any node connected to a switch in one plane must also be connected to a switch in the other plane.
- The default configuration for a system with SP Switch2 switches is as follows:
  - The system must contain a number of switches divisible by the number of planes in the system. In other words, a one-plane system can have any number of switches. A two-plane system must have an even number of switches.
  - In a two-plane system, all node switch boards must alternate between planes. For example, in a two-plane system, the first frame to contain a node switch board must contain a switch for plane 0, the second frame to contain a node switch board must contain a switch for plane 1, the third for plane 0, the fourth for plane 1, and so on.

- If the system contains multiple node switch board frames, where the frame contains one or more node switch boards in slots 1-16 of the frame, the switches must alternate between planes. For example, for a multiple node switch board frame in a two plane system, the switch in slot 0 must be in plane 0, the switch in slot 2 must be in plane 1, and so on.
- If the system contains intermediate switch board frames, where the frame contains one or more intermediate switch boards in slots 1-16 of the frame, the switches must be split among the planes. For example, if a system has two planes and eight intermediate switches, the first four switches (slots 2, 4,6 and 8) must be on plane 0 and the second four switches (slots 10, 12, 14 and 16) must be on plane 1.

4. You can override the default configuration. If you are not going to set up the system based on the default, you need to create an /etc/plane.info file to let the system know how the switches are to be configured.

Figure 2-42 on page 91 shows a sample configuration with a dual switch plane attachment.

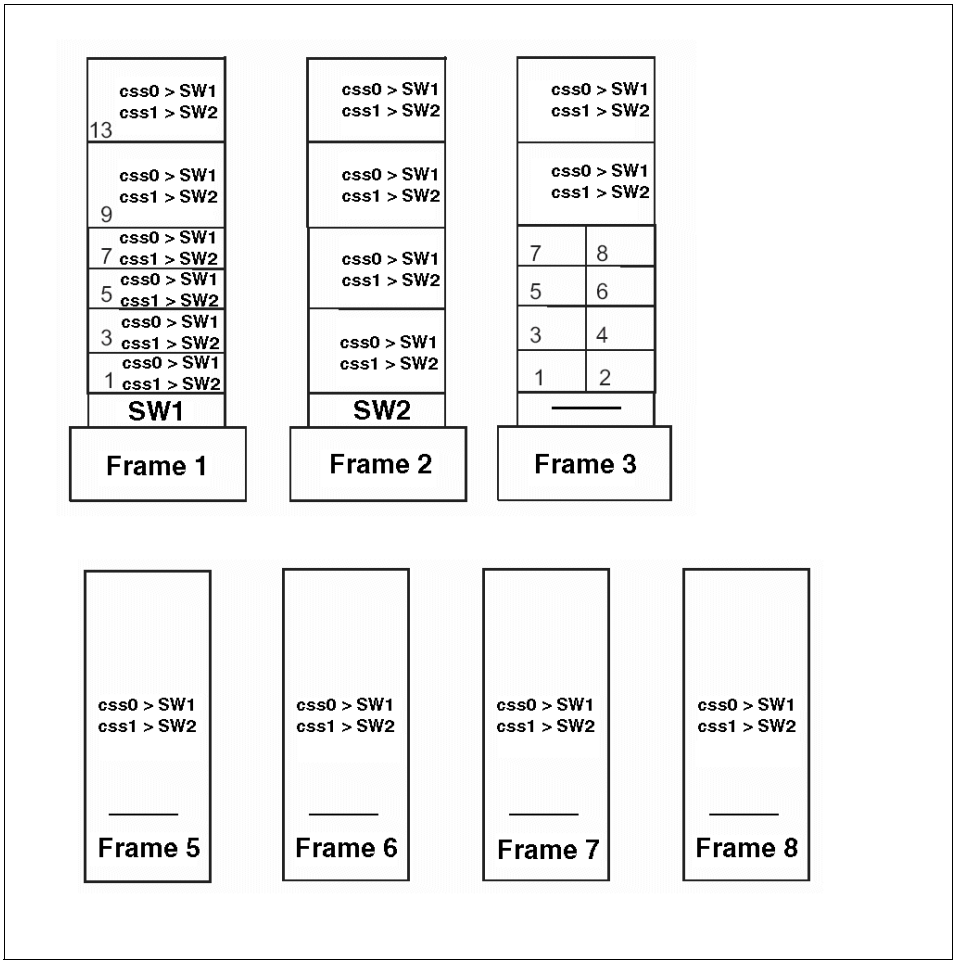


Figure 2-42 Nodes and SP-attached servers in a dual plane configuration

In this configuration, each node has two SP Switch2 adapters installed. The css0 connects to switch number 1 in Frame 1 and css1 connects to switch number 2 in Frame 2. Frames 5 to 8 are attached servers (pSeries p690). The css0 network is plane 0 and the css1 network is plane 1. A more detailed description of this configuration and its advantages is as follows:

- Frame 1 with SP Switch 2 SW1 contains four wide nodes in slots 1, 3, 5, and 7 and two high nodes in slots 9 and 13. These nodes collectively use six switch ports in SW1 and in SW2. Since slots 2, 4, 6, 8, 10, 11, 12, 14, 15, and 16 cannot be occupied by more nodes within the frame, ten switch ports remain available at this point. The high nodes in slots 9 and 13 connect to SP Expansion I/O Units in Frame 3, slots 1 and 2, respectively.

- Frame 2 with SP Switch SW2 has four high nodes at slots 1, 5, 9, and 13. Each of them connect to SP Expansion I/O Units in Frame 3, slots 3, 4, 5, and 6 respectively. Each node also connects to SW1 and SW2, so six switch ports now remain to be claimed on each switch.
- Frame 3 has no switch. It is a non-switched expansion frame. This frame has eight SP Expansion I/O Units, of which six are used by nodes in Frames 1 and 2. It also has two high nodes at slots 9 and 13 connected to the I/O units in slots 7 and 8 and two switch ports in SW1 and SW2, leaving four ports yet to be claimed on each switch.
- Frames 5 through 8 are SP-attached servers, each connected to one of the remaining four ports in each switch. The first of these has frame number 5.

We still have room for future upgrades in this configuration. For instance, by replacing the wide nodes in Frame 1 with two high nodes, another SP expansion frame (as Frame 4) can house two more high nodes and SP Expansion I/O Units.

Note: With SP Switch2, node numbers and switch port numbers are automatically generated.

2.16.4 The switch port numbering rule

In a switched system, the switch boards are attached to each other to form a larger communication fabric. Each switch provides some number of ports to which a node can connect (16 ports for an SP Switch/SP Switch2 and 8 ports for the SP Switch-8.) In larger systems, additional switch boards (intermediate switch boards) in the SP Switch frame are provided for switch board connectivity; such boards do not provide node switch ports.

Switch boards are numbered sequentially starting with 1 from the frame with the lowest frame number to that with the highest frame number. Each full switch board contains a range of 16 switch port numbers (also known as switch node numbers) that can be assigned. These ranges are also in sequential order with their switch board number. For example, switch board 1 contains switch port numbers 0 through 15.

Switch port numbers are used internally in PSSP software as a direct index into the switch topology and to determine routes between switch nodes.

Switch port numbering - SP Switch

Switch port numbering for an SP Switch

The SP Switch has 16 ports. Whether a node is connected to a switch within its frame or to a switch outside of its frame, you can use the following formula to determine the switch port number to which a node is attached:

```
switch_port_number = ((switch_number - 1) x 16) + switch_port_assigned
```

where `switch_number` is the number of the switch board to which the node is connected, and `switch_port_assigned` is the number assigned to the port on the switch board (0 to 15) to which the node is connected.

Figure 2-43 on page 94 shows the frame and switch configurations that are supported and the switch port number assignments in each node. Let us describe more details about each configuration.

In configuration 1, the switched frame has an SP Switch that uses all 16 of its switch ports. Since all switch ports are used, the frame does not support non-switched expansion frames.

If the switched frame has only wide nodes, it could use, at most, eight switch ports and, therefore, has eight switch ports to share with non-switched expansion frames. These expansion frames can be configured as in configuration 2 or configuration 3.

In configuration 4, four high nodes are mounted in the switched frame. Therefore, its switch can support 12 additional nodes in non-switched expansion frames. Each of these non-switched frames can house a maximum of four high nodes. If wide nodes are used, they must be placed in the high node slot positions.

A single PCI thin node can be mounted in a drawer. Therefore, it can be mounted in non-switched expansion frames. In this circumstance, it must be installed in the wide node slot positions (configuration 2) or high node slot positions (configuration 3 and 4).

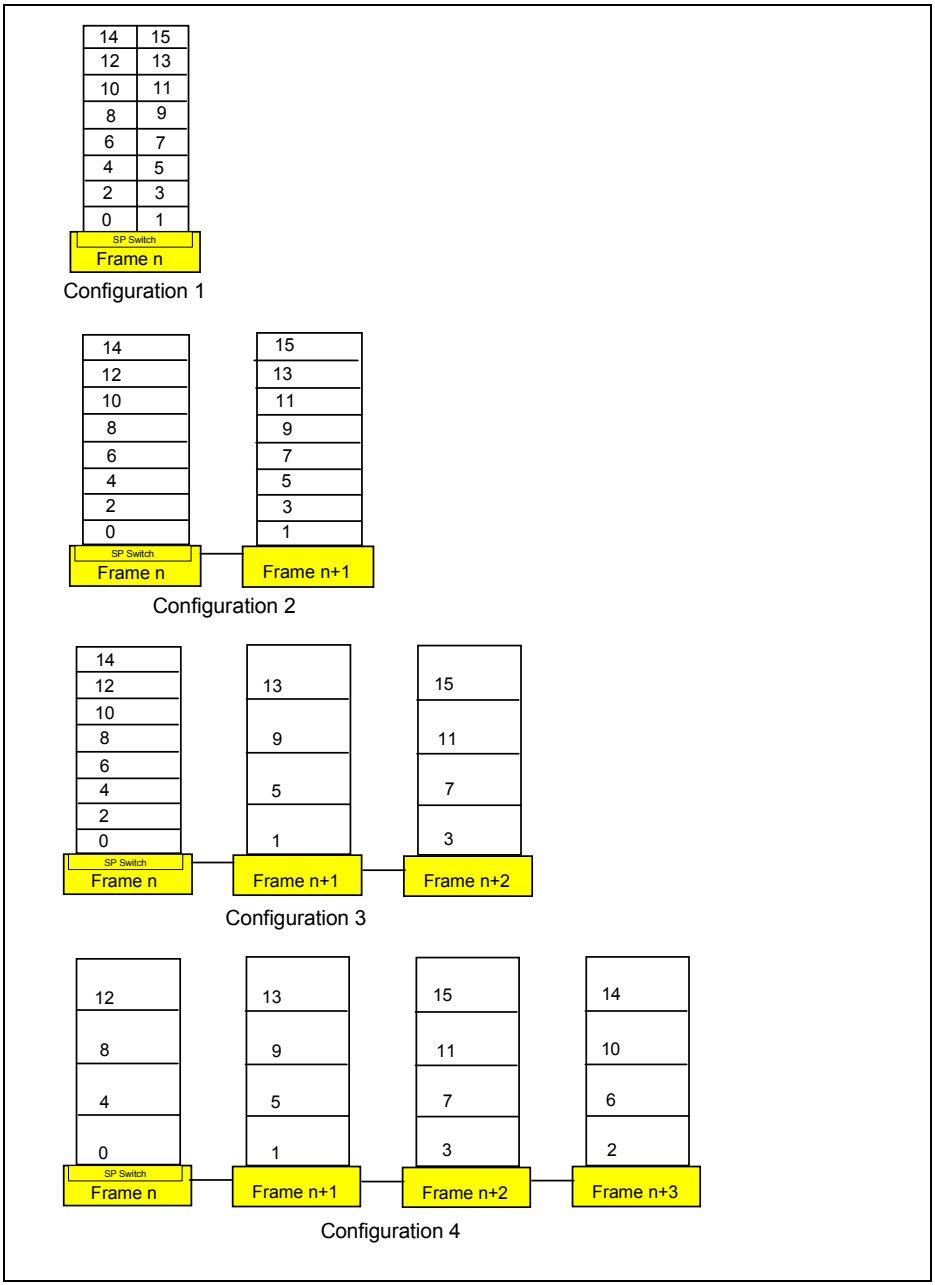


Figure 2-43 Switch port numbering for an SP Switch

Switch port numbering - SP Switch-8

Switch port numbering for an SP Switch-8

Node numbers for short and tall frames are assigned using the same algorithm. See 2.16.3, “The node numbering rule” on page 88.

An SP system with SP switch-8 contains only switch port numbers 0 - 7. The following algorithm is used to assign nodes their switch port numbers in systems with eight port switches:

1. Assign the node in slot 1 to switch_port_number = 0. Increment switch_port_number by 1.
2. Check the next slot. If there is a node in the slot, assign it the current switch_port_number, then increment the number by 1.

   Repeat until you reach the last slot in the frame or switch port number 7, whichever comes first.

Figure 2-44 shows sample switch port numbers for a system with a short frame and an SP Switch-8.

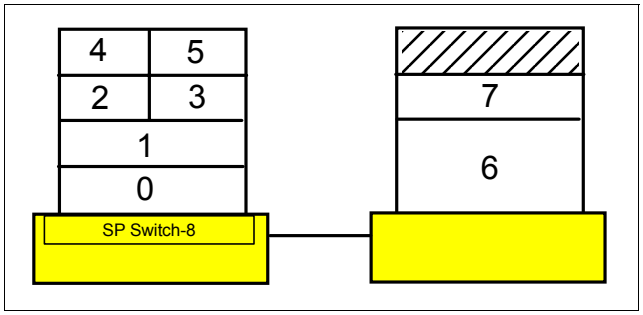


Figure 2-44 Example of switch port numbering for an SP Switch-8

Switch port numbering - SP Switch2

Switch port numbering for an SP Switch2

The switch port numbers for an SP Switch2 system are automatically assigned sequentially by the PSSP switch management component (CSS). As a node is assigned a CSS adapter, it is given the lowest available switch node number from 0 through 511. There is no correlation between the switch port number and any hardware connections.

Switch port number - switchless system

Switch port numbering for a switchless system

Perhaps you plan your Cluster 1600 system without a switch, but even if you do not plan to use an SP Switch you need to plan a switch network, whenever you plan to use any of the following:

- System partitioning
- An SP-attached server (now or in the future)

In any switchless system, a logical switch port number can be evaluated using frame numbers and slot numbers with the following formula:

```
switch_port_number = ((frame_number - 1) x 16) + slot_number - 1
```

Related documentation

The following documents should help you understand the concepts and examples covered in this guide in order to maximize your chances of success in the exam.

SP manuals

The manual *RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment,* GA22-7280 is a helpful hardware reference. It is included here to help you select nodes, frames, and other components needed, and ensure that you have the correct the physical configuration and environment.

RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment, GA22-7281 is a good reference to help plan and make decisions about what components to install and also which nodes, frames, and switches to use, depending on the purpose.

332 MHz Thin and Wide Node Service, GA22-7330 explains the configuration of 332 MHz thin and wide nodes.

SP redbooks

Inside the RS/6000 SP, SG24-5145 serves as an excellent reference for understanding the various SP system configurations you could have.

RS/6000 SP Systems Handbook, SG24-5596 is a comprehensive guide dedicated to the RS/6000 SP product line. Major hardware and software offerings are introduced and their prominent functions discussed.

2.17 Sample questions

This section provides a series of questions to help aid you in preparation for the certification exam. The answers to these questions can be found in Appendix A.

1. The SP Switch router node is an extension node. It can support multiple switch adapter connections for higher availability and performance. Which of the following is not a requirement of extension nodes?
   a. CWS
   b. PSSP 2.4 or higher on Primary node
   c. Primary node
   d. Backup node
2. Which of the following is not a true statement regarding the capability of an SP Switch2 over an SP Switch?
   a. Fault isolation
   b. Compatible with older SP Switch switches
   c. Improved bandwidth
   d. Higher availability
3. Which is a minimum prerequisite for PSSP Version 3 release 5?
   a. AIX 5L Version 5.1
   b. IBM C for AIX, Version 4.3
   c. Performance Toolbox Parallel Extensions (PTPE)
   d. IBM Performance Toolbox, Manager Component, Version 2.2
4. A customer is upgrading an existing 200 MHz high node to the new 332 MHz SMP thin node. The SP system contains an SP Switch. How many available adapter slots will the customer have on the new node?
   a. Two PCI slots. The Ethernet is integrated, and the SP Switch has a dedicated slot.
   b. Eight PCI slots. Two slots are used by an Ethernet adapter and the SP Switch adapter.
   c. Ten PCI slots. The Ethernet is integrated, and the SP Switch has a dedicated slot.
   d. Nine PCI slots. The Ethernet is integrated, and the SP Switch adapter takes up one PCI slot.
5. An SP-attached M/T 7040 server controlled by an HMC requires connectivity with the SP system in order to establish a functional and safe network. How many connections to the CWS are required?
   a. 1 - one Ethernet connection from HMC to the SP administrative LAN
   b. 3 - one Ethernet connection from HMC to the SP LAN, one RS-232 to the CWS and one RS-232 connection to the SP-attached server

c. 2 - one Ethernet connection from HMC to the SP LAN, one RS-232 connection to the CWS

d. 1 - one RS-232 connection from the HMC to the CWS

6. Which of the following is NOT a function provided by the control workstation?

   a. Authentication Service

   b. File Collection Service

   c. Boot/Install Service

   d. Ticket Granting Service

7. Which of the following is a minimum prerequisite for the CWS hardware?

   a. Two color graphic adapters

   b. SP Ethernet adapters for connections to the SP Ethernet

   c. 2 GB of disk storage

   d. 1 GB of main memory

8. Which SP-attached server is not using the integrated Ethernet and does not require the SP LAN adapter to be en0?

   a. M/T 7017

   b. M/T 7040

   c. M/T 7026

   d. M/T 7039

9. An SP-attached server M/T 7040 p690, fully configured, is part of a Cluster 1600. It is configured with 12 LPARs. Which of the following statements regarding LPARs is correct?

   a. Up to 32 LPARs can be configured and controlled by the CWS

   b. Each LPAR will show up as single frame in perspectives

   c. Only four more LPARs can be configured in order to not exceed the 16 LPARs per server limit

   d. 17 LPARs are configured and LPAR 6 can be deleted to satisfy the 16 LPARs per server limit

10. Which of the following statements regarding configuration rules is correct?

    a. The tall frame and short frames can be mixed within an SP system.

    b. A short frame supports only a single SP Switch-8 board.

    c. Tall frames does not support SP-Attached servers.

    d. If there is a single PCI thin node in a drawer, it must be installed in the even slot position (right side of the drawer).

2.18 Exercises

Here are some exercises that you may wish to perform:

1. Utilizing the study guide test environment (Figure 1) on page 3, describe the necessary steps to add an SP-attached server to the current environment.
2. Refer to the study guide test environment for the following exercise: Describe the necessary steps to add a third switch frame with one SMP high node and two SMP thin nodes to the current environment.
3. What are the necessary steps to make the first node on the second frame a boot/install server? Refer to study guide test environment (Figure 1 on page three). Assume that the first node on frame one is already a BIS.
4. Describe the configuration of figure (d) on page 62. Assume that you need to add two POWER3 thin nodes, what would be the locations?
5. The SP Switch router requires a minimum of three connections with your SP system. What are the required connections?

3

Cluster 1600 networking

Cluster 1600 is what used to be called an RS/6000 SP. The name has been changed to emphasize the fact that it now encompasses much more than frames with SP nodes, SP-attached servers, and an occasional SP switch. In fact, it is possible to have a Cluster 1600 complex comprised entirely of pSeries nodes. But this does not change the fact that one of the key components in any cluster design is that no matter how many building blocks we need and no matter which we use, they all need to be bound together to be able to communicate with each other. Figure 3-1 on page 102 is an example of a Cluster 1600 network configuration. This figure only shows the basic network needs for such a configuration. No additional adapters have been configured yet, only the SP Ethernet admin LAN, the RS-232 and switch network—this gives an idea why planning is very important.

This chapter covers some of the networking issues in a Cluster 1600 system managed by PSSP, and some of the issues when it comes to segmentation, name resolution, or routing, and what impact these things have on the IBM Parallel System Support Programs (PSSP) subsystems.

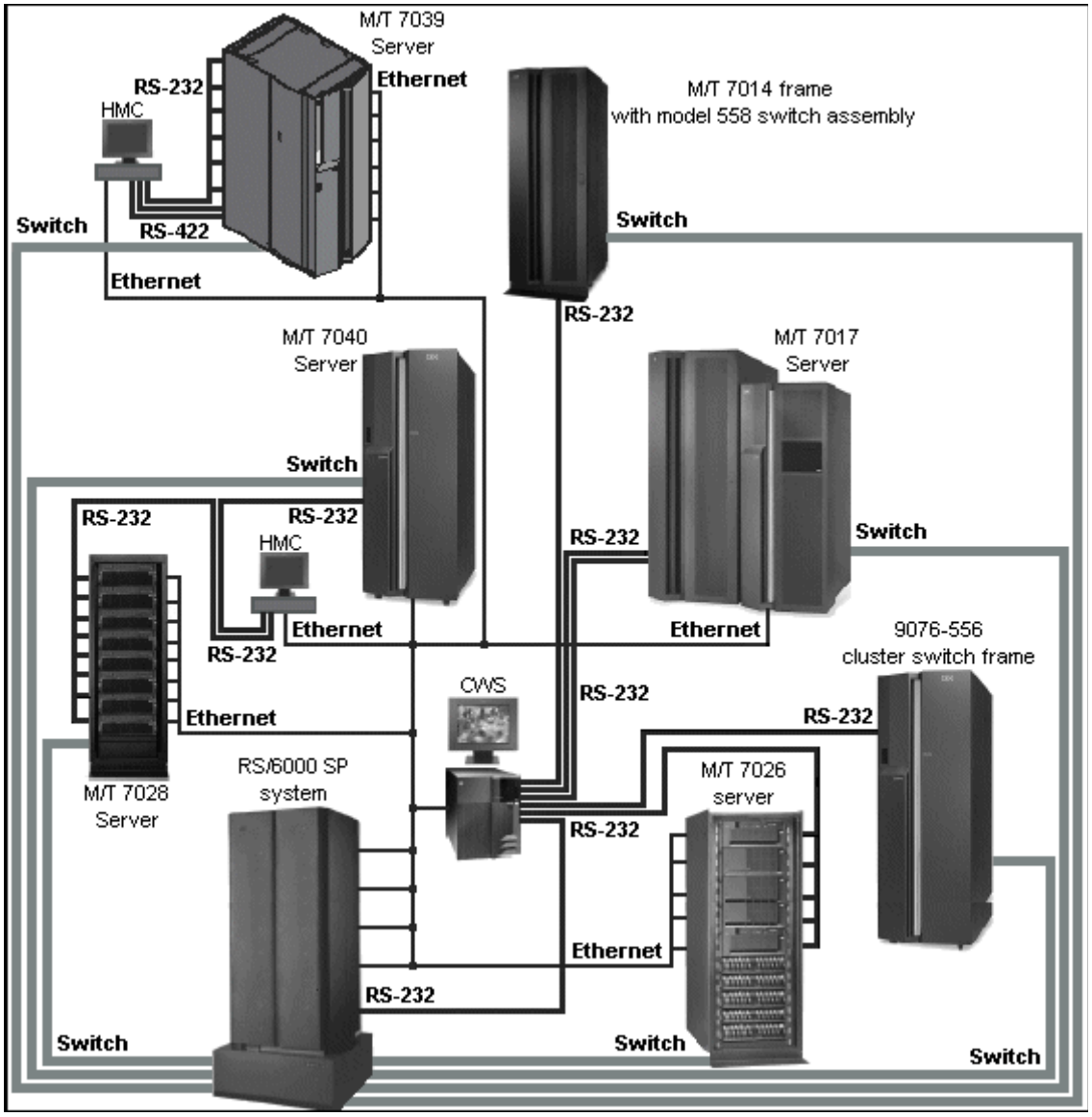


Figure 3-1 Cluster 1600 Managed by PSSP network configuration

3.1 Key concepts you should study

The concepts covered in this section will clarify some of the situations you might meet while dealing with the configuration of your Cluster 1600 network, as well as provide a good preparation for some of the questions in the certification exam. In order to improve your chances, you should become familiar with the following:

- How to create specific host names (both fully qualified and aliases), TCP/IP addresses, netmask values, and default routes.
- How to determine the Ethernet topology, segmentation, and routing across the system.
- How to determine TCP/IP addressing for the switch network.
- How to plan for network security.

3.2 Name, address, and network integration planning

Like in any other system configuration, network implementation requires planning. How many adapters will be used and the placement of these adapters? Which networks will you connect to? Will you be needing subnetting or routing? Will it be a secure or an unsecure network? The list goes on. This is why planning is so important.

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

In any Cluster 1600 configuration, each node needs to have an assigned IP address, a host name, and most likely a route. This is needed for each network connection and is also applicable for the control workstation (CWS) and the hardware management console (HMC).

You need to ensure that all the addresses you assign are unique within your site network. When it comes to host names, you need to plan how names and addresses will be resolved on your systems, that is, whether you will you be using DNS name servers, /etc/hosts files, or some other method. Remember that there are security issues to consider as well.

The following are suggested guidelines for designing the SP Ethernet topology for efficient network installs. Many of the configuration options will require additional network hardware beyond the minimal node and CWS requirements. For all other nodes, you must use the en0 adapter to connect each node to the

SP Ethernet admin LAN. The following requirements pertaining to the SP Ethernet admin LAN exist for all configurations:

- Each boot-install server's Ethernet adapter must be directly connected to each of the control workstations' Ethernet adapters.
- The Ethernet adapter must always be in the cluster nodes' lowest hardware slot of all Ethernets. This does not pertain to HMC-controlled servers.
- The NIM clients that are served by boot-install servers must be on the same subnet as the boot-install server's Ethernet adapter.
- NIM clients must have a route to the CWS over the SP Ethernet.
- The CWS must have a route to the NIM clients over the SP Ethernet.

3.2.1 Configure SP Ethernet admin LAN adapter

Independent of any of the network adapters, each machine has a host name, which is usually the name given to one of the network adapters in the machine. In this section we discuss the basic concept of configuring the Ethernet adapter on your CWS.

Add the CWS SP Ethernet admin Interface

Use **SMIT** or the **chdev** command to configure each Ethernet adapter connecting your frames to the CWS. For example, enter a command similar to the following to configure an SP Ethernet administrative LAN adapter:

```
chdev -l en0 -a netaddr=129.33.41.1 -a netmask=255.255.255.0 -a state=up
```

If the adapter is not yet defined or configured, use the **smit mkinet** or the **mkdev** command instead of **smit chinet** or **chdev** to specify a new IP host name and netmask values. For example, enter a command similar to the following to define and configure an SP Ethernet administrative LAN adapter:

```
mkdev -c if -s EN -t en -a netaddr=129.33.34.1 \
-a netmask=255.255.255.0 -a state=up -q -w en0
```

Verify the CWS interfaces

Verify each Ethernet adapter by pinging its IP address to see if you get a proper response. If you do not receive a response, debug the network problem, and reconfigure the adapter. For example:

```
ping -c 1 129.33.34.1
```

Set the host name of the CWS

Use **SMIT** or the **hostname** command to set the host name of the CWS:

```
hostname sp3n0
```

Important: The host name of the CWS is bound to many of the PSSP-used demons. Changing the host name on the CWS after configuring PSSP could cause problems.

Due to the strictness in the PSSP architecture, each node's en0 (first Ethernet adapter) needs to be connected to the SP LAN. When you do the network install on a node, this adapter will by default be your initial host name and reliable host name. The reliable host name is the name of your management adapter, which is used for management purposes (NIM, Kerberos, etc.). Changing the reliable host name adapter can be rather challenging, but if needed it is covered in *PSSP Administration Guide,* SA22-7348. The initial host name adapter, however, can be changed to whatever adapter you have on the node, such as the css adapter, which is then the "host name" of the node.

3.2.2 Set routes

Routing is very important in any larger system environment. PSSP supports multiple subnets, but all the nodes need to be able to access those subnets. Every node must have access to the control workstation, even when it is being installed from a boot/install server other than the CWS. By defining a route, you basically show the nodes adapter how to get to the other subnet through the gateway selected. The gateway is the IP address that is able to reach the other subnets.

Before configuring boot/install servers for other subnets, make sure the control workstation has routes defined to reach each of the additional subnets.

To set up static routes, you may use SMIT or the command line. To add routes using the command line, use the **route** command:

```
route add -net <ip_address_of_other_network> -netmask
<ip_address_of_gateway>
```

where:

<ip_address_of_other_network> is the IP address of the other network in your LAN. <ip_address_of_gateway> is the IP address of the gateway.

For example:

```
route add -net 192.168.15 -netmask 255.255.255.0 9.12.0.130
```

3.2.3 Host name resolution

TCP/IP provides a naming system that supports both flat and hierarchical network organization so that users can use meaningful, easily remembered names instead of 32-bit addresses.

In flat TCP/IP networks, each machine on the network has a file (/etc/hosts) containing the name-to-Internet-address mapping information for every host on the network.

When TCP/IP networks become very large, as on the Internet, naming is divided hierarchically. Typically, the divisions follow the network's organization. In TCP/IP, hierarchical naming is known as the domain name service (DNS) and uses the DOMAIN protocol. The DOMAIN protocol is implemented by the named daemon in TCP/IP.

The default order in resolving host names is:

1. BIND/DNS (named)
2. Network Information Service (NIS)
3. Local /etc/hosts file

The default order can be overwritten by creating a configuration file, called /etc/netsvc.conf, and specifying the desired order. Both default and /etc/netsvc.conf can be overwritten with the environment variable `nsorder`.

The /etc/resolv.conf file

The /etc/resolv.conf file defines the domain and name server information for local resolver routines. If the /etc/resolv.conf file does not exist, then BIND/DNS is considered not to be set up or running. The system will attempt name resolution using the local /etc/hosts file.

A sample /etc/resolv.conf file is:

```
# cat /etc/resolv.conf
domain msc.itso.ibm.com
search msc.itso.ibm.com itso.ibm.com
nameserver 9.12.1.30
```

In this example, there is only one name server defined, with an address of 9.12.1.30. The system will query this domain name server for name resolution. The default domain name to append to names that do not end with a period (.) is msc.itso.ibm.com. The search entry when resolving a name is `msc.itso.ibm.com` and `itso.ibm.com`

3.2.4 DNS

Domain Name Server (DNS) is the way that host names are organized on the Internet using TCP/IP. Host names are used to look up or resolve the name we know a system by and convert it to a TCP/IP address. All of the movement of data on a TCP/IP network is done using addresses, not host names; so, DNS is used to make it easy for people to manage and work with the computer network.

If your cluster configuration is on a site with many systems, you can use DNS to delegate the responsibility for name systems to other people or sites. You can also reduce your administration workload by only having to update one server in case you want to change the address of the system.

DNS uses a name space in a way similar to the directories and subdirectories we are used to. Instead of a "/" between names to show that we are going to the next level down, DNS uses a period or full stop.

In the same way that "/" denotes the root directory for UNIX, DNS has "." as the root of the name space. Unlike UNIX, if you leave out the full stop or period at the end of the DNS name, DNS will try various full or partial domain names for you. One other difference is that, reading left to right, DNS goes from the lowest level to the highest, whereas the UNIX directory tree goes from the highest to the lowest.

For example, the domain ibm.com is a subdomain of the .com domain. The domain itso.ibm.com is a subdomain of the ibm.com domain and the .com domain.

You can set up your system without DNS. This uses a file called /etc/hosts on each system to define the mapping from names to TCP/IP addresses. Because each system has to have a copy of the /etc/hosts file, this may become difficult to maintain, even for a small number of systems. Even though setting up DNS is more difficult initially, the administrative workload for three or four workstations may be easier than with /etc/hosts.

Maintaining a network of 20 or 30 workstations becomes just as easy as for three or four workstations. It is common for a Cluster 1600 system implementation to use DNS in lieu of /etc/hosts.

When you set up DNS, you do not have to match your physical network to your DNS setup, but there are some good reasons why you should. Ideally, the primary and secondary name servers should be the systems that have the best connections to other domains and zones.

All names and addresses of all IP interfaces on your nodes must be resolvable on the CWS and on independent workstations set up as authentication servers

before you install and configure the Cluster 1600 system. The Cluster 1600 managed by PSSP configuration uses only IPv4 addresses. Some PSSP components tolerate IPv6 aliases for IPv4 network addresses but not with DCE, HACMP, HACWS, or if using the SP Switch. For more information about the Cluster 1600 managed by PSSP support for IPv6, see the appendix on the subject in *PSSP Administration Guide,* SA22-7348.

Tip: Once you have set the host names and IP addresses on the CWS, you should not change them.

3.3 Networks

You can connect many different types of LANs to the cluster configuration, but regardless of how many you use, the LANs fall into one of the following categories.

3.3.1 The SP Ethernet admin LAN

In this section we describe the setup of that administrative Ethernet, which is often referred to as the SP Ethernet admin LAN or SP LAN.

The SP Ethernet is the administrative LAN that connects all nodes in one system running PSSP to the CWS. It is used for PSSP installation and communication among the CWS, boot-install servers, and other nodes in the network. For HMC-controlled servers, it can also be used to connect the HMC to the CWS for hardware control and monitoring.

In order for the PSSP installation to function, you must connect the SP Ethernet to the Ethernet adapter in the cluster's nodes' lowest hardware slot (of all Ethernet adapters) on that node. This adapter will become the management adapter. When a node is network booted, it selects the lowest Ethernet adapter from which to perform the install. This Ethernet adapter must be on the same subnet of an Ethernet adapter on the node's boot/install server.

In nodes that have an integrated Ethernet adapter, it is always the lowest Ethernet adapter. Be sure to maintain this relationship when adding Ethernet adapters to a node.

3.3.2 Frame and node cabling

The original 9076 SP frames included coaxial Ethernet cabling for the SP LAN, also known as *thin-wire* Ethernet or 10BASE-2. All nodes in the frame can be connected to that medium through the BNC connector of either their integrated

10 Mbps Ethernet or a suitable 10 Mbps Ethernet adapter using T-connectors. Access to the medium is shared among all connected stations and controlled by Carrier Sense, Even though this network is reliable, problem determination on this type of net is rather difficult. It is also very slow due to the fact that it only supports 10 Mbit half duplex. It also has a hard limit of 30 nodes per 10BASE-2 segment, and the total length must not exceed 185 meters, so these days most of us would choose the Ethernet 10/100 Mbit (100BASE-TX) together with an Ethernet switch. When using this solution, we will hardly run into any of the problems the 10BASE-2 gave us, such as collision.

The network install process uses the SP Ethernet for transferring the install image from the install server to the cluster nodes. Depending on the size of your system, this might become a well-known bottleneck, In that case you want to use Boot Install Servers (BIS).

The guidelines for designing the SP Ethernet topology for efficient network installs can be found in *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281. Keep in mind that many of the configuration options will require additional network hardware beyond the minimal node and CWS requirements, and there are also network addressing and security issues to consider.

3.3.3 SP LAN topologies

The network topology for the SP LAN mainly depends on the size of the system and should be planned on an individual basis. Usually the amount of traffic on this network is low, but when you do network installs it will be maximized. You might never hit any problems on a small system, but having a 128-node cluster, network planning is very important.

We strongly recommend that you provide additional network connectivity (through the SP Switch or any high-speed network) if any applications that perform significant communication among nodes are used (such as LoadLeveler®). This is due to the fact that all subnet routing will go through the SP Ethernet admin LAN adapter on the CWS and might very well cause a bottleneck. To avoid such problems, moving the subject applications to the CWS might solve the problem, since this would cut that traffic in half, but then the CWS must have the capacity to accommodate such an application.

Important: To avoid overloading the SP LAN by application traffic, it should be used only for SP node installations and system management, and applications should use any additional network connectivity.

If you connect the SP Ethernet to your external network, you must make sure that the user traffic does not overload the SP Ethernet network. If your outside network is a high-speed network such as FDDI or HIPPI, routing the traffic to the SP Ethernet can overload it. Guidelines for designing the SP Ethernet topology can be found in “Planning your system network section” in *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281.

In the following, only the SP LAN is considered. We show some typical network topologies with their advantages and limitations.

SP Ethernet admin LAN config

In relatively small cluster configurations, such as single frame systems or 4 - 6 pSeries nodes, the CWS and nodes typically share either a single thin-wire Ethernet (10-BASE-2) or a 10/100 Mbit switched environment as shown in Figure 3-2. In a switched environment, each node is individually cabled by 100BASE-TX Twisted Pair to ports of the Ethernet switch and operates in full duplex mode (some older SP nodes only support 10 Mbit half duplex).

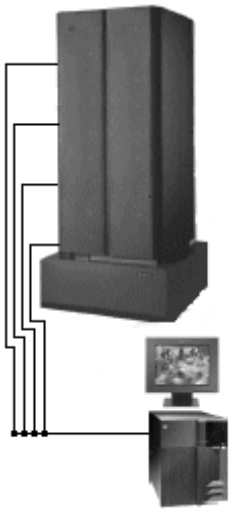
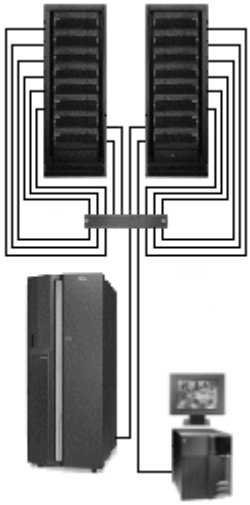
10-BASE-2 Shared Environment	10/100 Mbit Switched Environment
	

Figure 3-2 10-BASE-2 vs. 10/100 Mbit Switched

This configuration is characterized by the following properties:

- For small systems, you can use the CWS as the network install server.

- No routing is required since the CWS and all nodes share one subnet.
- The CWS acts as boot/install server for all nodes.
- Performance is limited to six to eight simultaneous network installs on a 10 Mbps HDX and 16 on a 100 Mbps TX network (a switched environment should be used due to the fact that a hub generates collisions).
- This setup is limited by the maximum number of 30 stations on a 10BASE-2 segment. In practice, not more than 16 to 24 stations should be connected to a single 10BASE-2 Ethernet segment. When using a 10/100 Mbps TX switched environment and considering only the network topology, the CWS should be able to install six to eight nodes in each Ethernet segment (port on the Ethernet switch) simultaneously since each Ethernet segment is a separate *collision domain*. Rather than the network bandwidth, the limiting factor most likely is the ability of the CWS itself in order to serve a very large number of NIM clients simultaneously, for example answering UPD bootp requests or acting as the NFS server for the mksysb images. To quickly install a large cluster system, it may still be useful to set up boot/install server nodes, but the network topology itself does not require boot/install servers. For an installation of all nodes of a large Cluster 1600 system, we advise the following:

1. Using the **spbootins** command, set up approximately as many boot/install server nodes as can be simultaneously installed from the CWS.
2. Install the BIS nodes from the CWS.
3. Install the non-BIS nodes from their respective BIS nodes. This provides the desired scalability for the installation of a whole, large Cluster 1600 configuration.
4. Using the **spbootins** command, change the non-BIS nodes configuration so that the CWS becomes their boot/install server. Do not forget to run **setup_server** to make these changes effective.
5. Reinstall the original BIS nodes. This removes all previous NIM data from them since no other node is configured to use them as boot/install server.

Using this scheme, the advantages of both a hierarchy of boot/install servers (scalable, fast installation of the cluster environment) and a flat network with only the CWS acting as a NIM server (less complexity, less disk space for BIS nodes) are combined. Future reinstallations of individual nodes (for example, after a disk crash in the root volume group) can be served from the CWS. Note that the CWS will be the only file collection server if the BIS nodes are removed, but this should not cause performance problems.

The configuration shown in Figure 3-1 on page 102 scales well to about 128 nodes. For larger systems, the fact that all the switched Ethernet segments form a single broadcast domain can cause network problems if operating system

services or applications frequently issue broadcast messages. Such events may cause broadcast storms, which can overload the network. For example, Topology Services from the RS/6000 Cluster Technology (RSCT) use broadcast messages when the group leader sends PROCLAIM messages to attract new members.

ARP cache tuning: Be aware that for cluster configurations with very large networks (and/or routes to many external networks), the default AIX settings for the Address Resolution Protocol (ARP) cache size might not be adequate. The ARP is used to translate IP addresses to Media Access Control (MAC) addresses and vice versa. Insufficient ARP cache settings can severely degrade your network's performance, in particular when many broadcast messages are sent. Refer to /usr/lpp/ssp/README/ssp.css.README for more information about ARP cache tuning.

In order to avoid problems with broadcast traffic, no more than 128 nodes should be connected to a single switched Ethernet subnet. Larger systems should be set up with a suitable number of switched subnets. To be able to network boot and install from the CWS, each of these switched LANs must have a dedicated connection to the CWS. This can be accomplished either through multiple uplinks between one Ethernet switch and the CWS or through multiple switches, each having a single uplink to the CWS. As discussed in the previous section, the limiting factor for the number of simultaneous network installations of nodes will probably be the processing power of the CWS, not the network bandwidth.

Heterogeneous 10/100 Mbps network

In many cases, an existing SP configuration will be upgraded with new nodes that have fast Ethernet connections, but older nodes should continue to run with 10 Mbps SP LAN connections. A typical scenario with connections at both 10 Mbps and 100 Mbps is shown in Figure 3-3 on page 113.

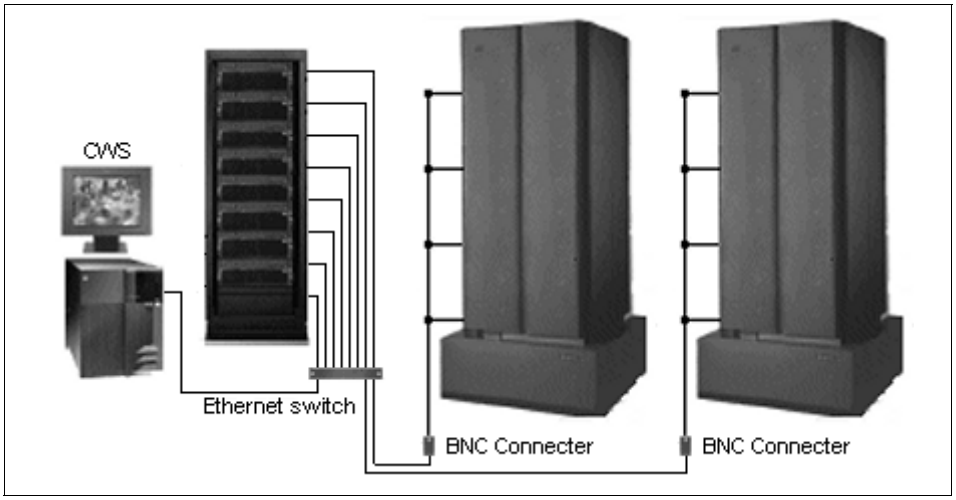


Figure 3-3 Heterogeneous 10/100 Mbps network

In this configuration, again an Ethernet Switch is used to provide a single LAN that connects to the CWS at 100 Mbps. One frame has new nodes with a 100 Mbps Ethernet. These nodes are individually cabled by 100BASE-TX Twisted Pair to ports of the Ethernet switch and operate in full duplex mode. Two frames with older nodes and 10BASE-2 cabling are connected to ports of the same Ethernet switch using BNC media converters. Ideally, a switching module with autosensing ports is used, which automatically detects the communication speed.

3.3.4 Additional LANs

The SP Ethernet can provide a means to connect all nodes and the CWS to your site networks. However, it is likely that you will want to connect your cluster nodes to a site network through other network interfaces. If the SP Ethernet is used for other networking purposes, the amount of external traffic must be limited. If too much traffic is generated on the SP Ethernet, the administration of the cluster nodes might be severely impacted. For example, problems might occur with network installs, diagnostic functions, and maintenance mode access.

Ethernet, Fiber Distributed Data Interface (FDDI), and token-ring are also configured by the PSSP. Other network adapters must be configured manually. These connections can provide increased network performance in user file serving and other network related functions. You need to assign all the addresses and names associated with these additional networks.

3.3.5 Switch network

In a cluster configuration, we optionally have another type of network called the Switch network. This network requires a switch device that gets mounted in a frame. The switch technology is covered in detail in 2.9, “SP Switch and SP Switch2 communication network” on page 50.

Switches are used to connect processor nodes, providing the message passing network through which they communicate with a minimum of four disjoint paths between any pair of nodes. In a Cluster 1600 system managed by PSSP, you can use only *one* type of switch, either SP Switch2 or SP Switch (not even in separate SP system partitions). For planning reasons remember that SP Switch2 or SP Switch require a frame for housing and in order to connect to the switch, and the cluster nodes require switch adapters.

If using IP for communication over the switch, each node needs to have an IP address and name assigned for the switch interface (the css0 adapter). If you plan to use Switch2 with two switch planes, you also need to have an IP address and name assigned for the css1 adapter and you have the option to use the ml0 aggregate IP interface. If hosts outside the switch network need to communicate over the switch using IP with nodes in the cluster system, those hosts must have a route to the switch network through one of the cluster nodes.

If you use Switch2, and all nodes are running PSSP 3.4 or later, you have optional connectivity, so some nodes can be left off the switch. However, this is not supported on SP Switch, where all nodes need to be attached

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

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

For more information about SP Switch and SP Switch2, see 2.9.2, “SP Switch hardware components” on page 54.

3.3.6 Subnetting considerations

All but the simplest Cluster 1600 system configurations will most likely include several subnets. Thoughtful use of netmasks in the planning of your network can economize the use of network addresses.

When using a subnet mask of 255.255.255.0, the first three bytes indicate the network you are on (A, B, or C) and the last byte is the host that you are on that network. Hosts .1 through .254 are available.

By using a subnet mask of 255.255.255.128, we can split that network into two halves, the first half containing the host addresses .1 through .126, the second half containing the host addresses .129 through .254.

As an example, if we are configuring a new cluster where none of the six subnets making up the SP Ethernet will have more than 16 nodes on them, how could the subnetting then be? If we used a netmask of 255.255.255.240, which provides 16 addresses, we only get 14 discrete addresses. However, if we used the netmask of 255.255.255.224, we would get 32 addresses, which would provide 30 discrete addresses per subnet. This would satisfy our need for 16 nodes on each subnet. Using 255.255.255.224 as a netmask, we can then allocate the address ranges as follows: 192.168.3.1-31, 192.168.3.33-63, and 192.168.3.65-96.

For example, if we used 255.255.255.0 as our netmask, we would have to use four separate Class C network addresses to satisfy the same wiring configuration (that is, 192.168.3.x, 192.168.4.x, 192.168.5.x, and 192.168.6.x). An example of SP Ethernet subnetting is shown in Figure 3-4 on page 116.

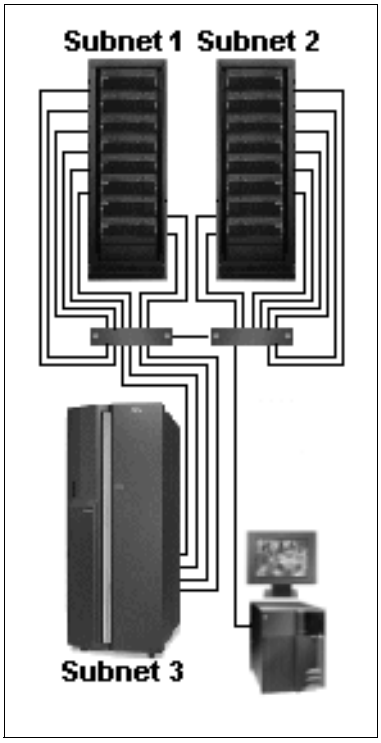


Figure 3-4 SP Ethernet subnetting example

3.4 Routing considerations

When planning routers, especially router nodes, several factors can help determine the number of routers needed and their placement in the Cluster 1600 configuration. The number of routers you need can vary depending on your network type (in some environments, router nodes might also be called gateway nodes).

For nodes that use Ethernet or Token Ring as the routed network, CPU utilization may not be a big problem. For nodes that use FDDI as the customer routed network, a customer network running at or near maximum bandwidth results in high CPU utilization on the router node. Applications, such as POE and the Resource Manager, should run on nodes other than FDDI routers. However, Ethernet and Token-Ring gateways can run with these applications.

For systems that use Ethernet or Token-Ring routers, traffic can be routed through the SP Ethernet. For FDDI networks, traffic should be routed across the

switch to the destination nodes. The amount of traffic coming in through the FDDI network can be up to ten times the bandwidth that the SP Ethernet can handle.

For bigger demands on routing and bandwidth, the SP Switch router can be a real benefit. For more information, refer to 2.4.1, “SP Switch Router” on page 34.

3.5 Using NIS in a Cluster 1600 configuration

Network Information System (formerly called the “yellow pages” or YP, and now NIS). NIS provides a distributed database service for managing the important administrative files, such as the passwd file and the hosts file. NIS’s main purpose is to centralize administration of commonly replicated files, letting you make a single change to the database rather than making changes on every node within a network environment.

NIS separates a network into three components: domains, servers, and clients.

An NIS domain defines the boundary where file administration is carried out. In a large network, it is possible to define several NIS domains to break the machines up into smaller groups. This way, files meant to be shared among five machines, for example, stay within a domain that includes the five machines and not all the machines on the network.

An NIS server is a machine that provides the system files to be read by other machines on the network. There are two types of servers: master and slave. Both keep a copy of the files to be shared over the network. A master server is the machine where a file may be updated. A slave server only maintains a copy of the files to be served. A slave server has three purposes:

- To balance the load if the master server is busy.
- To back up the master server.
- To enable NIS requests if there are different networks in the NIS domain. NIS client requests are not handled through routers; such requests go to a local slave server. It is the NIS updates between a master and a slave server that go through a router.

An NIS client is a machine that has to access the files served by the NIS servers.

Since NIS is not directly related to a Cluster 1600 managed by PSSP configuration we have created an appendix that covers the basic NIS terms. Therefore, for more information about NIS and how to configure it, refer to Appendix B, “NIS” on page 537.

3.6 Using AFS® in a Cluster 1600 configuration

Andrew File System (AFS) and Transarc began development at Carnegie Mellon's Information Technology Center in the mid 80s. AFS is a distributed file system that enables sharing files across both local area and wide area networks. AFS includes an authentication implementation that is based on Kerberos V4. Therefore, AFS authentication servers can be used in place of a Cluster 1600 managed by PSSP authentication server to provide credentials for Kerberos V4 principals.

Although AFS and SP Kerberos V4 authentication services can work together, there are differences in the way the authentication database is administered (for example, adding principals and changing passwords).

AFS uses a different set of servers, protocols, interfaces, and commands for Kerberos V4 database administration.

Some useful features of AFS:

- In conjunction with Kerberos, AFS provides a global authentication system (all passwords are verified with a site-wide database).
- Access Control Lists (ACLs) provide more flexibility in setting file access permissions than traditional UNIX file systems.
- Users can access AFS files at remote sites, if given appropriate permissions.

Since AFS is not directly related to a Cluster 1600 managed by PSSP configuration, we created an appendix that covers the basic AFS terms. Therefore, for more information about AFS, refer to Appendix C, “AFS as a Cluster 1600 Kerberos-based security system” on page 545.

3.7 Related documentation

The following documentation will help you understand the concepts and examples covered in this guide.

SP manuals

PSSP Administration Guide, SA22-7348 covers “Using a switch”.

RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment, GA22-7281 covers “Planning your network configuration”.

PSSP Installation and Migration Guide, GA22-7347 covers “Initializing to use AFS authentication”.

SP redbooks

IBM (e)server Cluster 1600 Managed by PSSP 3.5: What's New, SG24-6617

IBM Cluster 1600 and PSSP 3.4 Cluster Enhancements, SG24-6604

Managing IBM (e)server Cluster 1600 - Power Recipes for PSSP 3.4, SG24-6603

3.8 Sample questions

This section provides questions to help you prepare for the certification exam. The answers can be found in Appendix A, "Answers to sample questions" on page 521.

1. The Cluster 1600 managed by PSSP requires an Ethernet connection between the CWS and all nodes. Which of the following tasks do *not* use the SP Ethernet admin LAN?
   a. Network installation
   b. System management
   c. Event monitoring
   d. Hardware control
2. Setting up host name resolution is essential to all the PSSP components. The name associated with the en0 interface is known as:
   a. Initial host name
   b. Reliable host name
   c. host name
   d. Primary name
3. What is the default order for resolving host names if /etc/resolv.conf is present?
   a. A. /etc/hosts → DNS → NIS
   b. B. DNS → NIS → /etc/hosts
   c. C. NIS → DNS → /etc/hosts
   d. D. NIS → /etc/hosts → DNS
4. In a possible scenario with a segmented 100BASE-TX network, the CWS is equipped with additional Ethernet adapters. Nodes in each separate segment will need:
   a. A boot/install server for that segment

b. A route to the control workstation

c. A default route set to one of the nodes or a router on that segment

d. All of the above

5. Consider an Ethernet network where none of the six subnets making up the SP Ethernet admin LAN have more than 16 nodes on them. How many discrete addresses per subnet does a netmask of 255.255.255.224 provide?

   a. 16

   b. 32

   c. 30

   d. 8

6. Which network adapter must be manually configured?

   a. Ethernet

   b. FDDI

   c. ATM

   d. Token Ring

7. The default order for resolving host names can be overwritten by creating a configuration file and specifying the desired order. Which of the following is the correct location and name of the configuration file?

   a. /etc/netservice.conf

   b. /netservice.conf

   c. /etc/netsvc.conf

   d. netsvc.conf

8. Which of the following daemons is *not* used by NIS?

   a. ypserv

   b. ypbind

   c. ypupdated

   d. yppassword

9. Which of the following statements is a characteristic of an NIS slave server?

   a. Backs up other slave servers.

   b. Balances the load if the primary slave server is busy.

   c. Enables NIS requests if there are different networks in the NIS domain.

   d. Disables NIS request if there are different networks in the NIS domain.

3.9 Exercises

Here are some exercises and questions to further help you prepare:

1. On a test system that does not affect any users, practice setting up new static routes using the command line.

2. Which commands can be used to configure the SP Ethernet for the nodes in the SDR?

3. Which netmask can be used for the study test guide environment on page 3? What happens to the netmask if we add a third Ethernet segment to the environment?

4. On a test system that does not affect any users, use the environment variable `nsorder` to change the default order for resolving host names.

5. On a test system that does not affect any users, configure NIS.

4

I/O devices and file systems

This chapter provides an overview of internal and external I/O devices and how they are supported in the Cluster 1600 environments. It also covers a discussion on file systems and their utilization in the RS/6000 SP.

4.1 Key concepts you should study

Before taking the certification exam, make sure you understand the following concepts:

- Support for external I/O devices
- Possible connections of I/O devices, such as SCSI, RAID, and SSA
- Network File System (NFS)—how it works and how it is utilized in the RS/6000 SP, especially for installation
- The AFS and DFS™ file systems and their potential in RS/6000 SP environments

4.2 I/O devices

Anything that is not memory or CPU can be considered an input/output device (I/O device). I/O devices include internal and external storage devices as well as communications devices, such as network adapters, and, in general, any devices that can be used for moving data.

4.2.1 External disk storage

If external disk storage is part of your system solution, you need to decide which of the external disk subsystems available for the SP best satisfy your needs.

Disk options offer the following tradeoffs in price, performance, and availability:

- For availability, you can use either a RAID subsystem with RAID 1 or RAID 5 support, or you can use mirroring.
- For best performance when availability is needed, you can use mirroring or RAID 1, but these require twice the disk space.
- For low cost and availability, you can use RAID 5, but there is a performance penalty for write operations, One write requires four I/Os: A read and a write to two separate disks in the RAID array. An N+P (parity) RAID 5 array, comprised of N+1 disks, offers N disks worth of storage; therefore, it does not require twice as much disk space.

  Also, use of RAID 5 arrays and hot spares affect the relationship between *raw storage* and *available and protected storage*. RAID 5 arrays, designated in the general case as N+P arrays, provide N disks worth of storage. For example, an array of eight disks is a 7+P RAID 5 array providing seven disks worth of available protected storage. A hot spare provides no additional usable storage but provides a disk that quickly replaces a failed disk in the

RAID 5 array. All disks in a RAID 5 array should be the same size; otherwise, disk space will be wasted.

- For low cost when availability due to disk failure is not an issue, you can use what is known as JBOD (Just a Bunch of Disk).

After you choose a disk option, be sure to get enough disk drives to satisfy the I/O requirements of your application, taking into account whether you are using the Recoverable Virtual Shared Disk optional component of PSSP, mirroring, or RAID 5, and whether I/O is random or sequential.

Table 4-1 has more information on disk storage choices.

Table 4-1 Disk storage available for pSeries

Disk Storage	Description
2102	The IBM Fibre Channel RAID Storage Server is an entry-level to mid-range solution for customers having up to two servers requiring from 100 GB to 2 terabytes of RAID protected storage. It provides mission critical, easy to manage, high availability storage with redundancy in the controllers, power supplies, cooling units, and RAID disk storage. Customers with storage needs that are supported across multiple server platforms that require Fibre Channel attachment should consider the Fibre Channel RAID Storage Server. The IBM Enterprise Storage Server® and the IBM 7133 Serial Disk Systems should be considered for large enterprise-wide heterogeneous platforms, or for RS/6000 clustered environments where appropriate adapters may offer the right performance.
2104	The 2104 Expandable Storage Plus (Exp Plus) disk enclosure line provides greater capacity, more connectivity options (including Ultra3 LVD), and higher performance -- all at an affordable price. If you rely on eServer pSeries and RS/6000 servers and require external SCSI disk capacity, this may be the disk enclosure you've been waiting for: Up to 1TB of internal disk space accommodated in either one SCSI bus with all 14 disk drives populated (9.1, 18.2, 36.4 and 73.4GB disk drives), or in a dual host environment with a split SCSI bus and up to 7 SCSI drives in each bus.
2105	The 2105 Enterprise Storage Server family (ESS) offers up to 48 9.1 GB, 18.2 GB, 36.4 GB, or 72.8 GB disk eight-packs providing up to 22.4 TB of RAID-5 effective capacity. Support for the intermix of FICON™ (F10, F20), ESCON®, Fibre Channel and SCSI attachments, makes the ESS a natural fit for storage and server consolidation requirements.

Disk Storage	Description
7207	The IBM 7027 High Capacity Storage Drawer provides up to 30 disk bays plus 3 media bays in a rack package. In its base configuration, the 7027 comes with four 2.2 GB disk drives installed and twenty bays available. Six more disk bays plus three media bays can be added, offering a total capacity of up to 30 1-inch high disk drives (2.2 GB or 4.5 GB) or 15 of the 1.6-inch (9.1 GB) disk drives per drawer. Maximum capacity is increased to 136.5 GB of external storage when fully populating the drawer with the 9.1 GB disk drives. All supported disk drives are hot-swappable.The optional media bays can be filled with various tape and CD-ROM features.
7131	There are two different 7131 models to provide either SCSI-2 or SSA attachment. Up to five internal disks from 2.2 GB to 9.1 GB can be installed. These disks are hot-plug disks. Two additional media slots for CD-ROM, tape drive, or disks can be populated, but they are not hot-swap. When these non-hot-swap internal media bays are populated with disks, a total of 63.7 GB can be achieved.
7133	If you require high performance, the 7133 Serial Storage Architecture (SSA) Disk might be the subsystem for you. SSA provides better interconnect performance than SCSI and offers hot-pluggable drives, cables, and redundant power supplies. RAID 5, including hot spares, is supported on some adapters, and loop cabling provides redundant data paths to the disk. Two loops of up to 48 disks are supported on each adapter. However, for best performance of randomly accessed drives, you should have only 16 drives (one drawer or 7133) in a loop. The maximum internal storage is 582 GB.
7137	The 7137 subsystem supports both RAID 0 and RAID 5 modes. It can hold from 4 to 70 GB of data (61.8 GB maximum in RAID 5 mode). The 7137 is the low end model of RAID support. Connection is through SCSI adapters. If performance is not critical, but reliability and low cost are important, this is a good choice
3542 FaStT200	The IBM 3542 FAStT200 Storage Servers are the second generation of FAStT solutions to provide additional cost-effective offerings where moderate levels of performance, high availability, and scalability are needed. With either one or two hot-plug RAID adapters and up to 10 internal disks with up to 1.46 TB total disk space, this Fibre Channel-attached storage device offers scalability and high availability. The maximum distance with Fibre Channel attach can be 10 km.
3552 FaStT500	The IBM 3552 TotalStorage® FAStT500 Storage Server Model 1RU supports up to 220 disk drives with a capacity of up to 16 TBs; RAID 0, 1, 3, 5, and 10; 512 MB write-back cache, 256 MB per controller; with battery backup standard.

Disk Storage	Description
1742 FaStT700	The fibre array storage technology (FAStT) family of products is ideal for creating your storage area network (SAN), whether you are just starting out or are expanding your existing network. RAID levels 0, 1, 3, 5, and 10 are supported. The FAStT products are very scalable (from 36 GB to 16 TB), easy to install and manage, and provide world class performance and reliability, all at a very attractive price.

In summary, to determine what configuration best suits your needs, you must be prepared with the following information:

- The amount of storage space you need for your data
- A protection strategy (mirroring, RAID 5), if any
- The I/O rate you require for storage performance
- Any other requirements, such as multi-host connections, or whether you plan to use the Recoverable Virtual Shared Disk component of PSSP, which needs twin-tailed disks

You can find up-to-date information about the available storage subsystems on the Internet at:

`http://www.storage.ibm.com`

Especially for the pSeries, go to:

`http://www.storage.ibm.com/products_pseries.html`

Figure 4-1 on page 128 shows external device configurations that can be connected to an SP system.

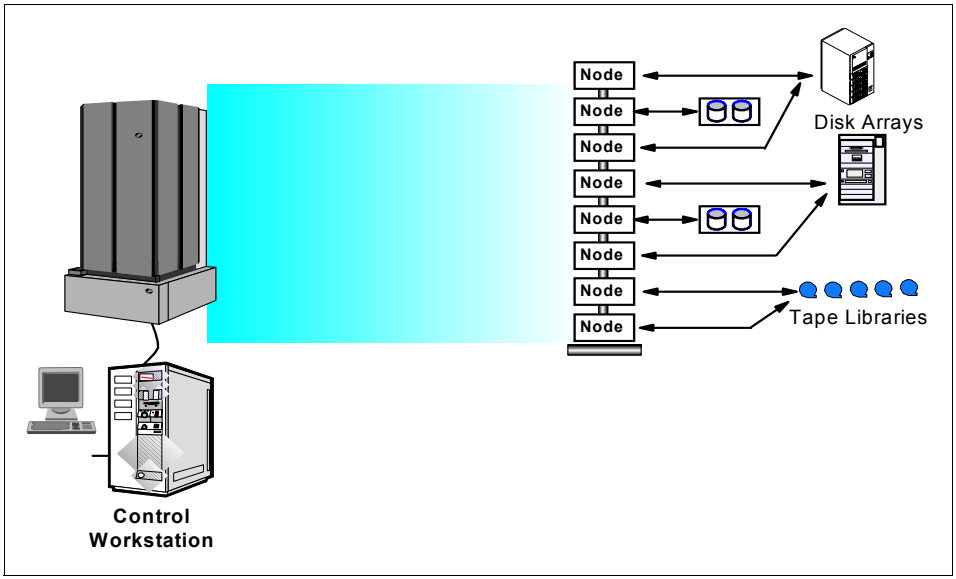


Figure 4-1 External devices

4.2.2 Internal I/O adapters

For the Cluster 1600-supported machines, several PCI adapters are available. These adapters are for communication and storage. The latest machines also make it possible to hot-plug the adapters, which enables you to replace, install, or remove an adapter while the machine is up and running. For a detailed overview of all the current PCI adapters available for any of the SP-attached servers or SP nodes, refer to *RS/6000 and IBM @server pSeries PCI Adapter Placement Reference,* SA38-0538 and to *RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment,* GA22-7280.

4.3 Multiple rootvg support

The concept called *Multiple Rootvg* or *Alternate Root Volume Group* provides the ability to boot a separate volume group on a node. To do this, an SDR class called Volume_Group was created in PSSP 3.1 to store the data. These additional volume groups allow booting of a separate version of the operating system on the node. Obviously, before using this alternative, you must do as many installations as you need. Each installation uses a different Volume_Group name created at the SDR level.

Multiple rootvg requirements

One way to significantly increase the availability of your SP system is to establish redundant copies of the operating system image using the disk mirroring feature of AIX. Mirroring the root volume group means that there will be multiple copies of the operating system image available to a workstation or node. Mirrored root volume groups are set up such that if a disk in the root volume group fails, the system will continue to run without interruption to your application.

We encourage you to mirror your root volume group. When you install a node, you choose how many copies to make of the root volume group. AIX supports one (the original), two (the original plus a mirror), or three (the original plus two mirrors) copies of a volume group. We recommend that you mirror the root volume group by choosing at least two copies.

PSSP provides commands to facilitate mirroring on the SP system. To mirror a root volume group, you need one disk or set of disks for each copy, depending on how many are needed to contain one complete system image. The POWER3 High Nodes are delivered with disk pairs as a standard feature. The extra disk should be used for a mirror of the root volume group when you install your system. For more information, refer to *PSSP Administration Guide,* SA22-7348.

Mirrored rootvg details

Although the name of these volume groups must be different in the SDR because they are different objects in the same class (the first one can be rootvg and the following othervg, for example), this name stays in the SDR and is not used directly by NIM to install the node. Only the attribute Destination Disks is used to create the rootvg node volume group.

If your node has two (or more) available rootvgs, only one is used to boot. It is determined by the bootlist of the node. Because the user determines which version of the operating system to boot, another concept appears with PSSP 3.1: the possibility to change the bootlist of a node directly from the CWS by using the new **`spbootlist`** command.

The operating system determines which copy of each operating system's logical volume is active, based on availability.

Prior to PSSP 3.1, the RS/6000 SP attributes, such as operating system level, PSSP level, installation time, and date, were associated with the Node object in the SDR.

In PSSP 3.1, or later, these attributes are more correctly associated with a volume group. A node is not at AIX 4.3.2, for example; a volume group of the node is at AIX 4.3.2. To display this information, a new option (**`-v`**) was added in the **`splstdata`** command.

Therefore, part of this feature is to break the connection between nodes and attributes more properly belonging to a volume group. For this reason, some information has been moved from the SMIT panel Boot/Install Server Information to the Create Volume Group Information or the Change Volume Group Information panel.

We now describe these features and the related commands in more detail.

4.3.1 The Volume_Group class

As explained, a new Volume_Group class has been created in PSSP 3.1. The following is a list of attributes:

- node_number
- vg_name (volume group name)
- pv_list (one or more physical volumes)
- quorum (quorum is true or false)
- copies (1, 2, or 3)
- install_image (name of the mksysb)
- code_version (PSSP level)
- lppsource_name (which lppsource)
- boot_server (which node serves this volume group)
- last_install_time (time of last install of this volume group)
- last_install_image (last mksysb installed on this volume group)
- last_bootdisk (which physical volume to boot from)

The attributes pv_list, install_image, code_version, lppsource_name, and boot_server have been duplicated from the Node class to the Volume_Group class. New SMIT panels associated with these changes are detailed in the following sections.

The node object

The new Volume_Group class uses some attributes from the old node class. The following list describes the changes made to the Node object:

- A new attribute is created: selected_vg.
- selected_vg points to the current Volume_Group object.
- The node object retains all attributes.

- Now the node attributes common to the Volume_Group object reflect the current volume group of the node.
- The Volume_Group objects associated with a node reflect all the possible volume group states of the node.

Note: All applications using the node object remain unchanged with the exception of some SP installation code.

File system planning

Plan ahead for expected growth of all your file systems. Also, monitor your file system growth periodically and adjust your plans when necessary. When the AIX 5L 32-bit kernel is enabled, the default file system is JFS. When the AIX 5L 64-bit kernel is enabled, the default file system is JFS2. It is possible to switch kernels but maintain the original file system. If a kernel switch is done, existing file systems remain the same, but all new ones are created as JFS for the 32-bit kernel and JFS2 for the 64-bit kernel.

Volume_Group default values

When the SDR is initialized, a Volume_Group object for every node is created.

By default, the vg_name attribute of the Volume_Group object is set to rootvg, and the selected_vg of the node object is set to rootvg.

The other default values are as follows:

- The default install_disk is hdisk0.
- Quorum is true.
- Mirroring is off; copies are set to 1.
- There are no bootable alternate root volume groups.
- All other attributes of the Volume_Group are initialized according to the same rules as the node object.

4.3.2 Volume group management commands

After describing the new volume group management features available in PSSP 3.1 or later, let us now describe the commands used to create, change, delete, mirror, and unmirror Volume_Group objects. Also, changes to existing commands in previous PSSP versions (previous to PSSP 3.1) are described.

spmkvgobj

All information needed by NIM, such as lppsource, physical disk, server, mksysb, and so forth, is now moved from Boot/Install server Information to a new panel accessible by the fast path **createvg_dialog** as shown in Figure 4-2.

```
Create Volume Group Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Start Frame                                        [] #
  Start Slot                                         [] #
  Node Count                                         [] #

  OR

  Node List                                          [10]

  Volume Group Name                                  [rootvg]
  Physical Volume List                               [hdisk0,hdisk1]
  Number of Copies of Volume Group                    1 +
  Boot/Install Server Node                           [0] #
  Network Install Image Name
```

Figure 4-2 New SMIT panel to create a volume group

The associated command of this SMIT panel is **spmkvgobj**, whose options are:

- -r vg_name
- -l node_list
- -h pv_list
- -i install_image
- -v lppsource_name
- -p code_version
- -n boot_server
- -q quorum
- -c copies

The following command built by the previous SMIT panel is a good example of the use of **spmkvgobj**:

```
/usr/lpp/ssp/bin/spmkvgobj -l '10' -r 'rootvg' -h 'hdisk0,hdisk1' -n '0' -i
'bos. obj.mksysb.aix432.090898' -v 'aix432' -p 'PSSP-3.1'
```

Here is more information about the **-h** option: For PSSP levels prior to PSSP 3.1, two formats were supported to specify the SCSI disk drive and are always usable:

- Hardware location format

  00-00-00-0,0 to specify a single SCSI disk drive, or
  00-00-00-0,0:00-00-00-1,0 to specify multiple hardware locations (in that case, the colon is the separator).

- Device name format

  hdisk0 to specify a single SCSI disk drive, or hdisk0, hdisk1 to specify multiple hardware locations (in that case, the comma is the separator).

  You must not use this format when specifying an external disk because the relative location of hdisks can change depending on what hardware is currently installed. It is possible to overwrite valuable data by accident.

A third format is now supported to be able to boot on SSA external disks, which is a combination of the parent and *connwhere* attributes for SSA disks from the Object Data Management (ODM) CuDv. In the case of SSA disks, the parent always equals ssar. The connwhere value is the 15-character unique serial number of the SSA drive (the last three digits are always 00D for a disk). This value is appended as a suffix to the last 12 digits of the disk ID stamped on the side of the drive. If the disk drive has already been defined, the unique identity may be determined using SMIT panels or by following these two steps:

1. Issue the command:

   ```
   lsdev -Ccpdisk -r connwhere
   ```

2. Select the 15-character unique identifier whose characters 5 to 12 match those on the front of the disk drive.

   For example, to specify the parent-connwhere attribute, you can enter:

   ```
   ssar//0123456789ABOOD
   ```

   Or, to specify multiple disks, separate using colons as follows:

   ```
   ssar//0123456789ABOOD:ssar//0123456789FGOOD
   ```

Important: The *ssar* identifier must have a length of 21 characters.Installation on external SSA disks is supported in PSSP 3.1 or later.

spchvgobj

After a Volume_Group has been created by the **spmkvgobj** command, you may want to change some information. Use the **spchvgobj** command or the new SMIT panel (fastpath is changevg_dialog) shown in Figure 4-3.

Note: In general, quorum should be turned off in a mirrored environment and turned on in an environment without mirroring. Quorum is turned on by default during node installation unless you specify **-q** false with the **spchvgobj** command before installation when you first set the mirroring options.

This command uses the same options as the **spmkvgobj** command. The following is an example built by the SMIT panel:

```
/usr/lpp/ssp/bin/spchvgobj -l '1' -r 'rootvg' -h 'hdisk0,hdisk1,hdisk2' -c
'2' -p 'PSSP-3.1'
```

```
Change Volume Group Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Start Frame                                        [] #
  Start Slot                                         [] #
  Node Count                                         [] #

  OR

  Node List                                          [1]

  Volume Group Name                                  [rootvg]
  Physical Volume List                               [hdisk0,hdisk1,hdisk2]
  Number of Copies of Volume Group                    2 +
  Set Quorum on the Node +
  Boot/Install Server Node                           [] #
  Network Install Image Name                         []
```

Figure 4-3 New SMIT panel to modify a volume group

Note: To verify the content of the Volume_Group class of node 1, you can issue the following SDR command:

```
SDRGetObjects Volume_Group node_number==1 vg_name pv_list copies.
```

sprmvgobj

To be able to manage the Volume_Group class, a third command to remove a Volume_Group object that is not the current one has been added: **sprmvgobj**.

This command accepts the following options:

```
-r vg_name
-l node_list
```

Regarding SMIT, the Delete Database Information SMIT panel has been changed to access the new SMIT panel named Delete Volume Group Information (the fastpath is deletevg_dialog).

Refer to Figure 4-4 for details.

```
Delete Volume Group Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Start Frame                                        []
#
  Start Slot                                         []
#
  Node Count                                         []
#

  OR

  Node List                                           [1]
```

Figure 4-4 New SMIT panel to delete a volume group

The following is an example built by the SMIT panel used in Figure 4-4:

```
/usr/lpp/ssp/bin/sprmvgobj -l '1' -r 'rootvg2'
```

Use of spbootins in PSSP

The **spbootins** command sets various node attributes in the SDR (code_version, lppsource_name, and so forth).

With the **spbootins** command, you can select a volume group from all the possible volume groups for the node in the Volume_Group class.

Attributes shared between the node and Volume_Group objects are changed using a new set of Volume_Group commands, not by using **spbootins**.

The **spbootins** command is as follows:

```
spbootins
   -r <install|diag|maintenance|migrate|disk|customize>
   -l <node_list>
   -c <selected_vg>
   -s <yes|no>
```

Figure 4-5 shows the new SMIT panel to issue **spbootins** (the fastpath is server_dialog).

```
Boot/Install Server Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Start Frame                                        [] #
  Start Slot                                         [] #
  Node Count                                         [] #

  OR

  Node List                                          [10]

  Response from Server to bootp Request               install +
  Volume Group Name                                  [rootvg]
  Run setup_server?                                   yes +
```

Figure 4-5 New SMIT panel to issue the spbootins command

You get the same result by issuing the following from the command line:

```
spbootins -l 10 -r install -c rootvg -s yes
```

Note that the value `yes` is the default for the `-s` option; in this case, the script setup_server is run automatically.

spmirrorvg

This command enables mirroring on a set of nodes given by the option:

```
-l node_list
```

You can force (or not force) the extension of the volume group by using the `-f` option (available values are: `yes` or `no`).

This command takes the volume group information from the SDR updated by the last **`spchvgobj`** and **`spbootins`** commands.

Note: You can add a new physical volume to the node rootvg by using the **`spmirrorvg`** command; the following steps give the details:

1. Add a physical disk to the actual rootvg in the SDR by using **`spchvgobj`** without changing the number of copies.
2. Run **`spmirrovg.`**

Figure 4-6 shows the new SMIT panel to issue **`spmirrorvg`** (the fastpath is start_mirroring).

```
 Initiate Mirroring on a Node

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Start Frame                                        []
#
  Start Slot                                         []
#
  Node Count                                         []
#

  OR

  Node List                                           [1]

  Force Extending the Volume Group?                    no             +
```

Figure 4-6 New SMIT panel to initiate the spmirrorvg command

The following is an example built by the SMIT panel in Figure 4-6:

```
/usr/lpp/ssp/bin/spmirrorvg -l '1''
```

For more detail regarding the implementation of mirroring root volume groups, refer to Appendix B of *PSSP Administration Guide,* SA22-7348.

Note: This command uses the **dsh** command to run the AIX-related commands on the nodes.

spunmirrorvg

This command disables mirroring on a set of nodes given by the option:

```
-l node_list
```

Figure 4-7 shows the new SMIT panel to issue **spunmirrorvg** (the fastpath is stop_mirroring).

```
Discontinue Mirroring on a Node

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Start Frame                                  []
#
  Start Slot                                   []
#
  Node Count                                   []
#

  OR
```

Figure 4-7 New SMIT panel to initiate the spunmirrorvg command

The following is the example built by the SMIT panel in Figure 4-7:

```
/usr/lpp/ssp/bin/spunmirrorvg -l '1''
```

Note: This command uses the **dsh** command to run the AIX-related commands on the nodes.

Use of splstdata in PSSP

splstdata can display information about Volume_Groups using the option **-v**.

Figure 4-8 on page 139 shows the information related to node 1 in the result of the command:

```
splstdata -v -l 1
```

```
List Volume Group Information

node# name            boot_server quorum copies   code_version
lppsource_name
         last_install_image       last_install_time   last_bootdisk
         pv_list
------------------------------------------------------------------------------
---
    1 rootvg           0          true      1         PSSP-3.1 aix432

         default                  Thu_Sep_24_16:47:50_EDT_1998 hdisk0

         hdisk0
    1 rootvg2          0          true      1         PSSP-3.1 aix432

         default                  Fri_Sep_25_09:16:44_EDT_1998 hdisk3
```

Figure 4-8 Example of splstdata -v

Determine your authentication methods using the command:

```
splstdata -p
```

Refer to Example 4-1 for the **splstdata -p** output.

Example 4-1 splstdata -p output

```
[c179s][/]> splstdata -p
List System Partition Information

System Partitions:
------------------
c179s

Syspar: c179s
------------------------------------------------------------------------------
syspar_name     c179s
ip_address      9.114.12.81
install_image   default
syspar_dir      ""
code_version    PSSP-3.4
haem_cdb_version 1048258006,608471295,0
auth_install    k4:std
auth_root_rcmd  k4:std
```

```
ts_auth_methods compat
auth_methods    k4:std
```

spbootlist

spbootlist sets the bootlist on a set of nodes by using the option:

```
-l node_list
```

This command takes the volume group information from the SDR updated by the last **spchvgobj** and **spbootins** commands.

See 4.3, “Multiple rootvg support” on page 128 for information on how to use this new command.

4.3.3 How to declare a new rootvg

Several steps must be done in the right order; they are the same as for an installation. The only difference is that you must enter an unused volume group name.

The related SMIT panel or commands are given in Figure 4-2 on page 132 and Figure 4-5 on page 136.

At this point, the new volume group is declared, but it is not usable. You must now install it using a Network Boot, for example.

How to activate a new rootvg

Several rootvgs are available on your node. To activate one of them, the bootlist has to be changed by using the **spbootlist** command or the related SMIT panel (the fastpath is bootlist_dialog) as shown in Figure 4-9 on page 141. Because the **spbootlist** command takes information from the node boot information given by **splstdata -b**, this information has to be changed by issuing the **spbootins** command. Once the change is effective, you can issue the **spbootlist** command.

Verify your node bootlist with the command:

```
dsh -w <node> 'bootlist -m normal -o'
```

Then reboot the node.

The following example shows the steps necessary to activate a new rootvg on node 1 (hostname is node01). We assume two volume groups (rootvg1 and rootvg2) have already been installed on the node. rootvg1 is the active rootvg.

1. Change the node boot information:

   ```
   spbootins -l 1 -c rootvg2 -s no
   ```

   Note: it is not necessary to run **setup_server**.

2. Verify:

   ```
   splstdata -b
   ```

3. Change the node bootlist:

   ```
   spbootlist -l 1
   ```

4. Verify:

   ```
   dsh -w node01 'bootlist -m normal -o'
   ```

5. Reboot the node:

   ```
   dsh -w node01 'shutdown -Fr'
   ```

Important: MCA nodes only: The key switch must be in the normal position.

```
Set Bootlist on Nodes

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                    [Entry Fields]
  Start Frame                                     [] #
  Start Slot                                      [] #
  Node Count                                      [] #

  OR

  Node List                                       []
```

Figure 4-9 SMIT panel for the spbootlist command

4.3.4 Booting from external disks

Support has been included in PSSP 3.1 for booting an SP node from an external disk. The disk subsystem can be either external Serial Storage Architecture (SSA) or external Small Computer Systems Interface (SCSI). The option to have an SP node without an internal disk storage device is now supported.

SSA disk requirements

Figure 4-10 and Figure 4-11 on page 143 show the SSA disk connections to a node.

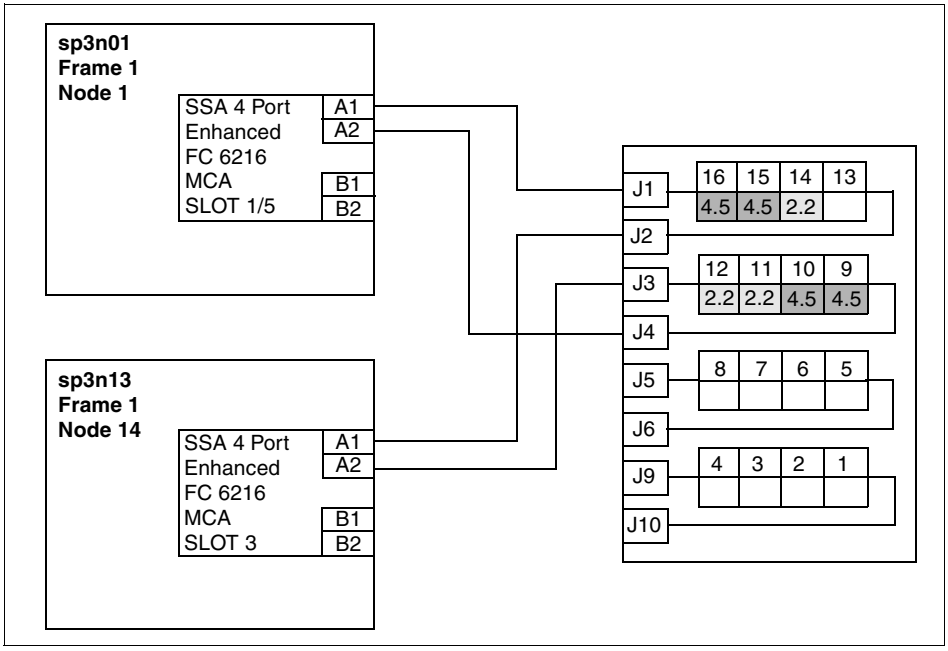


Figure 4-10 Cabling SSA disks to RS/6000 SP nodes

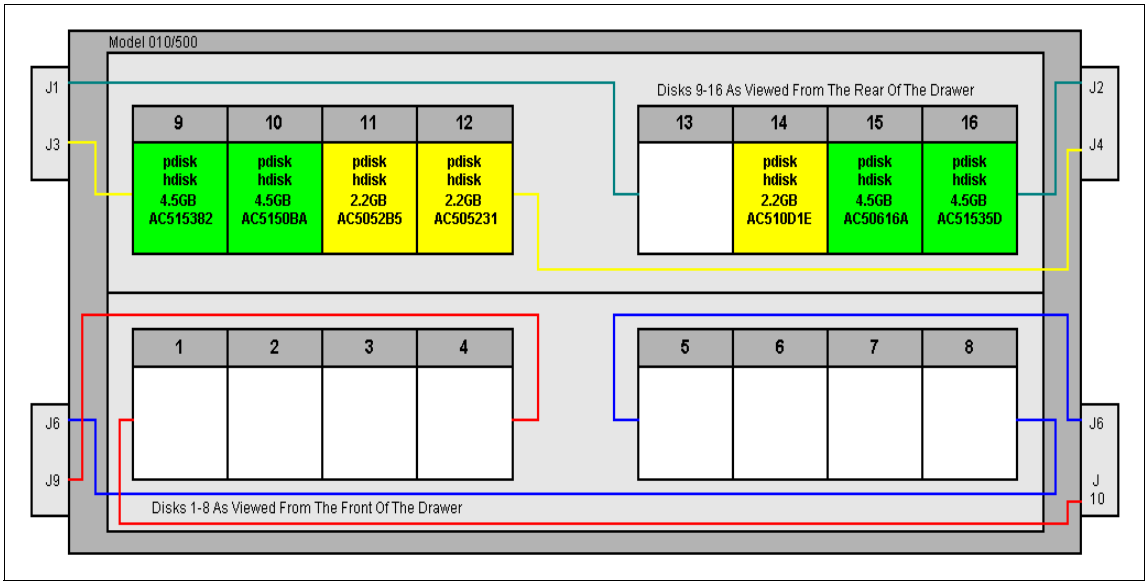


Figure 4-11 Connections on the SSA disks

All currently supported PCI-based Cluster 1600 node types can support an SSA boot. The adapter Advanced SerialRAID plus with feature code #6230 is the latest supported adapter. This adapter has a higher speed with up to 160 MB/s on all four SSA ports. An optical extender can also be used to attach systems with fiber cables that a further away. Only the latest 7133 Models D40 and T40 support the 40 MB/s per SSA port speed. Older systems may run with this adapter but only at 20 MB/s per port.

The SP-supported external SSA disk subsystems are:

- 7133 IBM Serial Storage Architecture Disk Subsystems Models D40 and T40
- 7133 IBM Serial Storage Architecture Disk Subsystems Models 010, 020, 500, and 600 (withdrawn from marketing)

SCSI disk requirements

The SP nodes and SP-attached servers can be booted from an external SCSI storage device such as M/T 2104. Table 4-2 on page 144 lists the nodes and the adapters for external disk booting.

Table 4-2 Available SCSI adapters for SCSI boot

	375/450 MHz POWER3 thin/wide node	332 MHz thin/wide node	POWER3 high node	375 MHz POWER3 high node
Feature #6203 Ultra3 SCSI	Yes	Yes	No	Yes
Feature #6204 Ultra SCSI DE	Yes	Yes	Yes	Yes
Feature #6205 Dual Channel Ultra2 SCSI	Yes	Yes	Yes	Yes
Feature #6206 Ultra SCSI SE	Yes	Yes	Yes	Yes
Feature #6207 Ultra SCSI DE	Yes	Yes	Yes	Yes
Feature #6208 SCSI-2 F/W SE	Yes	Yes	No	No
Feature #6209 SCSI-2 F/W DE	Yes	Yes	No	No

For more detailed information about supported PCI adapters, refer to *RS/6000 and IBM @server pSeries PCI Adapter Placement Reference,* SA38-0538.

Specifying an external installation disk

During the node installation process, external disk information may be entered in the SDR by first typing the SMIT fastpath `smitty node_data`. Depending on whether you have already created the Volume_Group, you must then choose Create Volume Group Information or Change Volume Group Information from the Node Database Information Window (related commands are spmkvgobj or spchvgobj). Alternatively, you may use the SMIT fastpath smitty changevg_dialog (refer to Figure 4-3 on page 134**)** to get straight there.

Figure 4-12 on page 145 shows the Change Volume Group Information window. In this, the user is specifying an external SSA disk as the destination for rootvg on node1. Note that you may specify several disks in the Physical Volume List field (refer to “spmkvgobj” on page 132 for more information on how to enter the data).

```
   Change Volume Group Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                  [Entry Fields]
  Start Frame                                        []
#
  Start Slot                                         []
#
  Node Count                                         []
#

  OR

  Node List                                           [1]

  Volume Group Name                                   [rootvg]
  Physical Volume List                                [ssar//0004AC50532100D]
```

Figure 4-12 SMIT panel to specify an external disk for SP node installation

When you press the Enter key in the Change Volume Group Information window, the external disk information is entered in the Node class in the SDR. This can be verified by running the **splstdata -b** command as shown in Figure 4-13 on page 146. This shows that the install disk for node 1 has been changed to ssar//0004AC50532100D.

Under the covers, `smitty changevg_dialog runs the` **spchvgobj** command. This command recognizes the external disk address formats. It may be run directly from the command line using this syntax:

```
spchvgobj -r rootvg -h ssar//0004AC50532100D -l 1
```

```
sp3en0{ / } splstdata -b -l 1

                  List Node Boot/Install Information


node#         hostname  hdw_enet_addr srvr     response
install_disk
      last_install_image   last_install_time  next_install_image
lppsource_name
                pssp_ver         selected_vg
------------------------------------------------------------------------------
---
    1 sp3n01.msc.itso.   02608CE8D2E1    0      install
ssar//0004AC510D1E00D
                default             initial             default
aix432
```

Figure 4-13 Output of the splstdata -b command

Changes to the bosinst.data file

When the changes have been made to the Node class in the SDR to specify an external boot disk, the node can be set to *install* with the **spbootins** command:

```
spbootins -s yes -r install -l 1
```

The **setup_server** command will cause the network install manager (NIM) wrappers to build a new bosinst.data resource for the node, which will be used by AIX to install the node.

The format of bosinst.data has been changed in AIX 4.3.2 to include a new member to the target_disk stanza specified as CONNECTION=. This is shown in Figure 4-14 on page 147 for node 1's bosinst.data file (node 1 was used as an example node in Figure 4-12 on page 145 and Figure 4-14 on page 147). NIM puts in the CONNECTION= member when it builds the file.

```
control_flow:
    CONSOLE = /dev/tty0
    INSTALL_METHOD = overwrite
    PROMPT = no
    EXISTING_SYSTEM_OVERWRITE = yes
    INSTALL_X_IF_ADAPTER = no
    RUN_STARTUP = no
    RM_INST_ROOTS = no
    ERROR_EXIT =
    CUSTOMIZATION_FILE =
    TCB = no
    INSTALL_TYPE = full
    BUNDLES =

target_disk_data:
    LOCATION =
    SIZE_MB =
    CONNECTION = ssar//0004AC505321OOD

locale:
    BOSINST_LANG = en_US
    CULTURAL_CONVENTION = en_US
    MESSAGES = en_US
    KEYBOARD = en_US
```

Figure 4-14 bosinst.data file with the new CONNECTION attribute

4.4 Global file systems

This section gives an overview of the most common *global* file systems. A global file system is a file system that resides locally on one machine (the file server) and is made globally accessible to many clients over the network. All file systems described in this section use UDP/IP as the network protocol for client/server communication (NFS Version 3 may also use TCP).

One important motivation to use global file systems is to give users the impression of a single system image by providing their home directories on all the machines they can access. Another is to share common application software that then needs to be installed and maintained in only one place. Global file systems can also be used to provide a large scratch file system to many machines, which normally utilizes available disk capacity better than distributing the same disks to the client machines and using them for local scratch space.

However, the latter normally provides better performance; so, a trade-off has to be made between speed and resource utilization.

Apart from the network bandwidth, an inherent performance limitation of global file systems is the fact that one file system resides completely on one machine. Different file systems may be served by different servers, but access to a single file, for example, will always be limited by the I/O capabilities of a single machine and its disk subsystems. This might be an issue for parallel applications where many processes/clients access the same data. To overcome this limitation, a *parallel* file system has to be used. IBM's parallel file system for the SP is described in 11.6, "IBM General Parallel File System (GPFS)" on page 384.

4.4.1 Network File System (NFS)

Sun Microsystem's Network File System (NFS) is a widely used global file system, available as part of the base AIX operating system. It is described in detail in Chapter 10, "Network File System" of *AIX Version 4.3 System Management Guide: Communications and Networks,* SC23-4127.

In NFS, file systems residing on the NFS server are made available through an *export* operation either automatically when the NFS start-up scripts process the entries in the /etc/exports file or explicitly by invoking the **exportfs** command. They can be mounted by the NFS clients in three different ways. A *predefined* mount is specified by stanzas in the /etc/filesystems file, an *explicit* mount can be performed by manually invoking the **mount** command, and *automatic* mounts are controlled by the **automount** command, which mounts and unmounts file systems based on their access frequency. This relationship is sketched in Figure 4-15 on page 149.

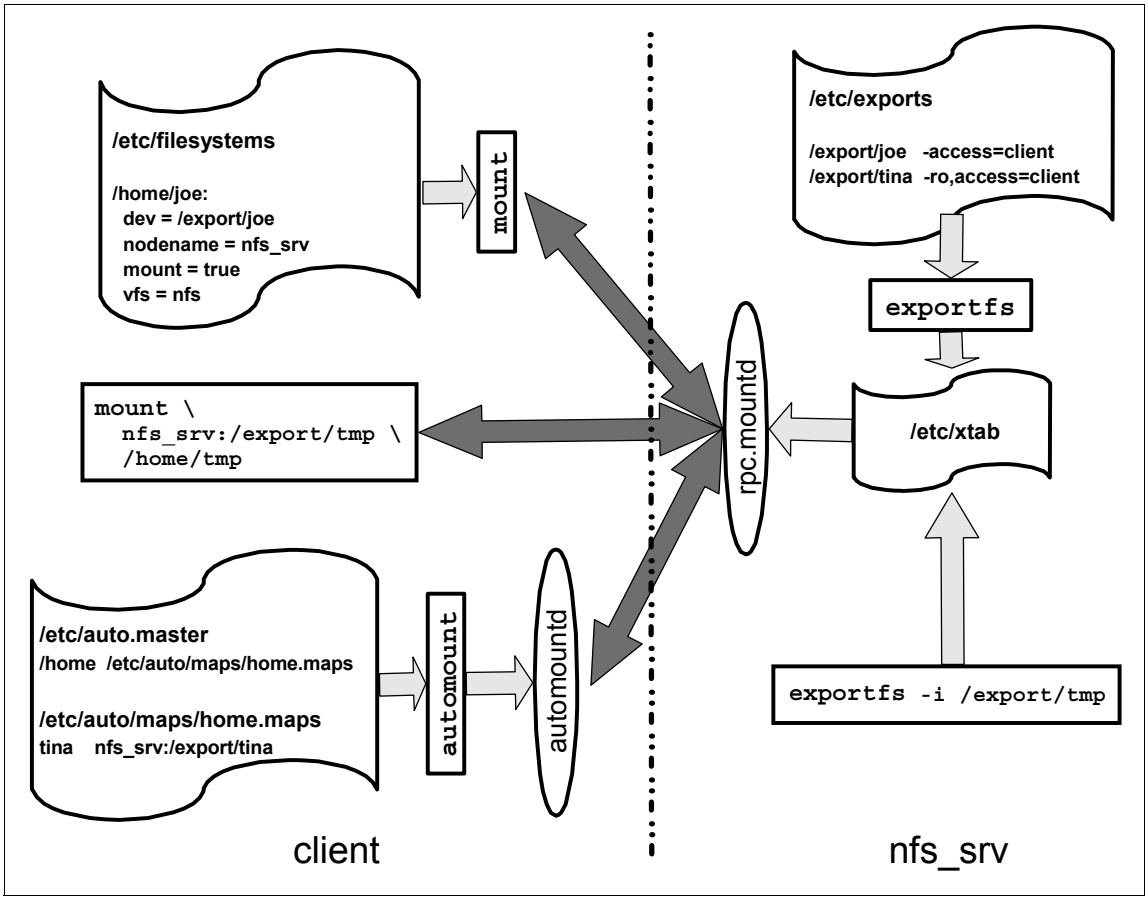


Figure 4-15 Conceptual overview of the NFS mounting process

The PSSP software uses NFS for network installation of the SP nodes. The CWS and boot/install servers act as NFS servers to make resources for network installation available to the nodes, which perform explicit mounts during installation. The SP accounting system also uses explicit NFS mounts to consolidate accounting information.

NFS is often used operationally to provide global file system services to users and applications. Among the reasons for using NFS is the fact that it is part of base AIX, it is well-known in the UNIX community, very flexible, and relatively easy to configure and administer in small to medium-sized environments. However, NFS also has a number of problems. They are summarized below to provide a basis for comparing NFS to other global file systems.

Performance NFS Version 3 contains several improvements over NFS Version 2. The most important change is that NFS Version 3 no longer limits the buffer size to 8 KB, thus improving its performance over high bandwidth networks. Other optimizations include the handling of file attributes and directory lookups and increased write throughput by collecting multiple write requests and writing the collective data to the server in larger requests.

Security Access control to NFS files and directories is by UNIX mode bits, that means by UID. Any root user on a machine that can mount an NFS file system can create accounts with arbitrary UIDs and, therefore, can access all NFS-mounted files. File systems may be exported read-only if none of the authorized users need to change their contents (such as directories containing application binaries), but home directories will always be exported with write permissions, as users must be able to change their files. An option for secure NFS exists, but is not widely used. Proprietary access control lists (ACLs) should not be used since not all NFS clients understand them.

Management A file system served by an NFS server cannot be moved to another server without disrupting service. Even then, clients mount it from a specific IP name/address and will not find the new NFS server. On all clients, references to that NFS server have to be updated. To keep some flexibility, alias names for the NFS server should be used in the client configuration. These aliases can then be switched to another NFS server machine should this be necessary.

Namespace With NFS, the client decides at which local mount point a remote file system is mounted. This means that there are no global, universal names for NFS files or directories since each client can mount them to different mount points.

Consistency Concurrent access to data in NFS is problematic. NFS does not provide POSIX single site semantics, and modifications made by one NFS client will not be propagated quickly to all other clients. NFS does support byte range advisory locking, but not many applications honor such locks.

Given these shortcomings, it is not recommended to use NFS in large production environments that require fast, secure, and easy-to-manage global file systems. On the other hand, NFS administration is fairly easy, and small environments with low security requirements will probably choose NFS as their global file system.

4.4.2 The DFS and AFS file systems

There are mainly two global file systems that can be used as an alternative to NFS. The Distributed File System (DFS) is part of the Distributed Computing Environment (DCE) from the Open Software Foundation (OSF), now known as the Open Group. The Andrew File System (AFS) from Transarc is the base technology from which DFS was developed; so, DFS and AFS are in many aspects very similar. Both DFS and AFS are not part of base AIX, they are available as separate products. Availability of DFS and AFS for platforms other than AIX differs but not significantly.

For reasons that will be discussed later, we recommend to use DFS rather than AFS except when an SP is to be integrated into an existing AFS cell. We, therefore, limit the following high-level description to DFS. Most of these general features also apply for AFS, which has a very similar functionality. After a general description of DFS, we point out some of the differences between DFS and AFS that justify our preference of DFS.

What is the Distributed File System?

DFS is a distributed application that manages file system data. It is an application of the Distributed Computing Environment (DCE) in the sense that it uses almost all of the DCE services to provide a secure, highly available, scalable, and manageable distributed file system.

DFS data is organized in three levels:

- Files and directories. These are the same data structures known from local file systems, such as the AIX Journaled File System (JFS). DFS provides a global namespace to access DFS files as described below.
- Filesets. A DFS *fileset* is a group of files and directories that are administered as a unit. Examples would be all the directories that belong to a particular project. User home directories may be stored in separate filesets for each user or may be combined into one fileset for a whole (AIX) group. Note that a fileset cannot be larger than an aggregate.
- Aggregates. An *aggregate* is the unit of disk storage. It is also the level at which DFS data is exported. There can be one or more filesets in an DFS aggregate. Aggregates cannot be larger than the logical volume in which they are contained.

The client component of DFS is the *cache manager*. It uses a local disk cache or memory cache to provide fast access to frequently used file and directory data. To locate the server that holds a particular fileset, DFS uses the *fileset location database (FLDB) server*. The FLDB server transparently accesses information about a fileset's location in the FLDB, which is updated if a fileset is created or moved to another location.

The primary server component is the *file exporter*. The file exporter receives data requests as DCE Remote Procedure Calls (RPCs) from the cache manager and processes them by accessing the local file systems in which the data is stored. DFS includes its own *Local File System (LFS*) but can also export other UNIX file systems (although with reduced functionality). It includes a *token manager* to synchronize concurrent access. If a DFS client wants to perform an operation on a DFS file or directory, it has to acquire a token from the server. The server revokes existing tokens from other clients to avoid conflicting operations. By this, DFS is able to provide POSIX single site semantics.

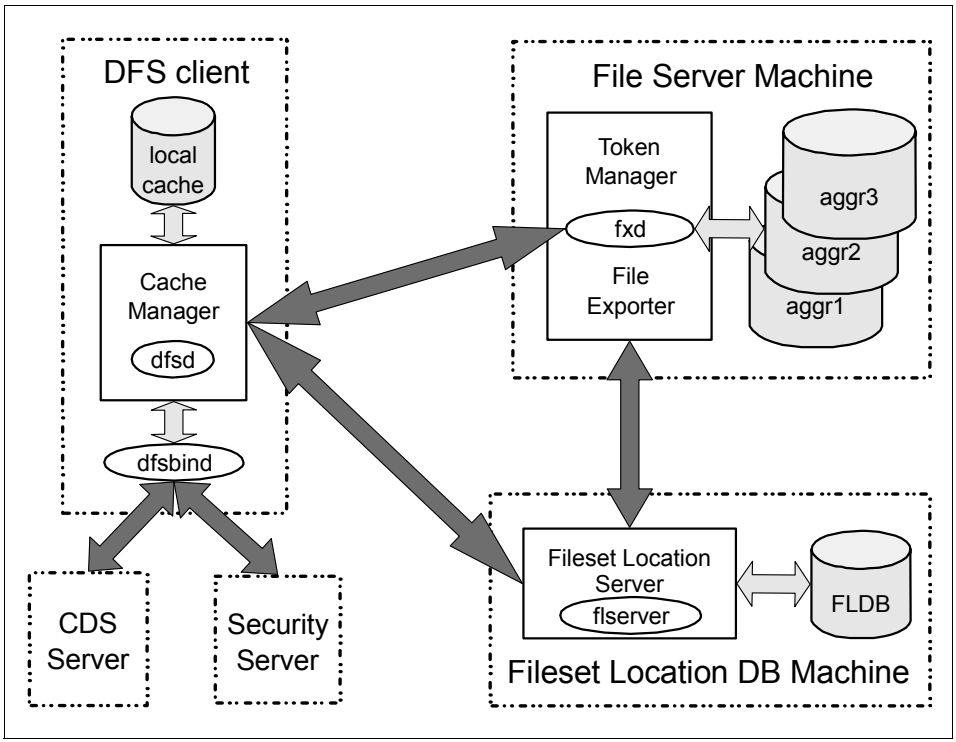


Figure 4-16 Basic DFS components

Figure 4-16 shows these DFS components. Note that this is an incomplete picture. There are many more DFS components like the replication server and various management services like the fileset server and the update server. More detailed information about DFS can be found in the product documentation *IBM DCE for AIX: Introduction to DCE* and *IBM DCE for AIX: FS Administration Guide and Reference,* which can be found at:

http://www.ibm.com/software/network/dce/library/publications/

.

The following list summarizes some key features of DCE/DFS and can be used to compare DFS with the discussion in 4.4.1, “Network File System (NFS)” on page 148.

Performance DFS achieves high performance through client caching. The client to server ratio is better than with NFS, although exact numbers depend on the actual applications. Like NFS, DFS is limited by the performance of a single server.

Security DFS is integrated with the DCE Security Service, which is based on Kerberos Version 5. All internal communication uses the authenticated DCE RPC, and all users and services that want to use DFS services have to be authenticated by logging in to the DCE cell (except when access rights are explicitly granted for unauthenticated users). Access control is by DCE principal. Root users on DFS client machines cannot impersonate these DCE principals. In addition, DCE Access Control Lists can be used to provide fine-grained control; they are recognized even in a heterogeneous environment.

Management Since fileset location is completely transparent to the client, DFS filesets can be easily moved between DFS servers. Using DCE’s DFS as the physical file system, this can even be done without disrupting operation. This is an invaluable management feature for rapidly growing or otherwise changing environments. The fact that there is no local information on fileset locations on the client means that administering a large number of machines is much easier than maintaining configuration information on all of these clients.

Namespace DFS provides a global, worldwide namespace. The file system in a given DCE cell can be accessed by the absolute path /.../cell_name/fs/, which can be abbreviated as /: (slash colon) within that cell. Access to foreign cells always requires the full cell name of that cell. The global name space ensures that a file will be accessible by the same name on every DFS client. The DFS client has no control over mount points; filesets are mounted into the DFS namespace by the servers. Of course, a client may use symbolic links to provide alternative paths to a DFS file, but the DFS path to the data will always be available.

Consistency Through the use of a token manager, DFS is able to implement complete POSIX single-site read/write semantics. If a DFS file is changed, all clients will see the modified data on their next access to that file. Like NFS, DFS does support byte range advisory locking.

Operation To improve availability, DFS filesets can be replicated; that is, read-only copies can be made available by several DFS servers. The DFS server processes are monitored and maintained by the DCE basic overseer server (BOS), which automatically restarts them as needed.

In summary, many of the problems related to NFS do not exist in DFS or have a much weaker impact. DFS is, therefore, more suitable for use in a large production environment. On the other hand, DCE administration is not easy and requires a lot of training. The necessary DCE and DFS licenses also cause extra cost.

Differences between DFS and AFS

Apart from the availability (and licensing costs) of the products on specific platforms, there are two main differences between DFS and AFS: The integration with other services and the mechanism to synchronize concurrent file access. The following list summarizes these differences:

Authentication AFS uses Kerberos Version 4 in an implementation that predates the final MIT Kerberos 4 specifications. DCE/DFS uses Kerberos Version 5. For both, the availability of other operating system services (such as Telnet or X display managers) that are integrated with the respective Kerberos authentication system depends on the particular platform.

Authorization DFS and AFS ACLs differ and are more limited in AFS. For example, AFS can only set ACLs on the directory level not on file level. AFS also cannot grant rights to a user from a foreign AFS cell; whereas, DFS supports ACLs for foreign users.

Directory Service DCE has the Cell Directory Service (CDS) through which a client can find the server(s) for a particular service. The DFS client uses the CDS to find the Fileset Location Database. There is no fileset location information on the client. AFS has no directory service. It relies on a local configuration file (/usr/vice/etc/CellServDB) to find the Volume Location Database (VLDB), the Kerberos servers, and other services.

RPC Both DFS and AFS use Remote Procedure Calls (RPCs) to communicate over the network. AFS uses Rx from Carnegie Mellon University. DFS uses the DCE RPC, which is completely integrated into DCE including security. AFS cannot use dynamic port allocation. DFS does so by using the RPC *endpoint map*.

Time Service	DFS uses the DCE Distributed Time Service. AFS clients use their cache manager and NTP to synchronize with the AFS servers.
Synchronization	Both DFS and AFS use a token manager to coordinate concurrent access to the file system. However, AFS revokes tokens from other clients when closing a file; whereas, DFS already revokes the token when opening the file. This means that DFS semantics are completely conforming with local file system semantics, whereas, AFS semantics are not. Nevertheless, AFS synchronization is better than in NFS, which does not use tokens at all.

It is obvious that DFS is well integrated with the other DCE core services; whereas, AFS requires more configuration and administration work. DFS also provides file system semantics that are superior to AFS. So, unless an existing AFS cell is expanded, we recommend that DFS is used rather than AFS to provide global file services.

4.5 Related documentation

This documentation will help you better understand the concepts and examples covered in this chapter. We recommend that you take a look at some of these books in order to maximize your chances of success in the SP certification exam.

SP manuals

RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment, GA22-7281. This manual gives detailed explanations on I/O devices.

RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment, GA22-7280. This book is the official document for supported I/O adapters.

SP redbooks

Inside the RS/6000 SP, SG24-5145. NFS and AFS concepts are discussed in this redbook.

IBM (e)server Cluster 1600 Managed by PSSP 3.5: What's New, SG24-6617.

4.6 Sample questions

This section provides a series of questions to help you prepare for the certification exam. The answers can be found in Appendix A, “Answers to sample questions” on page 521.

1. If you attach a tape drive to one of your nodes, which of the following statements are true?
   a. All nodes get automatic access to that tape drive.
   b. Tape access is controlled by the file collection admin file.
   c. Any node can be backed up to the tape unit through a named pipe using the switch to provide a high speed transport.
   d. The tape needs to be attached to the control workstation.
2. Not all node types can support SSA boot. Which of the following statements are true?
   a. Only external SP-attached servers support external SSA boot.
   b. Only PCI nodes support external SSA boot.
   c. Only MCA nodes support external SSA boot.
   d. All nodes support external SSA boot except SP-attached servers.
3. PSSP 3.1 or later supports multiple rootvg definitions per node. To activate an specific rootvg volume group, you have to:
   a. Issue the **spbootlist** command against the node.
   b. Issue the **spchvgobj** command against the node.
   c. Issue the **spbootins** command against the node.
   d. Issue the **spchvg** command against the node.
4. PSSP uses NFS for network installation and home directory services of the SP nodes. The control workstation and boot/install servers act as NFS servers to make resources for network installation available to the nodes. Which of the following statements are false?
   a. Home directories are served by the control workstation by default.
   b. Home directories are served by boot/install servers by default.
   c. The control workstation is always a NFS server.
   d. Boot/install servers keep local copies of PSSP software.
5. Which command enables mirroring on a set of nodes?
   a. spmirrorvg -l node_list
   b. spbootins -l <node_list>

c. sprmvgobj -l node_list

d. spmkvgobj -l nodelist

6. Which command displays information about Volume_Groups?

   a. splstdata -v -l <node #>

   b. spbootins -v <lppsouce_name>

   c. sprmvgobj -r vg_name

   d. spmkvgobj -h pv_list

7. When is NFS *not* recommended to be used as the global file system?

   a. Environments with low security requirements.

   b. In a large production environment.

   c. Environments where the administration is fairly easy.

   d. In small environments.

8. Which of the following statements regarding DFS data organization are *not* true?

   a. DFS data is organized in filesets.

   b. DFS data is organized in files and directories.

   c. DFS data is organized in distribution files.

   d. DFS is organized in aggregates.

9. When the SDR is initialized, a volume group is created for every node. By default, the vg_name attribute of the Volume_Group object is set to rootvg, and the selected_vg of the node is set to rootvg. Which of the following statements are default values?

   a. Quorum is false.

   b. The default install_disk is hdisk1.

   c. Mirroring is off, copies are set to 1.

   d. There are bootable, alternate root volume groups.

10. Which of the following commands can be used to be able to boot using SSA external disks?

    a. spbootins

    b. spmkvgobj

    c. spmirrorvg

    d. splstdata

4.7 Exercises

Here are some exercises you may wish to do:

1. On a test system that does not affect any users, upgrade to AIX 4.3.2 and PSSP 3.1.
2. On a test system that does not affect any users, list all the Volume_Group default values.
3. On a test system that does not affect any users, create a new volume group (rootvg1), then activate the new volume group. Hint: Check your level of AIX and PSSP before the exercise.
4. On a test system that does not affect any users, familiarize yourself with the various flags of the **spmkvgobj** command.
5. On a test system that does not affect any users, familiarize yourself with the various flags of the **spbootins** command.
6. On a test system that does not affect any users, familiarize yourself with the various flags of the **splstdata** command.

5

Cluster 1600 installation and administration

The SP-attached server support has been implemented since PSSP 3.1. At first only the RS/6000 Models 7017 S70, S7A, S80 and S85 were supported. Now with PSSP 3.5 the list of attached servers is much larger.

All these servers, either directly attached to the CWS or attached over an HMC, are too large to fit in a regular SP Frame. Clustering combines switched or non-switched (either SP Switch or SP Switch2) nodes. It includes logical nodes as LPARs, and physical nodes.The Cluster 1600 can contain up to a maximum of 128 logical nodes. Switch-only models, control workstations, and Hardware Management Consoles are not included. Therefore, a new Machine Type was developed to cover the Cluster 1600 machines. It is the 9078 Model 160.

The Cluster 1600 is a scalable system that may include:

- IBM eServer pSeries or RS/6000 servers
- SP system components
- Control workstations (CWS)
- Hardware Management Consoles (HMC)
- Management servers
- SP switches (SP Switch or SP Switch2)

- Industry standard network interconnects

5.1 Key concepts

Before taking the SP Certification exam, make sure you understand the following concepts:

- How the SP-attached servers are connected to the SP (control workstation and switch).
- What software levels are required to attach an SP-attached server.
- The difference between an SP-attached server and a dependent node.
- What are the node, frame, and switch numbering rules when attaching an SP-attached server?

5.2 Hardware attachment

In this section, we describe the hardware architecture of the SP-attached server and its attachment to the SP system, including areas of potential concern of the hardware or the attachment components.

5.2.1 Cluster 1600 overview

There are several pSeries models and the well-known RS/6000 SP nodes available that can be clustered together in a Cluster 1600 environment. For a quick overview, refer to Table 5-1.

Table 5-1 Cluster 1600 overview

Cluster 1600 nodes	HMC attach	Standard RS232 connection to CWS	Custom RS232 connection to CWS
M/T 9076			
POWER3 nodes and all other PCI nodes	No	No	Yes
M/T 7040			
IBM @server pSeries 670	Yes	Yes	No
IBM @server pSeries 690	Yes	Yes	No
M/T 7039[a]			

Cluster 1600 nodes	HMC attach	Standard RS232 connection to CWS	Custom RS232 connection to CWS
IBM @server pSeries 655	Yes	Yes	No
M/T 7038			
IBM @server pSeries 650	Yes	Yes	No
M/T 7028			
IBM @server pSeries 630	Yes	Yes	No
M/T 7026[b]			
IBM @server pSeries 660 models 6H1, 6H0 and 6M1	No	No	Yes
RS/6000 models M80 and H80	No	No	Yes
M/T 7017			
IBM @server pSeries 680	No	No	Yes
RS/6000 model S80, S7A and S70	No	No	Yes

a. M/T 7039 requires RS-422 connections between the HMC and the Bulk Power Controllers on the M/T 7040-W42 frame used with the 7039 server.
b. For M/T 7026 servers only, an SP Internal Attachment Adapter.

Overall, an Ethernet connection to the SP LAN may require an SP-supported card and a customer-supplied cable.

For more detailed information about the hardware specifications, refer to Chapter 2, “Validate hardware and software configuration” on page 7.

5.2.2 Cluster 1600 scaling limits and rules

The scalability of the Cluster 1600 is a very complex matter. Depending on which type of server is used and how many logical nodes are used, the scaling limits are:

- The type of servers installed

- Whether or not the system contains a switch
- The type of switch used
- Whether or not a server is divided into LPARs

We provide the scaling limits and rules for the Cluster 1600 managed by PSSP only. For the Cluster 1600 limits managed by PSSP 3.5, an operating system image or logical node can be:

- 7040 (pSeries 670, pSeries 690) LPAR or full system partition
- 7039 (pSeries 655) LPAR or full system partition
- 7038 (pSeries 650) LPAR or full system partition
- 7028 (pSeries 630 Model 6C4 only) LPAR or full system partition
- 7026 (pSeries 660 and RS/6000 H80/M80) server
- 7017 (S70, S7A, S80, pSeries 680) server
- 9076 SP nodes

The Cluster 1600 may be configured with multiple machine types. However, with PSSP, the cluster must meet all of the following limits for AIX operating system images installed. Any cluster system that exceeds these limits requires a special bid. Refer to Table 5-2 for the physical scaling limits and to Table 5-3 for the operating system scaling limits.

Table 5-2 Physical scaling limits

Server/node type	Maximum server/nodes per type	Maximum server/nodes per group	Maximum server/nodes per cluster
SP nodes	128	128	128
7026/7028/7039	64	64	N/A
7017	16	N/A	N/A
7040	32	N/A	N/A

Table 5-3 Operating system scaling limits

Server/node type	Maximum AIX images per server/node type	Maximum AIX images per server/node group	Maximum AIX images per cluster
SP nodes	128	128	128
7026/7028/7039	64	128	N/A

Server/node type	Maximum AIX images per server/node type	Maximum AIX images per server/node group	Maximum AIX images per cluster
7017	16	N/A	N/A
7040	48	N/A	N/A

Important: A Cluster 1600 with PSSP must meet *all* of the following limits:

- No more than 128 logical nodes from the machine type set {7040, 7039, 7038, 7028, 7026, 7017, 9076}
- No more than 32 servers from the machine type set {7040}
- No more than 64 servers from the machine type set {7040, 7039, 7038, 7028, 7026, 7017}
- No more than 16 servers from the machine type set {7017}
- No more than 128 9076 nodes

Cluster limits for supported pSeries server

For all the Cluster 1600-supported IBM eServer pSeries models several limits apply, depending on how many LPARs, physical servers, and so on. The HMC can control several different machine types. The weighting factor of each machine type differs. A pSeries p690 has twice the weighting factor of a pSeries 630. The rules that apply are shown in Table 5-4.

Table 5-4 Maximum rules for pSeries servers

Maximum values for server	SP Switch[a]	SP Switch2	SP Switch2 (total switched and non-switched, subject to other limits[a])	Industry standard connect
p670/p690 servers per cluster (POWER4)	32	32	32	32
p670/p690 servers per cluster (POWER4+)	N/A	32	32	32
p655/p650/p630 servers per cluster	N/A	64	64	64
LPARs per p690 server (POWER4)	8	16	16	16
LPARs per p690 server (POWER4+)	N/A	16	16	16

Maximum values for server	SP Switch[a]	SP Switch2	SP Switch2 (total switched and non-switched, subject to other limits[a])	Industry standard connect
LPARs per p670 server (POWER4)	4	4	16	16
LPARs per p670 server (POWER4+)	N/A	4	16	16
LPARs per p655/p630 server	N/A	2	4	4
LPARs per p650 server	N/A	2	8	8
LPARs per cluster	128	128	128	128
Number of switch planes supported per p670/p690 server	1	1 or 2[b]	1 or 2[b]	0
Number of switch planes supported per p655 server	N/A	1 or 2[b]	1 or 2[b]	0
Number of switch planes supported per p630/p650	N/A	1	1	0
Number of p690/p670 servers per HMC	8	8	8	8
Number of p655/650/630 servers per HMC	N/A	16	16	16
Number of LPARs per HMC	32	32	32	32

a. With the SP Switch, switched and non-switched logical nodes cannot be mixed.
b. Two-plane support implies two adapters per LPAR. Two-plane and single-plane configurations cannot be mixed.

5.2.3 External node attachment

The physical attachment of an external node is done with several connections that include:

- An Ethernet connection to the administrative LAN for system administration purposes
- RS-232 cable connections
  - For 7017 models – Two custom cables connecting the CWS to both the server SAMI port and to the S1 serial port
  - For 7026 models – One custom cable connecting the CWS to the Internal Attachment Adapter

- For 7040 models – One connection to the HMC (no RS-232 connection to the CWS)
- For 9076 models – One custom cable per frame connecting the CWS to the SP frame supervisor card

This section describes the attachment of the SP-attached server to the SP, highlighting the potential areas of concern that must be addressed before installation. The physical attachment is subdivided and described in three connections.

- Connections between the CWS and the SP-attached server are described in “CWS considerations” on page 173.
- Connections between the SP Frame and the SP-attached Server are described in “SP frame connections” on page 174.
- An optional connection between the SP Switch and the SP-attached server are described in “Switch connection (required in a switched SP system)” on page 174.

These connections are shown in Figure 5-1 on page 167.

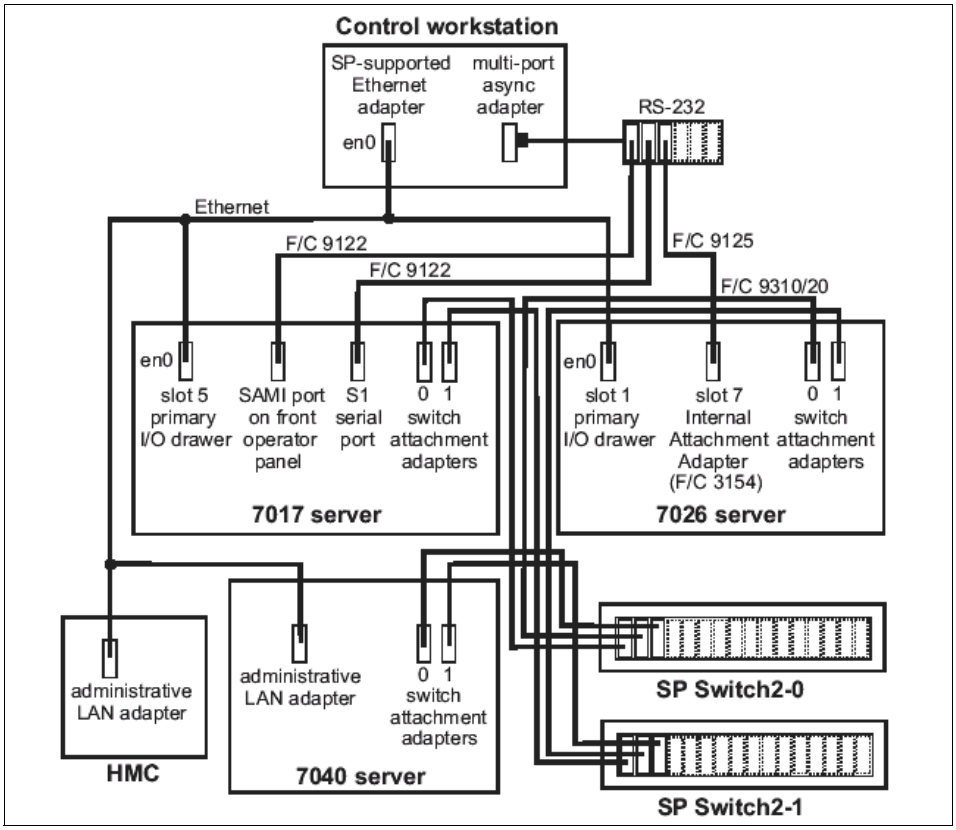


Figure 5-1 Cluster 1600 attachment overview

M/T 7039 attachment overview

M/T 7039 attachment overview

The 7039 pSeries 655 Model 651 node is located in a 7040-W42 frame. For the Cluster 1600 attachment, some special connections have to be made.

The M/T 7039 attachment needs the following connections:

- An administrative LAN Ethernet connection from the HMC and each server to the CWS for system administration purposes.
- Each server requires an RS-232 serial connection to the HMC.
- The server's M/T 7040-W42 frame requires two RS-422 serial connections with the HMC for power monitoring.

Since you can either have an SP Switch2 attachment or no SP Switch2 attachment, refer to Figure 5-2 for non-switched attachment and to Figure 5-3 on page 169 for switched attachment.

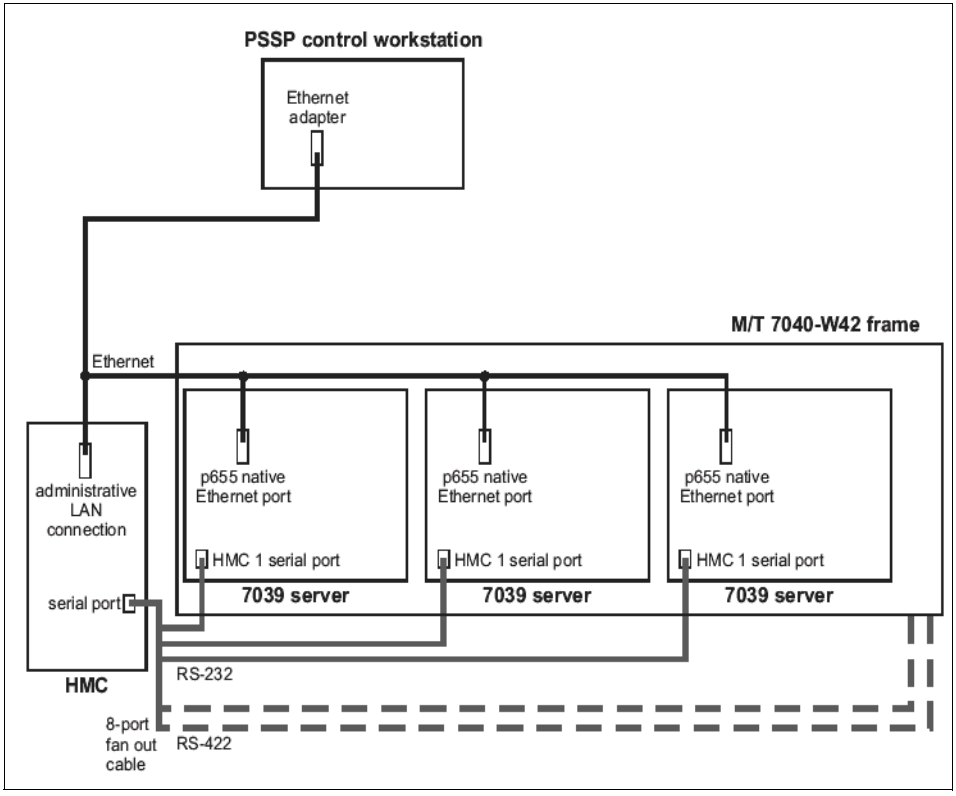


Figure 5-2 M/T 7039 p655 server for non switched attachment

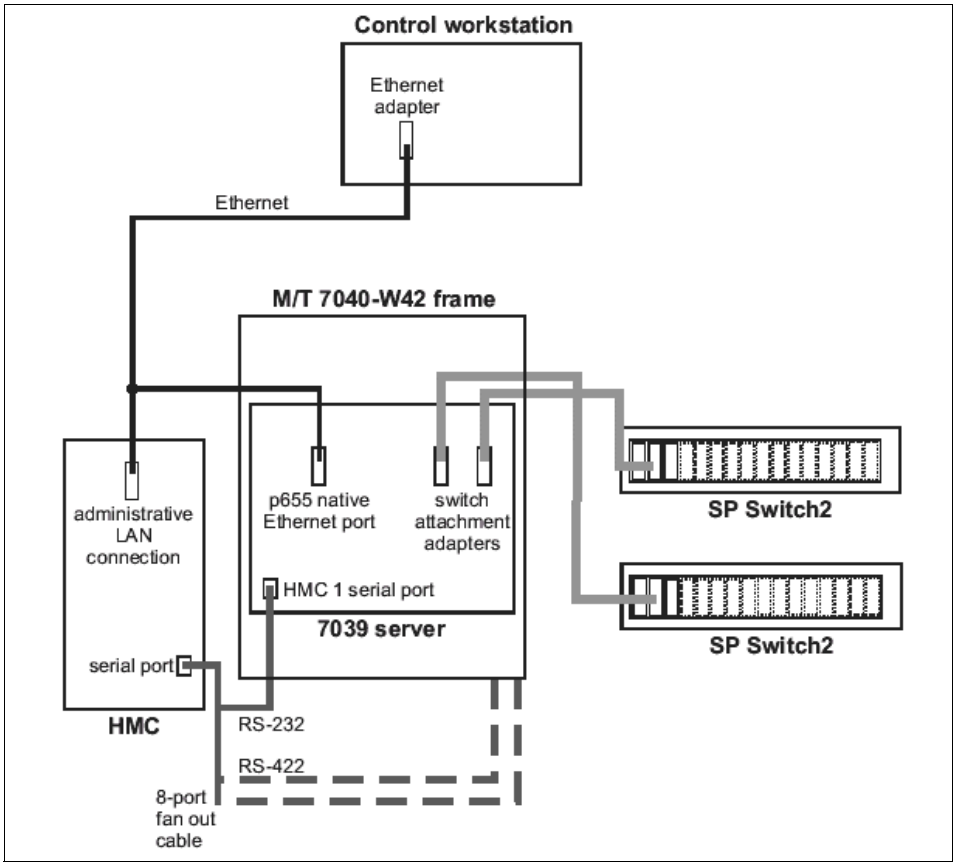


Figure 5-3 M/T 7039 p655 switched attachment with dual plane configuration

Note: All M/T 7039 cluster configurations require PSSP for control.

M/T 7028 attachment overview

M/T 7028 attachment overview

Depending on cluster control software, each M/T 7028 server in the Cluster 1600 requires an administrative LAN Ethernet connection to either the control workstation or management server for system administration purposes. In addition, each server also requires an RS-232 serial connection to the HMC.

Refer to Figure 5-4 on page 170 for an overview of an SP Switch2-configured 7028 p630 installation.

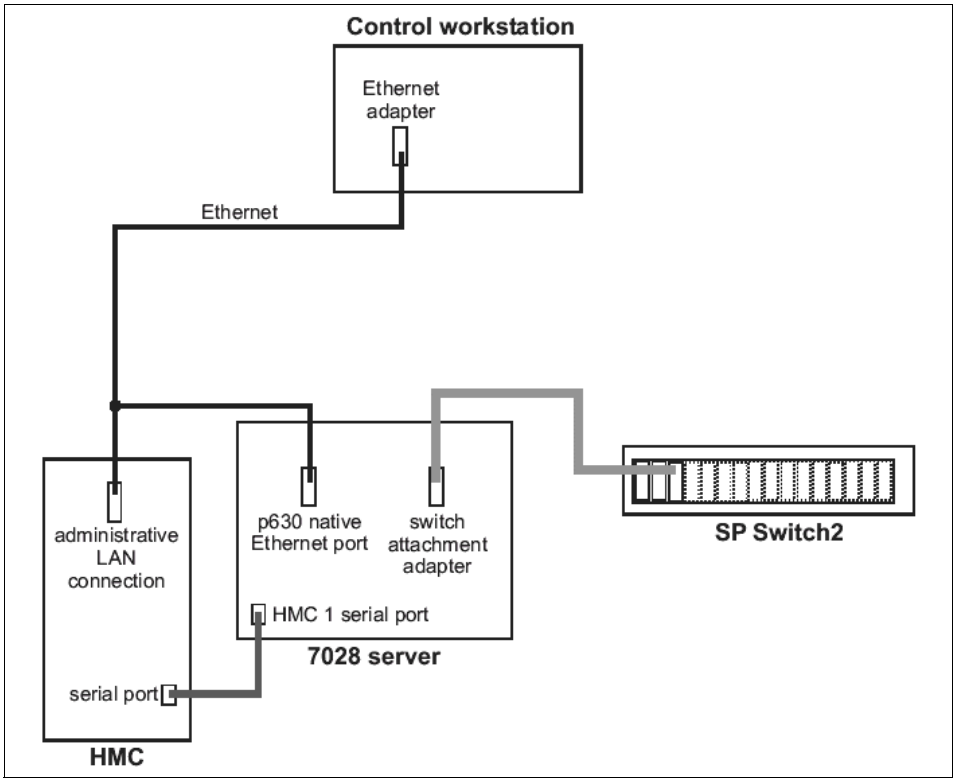


Figure 5-4 M/T 7028 in an SP Switch2 configuration

It is important to note that the size of the SP-attached servers like M/T 7017, 7040, and so on, prohibits them from being physically mounted in the SP frame. Since the SP-attached server is mounted in its own rack and is either directly attached to the CWS using RS-232 or via the HMC through the SP administrative LAN to the CWS, the SP system must view the SP-attached server as a frame. It is also viewed as a node. Because the PSSP code runs on the machine, it is managed by the CWS, and you can run standard applications on the SP-attached server. Therefore, the SP system views the SP-attached server as an object with both frame and node characteristics.

However, as the SP-attached server does not have *full* SP frame characteristics, it cannot be considered a standard SP expansion frame. Therefore, when assigning the server's frame number, you have to abide by the following rules:

- The SP-attached server cannot be the first frame in the SP system.
- The SP-attached server cannot be inserted between a switch-configured frame and any non-switched expansion frame using that switch. It can, however, be inserted between two switch-configured frames. Different

attachment configurations are described in 5.7, “Attachment scenarios” on page 207.

Once the frame number has been assigned, the server’s node numbers, which are based on the frame number, are automatically generated. The following rules are used.

Rules for M/T7017-S70, S7A, S80, p680, and 7026-p660

The node numbering rules for these machine types are:

- These SP-attached server types are viewed as single frames containing a single node.
- These servers occupy the slot 1 position.
- Each SP-attached server installed in the SP system subtracts one node from the total node count allowed in the system. However, because the server has frame-like features, it reserves 16 node numbers that are used in determining the node number of nodes placed after the attached server. The algorithm for calculating the node_number is:

```
node_number = (frame_number -1) * 16 + slot_number
```

Figure 5-5 on page 172 shows an example of node numbering with three frames.

For further information about the frame numbering issue, refer to Figure 5-14 on page 207.

Rules for HMC-controlled servers p690, p670, p655, p650, and p630

Note: An unpartitioned server has one LPAR and is seen by PSSP as one node. A partitioned server is seen by PSSP as one frame with as many nodes as there are LPARs. The number of these servers counts toward the total number of servers in one system.

- Each one of these servers is shown as a single frame.
- When no partitioning (LPAR) is configured, one node is shown for that frame.
- When partitioning is used, each LPAR is shown as a node in that frame. An LPAR functions like an SP node. But since originally only 16 nodes per frame are allowed in an SP frame, you can only configure up to 16 LPARs in p690, p690+ systems. No more than 16 LPARs will show up in a frame. There is no support from IBM for LPARs 17 to 32 configured on a system. Refer to “Cluster limits for supported pSeries server” on page 164.

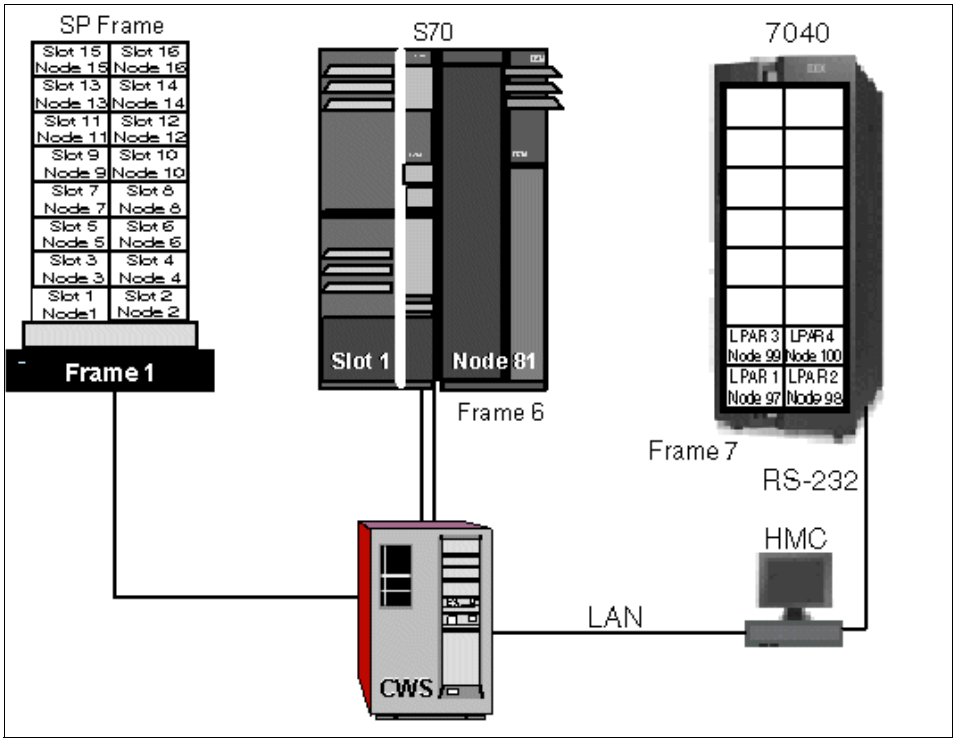


Figure 5-5 Node numbering

Important information on LPARs

Based on the preceding information, you could configure a p690+ server to have 17 LPARs with LPAR numbers 1 - 17. Realizing that the p690+ server can only be attached to the Cluster 1600 system managed by PSSP if, and only if, it has no more than 16 LPARs on it, you may think that you can delete any LPAR to meet the restriction.

However, that is not the case. If you deleted the LPAR with LPAR number 5, the p690+ server would only have 16 LPARs on it, but they would have LPAR numbers of 1 - 4 and 6 - 17. The LPAR with LPAR number 17 violates one of the conditions for p690+ attachment. Therefore, instead of deleting the LPAR with LPAR number 5, you must delete the LPAR with LPAR number 17 in order to properly configure the p690+ server for attachment to the cluster.

Control workstation connections

Control workstation connections

The SP-attached server does not have a frame or node supervisor card, which limits the full hardware, control, and monitoring capabilities of the server from the SP CWS (unlike other SP nodes). However, it does have some basic capabilities, such as power on/off.

M/T 7017 attachment

Three CWS connections to the SP-attached server are required for hardware control and software management:

- An Ethernet connection to the SP LAN for system administration purposes.
- A custom-built RS-232 cable connected from the S70 operator panel to a serial port on the CWS. It is used to emulate operator input at the operator panel. An S70-specific protocol is used to monitor and control the S70 hardware. This protocol is known as the Service and Manufacturing Interface (SAMI).
- A second custom-built RS-232 cable that must only use the S70 S1 serial port. This is used to support the s1term connectivity. This is a custom-built RS-232 cable, which is part of the order features, with a null modem and a *gender-bender*.

The other possible connection scenarios are shown in 5.2.3, “External node attachment” on page 165.

CWS considerations

In connecting the SP-attached server to the CWS, it is important to keep the following CWS areas of concern in mind:

- When connecting the SP-attached frame to the system, you need to make sure that the CWS has enough spare serial ports to support the additional connections. However, it is important to note that there is one restriction with the 16-port RS-232 connection. By design, it does not pass the required ClearToSend signal to the SAMI port of the SP-attached server, and, therefore, the *16-port RS-232 cannot be used* for the RS-232 connectivity to the SP-attached server. The eight-port and the 128-port varieties will support the required signal for connectivity to the SP-attached server.
- There are two RS-232 attachments for each S70/S7A/S80 SP attachment. The first serial port on the S70/S7A/S80 *must* be used for s1term connectivity.
- Floor placement planning to account for the effective usable length of the RS-232 cable.

The connection cables are 15 meters in length, but only 11.5 meters are effective. So, the SP-attached server must be placed at a distance where the RS-232 cable to the CWS is usable.

- Although M/T 7040, 7039, and 7028 servers do not require an RS-232 connection to the CWS, they do require an RS-232 connection between the servers and the HMC.
- In a HACWS environment, there will be no SP-attached server control from the backup CWS. In the case where a failover occurs to the backup CWS, hardmon and s1term support of the SP-attached is not available until failback to the primary CWS. The node will still be operational with switch communications and SP Ethernet support.

SP frame connections

The SP-attached server connection to the SP frame is as follows:

- 10 meter frame-to-frame electrical ground cable

  The entire SP system must be at the same electrical potential. Therefore, the frame-to-frame ground cables provided with the S70 server must be used between the SP system and the S70 server in addition to the S70 server electrical ground.

Frame considerations

In connecting the SP-attached server to the SP frame, it is important to have the following in mind:

- The SP system must be a *tall frame* because the 49-inch short *LowBoy* frames are not supported for the SP attachment.
- The tall frame with the eight-port switch is not allowed.
- The SP-attached server *cannot* be the first frame in the SP system. So, the first frame in the SP system must be an SP frame containing at least one node. This is necessary for the SDR_config code, which needs to determine whether the frame is with or without a switch.
- A maximum of 128 logical nodes are supported in one SP system. This means that if a switch is installed, there must be eight available switch connections in the SP system, one per SP-attached server.

For complete power planning information, refer to *RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment,* GA22-7280.

Switch connection (required in a switched SP system)

This is the required connection if the SP-attached server is to be connected to a switched SP system.

- The PCI SP Switch or SP Switch2 adapter, known as the SP system attachment adapter, of the SP-attached server connects to the 16-port SP switch through a 10-meter switch cable.

Switch considerations

In connecting the SP-attached server to the SP Switch, it is important to verify and follow the rules. Refer to 2.9, “SP Switch and SP Switch2 communication network” on page 50.

SP-attached server considerations

In connecting the SP-attached server to the SP system, it is important to have in mind the following potential concerns:

- Supported adapters

  All adapters currently supported in the SP environment are supported with the SP-attached servers. However, not all currently supported SP-attached server adapters are supported in the SP Switch-attached server environment. If the external node possesses adapters that are not currently supported in the SP environment, they *must* be removed from the SP-attached server.

  Since the variety of SP-attached servers is huge, the adapters that are supported in each system differ. Contact your IBM account representative. The latest information of PCI adapter placement is available in *RS/6000 and IBM @server pSeries: PCI Adapter Placement Reference SA38-0538,* and for SP internal nodes, in *RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment,* GA22-7280.

5.3 Cluster 1600 installation requirements

The Cluster 1600 installation requirements are shown here. Since the Cluster 1600 includes attached servers and SP nodes either communicating with SP Switch/SP Switch2, or no switch connections, several things have to be considered for the installation process. We will provide some information on that. The newer external nodes that are controlled by an HMC are only connected with an Ethernet connection to the CWS.

5.3.1 System specific requirements

The systems that can be integrated to a Cluster 1600 have slightly different requirements, which we discuss here.

M/T 7040, 7026, 7017, and 9076

- If you plan to use a preexisting server that is connected to an IPv6 network, you must remove the server from that network before installing it into a Cluster 1600 system.
- If your Cluster 1600 system is switch-configured, it must use either SP Switch or SP Switch2. A special switch adapter and connecting cable must be installed. For more information, refer to 2.9, “SP Switch and SP Switch2 communication network” on page 50
- All PCI adapters in both new and preexisting M/T 7026 and 7017 servers must be SP-supported.
- You must install up to three cable connections to the CWS. Since these cables have limited lengths, you must keep those lengths in mind because they limit the location of the servers in relation to the other system equipment.

Note: Placement of the servers is limited by the lengths of the following supplied cables:

- The 15-m (49 ft.) RS-232 cables
- Optional 10 or 20 m (33 or 66 ft.) switch cables

Approximately 3 m (10 ft.) of cable is needed for the vertical portion of the cable runs. Thus, the Cluster 1600 servers must be no more than 12 m (40 ft.) from the CWS.

M/T 7039 and M/T 7028

- If either M/T 7039 or M/T 7028 servers are used in your Cluster 1600 system and the system is going to be switch-configured, you must use an SP Switch2 or an industry standard Ethernet interconnect.
  - If SP Switch2 is used, one F/C 9126 must be ordered for each server connected to the switch. F/C 9126 is a specific code that reserves a switch port and assigns a 4032 switch interposer. F/C 9126 does not deliver any physical components.
  - If an industry standard Ethernet interconnect is used, the native ports on the p655 or p630 or optional adapters may be used to connect to the Ethernet network.

Note: Switch use is optional for M/T 7039 and M/T 7028 in a Cluster 1600. These systems may be configured without a switch.

- If PSSP is installed on the 7039 or 7028 servers, both a CWS and an HMC are required. This configuration requires the following:
  - An administrative LAN connection from the CWS to the HMC and to each server using customer supplied Ethernet cable.
  - An RS-232 connection between the HMC and each server.
  - M/T 7039 requires two RS-422 connections between the HMC and the frame.

5.3.2 Software requirements

The requirements for the different SP-attached servers and SP nodes is relevant for proper installation. Here we show you the recent requirements.

Software requirements for M/T 7040, 7026, 7017, and 9076

The supported software requirements are:

- AIX 5L 5.1 and PSSP 3.4 or AIX 5L 5.2 PSSP 3.5
- AIX 4.3.3 and PSSP 3.4

Software requirements for M/T 7039

For M/T 7039 the requirements are:

- AIX 5L for POWER V5.1 with the 5100-03 recommended maintenance package and PSSP 3.4 or PSSP 3.5

Software requirements for M/T 7028

The p630 servers, the HMC, and their associated control workstation require one of the following software levels when used as part of a Cluster 1600:

- AIX 5L 5.1 and PSSP 3.4 or PSSP 3.5
- AIX 5L 5.2 and PSSP 3.5

Note: Each Cluster 1600 system server requires its own PSSP license. PSSP is available as a CD.

5.4 Installation and configuration

The SP-attached server is treated as similarly as possible to a frame with a node. However, there are some important distinctions that have to be addressed during SP-attached server configuration, namely the lack of frame and node supervisor cards and support for two ttys instead of one as described in 5.2.3, “External node attachment” on page 165.

Information that is unique to the SP-attached server is entered in the configuration of this server. Once the administrator configures the necessary information about the SP-attached server processor in the SDR, then the installation should proceed the same as any standard SP node in the SP administrative network.

Configuration considerations

To configure an SP-attached server, do the following:

- Add two ttys on the CWS.
- Define the Ethernet adapter on the SP-attached server.
- In a switched system, configure the SP-attached server to the SP Switch.
- Frame definition of an SP-attached server:

  The rules for assigning the frame number of the SP-attached server are detailed in 5.2.3, “External node attachment” on page 165.

  The SP-attached server must be defined to PSSP using the **spframe** command and using the new options that are available for SP-attached servers for this command. Refer to Example 5-1.

Example 5-1 spframe command options

```
/usr/lpp/ssp/bin/spframe -p {hardware protocol}
-n {starting_switch_port}
[-r {yes|no}]   [-s {s1tty}]
start_frame frame_count starting_tty_port
Usage:  spframe [-p SP] [-r {yes|no}] [-m] [-o]
                start_frame frame_count starting_tty_port
        spframe -p SAMI [-n {starting_switch_port}] [-s {s1tty}]
                [-r {yes|no}] [-o] start_frame frame_count
starting_tty_port
        spframe -p CSP [-n {starting_switch_port}] [-r {yes|no}] [-o]
                start_frame frame_count starting_tty_porttion from
        spframe -p HMC -d {domain_name} -i {list_of_HMC_IPaddresses}
                [-n {starting_switch_port}] [-r {yes|no}] [-o] frame_number
```

Alternatively, you can use the smitty nonsp_frame_dialog menu as shown in Figure 5-6 on page 179.

```
 Non-SP Frame Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* Start Frame                                          [ ]
#
* Frame Count                                          [ ]
#
* Starting Frame tty port                               [/dev/tty0]
* Starting Switch Port Number                          [ ]
#
  s1 tty port                                           [ ]
* Frame Hardware Protocol                               [SAMI]
  Re-initialize the System Data Repository              no
+
```

Figure 5-6 Non-SP frame information

This menu will request frame number, tty ports, and switch port numbers. This will establish *hardmon* communications with the SP-attached server and create the frame object in the SDR. For all the various SP-attached servers you have these frame hardware protocols available:

- SP - Standard SP Frame protocol
- SAMI - RS/6000 S70/S7A/S80 and pSeries 680
- HMC - HMC-controlled servers (pSeries p690, 670, 655
- CSP - RS/6000 H80/M80, pSeries 660 Model 6H0/6H1/6M1

The number of tty port values you must define depends on the hardware protocol type you selected.

- HMC - Does not require a tty port value, but the HMC must be connected by the SP Ethernet administrative LAN or by the HMC trusted network.
- CSP - Requires one tty port value.
- SAMI - Requires two tty port values:
  - The servers that use the SAMI hardware protocol require two tty port values to define the tty ports on the CWS to which the serial cables connected to the server are attached. The tty port value defines the serial connection to the operator panel on these servers for hardware controls.

The s1 tty port value defines the connection to the serial port on the servers for serial terminal (s1term) support.

- Hardware Ethernet address collection

  The MAC address of the SP-attached server is retrieved by **sphrdwrad** in the same way as a normal SP node and placed in the SDR.

  Now that the SP-attached server is configured as an SP-attached server frame in the SDR, it is ready for standard configuration and installation as a normal node. Full instructions are defined in *PSSP Installation and Migration Guide,* GA22-7347*.*

- Boot/Install consideration

  The default setup for boot/install servers is that the CWS is the boot/install server for a single frame system. In a multiple frame system, the CWS installs the first node in each frame and defines this node as the boot/install server for the remaining nodes in its frame.

  If, however, the multiple frame system contains an SP-attached server, the CWS remains as the default boot/install server for the first node in each frame. The first node in each SP frame becomes the boot/install server with the exception of the SP-attached server, which is treated as a node instead of a frame.

- Installing the node

  The configuration and installation of the SP nodes and SP-attached servers are identical. All of the installation operations will be performed over the Ethernet with one of the tty lines providing the s1term capabilities and the other tty line providing the hardware control and monitoring functions.

- System partitioning consideration

  If the system has multiple partitions defined, and you wish to add an SP-attached server, you do not need to bring the system down to one partition, as the SP-attached server appears as a standard SP node to the system partition.

  Each SP-attached server has appropriate frame, slot values, and switch port numbers. These values are accommodated for existing attributes in the relevant frame, node, and Syspar_map SDR classes.

  When the SP-attached server frame/node is defined to the system with the **spframe** command, the switch port number to which the node is connected is identified. This number is also necessary in a switchless system to support system partitioning.

  If it is necessary to change the switch port number of the SP-attached server, then the node has to be deleted and redefined with a new switch port number.

Deleting this node should be done by deleting the frame to ensure that no inconsistent data is left in the SDR.

- If more than one partition exists, repartition to a single partition.
- Invoke **spdelframe** to delete the SP-attached server frame and node definitions.
- Recable the server to a new switch port.
- Invoke **spframe** to redefine the SP-attached server frame and node to specify the new switch port number.
- If the system was previously partitioned, repartition back to the system partitioning configuration.

- Considerations when integrating an existing SP-attached server:

  Perform the following steps to add an existing SP-attached server and preserve its current software environment:

  a. Physical attachment

     When integrating an existing SP-attached server node to your system, it is recommended (though not mandatory) that the frame be added to the end of your system to prevent having to reconfiguring the SDR. Different attachment scenarios are described in “Attachment scenarios” on page 207.

  b. Software levels

     If your SP-attached server is not at the required AIX level, upgrade to that level and ensure that PSSP code_version is set to PSSP-3.5.

  c. Customize node

     To perform a preservation install of an SP-attached server with PSSP software, the node must be set to *customize* instead of *install* in the SDR. For example:

     ```
     spbootins -r customize -l 33
     ```

  d. Mirroring

     If the root volume group of the SP-attached server has been mirrored, and the mirroring is to be preserved, the information about the existing mirrors must be recorded in the SDR; otherwise, the root volume group will be unmirrored during customization.

     For example, if the root volume group of the S70 Advanced Server has two copies on two physical disks in locations 30-68-00-0,0 and 30-68-00-2,0 with quorum turned off, enter the following to preserve the mirroring:

     ```
     spchvgobj -r rootvg -c 2 -q false -h 30-68-00-0,0:30-68-00-2,0 -l 33
     ```

To verify the information, enter:

```
splstdata -b -l 33
```

e. Set up the name resolution of the SP-attached server

For PSSP customization, the following must be resolvable on the SP-attached server:

- The control workstation host name.
- The name of the boot/install server's interface that is attached to the SP-attached server's en0 interface.

f. Set up routing to the control workstation host name

If a default route exists on the SP-attached server, it must be deleted. If it is not removed, customization will fail when it tries to set up the default route defined in the SDR. In order for customization to occur, a static route to the control workstation's host name must be defined. For example, the control workstation's host name is its Token Ring address, such as 9.114.73.76, and the gateway is 9.114.73.256:

```
route add -host 9.114.73.76 9.114.73.256
```

g. FTP the SDR_dest_info file

During customization, certain information will be read from the SDR. In order to get to the SDR, the /etc/SDR_dest_info file must be FTPed from the control workstation to the /etc/SDR_dest_info file of the SP-attached server ensuring the mode and ownership of the file is correct.

h. Verify perfagent

Ensure that perfagent.tools 2.2.32.x are installed on the SP-attached server.

i. Mount the pssplpp directory

Mount the /spdata/sys1/install/pssplpp directory from the boot/install server on the SP-attached server. For example, issue:

```
mount k3n01:/spdata/sys1/install/pssplpp /mnt
```

j. Install `ssp.basic`

Install spp.basic and its prerequisites onto the SP-attached server. For example:

```
installp /aXgd/mnt/PSSP-3.1 ssp.basic 2>&1 | tee /tmp/install.log
```

k. Unmount the pssplpp directory

Unmount the /spdata/sys1/install/pssplpp directory on the boot/install server from the SP-attached server. For example:

```
umount /mnt
```

l. Run pssp_script by issuing:

```
/usr/lpp/ssp/install/bin/pssp_script
```

m. Reboot

Perform a reboot of the SP-attached server.

For more details on installing nodes, refer to Chapter 9, “Frame and node installation” on page 301.

5.5 PSSP support

This section describes the PSSP software support for the SP-attached server. Of special interest is the fact that the SP-attached server does not use the SP node or frame supervisor cards. Hence, the software modifications and interface to the SP-attached server must simulate the architecture of the SP Frame Supervisor Subsystem such that the boundaries between an SP node and an SP-attached server node are minimal.

5.5.1 SDR classes

The SDR contains system information describing the SP hardware and operating characteristics. Several class definitions have changed to accommodate the support for SP-attached servers, such as frame, node, and Syspar_map classes. A new class definition has been added in PSSP 3.1, the NodeControl class.

The classes that contain information related to SP-attached servers are briefly described:

- Hardware Protocol

  For all the various SP-attached servers you have these frame hardware protocols available:

  - SP - Standard SP Frame protocol
  - SAMI - RS/6000 S70/S7A/S80 and pSeries 680
  - HMC - HMC-controlled servers (pSeries p690, 670, and 655)
  - CSP - RS/6000 H80/M80, pSeries 660 Model 6H0/6H1/6M1

- Frame class

  Currently, the frame class is used to contain information about each SP frame in the system. This information includes physical characteristics (number of slots, whether it contains a switch, and so forth), tty port, hostname, and the internal attributes used by the switch subsystem.

SP-attached server nodes do not have physical frame hardware and do not contain switch boards. However, they do have hardware control characteristics, such as tty connections and associated Monitor and Control Nodes (MACN). Therefore, an SDR Frame Object is associated with each SP-attached server node to contain these hardware control characteristics.

Two attributes have been added to the frame class: *hardware_protocol* and *s1_tty*.

The hardware_protocol attribute distinguishes the hardware communication method between the existing SP frames and the new frame objects associated with SP-attached server nodes. For these new nodes, the hardware communication methods are Service and Manufacturing Interface (SAMI), and CSP, which is the protocol used to communicate across the serial connection to the SP-attached server service processor. HMC protocol is used for the HMC-controlled servers. No direct serial connection is available between CWS and HMC.

The attribute s1_tty is used only for the SP-attached server nodes and contains the tty port for the S1 serial port connection established by the **s1term** command.

A typical example of a frame class with the new attributes and associated values is illustrated in Example 5-2.

Example 5-2 Example of a frame class with SP-attached servers

```
[c179s][/]> SDRGetObjects Frame
frame_number tty          frame_type   MACN         backup_MACN  slots
frame_in_config snn_index    switch_config hardware_protocol s1_tty
control_ipaddrs domain_name
           1 /dev/tty0    switch       c179s.ppd.pok.ibm.com "" 16 ""
""            ""          SP           ""           "" ""
           2 /dev/tty1    switch       c179s.ppd.pok.ibm.com "" 16 ""
""            ""          SP           ""           "" ""
          12 /dev/tty5    multinsb     c179s.ppd.pok.ibm.com "" 16 ""
""            ""          SP           ""           "" ""
           3 /dev/tty2    ""           c179s.ppd.pok.ibm.com "" 1 ""
""            ""          CSP          ""           "" ""
          11 ""           ""           c179s.ppd.pok.ibm.com "" 16 ""
""            ""          HMC          ""           9.114.213.98 c59ih04
           6 ""           ""           c179s.ppd.pok.ibm.com "" 16 ""
""            ""          HMC          ""           9.114.213.120 e159cec
          10 ""           ""           c179s.ppd.pok.ibm.com "" 16 ""
""            ""          HMC          ""           9.114.213.98 c59ih03
```

- **Node class**

  The SDR Node class contains node-specific information used throughout PSSP. Similarly, there will be an SDR Node object associated with the SP-attached server.

  SP frame nodes are assigned a node_number based on the algorithm described in 5.2.3, “External node attachment” on page 165.

  Likewise, the same algorithm is used to compute the node number of an SP-attached server frame node where the SP-attached server occupies the first and only slot of its frame. This means that for every SP-attached server frame node, 16 node numbers will be reserved, of which only the first one will ever be used.

  The node number is the key value used to access a node object.

  Some entries of the Node class example are shown in Table 5-5.

Table 5-5 Example of the Node class for SP nodes and SP-attached servers

Node Class	Nodes in an SP	Attached S70 node	Attached H80 server	Attached p690 server
Node number	1-15	17	any	any
Slot number	1-16	1(always)	1(always)	any
Switch node number	0-15	0-15	0-15	0-15
Switch chip	4-7	4-7	4-7	4-7
Switch number	1	1	1	1
Boot_device	disk	disk	disk	disk
Description	POWER3_SMP_High	7017-S70	7026-H80	7040-681
Platform	CHRP	CHRP	CHRP	CHRP
Hardware control type	179	10	12	13

  The platform attribute has a value of Common Hardware Reference Platform (CHRP) for the SP-attached server and the POWER3 SMP high node.

  The hardware_control_type key value is used to access the NodeControl class. The various attached servers have different types. Value 10 is for 7017-S70, value 12 is for 7026-H80, and value 13 is for HMC-controlled servers.

- **Syspar_map class**

  The Syspar_map class contains one entry for each switch port, assuming each frame would contain a switch.

  Because the SP-attached server has node characteristics, it has an entry in the Syspar_map class for that node with no new attributes.

  The `used` attribute of the Syspar_map class will be set to one for the SP-attached server node to indicate that there is a node available to partition. Since this node will be attached to the switch, switch_node_number will be set appropriately based on the switch port in an existing SP frame that the SP-attached server node is connected to.

  In a switchless system, switch_node_number is assigned by the administrator using the **spframe** command.

  An example of the Syspar_map class is shown in Example 5-3.

Example 5-3 Example of the Syspar_map class with SP-attached server

```
[c179s][/]> SDRGetObjects Syspar_map
syspar_name  syspar_addr  node_number  switch_node_number used
node_type

c179s        9.114.12.81            1           16            1 standard
c179s        9.114.12.81            5            1            1 standard
c179s        9.114.12.81            9            2            1 standard
c179s        9.114.12.81           13            9            1 standard
c179s        9.114.12.81           17            4            1 standard
c179s        9.114.12.81           21            5            1 standard
c179s        9.114.12.81           25           19            1 standard
c179s        9.114.12.81           29            7            1 standard
c179s        9.114.12.81           33           22            1 standard
c179s        9.114.12.81          145            0            1 standard
c179s        9.114.12.81          146           -1            1 standard
c179s        9.114.12.81           81            3            1 standard
c179s        9.114.12.81           82            6            1 standard
c179s        9.114.12.81           83            8            1 standard
c179s        9.114.12.81           84           14            1 standard
c179s        9.114.12.81           85           10            1 standard
c179s        9.114.12.81           86           11            1 standard
c179s        9.114.12.81           87           12            1 standard
c179s        9.114.12.81           88           13            1 standard
c179s        9.114.12.81          161           -1            1 standard
```

  The **SDR_config** command has been modified to accommodate these new SDR attribute values and now handles the assignment of switch_port_number for SP-attached server nodes.

- NodeControl class

  In order to support different levels of hardware control for different types of nodes, a new SDR class has been defined to store this information.

  The NodeControl class is a global SDR class that is not partition-sensitive. It contains one entry for each type of node that can be supported on an SP system. Each entry contains a list of capabilities that are available for that type of node. This is static information loaded during installation and is not changed by any PSSP code. This static information is required by the SDR_config script to properly configure the node.

  An example of the NodeControl class is given in Table 5-6.

Table 5-6 Example of the NodeControl class with the SP-attached server

Type	Capabilities	Slots_used	Platform_type	Processor_type
161	power,reset,tty,keySwitch,LCD,networkBoot	4	rs6k	MP
33	power,reset,tty,keySwitch,LCD,networkBoot	1	rs6k	UP
83	power,reset,tty,keySwitch,LCD,networkBoot	2	rs6k	UP
97	power,reset,tty,keySwitch,LCD,networkBoot	1	rs6k	UP
12	power,reset,tty,LCD,networkBoot	1	chrp	MP
115	power,reset,tty,keySwitch,LCD,networkBoot	2	rs6k	UP
81	power,reset,tty,keySwitch,LCD,networkBoot	2	rs6k	UP
145	power,reset	1	N/A	N/A
179	power,reset,tty,LCD,networkBoot	4	chrp	MP
177	power,reset,tty,LCD,networkBoot	1	chrp	MP
178	power,reset,tty,LCD,networkBoot	2	chrp	MP
13	power,reset,tty,LCD,networkBoot	1	chrp	MP
10	power,reset,tty,LCD,networkBoot	1	chrp	MP
65	power,reset,tty,keySwitch,LCD,networkBoot	1	rs6k	UP

Type	Capabilities	Slots_used	Platform_type	Processor_type
113	power,reset,tty,keySwitch,LCD,networkBoot	2	rs6k	UP

The key link between the Node class and the NodeControl class is the node type, which is an attribute stored in the SDR Node object. The SP-attached server 7040-690 has a node type value of 13 with hardware capabilities of power on/off, reset, tty, LCD, network Boot.

Perspectives routines and hardmon commands access this class to determine the hardware capabilities for a particular node before attempting to execute a command for a given node.

5.5.2 Hardmon

Hardmon is a daemon that is started by the System Resource Controller (SRC) subsystem that runs on the CWS. It is used to control and monitor the SP hardware (frame, switch, and nodes) by opening a tty that communicates using an internal protocol to the SP frame supervisor card through a serial RS-232 connection between the CWS and SP frame.

The SP-attached servers do not have a frame or node supervisor card that can communicate with the hardmon daemon. Therefore, a mechanism to control and monitor SP-attached servers is provided in PSSP.

Hardmon provides support for SP-attached servers in the following way:

- It discovers the existence of SP-attached servers.
- It controls and monitors the state of SP-attached servers, such as power on/off.

There are different ways to discover the SP-attached server:

- The eServer 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.
- 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.

Discover the SP-attached servers

For hardmon to discover the hardware, it must first identify the hardware and its capabilities. Today, for each frame configured in the SDRs frame class, hardmon opens a tty defined by the tty field. A two-way communication to the frame supervisor through the RS-232 interface occurs where hardmon sends hardware control commands and receives state data in the form of packets.

Since PSSP 3.1, two fields have been added to the SDR's frame class: *hardware_protocol* and *s1_tty*. These enable hardmon to determine the new hardware that is externally attached to the SP and also what software protocol must be used to communicate to this hardware.

Upon initialization, hardmon reads its entries in the SDR frame class and also examines the value of the hardware_protocol field to determine the type of hardware and its capabilities. If the value read is SP, this indicates that SP nodes are connected to hardmon through the SP's Supervisor subsystem. A value of SAMI is specific to the S70/S7A/S80 hardware since it is the SAMI software protocol that allows the communication, both sending messages and receiving packet data, to the S70/S7A/S80's Service Processor.

Another protocol is the HMC value that applies for all HMC-attached servers like p690, p670, p655, p650 and p630. For this type of machine the hmc daemon is started, but only one per attached HMC. The CSP protocol applies to the 7026-H80, and p660 servers. No additional daemon is started since hardmon communicates through the attached CSP serial card.

Discover the 7017 SP-attached server

Once hardmon recognizes the existence of one or more S70/S7A/S80s in the configuration, it starts a new process: the S70 daemon. One S70 daemon is started for each frame that has an SDR Frame class hardware_protocol value of SAMI. Now, hardmon can send commands and process packets or serial data as it would with normal SP frames. This is illustrated in Figure 5-7 on page 190.

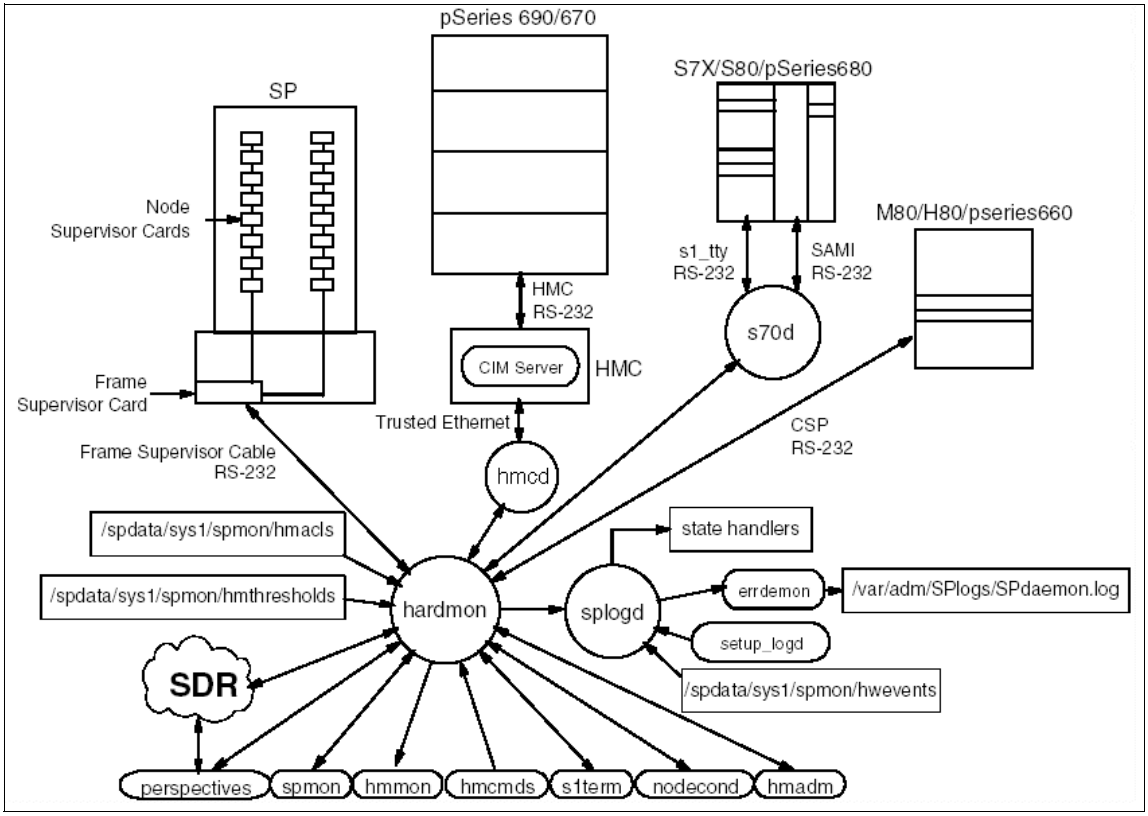


Figure 5-7 hardmon data flow

It is important to note that only hardmon starts the S70 daemon and no other invocation external to hardmon is possible. In addition, the parent hardmon daemon starts a separate S70 daemon for each S70 frame configured in the SDR frame class.

The S70 daemon starts with the following flags:

```
/usr/lpp/ssp/install/bin/S70d -d 0 2 1 8 /dev/tty2 /dev/tty1
```

where `-d` indicates the debug flag, `0` is the debug option, `2` is the frame number, `1` is the slot number (which is always 1), `8` is the file descriptor of the S70d's side of the socket that is used to communicate with hardmon, `/dev/tty2` is the tty that is used to open the SAMI/MI operator panel port, and `/dev/tty1` is the serial tty.

S70 daemon

The S70 daemon interfaces to the S70 hardware and emulates the frame and node supervisor by accepting commands from hardmon and responding with

hardware state information in the same way as the frame supervisor would. Its basic functions are:

- It polls the S70 for hardware changes in hardware status and returns the status to hardmon in the form of frame packet data.
- It communicates with the S70 hardware through the SAMI/MI interface.
- It accepts hardware control commands from hardmon to change the power state of the S70 and translates them into SAMI protocol, the language that the Manufacturing Interface (MI) understands. It then sends the command to the hardware.
- It opens the tty defined by the tty field in the SDR Frame class through which the S70 daemon communicates to the S70 serial connection.
- It supports an interface to the S70 S1 serial port to allow console connections through s1term.
- It establishes and maintains data handshaking in accordance with the S70 Manufacturing Interface (MI) requirements.

Data flow

Hardmon requests are sent to the S70 daemon where the command is handled by one of two interface components of the S70 daemon, the frame supervisor interface, or the node supervisor interface.

The frame supervisor interface is responsible for keeping current the state data in the frame's packet and formats the frame packet for return to hardmon. It will accept hardware control commands from hardmon that are intended for itself and *pass on* to the node supervisor interface commands intended to control the S70/S7A/S80 node.

The node supervisor interface polls state data from the S70/S7A/S80 hardware for keeping current the state data in the nodes' packet. The node supervisor interface will translate the commands received from the frame supervisor interface into S70/S7A/S80 software protocol and sends the command through to the S70/S7A/S80 service processor.

If the **hardmon** command is intended for the frame, the frame supervisor entity of the S70d handles it. If intended for the node, the node supervisor entity converts it to SAMI protocol and sends it out the SAMI/MI interface file descriptor as illustrated by Figure 5-8 on page 192.

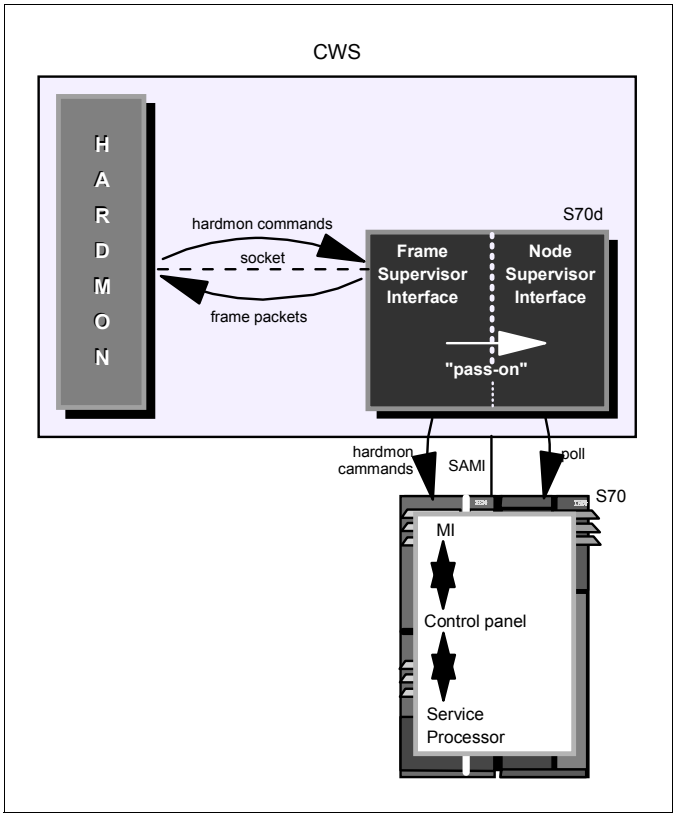


Figure 5-8 S70 daemon internal flow

The S70 daemon uses the SAMI protocol, which takes the form of 4-byte command words, to talk to the S70's Manufacturing Interface. This interface communicates with the S70's operator panel, which in turn communicates with the S70's Service Processor. It is the Service Processor that contains the instruction that acts upon the request. Data returned to the S70 daemon follows the reverse flow.

Monitoring of SP-attached server

For hardmon to monitor the hardware, it must first identify the hardware and its capabilities.

The hardware control type is determined from the SDR Node class as a hardware_control_type attribute. This attribute is the key into the NodeControl class. The NodeControl class indicates the hardware capabilities for monitoring. This relationship is illustrated in Figure 5-8.

Hardmon Resource Monitor Daemon

The Hardmon Resource Monitor Daemon (HMRMD) supports the Event Management resource variables to monitor nodes. With the new SP-attached servers, new resource variables are required to support their unique information.

There are four new hardmon variables that will be integrated into the Hardmon Resource Monitor for the SP-attached servers. They are SRChasMessage, SPCNhasMessage, src, and spcn. Historical states, such as nodePower, serialLinkOpen, and type, are also supported by the SP-attached servers. The mechanics involved with the definition of these variables are no different than those with previous variables and can be viewed through Perspectives and in conjunction with the Event Manager.

In order to recognize these new resource variables, the Event Manager must be stopped and restarted on the CWS and all the nodes in the affected system partition.

HMC daemon

The hmc daemon (hmcd) is the program that indirectly interfaces to the HMC controlled server hardware through the HMC Common Information Model (CIM) server. In fact, the hmcd acts as a frame and node supervisor and accepts all control commands from hardmon. Once hardmon recognizes the existence of p690 servers in its configuration, it starts the hmc daemon. Hardmon does it for each unique control_ipaddrs attribute in the SDR where the hardware_protocol is also set to HMC.

For more information about the new SDR attributes, refer to 5.5.1, “SDR classes” on page 183. Each hmcd belongs to a specific HMC connection. So, one hmcd daemon is responsible for multiple systems connected to this HMC. Example 5-4 shows the **ps** command output of an hmcd daemon running on the CWS.

Example 5-4 ps -ef output for hmcd

```
[c179s][/]> ps -ef |grep hmcd
    root 11646 29816  55   Jul 15      - 18552:15 /usr/lpp/ssp/install/bin/hmcd
-d 0 8436 9.114.213.120 1 e159cec 6 6
```

So you can get the following information from this output:

- hmcd -d 0 8436 - is the debug level (same as hardmon).
- 9.114.213.120 - is the IP address of the HMC.
- 1 - is the number of servers following.
- e159cec - is the domain name

- 6 6 - are the frame number and hardmon socket file descriptor.

Logfiles

The hmcd daemon logs its debug and error information in separate logfiles. The logfiles are located in /var/adm/SPlogs/spmon/hmcd. The naming conventions of the logfiles are as follows

- Logfile

  hmcd.<IP address>.log.<ddd>

  - IP address: IP address of the HMC
  - ddd: Julian date of the date the logfile was opened by the hmcd daemon

hmcd data flow

The data flow between the hardmon and hmc daemons is shown in Figure 5-9.

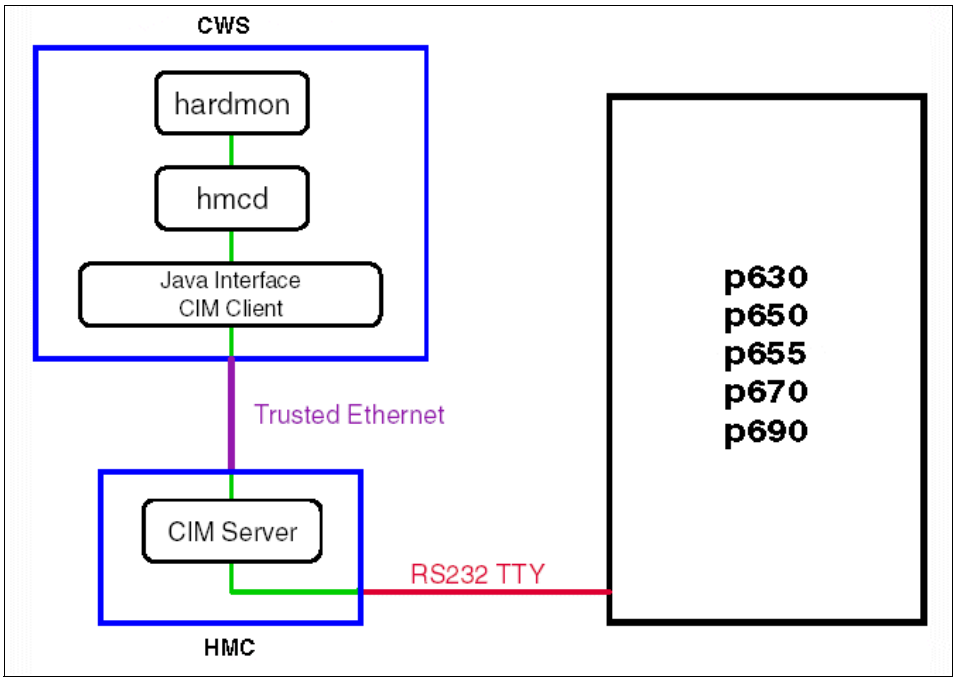


Figure 5-9 Data flow between hmcd and hardmon

The hardmon talks directly to hmcd. The hmcd gets the information from the CIM server running on the HMC. The connection between the CWS and the HMC is made via a trusted Ethernet network.

5.6 User interfaces

This section highlights the changes in the different user interface panels and commands that have been developed to represent the SP-attached server to the user.

5.6.1 Perspectives

As SP must now support nodes with different levels of hardware capabilities, an interface was architected to allow applications, such as Perspectives, to determine what capabilities exist for any given node and respond accordingly. This interface will be included with a new SDR table, the NodeControl class.

Note: The PSSP Perspectives function requires 24-bit graphics support on the xServer to which it is displayed. If displaying Perspectives on the control workstation display, then adapter GXT-135P or later, which supports both 24- and 8-bit, is acceptable.

The Perspectives interface needs to reflect the new node definitions: Those that are physically not located on an SP frame and those nodes that do not have full hardware control and monitoring capabilities.

There is a typical object representing the SDR frame object for the SP-attached server node in the Frame/Switch panel. This object has a unique pixmap placement to differentiate it from a high and low frame, and this pixmap is positioned according to its frame number in the Perspectives panel.

An example of the Perspectives representation of the SP-attached servers is shown in Figure 5-10 on page 196. In this case, we have a 7026 Model H80 shown as frame 3 and three pSeries 690s in LPAR mode as frames 6, 10 and 11.

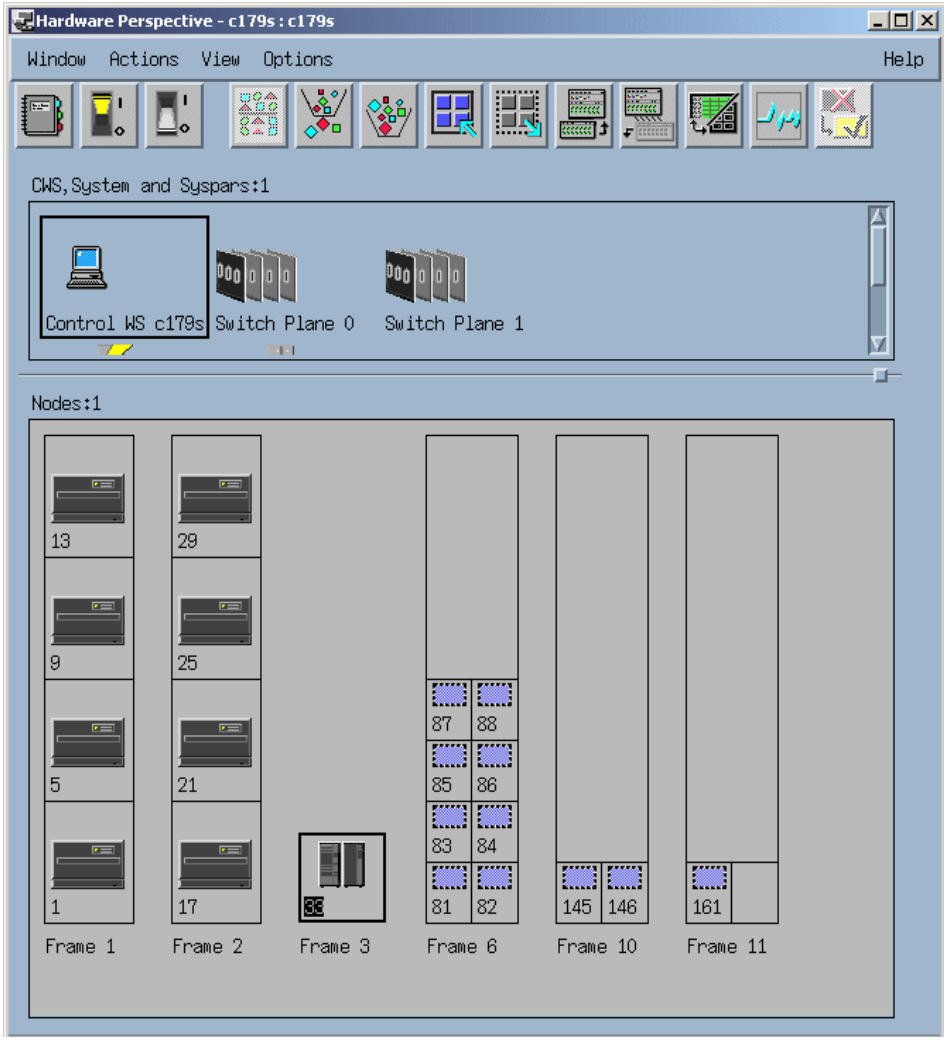


Figure 5-10 Example of Perspectives with an H80 and a p690 as attached servers

The monitored resource variables are handled the same as for standard SP nodes. Operations, status, frame, and node information are handled as for standard SP nodes.

Only the Hardware Perspective (sphardware) GUI is affected by the new SP-attached server nodes. The remaining panels, Partitioning Aid Perspective (spsyspar), Performance Monitoring Perspective (spperfmon), Event Perspective (spevent), and VSD Perspective (spvsd) are all similar to the sphardware Perspective node panel since they are based on the same class. Therefore, the

pixmap's placement will be similar to that of the sphardware Perspective node panel.

Event Manager

With the new SP-attached server nodes, new resource variables are required to support their unique information.

These new resource variables will be integrated into the Hardmon Resource Monitor for the SP-attached server:

- IBM.PSSP.SP_HW.Node.SRChasMessage
- IBM.PSSP.SP_HW.Node.SPCNhasMessage
- IBM.PSSP.SP_HW.Node.src
- IBM.PSSP.SP_HW.Node.spcn

In order to recognize these new resource variables, the Event Manager must be stopped and restarted on the CWS and all the nodes in the affected system partition.

System management

The various system management commands that display new SDR attributes for SP-attached servers are:

- `spmon`

  Example 5-5 shows the **`spmon -d -G`** output in an SP system that consists of an SP frame and an SP-attached server.

Example 5-5 output of the spmon -d -G command

```
[c179s][/spdata/sys1/syspar_configs/bin]> spmon -d -G
1.  Checking server process
    Process 29816 has accumulated 1 minutes and 56 seconds.
    Check successful

2.  Opening connection to server
    Connection opened
    Check successful

3.  Querying frame(s)
    7 frames
    Check successful

4.  Checking frames

      Controller Slot 17 Switch Switch   Power supplies
Frame Responds   Switch  Power  Clocking A     B     C     D
```

```
----- ---------- ------- ------ -------- ----- ----- ----- -----
   1       yes      yes     on      N_A     on    on    on   N/A
   2       yes      yes     on      N_A     on    on    on   N/A
   3       yes       no     N/A     N/A    N/A   N/A   N/A   N/A
   6       yes       no     N/A     N/A    N/A   N/A   N/A   N/A
  10       yes       no     N/A     N/A    N/A   N/A   N/A   N/A
  11       yes       no     N/A     N/A    N/A   N/A   N/A   N/A
  12       yes       no     N/A     N/A     on    on    on   N/A

5.  Checking nodes
------------------------------ Frame 1 ------------------------------
                        Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1     1  high   on     yes     N/A     no  LCDs are blank      no
  5     5  high   on     yes     N/A     no  LCDs are blank      no
  9     9  high   on     yes     N/A     no  LCDs are blank      no
 13    13  high   on     yes     N/A     no  LCDs are blank      no

          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1     1   yes    yes
  5     5   yes    yes
  9     9   yes    yes
 13    13   no     no

------------------------------ Frame 2 ------------------------------
                        Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1    17  high   on     yes     N/A     no  LCDs are blank      no
  5    21  high   on     yes     N/A     no  LCDs are blank      no
  9    25  high   on     yes     N/A     no  LCDs are blank      no
 13    29  high   on     yes     N/A     no  LCDs are blank      no

          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1    17   yes    yes
  5    21   yes    yes
  9    25   yes    yes
 13    29   yes    yes

------------------------------ Frame 3 ------------------------------
                        Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1    33 extrn   on     yes     N/A    N/A  LCDs are blank      no
```

```
          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1    33   no     yes

------------------------------ Frame 6 ------------------------------
                       Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1    81  thin   on     yes     N/A    N/A  LCDs are blank     N/A
  2    82  thin   on     yes     N/A    N/A  LCDs are blank     N/A
  3    83  thin   on     yes     N/A    N/A  LCDs are blank     N/A
  4    84  thin   on     yes     N/A    N/A  LCDs are blank     N/A
  5    85  thin  off     no      N/A    N/A  LCDs are blank     N/A
  6    86  thin  off     no      N/A    N/A  LCDs are blank     N/A
  7    87  thin  off     no      N/A    N/A  LCDs are blank     N/A
  8    88  thin  off     no      N/A    N/A  LCDs are blank     N/A

          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1    81   yes    yes
  2    82   yes    yes
  3    83   yes    yes
  4    84   yes    yes
  5    85   no     no
  6    86   no     no
  7    87   no     no
  8    88   no     no

----------------------------- Frame 10 -----------------------------
                       Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1   145  thin   on     yes     N/A    N/A  LCDs are blank     N/A
  2   146  thin   on     no      N/A    N/A  LCDs are blank     N/A

          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1   145   yes    yes
  2   146 noconn noconn

----------------------------- Frame 11 -----------------------------
                       Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1   161  thin   on     yes     N/A    N/A  LCDs are blank     N/A
```

```
          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1   161 noconn noconn

----------------------------- Frame 12 -----------------------------
                Clock Env
Slot Type  Power Input Error
---- ----- ----- ----- -----
  2   swit   on    0     no
  4   swit   on    0     no
```

- **splstdata**

  Example 5-6 is the output of **splstdata -n**. It shows seven frames. Example 5-7 on page 201 shows the output from **splstdata -f** where the H80 is shown as the third frame and the HMC-controlled machines (pSeries 690 and pSeries 655) are shown as frames 6, 8, 10, and 11.

  The SP frame has frame number 1 with four high nodes of node numbers 1, 5, 9, and 13, each occupying four slots.

  The SP-attached server has frame number 3, with one node, node_number 33, occupying one slot.

Example 5-6 splstdata -n output

```
[c179s][/spdata/sys1/syspar_configs/bin]> 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 c179n01.ppd.pok.i c179n01.ppd.pok.i ""                9.114.213.125
MP                                2 POWER3_SMP_High         1 true            ""
    5      1     5     4 c179n05.ppd.pok.i c179n05.ppd.pok.i ""                9.114.213.125
MP                                2 POWER3_SMP_High         1 true            ""
    9      1     9     4 c179n09.ppd.pok.i c179n09.ppd.pok.i ""                9.114.213.125
MP                                8 375_MHz_POWER3_         1 true            ""
   13      1    13     4 c179n13.ppd.pok.i c179n13.ppd.pok.i ""                9.114.213.125
MP                                2 POWER3_SMP_High         1 true            ""
   17      2     1     4 e179n01.ppd.pok.i e179n01.ppd.pok.i ""                9.114.213.125
MP                                8 POWER3_SMP_High         1 true            ""
   21      2     5     4 e179n05.ppd.pok.i e179n05.ppd.pok.i ""                9.114.213.125
MP                                2 POWER3_SMP_High         1 true            ""
   25      2     9     4 e179n09.ppd.pok.i e179n09.ppd.pok.i ""                9.114.213.125
MP                                2 POWER3_SMP_High         1 true            ""
```

```
   29      2    13     4 e179n13.ppd.pok.i e179n13.ppd.pok.i ""               9.114.213.125
MP                                2 POWER3_SMP_High         1 true           ""
   33      3     1     1 c179mn01.ppd.pok. c179mn01.ppd.pok. ""               9.114.213.125
MP                                4 7026-H80                1 true           ""
   81      6     1     1 e159rp01.ppd.pok. e159rp01.ppd.pok. ""               9.114.213.125
MP                                8 7040-681                1 true           e159rp01
   82      6     2     1 e159rp02.ppd.pok. e159rp02.ppd.pok. ""               9.114.213.125
MP                                8 7040-681                1 true           e159rp02
   83      6     3     1 e159rp03.ppd.pok. e159rp03.ppd.pok. ""               9.114.213.125
MP                                8 7040-681                1 true           e159rp03
   84      6     4     1 e159rp04.ppd.pok. e159rp04.ppd.pok. ""               9.114.213.125
MP                                6 7040-681                1 true           e159rp04
   85      6     5     1 e159rp05.ppd.pok. e159rp05.ppd.pok. ""               9.114.213.125
MP                                3 7040-681                1 true           e159rp05
   86      6     6     1 e159rp06.ppd.pok. e159rp06.ppd.pok. ""               9.114.213.125
MP                                3 7040-681                1 true           e159rp06
   87      6     7     1 e159rp07.ppd.pok. e159rp07.ppd.pok. ""               9.114.213.125
MP                                3 7040-681                1 true           e159rp07
   88      6     8     1 e159rp08.ppd.pok. e159rp08.ppd.pok. ""               9.114.213.125
MP                                3 7040-681                1 true           e159rp08
  145     10     1     1 c59rp01.ppd.pok.i c59rp01.ppd.pok.i ""               9.114.213.125
MP                                4 7039-651                1 true           c59rp01
  146     10     2     1 c59rp02.ppd.pok.i c59rp02.ppd.pok.i ""               9.114.213.125
MP                                1 7039-651                0 false          c59rp02
  161     11     1     1 ""                ""                ""               ""
MP                                1 ""                      0 false          c59ih04
```

Example 5-7 is the output of **splstdata -f**, which shows seven frames.

Example 5-7 splstdata -f output

```
[c179s][/spdata/sys1/syspar_configs/bin]> splstdata -f
                 List Frame Database Information

frame tty            s1_tty         type     protocol domain_name
control_ipaddrs
----- -------------- -------------- -------- -------- ------------
-----------------------------------------------
    1 /dev/tty0      ""             switch   SP       ""           ""

    2 /dev/tty1      ""             switch   SP       ""           ""

    3 /dev/tty2      ""             ""       CSP      ""           ""

    6 ""             ""             ""       HMC      e159cec
9.114.213.120

   10 ""             ""             ""       HMC      c59ih03      9.114.213.98
```

```
11 ""            ""            ""       HMC      c59ih04      9.114.213.98

12 /dev/tty5     ""            multinsb SP       ""           ""
```

Example 5-8 is the output of **spgetdesc -u -a**, which shows the hardware description obtained from the Node class.

Example 5-8 spgetdesc -u -a output

```
spgetdesc: Node 1 (c188n01.ibm.com) is a Power3_SMP_Wide.
spgetdesc: Node 5 (c188n05.ibm.com) is a 332_MHz_SMP_Thin.
spgetdesc: Node 9 (c188n09.ibm.com) is a 332_MHz_SMP_Thin.
spgetdesc: Node 13 (c188n13.ibm.com) is a Power3_SMP_Wide.
spgetdesc: Node 17 (c187-S70.ibm.com) is a 7017-S70.
```

5.6.2 Hardware Management Console

There are several methods to access the HMC GUI that controls the HMC-controlled servers. One possibility is to use Perspectives on the CWS that manages the Cluster 1600. To do so:

Access HMC with Perspectives

- At the SP Perspectives Launch Pad, select **Hardware: Manage SP-attached Servers**.

See Figure 5-11 on page 203 for the Perspectives Launch Pad.

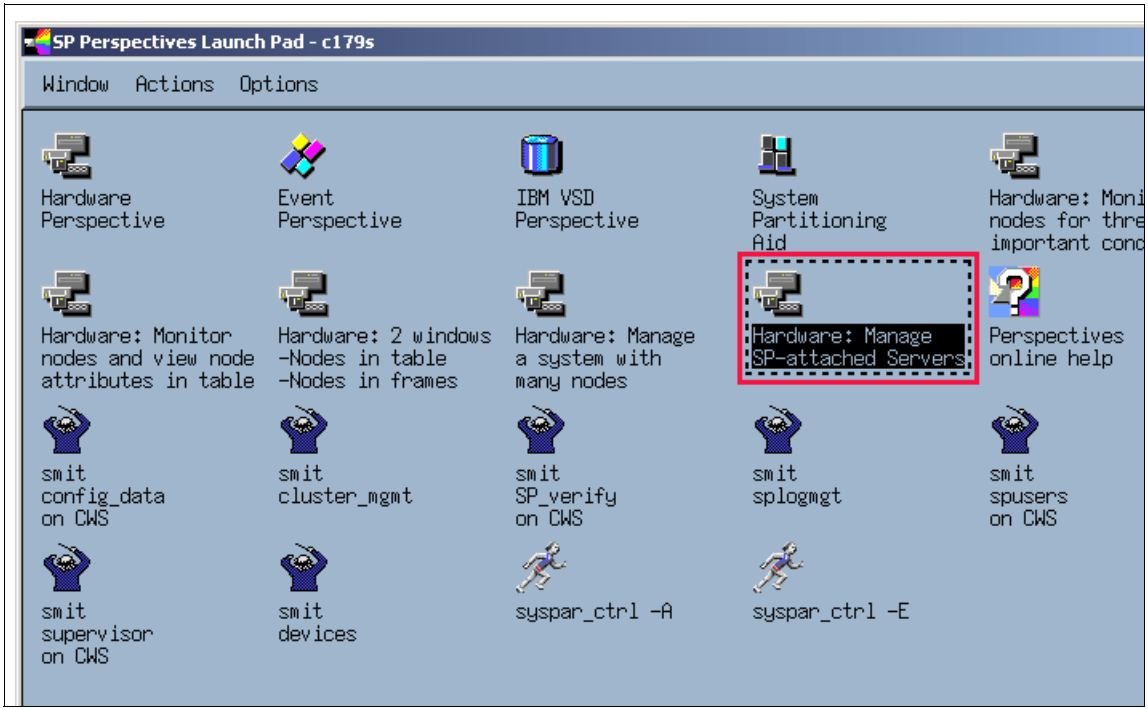


Figure 5-11 Perspectives Launch Pad

- Select the desired HMC controlled frame or node: select **Actions** → **Open HMC Interface**.

Refer to Figure 5-12 on page 204 for the described selection.

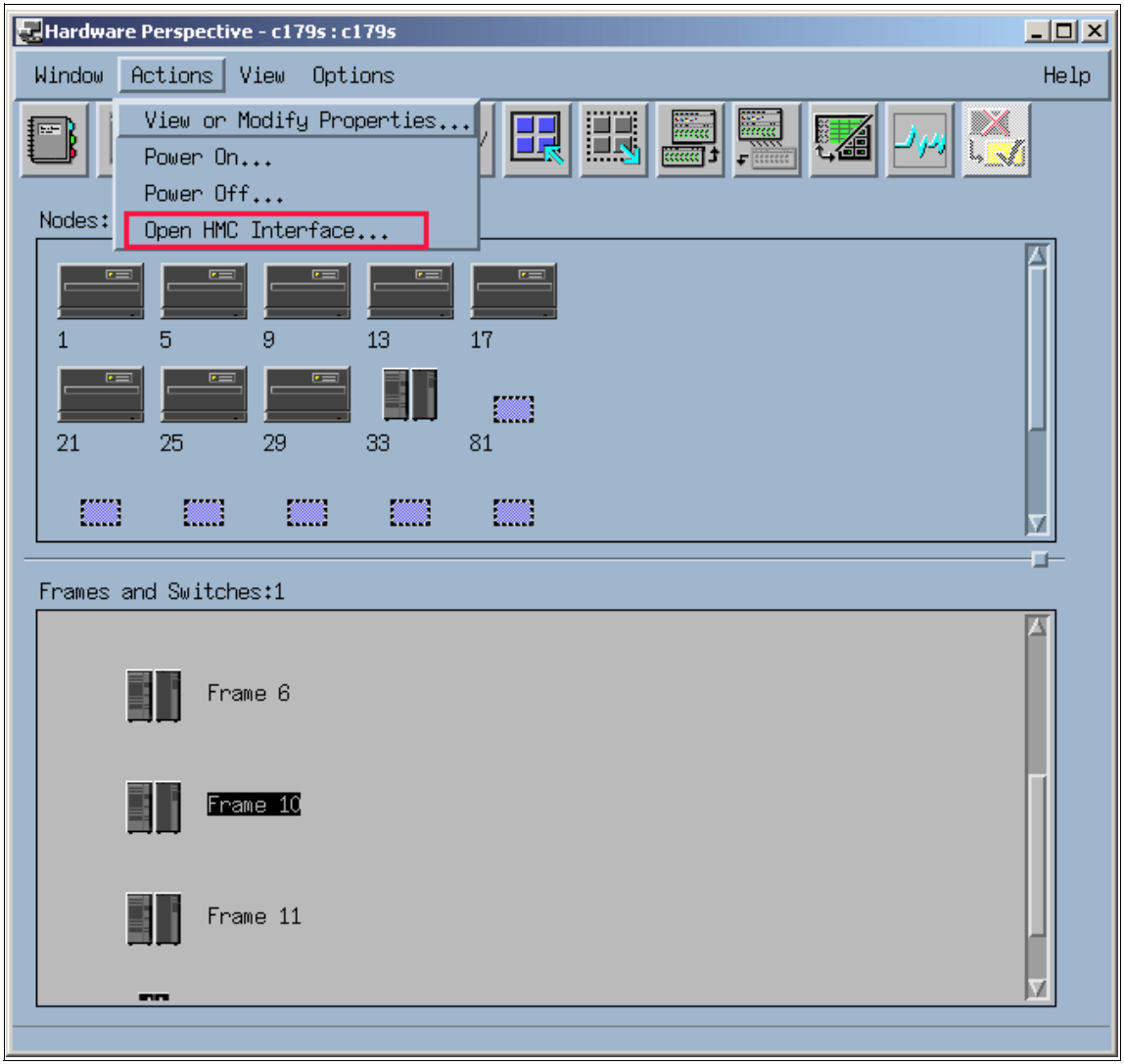


Figure 5-12 Open HMC Interface in Perspectives

The login screen for the HMC will then appear. You can log on with your *hscroot* user and manage the server, create or delete LPARs, and so on. See Figure 5-13 on page 205 for a look at the GUI interface.

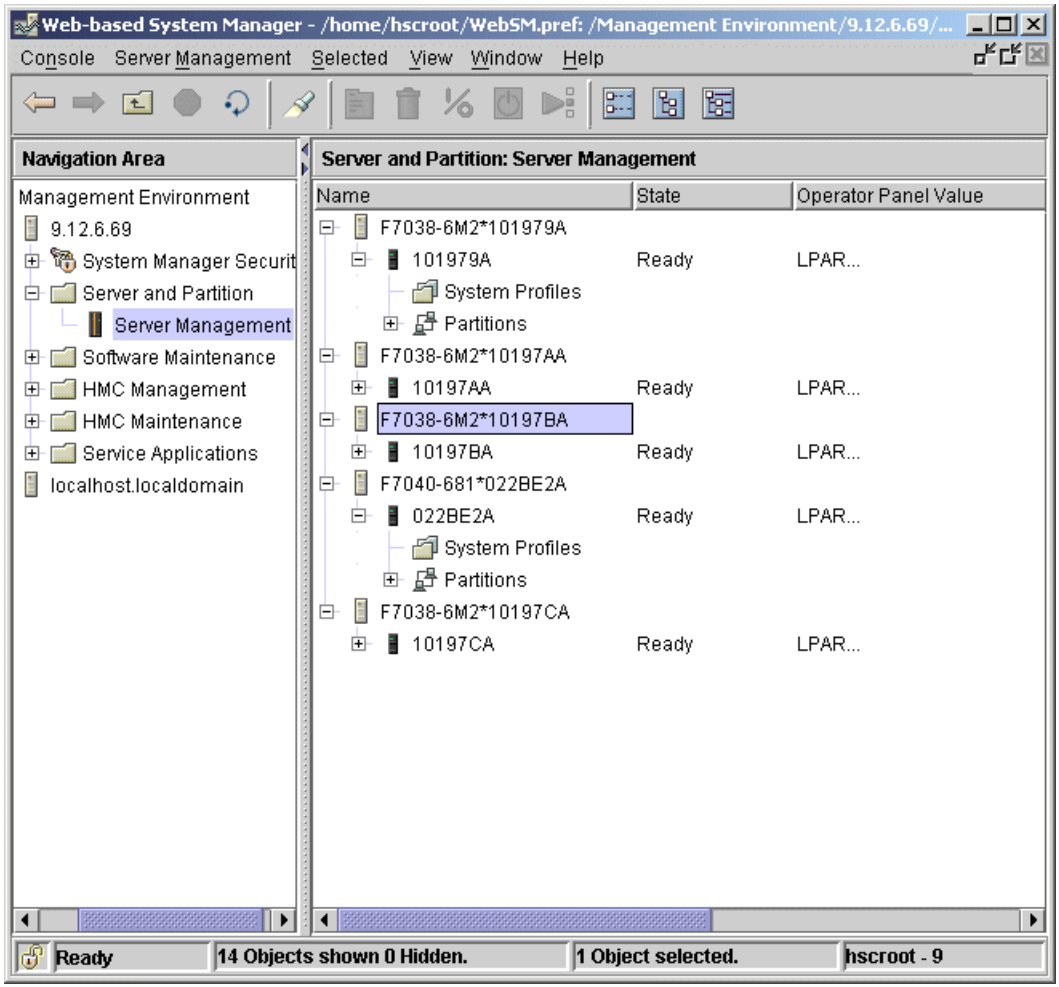


Figure 5-13 HMC GUI

The other method is to install the WebSM client on your workstation or PC and then access the HMC GUI through the network.

WebSM installation - access HMC GUI

Set up a Web server on an AIX 5.1 image (with APAR IY22854). Use the **`configassist`** command for this.

Remotely

- On the Windows® PC, start a Web browser.

- In the PC Web browser, enter the following URL for the AIX image you set up in the first step:

  `http://AIX_5.1_hostname/remote_client/remote_client.html`

- For SSL connectivity, you would then install the additional Remote Client security package:

  `http://AIX_5.1_hostname/remote_client/remote_client_security.html`

- If this is the first time you've accessed the Remote Client, you will be asked if you want to download the Remote Client environment. This will use a Java™ InstallShield application to install the Remote Client onto your PC.

- Install the Remote Client code. It will leave an icon on your desktop.

► On the windows PC, start a Web browser.

- In the PC Web browser, enter the following URL based on the HMC hostname:

  `http://HMC_hostname_or_IP_address/remote_client.html`

- Install the Remote Client code. It will leave an icon on your desktop.

Locally

► On the Web-Based System Manager server, go to the directory /usr/websm/remote_client. If this directory is not available or is empty, be sure you have the filesets installed that are shown in Table 5-7 on the Web-Based System Manager server.

Table 5-7 Needed filesets

Remote Client Source file	Fileset needed	Description
setup.exe	sysmgt.websm.webaccess	Contains the Remote Client
setupsec-ex.exe	sysmgt.websm.security	Needed to enable secure connections (export version)
setupsec-us.exe	sysmgt.websm.security-us	Needed to enable secure connections (U.S. version)

Copy the setup.exe or setupsec-ex.exe or setupsec-us.exe files onto removable media. Take the setup file from removable media to the Windows client and execute the setup scripts, according to the required security needs of the client.

5.7 Attachment scenarios

The following sections describe the various attachment scenarios of the SP-attached server to the SP system, but they do not show all the cable attachments between the SP frame and the SP-attached server.

Scenario 1: SP-attached server to a one-frame SP system

This scenario shows a single frame system with 14 thin nodes located in slots one through 14. The system has two unused node slots in position 15 and 16. These two empty node slots have corresponding switch ports that provide valid connections for the SP Attachment Adapter.

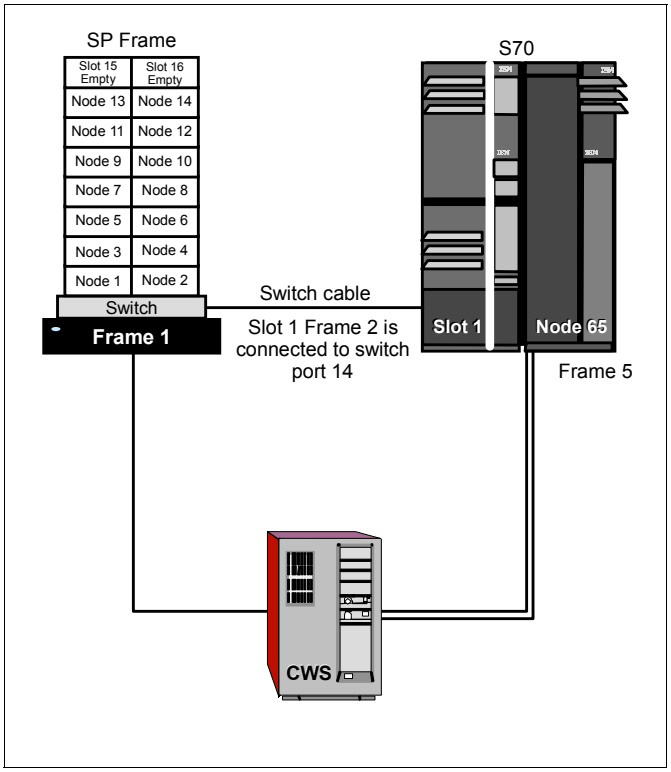


Figure 5-14 Scenario 1: SP-attached server and one SP frame

Scenario 2: SP-attached server to a two-frame SP system

This scenario (see Figure 5-15 on page 208) shows a two-frame system with four high nodes in each frame. This configuration uses eight switch ports and leaves eight valid switch ports available for future scalability. Therefore, it is important that the frame number assigned to the S70 must allow for extra non-switched frames (in this example, frames three and four), as the S70 frame must be

attached to the end of the configuration. On this basis, the S70 frame number must be at the very least 5 to allow for the two possible non-switch frames.

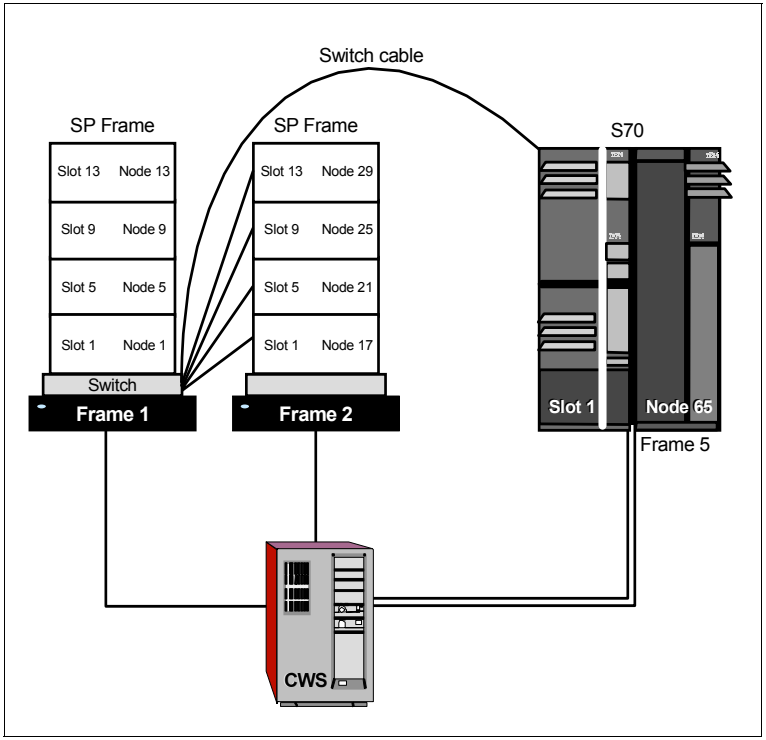


Figure 5-15 Scenario 2: SP-attached server to two SP frames

Note that the switch cable from frame 1 connects to the S70; for example, in this case, slot 1 frame 5 connects to switch port 3 of switch chip 5.

Scenario 3: One SP frame and multiple SP-attached servers

This scenario illustrates three important considerations:

1. The minimum requirement of one node in a frame to be able to attach one or more SP-attached servers to an SP system, since the SP-attached server cannot be the first frame in an SP environment.
2. It cannot interfere with the frame numbering of the expansion frames and, therefore, the SP-attached server is always at the end of the chain.
3. A switch port number must be allocated to each SP-attached server even though the SP system is switchless.

In this example, the first frame has a single thin node only, which is mandatory for any number of SP-attached servers. Frame 7 is an HMC-attached server M/T

7040 Model 690, which does not connect directly to the CWS. See Figure 5-16 for a overview.

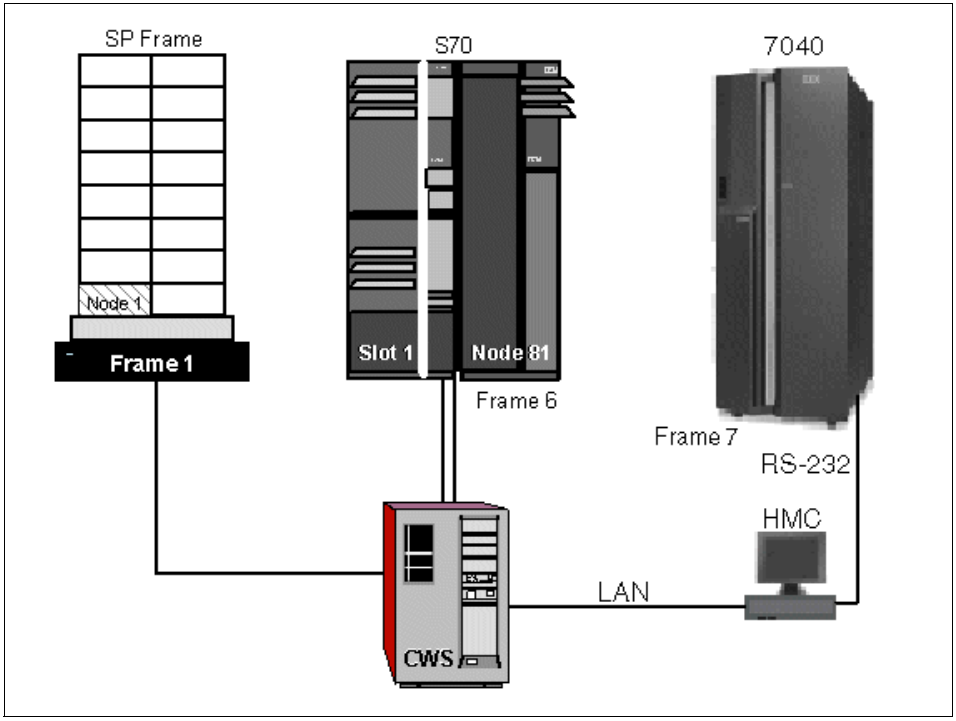


Figure 5-16 Scenario 3: SP frame and multiple SP-attached servers

Scenario 4: Non-contiguous SP-attached server configuration

Frame 1 and 3 of the SP system are switch-configured. Frame 3 is a non-switched expansion frame attached to frame 1. In this configuration, the SP-attached server could be given frame number 4, but that would forbid any future attachment of non-switched expansion frames to frame 1's switch. If, however, you assigned the SP-attached server frame number 15, your system could still be scaled using other switch-configured frames and non-switched expansion frames.

Frame 3 is another switch-configured frame, and the SP-attached server has previously been assigned frame number 10 for future scalability purposes.

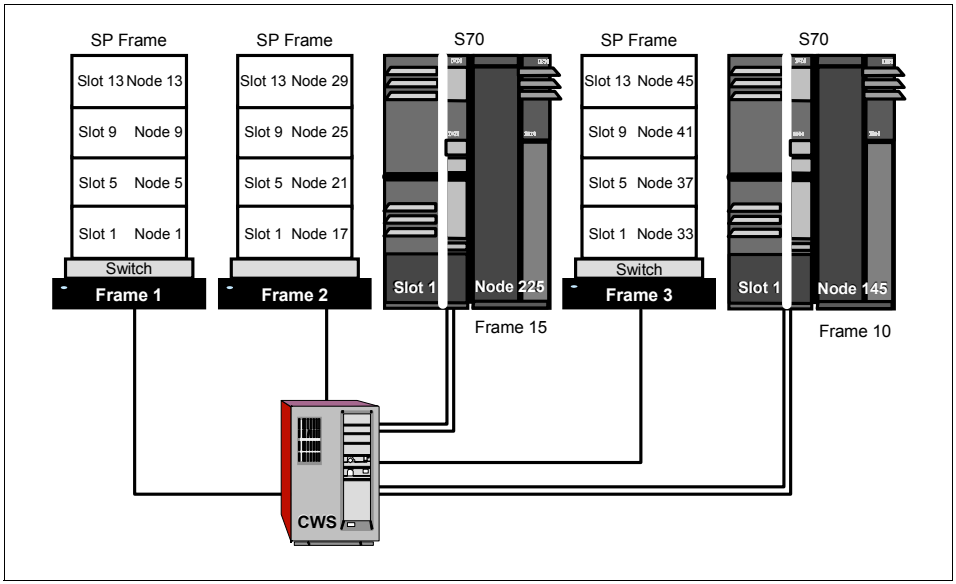


Figure 5-17 Scenario 4: Non-contiguous SP-attached server

For more information, see *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281.

5.8 Related documentation

These documents will help you understand the concepts and examples covered in this chapter in order to maximize your chances of success in the exam.

SP manuals

Chapter 15, “SP-attached Servers”, in *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281, provides some additional information regarding SP-attached servers.

SP redbooks

Chapter 4, “SP-attached Server Support”, in *PSSP 3.1 Announcement,* SG24-5332, provides some additional information on this topic.

5.9 Sample questions

This section provides a series of questions to aid you in preparation for the certification exam. The answers are in Appendix A, “Answers to sample questions” on page 521.

1. There must be three connections between the CWS and any SP-attached server. These are:
   a. A serial RS-232, an Ethernet and an SP Switch connection
   b. A serial RS-232, an Ethernet and a ground connection
   c. Two serial RS-232 and a ground connection
   d. Two serial RS-232 and an Ethernet connection
2. An SP-attached server is considered a node and also a frame. Which of the following statements are false?
   a. The node number for the SP-attached server is calculated based on the frame number.
   b. The frame number assigned to an SP-attached server cannot be 1.
   c. The SP-attached server cannot be installed between two switched frames.
   d. The SP-attached server cannot be installed between a switched frame and its expansion frames.
3. The SP-attached servers are considered standard nodes. However, there are some minor restrictions regarding system management. Which of the following statements are true?
   a. The SP-attached server does not have a frame or node supervisor card, which restrict the console access to a single session.
   b. The SP-attached server does not have a frame or node supervisor card, which limits the full hardware support, control, and monitoring capabilities of the server from the CWS.
   c. The CWS should have enough spare serial ports to connect the SP-attached server. Additional 16-port adapters may be required in order to provide the extra serial ports.
   d. The SP-attached server does not have a frame or node supervisor card, which restricts installation of SP-attached servers to one at a time.
4. The s70d daemon runs on the control workstation and communicates with the SP-attached server for hardware control and monitoring. Which of the following statements are false?
   a. The s70d is partition-sensitive; so, it will be one s70d daemon per SP-attached server per partition running on the CWS.

b. The s70d daemon is started and controlled by the hardmon daemon.

c. The s70d daemon uses SAMI protocol to connect to the SP-attached server's front panel.

d. One s70d daemon per SP-attached server runs on the CWS.

5. When connecting the SP-attached server to the SP frame, which of the following statements is *not* true?

   a. The tall frame with the eight port switch is not allowed.

   b. The SP system must be a tall frame because the 49 inch short LowBoy frames are not supported for the SP-attachment.

   c. A maximum of eight SP-attached servers are supported in one SP system.

   d. The SP-attached server can be the first frame in the SP system

6. Which of the following provides a mechanism to control and monitor SP-attached server Model 7026 p660?

   a. SAMI

   b. Hardmon

   c. HMC

   d. CSP

7. Which of the following is the minimum PSSP and AIX requirement for an SP-attached server Model 7028 p630?

   a. PSSP 3.4 and AIX 5L 5.1

   b. PSSP 3.4 and AIX 4.3.3

   c. PSSP 3.1 and AIX 4.3.2

   d. PSSP 3.5 and AIX 5L 5.2

8. Once the frame number has been assigned, the server's node numbers, which are based on the frame number, are automatically generated. Which of the following system defaults used for LPAR is true?

   a. Each SP-attached server's LPAR occupies one slot position.

   b. Each SP-attached server's LPAR is viewed as a single frame.

   c. Each SP-attached server's LPAR occupies the odd slot positions.

   d. Each SP-attached server's LPAR occupies the even slot positions.

9. For the S70 server, which of the following adapters is supported for SP-LAN communications?

   a. 100 Mbps AUI Ethernet

   b. 10 Mbps BNC

c. 100 Mbps BNC

d. ATM 155 TURBOWAYS® UTP

10. After configuring an SP-attached server into your SP system, which system management command displays information about the frames already installed?

    a. **splstframe**

    b. **spchvgobj**

    c. **spbootins**

    d. **splstdata**

5.10 Exercises

Here are some exercises you may wish to do:

1. Describe the necessary steps to change the switch port number of an SP-attached server.
2. What are the necessary steps to add an existing SP-attached server and preserve its current software environment?
3. Familiarize yourself with the various SP-attached server scenarios in 5.7, “Attachment scenarios” on page 207.

6

Cluster 1600 security

The intention of this chapter is not how to make your Cluster 1600 more secure nor explain the fundamentals of security. There are many great books out there that cover those topics and we provide references to some of them.

A Cluster 1600 system managed by PSSP is, practically speaking, nothing more than a bunch of pSeries nodes that are configured as a cluster and connected to one or more LANs. Therefore, it is exposed to the same security threats as any other network-connected workstation. No matter how much security you implement, it will not be effective unless it is implemented properly and administrated regularly.

You have to determine which security services software you need to obtain—in addition to the AIX and PSSP software—what is to be installed and configured, and at what point in the process. The Kerberos implementation that comes bundled with the PSSP software has been entrusted to perform authentication on the Cluster 1600 configuration, and the authenticated remote commands come with AIX. You need to plan and obtain, install, and configure any other security services that you choose to use.

6.1 Key concepts

Reading this chapter will not grant you any certification in security but it will cover some of the built-in security features that the Cluster 1600 managed by PSSP provides. Some of the key concepts of security in this environment are listed here in order of importance.

- Concepts of using Kerberos for authentication services on the cluster configuration. These include client/server activities, principals, realms, and tickets.
- Procedures for managing Kerberos that cover adding and deleting principals and authentication administrators.
- Concepts of sysctl as an PSSP Kerberos-based client/server system that runs commands remotely and in a parallel fashion.
- Procedures of sysctl authorization.
- Understanding how Kerberos provides better security services than standard AIX security.

6.2 Security-related concepts

A machine may be programmed to send information across the network impersonating another machine (which means assuming the identity of another machine or user). One way to protect the machines and users from being impersonated is to authenticate the packets that is travelling within the network.

Kerberos can provide such authentication services. However, the kerberized versions of **rsh** and **rcp** in PSSP only authenticate the connection, they do not encrypt the data that is actually sent after the authentications have been made. This means that, although the remote system is guaranteed to be the right one, the data that flows through the kerberized session to the network is still sent as clear readable text. Because of this, PSSP 3.4 introduced a feature called *secure remote command process*. We discuss this in 6.2.1, “Secure remote execution commands” on page 217.

In a Cluster 1600 environment managed by PSSP, three security mechanisms are supported:

- Standard AIX
- Kerberos V4
- DCE (using Kerberos V5)

Standard AIX

Standard AIX uses the normal UNIX facilities. When a host sends a request across the network, AIX only checks the originating IP address. Spoofing is possible by using the same IP address as the original system. Sniffing can, for example, be used to catch a password that was typed in during a Telnet session.

Kerberos V4

Kerberos V4 uses a secret key mechanism based on system IDs to identify all systems within its realm. The authentication process is encrypted, making it a secure way to prove the identity of your systems, but the actual data is not protected.

DCE

DCE is the most advanced method because it uses Kerberos V5. Kerberos V5 improvements over Kerberos V4 include:

- Use of an improved level of encryption
- More secure ticketing system (less vulnerable to sniffing)

These are related to three common defined security-related concepts. A brief description of these concepts and how they may be applied to the Cluster 1600 managed by PSSP system environment follows:

- Identification - This is a process by which one entity tells another who it is, that is, its identity. In the system environment, identification simply means a process that presents client identity credentials.
- Authentication - This is a process by which one entity verifies the identity of another. Identities and credentials are checked, but it does not add or restrict functions. In the system environment, authentication simply means a service requester's name and encrypted password are checked by using of available system utilities.
- Authorization - This process involves defining the functions that a user or process is permitted to perform. In the SP system environment, authorization simply means a service requester is granted permission to do a specific action, for example, execute commands remotely.

In a system environment, the server first identifies and authenticates the client and then checks its authorization for the function requested.

6.2.1 Secure remote execution commands

PSSP uses the AIX-authenticated remote commands (**rsh**, **rcp**, and **rlogin**), plus the **ftp** and **telnet** commands for system management. These commands are available for general use and support multiple authentication methods.

Some PSSP components rely on the ability to use the **rsh** command to issue remote commands as the root user from a node to be run on the control workstation (CWS), or from the CWS to be run on nodes.

Several components use the **rcp** command to copy files between the CWS and the nodes. In order to provide this capability, PSSP has effectively defined one root user across the entire system. Gaining access as root on any node implies that you can gain access as root on the CWS and all other nodes in the system. This may sound like a nightmare for some, especially for those where a single root user is not desirable, like in a server consolidation system. However, PSSP has support for restricted root access (RRA), which was introduced in PSSP 3.2.

The RRA option is available to limit the use of **rsh** and **rcp** within the Cluster 1600 PSSP system management. It restricts the PSSP and PSSP-related software from automatically using **rsh** and **rcp** commands as root user from a node. When restricted root access is active, any such actions can only be run from the control workstation or from nodes explicitly configured to authorize them. It restricts root **rsh** and **rcp** authorizations from the nodes to the CWS, but permits CWS-to-node **rsh** and **rcp** access. The drawback is that when restricted root access is enabled, programs like IBM Virtual Shared Disk (VSD) and GPFS that rely on the existence of a common PSSP root identity that can be authorized successfully under **rsh** and **sysctl** ACLs. So if you want to use restricted root access you cannot have VSD or GPFS enabled.

As of PSSP 3.4, you can have the PSSP software use a secure remote command process in place of the AIX **rsh** and **rcp** commands. You can acquire, install, and configure any secure remote command software of your choice that conforms to the IETF Secure Shell protocol. With restricted root access (RRA) and a secure remote command process enabled, the PSSP software has no dependency to internally issue **rsh** and **rcp** commands as a root user from the CWS to nodes, from nodes to nodes, or from nodes to the CWS.

When enabling this feature, commands such as **dsh**, **rsh** and **rcp** would get pointed to use the equivalent secure command, that is, **ssh** and **scp**, assuming that you are using secure remote command software. However, this is only supported for PSSP processes that use **dsh** in a background task. Implementing the secure remote process would make every **rsh** and **rcp** session encrypted (unless OpenSSH is implemented for authentication only, no data encryption).

.

Important: Secure remote process requires remote command software that conforms to the IETF Secure Shell protocol. Commands like **dsh**, **rsh**, and **rcp** will *only* use secure commands like **ssh** and **scp** if the PSSP processes use **dsh** in a background task.

Users exploiting **dsh** would have to set the needed environment variables in order to get **dsh** to use OpenSSH. In 6.2.2, “Using the secure remote command process” on page 220, you will find a small scenario (Example 6-1 on page 221) about how to execute a secure command using the secure remote command process with the needed environment variables configured. In our example, we use OpenSSH but you can use any secure remote command software that conforms to the Internet Engineering Task Force (IETF) Secure Shell working group protocol.

For more detailed information about restricted root access (RRA), as well as for the secure remote command process, see the chapter “Managing and using SP security services” in *PSSP Administration Guide,* SA22-7348, and *PSSP Installation and Migration Guide,* GA22-7347.

Since AIX 4.3 and prior releases, the commands **telnet** and **ftp**, as well as the r-commands **rcp**, **rlogin**, and **rsh**, have been enhanced to support multiple authentication methods (note that **rexec** is not included in this list). In earlier releases, the standard AIX methods were used for authentication and authorization, as follows:

telnet The **telnet** client establishes a connection to the server, which then presents the login screen of the remote machine, typically by asking for userid and password. These are transferred over the network and are checked by the server’s login command. This process normally performs both authentication and authorization.

ftp Again, userid and password are requested. Alternatively, the login information for the remote system (the server) can be provided in a $HOME/.netrc file on the local machine (the client), which is then read by the **ftp** client rather than querying the user. This method is discouraged since plain text passwords should not be stored in the (potentially remote) file system.

rexec Same as **ftp**. As mentioned above, use of $HOME/.netrc files is discouraged.

One of the main security concerns with this authentication for the above commands is the fact that passwords are sent in plain text over the network. They can easily be captured by any root user on a machine that is on the same network as where the connection is established.

rcp, rlogin, rsh The current user name (or a remote user name specified as a command line flag) is used, and the user is prompted for a password. Alternatively, a client can be authenticated by its IP name/address if it matches a list of trusted IP names/addresses that are stored in files on the server.

/etc/hosts.equiv lists the hosts from which incoming (client) connections are accepted. This works for all users except root (UID=0).

$HOME/.rhosts lists additional hosts, optionally restricted to specific userids, that are accepted for incoming connections. This is on a per-user basis and also works for the root user.

One primary security concern is *host impersonation*. It is relatively easy for an intruder to set up a machine with an IP name/address listed in one of these files and gain access to a system. Of course, if a password is requested rather than using $HOME/.rhosts or /etc/hosts.equiv files, this is sent in plain text over the network and eventually sniffed.

Remember that the .rhosts file also allows any user to grant password-free remote access on his/her account to any external computer without asking or informing the system administrator; things that are not considered good for security (nor for the ego of any system administrator). It would therefore make sense to remove any r commands, but in a Cluster 1600 managed by PSSP they are needed, and in any other cluster design they might be very useful. Experience shows that large networks have been penetrated in the past mainly because of hackers gaining access to these security files. However, the argument is pointless since both r commands, **ftp** and **telnet**, are just not the right applications to use when the network is untrusted, and this is why “security is trust”, and this is why we want to keep the SP Ethernet admin LAN trusted.

6.2.2 Using the secure remote command process

To determine whether you are using either AIX **rsh** or **rcp** or the secure remote command and copy method, the following environment variables are used. If no environment variables are set, the defaults are /bin/rsh and /bin/rcp.

You must be careful to keep these environment variables consistent. If setting the variables, all three should be set. The DSH_REMOTE_CMD and REMOTE_COPY_CMD executables should be kept consistent with the choice of the remote command method in RCMD_PGM:

- RCMD_PGM - remote command method, either **rsh** or **secrshell**
- DSH_REMOTE_CMD - remote command executable
- REMOTE_COPY_CMD - remote copy executable

In Example 6-1 on page 221 we run SYSMAN_test using a secure remote command method. SYSMAN_test saves its output in sm.errors in the current working directory. (The HN_METHOD=reliable environment variable can be used to run SYSMAN_test using the reliable_hostname instead of the default initial_hostname.)

Example 6-1 SYSMAN_test using secrshell

```
export RCMD_PGM=secrshell
export DSH_REMOTE_CMD=/bin/ssh
export REMOTE_COPY_CMD=/bin/scp
export HN_METHOD=reliable

SYSMAN_test -l sm.errors
```

For more information about PSSP security and secure remote command software, see “Control Workstation and Software Environment” in *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281, and *PSSP Installation and Migration Guide,* GA22-7347.

6.3 Defining Kerberos

Kerberos is a network authentication protocol, which is designed to provide strong authentication for client/server applications by using secret-key cryptography. It can also provide tools of authentication and strong cryptography over the network, which helps you secure your information across your cluster system so that data passing between workstations and servers is not corrupted either by accident or by tampering. However, in the PSSP implementation Kerberos is only used as a service for authenticating users in a network environment.

The authentication code that comes with PSSP is distributed in two separately installable options. The ssp.authent fileset contains only parts required on a system that is to be an authentication server. The remainder of Kerberos V4 authenticated services is distributed in the ssp.clients fileset.

Install ssp.clients on the CWS, even if you intend to use AFS or Kerberos authentication. The ssp.clients fileset needs to be installed on any other pSeries or RS/6000 workstation that you want to be an authentication server or from which you plan to perform system management tasks. In a cluster managed by PSSP all nodes will have ssp.clients installed.

For more detailed information about Kerberos planning, see “Security features of the PSSP software” in *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281.

Important: You must install ssp.clients on the control workstation, even if you intend to use AFS or Kerberos V4 authentication. If you want to use the control workstation as a Kerberos V4 authentication server, either primary or secondary, you must also install ssp.authent.

Authentication methods for a machine are selected by the AIX **chauthent** command and can be listed with the **lsauthent** command. These commands call the library routines set_auth_method() and get_auth_method(), which are contained in the library libauthm.a. Three options are available: **chauthent -std** enables standard AIX authentication; **chauthent -k5** and **chauthent -k4** enable Version 5 or 4 Kerberos authentication. More than one method can be specified, and authenticated applications or commands use them in the order specified by **chauthent** until one is successful (or the last available method fails, in which case access is denied). If standard AIX authentication is specified, it must always be the last method.

Note: In a Cluster 1600 managed by PSSP, the **chauthent** command should not be used directly. The authentication methods for cluster nodes and the CWS are controlled by the partition-based PSSP commands **chauthpar** and **lsauthpar**. Configuration information is stored in the Syspar SDR class in the auth_install, auth_root_rcmd, and auth_methods attributes.

6.4 Kerberos

The main reasons for using Kerberos are:

- Prevents unauthorized access to the system.
- Prevents non-encrypted passwords from being passed on the network.
- Provides security for remote commands, such as **rsh**, **rcp**, **dsh**, and **sysctl**. Descriptions of these commands are in Table 6-1.

The following table consists of basic Kerberos Terms.

Table 6-1 Basic Kerberos terms

Basic Kerberos Terms	Description
Principal	A Kerberos user or Kerberos ID. That is, a user who requires protected service.
Instance	The authority granted to the Kerberos user. Example for usage with a user: In root.admin, *root* is the principal, and *admin* is the instance which represents Kerberos authorization for administrative tasks. Example for usage with a service: In hardmon.sp3en0, *hardmon* represents the hardware monitor service, and *sp3en0* represents the machine providing the service

Realm	A collection of systems managed as a single unit. The default name of the realm on an SP system is the TCP/IP domain name converted to upper case. If DNS is not used, then the CWS hostname is converted to uppercase.
Authentication Server (Primary and secondary)	Host with the Kerberos database. This host provides the tickets to the principals to use. When running the setup_authent, program authentication services are initialized. At this stage, a primary authentication server must be nominated (this may be the CWS). A secondary authentication server may then be created later that serves as a backup server.
Ticket	An encrypted packet required for use of a Kerberos service. The ticket consists of the identity of the user. Tickets are by default stored in the /tmp/tkt<client's user ID> file.
Ticket-Granting Ticket (TGT)	Initial ticket given to the Kerberos principal. The authentication server site uses it to authenticate the Kerberos principal.
Service Ticket	Secondary ticket that allows access to certain server services, such as **rsh** and **rcp**.
Ticket Cache File	File that contains the Kerberos tickets for a particular Kerberos principal and AIX ID.
Service Keys	Used by the server sites to unlock encrypted tickets in order to verify the Kerberos principal.

6.4.1 Kerberos daemons

Kerberos authenticates information exchanged over the network. There are three daemons that deal with the Kerberos services. 6.4.2, "Kerberos authentication process" on page 224 illustrates the Kerberos authentication process.

The three Kerberos daemons are as follows:

kerberos This daemon only runs on the primary and secondary authentication servers. It handles getting ticket-granting and service tickets for the authentication clients. There may be more than one Kerberos daemon running on the realm to provide faster service, especially when there are many client requests.

kadmind This daemon only runs on the primary authentication server (usually the CWS). It is responsible for serving the Kerberos administrative

tools, such as changing passwords and adding principals. It also manages the primary authentication database.

kpropd This daemon only runs on secondary authentication database servers. When the daemon receives a request, it synchronizes the Kerberos secondary server database. The databases are maintained by the kpropd daemon, which receives the database content in encrypted form from a program, and kprop, which runs on the primary server.

6.4.2 Kerberos authentication process

Three entities are involved in the Kerberos authentication process: The client, the server, and the authentication database server. The following is an example of authentication:

1. The client (Host A) issues the **kinit** command that requests a ticket-granting ticket (TGT) to perform the **rcp** command on the destination host (Host B).

   For example:

   kinit root.admin and **rcp sp3en0:file**

2. The authentication database server (Host C) that is the Key Distribution Center (KDC) performs authentication tasks. If the information is valid, it issues a service ticket to the client (Host A).
3. The client (Host A) then sends the authentication and service ticket to the server (Host B).
4. The kshd daemon on the server (Host B) receives the request and authenticates it using one of the service keys. It then authorizes a Kerberos principal through the .klogin file to perform the task. The results of the **rcp** command are then sent to the client (Host A).

The Kerberos interaction is shown in Figure 6-1 on page 225.

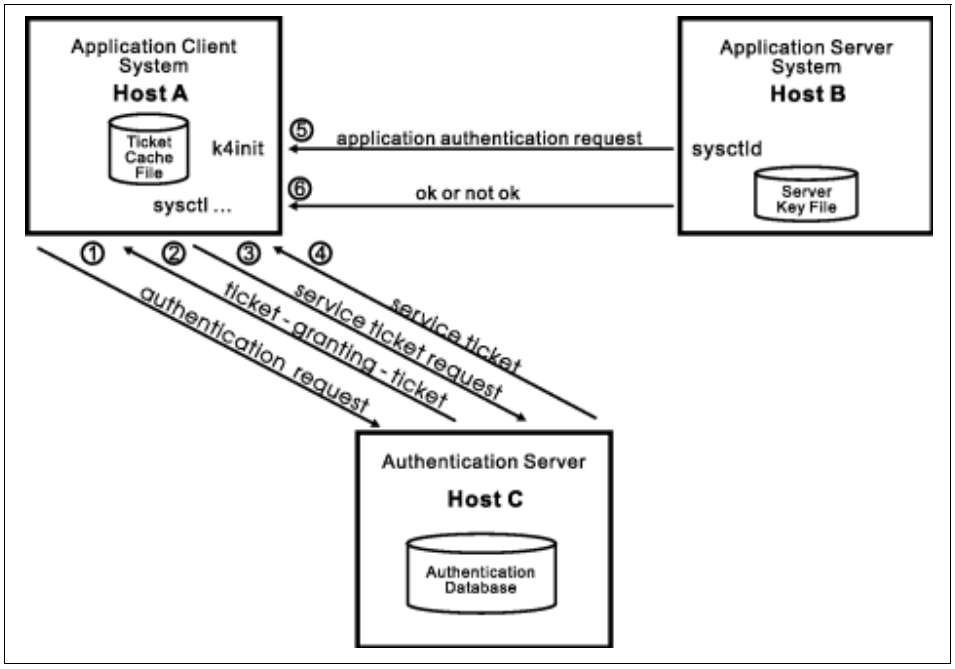


Figure 6-1 Overview of Kerberos V4 interaction with client and server

For more detailed information about Kerberos authentication, see the section "Kerberos V4 authentication" in *PSSP Administration Guide,* SA22-7348.

6.5 Kerberos paths, directories, and files

To avoid entering the complete path name each time you want to invoke a PSSP command, you should add the directories in the previous list to the user's path. For example, if using the Korn shell, add the directories to the path statement in his/hers .profile file, or in the system environment variable.

The following is an example of the roots user path statement:

```
PATH=$PATH:/usr/lpp/ssp/bin:/usr/lib/instl:/usr/sbin:
/usr/lpp/ssp/kerberos/bin
```

If you would like to use the man pages, set MANPATH as follows:

```
MANPATH=$MANPATH:/usr/lpp/ssp/man
```

Table 6-2 on page 226 displays the Kerberos directories and files on the Primary Authentication Server, which is usually the CWS.

Table 6-2 Kerberos directories and files on Primary Authentication Server

Directories and Files	Description
/.k	The master key cache file. Contains the DES key derived from the master password. The DES key is saved in /.k file using the **/usr/lpp/ssp/kerberos/etc/kstash** command. The kadmind daemon reads the master key from this file instead of prompting for the master password. After changing the master password, perform the following: Enter the **kstash** command to kill and restart kadmind daemon and to recreate the /.k file to store the new master key in it.
$HOME/.klogin	Contains a list of principals. For example, name.instance@realm. Listed principals are authorized to invoke processes as the owner of this file.
/tmp/tkt<uid>	Contains of the tickets owned by a client (user). The first ticket in the file is the TGT. The **kinit** command creates this file. The **klist** command displays the contents of the current cache file. The **kdestroy** command deletes the current cache file.
/etc/krb-srvtab	Contains the names and private keys of the local instances of Kerberos protected services. Every node and CWS, contains an /etc/krb-srvtab file that contains the keys for the services provided on that host. On the CWS the hardmon and rcmd service principals are in the file. They are used for SP system management and administration.
/etc/krb.conf	The first line contains the name of the local authentication realm. Subsequent lines specify the authentication server for a realm. For example, MSC.ITSO.IBM.COM MSC.ITSO.IBM.COM sp3en0.msc.itso.ibm.com admin server
/etc/krb.realms	Maps a host name to an authentication realm for the services provided by that host. Example of forms: host_name realm_name domain_name realm_name These are created by the setup_authent script on the primary authentication server.

Directories and Files	Description
/var/kerberos/database/*	This directory includes the authentication database created by setup_authent. Files residing in it include principal.pag and principal.dir; and access control lists for kadmin that are admin_acl.add, admin_acl.mod, and admin_acl.get.
/var/adm/SPlogs/kerberos/kerberos.log	This file records the kerberos daemon's process IDs and messages from activities.

Kerberos directories and files on the nodes are:

$HOME/.klogin

/etc/krb-srvtab

/etc/krb.conf

/etc/krb.realms

/tmp/tkt<uid>

6.6 Authentication services procedures

In this section we provide an overview of the required procedures to perform Kerberos authentication services.

1. Set up user accounts so that Kerberos credential can be obtained whenever a user logs in.
   - Add the name of the program that will execute the **kinit** command for the users in the /etc/security/login.cfg file. For example:

     ```
     program=/usr/lpp/ssp/kerberos/bin/k4init <program name>
     ```

   - Update the auth1 or auth2 attribute in the /etc/security/user file for each user account. For example:

     ```
     auth1=SYSTEM,Kerberos;root.admin
     ```

2. Log in to PSSP Kerberos authentication services.
   - Use the **k4init <principal>** command to obtain a ticket-granting ticket. For example, enter **k4init root.admin**.
   - Enter the password.
3. Display the authentication information.
   - Enter the command **k4list**.
4. Delete Kerberos tickets.

– Enter the command **k4destroy**.

– Verify that the tickets have been destroyed by entering the command **k4list** again.

Remember: You must define at least one user principal authorized to perform installation tasks. A system administrator, logged on as root, must assume the identity of this principal.

6.7 Kerberos passwords and master key

The following section describes the initial setup of passwords on the primary authenticator server:

During the installation stage, the **setup_authent** command is entered to configure the authentication services on the control workstation (CWS) and any other node connected to the cluster configuration. The **setup_authent** command gives an interactive dialog that prompts for two password entries:

- Master password (then the encrypted Kerberos master key will be written in the /.k file)
- Administrative principal's password

To change a principal's password:

Enter the **kpasswd** command to change a Kerberos principal's password. For example, to change the password of a current user, use this command.

To change the Kerberos master password:

1. Log in to Kerberos as the initial admin principal and enter the command:

   ```
   k4init root.admin
   ```

2. Change the password by entering the following command lines. The **kdb_util** command is used here to change the master key:

   ```
   kdbutil new_master_key /var/kerberos/database/newdb.$$
   kdb_util load /var/kerberos/database/newdb.$$
   ```

3. Replace the /.k file with the **kstash** command. This will store the new master key in the /.k file.

4. Kill and respawn the server daemons by entering the following command lines:

   ```
   stopsrc -s kerberos
   startsrc -s kerberos
   stopsrc -s kadmind
   startsrc -s kadmind
   ```

6.8 Kerberos principals

Kerberos principals are either users who use authentication services to run Kerberos-authenticated applications in the cluster system, or individual instances of the servers that run on the cluster nodes, the CWS, and on other nodes that have network connections to the Cluster 1600 configuration.

- User principals for system management

  A Cluster 1600 configuration managed by PSSP system must have at least one user principal defined. This user is the authentication database administrator who must be defined first so that other principals can be added later.

  When AFS authentication servers are being used, the AFS administrator ID already exists when the PSSP authentication services are initialized. When PSSP authentication servers are being used, one of the steps included in setting up the authentication services is the creation of a principal whose identifier includes the admin instance. It is suggested, but not essential, that the first authentication administrator also be the root user. For more detailed information about setting up authentication services, see “Relationship of PSSP Kerberos V4 and AFS” in *PSSP Administration Guide,* SA22-7348.

  Various installation tasks performed by root, or other users with UID 0, require the Kerberos authority to add service principals to the authentication database.

- Service principals used by PSSP components

  Two service names are used by the Kerberos-authenticated applications in an SP system:

  – hardmon is used by the System Monitor daemon on the CWS.

    The hardmon daemon runs only on the CWS. The SP logging daemon, splogd, can run on any other node in the cluster. Therefore, for each (short) network interface name on these nodes, a service principal is created with the name hardmon and the network name as the instance.

  – rcmd is used by sysctl.

    The remote commands can be run from, or to, any node on which the cluster system authenticated client services (ssp.clients) are installed. Therefore, for each (short) network interface name on all cluster nodes, the CWS, and other client systems, a service principal is created with the name rcmd and the network name as the instance.

6.8.1 Add a Kerberos principal

It is desirable to allow users to perform certain system tasks. Such users must be set up as Kerberos principals, and may include the following:

- Operators who use the **spmon** command to monitor system activities.
- Users who require extra security on the Print Management System when using it in open mode.
- System users who require partial root access. They may use the **sysctl** command to perform this. However, they must be set up as a Kerberos principal as well.

There are different ways to add Kerberos principals:

1. Use the **kadmin** command and its subcommand **add_new_key** (**ank** for short). This will always prompt for your administrative password.
2. Use the **kdb_edit** command. It allows the root user to enter this command without specifying the master key.
3. Use the **add_principal** command to allow a large number of principals to be added at one time.
4. Use the **mkkp** command to create a principal. This command is non-interactive and does not provide the capability to set the principal's initial password. The password must, therefore, be set by using the **kadmin** command and its subcommand **cpw**.
5. Add an Authentication Administrator.
   - Add a principal with an admin instance by using the **kadmin** command and its subcommand **add_new_key** (**ank** for short). For example:

     ```
     kadmin
     admin: add_new_key spuser1.admin
     ```

   - Add the principal identifier manually to one or more of the ACL files:

     /var/kerberos/database/admin_acl.add

     /var/kerberos/database/admin_acl.get

     /var/kerberos/database/admin_acl.mod

6.8.2 Change the attributes of the Kerberos principal

To change a password for a principal in the authentication database, a PSSP authentication database administrator can use either the **kpasswd** command or the kadmin program's **change_password** subcommand. You can issue these commands from any system running SP authentication services and that do not require a prior **k4init**.

To use the **kpasswd** command:

1. Enter the **kpasswd** command with the name of the principal whose password is being changed:

   ```
   kpasswd -n name
   ```

2. At the prompt, enter the old password.
3. At the prompt, enter the new password.
4. At the prompt, reenter the new password.

To use the kadmin program:

1. Enter the **kadmin** command:

   ```
   kadmin
   ```

   A welcome message and an explanation of how to ask for help are displayed.

2. Enter the **change_password** or **cpw** subcommand with the name of the principal whose password is being changed:

   ```
   cpw name
   ```

   The only required argument for the subcommand is the principal's name.

3. At the prompt, enter your admin password.
4. At the prompt, enter the principal's new password.
5. At the prompt, reenter the principal's new password.

To change your own admin instance password, you can use either the **kpasswd** command or the kadmin program's **change_admin_password** subcommand.

To use the **kpasswd** command:

1. Enter the **kpasswd** command with your admin instance name:

   ```
   kpasswd -n name.admin
   ```

2. At the prompt, enter your old admin password.
3. At the prompt, enter your new admin password.
4. At the prompt, reenter your new admin password.

To use the kadmin program:

1. Enter the **kadmin** command:

   ```
   kadmin
   ```

   A welcome message and explanation of how to ask for help are displayed.

2. Enter the `change_admin_password` or **cap** subcommand:

   ```
   cap
   ```

3. At the prompt, enter your old admin password.
4. At the prompt, enter your new admin password.
5. At the prompt, reenter your new admin password.

In addition to changing the password, you may want to change either the expiration date of the principal or its maximum ticket lifetime, though these are not so likely to be necessary. To do so, the root user on the primary authentication database system must use the **kdb_edit** command just as when adding new principals locally. The command finds if it already exists and prompts for changes to all its attributes starting with the password followed by the expiration date and the maximum ticket lifetime.

Use the **chkp** command to change the maximum ticket lifetime and expiration date for Kerberos principals in the authentication database. When logged into a system that is a Kerberos authentication server, the root user can run the **chkp** command directly. Additionally, any users who are Kerberos database administrators listed in the /var/kerberos/database/admin_acl.mod file can invoke this command remotely through a sysctl procedure of the same name.

The administrator does not need to be logged in on the server host to run **chkp** through sysctl but must have a Kerberos ticket for that admin principal (name.admin).

6.8.3 Delete Kerberos principals

There are two ways to delete principals. One is through the **rmkp** command, and another one is through the **kdb_util** command.

The following are the procedures to delete a principal with the **kdb_util** command and its subcommands.

1. The root user on the primary authentication server must edit a backup copy of the database and then reload it with the changed database. For example, in order to keep a copy of the primary authentication database in a file named slavesave in the /var/kerberos/database directory, enter the command:

   ```
   kdb_util dump /var/kerberos/database/slavesave
   ```

2. Edit the file by removing the lines for any unwanted principals.
3. Reload the database from the backup file by entering the command:

   ```
   kdb_util load /var/kerberos/database/slavesave
   ```

6.9 Server key

The server keys are located in the /etc/krb-srvtab file on the CWS and all the nodes. The file is used to unlock (decrypt) tickets coming in from client authentication.

- On the CWS, hardmon and rcmd service principals are in the file.
- On the nodes, rcmd service principals are in the file.
- The local server key files are created on the CWS by **setup_authent** during installation when authentication is first set up.
- The **setup_server** script creates server key files for nodes and stores them in the /tftpboot directory for network booting.
- Service Key information may be changed by using the command:

  ```
  ksrvutil change
  ```

- Service key information may be displayed by one of the following command lines. They will display information, such as the key version number, the service and its instance, and the realm name in some form.

  To view the local key file /etc/krb-srvtab, use:

  ```
  ksrvutil list
  k4list -srvtab
  krsvutil list -f /tftpboot/sp31n1-new-srvtab
  ```

6.9.1 Change a server key

A security administrator will decide how frequently service keys need to be changed.

The **ksrvutil** command is used to change service keys.

6.10 Using additional Kerberos servers

Secondary Kerberos authentication servers can improve security by providing backup to the primary authentication server and network load balancing. The kerberos and kprod daemons run on the secondary servers.

The tasks related to the Kerberos secondary servers are:

- Setting up a secondary Kerberos server
- Managing the Kerberos secondary server database

6.10.1 Set up and initialize a secondary Kerberos server

The following example provides the procedures to set up and initialize a secondary authentication server:

1. Add a line to the /etc/krb.conf file listing this host as a secondary server on the primary server.
2. Copy the /etc/krb.conf file from the primary authentication server.
3. Copy the /etc/krb.realms file from the primary server to the secondary server.
4. Run the setup_authent program following the prompt for a secondary server. (Note: It will also prompt you to log in as the same administrative principal name as defined when the primary server was set up.) The remainder of the initialization of authentication services on this secondary system takes place automatically.
5. After setup_authent completes, add an entry for the secondary authentication server to the /etc/krb.conf file on all SP nodes on which you have already initialized authentication.
6. If this is the first secondary authentication server, you should create a root crontab entry on the primary authentication server that invokes the script `/usr/kerberos/etc/push-kprop` that consists of the **kprop** command. This periodically propagates database changes from the primary to the secondary authentication server. Whenever the Kerberos database is changed, the **kprop** command may also be run to synchronize the Kerberos database contents.

6.10.2 Managing the Kerberos secondary server database

Both the kerberos and kpropd daemons run on the secondary authentication server and must be active all the time.

The kpropd daemon, which always runs on the secondary server, automatically performs updates on the secondary server database.

The kpropd daemon is activated when the secondary server boots up. If the kprod daemon becomes inactive, it may be automatically reactivated by the AIX System Resource Controller (SRC). That is, it may be restarted by using the **startsrc** command. The history of restarting the daemon is kept in the log file called /var/adm/SPlogs/kerberos/kprod.log.

6.11 SP services that utilize Kerberos

On a Cluster 1600 managed by PSSP, there are three different sets of services that use Kerberos authentication: The hardware control subsystem, the remote execution commands, and the sysctl facility. This section describes the authentication of these services and the different means they use to authorize clients that have been successfully authenticated.

6.11.1 Hardware control subsystem

The PSSP hardware control subsystem is implemented through the hardmon and splogd daemons, which run on the CWS and interface with the cluster hardware through the serial lines (RS-232). To secure access to the hardware, Kerberos authentication is used, and authorization is controlled through hardmon-specific Access Control Lists (ACLs). PSSP 3.1 and earlier releases only support Kerberos Version 4, not Version 5 authentication.

The following commands are the primary clients to the hardware control subsystem:

- **hmmon** - Monitors the hardware state.
- **hmcmds** - Changes the hardware state.
- **s1term** - Provides access to the node's console.
- **nodecond** - For network booting, uses **hmmon**, **hmcmds**, and **s1term**.
- **spmon** - Some parameters are used to monitor; some are used to change the hardware state. The **spmon -open** command opens an s1term connection.

Other commands, like **sphardware** from the SP Perspectives, communicate directly with an internal hardmon API that is also Kerberos Version 4 authenticated.

To Kerberos, the hardware control subsystem is a service represented by the principal name hardmon. PSSP sets up one instance of that principal for each network interface of the CWS, including IP aliases in case of multiple partitions. The secret keys of these hardmon principals are stored in the /etc/krb-srvtab file on the CWS. The **k4list -srvtab** command shows the existence of these service keys; see Example 6-2.

Example 6-2 k4list -srvtab

```
[c179s][/]> k4list -srvtab
Server key file:   /etc/krb-srvtab
Service         Instance        Realm      Key Version
-----------------------------------------------------
rcmd            c179s           PPD.POK.IBM.COM 1
```

```
hardmon         c179s           PPD.POK.IBM.COM 1
hardmon         c179cw          PPD.POK.IBM.COM 1
rcmd            c179cw          PPD.POK.IBM.COM 1
root            SPbgAdm         PPD.POK.IBM.COM 1
```

The client command in Example 6-2 performs a Kerberos Version 4 authentication. It requires that the user who invokes it has signed on to Kerberos with the **k4init** command and passes the user's Ticket-Granting Ticket to the Kerberos server to acquire a service ticket for the hardmon service. This service ticket is then presented to the hardmon daemon, which decrypts it using its secret key, stored in the /etc/krb-srvtab file.

Authorization to use the hardware control subsystem is controlled through entries in the /spdata/sys1/spmon/hmacls file. Example 6-3 shows the first 5 lines that are read by hardmon when it starts up. Since hardmon runs only on the CWS, this authorization file only exists on the CWS.

Example 6-3 `head -5 /spdata/sys1/spmon/hmacls`

```
[c179s][/]> head -5 /spdata/sys1/spmon/hmacls
c179s.ppd.pok.ibm.com root.admin a
c179s.ppd.pok.ibm.com root.SPbgAdm a
1 root.admin vsm
1 root.SPbgAdm vsm
1 hardmon.c179s vsm
```

Each line in the file lists an object, a Kerberos principal, and the associated permissions. Objects can either be host names or frame numbers. By default, PSSP creates entries for the CWS and for each frame in the system, and the only principals that are authorized are root.admin and the instance of hardmon for the SP Ethernet adapter. There are four different sets of permissions indicated by a single lowercase letter:

- v (Virtual Front Operator Panel) - Control/change hardware status.
- s (S1) - Access node's console through the serial port (s1term).
- m (Monitor) - Monitor hardware status.
- a (Administrative) - Use hardmon administrative commands.

Note that for the CWS, only administrative rights are granted. For frames, the monitor, control, and S1 rights are granted. These default entries should never be changed. However, other principals might be added. For example, a site might want to grant operating personnel access to the monitoring facilities without giving them the ability to change the state of the hardware or access the nodes' console.

Note: Refreshing hardmon:

When the hmacls file is changed, the **hmadm setacls** command must be issued on the CWS to notify the hardmon daemon of the change and cause it to reread that file. The principal that issues the **hmadm** command must have administrative rights in the original hmacls file; otherwise, the refresh will not take effect. However, hardmon can always be completely stopped and restarted by the root user. This will reread the hmacls file.

Care must be taken if any of the hardware monitoring or control commands are issued by users that are authenticated to Kerberos but do not have the required hardmon authorization. In some cases, an error message will be returned as shown in Example 6-4.

Example 6-4 `hmmon: 0026-614`

```
hmmon: 0026-614 You do not have authorization to access the Hardware
Monitor.
```

In other cases, no misleading error messages may be returned. This mostly happens when the principal is listed in the hmacls file but not with the authorization required by the command.

In addition to the above commands, which are normally invoked by the system administrator, two SP daemons are also hardmon clients: The splogd daemon and the hmrmd daemon. These daemons use two separate ticket cache files: /tmp/tkt_splogd and /tmp/tkt_hmrmd. Both contain tickets for the hardmon principal, which can be used to communicate with the hardmon daemon without the need to type in passwords.

6.11.2 Remote execution commands

In PSSP the authenticated r-commands in the base AIX operating system are used. They can be configured for multiple authentication methods, including the PSSP implementation of Kerberos Version 4. To allow applications that use the full PSSP paths to work properly, the PSSP commands **rcp** and **remsh**/**rsh** have not been removed but have been replaced by symbolic links to the corresponding AIX commands. This calling structure is shown in Figure 6-2 on page 238. We look at some of the implementation details of the security integration of the r-commands, focussing on the AIX **rsh** command and the corresponding rshd and krshd daemons.

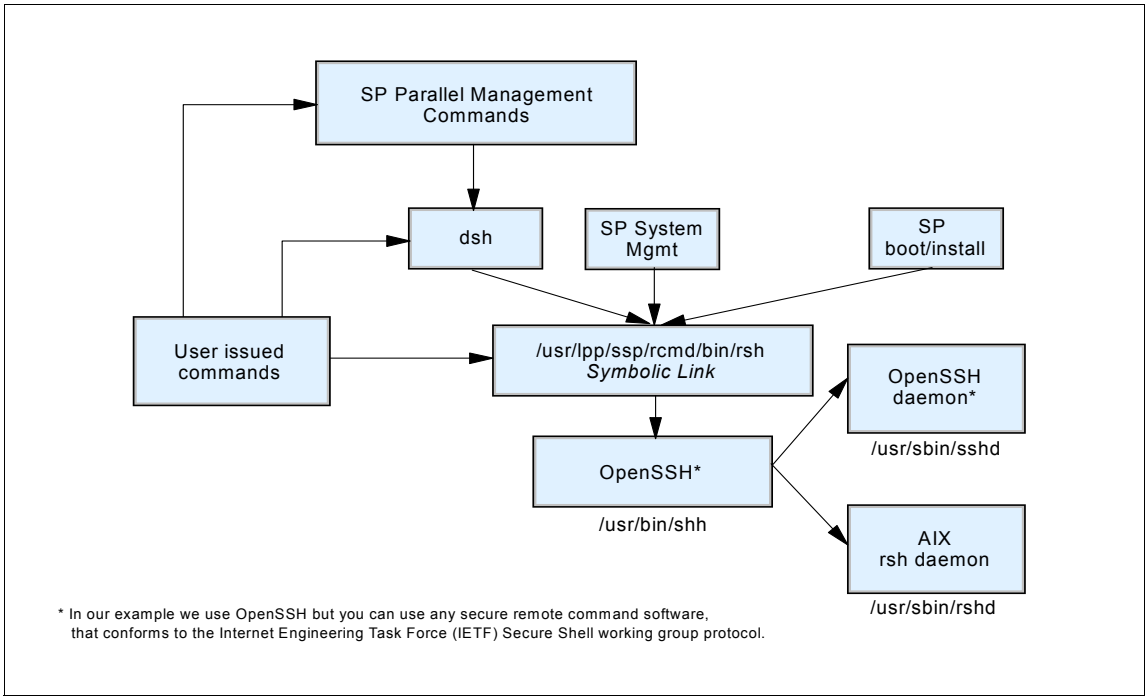


Figure 6-2 Remote shell structure in PSSP 3.5

Control flow in the rsh command

The full syntax of the AIX authenticated **rsh** command is:

```
/usr/bin/rsh RemoteHost [-n] [-l RemoteUser][-f|-F] [-k Realm] [Command]
```

Here, we assume that a command is present. When the **rsh** command is called, it issues the get_auth_method() system call, which returns the list of authentication methods that are enabled on the machine. It then attempts a remote shell connection using these methods, in the order they are returned, until one of the methods succeeds or all have failed.

K5MUTE: Authentication methods are set on a system level, not on a user level. This means that, for example, on an Cluster 1600 where Kerberos Version 4 and Standard AIX are set, a user's **rsh** command will produce a Kerberos authentication failure if that user has no Kerberos credentials. After that failure, the **rsh** attempts to use the standard AIX methods. The delay caused by attempting both methods cannot be prevented, but there is a means to suppress the error messages of failed authentication requests, which may confuse users: By setting the environment variable K5MUTE=1, these messages will be suppressed. Authorization failures will still be reported, however.

This is what happens for the three authentication methods:

STD When the standard AIX authentication is to be used, **rsh** uses the rcmd() system call from the standard C library (libc.a). The shell port (normally 514/tcp) is used to establish a connection to the /usr/sbin/rshd daemon on the remote host. The name of the local user, the name of the remote user, and the command to be executed are sent. This is the normal BSD-style behavior.

K5 For Kerberos Version 5 authentication, the kcmd() system call is issued (this call is not provided in any library). It acquires a service ticket for the /.:/host/<ip_name> service principal from the Kerberos Version 5 server over the Kerberos port (normally 88). It then uses the kshell port (normally 544/tcp) to establish a connection to the /usr/sbin/krshd daemon on the remote host. In addition to the information for STD authentication, kcmd() sends the Kerberos Version 5 service ticket for the rcmd service on the remote host for authentication. If the **-f** or **-F** flag of **rsh** is present, it also *forwards* the Ticket-Granting Ticket of the principal that invoked **rsh** to the krshd daemon. Note that ticket forwarding is possible with Kerberos Version 5 but not with Version 4.

K4 Kerberos Version 4 authentication is provided by the PSSP software. The system call spk4rsh(), contained in libspk4rcmd.a in the ssp.client fileset, is invoked by the AIX **rsh** command. It acquires a service ticket for the rcmd.<ip_name> service principal from the Kerberos Version 4 server over the kerberos4 port 750. Like kcmd(), the spk4rsh() subroutine uses the kshell port (normally 544/tcp) to connect to the /usr/sbin/krshd daemon on the remote host. It sends the STD information and the Kerberos Version 4 rcmd service ticket but ignores the **-f** and **-F** flags since Version 4 Ticket-Granting Tickets cannot be forwarded.

These requests are then processed by the rshd and krshd daemons.

The standard rshd daemon

The /usr/sbin/rshd daemon listening on the shell port (normally 514/tcp) of the target machine implements the standard, BSD-style **rsh** service. Details can be found in “rshd Daemon” in *AIX 5L Version 5.2 Commands Reference, Volume 4, N-R,* SC23-4118. The rshd daemon does the following:

- Does some health checks, such as verifying that the request comes from a well-known port.
- Verifies that the local user name (remote user name from the client’s view) exists in the user database and gets its UID, home directory, and login shell.
- Performs a chdir() to the user’s home directory (terminates if this fails).
- If the UID is not zero, rshd checks whether the client host is listed in /etc/hosts.equiv.
- If the previous check is negative, rshd checks whether the client host is listed in $HOME/.rhosts.
- If either of these checks succeeded, rshd executes the command under the user’s login shell.

Be aware that the daemon itself does not call the get_auth_method() subroutine to check if STD is among the authentication methods. The **chauthent** command simply removes the shell service from the /etc/inetd.conf file when it is called without the **-std** option; so, inetd will refuse connections on the shell port. But if the shell service is enabled again by editing /etc/inetd.conf and refreshing inetd, the rshd daemon honors requests even though **lsauthent** still reports that Standard AIX authentication is disabled.

The kerberized krshd daemon

The /usr/sbin/krshd daemon implements the kerberized remote shell service of AIX. It listens on the kshell port (normally 544/tcp) and processes the requests from both the kcmd() and spk4rsh() client calls.

In contrast to rshd, the krshd daemon actually uses get_auth_methods() to check if Kerberos Version 4 or 5 is a valid authentication method. For example, if a request with a Kerberos Version 4 service ticket is received, but this authentication method is not configured, the daemon replies with:

```
krshd: Kerberos 4 Authentication Failed: This server is not configured to
support Kerberos 4.
```

After checking if the requested method is valid, the krshd daemon then processes the request. This, of course, depends on the protocol version.

Handling Kerberos Version 5 requests

To authenticate the user, krshd uses the Kerberos Version 5 secret key of the host/<ip_hostname> service and attempts to decrypt the service ticket sent by the client. If this succeeds, the client has authenticated itself.

The daemon then calls the kvalid_user() subroutine, from libvaliduser.a, with the local user name (remote user name from the client's view) and the principal's name. The kvalid_user() subroutine checks whether the principal is authorized to access the local AIX user's account. Access is granted if one of the following conditions is true:

1. The $HOME/.k5login file exists and lists the principal (in Kerberos form).
2. The $HOME/.k5login file does not exist, and the principal name is the same as the local AIX user's name.

Case (1) is what is expected. But, be aware that case (2) above is quite counter-intuitive: It means that if the file does exist and is empty, access is denied, but if it does not exist, access is granted. This is the complete reverse of the behavior of both the AIX $HOME/.rhosts file and the Kerberos Version 4 $HOME/.klogin file. However, it is documented to behave this way (and actually follows these rules) in the kvalid_user() man page.

If the authorization check has passed, the krshd daemon checks if a Kerberos Version 5 TGT has been forwarded. If this is the case, it calls the **`k5dcelogin`** command, which upgrades the Kerberos TGT to full DCE credentials and executes the command in that context. If this **`k5dcelogin`** cannot be done because no TGT was forwarded, the user's login shell is used to execute the command without full DCE credentials.

DFS home directories: Note that this design may cause trouble if the user's home directory is located in DFS. Since the kvalid_user() subroutine is called by krshd before establishing a full DCE context via **`k5dcelogin`**, kvalid_user() does not have user credentials. It runs with the machine credentials of the local host and can only access the user's files if they are open to the other group of users. The files do not need to be open for the any_other group (and this would not help, either) since the daemon always runs as root and, therefore, has the hosts/<ip_hostname>/self credentials of the machine.

Handling Kerberos Version 4 requests

To authenticate the user, krshd uses the Kerberos Version 4 secret key of the rcmd.<ip_hostname> service and attempts to decrypt the service ticket sent by the client. If this succeeds, the client has authenticated itself.

The daemon then checks the Kerberos Version 4 $HOME/.klogin file and grants access if the principal is listed in it. This is all done by code provided by the PSSP software, which is called by the base AIX krshd daemon.

Note: PSPP still includes the **/usr/lpp/ssp/rcmd/bin/rcmdtgt** command, which can be used by the root user to obtain a ticket-granting ticket by means of the secret key of the rcmd.<localhost> principal stored in /etc/krb-srvtab.

NIM and remote shell

There is one important exception to keep in mind with respect to the security integration of the **rsh** command: When using boot/install servers, NIM will use a remote shell connection from the boot/install server to the CWS to update status information about the installation process that is stored on the CWS. This connection is made by using the rcmd() system call rather than the authenticated **rsh** command. The rcmd() system call always uses standard AIX authentication and authorization.

To work around this problem, PSSP uses the authenticated **rsh** command to temporarily add the boot/install server's root user to the .rhosts file of the CWS and removes this entry after network installation.

6.12 Sysctl is a PSSP Kerberos-based security system

The sysctl security system can provide root authority to non-root users based on their authenticated identity and the task they are trying to perform. Sysctl can also be run as a command line command. Usage of sysctl on Cluster 1600 systems is optional.

6.12.1 Sysctl components and process

The server daemon for the sysctl server is *sysctld*. It runs on all nodes and the CWS. It also contains built-in commands, configuration files, access control lists (ACL), and client programs.

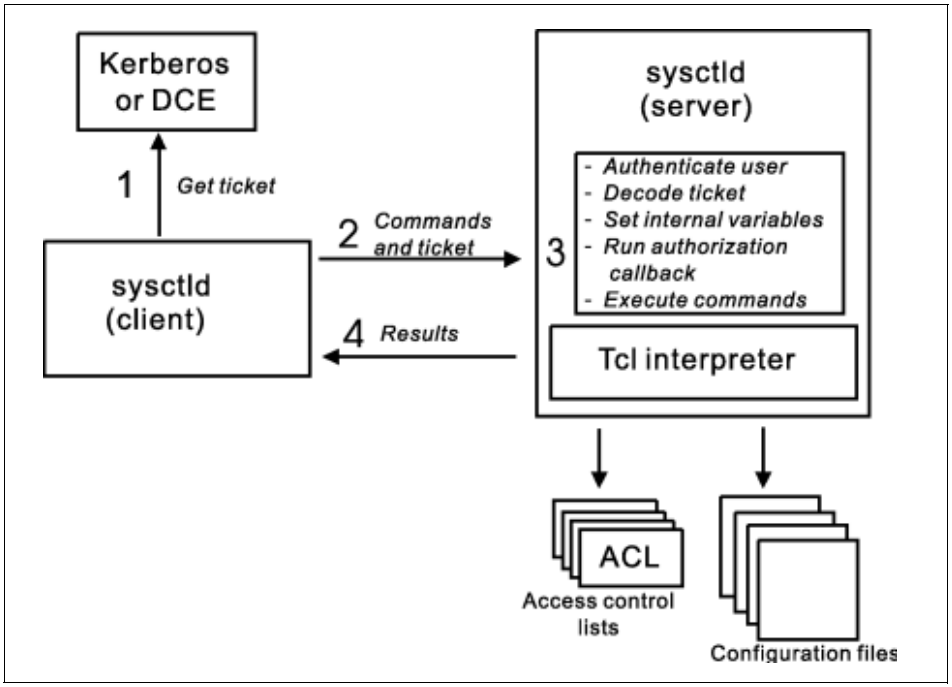


Figure 6-3 Sysctl elements

The following is the sysctl process depicted in Figure 6-3:

1. The sysctl client code gets its authentication information from the PSSP authentication services, Kerberos.
2. The sysctl client code sends the authentication information with the Service Tickets and commands to the specified sysctl server.
3. The server then performs the following tasks:
   - Authenticates the clients.
   - Decodes service the ticket.
   - Performs an authorization callback.
   - Executes commands as root.
4. **stdout** and **stderr** are sent back to the client. Output from each server is displayed with labeling indicating its source.

6.12.2 Terms and files related to the sysctl process

- Authorization callback - Once the client has been authenticated, the sysctl server invokes the authorization callbacks just before executing the commands.
- Access control lists (ACL) - These are text-based files that are used to give authority to specific users to execute certain commands.
- Configuration files - There are two main configuration files related to sysctl:
  - The /etc/sysctl.conf file configures the local sysctl server daemon by optionally creating variables, procedures, and classes; setting variables; loading shared libraries; and executing **sysctl** commands. The /etc/sysctl.conf file is on every machine that runs the sysctld daemon.
  - The /etc/sysctl.acl file contains the list of users authorized to access objects that are assigned the ACL authorization callback.
- Tcl-based set of commands - Access to this is provided by the sysctld daemon. These can be separated into the following three classes:
  - Base **Tcl** commands - These are the basic Tcl interpreter commands. They are also defined in the /etc/sysctl.conf file.
  - Built-in **sysctl** commands - These are Tcl-based IBM-written applications ready to be used by the sysctl programmer. These ACL processing commands include **acladd**, **aclcheck**, **aclcreate**, **acldelete**, and so on.
  - User-written scripts - These are programmer-written applications that use the base Tcl commands and built-in sysctl commands.

6.13 Related documentation

The following books provide further explanations of the key concepts discussed in this chapter.

SP manuals

- *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281

  The Planning for security section covers the planning task for:

  - Choosing authentication options
  - Understanding which security services software to acquire
  - DCE authentication
  - Kerberos V4
  - Standard AIX authentication

- *PSSP Administration Guide,* SA22-7348

  The Security section covers the security features of the PSSP software and includes conceptual information regarding:

  - SP security terminology
  - SP security concepts
  - Concepts specific to DCE
  - Concepts specific to Kerberos V4
  - Controlling remote execution by using sysctl

- *PSSP Installation and Migration Guide,* GA22-7347

  In discussing installing and configuring a new SP system, it covers the tasks for:

  - Setting authentication methods for AIX remote commands on the CWS
  - Initializing SP Kerberos V4 (optional)
  - Configuring DCE for the CWS (required for DCE)
  - Setting the authentication method for SP Trusted Services on the CWS

SP redbooks

Aditional AIX Security Tools on IBM pSeries, RS/6000, and SP/Cluster, SG24-5971

Exploiting RS/6000 SP Security: Keeping It Safe, SG24-5521

AIX 4.3 Elements of Security Effective and Efficient Implementation, SG24-5962

Other books

AIX 5L V 5.2 Security Guide, SC23-4860-01

AIX 5L V 5.2 Installation Guide and Reference, SC23-4389

PSSP Implementing a Firewalled RS/6000 SP System, GA22-7874

RS/6000 SP Cluster: The Path to Universal Clustering, SG24-5374

Other non-IBM books

SSH, The Secure Shell, O'Reilly

6.14 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers are in Appendix A, “Answers to sample questions” on page 521.

1. On the Cluster 1600 managed by PSSP, the AIX command **chauthent** should not be used directly because:
   a. It is not supported on Cluster 1600 environments.
   b. It does not provide Kerberos v4 authentication.
   c. The rc.sp script will reset any change made locally using the **chauthent** command.
   d. The rc.sp script will fail if the **chauthent** command is used on a node.
2. The /etc/krb-srvtab file contains:
   a. The ticket-granting ticket (TGT)
   b. The list of principals authorized to invoke remote commands
   c. The master key encrypted with the root.admin password
   d. The private Kerberos keys for local services
3. Which of the following is *not* a Kerberos client in a standard PSSP implementation?
   a. IBM SP Perspectives
   b. The hardmon daemon
   c. Remote shell (rsh)
   d. The system control facility (sysctl)
4. Which service names are used by the Kerberos-authenticated applications in an Cluster 1600 system? (Select two.)
   a. hardmon
   b. rsh
   c. rcp
   d. rcmd
5. Which of the following statements is used to add a Kerberos Principal?
   a. Remote shell (rsh).
   b. Use the **mkkp** command to create a principal.
   c. Use the **kerberos_edit** command.
   d. The system control facility (sysctl).
6. Which of the following Cluster 1600 managed by PSSP services does *not* use Kerberos authentication?
   a. The sysctl facility
   b. The hardware control subsystem

c. The remote execution commands

d. Service and Manufacturing Interface

7. Which of the following authentication methods is optional for a Cluster 1600 managed by PSSP configuration?

   a. STD

   b. K4

   c. AFS

   d. K5

8. Which of the following is *not* a Kerberos daemon?

   a. kpropd

   b. kinit

   c. kadmind

   d. kerberos

9. After changing the master password, the administrator enters the **kstash** command. Which of the following statements is true?

   a. The command will propagate the new password to the secondary authentication servers.

   b. The command deletes the current cache file.

   c. The command stores the new master key in the /kstash file.

   d. The command kills and restarts the kadmin daemon.

6.15 Exercises

Here are some exercises you may wish to do:

1. On a test system that does not affect any users, configure the Cluster 1600 managed by PSSP authentication services on the control workstation (CWS) and those nodes that are connected to the cluster. Change the principal's password, change the Kerberos master password, store the new master key, and stop and start the server daemons for the changes to take effect.
2. On a test system that does not affect any users, add a Kerberos principal.
3. On a test system that does not affect any users, change the attributes of the Kerberos principal.
4. Delete the above created Kerberos principal.
5. On a test system that does not affect any users, set up and initialize a secondary Kerberos server.

7

User and data management

This chapter covers user management that consists of adding, changing, and deleting users on the Cluster 1600 managed by PSSP, as well as how to control user login access using the SP user component.

Data and user management using the file collections facility is also covered. File collection provides the ability to have a single point of update and control of file images that will then be replicated across nodes.

AIX Automounters is briefly discussed. These allow users local access to any files and directories no matter which node they are logged in to.

NIS and AFS is widely used in cluster configurations. While no dependencies or adverse interactions between PSSP and NIS+ or AFS have been identified, the use of NIS+ and/or AFS on the Cluster 1600 managed by PSSP system has not been tested. Therefore, we only briefly cover these two topics.

7.1 Key concepts

The key concepts on user and data management are as follows, in order of importance:

- Considerations for administering SP users and SP user access control, and procedures for doing it.
- File collections and how this works in data management in the SP system.
- How to work with and manage file collections, and procedures to build and install them.
- The concepts of AIX Automounter and how it manages mounting and unmounting activities using NFS facilities.

7.2 Administering users on a Cluster 1600 system

Table 7-1 shows the issues and solutions concerning user and data management. You need to consider these when installing or configuring a Cluster 1600 managed by PSSP system.

Table 7-1 Issues and solutions when installing a Cluster 1600 PSSP system

Issues	Solutions
How to share common files across the SP system	File collections NIS
How to maintain a single user space	File collections NIS AMD AIX Automounter
Within a single user space, how to restrict access to individual nodes	Login control
Where should user's home directories reside?	Control Workstation (CWS) Nodes Other network systems
How does a user change access data?	AMD AIX Automounter
How does a user change the password?	File collections NIS
How to keep access to nodes secure	Kerberos AIX Security

SP User Management (SPUM) must be set up to ensure that there is a single user space across all nodes. It ensures that users have the same account, home directory, and environment across all nodes in the cluster configuration.

7.3 SP User data management

The following three options may be used to manage the user data on the Cluster 1600 managed by PSSP:

- SP User Management (SPUM)
- Network Information System (NIS)
- Manage each user individually on each machine on the network

The first two are more commonly used; SPUM is discussed in this chapter.

7.3.1 SP User Management (SPUM)

The following information is covered in this chapter:

- How to set up SPUM
- How to add, change, delete, and list SP users
- How to change SP user passwords
- SP user login and access control

7.3.2 Set up SPUM

Attributes need to be enabled, depending on which features you want to use in the SP User Management; you can choose one of the following:

- Automounter
- User administration
- File collection

You can activate SPUM by entering the SMIT fastpath **`smit site_env_dialog`**, depending on which attribute you want to use; set the fields to *true*:

- Automounter Configuration
- User Administration Interface
- File Collection Management

For more detailed information about the SP User Management, see the sections “Managing file collections”, “Managing user accounts“, or “Managing the automounter” in *PSSP Administration Guide,* SA22-7348, or look in *RS/6000 SP*

Planning, Volume 2: Control Workstation and Software Environment, GA22-7281.

7.3.3 Add, change, delete, and list SP users

Using the SP User Management commands, you can add and delete users, change account information, and set defaults for your users' home directories. Specify the user management options you wish to use in your site environment during the installation process, or change them later, either through SMIT panels or by using the **spsitenv** command, or through SMIT by entering **smit spmkuser**. The “Managing user accounts” section in the *PSSP Administration Guide,* SA22-7348 contains detailed instructions for entering site environment information.

The following are the steps for adding an SP user by entering **smit spmkuser**:

- Check the /usr/lpp/ssp/bin/spmkuser.default file for defaults for primary group, secondary groups, and initial programs.
- The user's home directory default location is retrieved from the SDR SP class, homedir_server, and homedir_path attributes.
- **spmkuser** only supports the following user attributes: id, pgrp, home (as in hostname: home_directory_path format), groups, gecos, shell, and login.
- A random password is generated and stored in the /usr/lpp/ssp/config/admin/newpass.log file.

The following example shows how to list SP users using the **spluser** command:

```
spluser spuser1
```

The output will be as shown in Example 7-1.

Example 7-1 spluser spuser1

```
spuser1 id=201 pgrp=1 groups=staff home=/u/spuser1 on
sp3en0:/home/sp3en0/spuser1 shell=/bin/ksh gecos= login=true
```

7.3.4 Change SP user passwords

The SP user passwords may be changed in the following manner:

- The user must log on to the system where the master password file is. Normally, it is on the CWS.
- Use the **passwd** command to change the password.
- The /etc/passwd and /etc/security /password files must be updated on all nodes.

7.3.5 Login control

It is advisable to limit access to the CWS. But users need CWS access to change their passwords in the pure SPUM. A script may be used to enable certain users to access CWS. This script is `/usr/lpp/ssp/config/admin/cw_restrict_login`.

In order to use this script, `/usr/lpp/ssp/config/admin/cw_allowed` must be edited to include the list of users who are permitted CWS login access. This file only allows one user per line starting at the left-most column, and no comments can be included on that file. Root user is not required to be listed in this file.

To make the script work, it must be included in /etc/profile on the CWS. If a restrictive login is to be removed, just comment out or delete the lines that were added in the /etc/profile file.

7.3.6 Access control

Due to the fact that interactive users have a potential negative impact on parallel jobs running on nodes, the **`spacs_cntrl`** command must be executed on each node where access control for a user or group of users must be set.

To restrict a user (for example, **`spuser1`**) on a particular node, enter `spac_cntrl block spuser1` on that node.

To restrict a group of users on a particular node, create a file with a row of user names (for example, name_list) and enter `spacs_cntrl -f name_list` on that node.

To check what **`spacs_cntrl`** is doing, enter `spacs_cntrl -v -l block spuser1`.

7.4 Configuring NIS

Although a Cluster 1600 is a machine containing multiple pSeries nodes or older SPs, you do not want to maintain a cluster as multiple computers but as one system. NIS is one of the tools that can make the daily operations of an SP simple and easy.

NIS is a distributed system that contains system configuration files. By using NIS, these files will look the same throughout a network, or in this case, throughout your SP machine. NFS and NIS are packaged together. Since the SP install image includes NFS, NIS comes along as well.

The most commonly used implementations of NIS are based upon the distribution maps containing the information from the /etc/hosts file and the user-related files /etc/passwd, /etc/group, and /etc/security/passwd.

NIS allows a system administrator to maintain system configuration files in one place. These files only need to be changed once, then propagated to the other nodes.

From a user's point of view, the password and user credentials are the same throughout the network. This means that a user only needs to maintain one password. When a user's home directory is maintained on one machine and made available through NFS, the user's environment is also easier to maintain.

From a cluster point of view, an NIS solution removes the PSSP restriction of changing users' passwords on the CWS. When you use file collections for system configuration file distribution, users have to change their password on the control workstation. When using NIS, you can control user password management across the entire cluster from any given node.

For more information about how to install and configure NIS, refer to Appendix B, "NIS" on page 537.

7.5 File collections

The Cluster 1600 managed by PSSP also has another tool that ensures that system configuration files look the same throughout your SP Management LAN. This tool is called File Collection Management.

File collections are sets of files or directories that simplify the management of duplicate or common files on multiple systems, such as cluster nodes. A file collection can be any set of regular files or directories. PSSP is shipped with a program called /var/sysman/supper, which is a Perl program that uses the Software Update Protocol (SUP) to manage the file collections.

When configuring the SDR, you are asked if you want to use this facility. When answered affirmatively, the control workstation configures a mechanism for you that will periodically update the system configuration files (you specify the interval). The files included in that configuration are:

- All files in the directory /share/power/system/3.2
- Some of the supper files
- The AMD files
- The user administration files (/etc/group, /etc/passwd, and /etc/security/group)
- /etc/security/passwd

In terms of user administration, the File Collection Management system is an alternative to using NIS for users who are not familiar with NIS or do not want to use it.

7.5.1 Terms and features of file collections

There are unique terms and features of file collections, which are covered in the following sections.

Terms used when defining file collections

- Resident - A file collection that is installed in its true location and that can be served to other systems
- Available - A file collection that is not installed in its true location but can be served to other systems

Unique features of file collections

Following are the unique features of file collections:

- Master Files

  A file collection directory does not contain the actual files in the collection. Instead, it contains a set of Master Files to define the collection. Some Master Files contain rules to define which files can belong in the collection and others contain control mechanisms, such as time stamps and update locks.

- The **supper** command interprets the Master Files

  You handle files in a collection with special procedures and the **supper** command, rather than with the standard AIX file commands. The **supper** command interprets the Master Files and uses the information to install or update the actual files in a collection. You can issue these commands in either a batch or interactive mode.

- /var/sysman/file.collections

  File collections require special entries in the /var/sysman/file.collections file, and you need to define them to the supper program. They also require a symbolic link in the /var/sysman/sup/lists file pointing to their Master File.

- Unique user ID

  File collections also require a unique, unused user ID for supman, the file collection daemon, along with a unique, unused port through which it can communicate.

  The default installation configures the user ID, supman_uid, to 102, and the port, supfilesrv_port, to 8431. You can change these values using **SMIT** or the **spsitenv** command.

- supman is the file collection daemon:

  The file collection daemon, supman, requires *read* access permission to any files that you want managed by file collections.

  For example, if you want a security file, such as /etc/security/limits, managed, you must add the supman ID to the security group. This provides supman with read access to files that have security group permission and allows these files to be managed across the cluster environment by file collections. You can add supman to the security group by adding supman to the security entry in the file /etc/groups.

7.5.2 File collection types

File collections can be primary or secondary. Primary file collections are used by the servers and also distributed out to the nodes. Secondary file collections are distributed from the server but not used by the server.

A primary file collection can contain a secondary file collection. For example, the power_system file collection is a primary file collection that consists of the secondary file collection, node.root. This means that power_system can be installed onto a boot/install server, and all of the files that have been defined within that file collection will be installed on that boot/install node, including those in node.root. However, the files in node.root would not be available on that node because they belong to a secondary file collection. They can, however, be served to another node. This avoids having to install the files in their real or final location.

Secondary file collection allows you to keep a group of files available on a particular machine to serve to other systems without having those files installed.

For example, if you want to have one .profile on all nodes and another .profile on the control workstation, consider using the power_system collection delivered with the IBM Parallel System Support Programs for AIX. This is a primary collection that contains node.root as a secondary collection.

- Copy .profile to the /share/power/system/3.2.
- directory on the control workstation.
- If you issue **`supper install power_system`** on the boot/install server, the power_system collection is installed in the /share/power/system/3.2 directory. Because the node.root files are in that directory, they cannot be executed on that machine but are available to be served from there. In this case, .profile is installed as /share/power/system/3.2/.profile.
- If you issue **`supper install node.root`** on a processor node, the files in the node.root collection are installed in the root directory and, therefore, can be

executed. Here, /share/power/system/3.2/.profile is installed from the file collection as /.profile on the node.

A secondary file collection is useful when you need a second tier or level of distributing file collections. This is particularly helpful when using boot/install servers within your Cluster 1600 or when partitioning the system into groups.

7.5.3 Predefined file collections

In PSSP there is a predefined collection of user-administration files: /etc/passwd and /etc/services.

PSSP is shipped with four predefined file collections:

- sup.admin
- user.admin
- power_system
- node.root

Information about each collection on a particular machine can be displayed by using the **supper status** command. You may issue the command anywhere. For example:

```
/var/sysman/supper status
```

sup.admin collection

The sup.admin file collection is a primary collection that is available from the control workstation, is resident (that is, installed), and available on the boot/install servers and resident on each processor node.

This file collection is important because it contains the files that define the other file collections. It also contains the file collection programs used to load and manage the collections. Of particular interest in this collection are:

- /var/sysman/sup, which contains the directories and Master Files that define all the file collections in the system.
- /var/sysman/supper, which is the Perl code for the supper tool.
- /var/sysman/file.collections, which contains entries for each file collection.

user.admin collection

The user.admin file collection is a primary collection that is available from the control workstation, resident, and available on the boot/install servers and resident on each processor node. This file collection contains files used for user management. When the user management and file collections options are turned on, this file collection contains the following files of particular interest:

- /etc/passwd
- /etc/group
- /etc/security/passwd
- /etc/security/group

The collection also includes the password index files that are used for login performance:

- /etc/passwd.nm.idx
- /etc/passwd.id.idx
- /etc/security/passwd.idx

power_system collection

The power_system file collection is used for files that are system dependent. It is a primary collection that contains one secondary collection called the node.root collection. The power_system collection contains no files other than those in the node.root collection.

The power_system collection is available from the control workstation and available from the boot/install servers. When the power_system collection is installed on a boot/install server, the node.root file collection is resident in the /share/power/system/3.2 directory and can be served from there.

node.root collection

This is a secondary file collection under the power_system primary collection. The node.root collection is available from the control workstation, resident, and available on the boot/install servers and resident on the processor nodes. It contains key files that are node-specific.

The node.root file collection is available on the control workstation and the boot/install servers under the power_system collection so that it can be served to all the nodes. You do not install node.root on the control workstation because the files in this collection might conflict with the control workstation's own root files.

7.5.4 File collection structure

The file collection servers are arranged in a hierarchical tree structure to facilitate the distribution of files to a large selection of nodes.

The control workstation is normally the Master Server for all of the default file collections. That is, a master copy of all files in the file collections originates from the control workstation. The /var/sysman/sup directory contains the Master Files for the file collections.

Figure 7-1 on page 259 shows the structure of the /var/sysman/sup directory, which consists of the Master Files for a file collection.

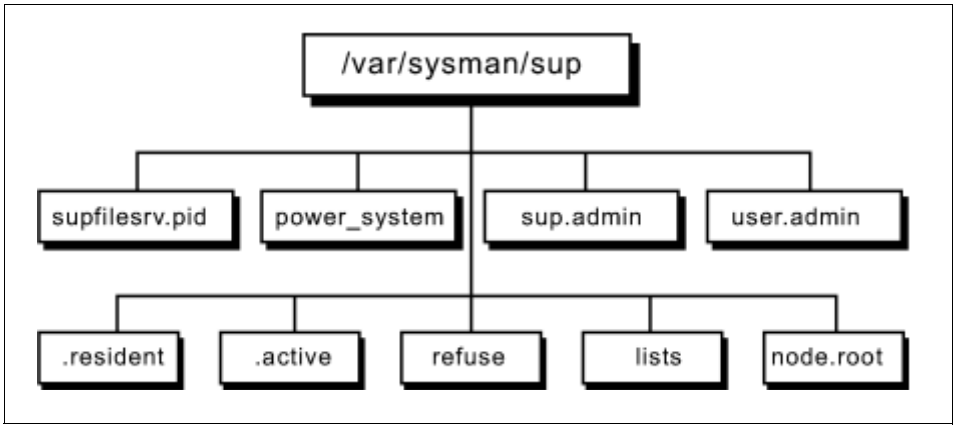


Figure 7-1 /var/sysman/sup files and directories

An explanation of these files and directories follows.

The files are:

.active Identifies the active volume group. It is not found on the control workstation.

.resident Lists each file collection in the Cluster 1600 managed by PSSP system. It is not found on the control workstation.

refuse Files are listed in this file for exclusion from updates.

supfilesrv.pid Consists of the process ID of the supfilesrv process.

The directories are:

lists Contains links to the list files in each file collection.

node.root Contains the Master Files in the node.root collection.

power_system Contains the Master Files in the power_system collection.

sup.admin: Contains the Master Files for the sup.admin collection.

user.admin: Contains the Master Files in the user.admin collection.

An example of an individual file collection with its directory and Master Files is illustrated in Figure 7-2 on page 260. It shows the structure of the /var/sysman/sup/sup.admin file collection.

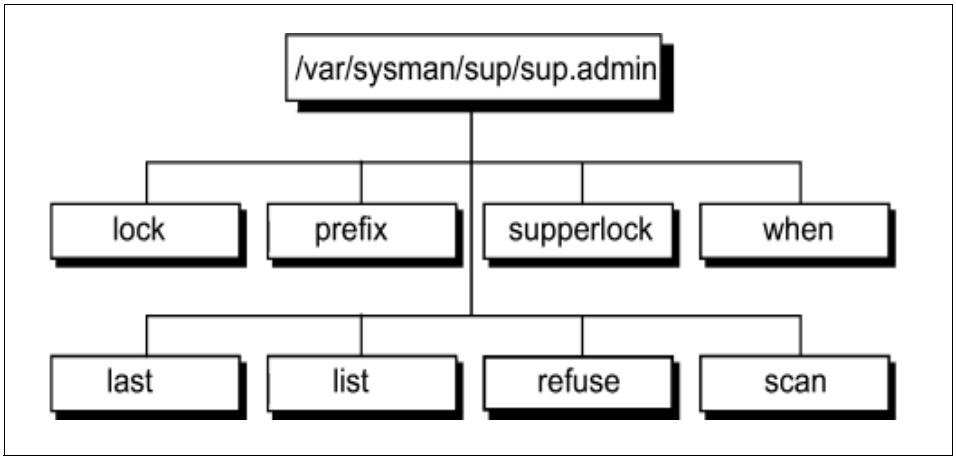


Figure 7-2 sup.admin master files

An explanation of these files follows.

last	Consists of a list of files and directories that have been updated.
list	Consists of a list of files that are part of the file collection.
lock	An empty lock file that prevents more than one update at the same time.
prefix	Consists of the name of a base directory for file references and the starting point for the supper scan process.
refuse	Consists of a list of files to be excluded from update.
scan	Consists of a list of files for the collection with their permission and time stamp.
supperlock	Created by supper to lock a collection during updates.
when	Contains the time for the last file collection update.
activate volume	Sets the active volume group. The active volume group must be set before installing a collection that requires a file system.
debug	Offers a choice of on or off. Turn debug messages on or off.
diskinfo	Shows available disk space and active volumes.
files collection	Shows all files associated with a resident collection.
install collection	Installs a collection.

log	Shows a summary of the last/current supper session.
offline collection	Disables updates of a collection.
online collection	Enables updates of a collection (this is the default).
quit	Exits the program.
remove collection	Removes a collection.
reset collection	Sets the last update time of a collection to the epoch.
rlog	Shows raw output of the last or current supper session.
scan collection	Runs a scan for a collection.
serve	Lists all collections this machine is able to serve.
status	Shows the current status of all available collections. The status information includes the names of all collections, whether they are resident on the local machine, and the name and size of the file system associated with each collection.
update collection	Updates a collection.
verbose	Offers a choice of on or off. Turn SUP output messages on or off.
when	Prints the last update time of all resident collections.
where	Shows the current servers for collections.
! command	Shell escape.

7.5.5 File collection update process

The file collection update process may be done in two ways:

- Set up file collection commands in the crontab file to run in a sequence.

  The actual update occurs on the Master Files on the control workstation.

  Issue the update command from the boot/install server to request file collection updates from the control workstation.

  Issue the update command from nodes to the boot/install server to obtain the required change to its files in the file collections.

- Issue the **`/var/sysman/super update user.admin`** command on each node. This can also be performed remotely with the **`rsh`** and **`rexec`** commands.

7.5.6 Supman user ID and supfilesrv daemon

The supman user ID should be a member of the security group, that is, add supman in security in the /etc/group file. This will allow it to have read access to any files to be managed by file collections.

The user ID for supman must be unique and unused. By default, it is 102.

The supfilesrv daemon resides on the master server only.

7.5.7 Commands to include or exclude files from a file collection

The following are commands to include files in or exclude files from a file collection:

upgrade Files to be upgraded unless specified by the **`omit`** or **`omitany`** commands.

always Files to be upgraded. This ignores **`omit`** or **`omitany`** commands.

omit Files to be excluded from the list of files to be upgraded.

omitany Wild card patterns may be used to indicate the range of exclusions.

execute The command specified is executed on the client process whenever any of the files listed in parentheses are upgraded.

symlink Files listed are to be treated as symbolic links.

7.5.8 Work and manage file collections

Working and managing file collections involves the following activities:

- Reporting file collection information
- Verifying file collection using the **`scan`** command
- Updating files in a file collection
- Adding and deleting files in a file collection
- Refusing files in a file collection

Brief explanations of these activities follow.

Reporting file collection information

The **`supper`** command is used to report information about file collections. It has a set of subcommands to perform file and directory management that includes verification of information and the checking of results when a procedure is being performed.

Table 7-2 provides a list of the supper subcommands or operands that can be used to report on the status and activities of the file collections.

Table 7-2 Brief description of supper subcommands

Supper Subcommands	Runs on	Reports on
where	Node Boot/Install Server	Current boot/install servers for collections
when	Node Boot/Install Server	Last update time of all resident collections
diskinfo	Boot/Install Server CWS	Available disk space and active volume on your machine
log	Node Boot/Install Server	Summary of the current or most recent supper session
rlog	Node Boot/Install Server	Raw output of the current or most recent supper session
status	Node Boot/Install Server CWS	Name, resident status, and access point of all available file collections, plus the name and estimated size of their associated file systems
files	Node Boot/Install Server	All the resident files resulting from a supper **update** or **install** command
serve	Boot/Install Server CWS	All the collections that can be served from your machine
scan	Node Boot/Install Server CWS	For verifying file collection. It creates a scan file in the /var/sysman/sup directory. The file consists of a list of files and directories in the file collection with the date installed and when it was last updated.
update	CWS	If a scan file is present, the **update** command reads it as an inventory of the files in the collection and does not do the directory search. If there is no scan file in the collection, the **update** command searches the directory, apply the criteria in the master files, and add the new file.
install	CWS	To install a collection.

Verifying file collections using scan

By running the **supper scan** command, a scan file is created in the /var/sysman/sup directory. The scan file:

- Consists of a list of all files and directories in the file collection.
- Shows permissions.
- Shows date installed and last updated.

Updating files in a file collection

Make sure changes are made to the files in the master file collection. If there is no /var/sysman/sup/scan file on the server, run the **supper scan** command.

Run the **supper update** command, first on any secondary server, then on the clients. The **supper update** command may be included in the crontab file to run regularly.

Supper messages are written to the following files: The /var/adm/SPlogs/filec/sup<mm>.<dd>.<yy>.<hh>.<mm> summary file and the /var/adm/SPlogs/filec/sup<mm>.<dd>.<yy>.<hh>.<mm>r detailed file.

Adding and deleting files in a file collection

Prior to performing addition or deletion of files in a file collection, consider the following:

- Make sure you are working with the Master Files.
- Add or delete files using standard AIX commands.
- Consider whether the files are in a secondary or primary collection.
- Check what the prefix, list, and refuse files contain.
- Check the prefix from the start point of the tree for file collection.
- If the file is not found in the tree structure, copy the file to it.
- If the entry is needed in the list file, add the entry to it.
- If there is no scan file on the master, run the **supper scan** command.
- Run the **supper update** command on the nodes.

Refuse files in a file collection

The refuse file allows you to customize the file collection at different locations. It is possible for you to create a file collection with one group of files and have different subsets of that group installed on the boot/install servers and the nodes.

The refuse file is created in the /var/sysman/sup directory on the system that will not be getting the files listed in the refuse file.

On a client system, the /var/sysman/sup/refuse file is a user-defined text file containing a list of files to exclude from all the file collections. This allows you to customize the file collections on each system. You list the files to exclude by their fully qualified names, one per line. You can include directories, but you must also list each file in that directory that you want excluded.

A system-created file contains a list of all the files excluded during the update process. If there are no files for this collection listed in the refuse file in the /var/sysman/sup directory, the refuse file in this directory will have no entries.

7.5.9 Modifying the file collection hierarchy

The default hierarchy of updates for file collections is in the following sequence:

1. Control Workstation (CWS)
2. Boot/install servers
3. Nodes

However, the default hierarchy can be changed. The following is an example of this:

- Original scenario

  CWS is the master server for the following two frames for the power_system file collection, and node.root is the secondary file collection associated with it.

  Frame 1: For the nodes and boot/install server A.

  Frame 2: For the nodes and boot/install server B.

- Change the hierarchy so that the boot/install server B on Frame 2 will become the master server for the power_system file collection:

  Take the boot/install server B off-line on Frame 2 by using the **`supper offline`** command. This will eliminate the logical path from the CWS to the boot/install server B for the power_system file collection.

- After the hierarchy change

  If a change is now made to node.root on the CWS, the boot/install server A and the nodes on Frame 1 will get updated, but boot/install server B and the nodes on Frame 2 will not get updated.

  If the same change is required on boot/install server B, then the update must be performed directly to the files on boot/install server B. Then the nodes on Frame 2 will get updated as well.

7.5.10 Steps in building a file collection

You may create your own file collection. You can build one for any group of files that you want to have identically replicated on nodes and servers in your system.

There are seven steps in building a file collection, and you must be root to perform all of them.

1. Identify the files you wish to collect. For example, it has been decided that program files (which are graphic tools) in the /usr/local directory are to be included on all nodes.
2. Create the file collection directory. In this case, create the /var/sysman/sup/tools directory.
3. Create master files that are list, prefix, lock, and supperlock. Pay attention to the list file that consists of rules for including and excluding files in that directory. Lock and supperlock files must be empty.
4. Add a link to the list file in the /var/sysman/sup directory. For example, in -s /var/sysman/sup/tools/list /var/sysman/sup/lists/tools.
5. Update the file.collections file. Add the name of the new file collection as either a primary or secondary collection.
6. Update the .resident file by editing the .resident file on your control workstation or your master server, and add your new collection, if you want to make it resident, or use the **`supper install`** command.
7. Build the scan file by running **`supper scan`**. The scan file only works with resident collections. It consists of an inventory list of files in the collection that can be used for verification and eliminates the need for supper to do a directory search on each update.

7.5.11 Installing a file collection

During initialization and customization processes, the required SP file collections are first built on the CWS and then installed on the boot/install servers and processor nodes. However, if you create your own, you have to install them on each server or node.

There are four steps involved, and you must be root to perform installation of a new file collection.

1. Update the sup.admin file collection that contains all the files that identify and control the file collections, such as the .collections and .resident files. Whenever changes are made to these two files, you need to update the sup.admin collection to distribute these updates. For example:

```
/var/sysman/supper update sup.admin.
```

2. Run the **supper install** command on each boot/install server or node that needs this collection. For example:

```
/var/sysman/supper install sup.admin
```

3. Add the **supper update** for new file collection to **crontabs**.
4. Run the **supper scan** command on the master.

7.5.12 Removing a file collection

The predefined file collections that come with the PSSP are required. Removing them will result in problems when running PSSP. Removing a file collection does not delete it. It removes it from the place where it was installed. To completely delete a file collection, you must remove it from every place it was installed.

There are two steps in removing a file collection:

1. Run **supper scan** to build a scan file. This will help to verify that none of the files in the file collection will be needed.
2. After verification, run the **supper remove <file collection>** command to remove the file collection.

7.5.13 Diagnosing file collection problems

The cross reference summary of common file collection problems and solutions is in 15.5, “Diagnosing file collection problems” on page 498.

7.6 SP user file and directory management

This section provides a description of the tool used for SP user files and directory management.

7.6.1 AIX Automounter

AIX Automounter is a tool that can make the Cluster 1600 system managed by PSSP appear as only one machine to both the end users and the applications by means of a global repository of storage. It manages mounting activities using standard NFS facilities. It mounts remote systems when they are used and automatically dismounts them when they are no longer needed.

The number of mounts can be reduced on a given system and has less probability of problems due to NFS file server outages.

On the Cluster 1600, the Automounter may be used to manage the mounting of home directories. It can be customized to manage other directories as well. It

makes system administration easier because it only requires modification of map files on the CWS to enable changes on a system-wide basis.

Following are the steps for Automounter initial configuration:

1. Use the **smit enter_data** command on the CWS to perform PSSP installation, which displays Site Environment Information.
2. Add users to the system.
3. Ensure the amd_config variable is set to *true* so that the automountd (which is the automounter daemon) will start.
4. Ensure that the usermgmt_config variable is set so that the maps for the users' home directories will be maintained.

The AIX Automounter reads automount map files to determine which directories to handle under a certain mount point. These map files are kept in the /etc/auto/map directory. The list of map files for the Automounter is stored in the /etc/auto.master file. The master files can also be accessed by means of NIS.

7.7 Related documentation

The following books are recommended readings to provide a broader view of user and data management.

SP manuals

PSSP Administration Guide, SA22-7348 covers "File collection types".

RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment, GA22-7281 covers "Understanding user account management choices".

PSSP Installation and Migration Guide, GA22-7347 covers "Initializing to use AFS authentication".

SP redbooks

IBM (e)server Cluster 1600 Managed by PSSP 3.5: What's New, SG24-6617I

7.8 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, "Answers to sample questions" on page 521.

1. Why can passwords not be changed directly on the nodes if SP User Management is being used?
   a. Because there is not a **passwd** command on the nodes.
   b. Because the /etc/passwd and /etc/security/passwd files are not present on the nodes.
   c. Because the /etc/passwd and /etc/security/passwd files get replaced with the files from the passwd file server.
   d. Who says it cannot be done?
2. What is the difference between an AIX user and an SP user?
   a. An AIX user is able to access a local resource in a node, while an SP user can only access SP-related resources.
   b. There is no difference, just terminology.
   c. SP users are global AIX users managed by the SP User Management facility on the Cluster 1600 managed by PSSP.
   d. SP users can Telnet to the control workstation, while AIX users cannot.
3. What is the command you would use to set access control on a node?
   a. spac_block
   b. cntrl_access
   c. restric_login
   d. spac_cntrl
4. What is the file collection that contains all the User Management related files?
   a. user_admin
   b. user.admin
   c. user.mgmt
   d. user_mgmt
5. Which of the following predefined file collections is *not* shipped with PSSP?
   a. node.root
   b. power_system
   c. user.admin
   d. supper.admin
6. Which command is used to report information about file collections?
   a. online_collection
   b. supper

c. offline_collection

d. update_collection

7. Which tool allows a system administrator to maintain system configuration files in one place?

   a. DFS

   b. NFS

   c. NIS

   d. AFS

8. What is the default hierarchy sequence of updates for file collections?

   a. Nodes/BIS/CWS

   b. CWS/Nodes/BIS

   c. BIS/CWS/Nodes

   d. CWS/BIS/Nodes

9. What must be considered prior to performing addition or deletion of files in a file collection?

   a. Run the **supper install** command on the nodes.

   b. If there is no scan file on the master, run the **supper status** command.

   c. Make sure you are working with the master files.

   d. If the entry is needed in the list file, add it to the /.k file.

10. Which tool can make the Cluster 1600 managed by PSSP system appear as only one machine to both the end users and the applications?

    a. NFS

    b. AIX Automounter

    c. NIS

    d. DFS

7.9 Exercises

Here are some exercises you may wish to perform:

1. On a test system that does not affect any users, build a file collection. You must be root to perform this exercise.
2. Install the file collection that you created in the previous exercise.
3. Display all the file collections on your system.

4. Remove the file collection you added in exercise 1. Can you remove the predefined file collection on your system? Explain.
5. On a test system that does not affect any users, add, change, list, and delete SP users.

Part 2

Installation and configuration

This part contains chapters describing the actual implementation of the steps for installing and configuring the control workstation, nodes, and switches. It also includes a chapter for system verification as a post-installation activity.

8

Configuring the control workstation

This chapter addresses various topics related to the initial configuration of the Control Workstation (CWS): Preparation of the environment, copy of the AIX and PSSP LPPs from the distribution media to the CWS, initialization of Kerberos services and the SDR. These topics are not listed in the chronological order of the CWS configuration process. Rather, they are gathered by categories: PSSP commands, configuration files, environment requirements, and LPP considerations.

Note: For a complete list of the supported CWS for the Cluster 1600 refer to Table 2-10 on page 38.

The installation process is divided into three areas:

- Environment preparation and SDR data entry.
- Boot/install server configuration. After this, nodes are ready to be installed over the network.
- Node installation.

In this chapter, we cover the CWS setup. We also provide an overview of the installation processes.

Note:

- PSSP 3.5 is supported on AIX 5L 5.1, Maintenance Level 03 or later.
- PSSP V3.5 does *not* support the AIX V4.3.3 operating system.
- 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.
- 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.

Before you install the SP system, you must perform several planning activities. These include completing worksheets or checklists that reflect your decisions regarding, but not limited to, the following:

- System partitioning (SP Switch only)
- Cluster 1600 and switch configuration
- System management options
- Boot/install servers and node relationships
- Location and type of authentication servers

It is essential that you plan your system carefully before attempting to install it. Refer to *RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment,* GA22-7280 and *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281 for planning details.

Always review the *READ THIS FIRST* document that accompanies the PSSP installation media for the latest information. Assuming that the PSSP CD-ROM is in /dev/cd0, issue the following command to view that document:

```
installp -iq -d /dev/cd0 all
```

Note: For the most up-to-date copy of this document, check the following site under *READ THIS FIRST*:

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

The summary of enhancements of the *READ THIS FIRST* is listed below:

- Additional resource information
- PSSP compatibility, limitations, restrictions, and prerequisites

- PSSP packaging
- Viewing the softcopy README documents from CD-ROM
- AIX 5L 5.1 minimal image content
- AIX 5L 5.2 minimal image content
- List of program materials

8.1 Key concepts

Before taking the RS/6000 SP certification exam, you should understand the following CWS configuration concepts:

- PSSP product packaging
- Mandatory and optional LPPs and filesets required
- Connectivity between the CWS and Cluster 1600 (for supported hardware, refer to Chapter 2, “Validate hardware and software configuration” on page 7
- Storage requirements and directory organization for PSSP software
- AIX system configuration files related to the SP system
- CWS configuration commands
- Source directory for the PSSP and AIX filesets
- Setup of Kerberos authentication services
- Finally configuring the CWS with the **`install_cw`** command
- Verifying the success of the configuration command

8.2 Summary of CWS configuration

This section presents a summary of the initial configuration of the CWS. It corresponds to Chapter 2, steps 1 to 31 of *PSSP Installation and Migration Guide,* GA22-7347.

The initial configuration of the CWS is the part of the PSSP installation where you prepare the environment before you start configuring the PSSP software. It consists of ten steps:

1. Updating the AIX system environment: we modify the PATH of the root user, change the maximum number of processes allowed by AIX, customize a few system files, such as /etc/services, and check that some system daemons are running.

2. Make sure that the AIX system is at the appropriate level (AIX, perfagent), and that it matches the prerequisites of the version of PSSP you are about to install.
3. You must check the physical connectivity between the CWS and the Cluster 1600 nodes. You *cannot* start configuring the Cluster 1600 nodes on the CWS until the physical installation is completed. You must then configure your TCP/IP network: Addresses, routes, name resolution, tuning of network parameters, and so on. The TCP/IP definition of all the Cluster 1600 servers must be completed on the CWS before initializing Kerberos services and before configuring the SDR. This step is critical to the success of the SP system installation. Refer to Chapter 3, “Cluster 1600 networking” on page 101 for more detail on the TCP/IP configuration step.
4. Allocate disk space on the CWS for storing the PSSP software, and restore it from the distribution media.
5. Installing PSSP prerequisites.
6. Moving required filesets from the PSSP directory to the respective lppsources.
7. Installation of filesets required for HMC-controlled server support.
8. Installation of the PSSP on the CWS using the **`installp`** command.
9. You must configure authentication services on the CWS either by using the Kerberos implementation distributed with PSSP or the AFS or secured remote command method.
10. Finally, initialize the SDR database that will be used to store all your SP system configuration information.

The tasks described in steps 8, 9, and 10 must be performed in this order after all other steps.

8.3 Key commands and files

The commands described in this section are to be used only on the CWS and not on the nodes.

8.3.1 setup_authent

This command has no argument. It configures the Kerberos authentication services for the SP system. It first searches the AIX system for Kerberos services already installed, checks for the existence of Kerberos configurations files, and then enters an interactive dialog where you are asked to choose and customize the authentication method to use for the management of the SP system. You can

choose to use the SP-provided Kerberos services, another already existing Kerberos V4 environment, or an AFS-based Kerberos service. If you choose the SP-provided Kerberos services, **setup_authent** will initialize the primary authentication server on the CWS.

Refer to chapter 2, steps 22 through 26 of the *PSSP Installation and Migration Guide,* GA22-7347*.*

8.3.2 chauthts

The **chauthts** command enables the authentication methods used by trusted services on the local host. Trusted services that support multiple methods will attempt to authenticate and authorize client requests using the methods in the order shown. Use this command to set the authentication methods on the control workstation at initial installation and on independent workstations.

8.3.3 k4init

The **k4init** command is used to obtain a Kerberos Version 4 authentication ticket and to authenticate a user's identify to the SP authentication service. All previous tickets are discarded. When you use the **k4init** command without options, it prompts for your principal name and password, and tries to authenticate your identity within the local realm. If the specified principal name and password are correct, **k4init** retrieves your initial ticket and puts it in the ticket file specified by your KRBTKFILE environment variable. If the KRBTKFILE variable is undefined, your ticket is stored in the /tmp/tktuid file, where UID specifies your user identification number.

Important: We strongly suggest using the **k4init** form. DCE includes the **kinit** command, so using **kinit** instead of **k4init** may produce unexpected results if DCE is installed.

Refer to chapter 2 and chapter 2 of *PSSP Administration Guide,* SA22-7348 for a detailed conceptual understanding of Kerberos.

8.3.4 install_cw

This command has no argument. It is used after the PSSP software has been installed on the CWS and after the Kerberos authentication services have been initialized. The command, **install_cw**, performs the following:

- Configures the control workstation.
- Installs PSSP SMIT panels.

- Starts SP daemons.
- Configures the SDR.
- Updates /etc/inittab and /etc/services.
- Sets node_number for the control workstation to 0 in the Object Data Management (ODM) databases.
- Creates the hardmon and hmacls files.
- Configures the system as one system partition. There is one system partition corresponding to the name of the control workstation in the SP object. It is called the default or persistent system partition because it always exists.
- Sets security attributes for the default system partition.

Before the installation of PSSP software on the CWS, you have to modify several AIX system files. These changes can be done in any order, as long as they are done before using the commands **setup_authent** and **install_cw**.

The SDR_test and the spmon_itest used in this chapter are discussed later in 10.3, “Key commands” on page 348.

8.3.5 .profile, /etc/profile, or /etc/environment

The root user (during installation) and any user chosen for SP system administration (during SP operation) need to have access to the PSSP commands. For each of these users, depending on your site policy, one of the files, $HOME/.profile, /etc/profile or /etc/environment has to be modified so that the PATH environment variable contains the directories where the PSSP and Kerberos commands are located.

For $HOME/.profile or /etc/profile, add the lines:

```
PATH=$PATH:/usr/lpp/ssp/bin:/usr/lib/instl:/usr/sbin:\
/usr/lpp/ssp/kerberos/bin
export PATH
```

For /etc/environment, add the line:

```
PATH=/usr/bin:/etc:/usr/sbin:/usr/ucb:/usr/bin/X11:/sbin:\
/usr/lpp/ssp/bin:/usr/lib/instl:/usr/lpp/ssp/kerberos/bin
```

8.3.6 /etc/inittab

This file is used to define several commands that are to be executed by the **init** command during an RS/6000 boot process.

On the CWS, you must make sure that this file starts the AIX System Resource Controller (SRC). The srcmstr entry of the CWS /etc/inittab must be uncommented and looks as follows:

```
srcmstr:2:respawn:/usr/sbin/srcmstr # System Resource Controller
```

/etc/inittab is also used to define which PSSP daemons are started at boot time. It is updated automatically during PSSP installation with the appropriate entries in the /etc/inetd.conf file.

8.3.7 /etc/inetd.conf

On the CWS, the inetd daemon configuration must contain the uncommented entries bootps and tftp. If they are commented prior to the PSSP installation, you must uncomment them manually. The PSSP installation scripts do not check or modify these entries.

8.3.8 /etc/rc.net

For improving networking performance, you can modify this file on the CWS to set network tunables to the values that fit your SP system by changing the parameters listed in Table 8-1.

Table 8-1 Control workstation network settings

Tunable	Recommended initial value	Description
thewall	This value is automatically sized by the system at boot.	The upper bound on the amount of real memory that can be used by the communications subsystem. The units are in 1 KB increments.
sb_max	163840	Upper limit on the size of the TCP and UDP buffers in mbufs of allocated space to a connection.
ipforwarding	1	Specifies whether the kernel should forward packets. A value of **1** forwards packets when they are not for the local system; a value of **0** prevents forwarding.
tcp_sendspace	65536	The default size of the TCP send window.
tcp_recvspace	65536	The default size of the TCP receive window.

Tunable	Recommended initial value	Description
upd_sendspace	32768	The default size of the UDP send buffer. The effective maximum is 65536 (64K).
upd_recvspace	65536	The default size of the UDP receive buffer.
tcp_mssdflt	1448	Maximum package size for remote network.
tcp_pmtu_discover	0	TCP MTU path discovery (AIX 4.3.1 or later).
udp_pmtu_dicover	0	UDP MTU path discovery (AIX 4.3.1 or later).

Using the no command

To display network tunable values, enter:

```
no -a
```

To change the value of tcp_mssdflt, enter:

```
no -o tcp_mssdflt=1448
```

When you change the network tunables, they take effect immediately. However, they are not preserved across a boot. To make the changes to the tunables effective across boots, add the **no -o** commands you used to change the network tunables to the last section of the /etc/rc.net file. Using the same syntax, place the commands under the line:

/usr/sbin/no -o extendednetstats=0 >>/dev/null 2>&1

as follows:

/usr/sbin/no -o tcp_mssdflt=1448 >>/dev/null 2>&1

If AIX 5.2 is installed, you must copy usr/lpp/ssp/install/config/nextboot.default to /etc/tunables/nextboot, and run **/usr/sbin/tunrestore**, which applies the nextboot file. See *AIX 5L V5.2 Performance Tools Guide and Reference,* SC23-4859 for details.

The rc.net file is also the recommended location for setting any static routing information. In particular, the CWS needs to have IP connectivity to each of the SP nodes or HMC network en0 adapter during the installation process. In the case where the CWS and all nodes' en0 adapters are not on the same Ethernet

segment (for example, when there are several frames), the rc.net file of the CWS can be modified to include a routing statement.

8.3.9 /etc/services

This file maps the names of network services to the well-known ports that they use to receive requests from their clients. This file may already contain entries relating to authentication services, since the file shipped with AIX contains entries used by DCE.

In addition to the DCE entries, many other reserved port names are in /etc/services, including an entry for the Kerberos V5 port, 88. The service name for this port is given as “kerberos”, which is also the name used by the standard MIT Kerberos V4 service. The port number usually assigned to the Kerberos V4 service is 750. In order to be consistent with and interoperate with AIX 3.2.5 systems running PSSP 1.2 with authentication services based on Kerberos V4, it was necessary to use the name “kerberos4”. You do not have to create an entry in the file for “kerberos4”, because the default port of 750/udp will be used if no “kerberos4” entry is found.

If you are not using AFS Version 3.4 authentication servers, you should only have to modify /etc/services if you are using some other service that uses one or more of the traditionally used (but not formally reserved) Kerberos V4 ports. They are:

- Service name: kerberos4 Port: 750/udp
- Service name: kerberos_admin Port: 751/tcp
- Service name: krb_prop Port: 754/tcp

You will also have to modify this file if you are using AFS Version 3.4 authentication servers. The kaserver in AFS Version 3.4 for AIX 4.1 accepts Kerberos V4 protocol requests using the well-defined udp port assigned to the “kerberos” service assigned to port 88 in the file distributed with the base operating system. MIT Kerberos V4, on which PSSP authentication services are based, uses a default port number of 750. PSSP authentication commands use the service name “kerberos4” to avoid this conflict with the Kerberos V5 service name, which is also used by DCE. For PSSP authentication commands to communicate with an AFS 3.4 kaserver, you must do either of the following:

- Stop the kaserver, redefine the udp port number for the “kerberos” service to 750 on the server system, then restart the kaserver.
- Add a statement to /etc/services that defines the udp port for the “kerberos4” service as 88 on the SP control workstation and on any other independent workstation that will be a client system for PSSP authenticated services.

There is a conflict in the use of port 88 by Kerberos V4 (as used by AFS) and by Kerberos V5 (assigned to DCE by AIX 4.1). The /etc/services file can be used to resolve this problem if you decide to use the AFS authentication services by adding the line:

```
kerberos4       88/udp # Kerberos V4 - added for PSSP
```

An alternative solution is to reconfigure the AFS authentication server to use another port (750).

8.4 Environment requirements

Before starting the installation of the PSSP software onto the CWS, prepare the hardware and software environment and pay attention to some rules and constraints.

8.4.1 Connectivity

During normal operation, the TCP/IP connectivity needed for user applications between the CWS and the Cluster 1600 can be provided through any type of network (Token Ring, FDDI, ATM) supported by the RS/6000 hardware. However, for the installation and the management of the Cluster 1600 from the CWS, there *must* exist an Ethernet network connecting all the Cluster 1600 servers to the CWS. This network may consist of several segments. In this case, the routing between the segments is provided either by one (or several) Cluster 1600 nodes with multiple Ethernet adapters, Ethernet hubs, or Ethernet routers.

Furthermore, Figure 2-1 on page 9 shows how the monitoring and control of the Cluster 1600 hardware from the CWS requires a serial connection between the CWS (*except for the servers connected to the CWS through the HMC)*, and *each* frame in the SP system and all other Cluster 1600 nodes. If there are many server nodes, there may not be enough built-in serial adapters on the CWS, and additional serial adapter cards may need to be installed on the CWS.

If SP-attached servers (RS/6000 S70, S7A or S80) are included in the SP system, *two* serial cables are needed to link the CWS to *each* of the servers. An Ethernet connection is also mandatory between the CWS and the SP-attached server configured on the en0 adapter.

Note: Refer to 2.10, “Peripheral devices” on page 67 for details on the exact PCI slot for the en0 adapter.

Refer to Table 8-2 for information on the required tty ports and the protocol used by each of them.

Table 8-2 tty and hardware protocol for Cluster 1600

SP model numbers	Protocol required	tty port values
IBM @server pSeries 690, 670, 655, 650, 630	HMC	Does *not* require a tty port value, but the Hardware Management Console (HMC) must either be connected to the SP Ethernet administrative LAN or to the HMC trusted network.
RS/6000 H80, M80, and IBM @server pSeries 660 (6H0, 6H1, 6M1)	CSP	1
RS/6000 S70, S7A, and S80 or IBM @server pSeries 680	SAMI	2
SP frames	SP	1

Important: All POWER4 pSeries models must use the HMC to integrate them into cluster 1600 environment.

Also, note that the CWS cannot be connected to an SP Switch (no css0 adapter in the CWS). The connectivity between the CWS, the frames, SP nodes, and the SP-attached servers through the serial links, and between the CWS and all the nodes and HMC through the Ethernet network, must be set up before starting the PSSP installation.

8.4.2 Disk space and file system organization

The volume group, logical volumes, and file systems must be created, and the AIX lppsource, mksysb images, and PSSP filesets *must* be placed in the correct location. Refer to *PSSP Installation and Migration Guide,* GA22-7347 for more information.

/spdata size and disk allocation

The /spdata directory contains, among other items, mksysb and installp file sets. We suggest that you create a separate volume group for the /spdata file system. These file sets require a minimum of 2 GB of disk space. You will require additional disk space if you need to support multiple AIX and PSSP release levels, and multiple mksysb images. If you have not done so already, use

RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment, GA22-7281, to help you estimate how much space you need to define. Keep in mind that this rule provides only a very rough estimate. As a point of comparison, the minimum system image (spimg) provided with PSSP is 130 MB versus an estimated 700 MB for the system images considered in this rule of thumb.

It is recommended, but not required, to dedicate a volume group of the CWS to the /spdata directory. The decision for creating such a volume group must take into account the backup strategy that you choose. The root volume group can be backed up using the **mksysb** command to create a bootable image, while other volume groups can be saved using the **savevg** command. Since there is no need of any file in the /spdata directory for restoring the CWS from a bootable image, the /spdata directory does not need to be part of the CWS **mksysb**.

Furthermore, the contents of the /spdata directory change when the systems installed on the Cluster 1600 are modified (the creation of new node system images). This is likely to be different from the time the content of the CWS rootvg changes. The schedules for the backup of the CWS volume group and for the /spdata directory will, therefore, be generally disjointed.

/spdata directory structure and naming convention

You must manually create the /spdata directory before beginning the PSSP installation with a minimum substructure consisting of the directories shown in Table 8-3.

Table 8-3 /spdata initial structure for PSSP-3.5

Directory	Description	mkdir command
/spdata/sys1/install/*name*/lppsource/installp/ppc	Location of required AIX file sets in installp format	**mkdir -p /spdata/sys1/install/*name*/lppsource/installp/ppc**
/spdata/sys1/install/*name*/lppsource/RPMS/ppc	Location of required AIX file sets in RPM format	**mkdir -p /spdata/sys1/install/*name*/lppsource/RPMS/ppc**
/spdata/sys1/install/images	Location of all required AIX mksysb images	**mkdir /spdata/sys1/install/images**
/spdata/sys1/install/pssplpp/*code_version*	Location of all SP installp file sets	**mkdir -p /spdata/sys1/install/pssplpp/code_version**
/spdata/sys1/install/pssp	Location of NIM configuration data files	mkdir /spdata/sys1/install/pssp

Where *name* is the new lppsource name for the nodes (such as aix520 if that is what you called the subdirectory with the AIX 5L 5.2 lppsource). Keep in mind that the setup_server program looks for this name later on during the installation process. By default, it is set to the string “default,” so that if you use that as your subdirectory name, you do not have to change the name here. code_version is the version of code of the form pssp-x.y. For example, PSSP-3.5.

The installable images (LPP) of the AIX systems must be stored in directories named /spdata/sys1/install/<name>/lppsource/installp/ppc (or /rpms) depending on the format of the files as described in Table 8-3 on page 286. You can set <name> to the name you prefer. However, it is recommended to use a name identifying the version of the AIX LPPs stored in this directory. The names generally used are aix51, aix52, and so on.

Attention: Except for <name>, the name of all directories listed in Figure 8-3 *must* be left unchanged.

In “/spdata size and disk allocation” on page 285, we mentioned one possibility of allocation of /spdata based on a backup strategy. We now present another possibility based on the contents of the subdirectories of /spdata. Instead of dedicating a volume group to /spdata, you can spread the information contained in the /spdata directory between the rootvg and another volume group (for example, let us call it spstdvg). All information that can be easily recreated is stored in spstdvg, while all information that is manually created during the installation of the SP system is stored in rootvg.

The advantage of this solution is to enable the backup of critical SP information along with the rest of the AIX system backup using the **mksysb** command, while all information that is not critical can be backed up independently with a different backup frequency (maybe only once at installation time). Practically, this implies that you create on the spstdvg volume group one file system for holding each directory as follows:

- /spdata/sys1/install/<name>/lppsource
- /spdata/sys1/install/images
- /spdata/sys1/install/pssplpp

These directories are then mounted over their mount point in rootvg.

Another advantage of this solution is that these directories contain most of the /spdata information. The remaining subdirectories of /spdata represent only around 30 MB. This solution, therefore, enables you to keep the size of the rootvg to a reasonable value for creating mksysb bootable images.

8.5 LPP filesets

The Cluster 1600 system requires at least the installation of AIX, Perfagent, and PSSP. Each of these products consists of many filesets, but only a subset of them are required to be installed. The following sections explain which filesets need to be installed depending on the configuration of the Cluster 1600 system.

Note: You must view READ THIS FIRST to know the fixes required for CWS as well as the nodes, if an mksysb is made manually for the nodes.

8.5.1 PSSP prerequisites

The PSSP software has prerequisites on the level of AIX installed on the CWS as well as on optional LPPs. These requirements are different for each release of PSSP. In AIX 5L 5.1 and later, changes are made to allow installation with installation tools other than **installp** to allow new installation media formats. With AIX 5L 5.1 and later, two new commands (**geninstall** and **gencopy**) are introduced, which call **installp** or **bffcreate**. Additional subdirectories are added into NIM LPP_SOURCE with AIX 5L 5.1. For NIM, instead of just putting everything in the LPP_SOURCE directory, appropriate subdirectories are created by the **gencopy** and **bffcreate** commands and the images are copied to those subdirectories based on the format of the install package. You must copy the AIX file sets under the /spdata/sys1/install/*name*/lppsource directory on your hard disk on the control workstation. The **bffcreate** command places the files into the appropriate subdirectory as follows:

- For AIX 5L 5.1 or later lppsource, all file sets in installp format must be placed in the following directory: /spdata/sys1/install/name/lppsource/installp/ppc.
- For AIX 5L 5.1 or later lppsource, all file sets in RPM format must be placed in the following directory: /spdata/sys1/install/name/lppsource/RPMS/ppc.

If you are copying the file sets over manually, you must create the appropriate directory structure for the lppsource. Refer to Table 8-3 on page 286.

If you are using NIM or the **bffcfreate** command to create your lppsource, you must designate on the command, regardless of the AIX version, that the target directory is /spdata/sys1/install/name/lppsource. For AIX 5L 5.1 systems, NIM and the **bffcreate** command create the appropriate subdirectories.

The AIX filesets and required AIX LPPs must exist in the correct directory. Links to filesets in other directories *are not allowed*. If you change the path name in any way, the installation *fails*. You can download all of the AIX filesets (a very large number) or only the minimal required AIX file sets (approximately 1 GB for AIX 5L 5.2).

The prefix.* syntax in the list refers to everything that starts with the prefix. For example, devices.* refers to all the filesets starting with devices. The minimum set of AIX components to be copied to the /spdata/sys1/install/<name>/lppsource directory can be found in chapter 2 step 16 of *PSSP Installation and Migration Guide,* GA22-7347. For the most recent update, refer to *READ THIS FIRST*.

Additional files you may want to add to your lppsource

- bos.acct.* is required if you plan to use PSSP accounting.
- bos.clvm.* is required if you plan to install IBM Virtual Shared Disk.
- bos.cpr.* is required to install LoadLeveler 3.1 or Parallel Environment 3.2.
- dce.* is required only if DCE is configured by PSSP anywhere on the system. You need the client portion of the DCE filesets because the installation code installs the DCE client code.

Additional files required on the CWS for HMC servers

- csm.client.*
- Java131.xml4j.*
- Java131.rte
- openCIMOM
  Copy the CIMOM fileset RPMS/noarch/openCIMOM-0.61-1.aix4.3.noarch.rpm from the AIX Toolbox for Linux Applications CD. It is labeled in the contents as openCIMOM. Copy the file set to the following location:

  /spdata/sys1/install/*name*/lppsource/RPMS/ppc/

The file set RPMS/noarch/openCIMOM-0.61.1.aix4.3.noarch.rpm is the minimum level of the fileset that is required. If your AIX Toolbox for Linux Applications CD contains a newer level of the fileset, such as RPMS/noarch/openCIMOM-0.7-1.aix5.1.noarch.rpm, you *should* copy the file set devices.chrp_lpar*.

Note: Be sure to obtain the latest levels and put the filesets in the correct lppsource, which is the installp/ppc/ subdirectory for installp packages and RPMS/ppc/ for the RPM files.

Refer to your disk usage planning in chapter 3 in the section determining install space requirements of *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281 to determine if you have allocated enough space to accomplish this task.

Allow at least 1-3 hours for moving all the filesets from media to disk. To copy the AIX LPP images, log in to the CWS as root and run **bffcreate** using SMIT or the command line. The following example shows the product media on cd0 and the selection of all LPPs. Using `all` may load unnecessary filesets into the directory:

```
bffcreate -qvX -t/spdata/sys1/install/name/lppsource -d /dev/cd0 all
```

If you are using AIX 5L 5.1 or later, you need to run the **inutoc** script as follows:

```
cd /spdata/sys1/install/name/lppsource/installp/ppc
inutoc .
```

In addition, the right level of perfagent must be installed on the CWS and copied to each /spdata/sys1/install/<name>/lppsource/installp/ppc directory as described in Table 8-4.

The perfagent.server fileset is part of the Performance Aide for AIX (PAIDE) feature of the Performance Toolbox for AIX (PTX®), a separate product. This product provides the capability to monitor your Cluster 1600 system's performance, collects and displays statistical data for SP hardware and software, and simplifies runtime performance monitoring of a large number of nodes.

The Performance Toolbox for AIX, Agent Component (PAIDE) is required. The correct level of AIX PAIDE (perfagent.tools) needs to be installed on the control workstation and copied to all of the lppsource directories. The perfagent.tools file set is part of AIX 4.3.3, AIX 5L 5.1, and AIX 5.2. The required level of perfagent is dependent upon the level of AIX and PSSP, as shown in Table 8-4.

Table 8-4 Perfagent filesets

AIX level	PSSP level	Required filesets
AIX 4.3.3	PSSP 3.4	perfagent.tools 2.2.33.*
AIX 5L 5.1	PSSP 3.4	perfagent.tools 5.1.0.*
AIX 5L 5.1	PSSP 3.5	perfagent.tools 5.1.0.*
AIX 5.2	PSSP 3.5	perfagent.tools 5.2.0.*

8.6 PSSP filesets installation on the CWS

The installation of the PSSP software on the CWS is an eight-step process.

8.6.1 Copy of the PSSP images

All PSSP software is first restored from the distribution media into the /spdata/sys1/install/pssplpp/PSSP-x.x directory using the **bffcreate** command. You can see the details on page 29 and 30 of *PSSP Installation and Migration Guide,* GA22-7347.You can use the following command to copy the PSSP images from the media:

```
bffcreate -d /dev/cd0 -t /spdata/sys1/install/pssplpp/PSSP-3.5 -X all
```

When **bffcreate** completes, rename ssp.3.5.0.0.I in /spdata/sys1/install/pssplpp/PSSP-3.5.

Then, do the following:

cd /spdata/sys1/install/pssplpp/PSSP-3.5

mv ssp.3.5.0.0.I pssp.installp

inutoc .

8.6.2 Move prerequisite files for PSSP 3.5

Several PSSP prerequisite files that are shipped on the PSSP 3.5 media must be moved to your AIX lppsource. Refer to Example 8-1 for more details.

Example 8-1 Moving prerequisite fileset

```
cd /spdata/sys1/install/pssplpp/PSSP-3.5
mv xlC.rte.* /spdata/sys1/install/name/lppsource/installp/ppc
mv xlC.aix50.* /spdata/sys1/install/name/lppsource/installp/ppc
mv ipfx.* /spdata/sys1/install/name/lppsource/installp/ppc
mv vacpp.ioc.* /spdata/sys1/install/name/lppsource/installp/ppc
mv vacpp.cmp.* /spdata/sys1/install/name/lppsource/installp/ppc
inutoc .
cd /spdata/sys1/install/name/lppsource/installp/ppc
inutoc .
```

8.6.3 Copy the minimum AIX image (mksysb)

The media shipped with the SP hardware contains the *spimg installp* image. With PSSP 3.5, the spimg installp image contains both the AIX 5L 5.1 32-bit kernel and the 64-bit kernel mksysb images. You may install any of these images for use on your nodes or use mksysb images of your own. You only need to install the AIX images that you intend to use.

Go to the images directory and issue the following command:

```
installp -a -d /dev/cd0 -X spimg
```

Note: With AIX 5L 5.1 Maintenance Package 02 (IY28102), the **installp -a -d /dev/cd0 -X spimg** command fails. To bypass this problem, first mount the CD-ROM using the **mount -v cdrfs -r /dev/cd0 /mnt** command. When using smit install_latest, select /mnt instead of /dev/cd0 as the input device. Alternately, use **installp** from the command line.

8.6.4 Install PSSP prerequisites

PSSP has prerequisites for certain filesets. Make sure that the bos.net (TCP/IP and NFS) and bos.net.uucp (for Kerberos V4 and secure file collection systems only) files are installed on your control workstation. Make sure that the perfagent.tools fileset is installed on your control workstation. This file should have been placed in the lppsource directory. For AIX 5L 5.1 and 5.2 issue the command:

```
installp -agXd /spdata/sys1/install/name/lppsource/installp/ppc \
perfagent.tools
```

8.6.5 Install the runtime files

PSSP 3.5 has prerequisites for runtime libraries from the VisualAge C++ product. For AIX 5L 5.1, they are:

- vacpp.ioc.aix50.rte 5.0.2.0
- xlC.aix50.rte 5.0.2.0

The vacpp.ioc.aix50.rte fileset is not part of the AIX installation package. These files and their associated prerequisites are placed in your AIX lppsource as shown in Example 8-1 on page 291. They must be installed now. There may be more recent levels of these files available. Check the AIX fix distribution service Web site at:

```
http://techsupport.services.ibm.com/server/support
```

Note: The vacpp.ioc.* filesets are not part of the VisualAge C++ 6.0 installation package. You must continue to install vacpp.ioc.aix50.rte 5.0.2.*x* and vacpp.ioc.rte 5.0.2.*x* even if you also install VisualAge C++ 6.0.

For AIX 5L 5.1 and 5.2 issue the command:

```
installp -agXd /spdata/sys1/install/name/lppsource/installp/ppc \
xlC.rte xlC.aix50.rte vacpp.ioc.aix50.rte
```

8.6.6 Install the RSCT files

RSCT is no longer shipped with the PSSP product set. It is now integrated into AIX 5L and shipped with the AIX CDs. AIX 5L Version 5.1 or later installs RSCT by default. Because PSSP 3.5 only runs on AIX 5L Version 5.1 or later, there is no need to package RSCT with PSSP. You must install the RSCT shipped with AIX 5L 5.1 or AIX 5.2. See the table in “Step 21: Install PSSP on the control workstation” of *PSSP Installation and Migration Guide,* GA22-7347 for a list of the RSCT filesets required on the control workstation. The following command installs RSCT filesets from the lppsource directory:

```
installp -agXd /spdata/sys1/install/name/lppsource/installp/ppc rsct
```

8.6.7 Install the HMC-controlled server files

Install the filesets required for HMC-controlled servers as mentioned in “Additional files required on the CWS for HMC servers” on page 289 using the following commands. For example, for AIX 5L 5.1:

```
installp -agXd /spdata/sys1/install/name/lppsource/installp/ppc/\
Java130.rte Java130.xml4j
```

For AIX 5L 5.2:

```
installp -agXd /spdata/sys1/install/name/lppsource/installp/ppc/\
Java131.rte Java131.ext.xml4j
```

Install the openCIMOM fileset that you copied to /spdata/sys1/install/name/lppsource/RPMS/ppc using the command:

```
/bin/rpm -i /spdata/sys1/install/name/lppsource/RPMS/ppc/\
openCIMOM-0.61-1.aix4.3.noarch.rpm
```

8.6.8 Installation of PSSP filesets on the CWS

This is the final step for installation of the PSSP filesets on the control workstation. Refer to Step 21 of *PSSP Installation and Migration Guide,* GA22-7347 for a complete list of filesets required for installation.

8.7 Setting the authentication services on the CWS

Before you proceed further, the authentication methods for remote AIX commands and the SP trusted services need to be set on the CWS.

8.7.1 Authentication setting on the CWS for remote commands

Set up the authentication methods for AIX remote commands on the CWS with the **chauthent** command. Select one or more authentication methods for the CWS. Your choices are k5, k4, or standard. This setting is used to determine initial security settings for PSSP when the install_cw script is run. Refer to Table 8-5 for valid authentication settings.

Table 8-5 Authentication settings for AIX remote command

Authentication service	Command used
DCE	chauthent -k5
Kerberos V4	chauthent -k4
Standard AIX	chauthent -std
DCE and Kerberos V4	chauthent -k5 -k4
DCE and standard AIX	chauthent -k5 -std
Kerberos and standard AIX	chauthent -k4 -std
DCE, Kerberos and standard AIX	chauthent -k5 -k4 -std

Note: PSSP 3.4 or later provides the ability to remove the dependency PSSP has on the **rsh** and **rcp** commands issued as root by enabling the use of a secure remote command method.

Depending on which authentication method you have selected, complete the following steps.

Secure remote command method

Follow the installation steps in chapter 5 of *An Introduction to Security in a CSM 1.3 for AIX 5L Environment,* SG24-6873. You can also find a detailed discussion in 6.2, “Security-related concepts” on page 216.

Note: You need to generate the public and private keys manually if the installation of the node is done from NIM with AIX 5L 5.1 or later.

Kerberos authentication method

If you selected the Kerberos method of authentication, the Kerberos needs to be initialized depending on the type of configuration you selected.

Initialize PSSP Kerberos V4 (optional)

Depending on what type of Kerberos V4 authentication server you use, SP, AFS, or another MIT Kerberos V4 implementation—you need to follow the table in chapter 2 of *PSSP Installation and Migration Guide,* GA22-7347.

PSSP authentication provides a program, /usr/lpp/ssp/bin/setup_authent, to initialize PSSP authentication services on the cluster nodes (including the control workstation and all the external nodes) for Kerberos V4 authentication servers and client systems. This program defines instances of the `hardmon` and `rcmd` authenticated services, and does one of the following:

- Creates a primary Kerberos V4 authentication server and database.
- Creates a secondary Kerberos V4 authentication server and database.
- Configures the CWS as a Kerberos V4 client.
- Initializes the CWS or other cluster nodes to use AFS authentication.

Run the **`setup_authent`** command and follow the directions described in chapter 2 step 23 of *PSSP Installation and Migration Guide,* GA22-7347.

Configuring DCE for the CWS

If you want PSSP to use DCE authentication services, do the following:

- Install DCE on the CWS.
- Update the spsec_overrides file (optional).
- Create DCE groups, organizations, principals, and accounts.
- Create SP administrative principals.
- Create CWS-specific keyfiles.

Refer to chapter 2 step 24 of *PSSP Installation and Migration Guide,* GA22-7347 for detailed installation procedures.

8.7.2 Setting the authentication method for PSSP trusted services

Depending on the authentication method selected, you can set the authentication method for PSSP trusted services on the CWS as shown in Table 8-6.

Table 8-6 Setting trusted services on the CWS

Authentication selected	Enter
DCE	chauthts dce
Kerberos V4	chauthts compat

Authentication selected	Enter
Both DCE and Kerberos V4	chauthts dce compat
None	chauthts

Obtaining credentials for Kerberos

You need to be an authenticated user before you can proceed further with the installation process. If DCE is selected, use the **dce_login** command, and if Kerberos is enabled use the **k4init** command. For example, you can obtain the credentials for Kerberos with the following command:

```
k4init root.admin
```

8.8 Configuring and verifying the CWS

The configuration of the CWS is a two-step process:

1. Use the **install_cw** command to finish installing PSSP on the CWS.
2. Run SDR and system monitor verification tests to ensure that the CWS has been configured successfully by issuing the commands shown in Table 8-7.

Table 8-7 SDR and system monitor verification

Verification	Command	smit
SDR	SDR_test	**smit SP_verify** and select the **System Data Repository** option.
System monitor	spmon_itest	**smit SP_verify** and select the **System Monitor Installation** option.

Note: After the tests are run, the system creates the spmon_itest.log in /var/adm/SPlogs/spmon and the SDR_test.log in /var/adm/SPlogs. View these files for problem analysis if the verification fails.

SP manuals

We recommend that you read the following documentation:

RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment, GA22-7281. Chapters 2, 3, and 5 provide detailed information about SP connectivity and storage requirements, and site information.

PSSP Installation and Migration Guide, GA22-7347. Chapter 2 describes in detail the installation of the CWS, the PSSP packaging, the system configuration files, and the authentication services.

PSSP Command and Technical Reference (2 Volumes), SA22-7351 contains a complete description of each CWS installation command listed in 8.3, “Key commands and files” on page 278.

SP redbooks

We recommend that you read the following redbooks:

RS/6000 SP: PSSP 3 Survival Guide, SG24-5344. Chapter 2 describes the logical flow of steps for the installation process.

An Introduction to Security in a CSM 1.3 for AIX 5L Environment, SG24-6873. Chapter 5 offers a detailed discussion of the secure remote execution method, and describes its implementation.

8.9 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, “Answers to sample questions” on page 521.

1. The install_cw script performs the initial customization of PSSP onto the control workstation, configures the default partition, and starts the SP daemons necessary for the following steps of the Cluster 1600 installation. Which of the following is *not* done by the install_cw script?
   a. Setting the Authentication method.
   b. Initializing the SDR.
   c. Configuring the system as one partition.
   d. Updating /etc/inittab and /etc/services.
2. Which Authentication method for SP trusted Services would you execute if the Authentication method selected was Kerberos V4?
   a. chauthts dce
   b. chauthts compat
   c. chauthts dce compat
   d. chauthts
3. When will you execute the command **SDR_test**?
   a. After installing PSSP filesets

b. After installing RSCT filesets

c. After setting the Authentication services on the CWS

d. After completion of the **install_cw** command

4. What is the recommended location for setting any static routing information?

   a. rc.network file

   b. rc file

   c. rc.route file

   d. rc.net file

5. What type of protocol will you specify for configuring a pSeries POWER4 server in the SDR?

   a. SP

   b. CSP

   c. SAMI

   d. HMC

6. Which of the following functions can the **setup_authent** command *not* do?

   a. Configure Primary Kerberos V4 server

   b. Configure a secondary Kerberos V4 server

   c. Configure the SDR

   d. Configure the CWS as a Kerberos V4 client

7. Which of the following statement is true when connecting an HMC server to the CWS?

   a. Two serial cables are needed between CWS and the server.

   b. An Ethernet connection is the only requirement.

   c. One serial and one Ethernet cable are needed to link the CWS to the server.

   d. One serial cable is needed to link the CWS to the server.

8. What is the location for the AIX 5L 5.1 and later filesets in the installp format on the CWS?

   a. /spdata/sys1/install/<name>/lppsource/installp/ppc

   b. /spdata/sys1/install/pssplpp/code_version

   c. /spdata/sys1/install/<name>/lppsource/rpms/ppc

   d. /spdata/sys1/<name>/lppsource

9. Which of the following filesets is *not* mandatory for installing HMC controlled server?

   a. Java131.rte

   b. Java131.ext.xml4j

   c. openCIMOM

   d. bos.clvm

8.10 Exercises

Here are some exercises you may wish to do:

1. On a test system that does not affect any users, perform the migration of the CWS to a new version of AIX and PSSP.
2. On a test system that does not affect any users, modify the network tunables on the CWS to the values that fit you SP system.
3. Familiarize yourself with the following key commands: **`setup_authent`** and **`install_cw`**. (Note: These commands are to be used only on the control workstation and not on the nodes.)
4. Familiarize yourself with the READ THIS FIRST document and various PSSP and AIX software requirements for the CWS. These requirements are different for each release of PSSP and AIX.
5. Familiarize yourself with the prerequisite LPPs for the HMC-controlled servers.
6. Familiarize yourself with the protocols required for hardware control of different types of nodes in Cluster 1600.

9

Frame and node installation

In Chapter 8, “Configuring the control workstation” on page 275, we presented the initial configuration of the CWS. This chapter addresses all of the other steps of installing a Cluster 1600 system from the configuration of the PSSP software on the CWS through the installation of AIX and PSSP on the Cluster 1600 nodes up to the first boot of nodes and switches.

Note: The term Cluster 1600 here refers to a combination of SP frames, SP-attached servers, and HMC-controlled servers, unless we specify one.

9.1 Key concepts

Before taking the RS/6000 SP certification exam, you should understand the following frame, node, and switch installation concepts:

- The structure of the SDR configuration information: Site information, frame information, and node information.
- The contents of the predefined subdirectories of /spdata.
- The files used for SDR configuration and SP frame, node, and switch installation.
- NIM concepts applied to the Cluster 1600 environment.
- The setup of boot/install servers (primary and secondary).
- Network installation concepts.
- Automatic and manual node conditioning.
- Cluster 1600 system customization.
- Cluster 1600 partitioning and its impact on commands and daemons.

9.2 Installation steps and associated key commands

This section presents the commands most widely used during a Cluster 1600 system configuration and installation.

To help you understand the use of each command, they are presented in association with the installation step in which they are used and in the order in which they are first used during the installation process. Some of these commands may be used several times during the initial installation and the upgrades of Cluster 1600 systems. In this case, we also provide information that is not related to the installation step but that you may need at a later stage.

Finally, this section is *not* intended to replace the Cluster 1600 manuals referenced in 9.4, “Related documentation” on page 342. You should refer to these manuals for a more thorough understanding of these commands before taking the SP certification exam.

9.2.1 Enter site environment information

At this stage, we suppose that the PSSP software has been loaded on the CWS and that the SDR has just been initialized (the last command executed on the CWS was **`install_cw).`** We are now at the beginning of the Cluster 1600 node customization and installation.

Note: Perform the steps to set the password for secure File Collections.The procedure can be found in chapter 2 step 32 of *PSSP Installation and Migration Guide,* GA22-7347.

The first task is to define, in the SDR, the site environment data used by the installation and management scripts. This can be done using the command line interface, **spsitenv**, or its equivalent SMIT window, Site Environment Information window (`smitty site_env_dialog`). This must be executed on the CWS only.

The **spsitenv** command defines all site-wide configuration information, as follows:

- The name of the default network install image
- Your method of time service, the name of your time servers, and the version of NTP in use
- Whether you want to have the SP services configure and manage the Automounter
- User Admin information
- Whether you want to use SP User Management
- Whether you want SP File Collection Management installed and where the daemon will run
- Whether you want to use SP Accounting
- Whether you use the default lpp_source NIM resource directory as the location of the AIX file sets
- Whether you use the base AIX locale installed on the control workstation
- Whether ASCII-only data can be written to the SDR or whether non-ASCII data is allowed
- SP model number and SP serial number
- Cluster model number and serial number
- Whether you want to force the system to be nonpartitionable
- Whether you want to run with restricted root access (RRA) enabled

  Note: Some applications (such as GPFS, Problem Management, and HACMP) cannot run with RRA enabled.

- Whether you want to run with a secure remote command method instead of the default **rsh** and **rcp** commands from the PSSP code

Important: Restricted root access *must* be enabled to use the secure remote command method.

See chapter 2 step 33 in *PSSP Installation and Migration Guide,* GA22-7347 to learn how to set up your system to run with the secure remote command method. Also refer to *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281 for more information.

Because of the number of parameters you must provide on the **spsitenv** command, we recommend that you use the SMIT interface rather than the command line.

In our environment, the site configuration is defined as shown in Example 9-1.

Example 9-1 Site environment information

```
Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                [Entry Fields]
  Default Network Install Image                     [bos.obj.aix520.02312]
  Remove Install Image after Installs                false                  +
NTP Installation                                    consensus              +
  NTP Server Hostname(s)                            []
  NTP Version                                        3                      +
Automounter Configuration                           true                   +
User Administration Interface                       true                   +
  Password File Server Hostname                     [c179s]
  Password File                                     [/etc/passwd]
  Home Directory Server Hostname                    [c179s]
  Home Directory Path                               [/home/c179s]
File Collection Management                          true                   +
 File Collection daemon uid                         [102]
 File Collection daemon port                        [8431]                   #
SP Accounting Enabled                               false                  +
 SP Accounting Active Node Threshold                [80]                     #
 SP Exclusive Use Accounting Enabled                 false                  +
 Accounting Master                                  [0]
Control Workstation LPP Source Name                 [aix52B_313C]
SP Administrative Locale                            en_US                  +
 SDR may contain ASCII data only                     true                  +
Root remote command access restricted               false                  +
Remote command method                               rsh                    +
Remote command executable                           []
Remote copy executable                              []
SP Model Number                                     [123]
SP Serial Number                                    [0200abcde]
```

```
Cluster Model Number                              [9078-160]
Cluster Serial Number                             []
Force Non Partitionable                            false                 +
BOTTOM]
```

9.2.2 Enter Hardware Management Console (HMC) information (HMC-controlled servers only)

You must perform two major steps if your SP system or clustered server system includes HMC-controlled servers.

1. Establish a trusted network between the CWS and HMC

   If you do not consider the SP Ethernet administrative LAN a trusted network and you want to establish one between the CWS and the HMC, follow the procedure described in chapter 2 step 34.1 Option A or Option B in *PSSP Installation and Migration Guide,* GA22-7347 to establish an HMC trusted network when your HMC-controlled server is not configured to your SP system.

2. Entering HMC information

   This is a two-step procedure:

   a. Configuring the HMC and the HMC controlled servers locally, before the HMC-controlled server is defined to PSSP. This involves various steps, discussed in brief:

      - Ensure that the HMC is installed and configured to operate either on the SP Ethernet administrative LAN network, or on the HMC trusted network.
      - Use the HMC User Management interface to define a user ID with the role of System Administrator and assign a password. This information will be required when defining the HMC to the CWS in the next section.
      - Ensure that the pSeries server (controlled by the HMC) is recognized by the HMC.
      - You can change the system name from the default name set by the HMC if required. Note the defined system name. This information will be required when entering the non-SP frame information for this server.

      Note: If the system name is changed in the future, the new name will then also need to be changed in the non-SP frame information stored in the SDR on the CWS.

- Use the HMC Partition Management interface to select the desired power-on mode for your system: full system partition (SMP) mode, logical partition standby mode, or physical partition mode.
- If you selected logical partition standby mode or physical partition mode, use the HMC Partition Management interface to create partitions and profiles as necessary.
- View the properties for each partition object and note the partition ID. Each *partition* is represented in PSSP as a *thin node*.
- Use the HMC maintenance: system configuration interface to enable the *remote virtual terminal*.

  Note: This function is required by operations such as the **s1term**, **sphrdwrad, spadaptr_loc**, and **nodecond** commands, and by PSSP administrative processes to transfer sensitive data files.

- Configure the security mode of the Object Manager to *Plain Socket*.

  Important: The PSSP hardware control and monitor functions fail if the Object Manager security mode is set to use SSL protocols.

- Set server policy to *Power off the system after all the logical partitions are powered off*. Use the HMC server and partition to disable this option for each POWER4 server managed by the HMC.

b. Configuring the CWS before entering non-SP frame information for your HMC-controlled servers.

   - Use the AIX **ping** command to verify that the CWS has network connectivity to each HMC.

     In an HMC-managed environment, the hardmon daemon does not communicate with the server hardware. It connects to the HMC through the daemon hmcd running on the CWS. To secure this connection, we need a user ID and password specified for hardmon. This must be done for every HMC we want to add to the Cluster 1600 system. Use the **sphmcid** command to store the user ID and password for the hardware monitor to use when establishing a remote client session with an HMC. When running this command, you will be prompted for the password to be stored with the specified user ID. Define the previously created HMC user ID to PSSP for hardmon as shown in Example 9-2.

Example 9-2 Setting hardmon authentication for HMC

```
sp4en0:/
root $ sphmcid sp4hmc hscroot
Password:
Verifying, please re-enter Password:
sphmcid: HMC entry updated.
```

Where sp4hmc is the HMC hostname and hscroot is the user_id.

- Define the switch node numbers for your nodes.

Refer to *Hardware Management Console for pSeries Installation and Operations Guide,* SA38-0590 and chapter 2 step 34.2 of *PSSP Installation and Migration Guide,* GA22-7347 for more details.

9.2.3 Enter frame information

After defining the site environment in the SDR, you must describe in the SDR the frames existing in your SP system and how they are numbered. Here we discuss how to enter frame configuration information into the SDR. This task is performed using either the command line interface, **spframe**, or its SMIT equivalent windows, SP Frame Information (`smitty sp_frame_dialog`) and non-SP Frame Information (`smitty nonsp_frame_dialog`). This task must be executed on the CWS only.

Since PSSP 3.4, this command also defines the hardware protocol used on the serial link (SP for SP nodes, SAMI for SP-attached servers, CSP for pSeries Power3 servers like 6H0, 6H1, 6M1, and HMC for pSeries Power4 servers) and the switch port to which a non-SP frame is attached. This command must be performed during the first installation of a Cluster 1600 system and also each time a new frame is added to the system. By specifying the **start_frame** argument for each frame in the SP system, it is possible to skip the frame number and to leave room for system growth and later addition of frames between the frames installed originally in the system.

In our environment, we define the first frame using the **spframe** command as follows:

```
spframe -r yes 1 1 /dev/tty0
```

This command enters information for one frame and indicates that frame 1 is connected to /dev/tty0 and reinitializes the SDR. The second frame will be defined later in Chapter 14, “RS/6000 SP reconfiguration and update” on page 437.

Enter non-SP frame information and reinitialize the SDR

SP-attached servers and HMC-controlled servers also require frame objects in the SDR. These frames are referred to as non-SP frames, and one object is required for each server attached to your SP. These objects have a non-SP hardware protocol associated with them, which instructs PSSP as to which method of hardware communication is to be used for controlling and monitoring the node associated with this frame object. Valid hardware protocol values for different pSeries models and the ttys required for them are discussed in Table 8-2 on page 285.

The servers that use the SAMI hardware protocol require two tty port values to define the tty ports on the CWS to which the serial cables connected to the server are attached. The tty port value defines the serial connection to the operator panel on these servers for hardware controls.The s1 tty port value defines the connection to the serial port on the servers for serial terminal (s1term) support. For more details on the physical connectivity, refer to Figure 5-1 on page 167.

Switch port numbers are required on SP Switch or switchless systems for each SP-attached server in your system. For HMC-controlled servers in an SP Switch or in a switchless system where the system is not forced to be non-partitioned, a switch node number is required for each logical partition (LPAR). These switch node numbers must be specified to PSSP through the /etc/switch.info file. See the switch.info file in chapter 2 of *PSSP Command and Technical Reference (2 Volumes),* SA22-7351 for detailed information on this file.

An example of the /etc/switch.info file might contain the following entries for an HMC-controlled server that will be defined as frame 5 with three LPARs attached to switch 2 in the system:

```
# Node_number Switch_node_number

65                  16
66                  17
67                  18
```

Important: 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.

Refer to chapter 2 from page 67 through page 69 of *PSSP Installation and Migration Guide,* GA22-7347 for entering the non-SP Frame installation. Also refer to chapter 6 in the section “Adding an HMC server” of *IBM (e)server Cluster*

1600 Managed by PSSP 3.5: What's New, SG24-6617 for viewing the sample output of the **splstdata** command at different stages of adding the SP-attached servers and the nodes.

9.2.4 Check the level of supervisor microcode

Once the frames have been configured, and before starting to configure the nodes, we recommend that you check to make sure that the SP frame microcode, known as supervisor code, is at the latest level supported by the PSSP being installed. The PSSP software contains an optional fileset, ssp.ucode, that must have been installed on the CWS to perform this operation.

The **spsvrmgr** command manages the supervisor code on the frames. It executes on the CWS only. It can be called from the command line or from SMIT. Each of the command line options is equivalent to one of the functions accessible from the SMIT RS/6000 SP Supervisor Manager window.

This command can be used to query the level of the supervisor code or to download supervisor code from the CWS onto the SP frame. We recommend that you use the SMIT panels to perform these operations. However, two commands can be used for system-wide checking and updating:

- `spsvrmgr -G -r status all`

  indicates if the supervisor microcode is up-to-date or needs to be upgraded.

- `spsvrmgr -G -u all`

  updates the supervisor microcode on all parts of the SP system.

  Since the **-u** option usually powers off the target of the **spsvrmgr** command, it is highly recommended to upgrade the SP system at the beginning of the SP system installation rather than later when the system is in production.

Note: You must have the latest version of ssp.ucode installed that is appropriate for your PSSP level before proceeding.

9.2.5 Check the previous installation steps

Since the complete PSSP and Cluster 1600 system installation is a complex process involving more than 85 steps, it is a good idea to perform some checking at several points of the process to insure that already executed steps were successful. Refer to Example 9-3 to verify that the System Monitor and Frame information was correctly installed.

Example 9-3 Verifying System Monitor and Frame information

```
spmon_ctest
```

After the tests are run, the system creates a log in /var/adm/SPlogs/spmon called spmon_ctest.log.

```
spmon -d -G
This command displays all the Frame and Nodes configured in the previous steps.
```

At this point in the installation, you can use **splstdata -f** and **splstdata -n** to verify that the frames are correctly configured in the SDR and that the **spframe** command correctly discovered the nodes in each frame.

9.2.6 Define the nodes' Ethernet information

Once the frame information is configured, and the microcode level is up-to-date, we define in the SDR the IP addresses of the en0 adapters of each of the nodes, as well as the type and default route for this Ethernet adapter.

This task is performed by the **spadaptrs** command, which executes only on the CWS, on the command line, or through its equivalent SMIT window, SP Ethernet Information (`smitty sp_eth_dialog`). The **spadaptrs** command can define adapters individually, by group, or by ranges. This step adds IP address-related information to the node objects in the SDR. It also creates adapter objects in the SDR for the SP Ethernet administrative LAN adapters on your nodes. This information is used during node customization and configuration.

For HMC-controlled servers, specifying the SP Ethernet adapter by its physical location code is suggested, especially if there is more than one Ethernet adapter present on the node. The physical location code for an adapter can be determined by running the **spadaptr_loc** command for that node. The command will return a list of all the SP-supported adapters installed on the node.

Example 9-4 on page 310 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.

Example 9-4 getting the Ethernet address with the spadaptr_loc command

```
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
```

Important:

- Do not try **spadaptr_loc** on the production system, since this command will power off the node. On the running LPAR, there is the **lsslot** command to show adapter location codes.
- The default route that you enter in this step is not the same as the default route on the node. The route that you enter here goes in the SDR node class. It is the route over which the node communicates with its boot/install server (for example, install, customize, and so on). The default route must be a valid path from the SP Ethernet administrative LAN adapter to the node's boot/install server and the CWS.

  The default route on the node is the route it will use for its network communications if there is no specific route to the destination. During the boot process, this is set to the default route in the SDR.

  In order for the route to remain set after customization, also set the route in /etc/inittab after the line that runs rc.sp. For the switch, set the route in /etc/inittab after the line that runs rc.switch.

We now enter information about the nodes attached to each Ethernet adapter using Perspectives, SMIT, or the **spadaptrs** command. Refer to Example 9-5 for configuring IPs for nodes with this command

Example 9-5 Command to configure IPs for 16 nodes.

```
spadaptrs -e 129.33.32.200 -t tp -d full-f 100 1 1 16 en0 \
129.33.32.1 255.255.255.192
```

This example configures an SP Ethernet administrative LAN adapter network of 16 nodes with IP addresses ranging from 129.33.32.1 to 129.33.32.16, a netmask of 255.255.255.192, and a default route of 129.33.32.200 for a twisted-pair Ethernet using 100 mbps full duplex for the communication transfer and rate:

Refer to Example 9-6 to configure IPs for HMC-controlled servers. The location of the Ethernet adapter was found when we ran the **spadaptr_loc** command as shown in Example 9-4 on page 310.

Example 9-6 Command to configure HMC-controlled servers

```
spadaptrs -P U0.1-P1-I1/E1 -t tp -d full -f 100 2 1 1 en 129.33.32.65 \
255.255.255.192
```

This example configures the adapter on the SP Ethernet administrative LAN adapter for the first logical partition of an HMC-controlled server. The adapter is a twisted-pair Ethernet adapter with communication transfer and rate set to 100 Mbps full duplex. The IP address is 129.33.32.65 with a netmask of 255.255.255.192. An HMC-controlled server is represented as frame 2, the node is assigned slot 1, and the adapter is located at physical location U0.1-P1-I1/E1 retrieved from the example Example 9-4 on page 310.

9.2.7 Discover or configure the Ethernet hardware address

Once the nodes' en0 IP addresses are known, the SDR must be loaded with the Hardware (MAC) address of these en0 adapters for future use by the bootp protocol during the installation of the AIX image onto each node through the network. This task is performed by **sphrdwrad**, only on the CWS, as a command or by using the SMIT Get Hardware Ethernet Address window (`smitty hrdwrad_dialog`).

You can provide this information, if you already know it, by creating the file /etc/bootptab.info (for more details, see 9.3.1, "/etc/bootptab.info" on page 334) to speed up the **sphrdwrad** command. For each node in the argument list of the command, **sphrdwrad** will look to find its hardware address in /etc/bootptab.info. If not found, it then queries the node through the hardware connection to the frame. In the latter case, the node will be powered down and powered up. If you are performing this step for an HMC-controlled server, you may already have the hardware Ethernet addresses available to you when you run the **spadaptr_loc** command as shown in Example 9-4 on page 310.

Note: Do not use the **sphrdwrad** command on a running node since it will be powered off.

The following examples show commands for acquiring the Ethernet addresses of nodes, and verifying the same.

- Acquiring the Ethernet addresses for all the nodes in an SP frame:

  ```
  sphrdwrad 1 1 rest
  ```

- Acquiring the Ethernet addresses for specified nodes. The -l flag specifies the nodes:

  ```
  sphrdwrad -l 10,12,17.
  ```

- You can verify whether the Ethernet addresses were placed in the SDR node object with the following command:

  ```
  splstdata -b
  ```

9.2.8 Configure additional adapters for nodes

In addition to the en0 adapters, the Cluster nodes can have other adapters used for IP communication: A second Ethernet adapter for connecting to a corporate backbone or to another segment of the SP Ethernet administrative network, Token Ring adapters, and so on.

The **spadaptrs** command is used to configure these additional adapters into the SDR. It executes on the CWS only, using the command line interface or the equivalent functions accessible from the SMIT Additional Adapter Database Information window (smitty add_adapt_dialog).

You can configure the IP address of individual adapters or range of adapters with **spadaptrs**; you can specify the type of adapter (Ethernet, Token Ring, and so on), and you can specify the subnet mask associated with the adapter, and so on.

Only Ethernet, Token Ring, FDDI, and Switch (css0) adapters can be configured using **spadaptrs**. Other types of adapters (PCA, ESCON) cannot be configured this way. You must either configure them manually after the nodes are installed, or write configuration code for them in the shell customization script firstboot.cust (see “firstboot.cust” on page 339).

Note: Ensure that all additional adapters listed previously are configured before performing the following operations:

- Node installation
- mksysb installation
- Node migration
- Node customization

You can use either the SMIT panel or the command line interface to configure the additional adapters. The following three examples show how to configure the adapters.

1. Configuring switch adapter css0

   The following **spadaptrs** command adds SDR information for a css (SP Switch and SP Switch2) network of 30 nodes (frame 1 slot 1 to frame 2 slot 16, with a wide node as the first node in each frame and the rest thin nodes, and a switch on each frame) with IP addresses from 129.33.34.1 to 129.33.34.30, and a netmask of 255.255.255.0. The IP addressing corresponds to the slots in the frame, with each wide node incrementing by 2 and each thin node incrementing by 1, and each high node by 4. If you specify the **-s** flag to skip IP addresses when you are setting the css switch

addresses, you must also specify **-n no** to not use switch numbers for IP address assignment, and **-a yes** to use ARP.

```
spadaptrs -s yes -n no -a yes 1 1 30 css0 129.33.34.1 255.255.255.0
```

2. Configuring the additional adapters

   The following **spadaptrs** command adds SDR information for an fi0 (FDDI adapter) network of 30 nodes (frame 1 slot 1 to frame 2 slot 16, with a wide node as the first node in each frame and the rest thin nodes) with IP addresses from 129.33.34.1 to 129.33.34.30, and a netmask of 255.255.255.0. The IP addressing corresponds to the slots in the frame, with each wide node incrementing by 2 and each thin node incrementing by 1.

   ```
   spadaptrs -s yes 1 1 30 fi0 129.33.34.1 255.255.255.0
   ```

3. Configuring the Ethernet adapter for the HMC server

   This following **spadaptrs** command adds SDR information for an additional Ethernet adapter for the second logical partition in an HMC-controlled server. The adapter is a twisted pair Ethernet adapter with full duplex, the speed set to 100 Mbps. The IP address is 129.33.35.66 with a netmask of 255.255.255.0. An HMC-controlled server is represented as frame 2, the node is assigned slot 1, and the adapter is located at the physical location U0.1-P1/E1 as found by the **spadaptr_loc** command in Example 9-4 on page 310.

   ```
   spadaptrs -P U0.1-P1/E1 -t tp -d full -f 100 2 1 1 en \
   129.33.35.66 255.255.255.0
   ```

Configure the aggregate IP interface for the nodes

To use the ml0 interface for running jobs over the switch (this step is optional for an SP Switch2 system), refer to Example 9-7 on page 314 to configure the aggregate IP.

Example 9-7 Adding aggregate IP address

```
To add an aggregate IP address of 9.114.66.20 and a network mask of
255.255.255.0 for device ml0 on node 7, enter:

spaggip -i css0,css1 -1 7 9.114.66.20 255.255.255.0
```

For the switch adapters, two options, **-a** and **-n**, make it possible to allocate IP addresses to switch adapters sequentially based on the switch node numbers. In our environment, we only need to define the second Ethernet adapter of node1 and the switch adapters (css0) of all SP nodes:

```
spadaptrs -s yes -n no -a yes 1 1 1 en1 192.168.31.11 255.255.255.0
spadaptrs -s yes -n no -a yes 1 1 12 css0 192.168.13.1 255.255.255.0
```

9.2.9 Assign initial host names to nodes

Once the SDR contains all IP information about the adapters of all nodes, you can change the host name of the nodes, also known as initial host name.This optional step is performed using **sphostnam,** on the CWS only, as a command or through the SMIT Hostname Information window (`smitty hostname_dialog`).

The default is to assign the long symbolic name of the en0 adapter as the host name of the node. If your site policy is different (for example, you may want to give to the node, as host name, the name of the adapter connected to your corporate network), you use **sphostnam** to change the initial host name. Again, like the previous one, this command applies either to one node or to a range or list of nodes. The following **sphostnam** command changes the format of the name and uses the short names but keeps the en0 adapter name as host name:

```
sphostnam -f short 1 1 12
```

9.2.10 PSSP security installation and configuration

This section describes how to configure and customize the PSSP selected authentication and authorization methods:

- Select security capabilities required on nodes.
- Create DCE host names.
- Update SDR with DCE Master Security and CDS server host names.
- Configure DCE Clients (Admin portion).
- Configure SP Trusted Services to use DCE authentication.
- Create SP Trusted Services DCE Keyfiles.
- Select Authorization Methods for AIX Remote Commands.
- Enable Authentication Methods for AIX Remote Commands.
- Enable Authentication Methods for SP Trusted Services.

Select security capabilities required on nodes

This step sets the security capabilities to be installed on the nodes. If DCE is selected, the DCE filesets are installed on the nodes, and the security, CDS, clients, and RPC are configured and started. The DCE filesets must be located in /spdata/sys1/install/name/lppsource/installp/ppc on the CWS to be installed automatically. If k4 is selected, various Kerberos V4 configuration files are installed. By default, AIX standard authentication is part of the AIX BOS and, therefore, no installation is required on the node. Example 9-8 sets the security capabilities to be installed on the nodes. The DCE steps are optional and only need to be executed if you are using the DCE.

Example 9-8 Setting authentication capability on nodes

```
spsetauth -p partition1 -i k4 std
To set partition "partition1" to have kerberos V4, and Standard AIX
as the set of authentication methods.
```

Refer to chapter 2 step 48 to 52 of *PSSP Installation and Migration Guide,* GA22-7347 for setting the authentication methods for DCE-enabled nodes.

Select authorization method for AIX remote commands

You now have to create the appropriate authorization files for use by root's remote commands, such as **rcp**, **rsh**, and so on, on the CWS. Possible methods are Kerberos 4, AIX standard authentication, and Kerberos 5. If some system partitions have authorization methods for AIX remote commands defined, the .klogin, .rhosts, and .k5login files will be created for each of the authorizations enabled. On the CWS, you can use either **spsetauth** or the SMIT Select Authorization Methods for Root access to Remote Commands window (`smitty spauth_rcmd`). Here we configured both Kerberos 4 and AIX standard methods for the default partition "partition1" as follows:

```
spsetauth -d -p partition1 k4 std
```

Note: A new option of *none* has been added to this menu. If *none* is selected, no other authorization methods can be selected at the same time for the selected system partition. The *none* option can be selected only if all nodes are at PSSP 3.4 or later. You need to enable the secure remote command method and the restricted root access (RRA) to be enabled.

Enable authentication methods for AIX remote commands

You can now choose the authentication methods used for System Management tasks. Valid methods are Kerberos 4, standard AIX, and Kerberos 5 (DCE). You perform this task only on the CWS using either **chauthpar** or the SMIT Select Authorization Methods for Root access to Remote Commands window (`smitty spauth_methods`). The **chauthpar** command in the following example enables the authentication with k4 and standard methods for partition "partition1" as follows:

```
chauthpar -c -p partition1 k4 std
```

Enable authentication methods for SP trusted services

This step enables the authentication method that will be used for the SP trusted services. The **chauthpts** command shows how to enable the authentication method as follows:

```
chauthpts -c -p partition1 compat
```

Note: You do *not* have to execute this step if you only have AIX standard security enabled.

If you enabled the DCE in the previous steps, you need to start the key management daemon with the **/usr/lpp/ssp/bin/spnkeyman_start** command.

After this you can add the extension node (optional). Refer to chapter 10 of *PSSP Installation and Migration Guide,* GA22-7347 for the installation procedure.

9.2.11 Start RSCT subsystems

After the SDR has been loaded with the frame and node information, IP addresses, symbolic names, and routes, you need to add and start the RSCT subsystems. Topology services (hats), host response (hr) are examples of RSCT subsystems.

Important: PSSP is one of the products using the old RSCT design.

RSCT subsystems are managed by the **syspar_ctrl** command and are listed in the file /usr/lpp/ssp/config/cmi/syspar_subsystems. The **syspar_ctrl** command controls the system partition-sensitive subsystems on the CWS and on the Cluster 1600 nodes. The following **syspar_ctrl** command (it is used only on the CWS since the nodes are still not installed) starts the system partition-sensitive subsystems:

```
syspar_ctrl -A
```

Refer to Example 9-9 to verify that all the required RSCT subsystems have been started successfully. For a default system partition “partition1” issue the following command:

```
lssrc -a | grep partition1
```

This should return the following output:

Example 9-9 Verifying success of syspar_ctrl

```
hags.partition1 hags 17134 active
hats.partition1 hats 22266 active
hr.partition1 hr 18228 active
haem.partition1 haem 21128 active
hagsglsm.partition1 hags 21338 active
haemaixos.partition1 haem 41000 active
Emonitor.partition1 emon inoperative
```

Note: To continue with the install, the subsystems listed in Example 9-9 *must* all be active (except for the Emonitor).

If a single subsystem is inactive, simply try starting that particular subsystem by issuing **syspar_ctrl -s** subsystem_name. For example:

- If the subsystem is hags, issue:

```
syspar_ctrl -s hags
```

- If more than one subsystem is inactive, stop and delete all of the RSCT subsystems with:

```
syspar_ctrl -D
```

- Then try to add and start all of the RSCT subsystems with:

```
syspar_ctrl -A
```

If you still have inactive RSCT subsystems, refer to chapters 23, 24, and 25 of *PSSP Diagnosis Guide,* GA22-7350 for diagnostic procedures.

Execution of this command on the CWS only starts the daemons on the CWS and not on any Cluster 1600 nodes. Since the daemons need to execute on all machines of the SP system for the subsystem to run successfully, **syspar_ctrl -A** must also be executed on each node when it is up. This is performed automatically at reboot time by the /etc/rc.sp script. For a complete description of the RSCT, refer to *RSCT for AIX 5L: Guide and Reference,* SA22-7889.

9.2.12 Set up nodes to be installed

Do this step if you want to change the default installation settings for any of the nodes. To find out the default settings of your nodes, use the **splstdata** command. This section describes the following:

- Changing the default boot/install information for the node objects in the SDR so that you can indicate a different boot/install server configuration to the Cluster 1600 system.
- Allows you to specify an alternate disk or disks to use when installing AIX on nodes.

Using multiple boot/install servers

If you want different nodes to be installed by a different boot/install server, you must specify the target nodes and which node will serve as the boot/install server. For example, the first node of your second frame, node 17, will be a boot/install server for the remaining 15 nodes in the second frame. You can use the **spchvgobj** command to enter this information into the SDR. The syntax used

with the **spchvgobj** command as shown below specifies a start frame of 2, a starting slot of 2, and a count of 15 nodes:

```
spchvgobj -r selected_vg -n 17 2 2 15
```

Note: You cannot export /usr or any directories below /usr because an NFS export problem will occur. If you have exported the /spdata/sys1/install/image directory or any parent directory, you must unexport it using the **exportfs -u** command before running **setup_server**. You need to do this because NIM attempts to export /spdata/sys1/install/images/bos.obj.ssp.*, where bos.obj.ssp.* is the install image during **setup_server** processing.

You can install multiple boot/install servers in your configuration depending on the number of nodes configured. Refer to chapter 2 step 60 “Using multiple Boot/Install server” of *PSSP Installation and Migration Guide,* GA22-7347 for the setup procedure of multiple boot/install server.

Note: We do not recommend using multiple boot/install servers in RRA, since it is not automatically supported by PSSP.

Selecting an installation disk

There are five ways to specify the disk or disks to use for installation.

The hardware location format

For example, to specify a single SCSI drive, enter:

00-00-00-0,0

or enter multiple hardware locations separated by colons:

00-00-00-0,0:00-00-00-1,0

The device names format

For example, to specify a single device name, enter:

hdisk0

or enter multiple device names separated by commas:

hdisk0,hdisk0

The parent and connwhere attributes format

To specify the parent-connwhere attribute:

ssar//0123456789ABCDE

Note: The parent-connwhere format should only be used for SSA drives. For more information on acquiring ssar numbers, see *IBM AIX: Kernel and Subsystems Technical Reference, Volume 2*.

The PVID format

For example:

00d4c45202be737f

To specify multiple disks by their PVID values, separate the specifications using colons:

00d4c45202be737f:00d4c452eb639a2c

The SAN target and logical unit identifier format

For example, if the SAN target worldwide port name for a fibre channel attached disk is 0x50060482bfd12c9c and the LUN ID is 0x8000000000000000, the SAN_DISKID specification will be:

0x50060482bfd12c9c//0x8000000000000000

To specify multiple fibre channel disks, separate the specifications using colons:

0x50060482bfd12c9c//0x8000000000000000:0x50060482bbffd7cb//0x0

Tip: Use the AIX **`lsattr -EH -l`** hdisk command to determine the worldwide port name and LUN ID for a disk.

The hardware location, SSA parent-connwhere, PVID, and SAN_DISKID formats can be used together as follows:

00-00-09-0,1:ssar//0123456789ABCDE:00d4c45202be737f

Selecting a kernel mode

With PSSP 3.5, you have the option of selecting your AIX kernel mode. This choice also determines the default file system, JFS, or Enhanced JFS. Your choices are:

- 32-bit kernel: Defaults to the Journal File System(JFS)
- 64-bit kernel: Defaults to the Enhanced Journal File System(JFS2)

The kernel mode is determined by the selected image. Two images are shipped with PSSP 3.5: The bos.obj.ssp.510 image uses the 32-bit kernel mode and the bos.obj.ssp.510_64 uses the 64-bit kernel mode. Be sure to install

bos.obj.ssp.510_64 only on nodes that support 64-bit kernel mode. Use the **spchvgobj** command to override the default for a particular node.

Note: Refer to the following Web site to view the various AIX processors that support the 64-bit kernel:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/aix52.htm

Specifying your own image

The default installation assumes that your nodes are not preinstalled. If you want to have them installed with your own install image, specify the following:

- Which nodes you are installing with your own install image
- The name of the installation image you are using, if you do not want to use the default image

Mirroring the root volume group

One way to significantly increase the availability of the SP system is to set up redundant copies of the operating system on different physical disks using the AIX disk mirroring feature. You can specify how many copies and which disks to use with the **spchvgobj** command.

9.2.13 spchvgobj

The **spchvgobj** command executes on the CWS only. It is equivalent to the SMIT Change Volume Group Information window (`smitty changevg_dialog`).

The PSSP installation scripts use a default configuration for the boot/install servers, the AIX image (mksysb) that will be installed on each node, and the disk where this image will be installed. This default is based on information that you entered in the site environment information panel. The default is to define the following as the boot/install servers:

- The CWS for a one-frame system
- The CWS and the first node of each frame in a multi-frame system

The default is to use rootvg as the default bootable volume group on hdisk0. If you wish to use a different configuration, you can use the SMIT equivalent as shown in Example 9-10, or the **spchvgobj** command as shown in “Changing SDR information for the nodes” on page 322 to specify the following:

  - Set of nodes
  - Bootable volume group
  - The names of disks to boot from

- Where to install the AIX image
- The number of mirrored disks
- The name of the boot/install server
- Where to fetch the AIX image
- The name of the installable image
- The name of the AIX lppsource directory
- The level of PSSP to be installed on the nodes
- The state of the quorum on the nodes

Example 9-10 Sample SMIT screen of the spchvgobj

```
Change Volume Group Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                 [Entry Fields]
  Start Frame                                        [1]
#
  Start Slot                                         [1]
#
  Node Count                                         [16]

  OR

  Node List                                          []

  Volume Group Name                                  [rootvg]
  Physical Volume List                               [00-07-00-0,0]
  Number of Copies of Volume Group                   [2]                      +
  Boot/Install Server Node                           [0] #
  Network Install Image Name                         [bos.obj.ssp.sample]
  LPP Source Name                                    [aix510]
PSSP Code Version                                   PSSP-3.5              +
  Set Quorum on the Node                             True +
[BOTTOM]
```

Changing SDR information for the nodes

You can use the spchvgobj command using the hardware location format for disk locations 00-07-00-0,0 and 00-07-00-1,0 for node 9 and set the number of copies to two. For example:

```
spchvgobj -r selected_vg -h 00-07-00-0,0:00-07-00-1,0 -1 9 -c
```

If you need to change the lpp_source NIM resource name from default to a new lpp_source NIM resource name such as aix510 for nodes 1 through 16, issue:

```
spchvgobj -r selected_vg -v aix510 1 1 16
```

If you need to change the install_image_name from default to a new install_image_name such as bos.obj.ssp.510 for nodes 17, 18, 21, 22, issue:

```
spchvgobj -r selected_vg -i bos.obj.ssp.510 -v aix51 -l 17,18,21,22
```

9.2.14 Verify all node information

Table 9-1 verifies that all the node information has been correctly entered into the SDR.

Table 9-1 Verifying node information in SDR.

To Display SDR	Enter
Site environment data	**splstdata -e**
Frame data	**splstdata -f**
Node data	**splstdata -n**
Adapter data	**splstdata -a**
Boot/install data	**splstdata -b**
SP expansion I/O data	**splstdata -x**
SP security settings	**splstdata -p**
Switch data	**splstdata -s**

At this point, you can optionally verify extension node information. Detailed descriptions can be found in chapter 10 of *PSSP Installation and Migration Guide,* GA22-7347.

9.2.15 Change the default network tunable values

When a node is installed, migrated, or customized (set to customize and rebooted), and that node's boot/install server does not have a /tftpboot/tuning.cust file, a default file of system performance tuning variable settings in /usr/lpp/ssp/install/config/tuning.default is copied to /tftpboot/tuning.cust on that node. For AIX 5L 5.2, if a node's boot/install server does not have a /tftpboot/nextboot file, a default file of system performance tuning variable settings in /usr/lpp/ssp/install/config/nextboot.default is copied to /tftpboot/nextboot on that node. The three alternate tuning files that contain initial performance tuning parameters for three different SP environments are:

1. /usr/lpp/ssp/install/config/tuning.commercial

   Commercial contains initial performance tuning parameters for a typical commercial environment.

2. /usr/lpp/ssp/install/config/tuning.development

   Development contains initial performance tuning parameters for a typical interactive/development environment.

3. /usr/lpp/ssp/install/config/tuning.scientific

   Scientific contains initial performance tuning parameters for a typical engineering/scientific environment.

You can also create your own alternate tuning file. Refer to chapter 2, step 63 of *PSSP Installation and Migration Guide,* GA22-7347 for additional information.

Note: Because an AIX 5L 5.2 node may be a boot/install server for a node earlier than AIX 5L 5.2, both the tuning.cust and nextboot files are transferred to the nodes.2

For the latest performance and tuning information, refer to the following Web site:

`http://techsupport.services.ibm.com/server/spperf.`

9.2.16 Perform additional node customization

Do this step to perform additional customization, such as:

- Adding installp images
- Configuring host name resolution
- Setting up NFS, AFS, or NIS
- Configuring adapters that are not configured automatically
- Modifying TCP/IP configuration files
- Setting time zones

PSSP provides the flexibility to run customer-supplied scripts during node installation. Following are the scripts, two of which are discussed in detail in 9.3.2, “/tftpboot” on page 335.

script.cust

This script is run from the PSSP NIM customization script (pssp_script) after the node’s AIX and PSSP software have been installed, but before the node has been rebooted. This script is run in a limited environment where not all services are fully configured. Because of this limited environment, you should restrict your

use of script.cust to functions that must be performed prior to the post-installation reboot of the node.

firstboot.cust

This script is run during the first boot of the node immediately after it has been installed. This script runs in a more “normal” environment where most all services have been fully configured. It is a preferred location for node customization functions that do not require a reboot of the node to become fully enabled.

firstboot.cmds

When in restricted root access mode and secure remote command mode, this sysctl script is run on the CWS during node installation to copy critical files from the CWS to the nodes. It is enabled in the firstboot.cust script. See the firstboot.cmds and firstboot.cust files for information on how to set up and enable this script for sysctl.

Note: There are special considerations to take into account if you are going to install nodes with secure remote command methods enabled.

Refer to chapter 2, step 64 of *PSSP Installation and Migration Guide,* GA22-7347 for additional information and consideration.

9.2.17 spbootins

The **spbootins** command executes on the CWS only. It is equivalent to the SMIT Boot/Install Server Information window (`smitty server_dialog`).

As mentioned in 9.2.13, “spchvgobj” on page 321, most of the boot/install server configuration in PSSP is associated with a volume group. The **spbootins** command is used to define in the SDR a set of SP nodes, the volume group on which they will boot, and how they will perform their next boot (from a server or from their disk). To specify that all nodes in frame one, at their next reboot, are to load the AIX image from their respective boot/install server and to ask not to run setup_server, run the following command:

```
spbootins -s no -r install 1 1 12
```

Note that the `-s` flag if set to `no` will not execute the set_server command.

9.2.18 Setting the switch

This section describes the procedure to set the switch. We recommend that you refer to chapter 14, “Using a Switch,” of *PSSP Administration Guide,* SA22-7348 for detailed information on setting up a switch. You also need to refer to all the

commands discussed in this section in the *PSSP Command and Technical Reference (2 Volumes),* SA22-7351 guide. There are four major steps in setting the switch.

Setting the switch topology in SDR

If a switch is part of the SP system, you now have to store the switch topology into the SDR. Sample topology files are provided with PSSP in the /etc/SP directory. These samples correspond to most of the topologies used by customers. If none of the samples match your real switch topology, you have to create one using the partitioning tool provided with PSSP (system partitioning aid available from the Perspectives Launch Pad). Once this file is created, it must be *annotated* and stored in the SDR (here, annotated means that the generic topology contained in the sample file is customized to reflect information about the real switch connections using the information stored in the SDR).

This task is performed using the **Eannotator** and **Etopology** commands on the CWS or by using the equivalent SMIT panels, for annotating (`smitty annotator`) and for Storing a topology file (`smitty etopology_store`). For example, let us consider that we have one NSB and no ISB and that we use the **Eannotator** command as shown below to annotate the switch topology before storing it in SDR:

```
Eannotator -F /etc/SP/expected.top.1nsb.0isb.0 -f
/etc/SP/expected.top/annotated -O no
```

The `-O no` option does not store the topology file into the SDR.

The following **Etopology** command is used for storing the switch topology file in the SDR.

```
Etopology /etc/SP/expected.top/annotated
```

Verify the switch primary and primary backup nodes

After choosing the switch topology, you can change the default primary and primary backup nodes using the **Eprimary** command or the SMIT Set Primary/Primary Backup Node window (`smitty primary_node_dialog`).

Note: 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. For full details on how the primary and backup nodes are selected and how failover is handled, see chapter 14 in the section "Selecting the Primary and Primary Backup nodes" of *PSSP Administration Guide,* SA22-7348. Also refer to chapter 4 section 4.3, "Eprimary modifications" in *IBM (e)server Cluster 1600 Managed by PSSP 3.5: What's New,* SG24-6617.

Run the **Eprimary** command to use node 5 instead of the default value, node1, as the primary node, as follows:

```
Eprimary -init 5
```

Set the clock source for all switches

Note: SP Switch2 does not require you to select switch clocks. This section applies only to SP Switch.

The last step in the configuration of the switch is to choose a clock source for all switches and to store this information in the SDR. This is done using the **Eclock** command on the CWS. Sample clock topology files are provided in the SDR. You can choose to use one of them or let the system decide for you. For example, if your SP system has no node switch boards and no intermediate switch boards, select /etc/SP/Eclock.top.6nsb.4isb.0 as an Eclock topology file. Enter:

```
Eclock -f /etc/SP/Eclock.top.0nsb.0isb.0
```

To verify the switch configuration information, enter:

```
splstdata -s
```

Setting up a system partition

This step is optional. The PSSP installation code sets up a default system partition configuration to produce an initial, single-system partition including all nodes in the system. This system partition is created automatically. You can proceed to the next section, if you do not want to partition your system.

If you want to partition the system, you can select an alternate configuration from a predefined set of system partitions to implement before booting the nodes or you can use the System Partitioning Aid to generate and save a new layout. Follow the procedure described in chapter 16, "Managing system partitions," of

PSSP Administration Guide, SA22-7348, and also refer to information in the “The System Partitioning Aid” section of the “Planning SP system partitions” chapter in *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281.

9.2.19 Configuring the CWS as boot/install server

The SDR now contains all required information to create a boot/install server on the CWS. The **setup_server** command configures the machine where it is executed (CWS or SP node) as a boot/install server. This command has no argument. It executes on the CWS and any additional boot/install servers. It is equivalent to clicking **Run setup_server Command** in the SMIT Enter Database Information window (`smitty enter_data`).

At this point, only the CWS will be configured since the other nodes are still not running. On the CWS, this command could have been executed automatically if you had specified the **-s yes** option when running **spbootins** as shown in 9.2.17, “spbootins” on page 325. Since we did not use this option previously in our environment, we have to execute **setup_server**.

On an additional boot/install server node, **setup_server** is automatically executed immediately after installation of the node if it has been defined as a server during the SDR configuration. Since this step can take a long time to complete, we recommend that after the server node installation, you check the /var/adm/SPlogs/sysman/<node>.console.log file. It contains information about the progress of the **setup_server** operation. This operation must be successfully completed before you try to install any client node from the server.

The command **setup_server** is a complex Perl program. It executes a series of configuration commands, called wrappers, that perform various tasks, such as configuring PSSP or setting the NIM environment as shown in Figure 9-1 on page 329. This is a simplified control flow sequence of the **setup_server** command.

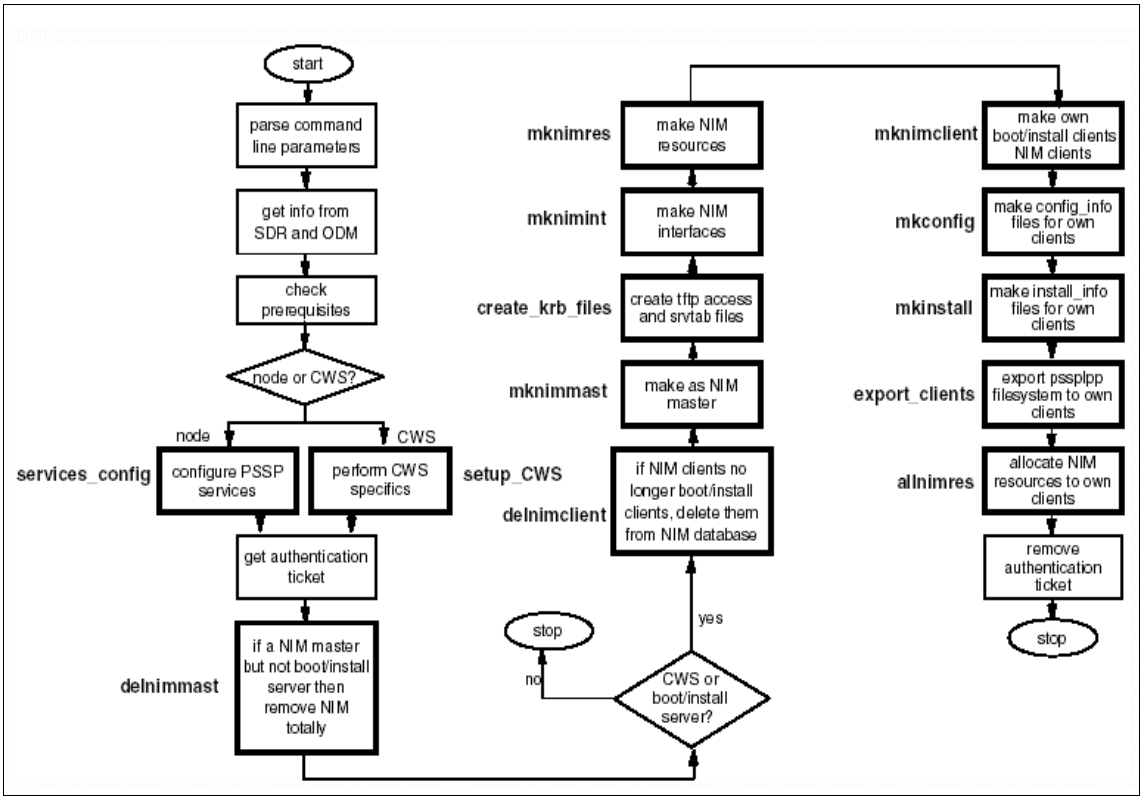


Figure 9-1 setup_server flow diagram

We now describe the step-by-step process of how the **setup_server** command calls a number of other commands as shown in Figure 9-1:

1. Get information from SDR and ODM.
2. Check prerequisites.
3. Configure PSSP services on this node: **services_config**.
4. If running on CWS, then perform CWS-specific tasks: **setup_CWS**.
5. Get an authentication ticket: **k4init**.
6. If running on a NIM master, but not a boot/install server, then unconfigure NIM and uninstall NIM filesets: **delnimmast -l <node_number>**.
7. If not running on the CWS or boot/install server, then exit.
8. If any NIM clients are no longer boot/install clients, then delete them from the NIM configuration database: **delnimclient -s <server_node_numb>**.
9. Make this node a NIM master: **mknimmast -l <node_number>**.

10. Create tftp access and srvtab files on this master: **create_krb_files**.
11. Make NIM interfaces for this node: **mknimint -l <node_number>**.
12. Make the necessary NIM resources on this master:

    mknimres -l <node_number>.
13. Make NIM clients of all of this node's boot/install clients:

    mknimclient -l <client_node_list>.
14. Make the config_info files for this master's clients: **mkconfig**.
15. Make the install_info files for this master's clients: **mkinstall**.
16. Export pssplpp file system to all clients of this server: **export_clients**.
17. Allocate the necessary NIM resources to each client:

    allnimres -l <client_node_list>.
18. Remove the authentication ticket.

9.2.20 Verify that the System Management tools were correctly installed

The following command directs you to run a verification test that checks for correct installation of the System Management tools on the CWS:

```
SYSMAN_test
```

After the tests are run, the system creates a log in /var/adm/SPlogs called SYSMAN_test.log.

Note: After performing this step, you can ignore any messages that you receive about the number of nodes tested. Since nodes are not available during this operation, they will not be tested.

For complete details of what the **SYSMAN_test** command checks on the CWS, the boot/install servers, the Cluster 1600 nodes, and also for diagnostic procedures if the system management verification fails, refer to chapter 29 of *PSSP Diagnosis Guide,* GA22-7350.

9.2.21 Network boot the boot/install server and nodes

After configuring the switch, and verifying that the system management tools were installed correctly, we are finally ready to install the Cluster nodes. This operation is two-fold. In the first stage, all additional boot/install servers are

installed through the Ethernet network from the CWS. In the second stage, all remaining nodes are installed from their boot/install servers.

Note: If you have more than eight boot/install servers on a single Ethernet segment, you should network boot those nodes in groups of eight or less.

There are two ways of network booting nodes.

1. Network Boot

   Table 9-2 shows how to install the nodes and monitor the installation progress through Perspectives as well as with the command line interface.

Table 9-2 Boot/install servers installation and monitoring the progress

If using	Do this
Perspectives	**SELECT** The Hardware Perspective icon by double-clicking **SELECT** The Nodes pane **SELECT** Actions -> LCD and LED display The LCD and LED display appears. **SELECT** Nodes to be netbooted **SELECT** Actions -> Network Boot **SELECT** Apply All selected nodes are booted. If you had the Hardware Perspective up before you added the required node information, you should delete and re-add the Nodes pane. If you had the Hardware Perspective up before you partitioned your system, you should delete and re-add the CWS/System/Syspars pane and then delete and re-add the Nodes pane.
nodecond	Enter: **nodecond frame_id slot_id &** Enter: **spmon -Led nodenode_number** or **spled &** to check the LCD and LED display for each node.

The **nodecond** command executes on the CWS In parallel with the **spmon** or **spled** command as shown in Table 9-2 on page 331. You can now initiate the boot and system installation on the target node. This phase is called *node*

conditioning, and it is done with the **nodecond** command. It is executed on the CW S for all nodes even if their boot/install server is *not* the CWS.

Once started, this command does not require any user input. It can, therefore, be started as a shell background process. If you have several nodes to install from the control workstation, you can start one **nodecond** command for each of them from the same shell. However, for performance reasons, it is not recommended to simultaneously install more than eight nodes from the same boot/install server.

After all boot/install servers have been installed, you can condition the remaining nodes. It is also possible to perform the installation using the Perspectives graphical user interface as shown in Table 9-2 on page 331. In the Hardware Perspective window, select the node you want to install, and then you need only click **Network Boot...** on the Action menu.

2. Manual (hand-conditioning)

   In some cases, you may want to perform the node installation manually using **nodecond** rather than automatically. Manual node conditioning is a more complex task, consisting of several steps. Since it greatly depends on the hardware type of the node, these steps differ for each category of nodes.

   For an SMP (thin and wide) node, perform the following steps:

   a. Power the node off by invoking **spmon**.
   b. From the SP Perspectives Launch Pad, select **Hardware Perspectives**.
   c. Click the processor nodes you are going to network boot.
   d. On the System Management Services menu:
      - Enter 2 to select multiboot.
      - Enter 4 to select boot devices.
      - Enter 3 to select the 1st boot device.
      - Select the Ethernet adapter to use for boot.

After the network installation of all of your boot/install servers is completed, run the following command on the CWS to refresh the system partition-sensitive subsystems:

```
/usr/lpp/ssp/bin/syspar_ctrl -r
```

Refer to 9.2.20, “Verify that the System Management tools were correctly installed” on page 330.

Network boot the remaining nodes with a procedure similar to Table 9-2 on page 331. After the network installation of all of your remaining nodes is complete, run the following command on the CWS to refresh the system partition-sensitive subsystems:

```
/usr/lpp/ssp/bin/syspar_ctrl -r
```

9.2.22 Verify node installation

If you have set up system partitions, select a global view to do this step. Run the spmon command to verify the node installation as follows:

```
spmon -d -G
```

Check hostResponds and powerLED indicators for each node. The detailed output of this command can be found in Example 10-4 on page 356.

9.2.23 Enable s1_tty on the SP-attached server (SAMI protocol only)

If you just installed an SP-attached server, you must ensure that the s1_tty is enabled on the server. On the SP-attached server, determine which tty is mapped to 01-S1-00-00 and enable the login with the **chdev** command as follows:

```
chdev -l tty0 -a login=enable
```

9.2.24 Update authorization files in restricted mode for boot/install servers (optional)

Once the node is installed, or as part of firstboot.cust, the remote command authorization files on the node serviced by the non-CWS boot/install server need to be updated depending on the setting of auth_root_rcmd.

9.2.25 Run verification tests on all nodes

Refer to 9.2.20, “Verify that the System Management tools were correctly installed” on page 330.

9.2.26 Check the system

At this point, we recommend that you check the Cluster system using the **SYSMAN_test** command.

9.2.27 Start the switch

Once all nodes have been installed and booted, you can start the switch. This is performed using the **Estart** command on the CWS or clicking **Start Switch** in the SMIT Perform Switch Operation Menu. To start the switch, do the following:

1. The fault-handling daemon for the SP Switch (fault_service_Worm_RTG_SP) or for the SP Switch2 (fault_service_Worm_CS), which checks, initializes, and prepares the switch for operation, must be running on the primary node. If the daemon is not running, use the **rc.switch** command to start it.
2. On the CWS, run the **Estart** command. The output of this command may appear as shown in Example 9-11.

Example 9-11 Sample output of the Estart command

```
Switch initialization started on tserv10.hpssl.kgn.ibm.com
Initialization successful for 16 nodes
Switch initialization completed
```

9.2.28 Verify that the switch was installed correctly

The following command is used to do a verification test to ensure that the switch is installed completely:

```
CSS_test
```

After the tests are run, the system creates a log in /var/adm/SPlogs called CSS_test.log. If the verification test fails, see the section on “Diagnosing switch problems” in *PSSP Diagnosis Guide,* GA22-7350.

Finally you need to check the switchResponds and powerLED indicators for each node by executing the following command:

```
spmon -d -G
```

Refer to the 10.3.5, “Monitoring hardware status” on page 355 for a sample output of the **spmon** command.

9.3 Key files

As for the commands presented previously, this section only presents the major system files used by PSSP.

9.3.1 /etc/bootptab.info

The bootptab.info file specifies the hardware (MAC) address of the en0 adapter of cluster nodes. It is used to speed up the execution of the **sphrdwrad** command. Each line contains the information for one node and is made of two parts: The node identifier and the MAC address.

The node identifier can be either the node number or a pair <frame_number>,<slot> separated by a comma with no blanks.

The MAC address is separated from the node identifier by a blank. It is formatted in hexadecimal with no period (.) or colon (:). The leading 0 of each part of the MAC address must be present.

In our environment, the /etc/bootptab.info file could be the example shown in Example 9-12.

Example 9-12 Example of /etc/bootptab.info

```
> cat /etc/bootptab.info
1 02608CF534CC
5 10005AFA13AF
1,610005AFA1B12
1,7 10005AFA13D1
8 10005AFA0447
9 10005AFA158A
10 10005AFA159D
11 10005AFA147C
1,12 10005AFA0AB5
1,13 10005AFA1A92
1,14 10005AFA0333
1,15 02608C2E7785
>
```

`1,14 10:0:5A:FA:03:33` is not a valid entry even if the second string is a usual format for MAC addresses.

9.3.2 /tftpboot

The /tftpboot directory exists on the CWS, the boot/install server, and on the Cluster 1600 nodes.

On the CWS and other boot/install servers, this directory is used as a repository for files that will be distributed to the client nodes during their installation. On the client nodes, the directory is used as a temporary storage area where files are downloaded from the boot/install server /tftpboot directory.

The customization of the boot/install server (with **`setup_server`**) creates several files in /tftpboot:

- <spot_name>.<archi>.<kernel_type>.<network>
- <hostname>-new-srvtab
- <hostname>.config_info

- <hostname>.install_info

You can also manually add customization scripts to the /tftpboot directory:

- tuning.cust
- script.cust
- firstboot.cust

In our environment, the /tftpboot directory of the CWS contains the files listed in Figure 9-2.

```
[sp3en0:/]# ls -al /tftpboot
total 7722
drwxrwxr-x 3 root     system       512 Dec 15 15:33 .
drwxr-xr-x 22 bin      bin         1024 Dec 15 14:58 ..
-rw-r--r-- 1 bin      bin        11389 Dec 03 12:03 firstboot.cust
drwxrwx---   2 root     system       512 Nov 12 16:09 lost+found
-r-------- 1 nobody   system       118 Dec 12 13:15 sp3n01-new-srvtab
-rw-r--r--   1 root     system       254 Dec 12 13:15
sp3n01.msc.itso.ibm.com.config_info
-rw-r--r--   1 root     system       795 Dec 12 13:15
```

Figure 9-2 Contents of the CWS /tftpboot directory

We will now describe in more detail the role of each of these files.

<spot_name>.<archi>.<kernel_type>.<network>

Files with this format of name are bootable images. The naming convention is:

- <spot_name>

  This the name of the spot from which this bootable image has been created. It is identical to the name of a spot subdirectory located under /sdpata/sys1/install/<aix_level>/spot. In our environment, the spot name is spot_aix432.

- <archi>

  This specifies the machine architecture that can load this bootable image. It is one of **`rs6k`**, **`rspc`**, or **`chrp`**.

- <kernel_type>

  This refers to the number of processors of the machine that can run this image. It is either `up` for a uniprocessor or `mp` for a multiprocessor.

- <network>

This depends on the type of network adapter through which the client machine will boot on this image. It can be ent, tok, fddi, or generic.

These files are created by the **setup_server** command. Only the images corresponding to the spot_name, architecture, and kernel type of the nodes defined to boot from the boot/install server will be generated, not all possible combinations of these options.

For each node, the tftpboot directory contains a symbolic link to the appropriate bootable image. You can see an example of this in Figure 9-2 on page 336 where this file is called spot_aix432.rs6k.mp.ent.

<hostname>-new-srvtab

These files are created by the **create_krb_files** wrapper of setup_server. <hostname> is the reliable host name of a node. For each client node of a boot/install server, one such file is created in the server /tftpboot directory.

This file contains the passwords for the rcmd principals of the SP node. Each node retrieves its <hostname>-new-srvtab file from the server and stores it in its /etc directory as krb-srvtab.

<hostname>.install_info

These files are created by the **mkinstall** wrapper of setup_server. <hostname> is the reliable host name of an SP node. For each client node of a boot/install server, one such file is created in the server /tftpboot directory.

This file is a shell script containing mainly shell variables describing the node en0 IP address, host name, boot/install server IP address, and hostname.

After the node AIX image has been installed through the network, the `pssp_script` script downloads the <hostname>.install_info file into its own /tftpboot directory, and it executes this shell to define the environment variable it needs to continue the node customization.

This file is also used by other customization scripts, such as psspfb_script.

<hostname>.config_info

These files are created by the **mkconfig** wrapper of setup_server. <hostname> is the reliable host name of an SP node. For each client node of a boot/install server, one such file is created in the server /tftpboot directory.

This file contains node configuration information, such as node number, switch node information, default route, initial hostname, and CWS IP information.

After the pssp_script script has executed the <hostname>.install_info scripts, it downloads the <hostname>.config_info file into the node /tftpboot directory and configures the node using the information in this file.

tuning.cust

The tuning.cust file is a shell script that sets tuning options for IP communications. A default sample file is provided with PSSP in /usr/lpp/ssp/samples/tuning.cust. Three files are also provided that contain recommended settings for scientific, commercial, or development environments (in /usr/lpp/ssp/install/config).

Before starting the installation of the nodes, you can copy one of the three pre-customized files into the /tftpboot directory of the CWS, or you can provide your own tuning file. Otherwise, the default sample will be copied to /tftpboot by the installation scripts.

During the installation of additional boot/install servers, the tuning.cust file will be copied from the CWS /tftpboot directory to each server /tftpboot directory.

During the installation of each node, the file will be downloaded to the node. It is called by the /etc/rc.net file; so, it will be executed each time a node reboots.

You should note that tuning.cust sets ipforwarding=1. So, you may want to change this value for nodes that are not IP gateways directly in the /tftpboot/tuning.cust on the node (not on boot/install servers).

script.cust

The script.cust file is a shell script that will be executed at the end of the node installation and customization process before the node is rebooted. The use of this file is optional. It is a user provided customization file. You can use it to perform additional customization that requires a node reboot to be taken into account.

Typically, this script is used to set the time zone, modifying paging space, and so on. It can also be used to update global variables in the /etc/environment file.

A sample script.cust file is provided with PSSP in /usr/lpp/ssp/samples. If you want to use this optional script, you must first create it in the /tftpboot directory of a boot/instal server by either providing your own script or copying and modifying the sample script. During node installation, the file is copied from the boot/install server onto the node.

You can either create one such file in the /tftpboot of the CWS, in which case it will be used on all nodes of the SP system, or you can create a different one in

the /tftpboot of each boot/install server to have a different behavior for each set of node clients to each server.

firstboot.cust

The firstboot.cust file is a shell script that will be executed at the end of the node installation and customization process after the node is rebooted. The use of this file is optional. It is a user provided customization file. This is the recommended place to add most of your customization.

This file should be used for importing a volume group, defining a host name resolution method used on a node, or installing additional software.

It is installed on the nodes in the same way as script.cust: It must be created in the /tftpboot directory of a boot/install server and is automatically propagated to all nodes by the node installation process.

Note: At the end of the execution of the firstboot.script, the host name resolution method (/etc/hosts, NIS, DNS) *must* be defined and able to resolve all IP addresses of the SP system: CWS, nodes, the Kerberos server, and the NTP server. If it is not, the reboot process will not complete correctly. If you do not define this method sooner, either by including configured file in the mksysb image or by performing the customization in the script.cust file, you must perform this task in the firstboot.cust file.

9.3.3 /usr/sys/inst.images

This directory is the standard location for storing an installable LPP image on an AIX system when you want to install the lpp from disk rather than from the distribution media (tape, CD). You can, for example, use it if you want to install on the CWS another product than AIX and PSSP.

This directory is *not* used by the PSSP installation scripts.

9.3.4 /spdata/sys1/install/images

The /spdata/sys1/install/images directory is the repository for all AIX installable images (mksysb) that will be restored on the SP nodes using the PSSP installation scripts and the NIM boot/install servers configured during the CWS installation.

This directory must exist on the CWS, and its name must be kept unchanged. The SP nodes installation process will not work if the AIX mksysb images are stored in another directory of the CWS.

If you want to use the default image provided with PSSP (spimg), you must store it in the /spdata/sys1/install/images directory.

If all nodes have an identical software configuration (same level of AIX and LPPs), they can share the same mksysb image independently from their hardware configuration.

If your SP system has several boot/install servers, the installation script will automatically create the /spdata/sys1/install/images directory on the boot/install servers and load it with the mksysb images needed by the nodes that will boot from each of these servers.

9.3.5 /spdata/sys1/install/<name>/lppsource

For each level of AIX that will be running on a node in the SP system, there must exist on the CWS an /spdata/sys1/install/<name>/lppsource/installp/ppc directory for the filesets in installp format. For the filesets in the RPM format, the directory will be /spdata/sys1/install/<name>/lppsource/rpms/ppc. The recommended rule is to set the relative pathname <name> to a name significantly indicating the level of AIX: aix433, aix51, aix52. However, this is not required, and you may choose whatever name you wish.

This directory must contain the AIX LPP images corresponding to the AIX level. In addition, this directory must contain the perfagent code corresponding to the AIX level, the runtimes, the RSCT filesets and the HMC prerequisite filesets. Starting with AIX release 4.3.2, perfagent.tools is part of AIX and not Performance Aide for AIX (PAIDE), as it used to be in previous AIX releases.

If the Cluster 1600 system contains several boot/install servers, this directory will only exist on the CWS. It will be known as a NIM resource by all servers but will be defined as hosted by the CWS. When a node needs to use this directory, it mounts it directly from the CWS, irrespective of which NIM master it is pointing at.

9.3.6 /spdata/sys1/install/pssplpp/PSSP-x.x

For each level of PSSP that will be used by either the CWS or a node in the SP system, there must exist on the CWS a /spdata/sys1/install/pssplpp/PSSP-x.x directory where PSSP-x.x is one of PSSP-3.2, PSSP-3.4 or PSSP-3.5.

During the first step of the PSSP software installation on the CWS (refer to 8.6.1, “Copy of the PSSP images” on page 291), the PSSP source images must be installed using **`bffcreate`** into these directories.

If the cluster system contains more than one boot/install server, the installation scripts will create the /spdata/sys1/install/pssplpp/PSSP-x.x directories on each server and load them with the PSSP LPP filesets.

9.3.7 /spdata/sys1/install/pssp

You can create this directory manually on the CWS in the first steps of the PSSP installation. The CWS installation script will then store in this directory several files that will be used later during the node installation through the network.

/spdata/sys1/install/pssp is also automatically created on the additional boot/install servers and populated with the following files:

- pssp_script

  pssp_script is executed on each node by NIM after the installation of the mksysb on the node and before NIM reboots the node. It is run under a single user environment with the RAM file system in place. It installs required LPPs (such as PSSP) on the node and does post-PSSP installation setup. Additional adapter configuration is performed after the node reboot by psspfb_script.

  You should not modify this script. User customization of the node should be performed by other scripts: tuning.cust, script.cust, or firstboot.cust (refer to 9.3.2, “/tftpboot” on page 335).

- bosinst_data

  The bosinst_data, bosinst_data_prompt, and bosinst_data_noprompt are NIM control files created by the installation of PSSP on the CWS. They are used during NIM installation of each Cluster node. They contain configuration information, such as the device that will be used as the console during node installation, locale information, and the name of the disk where to install the system. For further information, refer to *AIX 5L V 5.2 Installation Guide and Reference,* SC23-4389.

9.3.8 image.data

In a mksysb system image, the image.data file is used to describe the rootvg volume group. In particular, it contains the size of the physical partition (PPSIZE) of the disk from which the mksysb was created. You usually do not need to modify this file. However, if the mksysb is to be restored on a node where the PPSIZE is different from the PPSIZE defined in the image.data file, you may need to manually create a NIM image data resource and allocate it to the node that needs to be installed.

9.4 Related documentation

For complete reference and ordering information for the documents listed in this section, see “Related publications” on page 553.

Pssp:3.5 manuals

PSSP Administration Guide, SA22-7348 chapters 9, 10, 14, and 16 provide detailed information about the services that may be configured in the SP system: Time Server, Automounter, Security, Switch and System partitions.These chapters are useful for filling in the site environment information details as shown in Example 9-1 on page 304. Chapter 14 provides information on setting up the switch.

PSSP Installation and Migration Guide, GA22-7347, for PSSP 2.5. In Chapter 2, steps 32 to 85 detail the complete installation process. Appendixes B and E describe the SP daemons and the SDR structure.

PSSP Command and Technical Reference (2 Volumes), SA22-7351 contains a complete description of each command listed in 9.2, “Installation steps and associated key commands” on page 302.

PSSP Diagnosis Guide, GA22-7350 provides details for solving problems related to NIM, node installation, SDR, SP2® Switch, HMC-controlled server problems, remote commands, and Perspectives.

SP redbooks

RS/6000 SP: PSSP 3 Survival Guide, SG24-5344. Chapter 2 contains practical tips and hints about specific aspects of the installation process.

IBM (e)server Cluster 1600 Managed by PSSP 3.5: What’s New, SG24-6617. Chapter 4 contains PSSP 3.5 enhancements. Chapter 3 gives an overview of the new RSCT components. Chapter 6 gives you a feel of how the HMC-controlled servers are viewed with some screen shots.

Others

Hardware Management Console for pSeries Installation and Operations Guide, SA38-0590 for configuration of the HMC controlled servers.

AIX 5L V 5.2 Installation Guide and Reference, SC23-4389 has details on AIX configuration files and NIM.

9.5 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, “Answers to sample questions” on page 521.

1. In a cluster system, which are true statements regarding a node's initial hostname and reliable hostname as defined in the SDR? (Note: Two options are correct.)
   a. The initial hostname is the standard TCP/IP hostname associated with one of the available TCP/IP interfaces on the node.
   b. The initial hostname refers to an SP node or CWS's hostname prior to PSSP installation on the node or CWS.
   c. The reliable hostname is the TCP/IP interface name associated with the node's en0 interface.
   d. The reliable hostname is the TCP/IP interface name associated with the interface on an SP node that responds the most quickly to heartbeat packets over a period of time.
2. How do you configure node 1 to be a boot/install server?
   a. Run the **setup_server** script on node 1.
   b. Install NIM, and then run **setup_server** on node 1.
   c. Change the boot/install server field in the SDR for some nodes and then run **setup_server**.
   d. Change the boot/install server field in the SDR for some nodes, and then run the **spbootins** command to set those nodes to install.
3. Which command is used to configure additional adapters for nodes?
   a. **sphrdwrad**
   b. **spethernt**
   c. **spadapters**
   d. **spadaptrs**
4. Which of the following daemons does the syspar_crtl -A command *not* start?
   a. hats
   b. spconfigd
   c. kerberos
   d. pman
5. Which of the following statements is true about the /tftpboot directory?

a. The directory only exists on the boot/install server and on the SP client nodes.

b. You cannot add customization scripts to the directory.

c. The customization of the boot/install server creates several files in the /tftpboot directory.

d. On the client nodes, the directory is used as a permanent storage area.

6. Which of the following commands is used to verify configuration information stored in the SDR about the frames and the nodes?

   a. **`splstconfig`**

   b. **`splstdata`**

   c. **`spframeconfig`**

   d. **`spnodeconfig`**

7. Which of the following commands configures the machine where it is executed, as a boot/install server?

   a. **`setup_cws`**

   b. **`setup_nodes`**

   c. **`server_setup`**

   d. **`setup_server`**

8. Which of the following two commands customize the switch topology file and store it in the SDR?

   a. **`Eannotator`**

   b. **`Estart`**

   c. **`Eprimary`**

   d. **`Etopology`**

9. Which of the methods listed below is *not* the format for selecting an installation disk for a node?

   a. The hardware location format

   b. The device names format

   c. The PVID format

   d. The Volume group format.

9.6 Exercises

Here are some exercises you may wish to do.

1. On a test system that does not affect any users, check if the supervisor microcode is up to date or needs to be upgraded. Which command will upgrade the supervisor microcode? What does the -u flag do to the target node?
2. What is the role of the hmcd daemon?
3. Familiarize yourself with the steps involved in the installation of the HMC-controlled server.
4. What is the role of the /tftpboot directory? Which customization scripts can be manually added to the /tftpboot directory?
5. Familiarize yourself with the following key files: /etc/bootptab.info, tuning.cust, script.cust, and firstboot.cust.
6. Familiarize yourself with the **splstdata, spmon** and **spchvgobj** commands with all the major options.

10

Verification commands and methods

This chapter presents some of the commands and methods available to the Cluster 1600 administrator to check whether the system has been correctly configured, initialized, and started.

10.1 Key concepts

Before taking the SP certification exam, you should understand the following concepts related to verifying and checking a Cluster 1600 system:

- The **splstdata** command
- The various components of a Cluster 1600 system and the different verification methods that apply to each of them.
- PSSP daemons, system partition-sensitive daemons, as well as Switch daemons that must be running, and how to check if they are alive.

10.2 Introduction to Cluster 1600 system checking

Several options are available to the Cluster 1600 user or administrator who wishes to verify that the system has been successfully installed and is running correctly:

- Commands and SMIT menus
- Graphical interfaces
- Logs

Section 10.3, “Key commands” on page 348 presents the commands that are available for checking various aspects of a Cluster 1600 system. Section 10.4, “Graphical user interface” on page 361 gives a few hints about the use of Perspectives. Section 10.5, “Key daemons” on page 362 focuses on the daemons that are important to monitor a Cluster 1600 system. Section 10.6, “SP-specific logs” on page 365 lists the logs that are available to the user to check the execution of commands and daemons.

10.3 Key commands

PSSP comes with several commands for checking the system. But some AIX commands are also useful to the Cluster 1600 user. We present here the AIX and PSSP commands most widely used for this purpose.

10.3.1 Verify installation of software

During the CWS and the Cluster 1600 system installation, your first verification task consists of checking that the AIX and PSSP software was successfully installed, and that the basic components are configured correctly before you start entering configuration data specific to your environment.

Checking software levels: lslpp

The **lslpp** command is the standard AIX command to check whether an LPP has been installed and to verify its level. You should use it on the CWS after installation of AIX and after you have installed (**installp**) the PSSP software from the /spdata/sys1/install/pssplpp/PSSP-x.x directory. At this point, you should check the consistency between the level of AIX, perfagent, and PSSP:

```
lslpp -La bos* devices* perf* X11* xlC* ssp* rsct* | more
```

You should also verify that you have installed all PSSP filesets corresponding to your cluster hardware configuration and to the options you wish to use (HMC, VSD, RVSD, and so on).

Checking the SDR initialization: SDR_test

Immediately after initialization of the SDR (**install_cw**), you should test that the SDR is functioning properly using the **SDR_test** command as shown in 8.8, “Configuring and verifying the CWS” on page 296. This command can also be used later, during operation of the clustered system, if you suspect problems with the SDR. A return code of zero indicates that the test completed as expected; otherwise it returns the numbers of the errors. If you do not specify the **-l** flag, error messages are recorded in /var/adm/SPlogs/SDR_test.log for root users. For non-root users, error messages are recorded in /tmp/SDR_test.log. The SDR_test file is located in /usr/lpp/ssp/bin.

Checking the System Monitor installation: spmon_itest

The **install_cw** command also installs the System Monitor (spmon) on the CWS. At the same time that you test the SDR initialization, you can also test whether the spmon is correctly installed by using the **spmon_itest** command as shown in 8.8, “Configuring and verifying the CWS” on page 296. A return code of zero indicates that the test completed as expected; otherwise it returns the numbers of the errors. If errors are detected, more detailed information is recorded in the log file. If you do not specify the **-l** flag, error messages are recorded in /var/adm/SPlogs/spmon_itest.log. The spmon_itest file is located in /usr/lpp/ssp/bin.

Checking the System Monitor configuration: spmon_ctest

After the Cluster 1600 hardware has been discovered by the CWS (**spframe**), you can check whether the system monitor and the Perspectives have been correctly configured with the information about the cluster frames and nodes hardware with the **spmon_ctest** command. This command also checks whether the hardmon daemon is running, the serial RS232 links to the frames and nodes are properly connected, the HMC is configured, the CWS can access the frames and nodes hardware through these connections, and the hardware information has been stored in the SDR. A return code of zero indicates that the test completed

as expected; otherwise it returns the number of errors. The error messages are recorded in /var/adm/SPlogs/spmon_ctest.log. The **spmon_ctest** file is located in /usr/lpp/ssp/bin.

Checking LPP installation on all nodes: lppdiff

After complete installation of an SP system, or any time during the life of the SP system, you may need to check the level of software installed on all, or a subset, of nodes. The **lppdiff** command is an easier alternative to the use of **dsh lslpp** since it sorts and formats the output by filesets. It can be used to list any filesets and is not limited to PSSP. The **lppdiff** command is located in /usr/lpp/ssp/bin.

For example, to check all PSSP related filesets, you can use:

```
lppdiff -Ga ssp* rsct*
```

Checking PSSP level: splst_versions

If you only need to look for the PSSP versions installed on the nodes, and not for all the detailed information returned by **lppdiff**, you can use the **splst_versions** command. For example, in our environment, we can get this information for each node; see Figure 10-1. The **splst_version** command is located in /usr/lpp/ssp/bin.

```
[sp3en0:/usr/lpp/ssp]#  splst_versions -tG
1 PSSP-3.4
5 PSSP-3.5
6 PSSP-3.5
7 PSSP-3.4
8 PSSP-3.4
9 PSSP-3.5
10 PSSP-3.5
11 PSSP-3.5
12 PSSP-3.5
13 PSSP-3.5
14 PSSP-3.5
15 PSSP-3.5
[sp3en0:/usr/lpp/ssp]#
```

Figure 10-1 PSSP versions installed on each node

Checking Sysman components: SYSMAN_test

The **SYSMAN_test** command is a very powerful test tool. It checks a large number of cluster system management components. We present it in this section since it is recommended to execute this command after installation of the CWS and before the installation of the node as described in Chapter 9, “Frame and node

installation" on page 301. However, its use is not limited to the installation phase of the life of your Cluster 1600 system; it can provide valuable information during normal operation.

The **SYSMAN_test** command is executed on the CWS, but it does not restrict its checking to components of the CWS. If nodes are up and running, it can also perform several tests on them as described in Chapter 9, "Frame and node installation" on page 301. Subsets of the components checked by **SYSMAN_test** are: ntp, automounter, file collection, user management, nfs daemons, /.klogin file, and so on.

The output of **SYSMAN_test,** using the **-v** (verbose) option, is generally large. We therefore recommend that you redirect the output to a file to prevent flooding the screen with messages that display too fast and then use a file browser or editor to look at the results of the command. An alternative is to look at file /var/adm/SPlogs/SYSMAN_test.log, but this file does not contain all the information provided by the verbose option.

The **SYSMAN_test** command is located in /usr/lpp/ssp/bin. Following is an example of successful completion of the **SYSMAN_test** on the CWS as well as all the nodes. This command was run when all the configured nodes were on.

```
[c179n01][/var/adm/SPlogs]> SYSMAN_test
SYSMAN_test: Verification succeeded.
[c179n01][/var/adm/SPlogs]>
```

Note: PSSP 3.4 and later provides the ability to run commands like **SYSMAN_test** using secure remote command and secure remote copy methods. Refer to 6.2.2, "Using the secure remote command process" on page 220 for setting the right environment for the secure command method.

ssp.css: Switch code CSS_test

The last command we present to check system installation is **CSS_test**. There is no point to use it on a switchless system.

The **CSS_test** command can be used to check that the ssp.css LPP has been correctly installed. In particular, **CSS_test** checks for inconsistencies between the software levels of ssp.basic and ssp.css. This is why we present this command in this section. However, it is also useful to run it on a completely installed and running system where the switch has been started, since it will also check that communication can be performed over the switch between the cluster nodes. PSSP 3.4 and later provides the ability to run commands using secure remote command and secure remote copy methods. Refer to 6.2.2, "Using the secure remote command process" on page 220 for setting the environment variables to run this command with the secure remote command method. The **CSS_test** command is located in /usr/lpp/ssp/bin.

10.3.2 Verify system partitions

Two commands are particularly useful for checking the cluster system partitions. Both are located in /usr/lpp/ssp/bin.

Listing existing partition: splst_syspars

The first of these commands, **splst_syspar**, only lists the existing partitions in the Cluster 1600 system. Using its only option, **-n**, you can obtain either the symbolic or the numeric value of the partition:

```
[sp3en0:/usr/lpp/ssp]# splst_syspars -n
sp3en0
[sp3en0:/usr/lpp/ssp]# splst_syspars
192.168.3.130
[sp3en0:/usr/lpp/ssp]#
```

Verifying system partitions: spverify_config

The **spverify_config** command is used to check the consistency of the information stored in the SDR regarding the partitions defined in the cluster system. The command is not valid on a system with SP Switch2 or on a switchless clustered server system. It is only to be used when the system has more partitions than the initial default partition.

10.3.3 Verifying the authentication services

Here we discuss two commands used to verify the PSSP authentication settings. They are located in the /bin directory.

lsauthpar

The **lsauthpar** command lists the remote command authentication methods that are active for the system partition. This command verifies the settings in the System Data Repository (SDR), reporting spurious settings; to make corrections, use the **chauthpar** command. This command also provides an option to check whether all nodes in the applicable system partition have the correct setting. No remote verification occurs unless the SDR setting is valid. Example 10-1 shows the output of the **lsauthpar** command.

Example 10-1 lsauthpar output

```
$lsauthpar
Kerberos 5
Kerberos 4
Standard Aix
```

lsauthpts

The **lsauthpts** command lists the trusted services authentication methods that are active for the system partition. This command always verifies the correctness of the setting in the System Data Repository, reporting spurious settings to the user, who can use the **chauthpts** command to make corrections. This command also provides an option to check whether all nodes in the applicable system partition have the correct setting. No remote verification occurs unless the SDR setting is valid. Example 10-2 shows the output if the trusted service was set to *compat*.

Example 10-2 lsauthpts command output

```
$lsauthpts
Compatibility
```

10.3.4 Checking subsystems

These are some useful commands for checking the different PSSP subsystems.

Checking subsystems: lssrc

The **lssrc** command is not part of PSSP. It is a standard AIX command, part of the System Resource Controller feature of AIX. It is used to get the status of a subsystem, a group of subsystems, or a subserver.

In an clustered environment, it is especially used to obtain information about the status of the system partition-sensitive subsystems. To check whether these subsystems are running on the CWS, you can use the **lssrc** command with the **-a** option to get the status of all AIX subsystems, and then filter (**grep**) the result on the partition name. Figure 10-2 lists the results in our environment.

```
[sp3en0:/]# lssrc -a | grep sp3en0
 sdr.sp3en0       sdr              9032    active
 hats.sp3en0      hats             15144   active
 hags.sp3en0      hags             21984   active
 hagsglsm.sp3en0  hags             104768   active
 haem.sp3en0      haem             17620   active
 haemaixos.sp3en0 haem             105706   active
 hr.sp3en0        hr               37864   active
 pman.sp3en0      pman             102198   active
 pmanrm.sp3en0    pman             25078   active
 Emonitor.sp3en0  emon                     inoperative
[sp3en0:/]#
```

Figure 10-2 Listing status of system partition-sensitive subsystems on the CWS

You can also use the **lssrc** command on Cluster 1600 nodes or to get detailed information about a particular subsystem. Figure 10-3 shows a long listing of the status of the topology services subsystem on one of the cluster nodes.

```
 [sp3n06.msc.itso.ibm.com:/]# lssrc -l -s hats
Subsystem         Group            PID     Status
 hats             hats             7438    active
 Network Name   Indx Defd Mbrs St Adapter ID      Group ID
 SPether        [ 0]   13   13  S 192.168.31.16   192.168.31.115
 SPether        [ 0]              0x4666fc36      0x46744d3b
 HB Interval = 1 secs. Sensitivity = 4 missed beats
 SPswitch       [ 1]   12   12  S 192.168.13.6    192.168.13.15
 SPswitch       [ 1]              0x4667df4c      0x46682bc7
 HB Interval = 1 secs. Sensitivity = 4 missed beats
   2 locally connected Clients with PIDs:
 haemd(  9292) hagsd(  8222)
   Configuration Instance = 912694214
   Default: HB Interval = 1 secs. Sensitivity = 4 missed beats
   CWS = 192.168.3.130
[sp3n06.msc.itso.ibm.com:/]#
```

Figure 10-3 Listing topology services information on node sp3n06

syspar_ctrl -E

The **syspar_ctrl** command is the PSSP command providing control of the system partition-sensitive subsystems. In 9.2.11, “Start RSCT subsystems” on page 317, we saw that the **-A** option of this command adds and starts the subsystems.

The **syspar_ctrl -E** command displays (examines) all supported subsystems and reports on the lists of subsystems it can manage.

You can then use the other options of **syspar_ctrl** to stop, refresh, start, or delete subsystems that were reported as manageable by **syspar_ctrl -E**.

10.3.5 Monitoring hardware status

This monitoring of the Cluster 1600 nodes (except HMC) is done through the RS-232 line that connects the control workstation and each frame. From the CWS, the hardmon daemon uses a designated tty to connect to each frame supervisor card. For the HMC-controlled servers where there are no serial connections, the monitoring is done through the Ethernet connection with the help of the **hmcd** daemon.

Checking hardware connectivity: spmon_ctest

The **spmon_ctest** command runs on the CWS and performs many checks. We present it here since it tests hardware connectivity between the CWS and the cluster nodes. However, it also checks that the hardmon and SDR daemons are running, that it can communicate with all the cluster frames, and that the system monitor was correctly configured. Example 10-3 shows the successful execution of the command.

Example 10-3 spmon_ctest command output

```
[c179s][/]> spmon_ctest
spmon_ctest: Start spmon configuration verification test
spmon_ctest: Verification Succeeded
[c179s][/]>
```

We recommend that you use this command each time a new frame or node has been added to a Cluster 1600 system, after using the **spframe** command, and to check that the new nodes have been correctly discovered by PSSP and that they have been taken into account in the SDR. If you do not specify the **-l** flag, error messages are recorded in /var/adm/SPlogs/spmon_ctest.log.

Monitoring hardware activity: spmon -d

The **spmon** command is a monitoring and control command. Here we discuss only the verification and monitoring part. To execute the **spmon** command in monitor mode you must be authorized to access the hardware monitor subsystem and must be granted monitor permission for the hardware objects (frames, slots) specified in the command. It can retrieve and display information about the hardware component of the Cluster 1600 system as well as act on them. We present only a few options here.

The **spmon -d -G** command displays a summary of the hardware status of all components: Frames, nodes, and switches. It checks that the hardmon daemon is running and then reports on the power status, key setting, LEDs, hostresponds and switchresponds, and so on. Example 10-4 shows the result of this command on our CWS. Here Frame 3 is a 7026 H80 server. Frames 6, 10, and 11 are the HMC-controlled servers, with 8, 2, and 1 LPARs, respectively.

Note: LPARs are represented as thin nodes in the PSSP.

Example 10-4 spmon -d -G output

```
[c179s][/]> spmon -d -G
1.  Checking server process
    Process 23480 has accumulated 0 minutes and 0 seconds.
    Check successful

2.  Opening connection to server
    Connection opened
    Check successful

3.  Querying frame(s)
    7 frames
    Check successful

4.  Checking frames

      Controller Slot 17 Switch Switch   Power supplies
Frame Responds   Switch  Power  Clocking A     B     C     D
----- ---------- ------- ------ -------- ----- ----- ----- -----
   1      yes      yes     on      N_A     on    on    on   N/A
   2      yes      yes     on      N_A     on    on    on   N/A
   3      yes       no     N/A     N/A    N/A   N/A   N/A   N/A
   6      yes       no     N/A     N/A    N/A   N/A   N/A   N/A
  10      yes       no     N/A     N/A    N/A   N/A   N/A   N/A
  11      yes       no     N/A     N/A    N/A   N/A   N/A   N/A
  12      yes       no     N/A     N/A     on    on    on   N/A

5.  Checking nodes
------------------------------ Frame 1 ------------------------------
                         Host     Key     Env   Front Panel       LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED           Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1     1  high   on     yes     N/A     no  LCDs are blank       no
  5     5  high   on     yes     N/A     no  LCDs are blank       no
  9     9  high   on     yes     N/A     no  LCDs are blank       no
 13    13  high   on     yes     N/A     no  LCDs are blank       no

          Switch Responds (per plane)
```

```
Slot Node 0      1
---- ---- ------ ------
  1     1   yes    yes
  5     5   yes    yes
  9     9   yes    yes
 13    13   yes    yes

------------------------------ Frame 2 ------------------------------
                       Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1    17  high   on     yes     N/A     no  LCDs are blank      no
  5    21  high   on     yes     N/A     no  LCDs are blank      no
  9    25  high   on     yes     N/A     no  LCDs are blank      no
 13    29  high   on     yes     N/A     no  LCDs are blank      no

          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1    17   yes    yes
  5    21   yes    yes
  9    25   yes    yes
 13    29   yes    yes

------------------------------ Frame 3 ------------------------------
                       Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1    33 extrn   on     yes     N/A    N/A  LCDs are blank      no

          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1    33   no     yes

------------------------------ Frame 6 ------------------------------
                       Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1    81  thin   on     no      N/A    N/A  0a27               N/A
  2    82  thin   on     no      N/A    N/A  0a69               N/A
  3    83  thin   on     no      N/A    N/A  0a69               N/A
  4    84  thin   on     no      N/A    N/A  0a69               N/A
  5    85  thin  off     no      N/A    N/A  LCDs are blank     N/A
  6    86  thin  off     no      N/A    N/A  LCDs are blank     N/A
  7    87  thin  off     no      N/A    N/A  LCDs are blank     N/A
  8    88  thin  off     no      N/A    N/A  LCDs are blank     N/A

          Switch Responds (per plane)
```

```
Slot Node 0      1
---- ---- ------ ------
  1    81   no     no
  2    82   no     no
  3    83   no     no
  4    84   no     no
  5    85   no     no
  6    86   no     no
  7    87   no     no
  8    88   no     no

------------------------------ Frame 10 ------------------------------
                       Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1   145  thin   on     yes      N/A    N/A  LCDs are blank     N/A
  2   146  thin   on     no       N/A    N/A  LCDs are blank     N/A

          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1   145   yes    yes
  2   146 noconn noconn

------------------------------ Frame 11 ------------------------------
                       Host     Key     Env   Front Panel      LCD/LED
Slot Node Type  Power Responds Switch  Error LCD/LED          Flashes
---- ---- ----- ----- -------- ------- ----- ---------------- -------
  1   161  thin   on     no       N/A    N/A  LCDs are blank     N/A

          Switch Responds (per plane)
Slot Node 0      1
---- ---- ------ ------
  1   161 noconn noconn

------------------------------ Frame 12 ------------------------------
                 Clock Env
Slot Type  Power Input Error
---- ----- ----- ----- -----
  2   swit   on    0     no
  4   swit   on    0     no
```

You can also query specific hardware information using the query option of **spmon**. For example, you can get the Power LED status of node 17:

```
>spmon -q node17/powerLED/value
1
```

This option is generally used when writing script. For interactive use, it is easier to use the graphical tools provided by PSSP (see 10.4, “Graphical user interface” on page 361).

10.3.6 Monitoring node LEDs: spmon -L, spled

If you only wish to remotely look at the LEDs on the front panel of nodes, there are alternatives to the **spmon -d** command:

- **spmon -L <node>** retrieves the current value of the LED display for one node.
- **spled** opens a graphical window on your X terminal and starts monitoring and displaying in this window the values of the LEDs for all nodes. The window stays open until you terminate the **spled** process.

10.3.7 Extracting SDR contents

The SDR is the main repository for information about an SP system. It is, therefore, important that you know how to manage the information it contains. Many commands are available for this purpose. We only present two of these commands here. We strongly encourage you to refer to *PSSP Command and Technical Reference (2 Volumes),* SA22-7351, and to read about these two commands, as well as about all commands whose names start with SDR.

SDRGetObjects

The **SDRGetObjects** command extracts information about all objects in a class. For example, you can list the reliable hostname of all SP nodes:

```
[c179s][/]> SDRGetObjects Node reliable_hostname
reliable_hostname
c179n01.ppd.pok.ibm.com
c179n05.ppd.pok.ibm.com
c179n09.ppd.pok.ibm.com
c179n13.ppd.pok.ibm.com
e179n01.ppd.pok.ibm.com
e179n05.ppd.pok.ibm.com
e179n09.ppd.pok.ibm.com
e179n13.ppd.pok.ibm.com
c179mn01.ppd.pok.ibm.com
c59rp01.ppd.pok.ibm.com
c59rp02.ppd.pok.ibm.com
e159rp01.ppd.pok.ibm.com
e159rp02.ppd.pok.ibm.com
e159rp03.ppd.pok.ibm.com
e159rp04.ppd.pok.ibm.com
e159rp05.ppd.pok.ibm.com
e159rp06.ppd.pok.ibm.com
```

```
e159rp07.ppd.pok.ibm.com
e159rp08.ppd.pok.ibm.com
c59ih04.ppd.pok.ibm.com
[c179s][/]>
```

The output of **SDRGetObjects** can be long when you display information about all objects that are defined in a class. You can, therefore, use the == option of this command to filter the output: The command will only display a result for objects that satisfy the predicate specified with ==. For example, to display the node number and name of the lppsource directory used by only the multiprocessor nodes in our environment:

```
[sp3en0:/]# SDRGetObjects Node processor_type==MP node_number
lppsource_name
node_number  lppsource_name
           1 aix432
```

splstdata

The **SDRGetObjects** command is very powerful and is often used in cluster management script files. However, its syntax is not very suitable for everyday interactive use by the SP administrator since it requires that you remember the exact spelling of classes and attributes. PSSP provides a front end to **SDRGetObjects** for the most often used queries: **splstdata**. This command offers many options. We have already presented options **-a**, **-b**, **-f**, and **-n** in 9.2.14, “Verify all node information” on page 323. You must also know how to use:

splstdata -v	To display volume group information
splstdata -h	To extract hardware configuration information
splstdata -i	To display node IP configuration

10.3.8 Checking IP connectivity: ping/telnet/rlogin

The availability of IP communication between the CWS and the Cluster 1600 nodes is critical for the successful operation of the clustered system. However, PSSP does not provide any tool to check the TCP/IP network since there is nothing specific to the cluster in this area. Common TCP/IP commands can be used in the cluster environment: **ping**, **telnet**, **rlogin**, **traceroute**, **netstat**, **arp**, and so on. These commands will return information for all IP connections, including the SP Ethernet service network and the Switch network if it has been configured to provide IP services. For example, running the **arp -a** command on node 6:

```
[sp3n06.msc.itso.ibm.com:/]# arp -a
  ? (192.168.13.4) at 0:3:0:0:0:0
  sp3sw05.msc.itso.ibm.com (192.168.13.5) at 0:4:0:0:0:0
  sp3sw07.msc.itso.ibm.com (192.168.13.7) at 0:6:0:0:0:0
  sp3n01en1.msc.itso.ibm.com (192.168.31.11) at 2:60:8c:e8:d2:e1 [ethernet]
```

```
    sp3n05.msc.itso.ibm.com (192.168.31.15) at 10:0:5a:fa:13:af [ethernet]
    sp3n07.msc.itso.ibm.com (192.168.31.17) at 10:0:5a:fa:13:d1 [ethernet]
  [sp3n06.msc.itso.ibm.com:/]#
```

shows that IP communications have already been established between node 6 and node 7 through the Ethernet network as well as through the switch.

10.3.9 SMIT access to verification commands

Many of the commands listed previously can be accessed through SMIT.

Each option of the **splstdata** can be called from an entry in the SMIT List Database Information window (`smitty list_data`) or one of its subwindows.

Figure 10-4 present the SMIT RS/6000 SP Installation/Configuration Verification window (`smitty SP_verify`). The first six entries in this window respectively correspond to **spmon_itest, spmon_ctest**, **SDR_test**, **SYSMAN_test**, **CSS_test**, and **spverify_config.** The last one corresponds to commands we did not mention in this section: **st_verify**.

```
 RS/6000 SP Installation/Configuration Verification

Move cursor to desired item and press Enter.

  System Monitor Installation
  System Monitor Configuration
  System Data Repository
  System Management
  Communication Subsystem
  System Partition Configuration
  Job Switch Resource Table Services Installation
```

Figure 10-4 SMIT verification window

10.4 Graphical user interface

PSSP provides an alternative to the use of the command line interface or the SMIT panels for monitoring a system. You can use the graphical interface for that purpose. These interfaces are started by the command **perspectives**.

It is impossible in a book such as this study guide to provide a complete description of all the features of the new PSSP Perspectives User Interface. All monitoring and control functions needed to manage a Cluster 1600 system can

be accessed through this interface. We therefore recommend that you refer to chapter 19 of *PSSP Administration Guide,* SA22-7348 and *SP Perspectives: A New View of Your SP System,* SG24-5180, for further information about this tool. Another good source of information is the Perspectives online help available from the Perspectives Launch Pad.

The Perspectives initial panel, Launch Pad, is customizable. You can add icons to this panel for the actions you use often. By default, the Launch Pad contains shortcuts to some of the verification commands we presented in previous sections:

- Monitoring of **hostsResponds**, **switchResponds**, **nodePowerLEDs**
- **SMIT SP_verify**
- **syspar_ctrl -E**

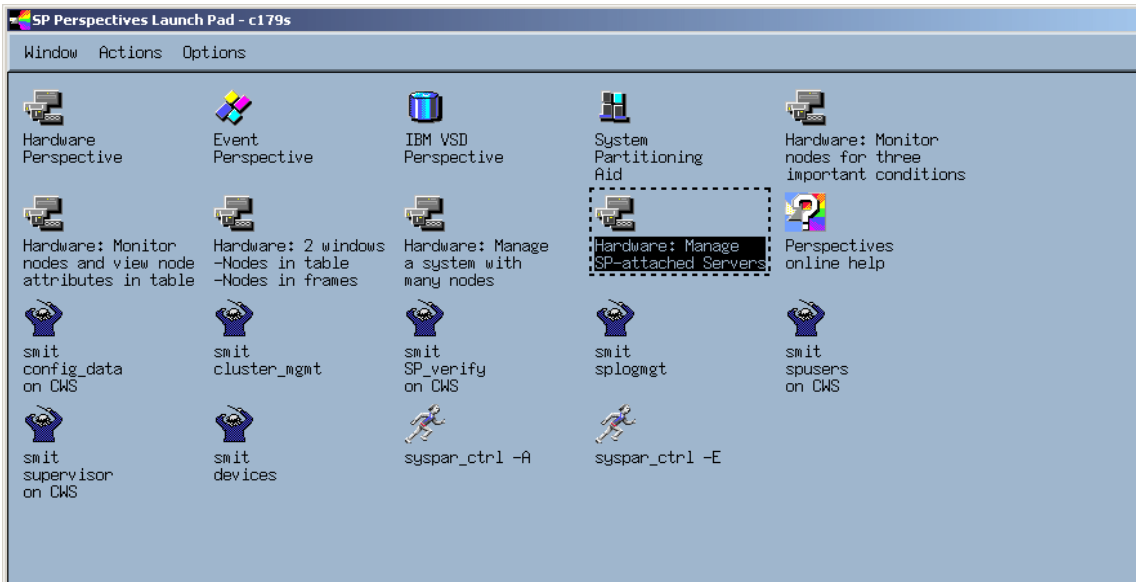


Figure 10-5 Perspectives Launch Pad

If you decide to perform most of your cluster monitoring through the Perspectives tools, we recommend that you add your favorite tools to the Launch Pad.

10.5 Key daemons

The management of a Cluster 1600 system relies heavily on the availability of several daemons. It is important that you understand the role of these daemons.

Furthermore, you should know, for the most important daemons, how they are started, how to check whether they are running, and how they interact.

The Cluster 1600-related daemons are listed in Table 10-1.

Table 10-1 Cluster 1600 daemons

Function	Daemon used
Hardware monitoring	hardmon, S70d, hmcd
SDR	sdrd
Switch fault handling	fault_service_Worm_RTG_SP, also known as the Worm
Switch management	cssadm, css.summlog
System partition-sensitive daemons	haemd, hagsd, hagsglsmd, hatsd, hrd
Kerberos daemons	kadmind, kerberos, kpropd
Event and Problem management	pmand, pmanrmd
SP SNMP trap generator	sp_configd
Hardware events logging	splogd
SNMP manager	spmgrd
File collection	supfilesrv
Job Switch Resource Table Services	Job Switch Resource Table Services
Sysctl	sysctld
Network Time Protocol	xntpd

We now provide a brief description of some of these daemons.

10.5.1 Sdrd

The sdrd daemon runs on the CWS. It serves all requests from any client application to manipulate SDR information. It is managed using the AIX SRC commands. The sdrd is started at CWS boot time. This daemon must be running before any SP management action can be performed. Appendix E of *PSSP Administration Guide,* SA22-7348 gives you details on the SDR.

You can use any of the following commands to check that the sdrd is running:

```
ps -ekf | grep sdrd
```

```
lssrc -g sdr
SDR_test
splstdata -e
```

10.5.2 Hardmon

The hardmon daemon runs on the CWS. It uses different protocols to communicate with each type of the Cluster 1600 servers. It manages the serial port of the CWS that are connected to the SP frames with the help of the SP protocol. In the case of the Enterprise server (S70, S7A, S80, p680) that uses SAMI protocol, the hardmon uses the s70d daemon to make connections for every attached server. The pSeries (POWER3) servers communicate with hardmon through the CSP protocol. For the SP-attached POWER3 servers, using the CSP protocol, no other daemon is necessary to translate the communication protocol for hardmon. There is one additional hmcd daemon running for each HMC server on the CWS to provide communication between the hardmon daemon and the HMC server through the Ethernet. Refer to Figure 5-7 on page 190 for details about hardmon communications.

No management of the cluster hardware can be performed until the hardmon daemon is running. It is, therefore, important that you verify that this daemon is always running on the CWS. You can check hardmon with one of the following commands:

```
ps -ekf | grep hardmon
lssrc -s hardmon
spmon_ctest
```

10.5.3 Worm

The worm runs on all cluster nodes equipped with a switch. It is started by the **rc.switch** script, which is started at node boot time. The worm must be running on the primary node before you can start the switch with the **Estart** command. We recommend that you refer to Chapter 14 of the *PSSP Administration Guide,* SA22-7348 for more details about the Switch daemons.

10.5.4 Topology Services, Group Services, and Event Management

The Topology Services, Group Services, and Event Management subsystems are managed by the PSSP **syspar_ctrl** command (refer to 10.3.2, “Verify system partitions” on page 352).

These subsystems are closely related. The Topology Services provide information about the cluster systems to the Group Services. Event Management subsystems rely on information provided by the Topology Services subsystem to

offer their own services to other client applications. VSD, RVSD, and GPFS are examples of client applications of the Topology Services, Group Services, and Event Management subsystems.

We recommend that you refer to chapters 23, 24, and 25 of *PSSP Administration Guide,* SA22-7348 for details about these subsystems.

10.6 SP-specific logs

Since the Cluster 1600 systems are complex, the amount of data that a Cluster 1600 administrator may need to look at to manage such systems is far beyond what can reasonably be gathered in one file or displayed on one screen.

The various components of PSSP, therefore, store information about their processing in various logs. PSSP generates information in about 30 log files. A complete list of all these logs can be found on pages 74-79 of *PSSP Diagnosis Guide,* GA22-7350.

Most of the cluster-related logs can be found in /var/adm/SPlogs on the CWS and on the cluster nodes. A few other logs are stored in /var/adm/ras.

You generally only look at logs for problem determination. For the purpose of this chapter (verifying the PSSP installation and operation), we only mention the /var/adm/SPlogs/sysman directory. On each node, this directory contains the trace of the AIX and PSSP installations, their configuration, and the execution of the customization scripts described in 9.3.2, “/tftpboot” on page 335. We recommend that you look at this log after the installation of a node to check that it has successfully completed. The installation of a node involves the execution of several processes that are not linked to a terminal (scripts defined in /etc/inittab, for example). You may not notice that some of these scripts have failed if you do not search for indications of their completion in the /var/adm/SPlogs/sysman directory.

10.7 Related documentation

For complete reference and ordering information for the documents listed in this section, see “Related publications” on page 553.

SP manuals

Refer to the related documents for Version 3.5 of PSSP:

PSSP Installation and Migration Guide, GA22-7347. The installation of a Cluster 1600 system is a long process involving many steps (up to 85). Therefore,

several verifications can be performed during installation to ensure that the already executed steps have been completed correctly. Chapter 2, “Validate hardware and software configuration” on page 7 of this guide documents the use of these verification methods at various stages of the Cluster 1600 installation.

PSSP Command and Technical Reference (2 Volumes), SA22-7351 contains a complete description of each command listed in 10.3, “Key commands” on page 348.

Chapter 29 of *PSSP Diagnosis Guide,* GA22-7350 describes, in detail, the verification of System Management installation using the **SYSMAN_test** command.

Chapter 14 of *PSSP Administration Guide,* SA22-7348 describes the switch related daemons, while chapters 23, 24, and 25 provide detailed information about the partition-sensitive subsystems and their daemons.

SP redbooks

Chapter 5 in *RS/6000 SP Monitoring: Keeping It Alive*, SG24-4873 provides a detailed description of the Perspectives graphical user interface.

SP Perspectives: A New View of Your SP System, SG24-5180, is entirely dedicated to explaining the use of Perspectives, but only addresses Version 3.1 of PSSP.

10.8 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, “Answers to sample questions” on page 521.

1. PSSP provides several tools and scripts for checking components and verifying that they are working properly. Which command can be used to verify that the SDR has been properly set up and that it is working fine?
   a. **test_SDR**
   b. **SDR_itest**
   c. **SDR_ctest**
   d. **SDR_test**
2. How do you obtain frame, switch, and node hardware information in PSSP 3.5?
   a. Run the command **spmon -g**.
   b. Run the command **SDRGetObjects Hardware**.

c. Run the command **spmon -d**.

d. Run the command **spmon -G -d**.

3. The hardmon daemon runs on the CWS only. Which of the following statements is false?

   a. For hardware control of the various Cluster 1600 nodes it uses different types of protocols.

   b. It is a partition-sensitive daemon.

   c. It requires read/write access to each tty connected to frames.

   d. It logs information in the /var/adm/SPlogs/hardmon directory.

4. Which of the following worm characteristics is *false*?

   a. The worm runs on all SP nodes in an SP system equipped with a switch.

   b. The worm is started by the rc.switch script.

   c. The worm must be started manually.

   d. The worm must be running on the primary node before you can start the switch.

5. Which command checks a large number of SP system management components?

   a. **spmon_ctest**

   b. **SYSMAN_test**

   c. **test_SYSMAN**

   d. **css_test**

6. Which of the following commands lists the remote authentication settings and the trusted services that are active for a system partition?

   a. **chauthpar**

   b. **chauthpts**

   c. **lsauthpar**

   d. **lsauthpts**

10.9 Exercises

Here are some exercises you may wish to perform:

1. Familiarize yourself with the different verification and monitoring commands documented in the chapter.

2. Use the various flags of the **splstdata** command to extract data from the SDR.
3. Familiarize yourself with some of the key daemons documented in 10.5, “Key daemons” on page 362.

11

Cluster 1600-supported products

Cluster 1600 managed by PSSP

Parallel System Support Programs (PSSP) is a collection of cluster software tools that provide a solid foundation on which to scale workloads and manage more than a hundred clustered servers. Built on over a decade of RS/6000 SP clustering experience, PSSP offers high performance and scalability, and easy to use systems management. It provides “cluster-aware” tools for hardware (hardmon), device management, security administration (Kerberos), error logging, problem management (pman), just to mention a few. Everything is controlled from a single point-of-control, the control workstation (CWS).

PSSP includes several components that help deliver extreme scalability for parallel applications in high performance computing, including:

- Communications Subsystem - software that supports SP Switch and SP Switch2, including device drivers, switch configuration and initialization, fault handling, and switch adapter diagnostics
- IBM Virtual Shared Disk (VSD) - an application programming interface used to create logical disk volumes that can be accessed by serial or parallel applications running on any server in the cluster (regardless of the physical location of the disk drives from which, or to which, the data blocks are being transferred). We briefly cover the terminology of VSD in 11.3, “IBM Virtual Shared Disks” on page 380.

- IBM Recoverable Virtual Shared Disk (RVSD) - software that provides recovery from failure of virtual shared disk server nodes and is designed to ensure continuous access to data and file systems from anywhere in the cluster. We briefly cover the terminology in 11.5, “IBM Recoverable Virtual Shared Disk” on page 382.
- PSSP, along with the IBM Parallel Environment for AIX, supports the Message Passing Interface standard for the development of distributed memory parallel applications. See 11.7, “IBM Parallel Environment” on page 387.
- The newest version of PSSP, Version 3.5 for AIX 5L, provides investment protection for existing SP and Cluster 1600 customers. It lets them add new pSeries servers to clusters that also include older pSeries or RS/6000 servers, SP nodes, SP Switch or SP Switch2. PSSP Version 3.5 provides full support of the AIX 5L 64-bit kernel and device drivers.

A large number of software products are supported in the Cluster 1600 configuration. Although most of these products are not essential for every cluster configuration, they are commonly found in customer environments.

Important: At the time of writing there was some limitation to what was supported. On an AIX 5L V5.2 system running PSSP 3.5, some licensed programs were currently *not* supported:

- IBM General Parallel File System for AIX (GPFS)
- IBM LoadLeveler for AIX 5L (LoadLeveler)
- IBM Parallel Environment for AIX (PE)
- IBM Engineering and Scientific Subroutine Library (ESSL) for AIX
- IBM Parallel Engineering and Scientific Subroutine Library (Parallel ESSL) for AIX

In addition, a system with a switch will run in IP mode only on AIX 5L 5.2.

In this chapter we provide the basic concepts for understanding and configuring some of these cluster products. In preparation for the certification exam, you should understand how the following products work and what solutions they provide:

- IBM LoadLeveler

  Distributed, network-wide job management
- High Availability Cluster Multi-Processing
  - HACWS, High Availability Control Work Station
  - HACMP or HACMP/ES

- IBM Virtual Shared Disks

  Create and manage virtual shared disks, Concurrent and Recoverable
- IBM General Parallel File System for AIX (GPFS)

  Data file system for parallel access to large files
- IBM Parallel Environment for AIX (PE)

  Parallel application development and execution, message passing, and parallel task communications
- Engineering and Scientific Subroutine Library (ESSL)

  Collection of mathematical subroutines that provide optimum performance for floating-point engineering and scientific applications

11.1 LoadLeveler

LoadLeveler is a program designed to automate workload management. In essence, it is a scheduler that also has facilities to build, submit, and manage jobs. The jobs can be processed by any one of a number of machines, which together are referred to as the LoadLeveler cluster. Any pSeries machine may be part of a cluster, although LoadLeveler is most often run in the SP or cluster environment. A sample LoadLeveler cluster is shown in Figure 11-1.

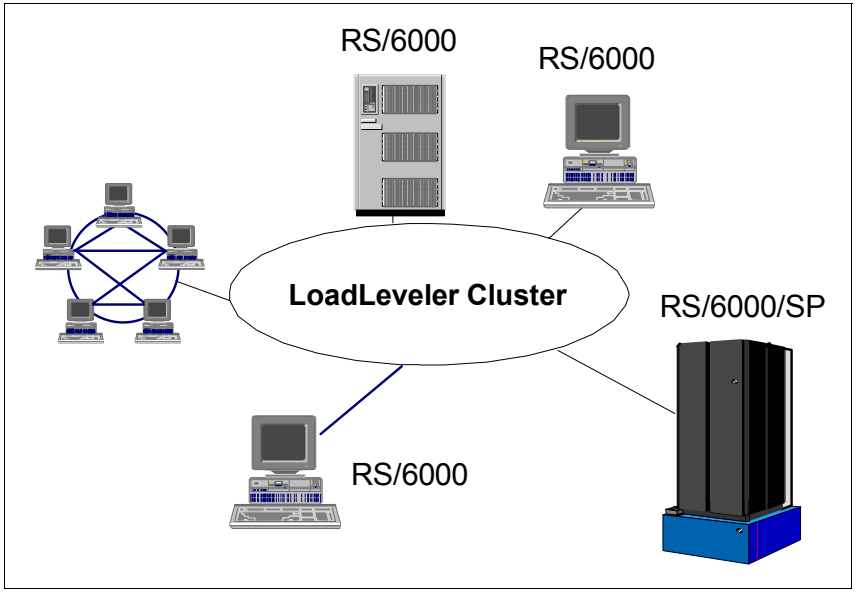


Figure 11-1 Example LoadLeveler configuration

Important concepts in LoadLeveler are:

Cluster - A group of machines that are able to run LoadLeveler jobs. Each member of the cluster has the LoadLeveler software installed.

Job - A unit of execution processed by LoadLeveler. A serial job runs on a single machine. A parallel job is run on several machines simultaneously and must be written using a parallel language Application Programming Interface (API). As LoadLeveler processes a job, the job moves into various job states, such as Pending, Running, and Completed.

Job Command File - A formal description of a job written using LoadLeveler statements and variables. The command file is submitted to LoadLeveler for scheduling of the job.

Job Step - A job command file specifies one or more executable programs to be run. The executable and the conditions under which it is run are defined in a single job step. The job step consists of several LoadLeveler command statements.

Figure 11-2 on page 373 schematically illustrates a series of job steps. In this figure, data is copied from tape job step 1. Depending on the exit status of this operation, the job is either terminated or continues on to job step 2. Again, LoadLeveler examines the exit status of job step 2 and either proceeds on to job step 3, which, in this example, formats and prints the data that the user requires, or terminates.

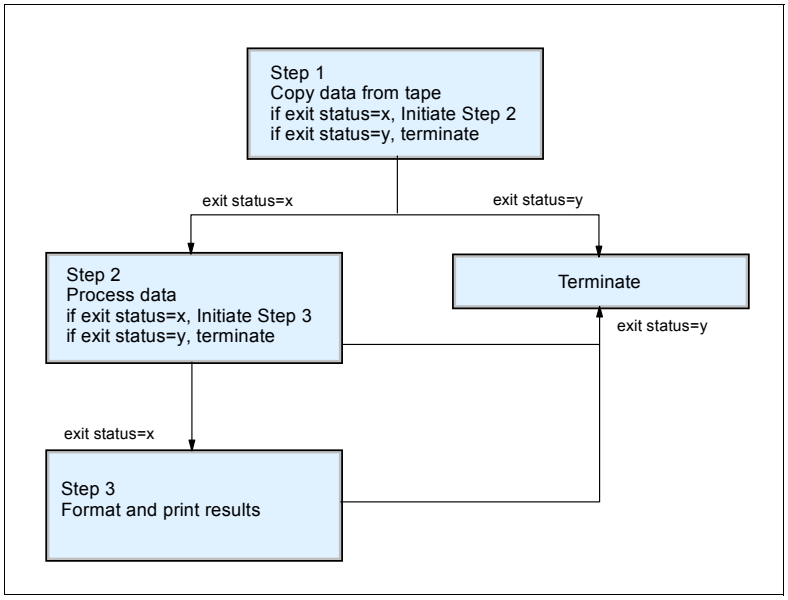


Figure 11-2 LoadLeveler job steps

Each machine in the LoadLeveler cluster performs one or more roles in scheduling jobs. These roles are as follows:

- Scheduling Machine

  When a job is submitted, it gets placed in a queue managed by a scheduling machine. This machine contacts another machine that serves as the central manager for the entire LoadLeveler cluster. The scheduling machine asks the central manager to find a machine that can run the job, and also keeps persistent information about the job. Some scheduling machines are known as public scheduling machines, meaning that any LoadLeveler user can access them. These machines schedule jobs submitted from submit-only machines, which are described below.

- Central Manager Machine

  The role of the Central Manager Machine is to examine a job's requirements and find one or more machines in the LoadLeveler cluster that will run the job. Once it finds the machines, it notifies the scheduling machine.

- Executing Machine

  The machine that runs the job is known as the Executing Machine.

- Submitting Machine

  This type of machine is known as a submit-only machine. It participates in the LoadLeveler cluster on a limited basis. Although the name implies that users

of these machines can only submit jobs, they can also query and cancel jobs. Users of these machines also have their own Graphical User Interface (GUI) that provides them with the submit-only subset of functions. The submit-only machine feature allows workstations that are not part of the LoadLeveler cluster to submit jobs to the cluster.

Once LoadLeveler examines a job to determine its required resources, the job is dispatched to a machine to be processed. Arrows 2 and 3 in Figure 11-3 indicate that the job can be dispatched to either one machine, or—in the case of parallel jobs—to multiple machines. Once the job reaches the executing machine, the job runs.

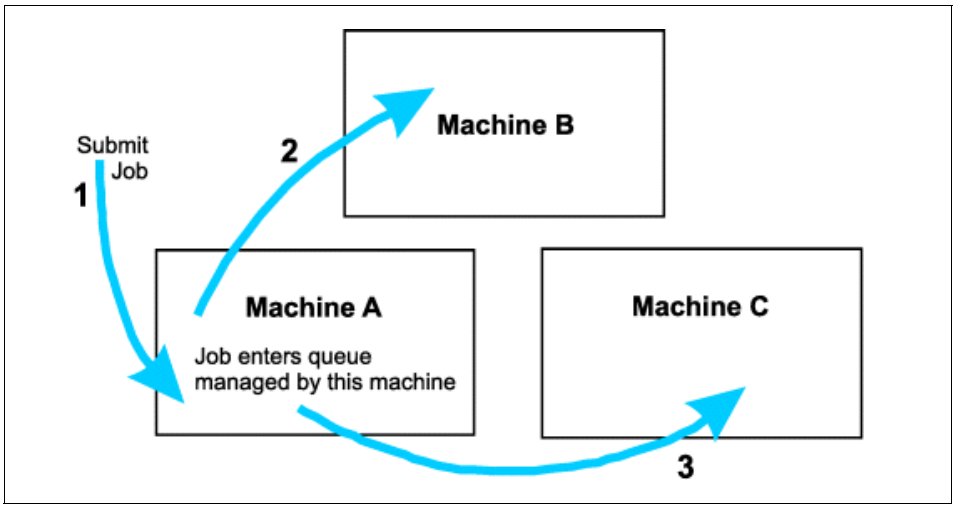


Figure 11-3 Job queues

Jobs do not necessarily get dispatched to machines in the cluster on a first-come, first-serve basis. Instead, LoadLeveler examines the requirements and characteristics of the job and the availability of machines, and then determines the best time for the job to be dispatched.

Tip: Keep in mind that one machine can assume multiple roles.

Jobs do not get dispatched to the executing machines on a first-come, first-served basis unless LoadLeveler is specifically configured to run that way, that is, with a first-in first-out (FIFO) queue. Instead, the negotiator calculates a priority value for each job called SYSPRIO that determines when the job will run. Jobs with a high SYSPRIO value run before those with a low value.

The system administrator can specify several parameters that are used to calculate SYSPRIO. Examples of these are: How many other jobs the user already has running, when the job was submitted, and what priority the user has assigned to it. The user assigns priorities to his own jobs by using the user_priority keyword in the job command file.

SYSPRIO is referred to as a job's *system priority*; whereas the priority that a user assigns his own jobs is called *user priority*. If two jobs have the same SYSPRIO calculated for them by LoadLeveler, then the job that runs first is the job that has the higher user priority.

Understanding the SYSPRIO expression: SYSPRIO is evaluated by LoadLeveler to determine the overall system priority of a job. A system priority value is assigned when the negotiator adds the new job to the queue of jobs eligible for dispatch.

The SYSPRIO expression can contain class, group, and user priorities, as shown in the following example:

SYSPRIO: (ClassSysprio * 100) + (UserSysprio * 10) + (GroupSysprio * 1) - (QDate)

The priority of a job in the LoadLeveler queue is completely separate and must be distinguished from the AIX *nice* value, which is the priority of the process the executable program is given by AIX.

LoadLeveler also supports the concept of *job classes*. These are defined by the system administrator and are used to classify particular types of jobs. For example, we define two classes of jobs that run in the clusters called *night* jobs and *day* jobs. We might specify that executing machine A, which is very busy during the day because it supports a lot of interactive users, should only run jobs in the night class. However, machine B, which has a low workload during the day, could run both. LoadLeveler can be configured to take job class into account when it calculates SYSPRIO for a job.

As SYSPRIO is used for prioritizing jobs, LoadLeveler also has a way of prioritizing executing machines. It calculates a value called MACHPRIO for each machine in the cluster. The system administrator can specify several different parameters that are used to calculate MACHPRIO, such as load average, number of CPUs, the relative speed of the machine, free disk space, and the amount of memory.

Machines may be classified by LoadLeveler into pools. Machines with similar resources, for example a fast CPU, might be grouped together in the same pool so that they could be allocated CPU-intensive jobs. A job can specify as one of

its requirements that it will run on a particular pool of machines. In this way, the right machines can be allocated the right jobs.

11.2 HACWS

HACWS is an optional collection of components that implement a backup CWS for a Cluster 1600 managed by PSSP. The backup CWS takes over when the primary CWS requires upgrade service or fails. The HACWS components are:

- A second pseries machine supported for CWS use

  Refer to *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281 for more information about which are supported

- The HACWS connectivity feature (#1245) ordered against each frame in the system

  This furnishes a twin-tail for the RS-232 connection so that both the primary and backup CWSs can be physically connected to the frames.

- HACMP for AIX installed on each CWS

  HACWS is configured as a two-node rotating HACMP cluster.

- The HACWS feature of PSSP

  This software provides SP-specific cluster definitions and recovery scripts for CWS failover. This feature is separately orderable and priced and does not come standard with PSSP.

- Twin-tailed external disk

  This is physically attached to each CWS to allow access to data in the /spdata file system.

An HACWS cluster is depicted in Figure 2-14 on page 41.

If the primary CWS fails, the backup CWS can assume all CWS functions with the following exceptions:

- Updating passwords (if SP User Management is in use)
- Adding or changing SP users
- Changing Kerberos keys (the backup CWS is typically configured as a secondary authentication server)
- Adding nodes to the system
- Changing site environment information

HACMP or HACMP/ES is run on the CWS only if HACWS is being used. HACMP/ES is run on the nodes.

HACMP ensures that critical resources are available for processing. HACMP has several features that you can choose to use independently. You can run HACMP on the Cluster 1600 system managed by PSSP with or without the Enhanced Scalability feature (HACMP/ES).

HACMP/ES builds on the Event Management and Group Services components of RSCT to scale HACMP function.

IBM High Availability Cluster Multi-Processing / ES

IBM High Availability Cluster Multi-Processing for AIX Enhanced Scalability (HACMP/ES) was developed to exploit SP High Availability Infrastructure (HAI) of IBM Parallel System Support Programs for AIX (PSSP). When HACMP/ES was introduced, it only supported the SP. However, these days our configurations are a combination of different SP systems and stand-alone IBM RS/6000 or pSeries workstations. Newer clusters are likely to include the IBM pSeries 690 (Regatta). HACMP/ES clusters can contain up to 32 nodes.

HACMP/ES provides an environment that insures that mission-critical applications can recover quickly from hardware and software failures. HACMP/ES is a high availability system that ensures that critical resources are available for processing.

While some of the AIX 5L features eliminate specific components as “single points of failures,” the system is still vulnerable to other component failures, such as a disk array. By itself it could be made a highly available component, but if the processor to which it is connected fails, then the disk array also becomes unavailable.

If you are doing mission-critical processing, you need something that ensures total system availability. High availability combines software with hardware to minimize downtime by quickly restoring services when a system, component, or application fails. While not instantaneous, the restoration of service is rapid, usually 30 to 300 seconds.

IBM High Availability Geographic Cluster (HAGEO)

HAGEO is an extension of the HACMP software. HACMP ensures that the computing environment within a site remains highly available. HAGEO ensures that the critical data remains highly available even if an entire site fails or is destroyed by a disaster.

The first thing to notice is that an HAGEO cluster is just like an HACMP cluster. An HAGEO cluster can consist of two to eight nodes, just as a HACMP cluster. However, in HAGEO, these nodes are now spread between two sites. The sites are connected by one or more geographic networks, even though you may be

inclined to think of this as two HACMP clusters connected by geographic networks.

When you configure the HAGEO cluster, you define a new component: The HAGEO site. You define the nodes as belonging to one of the sites, and you include the user ID of the person to be notified in case of a site-related event. You also define one site as the dominant site and include information about the backup communications link between sites you have set up. The dominant site is the site you choose to leave running if site isolation occurs. In each site, the two nodes have a set of shared disks physically cabled together in the same way as an HACMP cluster. In most cases, these shared disks contain the data that will be geographically mirrored from one site to another.

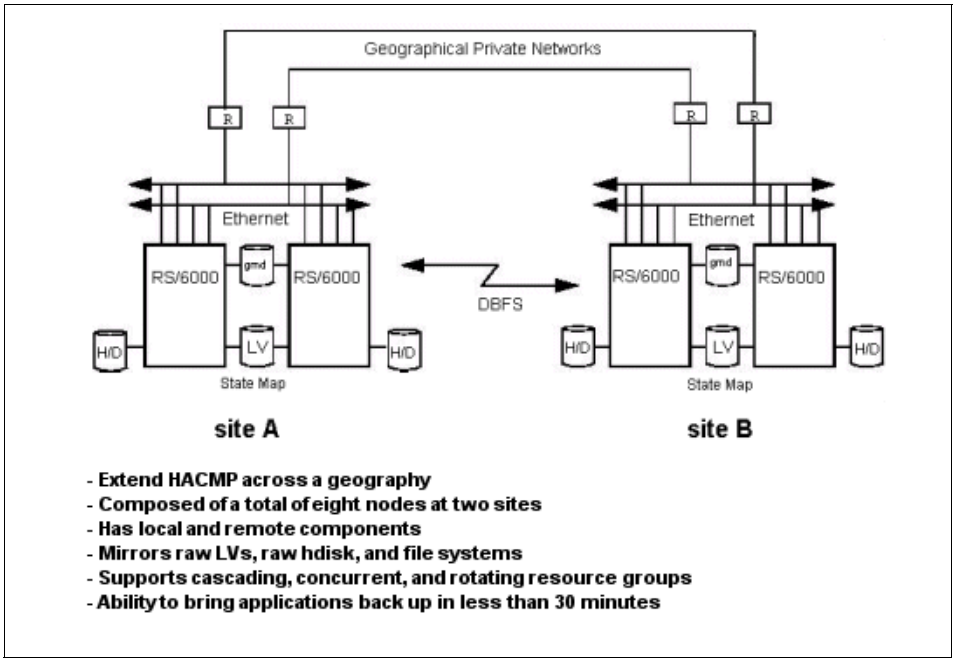


Figure 11-4 Site protection with HAGEO

HAGEO provides real-time mirroring of customer data between systems connected by local or point-to-point networks, bringing disaster recovery capability to a cluster of IBM pSeries or RS/6000 nodes placed in two widely separated geographic locations. HAGEO automatically responds to site and communication failures and provides for automatic site takeover; refer to Figure 11-4. Tools are available for data synchronization after an outage, configuration management, capacity planning, performance monitoring, and problem determination.

IBM Geographic Remote Mirror (GeoRM)

The geographic remote mirroring capability is available alone, as a separate feature without the failure detection and automatic takeover capability of HAGEO. GeoRM allows customer data to be mirrored in real time between geographically dispersed locations using LANs or point-to-point networks, with no limitation on the distance between locations.

The basic GeoRM configurations are as follows:

- GeoRM active-backup

  All applications run at the active site and are mirrored to the backup site. One variation of this configuration includes having one or more applications run on machines in dispersed geographic locations; these are all configured as belonging to one active site. These applications are mirrored to one or more machines at the backup site. Figure 11-5 shows a 4-machine GeoRM configuration.

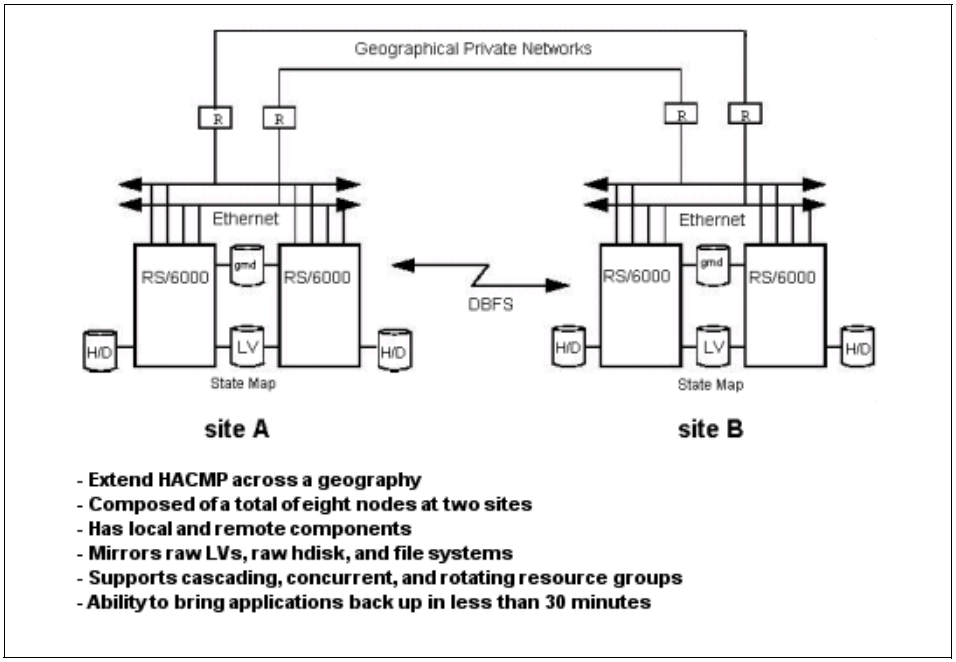


Figure 11-5 GeoRM configuration

- GeoRM mutual backup

  Applications run at each site and are mirrored at the other site. All configurations provide a rapid remote backup for critical data. The data at the backup site is guaranteed to be up-to-date with the data entered at the active

site until the failure of an active site or one of its components interrupts the geo-mirroring process. Basically, machines at the backup sites serve as storage devices

11.3 IBM Virtual Shared Disks

The IBM Virtual Shared Disks (VSD) is a subsystem that lets application programs that are running on different nodes of a system partition access a raw logical volume as if it were local at each of the nodes. Each virtual shared disk corresponds to a logical volume that is actually local at one of the nodes, which is called the server node. The IBM VSD subsystem routes I/O requests from the other nodes, called client nodes, to the server node and returns the results to the client nodes as shown in Figure 11-6 on page 381.

The I/O routing is done by the IBM VSD device driver that interacts with the AIX Logical Volume Manager (LVM). The device driver is loaded as a kernel extension on each node, so that raw logical volumes can be made globally accessible.

The configuration has two different locations with two machines at site S1 and one machine at site S2. The two machines at site S1 can be geographically separated, even though they are considered to be one site by the GeoRM software. The three machines have access to the geographical networks through the routers. Each machine has its own disks. When you configure GeoRM, you establish a one-to-one pairing between GeoMirror devices on machines alpha and delta and a one-to-one pairing between GeoMirror devices on machines beta and delta. Data entered at either machine at site alpha is sent over a geographical network and mirrored on a corresponding disk at site delta.

A second variation of this configuration includes having an application running on two or more machines in the same location, sharing a disk to facilitate handling local machine failure or maintenance while keeping the application and geo-mirroring running. The disks at site S1 are shared between the machines alpha and beta.

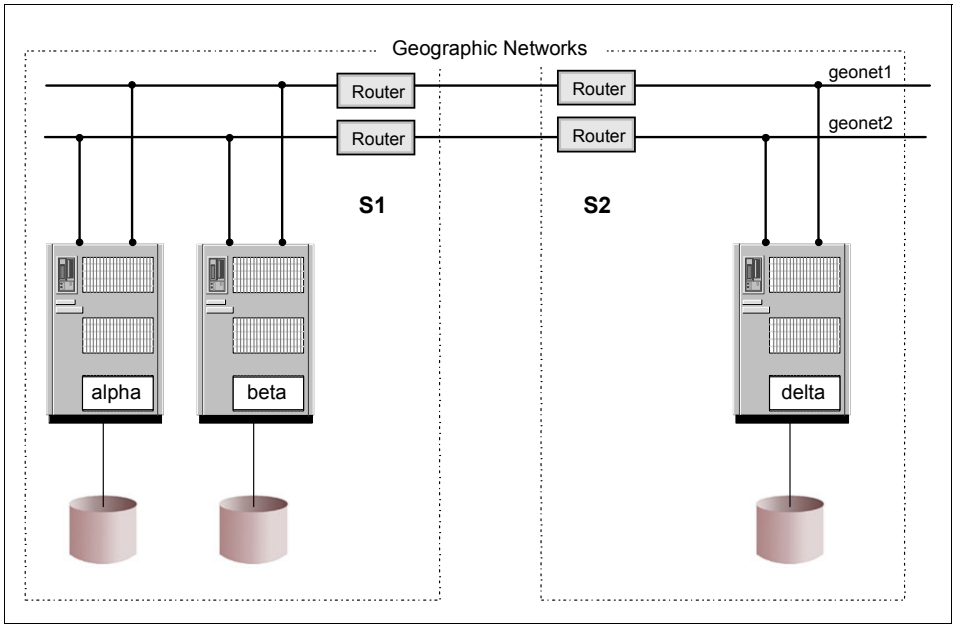


Figure 11-6 An IBM Virtual Shared Disk IP network implementation

11.4 IBM concurrent virtual shared disks

This subsystem also includes concurrent disk access, which allows you to use multiple servers to satisfy disk requests by taking advantage of the concurrent disk access environment supplied by AIX. In order to use this environment, VSD uses the services of Concurrent LVM (CLVM), which provides the synchronization of LVM and the management of concurrency for system administration services. Concurrent disk access extends the physical connectivity of multi-tailed concurrent disks beyond their physical boundaries. You can configure volume groups with a list of Virtual Shared Disk servers. Nodes that are not locally attached will have their I/O distributed across these servers. For example, in Figure 11-7 on page 382, node 1 is not attached to any disks. To access VG1, you could use nodes 2 or 3. If you want to access VG2, you can do this through node 3 only (or node 4 if/when node 3 fails).

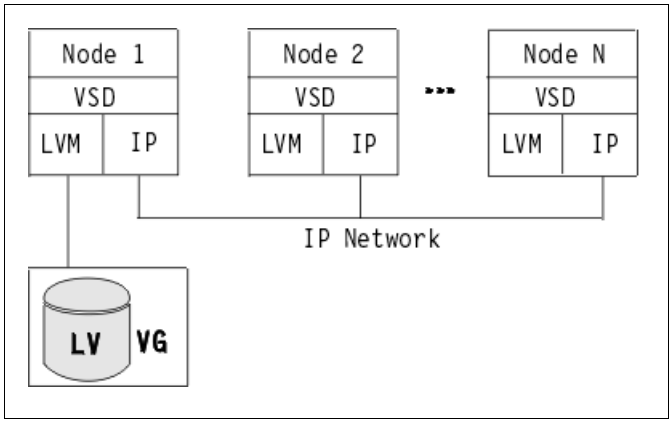


Figure 11-7 An IBM Concurrent Virtual Shared Disk IP network implementation

11.5 IBM Recoverable Virtual Shared Disk

IBM Recoverable Virtual Shared Disk (RVSD) is a subsystem that provides recoverability of your virtual shared disks if a node, adapter, or disk failure occurs.

To understand the value of the IBM RVSD component, consider a system without it. The virtual shared disk function lets all nodes in the system partition access a given disk, even though that specific disk is physically attached to only one node. If the server node should fail, access to the disk is lost until the server node is rebooted by the administrator.

By using RVSD and twin-tailed disks or disk arrays, you can allow a secondary node to take over the server function from the primary node when certain types of failure occur.

A twin-tailed disk is a disk or group of disks that are attached to two nodes of a Cluster 1600.

For recoverability purposes, only one of these nodes serves the disks at any given time. The secondary or backup node provides access to the disks if the primary node fails, is powered off, or if you need to change the server node temporarily for administrative reasons. Both must be in the same system partition.

RVSD automatically manages your virtual shared disks by detecting error conditions such as node failures, adapter failures, and hardware interruptions at the disk (EIO errors), and then switching access from the primary node to the

secondary node. This means that your application can continue to operate normally. RVSD also allows you to cut off access to virtual shared disks from certain nodes and to dynamically change the server node using the **fencevsd** and **vsdchgserver** commands (or using the GUI).

With RVSD, you can recover more easily from node failures and have continuous access to the data on the twin-tailed disks.

Virtual Shared Disk recovery with twin-tailed disks

The following three figures show a simple system with one twin-tailed RVSD configuration.

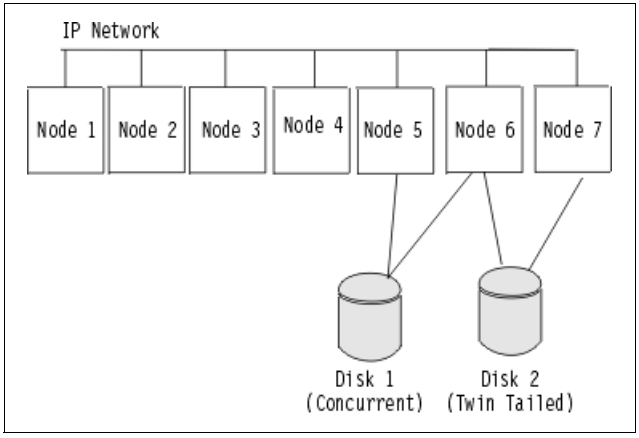


Figure 11-8 Basic one twin-tailed RVSD configuration

Figure 11-8 shows the basic configuration. The primary node is the server of information on the disk. In Figure 11-9 the secondary node is acting as the server following a failure of the primary node.

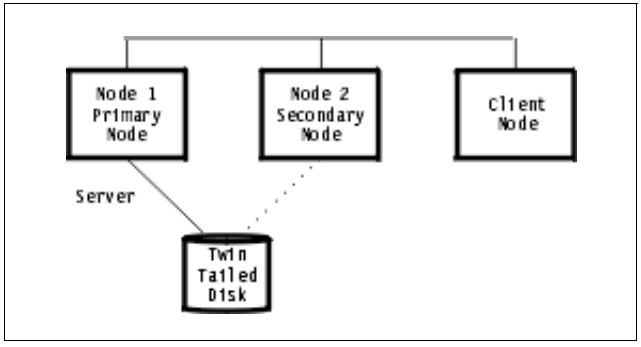


Figure 11-9 The secondary node serves after a primary node failure

In Figure 11-10, the primary node is again the server and has automatically taken over the disk from the secondary node.

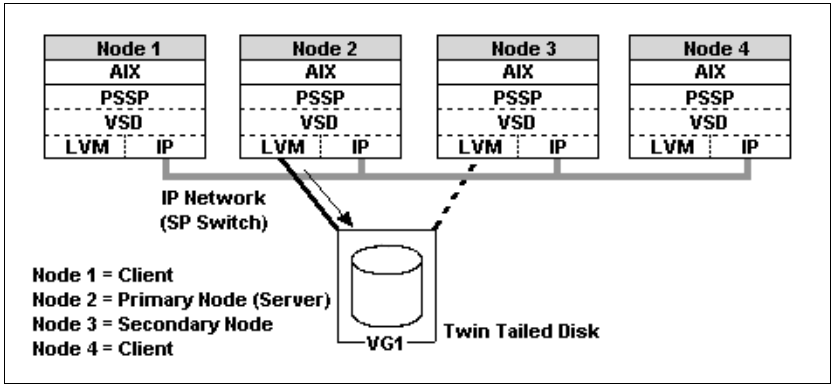


Figure 11-10 The primary node is the server again after recovery

Recovery is transparent to applications that have been enabled for recovery. There is no disruption of service, only a slight delay while takeover occurs. RVSD provides application programming interfaces, including recovery scripts and C programs, so you can enable your applications to be recoverable.

11.6 IBM General Parallel File System (GPFS)

General Parallel File System (GPFS) is a scalable cluster file system for cluster systems, including the Linux (Cluster 1350) and the AIX (Cluster 1600) systems. GPFS provides a standard, robust file system for serial and parallel applications on the Cluster 1600 managed by PSSP, where it exploits VSD technology and the Kerberos-based security features. From a user's view, it resembles NFS, but unlike NFS, the GPFS file system can span multiple disks on *multiple nodes*.

GPFS was originally developed for large-scale multimedia. Later, it was extended to support the additional requirements of parallel computing. GPFS supports file systems of several tens of terabytes and has run at I/O rates of several gigabytes per second.

GPFS was one of the key system software components of the ASCI White Supercomputer. On ASCI White, GPFS allowed parallel programs running on 492 computed nodes to read and write individual files at data rates of up to 7 GB/s in a 75 TB file system. This prodigious I/O rate and storage capacity is required to provide data to the simulation codes and to allow output to be stored at the rate required by the machine's tremendous computational power.

GPFS is particularly appropriate in an environment where the aggregate peak need for data exceeds the capability of a distributed file system server. It is not appropriate for those environments where hot backup is the main requirement or where data is readily partitioned along individual node boundaries.

In addition to high-speed parallel file access, GPFS provides fault-tolerance, including automatic recovery from disk and node failures. Its robust design makes GPFS appropriate for commercial applications such as large Web servers, data mining and digital libraries.

A user sees a GPFS file system as a normal file system. Although it has its own support commands, usual file system commands such as **`mount`** and **`df`**, work as expected on GPFS. GPFS file systems can be flagged to mount automatically at boot time. GPFS supports relevant X/OPEN standards with a few minor exceptions. Large NFS servers, constrained by I/O performance, are likely candidates for GPFS implementations.

In an AIX cluster environment, GPFS is designed to operate with the following:

- AIX 5 L, providing:
  - The basic operating system services and the routing of file system calls requiring GPFS data
  - The LVM subsystem for direct disk management
  - Persistent reserve for transparent failover of disk access in the event of disk failure
- Reliable Scalable Cluster Technology (RSCT) subsystem of AIX 5L

  This provides the capability to create, modify, and manage an RSCT peer domain as shown in Figure 11-11 on page 386:
  - The Resource Monitoring and Control (RMC) component - establishing the basic cluster environment, monitoring the changes within the domain, and enabling resource sharing within the domain.
  - The Group Services component - coordinating and synchronizing the changes across nodes in the domain, thereby maintaining the consistency in the domain.
  - The Topology Services component - providing network adapter status, node connectivity, and a reliable messaging service.

  The configuration manager employs the above subsystems to create, change, and manage the RSCT peer domain.

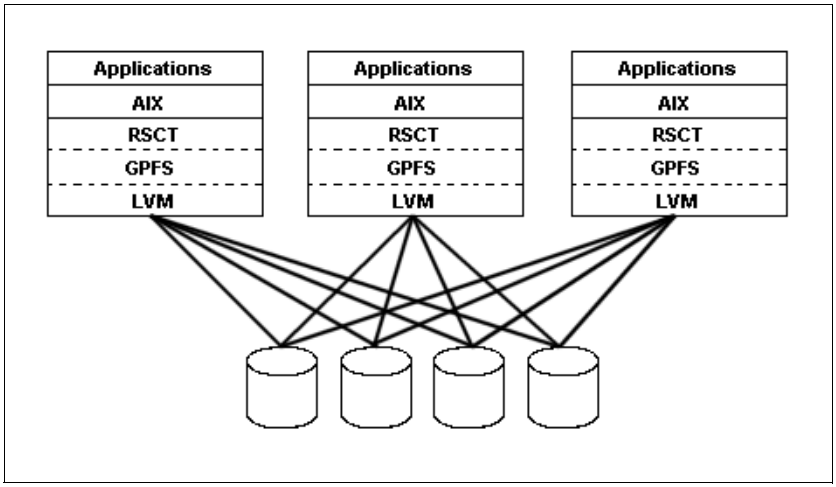


Figure 11-11 An AIX RSCT peer domain environment

- The HACMP/ES program product, providing:
  - The basic cluster operating environment
  - The Group Services component - coordinating and synchronizing the changes across nodes in the HACMP cluster, thereby maintaining the consistency in the cluster. Refer to Figure 11-12.
  - The Topology Services component - providing network adapter status, node connectivity, and a reliable messaging service.

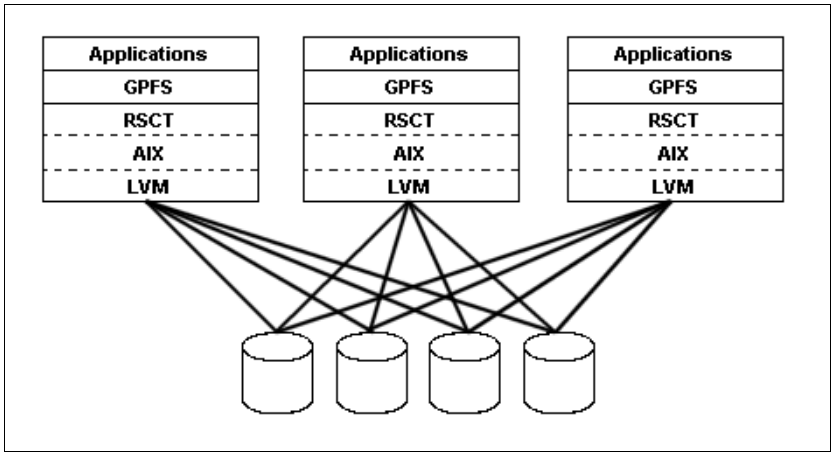


Figure 11-12 An AIX HACMP environment

  - Or, the IBM Virtual Shared Disk Subsystem component of PSSP. Figure 11-13 shows an example of GPFS in a PSSP environment, a configuration that is widely used in a Cluster 1600 managed by PSSP. For the RSCT and HACMP solution, PSSP is not needed.

- IBM Virtual Shared Disk Subsystem component of PSSP, providing:
  - The Group Services component - coordinating and synchronizing changes across nodes in the nodeset, thereby maintaining the consistency. It also provides sequencing of recovery on multiple nodes and initialization information on the IBM Virtual Shared Disk.
  - The Topology Services component - providing network adapter status, node connectivity, and a reliable messaging service.
  - The IBM Virtual Shared Disk component - for disk driver level support for GPFS cluster-wide disk accessibility.
  - The IBM Recoverable Virtual Shared Disk component - for the capability to fence a node from accessing certain disks, which is a prerequisite for successful recovery of that node. It also provides for transparent failover of disk access in the event of the failure of a disk server.

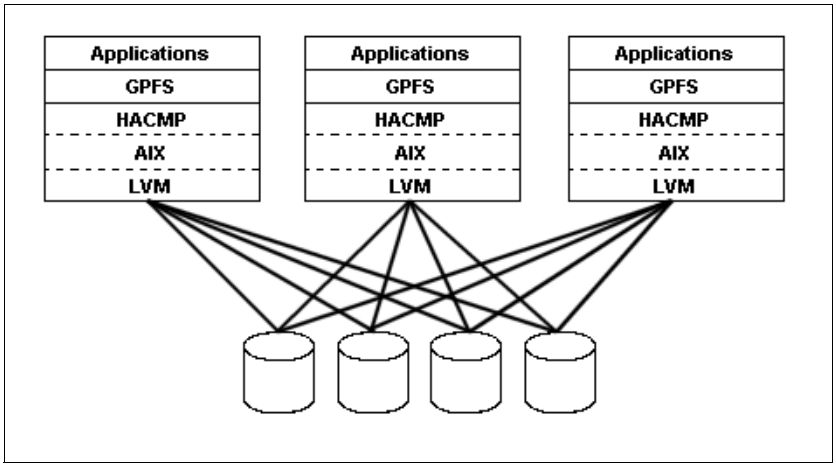


Figure 11-13 GPFS in a PSSP environment (not recommended)

11.7 IBM Parallel Environment

The IBM Parallel Environment (PE) for AIX provides support for parallel application development on a Cluster 1600 system managed by PSSP (SP system), on a single RS/6000 processor, or a TCP/IP-networked cluster of IBM RS/6000 or pSeries processors. The PE licensed program contains tools to

support the development and analysis of parallel applications written in Fortran, C, or C++. It also provides a user-friendly runtime environment for their execution. Parallel Environment supports the Message Passing Library (MPL) subroutines, the Message Passing Interface (MPI) standard, Parallel ESSL, and the Communications Low-Level Applications Programming Interface (LAPI). LAPI is designed to provide optimal communication performance on SP Switch2 or SP Switch.

PE consists of the following:

- The Parallel Operating Environment (POE), for submitting and managing jobs
- Message passing libraries (MPL and MPI), for communication among the tasks that make up a parallel program
- A parallel debugger (pdbx), for debugging parallel programs
- Parallel utilities, for easing file manipulation
- Xprofiler, for analyzing a parallel application's performance
- PE Benchmarker, a suite of applications and utilities you can use to analyze program performance

11.8 Engineering and Scientific Subroutine Library

The IBM Engineering and Scientific Subroutine Library for AIX (ESSL) family of products is a state-of-the-art collection of mathematical subroutines. Running on pSeries and RS/6000 nodes, the ESSL family provides a wide range of high-performance mathematical functions for a variety of scientific and engineering applications:

- ESSL contains over 400 high-performance mathematical subroutines tuned for IBM UNIX hardware.
- Parallel ESSL contains over 100 high-performance mathematical subroutines specifically designed to exploit the full power of parallel processing.

Parallel ESSL subroutines make it easier for developers, especially those not proficient in advanced parallel processing techniques, to create or convert applications to take advantage of the parallel processors of a Cluster 1600 system managed by PSSP.

Parallel ESSL accelerates applications by substituting comparable math subroutines and in-line code with high performance, highly-tuned subroutines. They are tuned for optimal performance on a system with SP Switch2 or SP Switch (16-port or 8-port). Both new and current numerically intensive applications written in Fortran, C, or C++ can call Parallel ESSL subroutines. New applications can be designed to take advantage of complete Parallel ESSL

capabilities. Existing applications can be enabled by replacing comparable routines and in-line code with calls to Parallel ESSL subroutines. Parallel ESSL supports the Single Program Multiple Data programming model and provides subroutines in six major areas of mathematical computations:

- Level 2 PBLAS
- Level 3 PBLAS
- Linear Algebraic Equations
- Eigensystem Analysis and Singular Value Analysis
- Fourier Transforms
- Random Number Generation

11.9 Related documentation

The concepts that need to be understood in this section are not the ones related to installation or configuration, but a general understanding of the functionality of these products is advised.

SP manuals

RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment, GA22-7281

PSSP Installation and Migration Guide, GA22-7347

PSSP Administration Guide, SA22-7348

IBM (e)server Cluster 1600: Planning, Installation, and Service, GA22-7863

PSSP Managing Shared Disks, SA22-7349

SP redbooks

RS/6000 SP Cluster: The Path to Universal Clustering, SG24-5374

Managing IBM (e)server Cluster 1600 - Power Recipes for PSSP 3.4, SG24-6603

IBM (e)server Cluster 1600 Managed by PSSP 3.5: What's New, SG24-6617

IBM Cluster 1600 and PSSP 3.4 Cluster Enhancements, SG24-6604

11.10 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, “Answers to sample questions” on page 521.

1. In planning for the use of LoadLeveler to run periodic batch jobs across several nodes, one requirement that is key to the use of LoadLeveler states that a flat UID namespace is required across all nodes in a LoadLeveler cluster. Why is this?
   a. LoadLeveler runs different jobs from a variety of client machines to a number of server machines in the defined LoadLeveler cluster and, due to standard UNIX security requirements, must be able to depend on the UID being consistent across all to nodes defined to the cluster.
   b. If such a namespace is not established, LoadLeveler will not be able to properly distinguish one UID from another, which may disrupt its capabilities for managing parallel jobs.
   c. LoadLeveler runs different jobs from a variety of client machines to a number of server machines in the defined LoadLeveler cluster and, due to standard hostname resolution differences between machines, depends on the /etc/hosts file being present even if DNS is implemented.
   d. A flat UID namespace is optional, but more efficient load-balancing can be achieved using this approach.
2. An HACWS environment requires which of the following to connect the two CWSs to the frame?
   a. An SCSI Target Mode Cable
   b. An additional Ethernet adapter for the frame supervisor card
   c. A Y-cable to link the two serial cables to the one port
   d. A null-modem cable
3. In a HACWS environment, if the primary control workstation fails, the backup CWS assumes all functions of the primary CWS. Which of the following functions is an exception to the previous statement?
   a. Authentication server
   b. Boot/install server
   c. Hardware monitoring
   d. Adding or changing SP users

4. Assuming the Working Collective is set to all nodes in the VSD cluster, which command would most satisfactorily determine whether the VSDs are up and running on all the VSD nodes?

   a. **dsh statvsd -a**

   b. **dsh lsvsd -l | pg**

   c. **dsh vsdatalst -a | pg**

   d. **SDRGetObjects VSD_Table CState==active**

5. You are in charge of installing, configuring, and starting a simple VSD configuration. Which of the following better describes the steps you will execute in order to get this done?

   a. 1) Create volume groups.
      2) Create logical volumes.
      3) Create virtual shared disks.
      4) Activate virtual shared disks.

   b. 1) Install the VSD and RSVD software.
      2) Designate the VSD nodes.
      3) Create virtual shared disks.
      4) Configure virtual shared disks.
      5) Start virtual shared disks.

   c. 1) Install the VSD and RSVD software.
      2) Designate the VSD nodes.
      3) Create virtual shared disks.
      4) Configure virtual shared disks.
      5) Prepare the virtual shared disks.
      6) Start virtual shared disks.

   d. 1) Install the VSD and RSVD software.
      2) Set authorization.
      3) Designate the VSD nodes.
      4) Create virtual shared disks.
      5) Configure virtual shared disks.
      6) Start virtual shared disks.

6. What is the definition of a GPFS node?

   a. It is the server node that provides token management.

   b. It is the node that has GPFS up and running.

   c. It is the node that provides the data disks for the file system.

   d. It is the server node that has GPFS and VSD up and running.

7. Which of the following attributes is **mmchconfig** capable of changing?

   a. MaxFiles to Cache

b. Quota Enforcement

c. Default Data Replication

d. Mount Point

8. Which of the following is *not* a striping algorithm implemented by GPFS?

   a. Round Robin

   b. Balanced Random

   c. Random

   d. Balanced Round Robin

9. Even though your installation does not have twin-tailed disks or SSA loops for multi-host disk connection, what does GPFS require?

   a. VSD

   b. RVSD

   c. NIS

   d. DNS

11.11 Exercises

Here are some exercises you may wish to perform:

1. Explore the work flow characteristics of LoadLeveler.
2. What are the components of the HACWS?
3. What are the necessary steps to create a VSD?
4. What decisions have to be made before creating a GPFS FS?
5. Familiarize yourself with the **`mmchconfig`** command. Which attribute is capable of changing?
6. Explain the capabilities of the **`mmfsck`** command.
7. Explore which FS attribute the **`mmchfs`** command is capable of changing.

Part 3

Application enablement

This part contains chapters for the planning and configuration of additional products that are present in most of the Cluster 1600 installations. This includes the IBM Virtual Shared Disk and the IBM Recoverable Virtual Shared Disk, as well as GPFS, and a section dedicated to problem management tools available in PSSP.

12

Problem management tools

This chapter provides an overview and several examples for problem management by using the tools available in PSSP. By problem management, we understand problem notification, log consolidation, and automatic recovery.

This chapter covers this by first giving an explanation of the technology used by all the problem management tools available in PSSP. It then describes two ways of using these tools and setting up monitors for critical components, such as memory, file system space, and daemons. The first method is using the command line interface through the problem management subsystem (PMAN); the second method is using the graphical user interface (Perspectives).

12.1 Key concepts

Before taking the exam, make sure you understand the following concepts:

- What is a resource monitor?
- What is the configuration data for the Event Management subsystem, and where is it stored?
- How to manage the Event Management daemons.
- How to get authorization to use the Problem Management subsystem.
- How to use the **pmandef** command.
- How to define conditions and events through SP Event Perspectives.

12.2 AIX service aids

Basically, every node (and the control workstation) is a pSeries AIX machine. This means that all the problem determination tools available for standard pSeries machines are also available for cluster nodes and CWSs.

AIX provides facilities and tools for error logging, system tracing, and system dumping (creation and analysis). Most of these facilities are included in the bos.rte fileset in AIX and, therefore, are installed on every node and control workstation automatically. However, some additional facilities, especially tools, are included in an optionally installable package called bos.sysmgt.serv_aid that should be installed in your nodes and control workstations.

12.2.1 Error logging facility

The AIX error logging facility records hardware and software failures or informational messages in the error log. All of the AIX and PSSP subsystems will use this facility to log error messages or information about changes to state information.

By analyzing this log, you can get an idea of what went wrong, when, and possibly why. However, due to the way information is presented by the **errpt** command, it makes it difficult to correlate errors within a single machine. This is much worse in the cluster where errors could be caused by components on different machines. We get back to this point later in this chapter.

The errdemon daemon keeps the log file updated based on information and errors logged by subsystems through the errlog or errsave facilities if they are running at kernel level. In any case, the errdemon daemon adds the entries in the error log on a first-come-first-serve basis.

This error log facility also provides a mechanism through which you could create a notification object for specific log entries. You could instruct the errdemon daemon to send you an e-mail every time there is a hardware error. The section “Using the AIX Error Log Notification Facility” in *PSSP Diagnosis Guide,* GA22-7350, provides excellent examples on setting up notification methods.

Log analysis is not bad. However, log monitoring is much better. You do not really want to go and check the error log on every node in your 128-node installation. Probably what you do is to create some notification objects in your nodes to instruct the errdemon daemon on those nodes to notify you in case of any critical error getting logged into the error log.

PSSP provides facilities for log monitoring and error notification. This differs from AIX notification in the sense that although it uses the AIX notification methods, it provides a global view of your system; so you could, for example, create a monitor for your AIX error log on all your nodes at once with a single command or a few clicks.

12.2.2 Trace facility

A trace facility is available through AIX. However, it comes in an optional fileset called bos.sysmgt.trace. Although the base system (bos.rte) includes minimal services for tracing, you need to install this optional component if you want to activate the trace daemon and generate trace reports.

If you get to the point where a trace is needed, it is probably because all the *conventional* methods have failed. Tracing is a serious business; it involves commitment and dedication to understand the trace report.

Tracing basically works in a two-step mode. You turn on the trace on selected subsystems and/or calls, and then you analyze the trace file through the report tools.

The events that can be included or excluded from the tracing facility are listed in the /usr/include/sys/trchkid.h header file. They are called *hooks* and *sub-hooks*. With these hooks, you can tell the tracing facility which specific event you want to trace. For example, you could generate a trace for all the CREAT calls that include file creations.

To learn more about tracing, refer to chapter 11, “Trace Facility,” in *AIX V4.3 Problem Solving Guide and Reference*, SC23-4123.

12.2.3 System dump facility

AIX generates a system dump when a severe error occurs. A system dump can also be user-initiated by users with root authority. A system dump creates a picture of your system's memory contents.

In AIX5L Version 5.2, the default location for the system dump is the paging space (hd6). It means that when the system is started up again, the dump needs to be moved to a different location. By default, the final location of a system dump is the /var/adm/ras directory, which implies that the /var file system should have enough free space to hold this dump. The size of the dump depends on your system memory and load. It can be obtained (without causing a system dump) by using the **sysdumpdev -e** command. This command only gives an estimate of the dump space required depending on the processes running at that time in the memory.

If there is not enough space in /var/adm/ras for copying the dump, the system will asks you what to do with this dump (throw it away, copy it to tape, and so on). This is changed for cluster nodes since they usually do not have people staring at the console because there is no console (at least not a physical console, *except for the HMCs*). The primary dump device is not hd6 but hd7 (a special dump device); so when the machine boots up, there is no need for moving the dump since the device is not being used for anything else. Although your nodes are running AIX5L v5.2, the primary dump device should be hd6 (paging space). The /etc/rc.sp script will change it back to /dev/hd7 on every boot.

A system dump certainly can help a lot in determining the cause. A good system dump in the right hands can point to the guilty component. Keep in mind that a system dump is a copy of selected areas of the kernel. These areas contain information about the processes and routines running at the moment of the crash. However, for the operating system, it is easier to keep this information in memory address format. So, for a good system dump analysis you will need the table of symbols that can be obtained from the operating system executable (/unix). Therefore, always save your system dumps along with the /unix corresponding to the operating system executable where the dump was produced. Support people will thank you.

For more information on AIX system dump, refer to chapter 12, "System Dump Facility," on page 81 of *AIX V4.3 Problem Solving Guide and Reference,* SC23-4123.

12.3 PSSP service aids

PSSP provides several tools for problem determination. Therefore, in this complex environment, you require company. The facilities that PSSP provides range from log files being present on every node and the CWS to Perspectives that utilize the Reliable Scalable Cluster Technology (RSCT).

12.3.1 SP log files

Besides errors and information being logged into the AIX error log, most of the PSSP subsystems write to their own log files where, usually, the information you need for problem isolation and problem determination resides.

Since some components run only on the CWS (such as the SDR daemon, the host respond daemon, the switch admin daemon, and so on), others run only on nodes (such as the switch daemon). This needs to be taken into consideration in the search for logs. *PSSP Diagnosis Guide,* GA22-7350, contains a complete list of PSSP log files and their location.

Unfortunately, there is not a common rule for analyzing log files. They are very specific to each component, and, in most cases, are created as internal debugging mechanisms and not for public consumption.

In this redbook, we cover some of these log files and explain how to read them. The only official logging information is the AIX error log. However, nothing is stopping you from reading these log files. As a matter of fact, these PSSP log files sometimes are essential for problem determination.

All the PSSP log files are located in the /var/adm/SPlogs directory. All the RSCT log files are located in the /var/ha/log directory. So, considering that these two locations reside on the /var file system, make sure you have enough free space for holding all the logged information. Refer to *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281, for details on disk space requirements.

12.4 Event Management

Event Management (EM) provides for comprehensive monitoring of hardware and software resources in the system. A resource is simply an entity in the system that provides a set of services. CPUs execute instructions, disks store data, and database subsystems enable applications. You define what system events are of interest to your application, register them with EM, and let EM efficiently monitor the system. Should the event occur, EM will notify your application. Figure 12-1 on page 400 illustrates EM's functional design.

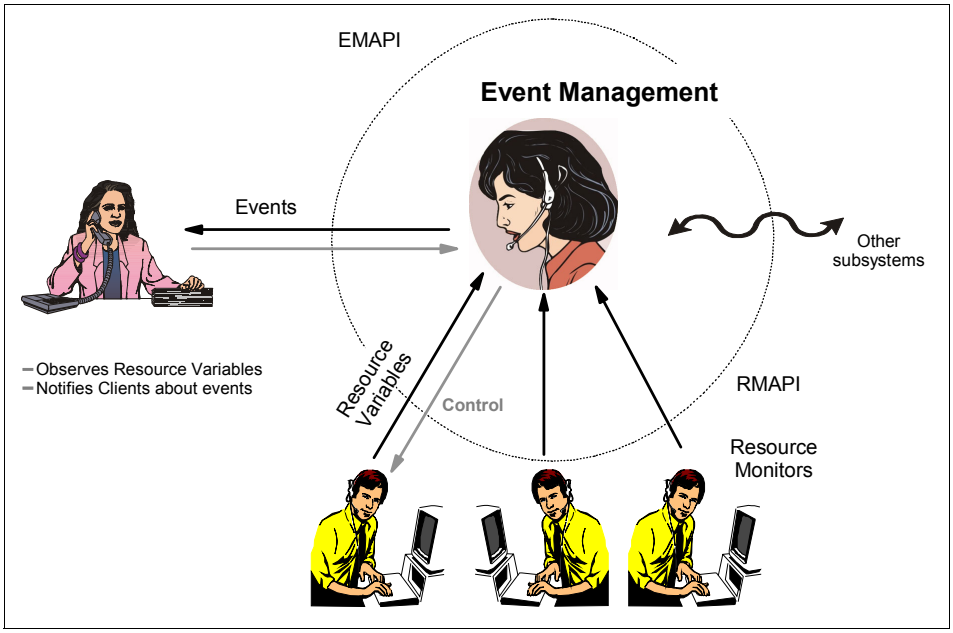


Figure 12-1 EM design

EM gathers information on system resources using Resource Monitors (RMs). RMs provide the actual data on system resources to the event-determining algorithms of EM. RMs are integral to EM, but how do RMs get their data? Data-gathering mechanisms vary according to platform (for example, sampling CPU data in an AIX environment is implemented completely different than in a Windows NT® environment). Cluster 1600-specific implementation of resource data-gathering mechanisms is described later.

EM is a distributed application, implemented by the EM daemon (haemd), running on each node and the CWS. Similar to Topology Services (TS) and Group Services (GS), EM is partition-sensitive; thus, the CWS may run multiple instances of haemd. To manage its distributed daemons, EM exploits GS. GS serves applications, such as EM. Because EM must communicate reliably among its daemons, it uses the Reliable Messaging information built from TS. This is shown in Figure 12-2 on page 401.

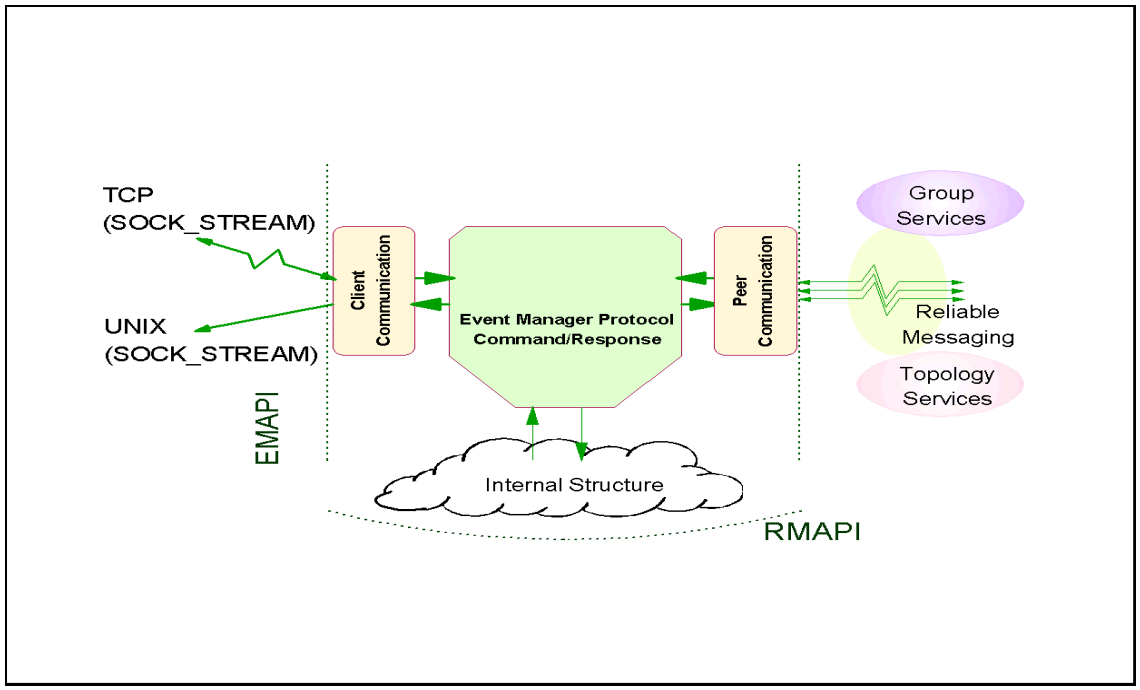


Figure 12-2 EM client and peer communication

EM receives resource data across the Resource Monitor Application Programming Interface (RMAPI). Clients communicate with EM through the Event Manager Application Programming Interface (EMAPI). An EM client can comprise many processes spread across nodes in a partition. A local process, that is, one executing on the same node as a given EM daemon, uses reliable UNIX domain sockets to talk to EM. On the CWS, a local process connects to the EM daemon that is running in the same system partition as the overall client. In this manner, the client can get events from anywhere in its partition.

To remote clients, that is, clients executing in a separate partition or outside the SP entirely, use TCP/IP sockets, which is a less reliable method because of the protocol that cannot always properly deal with crashed communication sessions between programs. Remote clients usually connect only to the EM daemon on the CWS. When connecting, a remote client specifies the name of the target partition on the call to the EMAPI. The remote client will then connect to the EM daemon on the CWS that is running in the target partition. A client could connect directly to any EM daemon in the target partition and get the same events, but you would need an algorithm to determine the target node. It is easier to just connect to the appropriate daemon on the CWS.

12.4.1 Resource monitors

Resource monitors are programs that observe the state of specific system resources and transform this state into several resource variables. The resource monitors periodically pass these variables to the Event Manager daemon. The Event Manager daemon then applies expressions, which have been specified by EM clients, to each resource variable. If the expression is true, an event is generated and sent to the appropriate EM client. EM clients may also query the Event Manager daemon for the current values of the resource variables.

12.4.2 Configuration files

Resource variables, resource monitors, and other related information, are specified in several System Data Repository (SDR) object classes. Information stored in these SDR classes is then translated into a binary form that can be easily used by the Event Management subsystem.

This EM database, call Event Management Configuration Database (EMCDB), is produced by the **haemcfg** command from the information in the SDR. The format of the EMCDB is designed to permit quick loading of the database by the Event Manager daemon and the Resource Monitor API (RMAPI). It also contains configuration data in an optimized format to minimize the amount of data that must be sent between the Event Manager daemons and between an Event Manager daemon and its resource monitors.

When the SDR data is compiled, the EMCDB is placed in a staging file. When the Event Manager daemon on a node or the control workstation initializes, it automatically copies the EMCDB from the staging file to a run-time file on the node or the control workstation. The run-time file is called /etc/ha/cfg/em.domain_name.cdb when domain_name is the system partition name.

Each time you execute the **haemcfg** command, or recreate the Event Management subsystem with the **syspar_ctrl** command, a new EMCDB file is created with a new version number. The new version number is stored in the Syspar SDR class, as shown in Figure 12-3.

```
[sp3en0:/]# SDRGetObjects Syspar haem_cdb_version
haem_cdb_version
913591595,334861568,0
```

Figure 12-3 EMCDB version stored in the syspar class

To check the version number of the run-time version, you can use the following commands:

From the CWS:

```
lssrc -ls haem.domain_name
```

From a node:

```
lssrc -ls haem
```

Because the Event Management subsystem is a distributed subsystem, all the Event Manager daemons have to use the same configuration information provided by the EMCDB. Using the same EMCDB version is vital.

The way in which Event Manager daemons determine the EMCDB version has important implications for the configuration of the system. To place a new version of the EMCDB into production (that is, to make it the runtime version), you must stop each Event Manager daemon in the domain after the **haemcfg** command is run. Stopping the daemons dissolves the existing peer group. Once the existing peer group is dissolved, the daemon can be restarted. To check whether the peer group has been dissolved, use the following command:

```
/usr/sbin/rsct/bin/hagsgr -s hags.domain_name | grep ha_em_peers
```

where *domain_name* is added only if the command runs on the control workstation. The output from these commands should be null.

Once the peer group is dissolved, the daemons can be restarted. As they restart, the daemons form a new peer group.

12.5 Problem management

The Problem Management subsystem (PMAN) is a facility used for problem determination, problem notification, and problem solving. It uses the RSCT infrastructure for monitoring conditions on behalf of authorized users and then generates actions accordingly.

The PMAN subsystem consists of three components:

- pmand

  This daemon interfaces directly with the Event Manager daemon to register conditions and to receive notifications. This daemon runs on every node and the control workstation, and it is partition-sensitive (the CWS may have more than one daemon running in case of multiple partitions).

- pmanrmd

  This is a resource monitor provided by PMAN to *feed* Event Management with 16 additional user-defined variables. You can program this resource monitor to periodically run a command or execute a script to update one of these

variables. Refer to “Monitoring a log file” on page 405 for an example of how to use this facility.

- sp_configd

  Through this daemon, PMAN can send Simple Network Management Protocol (SNMP) traps to SNMP managers to report predefined conditions.

12.5.1 Authorization

In order to use the Problem Management subsystem, users need to obtain a Kerberos principal, and this principal needs to be listed in the access control list (ACL) file for the PMAN subsystem. This ACL file is managed by the sysctl subsystem and is located in /etc/sysctl.pman.acl. The content of this file is as follows:

```
#acl#
# These are the kerberos principals for the users that can configure
# Problem Management on this node.  They must be of the form as indicated
# in the commented out records below.  The pound sign (#) is the comment
# character, and the underscore (_) is part of the "_PRINCIPAL" keyword,
# so do not delete the underscore.
#_PRINCIPAL root.admin@PPD.POK.IBM.COM
#_PRINCIPAL joeuser@PPD.POK.IBM.COM
_PRINCIPAL root.admin@MSC.ITSO.IBM.COM
```

In this case, the principal authorized to use the Problem Management subsystem is *root.admin* in the MSC.ITSO.IBM.COM realm.

Each time you make a change to this file, the sysctl subsystem must be refreshed. To refresh the sysctl subsystem, use the following command:

```
refresh -s sysctld
```

The **pmandef** command has a very particular syntax; so, if you want to give it a try, take a look at *PSSP Command and Technical Reference (2 Volumes),* SA22-7351, for a complete definition of this command. Chapter 26, “Using the Problem Management Subsystem,” of *PSSP Administration Guide,* SA22-7348, contains several examples and a complete explanation of how to use this facility.

Finally, the /usr/lpp/ssp/install/bin/pmandefaults script is an excellent starting point for using the PMAN subsystem. It has several examples of monitors for daemons, log files, file systems, and so forth.

Monitoring a log file

Now we know that the PMAN subsystem provides 16 resource variables for user-defined events. In this section, we use one of these variables to monitor a specific condition that is not monitored by default for the PSSP components.

Let us assume that you want to get a notification on the console's screen each time there is an authentication failure for remote execution. We know that the remote shell daemon (rshd) logs these errors to the /var/adm/SPlogs/SPdaemon.log log; so we can create a monitor for this specific error.

First, we need to identify the error that gets logged into this file every time somebody tries to execute a remote shell command without the corresponding credentials. Let us try and watch the error log file:

```
Feb 27 14:30:16 sp3n01 rshd[17144]: Failed krb5_compat_recvauth
Feb 27 14:30:16 sp3n01 rshd[17144]: Authentication failed from
sp3en0.msc.itso.ibm.com: A connection is ended by software.
```

From this content we see that `Authentication failed` seems to be a good string to look for. So, the idea here is to notify the operator (console) that there was a failed attempt to access this machine through the remote shell daemon.

Now, there is a small problem to solve. If we are going to check this log file every few minutes, how do we know if the log entry is new, or if it was already reported? Fortunately, the way user-defined resource variables work is based on strings. The standard output of the script you associate with a user-defined resource variable is stored as the value of that variable. This means that if we print out the last Authentication failed entry every time, the variable value will change only when there is a new entry in the log file.

Let's create the definition for a user-defined variable. To do this, PMAN needs a configuration file that has to be loaded to the SDR using the **pmanrmloadSDR** command.

PSSP provides a template for this configuration file known as *pmanrmd.conf*. It is located in the /spdata/sys1/pman directory on the CWS. Let us make a copy of this file and edit it:

```
TargetType=NODE_RANGE
Target=0-5
Rvar=IBM.PSSP.pm.User_state1
SampInt=60
Command=/usr/local/bin/Guard.pl
```

In this file, you can define all 16 user-defined variables (there must be one stanza per variable). In this case, we have defined the IBM.PSSP.pm.User_state1 resource variable. The resource monitor (pmanrmd) updates this variable every 60 seconds as specified in the sample interval (SampInt). The value of the variable will correspond to the standard output of the /usr/local/bin/Guard.pl script. Let us see what the script does.

```
#!/usr/lpp/ssp/perl5/bin/perl

my $logfile="/var/adm/SPlogs/SPdaemon.log";
my $lastentry;


open (LOG,"cat $logfile|") ||
        die "Ops! Can't open $logfile: $!\n";

while (<LOG>) {
   if(/Authentication failed/) {
        $lastentry = $_;
        }
   }

print "$lastentry";
```

The script printed out the `Authentication failed` entry from the log file. If there is no new entry, the old value will be the same as the new value; so, all we have to do is to create a monitor for this variable that gets notified every time the value of this variable changes. Let us take a look at the monitor's definition:

```
[sp5en0:/]# /usr/lpp/ssp/bin/pmandef -s authfailed \
-e 'IBM.PSSP.pm.User_state1:NodeNum=0-5:X@0!=X@P0' \
-c "/usr/local/bin/SaySomething.pl" \
-n 0
```

This command defines a monitor, through PMAN, for the IBM.PSSP.pm.User_state1 resource variable. The expression `X@0!=X@P0` means that if the previous value (`X@P0`) is different from the current value (`X@0`), then the variable has changed. The syntax for this variable is different because these user-defined variables are structured byte strings (SBS); so to access the value of this variable, you have to index this structure. However, these user-defined variables have only one field; so, only the index `0` is valid.

You can get a complete definition of this resource variable (and others) with the following command:

```
[sp5en0:/]# haemqvar "" IBM.PSSP.pm.User_state1 "*"|more
```

This command gives you a very good explanation along with examples of how to use it.

Now that we have subscribed our monitor, let us see what the /usr/local/bin/SaySomething.pl script does.

```
#!/usr/lpp/ssp/perl5/bin/perl

$cwsdisplay = "sp5en0:0";
$term="/usr/dt/bin/aixterm";
$cmd = "/usr/local/bin/SayItLoud.pl";
$title = qq/\"Warning on node $ENV{'PMAN_LOCATION'}\"/;
$msg = $ENV{'PMAN_RVFIELD0'};
$bg = "red";
$fg = "white";
$geo = "60x5+200+100";

$execute = qq/$term -display $cwsdisplay -T $title -geometry $geo -bg $bg -fg
$fg -e $cmd $msg/;

system($execute);
```

This script will open a warning window with a red background notifying the operator (it is run on node 0, the CWS) about the intruder.

The script /usr/local/bin/SayItLoud.pll displays the error log entry (the resource variable value) inside the warning window. Let's take a look at this script:

```
#!/usr/lpp/ssp/perl5/bin/perl

print "@ARGV\n";
print "------ Press Enter ------\n";
<STDIN>
```

Now that the monitor is active, let us try to access one of the nodes. We destroy our credentials (the **kdestroy** command), and then we try to execute a command on one of the nodes:

```
[sp5en0:/]# kdestroy
[sp5en0:/]# dsh -w sp5n01 date
sp5n01: spk4rsh: 0041-003 No tickets file found.  You need to run "k4init".
sp5n01: rshd: 0826-813 Permission is denied.
dsh:  5025-509 sp5n01 rsh had exit code 1
```

After a few seconds (a minute at most) we receive the warning window shown in the following warning message (Figure 12-4) at the CWS:

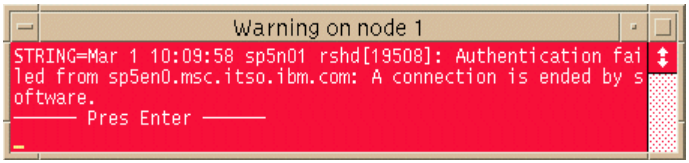


Figure 12-4 User-defined resource variables - Warning window example

The example shown here is very simple. It is not intended to be complete, but to show how to use these user-defined resource variables.

Information sent by the Problem Management subsystem in a notification can be logged into different repositories for further analysis. The **notify_event** script captures event information and mails it to the user running the command on the local node.

The **log_event** script captures event information and logs it to a wraparound file. The syntax for the **log_event** script is:

```
/usr/lpp/ssp/bin/log_event <log_filename>
```

The **log_event** script uses the AIX **alog** command to write to a wraparound file. The size of the wraparound file is limited to 64 K. The **alog** command must be used to read the file. Refer to the AIX alog man page for more information on this command.

12.6 Event Perspective

The SP Perspectives are a set of applications, each of which has a graphical interface (GUI), that enable you to perform monitoring and system management tasks for your SP system by directly manipulating icons that represent system objects.

Event Perspective is one of these applications. It provides a graphical interface to Event Management and the Problem Management subsystems.

Through this interface, you can create monitors for triggering events based on defined conditions and generate actions by using the Problem Management subsystem when any of these events is triggered.

12.6.1 Defining conditions

The procedure for creating monitors is straightforward. A condition needs to be defined prior to the creation of the monitor.

Conditions are based on resource variables, resource identifiers, and expressions, which, at the end, is what the Event Manager daemon evaluates.

To better illustrate this point, let's define a condition for a *file system full*. This condition will later be used in a monitor. The following steps are required for creating a condition:

1. Decide what you want to monitor. In this step, you need to narrow down the condition you want to monitor. For example: We want to monitor free space in the /tmp file system. Then we have to decide on the particular resource we want to monitor, and the condition. We should also think of where in the SP system we want to monitor free space in /tmp. Let us decide on that later.
2. Identify the resource variable. Once you have decided the condition you want to monitor, you need to find the variable that represents the particular resource associated to the condition. In our case, free space in a file system.

   PSSP provides some facilities to determine the right variable. This command provides you with information on how to use it. Let's use this command, **haemqvar**.

   We can use **haemqvar** to list all the variables related to file systems as follows:

```
[sp3en0:/]# haemqvar -d IBM.PSSP.aixos.FS "" "*"
IBM.PSSP.aixos.VG.free   Free space in volume group, MB.
IBM.PSSP.aixos.FS.%totused   Used space in percent.
IBM.PSSP.aixos.FS.%nodesused   Percent of file nodes that are used.
```

   In this case, we have listed the variables in the IBM.PSSP.aixos.FS class. You may use the same format to list other classes.

   In particular, we are interested in the IBM.PSSP.aixos.FS.%totused variable, which represents exactly what we want to monitor.

3. Define the expression. In order to define the expression we will use in our condition, we need to know how we use this variable. In other words, what are the resource identifiers for this variable. So let us use the **haemqvar** command again; but this time, let us query the specific variable and get a full description, as shown in Figure 12-5 on page 410.

```
[sp3en0:/]# haemqvar  "IBM.PSSP.aixos.FS" IBM.PSSP.aixos.FS.%totused "*"
Variable Name:  IBM.PSSP.aixos.FS.%totused
Value Type:     Quantity
Data Type:      float
Initial Value:  0.000000
Class:          IBM.PSSP.aixos.FS
Locator:        NodeNum
Variable Description:
    Used space in percent.

    IBM.PSSP.aixos.FS.%totused represents the percent of space in a file
    system that is in use. The resource variable's resource ID specifies
    the names of the ldescriptogical volume (LV) and volume group (VG) of the
file
    system, and the number of the node (NodeNum) on which the file system
    resides.
...lines not displayed...
   The lsvg command can be used to list, and display information about
    the volume groups defined on a node. For example:

      # lsvg | lsvg -i -l
      spdata:
      LV NAME     TYPE     LPs   PPs   PVs  LV STATE      MOUNT POINT
      spdatalv    jfs      450   450   1    open/syncd    /spdata
      loglv00     jfslog   1     1     1    open/syncd    N/A
      rootvg:
      LV NAME     TYPE     LPs   PPs   PVs  LV STATE      MOUNT POINT
      hd6         paging   64    64    1    open/syncd    N/A
      hd5         boot     1     1     1    closed/syncd  N/A
      hd8         jfslog   1     1     1    open/syncd    N/A
      hd4         jfs      18    18    1    open/syncd    /
      hd2         jfs      148   148   1    open/syncd    /usr
      hd9var      jfs      13    13    1    open/syncd    /var
      hd3         jfs      32    32    1    open/syncd    /tmp
      hd1         jfs      1     1     1    open/syncd    /home

...lines not displayed...
    When enough files have been created to use all the available
    i-nodes, no more files can be created, even if the file system
```

Figure 12-5 Resource variable query (partial view)

This command gives us a complete description of the variable and also tells us how to use it in an expression. Our expression would be: `X>90`.

We could use a rearm expression in our condition. A rearm expression is optional; it defines a second condition that Event Manager will switch to when

the main expression triggers. In our example, a rearm expression would be `X<60`. This means that after the file system is more than 90 percent used, Event Manager sends us a notification, and then continues monitoring the file system; but now it will send us a notification when the space used falls below 60 percent.

4. Create the condition. To do this, let us move the focus to the conditions panel on Event Perspective and then select **Actions** → **Create...** as shown in Figure 12-6.

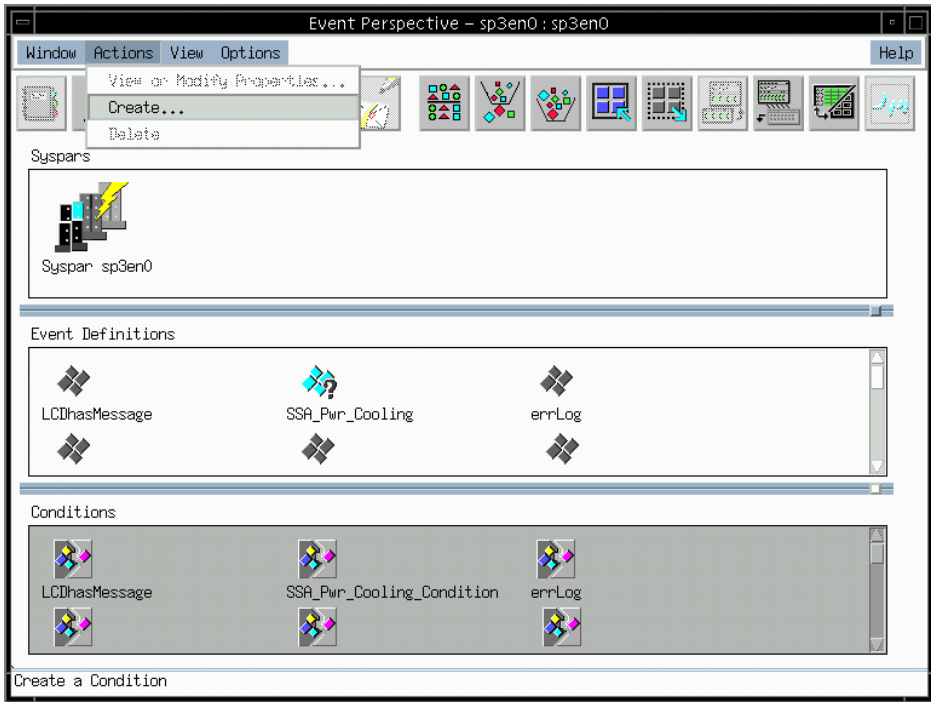


Figure 12-6 Create condition option from Event Perspectives

Once you click **Actions** → **Create...**, you will be presented with the Create Condition panel shown in Figure 12-7 on page 412.

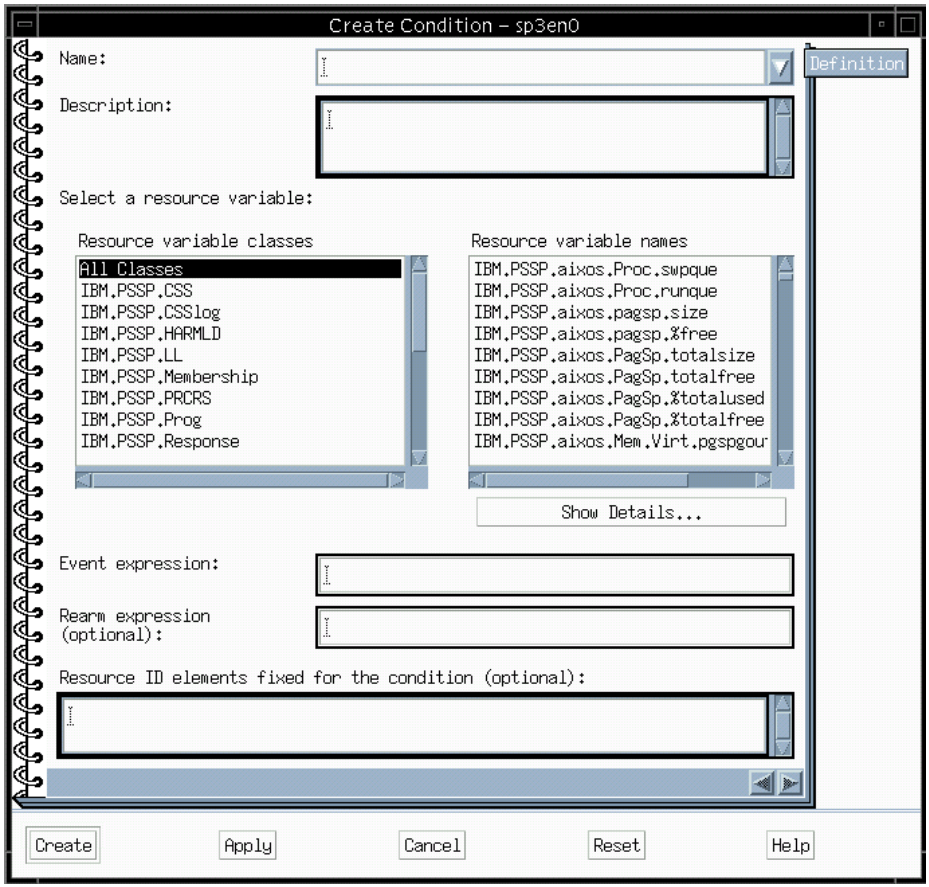


Figure 12-7 Create the Condition panel

As you can see in the Create Condition panel, there are two initial input boxes for the name (Name) of the condition and the description (Description). For our example, let's name the condition File_System_Getting_Full and give a brief description, such as `The file system you are monitoring is getting full. Better do something!` This is shown in Figure 12-8 on page 413.

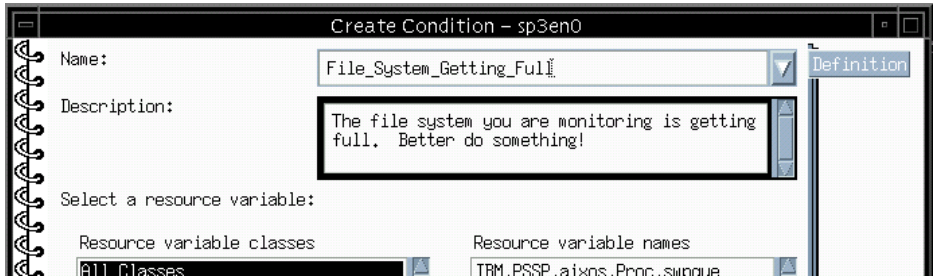


Figure 12-8 Defining name and description of a condition

Now we select the resource variable class (**IBM.PSSP.aixos.FS**) and the resource variable (**IBM.PSSP.aixos.FS.%totused**), followed by the expression, and then rearm the expression we defined in the previous step. This is shown in Figure 12-9.

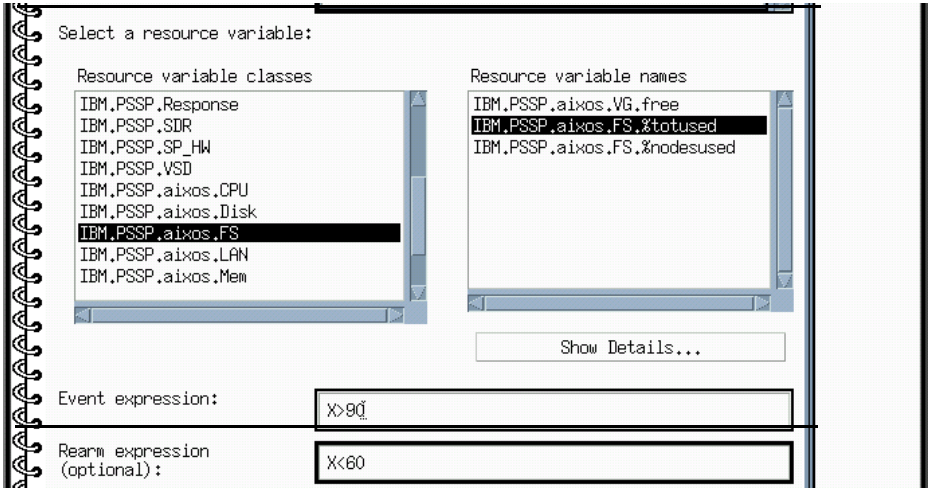


Figure 12-9 Selecting resource variable and defining expression

If you click **Show Details...**, it will present you the same output we got with the **haemqvar** command. We will leave the last input box empty, which represents the resources ID that you want to fix. For example, this resource variable (IBM.PSSP.aixos.FS.%totused) has two resource IDs. One is the volume group name (VG); the other is the logical volume name (LV). By using the last input box, we can fix one of the two resource IDs to a specific file system; this condition can be applied to that particular file system only. However, leaving this input blank enables us to use this condition in any monitor.

Once the condition has been created, an icon will appear in the Conditions panel as shown in Figure 12-10.

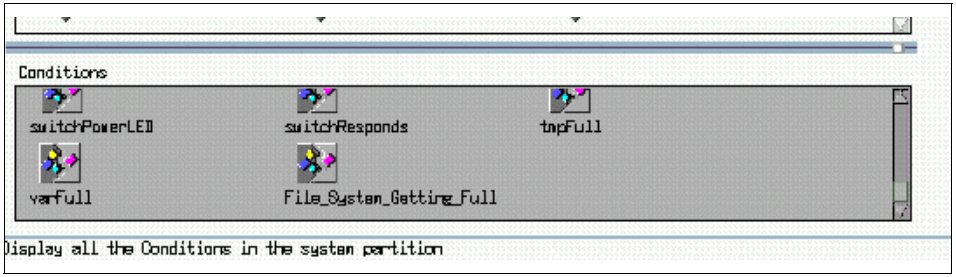


Figure 12-10 Conditions panel - New condition

12.7 Related documentation

This documentation will help you get more detailed information on the different topics covered in the chapter. Also, remember that good hands-on experience may reduce the amount of preparation for the IBM eServer Cluster 1600 Certification exam.

SP manuals

The only SP manual that can help you with this is the *PSSP Administration Guide,* SA22-7348. There is a section dedicated to availability and problem management. We recommend that you read at least chapter 26. Chapter 19 will help you understand the SP Perspectives.

SP redbooks

There are several books that cover the topics in this chapter. We recommend three of them.

- Chapters 2 and 3 of *RS/6000 SP Monitoring: Keeping it Alive,* SG24-4873, will give you a good understanding of the concepts involved.
- *Inside the RS/6000 SP,* SG24-5145 contains an excellent description of the Event Management and Problem Management subsystems.

12.8 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, “Answers to sample questions” on page 521.

1. The log_event utility provided with the Problem Management subsystem writes event information:
   a. To the SDR
   b. To the AIX error log
   c. To the /var/adm/SPlogs/pman/log directory
   d. To a wraparound file using the AIX **alog** command
2. The problem management subsystem (PMAN) requires Kerberos principals to be listed in its access control list file in order to function. Which file needs to be updated for getting access to PMAN functionality?
   a. /etc/sysctl.acl
   b. /etc/syscal.cmds.acl
   c. /etc/pman.acl
   d. /etc/sysctl.pman.acl
3. Which command would you use if you want to see a resource variable definition?
   a. `SDRGetObjects EM_Resource_Variable`
   b. `lssrc -ls haem.sp3en0 -a <variable name | *>`
   c. `haemqvar "<variable class | *>" "<variable name | *>" "<instance | *>"`
   d. `lsresvar -l <resource variable name>`
4. Although the base system (bos.rte) includes minimal services for tracing, which of the following optional filesets do you need to install if you want to activate the trace daemon and generate trace reports?
   a. bos.trace.sysmgt
   b. bos.rte.sysmgt
   c. bos.sysmgt.rte
   d. bos.sysmgt.trace
5. Where is the location of all the PSSP log files?
   a. /var/adm/logs
   b. /var/adm/SPlogs

c. /var/ha/logs

d. /var/SPlogs

6. Event Management (EM) provides comprehensive monitoring of hardware and software resources in the system. Which of the following does EM use to gather information on system resources?

   a. Event Management Monitors

   b. System Monitors

   c. Trace Monitors

   d. Resource Monitors

7. Which of the following is a PMAN subsystem component?

   a. sp_configd

   b. sp_configp

   c. spmand

   d. spmanrmd

8. What is the correct order for defining a condition with SP Event Perspectives?

   a. Decide what you want to monitor, identify the resource variable, define the expression, and create the condition.

   b. Decide what you want to monitor, define the expression, identify the resource variable, and create the condition.

   c. Define the expression, decide what you want to monitor, identify the resource variable, and create the condition.

   d. Define the expression, create the condition, decide what you want to monitor, and identify the resource variable.

12.9 Exercises

Here are some exercises you may wish to perform:

1. On a test system that does not affect any users, define a condition or event to monitor using Event Perspectives.
2. What does the PMAN facility provide? On a test system that does not affect any users, set up and test a monitor that will send a notification to the console's screen each time there is an authentication failure for remote execution.

Part 4

On-going support

This part contains chapters dedicated to software maintenance, system reconfiguration including migration, and problem determination procedures and checklists.

13

PSSP software maintenance

This chapter discusses how to maintain backup images for the CWS and Cluster 1600 nodes, as well as how to recover the images you created. In addition, we discuss how to apply the latest PTFs for AIX and PSSP. We provide the technical steps for information based on the environment we set at the beginning of this book. Finally, we discuss the overview of software migration and coexistence.

13.1 Key concepts

This section discusses key concepts on how to maintain the software on the SP. You should understand the following:

- How to create and manage backup images for CWS and the cluster nodes.
- How to restore CWS or the nodes, and what are the necessary procedures after restoring.
- How to apply the PTFs and what the required tasks are for AIX and PSSP on CWS and the nodes.
- What are the effects of the PTFs you applied on your PSSP system?
- The concept of software migration and coexistence in supported environments.
- What are the changes made from PSSP V2 to PSSP V3?

13.2 Backup of the CWS and cluster node images

Maintaining a good copy of backup images is as important as initial implementation of your PSSP system. Here we discuss how to maintain the CWS backup image and how to efficiently create cluster node images with a scenario we set up in our environment.

13.2.1 Backup of the CWS

The backup of the CWS is the same as the strategy you use for stand-alone servers because it has its own tape device to use for backup. In AIX, we usually back up the system with the command: **mksysb -i <device_name>**.

Remember that the **mksysb** command backs up *only* rootvg data. Thus, data other than rootvg should be backed up with the command **savevg** or another backup utility, such as sysback.

13.2.2 Backup of the cluster node images

In scientific or parallel computing environments, we may only need one copy of node images across the PSSP complex because, in most cases, all node images are identical. However, in commercial or server consolidation environments, we usually maintain a separate copy of a node image per application or even per node in the cluster. Therefore, you need to understand your environment and set up the cluster node backup strategy.

In general, it is recommended that you keep the size of the node's image as small as possible so that you can recover images quickly and manage the disk space needed. It is also recommended that user data should be separate from rootvg so that you can maintain a manageable size of node images. Here, the node image is the operating system image, not the user data image. For user data, you should consider another strategy, such as TSM, for backup.

Also, remember that the node image you create is a file and is not bootable, so you should follow the network boot process, as discussed in 9.2.21, "Network boot the boot/install server and nodes" on page 330, to restore it.

Depending upon your environment, there are many ways you can set up a cluster node backup. Here we introduce the way we set it up in our environment.

13.2.3 Case scenario: How do we set up node backup?

In our environment, we set up sp3n01 as the boot/install server. Thus, we created the same /spdata directory structure as CWS. Assuming that all nodes have different images, we needed to create individual node images. We NFS-mounted the boot/install server node's /spdata/sys1/install/images directory to all nodes and the CWS's /spdata/sys1/install/images directory to the boot/install server node. We then ran

```
mksysb -i /<mount_point>/bos.obj.<hostname>.image
```

on all nodes, including the boot/install server node. In this way, all node images were created on each /spdata/sys1/install/images directory, as shown in Figure 13-1 on page 422.

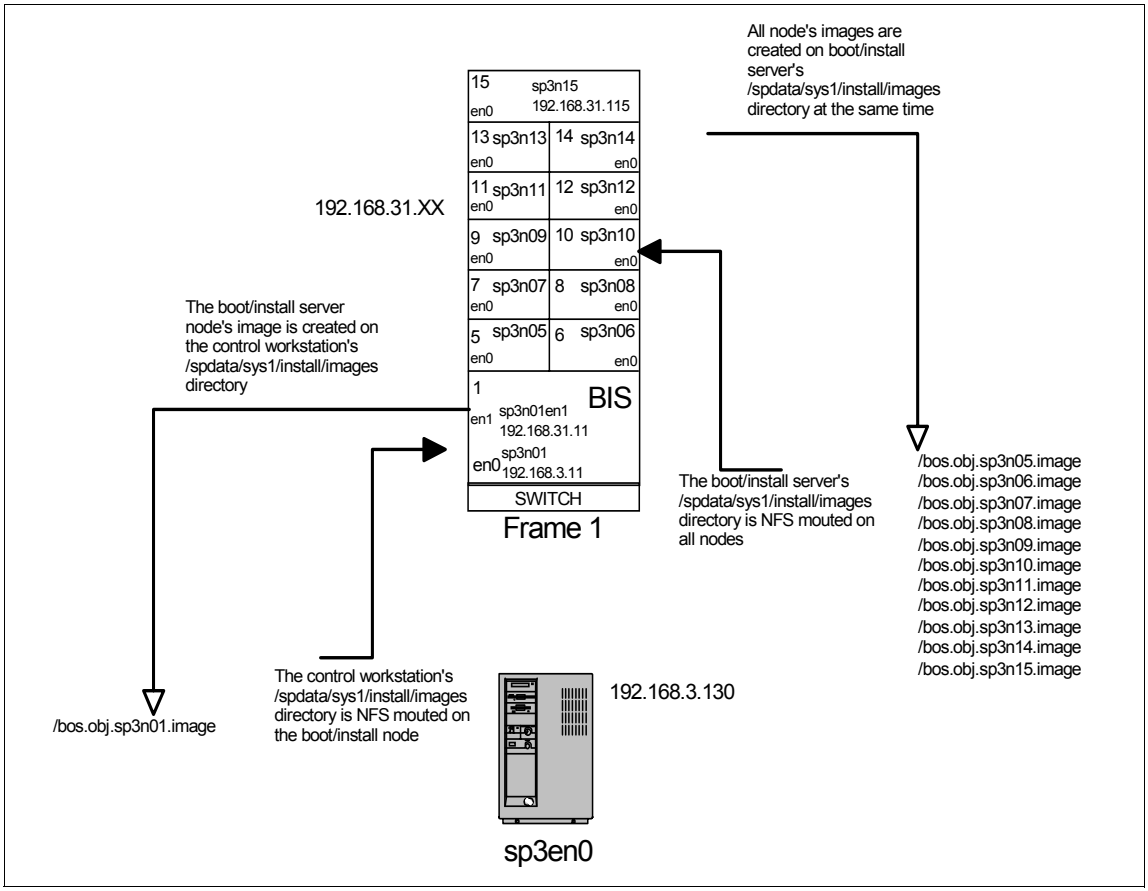


Figure 13-1 Mechanism of SP node backup in boot/install server environment

Of course, you can write scripts to automate this process. Due to the nature of this book, we only introduce the mechanism for the node backup strategy.

13.3 Restoring from mksysb image

In the following sections, we discuss the recovery of the CWS and nodes when a system has crashed.

13.3.1 Restoring the CWS

You may have problems when you do software maintenance. Here, we discuss how you can recover the CWS from a recent backup tape you created. Restoring

the CWS is similar to recovering any pSeries server except that you need some post activity.

To restore an image of the CWS, do the following:

1. Execute the normal procedure used to restore any pSeries server.
2. Issue the **/usr/lpp/ssp/bin/install_cw** command.

   When an mksysb image is made from an existing CWS, there are certain ODM attributes that are not saved, such as node_number information. This script creates the proper node_number entry for the CWS in the ODM. It also executes some other functions, as explained in 8.3.4, “install_cw” on page 279.
3. Verify your CWS as explained in 8.8, “Configuring and verifying the CWS” on page 296.

13.3.2 Restoring the node

The procedure that is used to restore the mksysb image to a node is similar to the installation process using NIM. You have to change some parameters in the original environment.

The first step is to put the image that you want to restore in the /spdata/sys1/install/images directory. Then you have to change the network environment for that node. To do this, enter the following commands:

```
# spchvgobj -r rootvg -i <image name> -l <node_number>
# spbootins -r install -l <node_number>
```

As an example, to restore node 5 with an image called image.sp3n05:

```
# spchvgobj -r rootvg -i bos.obj.sp3n05.image -l 5
# spbootins -r install -l 5
```

You can verify the environment with the following command:

```
# splstdata -b -l 5
```

Check the fields `response` and `next_install_image`.

Now network-boot the node to restore the correct image. You can do this in another node, different from the original, without worrying about the node number and its specific configuration. After the node is installed, **pssp_script** customizes it with the correct information.

13.4 Applying the latest AIX and PSSP PTFs

This section is to be used for applying Program Temporary Fixes (PTFs) for AIX, PSSP, and other Licensed Program Products (LPPs) in the SP.

13.4.1 On the CWS

This section briefly describes how to apply AIX and PSSP PTFs on the CWS.

Applying AIX PTFs

Here are the steps for applying AIX PTFs:

1. Create an mksysb backup image of the CWS.
2. Check that the tape is OK by listing its contents with the command:

   ```
   smitty lsmksysb
   ```

3. Copy the PTFs to the lppsource directory /spdata/sys1/install/aix52/lppsource/installp/ppc.
4. Create a new .toc file by executing the commands:

   ```
   # cd /spdata/sys1/install/aix52/lppsource/installp/ppc
   # inutoc
   ```

5. Update the new PTFs to the CWS using SMIT:

   ```
   # smitty update_all
   ```

6. Then update the SPOT with the PTFs in the lppsource directory using the command:

   ```
   # smitty nim_res_op
   ```

   with the following as input to the menu:

   Resource name: `spot_aix52`

   Network Install Operation to perform: `update_all`

If the status of the installation is OK, then you are done with the update of the AIX PTFs on the CWS.

Applying PSSP PTFs

The steps for applying PSSP PTFs are as follows:

1. Create an mksysb backup image of the CWS. Always check that the tape is OK by listing its contents with the command:

   ```
   smitty lsmksysb
   ```

2. Copy the PTFs to the directory /spdata/sys1/install/pssplpp/PSSP-3.5 for PSSP 3.5.
3. Create a new .toc file by issuing the following commands:

```
# cd /spdata/sys1/install/pssplpp/PSSP-3.5
# inutoc
```

4. Check the READ THIS FIRST paper that comes with any updates to the PSSP and the .info files for the prerequisites, corequisites, and any precautions that need to be taken for installing these PTFs. Check the filesets in the directory you copied to make sure that all the required filesets are available.
5. Update the new PTFs to the CWS using:

```
# smitty update_all
```

Note: In many cases, the latest PSSP PTFs include the microcode for the supervisor card. We strongly recommend that you check the state of the supervisor card after applying the PSSP PTFs. Be sure to have the latest software level on the HMC.

13.4.2 To the nodes

There are many ways to install PTFs on the nodes. If you have a server consolidation environment and different filesets installed on each node, it will be difficult to create one script to apply the PTFs to all the nodes at once. However, here we assume that we have installed the same AIX filesets on all the nodes. Thus, we apply the PTFs to one test node, create a script, and then apply the PTFs to the rest of the nodes.

Important: *Before* you apply the latest PTFs to the nodes, make sure you apply the same level of PTFs on the CWS and boot/install server nodes.

Applying AIX PTFs

This method is to be used for installing the PTFs on a node by using the SMIT and **dsh** commands.

For any of the options you choose, it is better to install the PTFs on one node and do the testing before applying them to all the nodes. In our scenario, we selected sp3n01 as the test node.

1. Log in as root and mount the lppsource directory of the CWS in sp3n01 with the command:

```
# mount sp3en0:/spdata/sys1/install/aix52/lppsource/installp/ppc \ /mnt
```

2. Apply the PTFs using the command:

   ```
   # smitty update_all
   ```

   First, run this with the `PREVIEW only` option set to `yes` and check that all prerequisites are met. If it is OK, then go ahead and install the PTFs with the `PREVIEW only` option changed back to `no`.

3. Unmount the directory you had mounted in step1 using the command:

   ```
   # umount /mnt
   ```

4. If everything runs OK on the test node, then prepare the script from the /smit.script file for the rest of the nodes. As an example, you may create the following script:

```
#/use/bin/ksh!
# Name of the Script:ptfinst.ksh
#
mount sp3en0:/spdata/sys1/install/aix52/lppsource /mnt
/usr/lib/instl/sm_inst installp_cmd -a -d '/mnt' -f '_update_all' '-c' '-N'
'-g' '-X'
umount /mnt
```

5. Change the file mode to executable and owned by the root user:

   ```
   # chmod 744 /tmp/ptfinst.ksh
   # chown root.system /tmp/ptfinst.ksh
   ```

6. Copy to the rest of the nodes with the command:

   ```
   # hostlist | pcp -w - /tmp/ptfinst.ksh /tmp
   ```

7. Execute the script using **dsh** except on the test node.

While installing the PTFs, if you get any output saying that a reboot is required for the PTFs to take effect, you should reboot the node. If you have a switch, then before rebooting a node, you may need to fence it using the command:

```
# Efence -autojoin sp3n01
```

Applying PSSP PTFs

Applying PSSP PTFs to the nodes can be done with the same methods we used for applying AIX PTFs. Before applying the PTFs, make a backup image for the node.

Follow the same procedure except for step 1; you need to mount the PSSP PTFs directory instead of the lppsource directory. The command is:

```
# mount sp3en0:/spdata/sys1/install/pssplpp/PSSP-3.5 /mnt
```

When updating the ssp.css fileset of PSSP, you must reboot the nodes for the kernel extensions to take effect.

We recommend that you make another backup image after you have applied the PTFs.

13.5 Software migration and coexistence

In earlier chapters, we discussed what is available in AIX and PSSP software levels. This section discusses the main changes driven by PSSP 3.5 when you migrate your system from PSSP 3.1 and AIX 4.3.2.

Because migration of your CWS, your nodes, or both, is a complex task, you must do careful planning before you attempt to migrate. Thus, a full migration plan involves breaking your migration tasks down into distinct, verifiable (and recoverable) steps, and planning of the requirements for each step. A well-planned migration has the added benefit of minimizing system downtime.

13.5.1 Migration terminology

An AIX level is defined as <Version>.<Release>.<Modification>. A migration is a process of changing to a newer version or release, while an update is a process of changing to a new modification level. In other words, if you change the AIX level from 5.1 to 5.2, it is a migration, while if you change the AIX level from 5.2 ML-01 to 5.2 ML-02, it is an update. However, all PSSP level changes are updates.

13.5.2 Supported migration paths

In PSSP 3.5, the only supported paths are those shown in Table 13-1 on page 428. If your current system, CWS, or any node is running at a PSSP or AIX level not listed in the From column of Table 13-1 on page 428, you must update to one of the listed combinations before you can migrate to PSSP 3.5. Refer to *PSSP Installation and Migration Guide,* GA22-7347, for migration procedure details, and *IBM (e)server Cluster 1600 Managed by PSSP 3.5: What's New,* SG24-6617, page 130, for supported migration paths.

Table 13-1 Supported migration paths from PSSP 3.1

From PSSP Level	From AIX Level	To PSSP Level	To AIX Level
2.2	4.1.5 4.2.1	3.1	4.3.2
2.3	4.2.1 4.3.2	3.1	4.3.2
2.4	4.2.1 4.3.2	3.1	4.3.2

You can migrate the AIX level and update the PSSP levels at the same time. However, we recommend that you migrate the AIX level first without changing the PSSP level and verify system stability and functionality. Then, update the PSSP.

However, even if you have found your migration path, some products or components of PSSP have limitations that might restrict your ability to migrate:

- Switch Management
- RS/6000 Cluster Technology
- Performance Toolbox Parallel Extensions
- High Availability Cluster Multi-Processing
- IBM Virtual Shared Disk
- IBM Recoverable Virtual Shared Disk
- General Parallel File System
- Parallel Environment
- LoadLeveler
- Parallel Tools
- PIOFS, CLIO/S, and NetTAPE
- Extension node support

For more information about these limitations, refer to *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281.

13.5.3 Migration planning

In many cases, we recommend the migration rather than a new install because the migration preserves all local system changes you have made, such as:

- Users and groups (the settings for the users, such as passwords, profiles, and login shells)
- File systems and volume groups (where names, parameters, sizes, and directories are kept)
- PSSP setup (AMD, File Collections)
- Network setup (TCP/IP, SNA)

Before migrating, you may want to create one or more system partitions. As an option, you can create a production system partition with your current AIX and PSSP level software and a test system partition with your target level of AIX and PSSP 3.5 level software.

Before you migrate any of your nodes, you *must* migrate your CWS and boot/install server node to the latest level of AIX and PSSP of any node you wish to serve. After these general considerations, we now give some details of the migration process at the CWS level and then at the node level.

13.5.4 Overview of a CWS PSSP update

This section briefly describes what is new in PSSP 3.1 for updating the CWS. For further information, refer to *PSSP Installation and Migration Guide,* GA22-7347.

We describe the main steps in the installation process, but with the migration goal in mind. We also assume the migration of the CWS to AIX 4.3.2 has been done successfully.

1. Create the required /spdata directory, such as /spdata/sys1/install/aix432/lppsource and /spdata/sys1/install/pssplpp/PSSP-3.1.

   ```
   # mkdir -p /spdata/sys1/install/aix432/lppsource
   # mkdir -p /spdata/sys1/install/pssplpp
   ```

2. Copy the AIX LPP images and others required for AIX LPPs from AIX 432 media to /spdata/sys1/install/aix432/lppsource on the CWS.
3. Verify the correct level of PAIDE (perfagent).

   The perfagent.server fileset must be installed and copied to all of the lppsource directories on CWS of any SP that has one or more nodes at PSSP 2.4 or earlier.

   The perfagent.tools fileset is part of AIX 4.3.2. This product provides the capability to monitor the performance of your SP system, collects and displays statistical data for SP hardware and software, and simplifies runtime performance monitoring of a large number of nodes. This fileset must be

installed and copied to all of the lppsource directories on CWS of any SP that has one or more nodes at PSSP 3.1.

4. Copy the PSSP images for PSSP 3.1 into the /spdata/sys1/install/pssplpp/PSSP-3.1 directory, rename the PSSP package to pssp.installp and create the .toc file:

```
# bffcreate -qvx -t /spdata/sys1/install/pssplpp/PSSP-3.1 -d /dev/rmt0 all
# cd /spdata/sys1/install/pssplpp/PSSP-3.1
# mv ssp.usr.3.1.0.0 pssp.installp
# inutoc .
```

5. Copy an installable image (mksysb format) for the node into /spdata/sys1/install/images.

6. Stop the daemons on the CWS and verify.

```
# syspar_ctrl -G -k
# stopsrc -s sysctld
# /etc/amd/amq (PSSP 2.2 users only)(see note)
# stopsrc -s splogd
# stopsrc -s hardmon
# stopsrc -g sdr
```

   Issue the **lssrc -a** command to verify that the daemons are no longer running on the CWS.

7. Install PSSP on the CWS.

   The PSSP 3.1 filesets are packaged to be installed on top of previously supported releases. You may install all filesets available or minimum filesets in the PSSP 3.1 package.

   To properly set up the PSSP 3.1 on the CWS for the SDR, hardmon, and other SP-related services, issue the following command:

```
# install_cw
```

8. Update the state of the supervisor microcode.

   Check which supervisors need to be updated by using SMIT panels or by issuing the **spsvrmgr** command:

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

   In case an action is required, you can update the microcode by issuing the command:

```
# spsvrmgr -G -u <frame_number>:<slot_number>
```

9. Refresh all the partition-sensitive subsystem daemons.

10. Migrate shared disks.

    If you already use Virtual Shared Disk (VSD), you have some preparation to do.

13.5.5 Overview of node migration

You cannot migrate the nodes until you have migrated the CWS and boot/install servers to your target AIX level (4.3.2) and PSSP 3.1. You can migrate the nodes to your AIX level and PSSP 3.1 in one of three ways:

- Migration install

  This method preserves all the file systems except /tmp as well as the root volume group, logical volumes, and system configuration files. This method requires the setup of AIX NIM on the new PSSP 3.1 CWS and boot/install servers. This applies only to migrations when an AIX version or release is changing.

- mksysb install

  This method erases all existence of current rootvg and installs your target AIX level and PSSP 3.1 using an AIX 4.3.2 mksysb image for the node. This installation requires the setup of AIX NIM on the new PSSP 3.1 CWS or boot/install servers.

- Upgrade

  This method preserves all occurrences of the current rootvg and installs AIX PTF updates using the `installp` command. This method applies to AIX modification level changes or when the AIX level is not changing, but you are updating to a new level of PSSP.

To identify the appropriate method, you must use the information in Table 8 in *PSSP Installation and Migration Guide,* GA22-7347.

Although the way to migrate a node has not changed with PSSP 3.1, we point out here how the PSSP 3.1 enhancements can be used when you want to migrate.

1. Migration install of nodes to PSSP 3.1

   Set the bootp_response parameter to migrate for the node you migrate with the new PSSP 3.1 commands (**`spchvgobj`** and **`spbootins`**).

   If we migrate the nodes 5 and 6 from AIX4.2.1 and PSSP 2.4 to AIX 4.3.2 and PSSP 3.1, we issue the following commands assuming the lppsource name directory is /spdata/sys1/install/aix432/lppsource:

   ```
   # spchvgobj -r rootvg -p PSSP-3.1 -v aix432 -l 5,6
   # spbootins -r migrate -l 5,6
   ```

   The SDR is now updated and setup_server will be executed. Verify this with the command:

   ```
   splstdata -G -b -l <node_list>
   ```

Finally, a shutdown followed by a network boot will migrate the node. The AIX part will be done by NIM, whereas the script **pssp_script** does the PSSP part.

2. mksysb installation of nodes

   This is the node installation we discussed in Chapter 9, “Frame and node installation” on page 301.

3. Update to a new level of PSSP and to a new modification level of AIX.

   If you are on AIX 4.3.1 and PSSP 2.4 and you want to go to AIX 4.3.2 and PSSP 3.1, you must first update the AIX level of the node by mounting the aix432 lppsource directory from the CWS on your node and running the **installp** command.

   Then, after you have the right AIX level installed on your node, set the bootp_response parameter to customize with the new PSSP 3.1 commands (**spchvgobj** and **spbootins**) for nodes 5 and 6:

   ```
   # spchvgobj -r rootvg -p PSSP-3.1 -v aix432 -l 5,6
   # spbootins -r customize -l 5,6
   ```

   Then copy the pssp_script file from the CWS to the node:

   ```
   # pcp -w <node> /spdata/sys1/install/pssp/pssp_script \
   /tmp/pssp_script
   ```

   After the copy is done, execute the pssp_script that updates the node’s PSSP to the new PSSP 3.1 level.

13.5.6 Coexistence

PSSP 3.1 can coexist with PSSP 2.2 and later. Coexistence is the ability to have multiple levels of AIX and PSSP in the same partition.

Table 13-2 shows what AIX levels and PSSP levels are supported by PSSP 3.1 in the same partition. Any combination of PSSP levels listed in this table can coexist in a system partition. So, you can migrate to a new level of PSSP or AIX one node at a time.

Table 13-2 Possible AIX or PSSP combinations in a partition

AIX Levels	PSSP Levels
AIX 4.1.5 or AIX 4.2.1	PSSP 2.2
AIX 4.2.1 or AIX 4.3.2	PSSP 2.3
AIX 4.2.1 or AIX 4.3.2	PSSP 2.4
AIX 4.3.2	PSSP 3.1

Some PSSP components and related LPPs still have some limitations. Also, many software products have PSSP and AIX dependencies.

13.6 Related documentation

This study guide only provides key points; so, it is recommended that you review the following reference books for details.

SP manuals

Refer to *PSSP Installation and Migration Guide,* GA22-7347. For details on how to boot from the mksysb tape, read the *AIX V4.3 Quick Install and Startup Guide*, SC23-4111.

SP redbooks

RS/6000 SP Software Maintenance, SG24-5160 provides everything you need for software maintenance. We strongly recommend that you read this for real production work. For the sections on backup and PTFs, refer to chapters 7 and 8. For the section on software migration, read chapter 2 of *PSSP 3.1 Announcement,* SG24-5332.

13.7 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, “Answers to sample questions” on page 521.

1. You have applied the latest PSSP fixes to the CWS. A message posted during fix installation states that a microcode update for high nodes is included in this fix. You query the status of your high node supervisor microcode and get the following output:

```
Frame  Slot  Supervisor  Media        Installed     Required
                                 State       Versions      Version
Action
                   _____  ____  _________  ____________  ____________
____________
                   1      9     Active     u_10.3a.0612  u_10.3a.0614
Upgrade
                                            u_10.3a.0614
                                            u_10.3a.0615
```

   What command is used to update the supervisor microcode on the high nodes?

a. **spucode**

b. **spsvrmgr**

c. **spmicrocode**

d. **sphardware**

2. You have applied the latest PSSP fixes to the CWS. What is a recommended task to perform?

   a. Check the state of all supervisor microcode.

   b. Delete and re-add all system partition-sensitive daemons.

   c. Stop and restart the NTP daemon on all nodes.

   d. Remove and reacquire the administrative Kerberos ticket.

3. Which of the following is a supported migration path?

   a. AIX 3.2.5/PSSP 2.1 → AIX 4.2.1/PSSP 3.1

   b. AIX 4.2.1/PSSP 2.4 → AIX 4.3.2/PSSP 3.1

   c. AIX 4.1.4/PSSP 2.1 → AIX 4.3.2/PSSP 2.2

   d. AIX 4.1.5/PSSP 2.3 → AIX 4.1.5/PSSP 2.4

4. Which of the following is *not* a supported migration path?

   a. AIX 4.2.1/PSSP 2.2 → AIX 4.3.2/PSSP 3.1

   b. AIX 4.3.2/PSSP 2.4 → AIX 4.3.2/PSSP 3.1

   c. AIX 4.1.5/PSSP 2.2 → AIX 4.3.2/PSSP 2.2

   d. AIX 4.1.5/PSSP 2.2 → AIX 4.2.1/PSSP 2.3

5. Which of the following commands is part of the procedures to restore an image of the CWS?

   a. spbootins -r install -1 <mksysb image name> -l

   b. /usr/lpp/ssp/bin/install_cw

   c. mksysb -i /<mount_point>/bos.obj.<hostname>.image

   d. spbootins -r install -l <node_number>

13.8 Exercises

Here are some exercises you may wish to do:

1. On a test system that does not affect any users, perform a CWS backup, and a node backup.
2. Apply the latest AIX and PSSP PTFs to the CWS.

3. Apply the latest AIX and PSSP PTFs to a node.
4. Perform another backup for the CWS and the node.
5. Migrate the CWS to AIX 4.3.2. Then update the CWS PSSP level to PSSP 3.1.
6. After performing exercise 4, migrate the node to AIX 4.3.2 and PSSP 3.1.
7. Restore the CWS to its original state. Then restore the node to its original state.

14

RS/6000 SP reconfiguration and update

Most commercial installations start with a small number of nodes and expand their environment as time goes by or new technology becomes available. In Chapter 7, “User and data management” on page 249 and Chapter 8, “Configuring the control workstation” on page 275 we discussed the key commands and files used for initial implementation based on our environment.

In this chapter, we discuss the procedures used to reconfigure an SP system, such as adding frames, nodes, and switches, which are the most frequent activities you may face. Then we describe the required activities used to replace an existing MCA-based uniprocessor node to PCI-based 332 MHz SMP node.

14.1 Key concepts

This section gives the key concepts for the certification exam on reconfiguration and migration of RS/6000 SP. You should understand the following:

- The types of SP nodes and what the differences are among the nodes.
- What the procedures are when you add new frames or SP nodes, as well as the software and hardware requirements.
- How to reconfigure the boot/install server when you set up a multiframe environment.
- How to replace existing MCA-based uniprocessor nodes or SMP nodes to the new PCI-based 332 MHz SMP node along with its software and hardware requirements and procedures.
- The technology updates on PSSP V3.

14.2 Environment

This section describes the environment for our SP system. From the initial SP system, we added a second switched frame and added one high node, four thin nodes, two Silver nodes, and three wide nodes, as shown in Figure 14-1 on page 439.

In Figure 14-1, sp3n17 is set up as the boot/install server. The Ethernet adapter (en0) of sp3n17 is cabled to the same segment (subnet 3) of the en0 of sp3n01 and CWS. The en0's of the rest of the nodes in frame 2 are cabled with the en1 of sp3n17 so that they will be in the same segment (subnet 32).

Thus, we install sp3n17, which is a boot/install server, first from CWS. Then we install the rest of the node from sp3n17. In the following sections, we summarize the steps for adding frames, nodes, and SP switches from *PSSP Installation and Migration Guide,* GA22-7347, even though the physical installation was done at the same time.

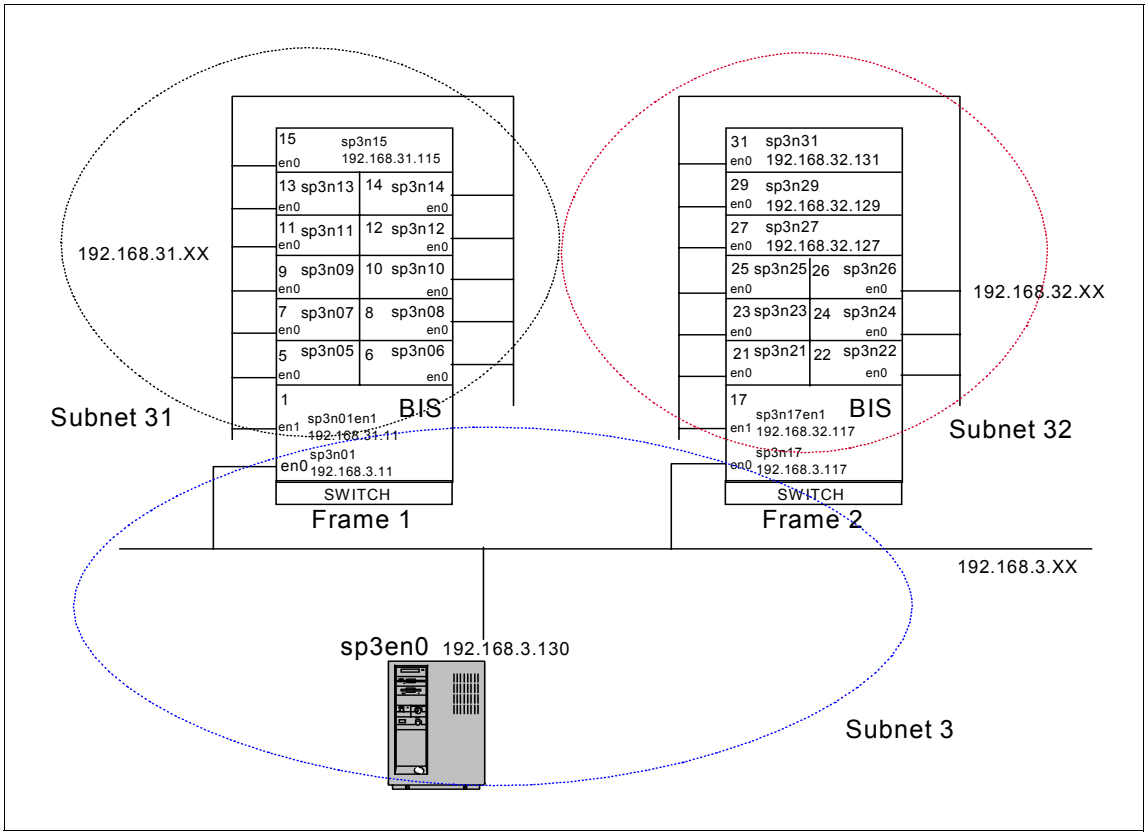


Figure 14-1 Environment after adding a second switched frame and nodes

14.3 Adding a frame

In our environment, we assigned sp3n17 as the boot/install server node. Thus, we added en0 of sp3n17 on subnet 3 and en1 of sp3n17 on subnet 32 so that en1 will be a gateway to reach the CWS from the nodes in frame 2.

With this configuration, we summarized the steps as follows:

Note: You should obtain a valid Kerberos ticket by issuing the **`klist`** or **`k4init`** command from the RS/6000 SP authentication services to perform the following tasks.

1. Archive the SDR on the CWS. Each time you reconfigure your system, it is strongly recommended to back up the SDR with the command:

```
[sp3en0:/]# SDRArchive
SDRArchive: SDR archive file name is
/spdata/sys1/sdr/archives/backup.98350.1559
```

In case something goes wrong, you can simply restore with the command:

```
SDRRestore <archive_file>
```

2. Un-partition your system (optional) from the CWS.

   If your existing system has multiple partitions defined and you want to add a frame that has a switch, you need to bring the system down to one partition by using the **Eunpartition** command before you can add the additional frame.

3. Connect the frame with RS-232 and recable the Ethernet adapters (en0), as described in 14.2, “Environment” on page 438, to your CWS.

4. Configure the RS-232 control line.

   Each frame in your system requires a serial port on the CWS configured to accommodate the RS-232 line. Note that SP-attached servers require two serial lines. Define **tty1** for the second Frame:

   ```
   [sp3en0:/]# mkdev -c tty -t 'tty' -s 'rs232' -p 'sa1' -w 's2'
   ```

5. Enter frame information and reinitialize the SDR.

   For SP frames, this step creates frame objects in the SDR for each frame in your system. At the end of this step, the SDR is reinitialized resulting in the creation of node objects for each node attached to your frames.

   Note: You must perform this step once for SP frames and once for non-SP frames (SP-attached servers). You do not need to reinitialize the SDR until you are entering the last set of frames (SP or non-SP).

   Specify the **spframe** command with **-r yes** to reinitialize the SDR (when running the command for the final series of frames), a starting frame number, a frame count, and the starting frame's tty port.

   In our environment, we enter information for two frames (frame 1 and frame 2) and indicate that frame 1 is connected to /dev/tty0 and frame 2 to /dev/tty1 and reinitialize the SDR:

   ```
   [sp3en0:/]# spframe -r yes 1 2 /dev/tty0
   0513-044 The stop of the splogd Subsystem was completed successfully.
   0513-059 The splogd Subsystem has been started. Subsystem PID is 111396.
   ```

   Note: If frames are not contiguously numbered, repeat this step for each series of contiguous frames.

As a new feature of PSSP 3.1, SP-attached servers are supported. For non-SP frames, SP-attached servers also require frame objects in the SDR as non-SP frames, and one object is required for each S70, S70 Advanced, or S80 attached to your SP.

The S70, S70 Advanced, and S80 require two tty port values to define the tty ports on the CWS to which the serial cables connected to the server are attached. The spframe tty port value defines the serial connection to the operator panel on the S70, S70 Advanced, and S80 hardware controls. The s1 tty port value defines the connection to the serial port on the S70, S70 Advanced, and S80 for serial terminal (s1term) support. A switch port value is required for each S70, S70 Advanced, or S80 attached to your SP.

Specify the **spframe** command with the **-n** option for each series of contiguous non-SP frames. Specify the **-r yes** option when running the command for the final series of frames.

If you have two S70 servers (frames 3 and 4), then the first server has the following characteristics:

Frame Number: 3

tty port for operator panel connection: /dev/tty2

tty port for serial terminal connection: /dev/tty3

Switch port number: 14

The second server has the following characteristics:

Frame Number: 4

tty port for operator panel connection: /dev/tty4

tty port for serial terminal connection: /dev/tty5

Switch port number: 15

To define these servers to PSSP and reinitialize the SDR, enter:

```
# spframe -r yes -n 14 3 2 /dev/tty2
```

Note: The SP-attached server in your system will be represented with the node number corresponding to the frame defined in this step. Continue with the remaining installation steps to install the SP-attached server as an SP node.

6. Verify frame information with the command: **splstdata -f** or **spmon -d**.

The output looks as follows:

```
[sp3en0:/]# splstdata -f
                 List Frame Database Information

frame#              tty           s1_tty      frame_type  hardware_protocol
------------------------------------------------------------------------------
     1        /dev/tty0               ""          switch                 SP
     2        /dev/tty1               ""          switch                 SP

[sp3en0:/]# spmon -d
1.  Checking server process
    Process 16264 has accumulated 0 minutes and 0 seconds.
    Check ok
2.  Opening connection to server
    Connection opened
    Check ok
3.  Querying frame(s)
    2 frame(s)
    Check ok
4.  Checking frames
    This step was skipped because the -G flag was omitted.
5.  Checking nodes
--------------------------------- Frame 1
-----------------------------------
Frame  Node    Node         Host/Switch  Key     Env   Front Panel
LCD/LED is
Slot   Number  Type  Power   Responds    Switch  Fail    LCD/LED
Flashing
------------------------------------------------------------------------------
----
  1       1    high    on   yes  yes     normal   no   LCDs are blank   no
  5       5    thin    on   yes  yes     normal   no   LEDs are blank   no
  6       6    thin    on   yes  yes     normal   no   LEDs are blank   no
  7       7    thin    on   yes  yes     normal   no   LEDs are blank   no
  8       8    thin    on   yes  yes     normal   no   LEDs are blank   no
  9       9    thin    on   yes  yes     normal   no   LEDs are blank   no
 10      10    thin    on   yes  yes     normal   no   LEDs 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
 14      14    thin    on   yes  yes     normal   no   LEDs are blank   no
 15      15    wide    on   yes  yes     normal   no   LEDs are blank   no

--------------------------------- Frame 2
-----------------------------------
Frame  Node    Node         Host/Switch  Key     Env   Front Panel
LCD/LED is
```

```
Slot   Number  Type  Power   Responds    Switch  Fail    LCD/LED
Flashing
-------------------------------------------------------------------------
----
  1      17    high    on    no  notcfg  normal   no   LCDs are blank   no
  5      21    thin    on    no  notcfg  normal   no   LEDs are blank   no
  6      22    thin    on    no  notcfg  normal   no   LEDs are blank   no
  7      23    thin    on    no  notcfg  normal   no   LEDs are blank   no
  8      24    thin    on    no  notcfg  normal   no   LEDs are blank   no
  9      25    thin    on    no  notcfg  N/A      no   LCDs are blank   no
 10      26    thin    on    no  notcfg  N/A      no   LCDs are blank   no
 11      27    wide    on    no  notcfg  normal   no   LEDs are blank   no
 13      29    wide    on    no  notcfg  normal   no   LEDs are blank   no
 15      31    wide    on    no  notcfg  normal   no   LEDs are blank   no
```

Note that SP-attached servers are represented as a one-node frame. If an error occurred, the frame must be deleted using the **spdelfram** command prior to reissuing the **spframe** command. After updating the RS-232 connection to the frame, you should reissue the **spframe** command.

14.4 Adding a node

In our environment, we add one high node as 2nd boot/install server, four thin nodes, two Silver nodes, and three wide nodes as shown in Figure 14-1 on page 439. Assume that all nodes were installed when the frame was installed. Thus, the following steps are the continuation of 14.1, “Key concepts” on page 438. After we enter all nodes information into the SDR, we will install sp3n17 first and then install the rest of the nodes.

1. Gather all information that you need:
   - Hostnames for all nodes
   - IP addresses for all nodes
   - Default gateway information, and so on
2. Archive the SDR with the command: **SDRArchive**
3. Update the /etc/hosts file or DNS map with new IP addresses on the CWS. Note that if you do not update the /etc/hosts file now, the **spethernt** command will fail.
4. Check the status and update the state of the supervisor microcode with the command: **spsvrmgr**.

The output looks like this:

```
[sp3en0:/]# spsvrmgr -G -r status all

spsvrmgr: Frame  Slot  Supervisor  Media          Installed     Required
                       State       Versions       Version       Action
          _____  ____  __________  _____________  ____________  ____________
          1      0     Active      u_10.1c.0709   u_10.1c.070c  None
                                   u_10.1c.070c
                 ____  __________  _____________  ____________  ____________
                 1     Active      u_10.3a.0614   u_10.3a.0615  None
                                   u_10.3a.0615
                 ____  __________  _____________  ____________  ____________
                 17    Active      u_80.19.060b   u_80.19.060b  None
          _____  ____  __________  _____________  ____________  ____________
          2      0     Active      u_10.3c.0709   u_10.3c.070c  None
                                   u_10.3c.070c
                 ____  __________  _____________  ____________  ____________
                 1     Active      u_10.3a.0614   u_10.3a.0615  None
                                   u_10.3a.0615
                 ____  __________  _____________  ____________  ____________
                 9     Active      u_10.3e.0704   u_10.3e.0706  None
                                   u_10.3e.0706
                 ____  __________  _____________  ____________  ____________
                 10    Active      u_10.3e.0704   u_10.3e.0706  None
                                   u_10.3e.0706
                 ____  __________  _____________  ____________  ____________
                 17    Active      u_80.19.060b   u_80.19.060b  None
```

In our environment, there is no *Required Action* needed to be taken. However, if you need to update the microcode of the frame supervisor of frame 2, enter:

```
# spsvrmgr -G -u 2:0
```

5. Enter the required en0 adapter Information with the command: **spethernt**.

```
[sp3en0:/etc]# spethernt -s no -l 17 192.168.3.117 255.255.255.0
192.168.3.130
[sp3en0:/etc]# spethernt -s no -l 21 192.168.32.121 255.255.255.0
192.168.32.117
[sp3en0:/etc]# spethernt -s no -l 22 192.168.32.122 255.255.255.0
192.168.32.117
[sp3en0:/etc]# spethernt -s no -l 23 192.168.32.123 255.255.255.0
192.168.32.117
[sp3en0:/etc]# spethernt -s no -l 24 192.168.32.124 255.255.255.0
192.168.32.117
[sp3en0:/etc]# spethernt -s no -l 25 192.168.32.125 255.255.255.0
192.168.32.117
```

```
[sp3en0:/etc]# spethernt -s no -l 26 192.168.32.126 255.255.255.0
192.168.32.117
[sp3en0:/etc]# spethernt -s no -l 27 192.168.32.127 255.255.255.0
192.168.32.117
[sp3en0:/etc]# spethernt -s no -l 29 192.168.32.129 255.255.255.0
192.168.32.117
[sp3en0:/etc]# spethernt -s no -l 31 192.168.32.131 255.255.255.0
192.168.32.117
```

If you are adding an extension node to your system, you may want to enter the required node information now. For more information, refer to chapter 9 of *PSSP Installation and Migration Guide,* GA22-7347.

6. Acquire the hardware Ethernet addresses with the command: **sphrdward**.

   This step gets hardware Ethernet addresses for the en0 adapters for your nodes from the nodes themselves and puts them into the *Node Objects* in the SDR. This information is used to set up the /etc/bootptab files for your boot/install servers.

   To get all hardware Ethernet addresses for the nodes specified in the node list (the **-l** flag), enter:

   ```
   [sp3en0:/]# sphrdwrad -l 17,21,22,23,24,25,26,27,29,31
   ```

   A sample output looks like the following:

   ```
   Acquiring hardware Ethernet address for node 17
   Acquiring hardware Ethernet address for node 21
   Acquiring hardware Ethernet address for node 22
   Acquiring hardware Ethernet address for node 23
   Acquiring hardware Ethernet address for node 24
   Acquiring hardware Ethernet address for node 25
   Acquiring hardware Ethernet address for node 26
   Acquiring hardware Ethernet address for node 27
   Acquiring hardware Ethernet address for node 29
   Acquiring hardware Ethernet address for node 31
   Hardware ethernet address for node 17 is 02608C2E86CA
   Hardware ethernet address for node 21 is 10005AFA0518
   Hardware ethernet address for node 22 is 10005AFA17E3
   Hardware ethernet address for node 23 is 10005AFA1721
   Hardware ethernet address for node 24 is 10005AFA07DF
   Hardware ethernet address for node 25 is 0004AC4947E9
   Hardware ethernet address for node 26 is 0004AC494B40
   Hardware ethernet address for node 27 is 02608C2E7643
   Hardware ethernet address for node 29 is 02608C2E7C1E
   Hardware ethernet address for node 31 is 02608C2E78C9
   ```

Note: Do not do this step on a production system because it shuts down the nodes. Select only the new nodes you are adding. All the nodes you select are powered off and back on. The nodes for which you are obtaining Ethernet addresses must be physically powered on when you perform this step. No ttys can be opened in write mode.

7. Verify the Ethernet addresses with the command: **splstdata -b**.

   ```
   [sp3en0:/]# splstdata -b
   ```

   A sample output looks as follows:

```
List Node Boot/Install Information

node#         hostname  hdw_enet_addr srvr     response
install_disk
      last_install_image   last_install_time  next_install_image
lppsource_name
                pssp_ver         selected_vg
------------------------------------------------------------------------
----
    1 sp3n01.msc.itso.   02608CF534CC    0         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_11:18:20     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
    5 sp3n05.msc.itso.   10005AFA13AF    1         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_15:59:40     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
    6 sp3n06.msc.itso.   10005AFA1B12    1         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_15:59:56     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
    7 sp3n07.msc.itso.   10005AFA13D1    1         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_16:05:20     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
    8 sp3n08.msc.itso.   10005AFA0447    1         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_15:53:33     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
    9 sp3n09.msc.itso.   10005AFA158A    1         disk
hdisk0
```

```
         bos.obj.ssp.432 Thu_Dec__3_15:56:28     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
   10 sp3n10.msc.itso.   10005AFA159D    1         disk
hdisk0
         bos.obj.ssp.432 Fri_Dec__4_10:25:44     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
   11 sp3n11.msc.itso.   10005AFA147C    1         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_15:59:57     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
   12 sp3n12.msc.itso.   10005AFA0AB5    1         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_15:55:29     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
   13 sp3n13.msc.itso.   10005AFA1A92    1         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_16:07:48     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
   14 sp3n14.msc.itso.   10005AFA0333    1         disk
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_16:08:31     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
   15 sp3n15.msc.itso.   02608C2E7785    1      install
hdisk0
         bos.obj.ssp.432 Thu_Dec__3_16:05:03     bos.obj.ssp.432
aix432
                PSSP-3.1              rootvg
   17 sp3n17.msc.itso.   02608C2E86CA    0      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
   21 sp3n21.msc.itso.   10005AFA0518   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
   22 sp3n22.msc.itso.   10005AFA17E3   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
```

```
   23 sp3n23.msc.itso.   10005AFA1721   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
   24 sp3n24.msc.itso.   10005AFA07DF   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
   25 sp3n25.msc.itso.   0004AC4947E9   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
   26 sp3n26.msc.itso.   0004AC494B40   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
   27 sp3n27.msc.itso.   02608C2E7643   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
   29 sp3n29.msc.itso.   02608C2E7C1E   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
   31 sp3n31.msc.itso.   02608C2E78C9   17      install
hdisk0
                 initial             initial             default
default
                PSSP-3.1              rootvg
```

8. Configure additional adapters for nodes to create adapter objects in the SDR with the command **spadaptrs**. You can only configure Ethernet (en), FDDI (fi), Token Ring (tr), and css0 (applies to the SP Switch) with this command. To configure adapters, such as ESCON and PCA, you must configure the adapter manually on each node using **dsh** or modify the firstboot.cust file.

   For en1adapter, enter:

   ```
   [sp3en0:/]# spadaptrs -s no -t bnc -l 17 en1 192.168.32.117 255.255.255.0
   ```

For the css0 (SP Switch) adapter, the output looks as follows:

```
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 17 css0 192.168.13.17
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 21 css0 192.168.13.21
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 22 css0 192.168.13.22
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 23 css0 192.168.13.23
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 24 css0 192.168.13.24
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 25 css0 192.168.13.25
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 26 css0 192.168.13.26
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 27 css0 192.168.13.27
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 29 css0 192.168.13.29
255.255.255.0
[sp3en0:/]# spadaptrs -s no -n no -a yes -l 31 css0 192.168.13.31
255.255.255.0
```

If you specify the **-s** flag to skip IP addresses when you are setting the css0 switch addresses, you must also specify **-n no** to not use switch numbers for IP address assignment and **-a yes** to use ARP.

Note: The command **spadaptrs** is supported by only two adapters for the Ethernet (en), FDDI (fi), and Token Ring (tr) in PSSP V2.4 or earlier. However, with PTFs (ssp.basic.2.4.0.4) on PSSP 2.4 or PSSP3.1, it is changed to support as many adapters as you can have in the system.

9. Configure initial host names for nodes to change the default host name information in the SDR node objects with the command **sphostnam**. The default is the long form of the en0 host name, which is how the **spethernt** command processes defaulted host names. However, we set the hostname as short name:

   ```
   [sp3en0:/]# sphostnam -a en0 -f short -l 17,21,22,23,24,25,26,27,29,31
   ```

10. Set up nodes to be installed.

Note: You cannot export /usr or any directories below /usr because an NFS export problem will occur. If you have exported the /spdata/sys1/install/image directory or any parent directory, you must unexport it using the **exportfs -u** command before running setup_server.

From the output of step 7, we need to change the image name and AIX version. In addition, we checked that the sp3n17 node points to the CWS as boot/install server, and all the rest of the nodes point to sp3n17 as boot/install server, which is the default in a multiframe environment. However, if you need to select a different node to be boot/install server, you can use the **-n** option of the **spchvgobj** command.

To change this information in SDR, enter:

```
[sp3en0:/]# spchvgobj -r rootvg -i bos.obj.ssp.432 -l
17,21,22,23,24,25,26,27,29,31
```

A sample output looks as follows:

```
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 17, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 21, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 22, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 23, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 24, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 25, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 26, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 27, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 29, volume group rootvg.
spchvgobj: Successfully changed the Node and Volume_Group objects for node
number 31, volume group rootvg.
spchvgobj: The total number of changes successfully completed is 10.
spchvgobj: The total number of changes which were not successfully
completed is 0.
```

Now run the command **spbootins** to run setup_server to configure the boot/install server. We first installed sp3n17, then the rest of the nodes later:

```
[sp3en0:/]# spbootins -r install -l 17
```

11. Refresh the system partition-sensitive subsystems on both the CWS and the nodes:

    ```
    [sp3en0:/]# syspar_ctrl -r -G
    ```

12. Verify all node information with the **splstdata** command with the options **-f, -n, -a,** or **-b**.

13. Change the default network tunable values (optional).

If you set up the boot/install server, and it is acting as a gateway to the CWS, the **ipforwarding** command must be enabled. To turn it on, issue:

```
# /usr/sbin/no -o ipforwarding=1
```

When a node is installed, migrated, or customized (set to customize and rebooted), and that node's boot/install server does not have a l/tftpboot/tuning.cust file, a default file of system performance tuning variable settings in /usr/lpp/ssp/install/config/tuning.default is copied to /tftpboot/tuning.cust on that node. You can override these values by following one of the methods described in the following list:

IBM supplies three alternate tuning files that contain initial performance tuning parameters for three different SP environments: /usr/lpp/ssp/install/config/tuning.commercial, tuning.development, and tuning.scientific.

Note: The S70, S70 Advanced, and S80 should not use the tuning.scientific file because of the large number of processors and the amount of traffic that they can generate.

To select the sample tuning file, issue the **cptuning** command to copy to /tftpboot/tuning.cust on the CWS and propagate from there to each node in the system when it is installed, migrated, or customized.

Note that each node inherits its tuning file from its boot/install server. Nodes that have as their boot/install server another node (other than the CWS) obtain their tuning.cust file from that server node; so, it is necessary to propagate the file to the server node before attempting to propagate it to the client node.

14. Perform additional node customization, such as adding installp images, configuring host names, setting up NFS, AFS, or NIS, and configuring adapters that are not configured automatically (optional).

    The **script.cust** script is run from the PSSP NIM customization script (pssp_script) after the node's AIX and PSSP software have been installed but before the node has been rebooted. This script is run in a limited environment where not all services are fully configured. Because of this limited environment, you should restrict the use of the script.cust function. The function must be performed prior to the post-installation reboot of the node.

    The **firstboot.cust** script is run during the first boot of the node immediately after it has been installed. This script runs better in an environment where most of the services have been fully configured.

15. Additional switch configuration (optional)

    If you have added a frame with a switch, perform the following steps:

a. Select a topology file from the /etc/SP directory on the CWS.

> **Note:** SP-attached servers never contain a node switch board; therefore, never include non-SP frames when determining your topology files.

b. Manage the switch topology files.

The switch topology file must be stored in the SDR. The switch initialization code uses the topology file stored in the SDR when starting the switch (**Estart**). When the switch topology file is selected for your system's switch configuration, it must be annotated with **Eannotator** then stored in the SDR with **Etopology**. The switch topology file stored in the SDR can be overwritten by having an expected.top file in /etc/SP on the primary node. **Estart** always checks for an expected.top file in /etc/SP before using the one stored in the SDR. The expected.top file is used when debugging or servicing the switch.

c. Annotate a switch topology file with the command: **Eannotator**.

Annotate a switch topology file before storing it in the SDR. Use the **Eannotator** command to update the switch topology file's connection labels with their correct physical locations. Use the **-O yes** flag to store the switch topology file in the SDR:

```
[sp3en0:/etc]# Eannotator -F /etc/SP/expected.top.2nsb.0isb.0 -f
/etc/SP/expected.top.annotated -O yes
```

d. Set the switch clock source for all switches with the command: **Eclock**.

For our environment, select **/etc/SP/Eclock.top.2nsb.0isb.0** as a topology file and enter:

```
[sp3en0:/]# Eclock -f /etc/SP/Eclock.top.2nsb.0isb.0
```

To verify the switch configuration information, enter:

```
[sp3en0:/]# splstdata -s
```

A sample output looks as follows:

```
List Node Switch Information

                       switch  switch  switch switch    switch
node# initial_hostname   node# protocol number   chip chip_port
-------------------------------------------------------------------------
--------
    1 sp3n01.msc.itso.       0       IP      1      5         3
    5 sp3n05.msc.itso.       4       IP      1      5         1
    6 sp3n06.msc.itso.       5       IP      1      5         0
    7 sp3n07.msc.itso.       6       IP      1      6         2
```

```
    8 sp3n08.msc.itso.       7       IP      1      6         3
    9 sp3n09.msc.itso.       8       IP      1      4         3
   10 sp3n10.msc.itso.       9       IP      1      4         2
   11 sp3n11.msc.itso.      10       IP      1      7         0
   12 sp3n12.msc.itso.      11       IP      1      7         1
   13 sp3n13.msc.itso.      12       IP      1      4         1
   14 sp3n14.msc.itso.      13       IP      1      4         0
   15 sp3n15.msc.itso.      14       IP      1      7         2
   17 sp3n17.msc.itso.      16       IP      2      5         3
   21 sp3n21.msc.itso.      20       IP      2      5         1
   22 sp3n22.msc.itso.      21       IP      2      5         0
   23 sp3n23.msc.itso.      22       IP      2      6         2
   24 sp3n24.msc.itso.      23       IP      2      6         3
   25 sp3n25.msc.itso.      24       IP      2      4         3
   26 sp3n26.msc.itso.      25       IP      2      4         2
   27 sp3n27.msc.itso.      26       IP      2      7         0
   29 sp3n29.msc.itso.      28       IP      2      4         1
   31 sp3n31.msc.itso.      30       IP      2      7         2

switch   frame    slot  switch_partition  switch  clock switch
number  number  number            number    type  input  level
-------------------------------------------------------------------------
--------
     1       1      17                 1     129      0
     2       2      17                 1     129      3

switch_part           topology           primary         arp
switch_node
     number           filename              name     enabled
nos._used
-------------------------------------------------------------------------
--------
           1   expected.top.an  sp3n05.msc.itso.         yes
no
```

16. Network boot the boot/install server node sp3n17.

    a. To monitor installation progress by opening the node's read-only console, issue:

       ```
       [sp3en0:/]# s1term 2 1
       ```

    b. To network boot sp3n17, issue:

       ```
       [sp3en0:/]# nodecond 2 1&
       ```

    Monitor /var/adm/SPlogs/spmon/nc/nc.<frame_number>.<node_number> and check the /var/adm/SPlogs/sysman/<node>.console.log file on the boot/install node to see if **setup_server** has completed.

17. Verify that system management tools were correctly installed on the boot/install servers. Now that the boot/install servers are powered up, run the verification test from the CWS to check for correct installation of the system management tools on these nodes.

    To do this, enter:

    ```
    [sp3en0:/]# SYSMAN_test
    ```

    After the tests are run, the system creates a log in /var/adm/SPlogs called SYSMAN_test.log.

18. After you install the boot/install server, run the command **spbootins** to run **setup_server** for the rest of the nodes:

    ```
    [sp3en0:/etc/]# spbootins -r install -l 21,22,23,24,25,26,27,29,31
    ```

    The sample output is as follows:

```
setup_server command results from sp3en0
-------------------------------------------------------
setup_server: Running services_config script to configure SSP services.This
may tak
e a few minutes...
rc.ntp: NTP already running - not starting ntp
0513-029 The supfilesrv Subsystem is already active.
Multiple instances are not supported.
/etc/auto/startauto: The automount daemon is already running on this
system.
setup_CWS: Control Workstation setup complete.
mknimmast: Node 0 (sp3en0) already configured as a NIM master.
create_krb_files: tftpaccess.ctl file and client srvtab files
created/updated
on server node 0.
mknimres: Copying /usr/lpp/ssp/install/bin/pssp_script to
/spdata/sys1/install/pssp
/pssp_script.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data.template to
/spdata/sys1
/install/pssp/bosinst_data.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data_prompt.template
to /spda
ta/sys1/install/pssp/bosinst_data_prompt.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data_migrate.template
to /spd
ata/sys1/install/pssp/bosinst_data_migrate.
mknimclient: 0016-242: Client node 1 (sp3n01.msc.itso.ibm.com) already
defined on s
erver node 0 (sp3en0).
mknimclient: 0016-242: Client node 17 (sp3n17.msc.itso.ibm.com) already
defined on
```

```
server node 0 (sp3en0).
export_clients: File systems exported to clients from server node 0.
allnimres: Node 1 (sp3n01.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 17 (sp3n17.msc.itso.ibm.com) prepared for operation: disk.
setup_server: Processing complete (rc= 0).
setup_server command results from sp3n01.msc.itso.ibm.com
------------------------------------------------------
setup_server: Running services_config script to configure SSP services.This
may tak
e a few minutes...
rc.ntp: NTP already running - not starting ntp
supper:  Active volume group rootvg.
Updating collection sup.admin from server sp3en0.msc.itso.ibm.com.
File Changes:  0 updated, 0 removed, 0 errors.
Updating collection user.admin from server sp3en0.msc.itso.ibm.com.
File Changes:  6 updated, 0 removed, 0 errors.
Updating collection power_system from server sp3en0.msc.itso.ibm.com.
File Changes:  0 updated, 0 removed, 0 errors.
Updating collection node.root from server sp3en0.msc.itso.ibm.com.
File Changes:  0 updated, 0 removed, 0 errors.
0513-029 The supfilesrv Subsystem is already active.
Multiple instances are not supported.
/etc/auto/startauto: The automount daemon is already running on this
system.
mknimmast: Node 1 (sp3n01.msc.itso.ibm.com) already configured as a NIM
master.
create_krb_files: tftpaccess.ctl file and client srvtab files
created/updated
on server node 1.
mknimres: Copying /usr/lpp/ssp/install/bin/pssp_script to
/spdata/sys1/install/pssp
/pssp_script.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data.template to
/spdata/sys1
/install/pssp/bosinst_data.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data_prompt.template
to /spda
ta/sys1/install/pssp/bosinst_data_prompt.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data_migrate.template
to /spd
ata/sys1/install/pssp/bosinst_data_migrate.
mknimclient: 0016-242: Client node 5 (sp3n05.msc.itso.ibm.com) already
defined on s
erver node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 6 (sp3n06.msc.itso.ibm.com) already
defined on s
erver node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 7 (sp3n07.msc.itso.ibm.com) already
defined on s
```

```
erver node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 8 (sp3n08.msc.itso.ibm.com) already
defined on s
erver node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 9 (sp3n09.msc.itso.ibm.com) already
defined on s
erver node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 10 (sp3n10.msc.itso.ibm.com) already
defined on
server node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 11 (sp3n11.msc.itso.ibm.com) already
defined on
server node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 12 (sp3n12.msc.itso.ibm.com) already
defined on
server node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 13 (sp3n13.msc.itso.ibm.com) already
defined on
server node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 14 (sp3n14.msc.itso.ibm.com) already
defined on
server node 1 (sp3n01.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 15 (sp3n15.msc.itso.ibm.com) already
defined on
server node 1 (sp3n01.msc.itso.ibm.com).
export_clients: File systems exported to clients from server node 1.
allnimres: Node 5 (sp3n05.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 6 (sp3n06.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 7 (sp3n07.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 8 (sp3n08.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 9 (sp3n09.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 10 (sp3n10.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 11 (sp3n11.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 12 (sp3n12.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 13 (sp3n13.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 14 (sp3n14.msc.itso.ibm.com) prepared for operation: disk.
allnimres: Node 15 (sp3n15.msc.itso.ibm.com) prepared for operation: disk.
setup_server: Processing complete (rc= 0).

setup_server command results from sp3n17.msc.itso.ibm.com
-------------------------------------------------------
setup_server: Running services_config script to configure SSP services.This
may tak
e a few minutes...
rc.ntp: NTP already running - not starting ntp
supper:  Active volume group rootvg.
Updating collection sup.admin from server sp3en0.msc.itso.ibm.com.
File Changes:  0 updated, 0 removed, 0 errors.
Updating collection user.admin from server sp3en0.msc.itso.ibm.com.
```

```
File Changes:  6 updated, 0 removed, 0 errors.
Updating collection power_system from server sp3en0.msc.itso.ibm.com.
File Changes:  0 updated, 0 removed, 0 errors.
Updating collection node.root from server sp3en0.msc.itso.ibm.com.
File Changes:  0 updated, 0 removed, 0 errors.
0513-029 The supfilesrv Subsystem is already active.
Multiple instances are not supported.
/etc/auto/startauto: The automount daemon is already running on this
system.
mknimmast: Node 17 (sp3n17.msc.itso.ibm.com) already configured as a NIM
master.
create_krb_files: tftpaccess.ctl file and client srvtab files
created/updated
on server node 17.
mknimres: Copying /usr/lpp/ssp/install/bin/pssp_script to
/spdata/sys1/install/pssp
/pssp_script.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data.template to
/spdata/sys1
/install/pssp/bosinst_data.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data_prompt.template
to /spda
ta/sys1/install/pssp/bosinst_data_prompt.
mknimres: Copying /usr/lpp/ssp/install/config/bosinst_data_migrate.template
to /spd
ata/sys1/install/pssp/bosinst_data_migrate.
mknimclient: 0016-242: Client node 21 (sp3n21.msc.itso.ibm.com) already
defined on
server node 17 (sp3n17.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 22 (sp3n22.msc.itso.ibm.com) already
defined on
server node 17 (sp3n17.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 23 (sp3n23.msc.itso.ibm.com) already
defined on
server node 17 (sp3n17.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 24 (sp3n24.msc.itso.ibm.com) already
defined on
server node 17 (sp3n17.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 25 (sp3n25.msc.itso.ibm.com) already
defined on
server node 17 (sp3n17.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 26 (sp3n26.msc.itso.ibm.com) already
defined on
server node 17 (sp3n17.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 27 (sp3n27.msc.itso.ibm.com) already
defined on
server node 17 (sp3n17.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 29 (sp3n29.msc.itso.ibm.com) already
defined on
```

```
server node 17 (sp3n17.msc.itso.ibm.com).
mknimclient: 0016-242: Client node 31 (sp3n31.msc.itso.ibm.com) already
defined on
server node 17 (sp3n17.msc.itso.ibm.com).
export_clients: File systems exported to clients from server node 17.
allnimres: Node 21 (sp3n21.msc.itso.ibm.com) prepared for operation:
install.
allnimres: Node 22 (sp3n22.msc.itso.ibm.com) prepared for operation:
install.
allnimres: Node 23 (sp3n23.msc.itso.ibm.com) prepared for operation:
install.
allnimres: Node 24 (sp3n24.msc.itso.ibm.com) prepared for operation:
install.
allnimres: Node 25 (sp3n25.msc.itso.ibm.com) prepared for operation:
install.
allnimres: Node 26 (sp3n26.msc.itso.ibm.com) prepared for operation:
install.
allnimres: Node 27 (sp3n27.msc.itso.ibm.com) prepared for operation:
install.
allnimres: Node 29 (sp3n29.msc.itso.ibm.com) prepared for operation:
install.
allnimres: Node 31 (sp3n31.msc.itso.ibm.com) prepared for operation:
install.
setup_server: Processing complete (rc= 0).
```

19. Network boot the rest of the nodes:

    ```
    [sp3en0:/]# nodecond 2 5&
    ```

    Monitor /var/adm/SPlogs/spmon/nc/nc.<frame_number>.<node_number> and check the /var/adm/SPlogs/sysman/<node>.console.log file on the boot/install node to see if setup_server has completed.

20. Verify node installation.

    To check the hostResponds and powerLED indicators for each node, enter:

    ```
    [sp3en0:/]# spmon -d -G
    ```

21. Start the switch with the following command after all nodes are installed:

    ```
    [sp3en0:/]# Estart
    Estart: Oncoming primary != primary, Estart directed to oncoming primary
    Estart:  0028-061 Estart is being issued to the primary node:
    sp3n05.msc.itso.ibm.com.
    Switch initialization started on sp3n05.msc.itso.ibm.com.
    Initialized 14 node(s).
    Switch initialization completed.
    ```

    If you have set up system partitions, do this step in each partition.

22. Verify that the switch was installed correctly by running a verification test to ensure that the switch is installed completely. To do this, enter:

```
[sp3en0:/]# CSS_test
```

After the tests are run, the system creates a log in /var/adm/SPlogs called CSS_test.log. To check the switchResponds and powerLED indicators for each node, enter:

```
[sp3en0:/]# spmon -d -G
```

23. Customize the node just installed:
    - Update .profile with proper PSSP command paths.
    - Get the Kerberos ticket with the command **k4init root.admin**, and so on.

14.5 Adding an existing S70 to an SP system

If you want to preserve the environment of your existing S70, S7A, or S80 server, perform the following steps to add an SP-attached server and preserve your existing software environment:

1. Upgrade AIX

   If your SP-attached server is not at AIX 4.3.2, you must first upgrade to that level of AIX before proceeding.

2. Set up name resolution of the SP-attached server

   In order to do PSSP customization, the following must be resolvable on the SP-attached server:

   - The CWS host name.
   - The name of the boot/install server's interface that is attached to the SP-attached server's en0 interface.

3. Set up routing to the CWS host name

   If you have a default route set up on the SP-attached server, you will have to delete it. If you do not remove the route, customization will fail when it tries to set up the default route defined in the SDR. In order for customization to occur, you must define a static route to the control workstation's host name. For example, the control workstation's host name is its token ring address, such as 9.114.73.76, and your gateway is 9.114.73.256:

   ```
   # route add -host 9.114.73.76 9.114.73.256
   ```

4. FTP the SDR_dest_info file

   During customization, certain information will be read from the SDR. In order to get to the SDR, you must FTP the /etc/SDR_dest_info file from the control workstation to the /etc/SDR_dest_info file on the SP-attached server and check the mode and ownership of the file.

5. Verify perfagent

   Ensure that perfagent.tools 2.2.32.x are installed in your SP-attached server.

6. Mount the /spdata/sys1/install/pssplpp directory on the boot/install server from the SP-attached server. For example, issue:

   ```
   # mount sp3en0:/spdata/sys1/install/pssplpp /mnt
   ```

7. Install **ssp.basic** and its prerequisites onto the SP-attached server:

   ```
   # installp -aXgd/mnt/PSSP-3.1 ssp.basic 2>&1 | tee /tmp/install.log
   ```

8. Unmount the /spdata/sys1/install/pssplpp directory on the boot/install server from the SP-attached server:

   ```
   # umount /mnt
   ```

9. Run the **pssp_script** by issuing:

   ```
   # /usr/lpp/ssp/install/bin/pssp_script
   ```

10. Perform a reboot:

    ```
    # shutdown -Fr
    ```

14.5.1 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.

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 14-2 on page 462 shows how the settings should look.

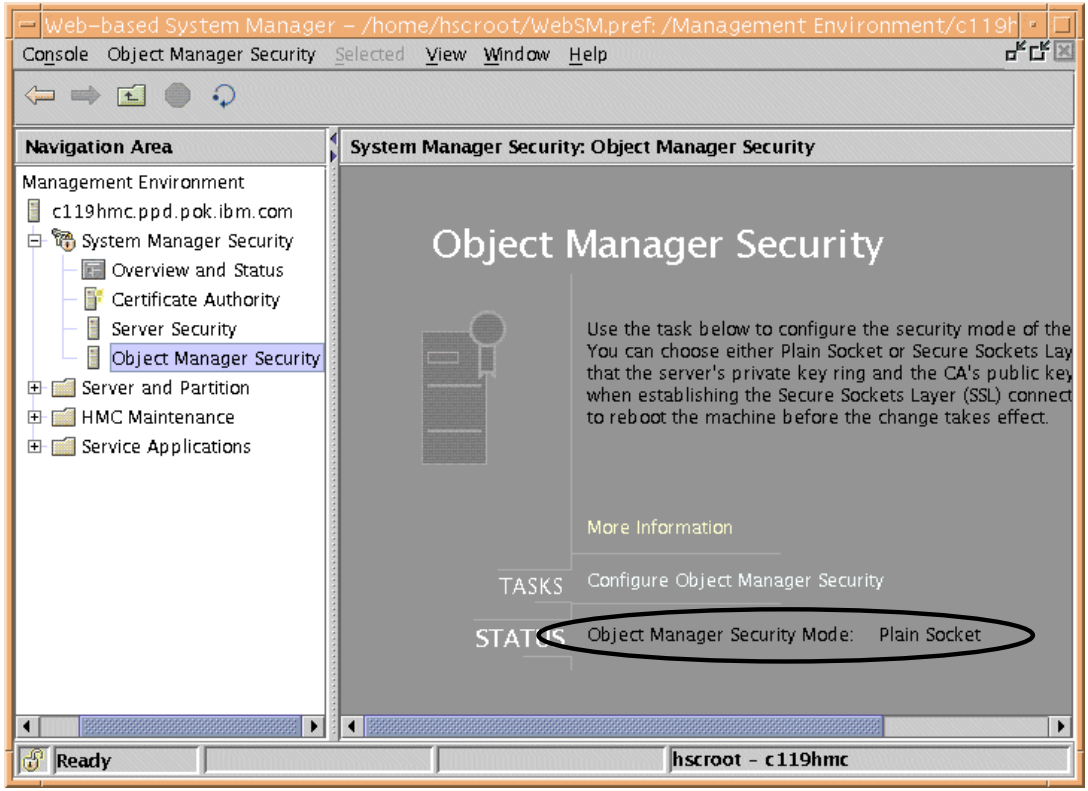


Figure 14-2 Setting the Object Manager Security

pSeries p630, 650, 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, p650 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; for the p655, RJ020829; and for the p650, 3K030515.

Example 14-1 shows how to list the firmware level installed in your machine.

Example 14-1 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 14-2.

Example 14-2 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 commands:

```
/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 the 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 an 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 14-3.

Example 14-3 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 14-3, the domain name for the server is Enterprise.

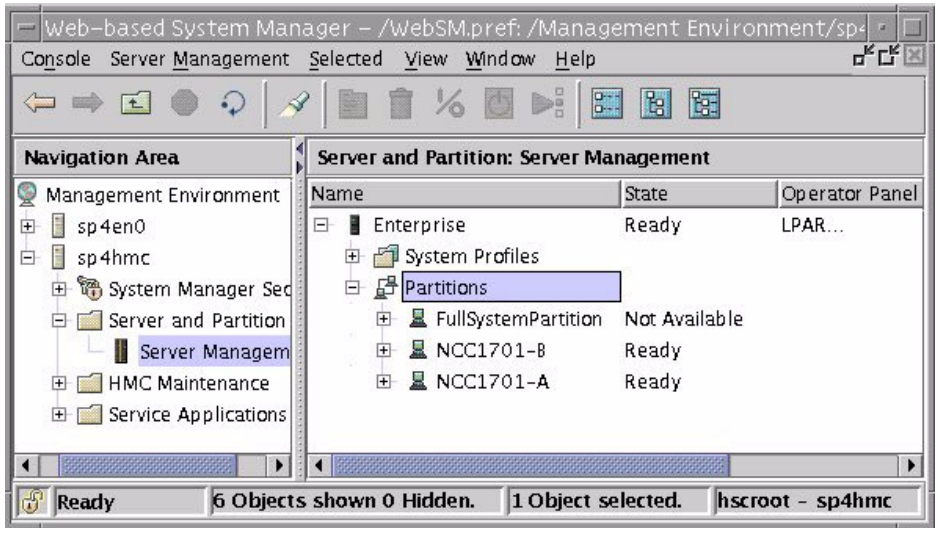


Figure 14-3 Web-based System Manager Console for HMC

Example 14-4 shows the syntax of the **spframe** command. Notice that there is no tty information for HMC frames.

Example 14-4 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 shown in Example 14-5. 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 14-5 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 also collects the hardware addresses. The nodes are powered off by the command. No tty connection can be open when running it. 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 14-6.

Example 14-6 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 14-4.

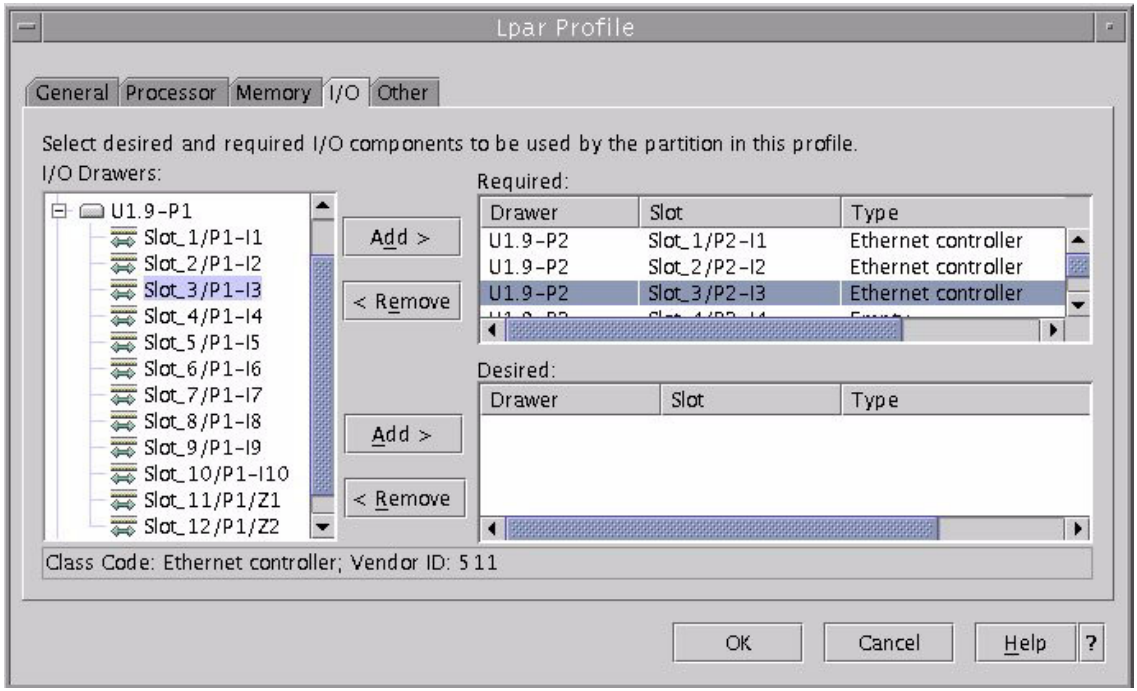


Figure 14-4 Hardware Management Console LPAR I/O profile

5. The hardware addresses can be queried with the **`netstat -I`** command or with **`lscfg -vl ent2`**, 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 use of the **`spadaptrs`** command 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 14-7 on page 469, 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 a 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 14-7 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 *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

14.6 Adding a switch

This section was already summarized as part of the previous section. However, here we introduce the following two cases when you add the SP Switch only:

- Adding a switch to a switchless system
- Adding a switch to a system with existing switches

14.6.1 Adding a switch to a switchless system

Perform the following to add a switch to a switchless system:

1. Redefine the system to a single partition.

   Refer to *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281, for more information.

2. Install the level of communication subsystem software (ssp.css) on the CWS with the **`installp`** command.

3. Install the new switch.

   Your IBM Customer Engineer (CE) performs this step. This step may include installing the switch adapters and installing a new frame supervisor card.

4. Create the switch partition class with the following command:

   ```
   # Eprimary -init
   ```

5. Check and update the state of the supervisor microcode with the command:

   ```
   spsvrmgr
   ```

6. Configure the switch adapters for each node with the **`spadaptrs`** command to create css0 adapter objects in the SDR for each new node.

7. Reconfigure the hardware monitor to recognize the new switch. Enter:

   ```
   # hmcmds -G setid 1:0
   ```

8. Update the System Data Repository (SDR).

   To update the SDR switch information, issue the following command:

   ```
   # /usr/lpp/ssp/install/bin/hmreinit
   ```

9. Set up the switch.

   Refer to step 15 in 14.4, “Adding a node” on page 443.

10. Refresh system partition-sensitive subsystems with the following command on the CWS after adding the switch:

    ```
    # syspar_ctrl -r -G
    ```

11. Set the nodes to *customize* with the following command:

    ```
    # spbootins -r customize -l <node_list>
    ```

12. Reboot all the nodes for node customization.
13. Start up the switch with **Estart** and verify the switch.

14.6.2 Adding a switch to a system with existing switches

Perform the following steps to add a switch to a system with existing switches:

1. Redefine the system to a single partition.
2. Install the new switch.

   Your IBM Customer Engineer (CE) performs this step. This step includes installing the switch adapters and installing new frame supervisors.

3. Check and update the state of the supervisor microcode with the command: `spsvrmgr`
4. Configure the adapters for each node with the **spadaptrs** command to create **css0** adapter objects in the SDR for each new node.
5. Set up the switch.

   Refer the step 15 in “Adding a node” on page 443.

6. Refresh system partition-sensitive subsystems with the following command on the CWS after adding the switch:

   ```
   # syspar_ctrl -r -G
   ```

7. Set the nodes to **customize** with the following command:

   ```
   # spbootins -r customize -l <node_list>
   ```

8. Reboot all the nodes for node customization.
9. Start up the switch with **Estart** and verify the switch.

14.7 Replacing to PCI-based 332 MHz SMP node

This migration scenario is summarized only for preparing for the exam and will not provide complete information for conducting an actual migration. However, this section provides enough information to understand the migration process.

14.7.1 Assumptions

- There is only one partition in the SP system.
- All nodes being upgraded are required to be installed with a current mksysb image. Note that logical names for devices on the new 332 MHz SMP node will most likely not be the same as on the legacy node. This is because the 332 MHz SMP node will be freshly installed and is a different technology.
- The node we are migrating is not a boot/install server node.
- HACWS is not implemented.
- Install AIX Version 4.3.2 and PSSP Version 3.1.

14.7.2 Software prerequisites

Getting the correct software acquired, copied, and installed can be a most complex task in any SP installation. Migrating to the 332 MHz SMP node and PSSP 2.4 or PSSP 3.1 is no exception. The basic facts surrounding proper software prerequisites and installation are:

- Required base level AIX filesets and all PTFs should be installed on the CWS and nodes.
- Required base level AIX filesets and all PTFs should be copied and available in /spdata/sys1/install/aix432/lppsource.
- Required base level AIX filesets should be built into the appropriate SPOT.
- Required AIX fixes should be used to customize the appropriate SPOT.
- Required PSSP fixes should be copied into the PSSP directory (/spdata/sys1/install/pssplpp/PSSP-3.1).

PSSP code

A general recommendation is to install all the latest level fixes during a migration. This includes both the CWS and the nodes. The fixes will not be installed by default even if properly placed in the /spdata/sys/install/pssplpp/PSSP-3.1 directory. You must explicitly specify `Install at latest available level` for the CWS and modify the /tftpboot/script.cust file to install the fixes on the nodes.

Mirroring considerations

Nodes with pre-PSSP V3.1 on which the rootvg VG has been mirrored are at risk of losing the mirroring on that node if the information regarding the mirroring is not entered into the SDR prior to migrating that node to PSSP 3.1. Failure to update this information in the SDR will result in the VG being unmirrored.

Migration and coexistence considerations

Table 14-1 shows service that must be applied to your existing SP system prior to migrating your CWS to PSSP 3.1. Coexistence also requires this service.

Table 14-1 Required service PTF set for migration

PSSP Level	PTF Set Required
PSSP 2.2	PTF Set 20
PSSP 2.3	PTF Set 12
PSSP 2.4	PTF Set 5

14.7.3 Control workstation requirements

The CWS has certain minimum memory requirements for PSSP 2.4 and PSSP 3.1. This does not take into account other applications that may be running on the CWS (not recommended for performance reasons).

AIX software configuration

The required AIX software level is 4.2.1 or 4.3.1 for PSSP 2.4 and 4.3.2 for PSSP 3.1. There are also some required fixes at either level that will need to be installed. Refer to the Software Requisite® section and *PSSP: Installation and Migration Guide,* GC23-3898, for documentation of these levels. AIX must be at a supported level before PSSP can be installed.

PSSP software configuration

PSSP 2.4 with PTF set 3 is the minimal required level of PSSP on the CWS in order to have 332 MHz SMP Nodes. Please refer to the Software Requisites section in *PSSP: Installation and Migration Guide,* GC23-3898, for the specific levels that are required.

NIM configuration

The NIM configuration on the CWS will also need to be updated to current levels. Refer to the Software Requisites section for information on the lppsource and SPOT configuration. Note that any additional base operating system filesets and related filesets that are installed on the existing nodes should be in the lppsource directory.

14.7.4 Node migration

This section summarizes the required steps to replace existing nodes with the new 332 MHz SMP nodes. This procedure can be done simultaneously on all nodes, or it can be performed over a period of time. The CWS will need to be

upgraded before any nodes are replaced. The majority of the time will be spent in preparation and migration of the CWS and nodes to current levels of software and the necessary backups for the nodes being replaced.

Phase I: Preparation on the CWS and existing nodes

a. Plan any necessary client and server verification testing.

b. Plan any external device verification (tape libraries, and so on).

c. Capture all required node documentation.

d. Capture all non-rootvg VG information.

e. A script may be written to back up the nodes. An example script is:

```
#/usr/bin/ksh
CWS=cws
DATE=$(date +%y%m%d)
NODE=$(hostname -s)
/usr/sbin/mount cws:/spdata/sys1/install/images /mnt
/usr/bin/mksysb -i /mnt/bos.obj.${NODE}.${DATE}
/usr/sbin/unmount /mnt
```

f. Create a full system backup for each node. Some example commands are:

```
# exportfs -i -o access=node1:node3,root=node1:node3 \
                    /spdata/sys1/install/images
# pcp -a /usr/local/bin/backup_nodes.ksh
# dsh -a /usr/local/bin/backup_nodes.ksh
```

g. Create system backup (rootvg) for the control workstation:

```
# mksysb -i /dev/rmt0
```

h. Copy required AIX filesets including PCI device filesets to the /spdata/sys1/install/aix432/lppsource directory.

i. Copy required AIX fixes including PCI device fixes to the /spdata/sys1/install/aix432/lppsource directory.

j. Copy PSSP to the /spdata/sys1/install/pssplpp/PSSP-3.1 directory.

k. Copy latest PSSP fixes to /spdata/sys1/install/pssplpp/PSSP-3.1 directory.

l. Copy coexistence fixes to /spdata/sys1/install/pssplpp/PSSP-3.1 directory if needed.

m. Create /spdata volume group backup:

```
# savevg -i /dev/rmt0 spdatavg
```

Phase II: Perform on the existing nodes

a. Perform the preparation steps.

b. Upgrade AIX as required on the CWS. Do not forget to update the SPOT if fixes were added to the lppsource directory. Perform a SDRArchive before backing up the CWS. Take a backup of the CWS after this is successfully completed.

c. Upgrade to the latest level of PSSP and latest fixes. If you plan on staying in this state for an extended period of time, you may need to install coexistence fixes on the nodes. These fixes allow nodes at earlier levels of PSSP to operate with a CWS at the latest level of PSSP. Take another backup of the CWS.

d. Verify operation of the upgraded CWS with the nodes. Perform a SDRArchive.

e. Upgrade PSSP and AIX (if needed) on the nodes that will be replaced by 332 MHz SMP nodes. Install the latest PSSP fixes.

f. Verify operation of the nodes and back up the nodes after successful verification. Archive the SDR through the **SDRArchive** command.

g. Shut down the original SP nodes that are being replaced.

h. Remove the node definitions for the nodes being replaced using the **spdelnode** command. This is to remove any of the old nodes from the SDR since the new configuration is guaranteed to be different. Now is the time to back up the /spdata volume group on the CWS.

i. Bring in and physically install the new nodes. You will move all external node connections from the original nodes to the new nodes.

Phase III: Rebuild SDR and install new 332 MHz SMP nodes

a. Rebuild SDR with all required node information on the CWS.

b. Replace old nodes with new 332 MHz SMP nodes. Be careful to cable networks, DASD, and tapes in the proper order (for example, ent1 on the old SP node should be connected to what will be ent1 on the new 332 MHz SMP Node).

c. Netboot all nodes being sure to select the correct AIX & PSSP levels.

d. Verify AIX and PSSP base code levels on nodes.

e. Verify AIX and PSSP fix levels on nodes and upgrade if necessary.

f. Verify node operation (`/usr/lpp/ssp/install/bin/node_number`, `netstat -in`).

g. You will need the node documentation acquired during the preparation step.

h. Perform any necessary external device configuration (tape drives, external volume groups, and so on).

i. Perform any necessary client and server verification testing.

j. Perform any external device verification (tape libraries, and so on).

k. Create a full system backup for nodes.

l. Create a system backup (rootvg) for the CWS.

m. Create a /spdata volume group backup.

14.8 Related documentation

We assume that you already have experience with the key commands and files from Chapter 7, “User and data management” on page 249 and Chapter 8, “Configuring the control workstation” on page 275. The following IBM manuals will help you with a detailed procedure for reconfiguring your SP system.

SP manuals

To reconfigure your SP system, you should have hands-on experience with initial planning and implementation. The manuals *RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment,* GA22-7280, and *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281, give you a good description of what you need. For details about reconfiguration of your SP system, you can refer to chapter 5 of the following two manuals: *PSSP: Installation and Migration Guide,* GC23-3898, and *PSSP Installation and Migration Guide,* GA22-7347.

Other sources

Migrating to the RS/6000 SP 332 MHz SMP Node, IBM intranet:

`http://dscrs6k.aix.dfw.ibm.com/`

14.9 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, “Answers to sample questions” on page 521.

1. In order to change the css0 IP address or hostname, you should (select more than one step):

   a. Delete and restore the NIM environment.

   b. Remove the css0 information from the SDR and reload it.

c. Change the values as required in the SDR and DNS/hosts environment.

d. Customize the nodes.

2. Your site planning representative has asked if the upgraded frame has any additional or modified environmental requirements. Therefore:

   a. The upgraded frame requires increased power.

   b. The upgraded frame has a decreased footprint.

   c. The upgraded frame is taller.

   d. The upgraded frame requires water cooling.

3. If you set up a boot/install server, and it is acting as a gateway to the control workstation, **`ipforwarding`** must be enabled. Which of the following commands will you issue to turn it on?

   a. /usr/sbin/no -ip ipforwarding=2

   b. /usr/sbin/no -l ipforwarding=1

   c. /usr/sbin/no -o ipforwarding=2

   d. /usr/sbin/no -o ipforwarding=1

4. Which of the following statements is *not* an assumption when replacing a node to PCI-based 332 MHz SMP node?

   a. Install AIX Version 4.3.2 and PSSP Version 3.1.

   b. HACWS is not implemented.

   c. The node we are migrating is not a boot/install server.

   d. There are two partitions in the SP system.

5. In order to update the microcode on the frame supervisor of frame 2, which of the following commands will you use?

   a. spsvrmgr -G -u 2:1

   b. spsvrmgr -G -u 2:0

   c. spsvrmgr -R -u 2:0

   d. spsvrmgr -R -u 2:1

14.10 Exercises

Here are some exercises you may wish to do:

1. Familiarize yourself with the steps required to add a frame.
2. Describe the steps to add an existing S70 to your SP environment.

3. Familiarize yourself with the steps required to add a node.
4. Explore the steps to add a switch to a switchless system.
5. Explore the steps to add a switch to a system with existing switches.
6. Familiarize yourself with the node migration steps to upgrade or replace an existing node.

15

Problem diagnosis

In this chapter, we discuss common problems related to node installation, SP user management, Kerberos, and SP Switches. In most of the sections, we start with the checklists, and the recovery for each problem is stated as actions rather than detailed procedures. Therefore, we recommend reading the related documents for detailed procedures to help you better understand each topic in order to resolve real world problems.

15.1 Key concepts

This section gives you the key concepts you have to understand when you prepare for the certification exam on diagnosing problems of the RS/6000 SP. You should understand:

- The basic SP hardware and software.
- The basic SP implementation process and techniques to resolve common problems.
- The overview of the **setup_server** wrapper, including NIM.
- The network boot process and how to interpret its LED for common problems.
- The mechanism of SP user management with automount and file collection and the techniques to resolve common problems.
- The basic concept of Kerberos, its setup, and the techniques to resolve common problems.
- The basic SP system connectivity and its related problems.
- The different features on the 604 high node and its problems.
- The basic SP switch operations and key commands.
- The basic techniques to resolve common SP switch problems.

15.2 Diagnosing node installation-related problems

We start with this section by introducing two types of common problems when installing the SP nodes: **setup_server** and network boot problems.

15.2.1 Diagnosing setup_server problems

The problems with **setup_server** are complicated and require reasonable understanding of each wrapper. Therefore, it is hard to make simple checklists. However, since the error messages are well indicated in the standard output while **setup_server** is running, you should carefully observe the messages and try to understand them in order to solve the problems. The probable causes for **setup_server** failure are usually three types, as follows:

- Kerberos problems
- SDR problems
- NIM-related problems

Kerberos problems in **setup_server** are usually related to the Kerberos ticket. Thus, we only discuss the problems with SDR and those that are NIM related.

Note that the **setup_server** script should run on the boot/install servers. If you have a boot/install server setup other than CWS, run **setup_server** through the **spbootins** command with **-s yes** (which is the default) on CWS, then **setup_server** will run on each boot/install server using **dsh** and return the progress message output on the CWS.

Problems with the SDR

The most common problem with the SDR on **setup_server** is that the information within the SDR is not correct. But, you should also verify the /etc/SDR_dest_info file and see if it is pointing to the correct partition IP address. Then, check all the information in the SDR with the command **splstdata** with various options. One important class of **setup_server** is **Syspar_map**. Check this with the command **SDRGetObjects Syspar_map** to find the problem.

Problems with NIM export

When **setup_server** executes, the **export_clients** wrapper exports the directories that are locations of the resources that the NIM client needs to perform the installation. Sometimes NIM cannot configure a NIM client when a NIM client definition is not entirely removed from the exported directories it manages. Here is an example of the successful export, by the **exportfs** command, of a NIM client, sp3n05, which is ready to be installed:

```
# exportfs
/spdata/sys1/install/pssplpp -ro
/spdata/sys1/install/pssp/noprompt
/spdata/sys1/install/pssp/pssp_script
/spdata/sys1/install/images/bos.obj.min.432
-ro,root=sp3n05.msc.itso.ibm.com
/export/nim/scripts/sp3n05.script -ro,root=sp3n01.msc.itso.ibm.com
/spdata/sys1/install/aix432/lppsource -ro
```

A problem occurs if the NIM client is listed in some of these directories, but the resource has not been allocated. This may happen if NIM has not successfully removed the NIM client in a previous NIM command.

To resolve this, you may follow the following procedure:

1. Check the exported file or directory with the command:

```
# exportfs
/spdata/sys1/install/pssplpp -ro
/spdata/sys1/install/aix432/lppsource -ro
/spdata/sys1/install/images/bos.obj.min.432 -ro,root=sp3n05
```

2. Un-export a file or directory with the **exportfs -u** command:

```
# exportfs -u /spdata/sys1/install/images/bos.obj.min.432
```

3. Verify that the exported directory has been removed from the export list:

```
# exportfs
/spdata/sys1/install/pssplpp -ro
/spdata/sys1/install/aix432/lppsource -ro
```

Once the NFS export has been corrected, you can issue **setup_server** on the NIM master to redefine the NIM client.

Problems with conflicting NIM Cstate and SDR

Before we discuss this problem, it is helpful to understand NIM client definition. Table 15-1 shows information on this.

Table 15-1 NIM client definition information

boot_response	Cstate	Allocations
install	BOS installation has been enabled.	spot psspspot lpp_source lppsource bosinst_data noprompt script psspscript mksysb mksysb_1
diag	Diagnostic boot has been enabled.	spot psspspot bosinst_data prompt
maintenance	BOS installation has been enabled.	spot psspspot bosinst_data prompt
disk or customize	Ready for a NIM operation.	
migrate	BOS installation has been enabled.	spot psspspot lpp_source lppsource bosinst_data migrate script psspscript mksysb mksysb_1

A NIM client may be in a state that conflicts with your intentions for the node. You may intend to install a node, but **setup_server** returns a message that the `nim -o bos_inst` command failed for this client. When **setup_server** runs on the NIM master to configure this node, it detects that the node is busy installing and does not reconfigure it. This can happen for several reasons:

- During a node NIM mksysb installation, the client node being installed was interrupted before the successful completion of the node installation.
- A node was booted in diagnostics or maintenance mode, and now you would like to reinstall it.

- The node was switched from one boot response to another.

Each of these occurrences causes the client to be in a state that appears that the node is still installing.

To correct this problem, check with the `lsnim -l <client_name>` command and issue the following command for the NIM client:

```
# nim -Fo reset <client_name>
```

It is recommended that you should always set back to **`disk`** when you switch boot response from one state to another.

Problems with allocating the SPOT resource

If you get error messages when you allocate the SPOT resources, follow these steps to determine and correct the problem:

1. Perform a check on the SPOT by issuing:

   ```
   # nim -o check spot_aix432
   ```

   This check should inform you if there is a problem.

2. If you are unable to determine the problem with the SPOT, you can update the SPOT by issuing:

   ```
   # nim -o cust spot_aix432
   ```

3. Deallocate resources allocated to clients with:

   ```
   # nim -o deallocate -a spot_aix432
   ```

4. Finally, remove the SPOT with:

   ```
   # nim -Fo remove spot_aix432
   ```

   and then run **`setup_server`** to re-create the SPOT.

Problems with creation of the mksysb resource

If **`setup_server`** cannot create the mksysb resource, verify that the specified mksysb image is in the /spdata/sys1/install/images directory.

Problems with creation of the lppsource resource

If **`setup_server`** is unable to create the lppsource resource, verify that the minimal required filesets reside in the lppsource directory:

```
# /spdata/sys1/install/aix432/lppsource
```

To successfully create the lppsource resource on a boot/install server, **`setup_server`** must acquire a lock in the lppsource directory on the CWS. Failure to acquire this lock may mean that the lock was not released properly. This lock

file contains the hostname of the system that currently has the lock and is located in /spdata/sys1/install/lppsource/lppsource.lock.

Log in to the system specified in the lock file and determine if **setup_server** is currently running. If it is not running, remove the lock file and run **setup_server** again on the system that failed to create the lppsource resource.

In another case of NIM allocation failures, you may get the following error messages:

```
0042-001 nim: processing error encountered on "master":
rshd: 0826-813 Permission is denied. rc=6.
0042-006 m_allocate: (From_Masster) rcmd Error 0
allnimres: 0016-254: Failure to allocate lpp_source resource
lppsource_defualt
from server (node_number) (node_name) to client (node_number) (node_name)
(nim -o allocate ; rc=1)
```

This failure is caused by incorrect or missing **rcmd** support on the CWS, in the ./rhosts file, for the boot/install server nodes. The ./rhosts file needs to have an entry for the boot/install server hostname when trying to execute the **allnimres** command. The **setup_server** command on the boot/install server node should correct this problem.

Problems with creation of the SPOT resource

If **setup_server** fails to create the SPOT resource, verify that the following resources are available:

1. Check if the file systems /, /tftpboot, and /tmp are full by using the command:
   `df -k`
2. Check the valid lppsource resource is available by using the command:

```
# lsnim -l lppsource
lppsource:
class = resources
type = lpp_source
server = master
location = /spdata/sys1/install/lppsource
alloc_count = 0
Rstate = ready for use
prev_state = unavailable for use
simages = yes
```

The `Rstate` is `ready for use`, and `simages` is `yes`.

If the `simages` attribute on the lppsource resource is `no`, then the required images for the support images needed to create the SPOT were not available in the lppsource resource.

If you have missing installp images from the lppsource directory, download from the AIX4.3 installation media to /spdata/sys1/install/aix432/lppsource. Then, remove the lppsource with `nim -o remove aix432` and run **`setup_server`**.

15.2.2 Diagnosing network boot process problems

This section describes the common problems on the network boot process. We introduce common checklists you need to perform, the summary of the network process, and diagnose common LED problems as examples.

Common checklists

When you have a problem with network booting, check the following:

- Check whether the cable is connected or not.
- Monitor the log file with:

  `# tail -f /var/adm/SPlogs/spmon/nc/nc.<frame_number>.<slot_number>`

  for any error.

  If the **`nodecond`** command keeps failing, try to follow the manual node conditioning procedure as shown in 9.2.21, “Network boot the boot/install server and nodes” on page 330.
- Check if there is any Kerberos error.
- Check if there is any SDR error.

Overview of network boot process

In order to resolve any network boot related problems, you may need to understand the flow of network boot process. Here, we summarize the network boot process after you issue the **`nodecond`** command.

- When **`nodecond`** exits, the node is in the process of broadcasting a bootp request.
  a. *LED 231* sends a bootp broadcast packet through the network.
  b. *LED 260* reaches the limit for not receiving a reply packet.
  c. Attempts to retrieve the boot image file.
  d. *LED 299* received a valid boot image.
  e. Continued to read the boot record from the boot image and create the RAM file system.
  f. Invokes: `/etc/init(/usr/lib/boot/ssh)`
  g. Invokes: `/sbin/rc.boot`
- After **`rc.boot`** is executed:
  a. Cannot execute `bootinfo_<platform>`, hang at LED C10.
  b. Remove unnecessary files from RAM file system.

c. Read IPL Control Block.
d. Cannot determine the type of boot, hang at LED C06.
e. *LED 600* executes **cfgmgr -fv**. Set IP resolution by /etc/hosts.
f. *LED 606* configures lo0,en0. If error, hang at LED 607.
g. *LED 608* retrieves `niminfo (/tftpboot/<reliable_hostname>)` file through tftp. If error, hang at LED 609.
h. Create /etc/hosts and Configure IP route. If error, hang at LED 613.
i. *LED 610* performs NFS mount of the SPOT file system. If error, hang at LED 611.
j. *LED 612* executes the **rc.bos_inst** script.
k. Change Mstate attribute of the NIM client object to: `in the processing of booting`
l. `LED 610` creates local mount point. If error, hang at LED 625. Attempt to NFS mount directory. If error, hang at LED 611. Clear the information attribute of the NIM client object.
m. *LED 622* links the configuration methods and executes **cfgmgr -vf** for the first and second phase.
n. Exit **/etc/rc.boot** for the first phase and start the second phase.
o. Set /etc/hosts for IP resolution and reload niminfo file.
p. Execute **rc.bos_inst** again.
q. Delete the **rc.boot** file.
r. Define the IP parameters.
s. Copy ODM objects for pre-test diagnostics.
t. Clear the information attribute of the NIM client object.
u. Invoke the **bi_main** script.

- After the **bi_main** script is invoked:

a. Invoke the initialization function and change the NIM Cstate attribute to `Base Operation System Installation is being performed`.
b. *LED C40* retrieves bosinst.data, image.data and preserve.list files and creates a file with the description of all the disks.
c. *LED C42* changes the NIM information attribute to `extract_diskette_data` and verify the existence of image.data.
d. Change the NIM information attribute to **setting_console** and set the console from the bosinst.data file. If error, hang at LED C45.
e. Change the NIM information attribute to **initialization**.
f. *LED C44* checks for available disks on the system.
g. *LED C46* validates target disk information.
h. *LED C48* executes the BOSMenus process.
i. *LED C46* initializes the log for bi_main script and sets the minimum values for LVs and file systems.

j. Prepare for restoring the operating system.
k. *LED C54* restores the base operating system.
l. *LED C52* changes the environment from RAM to the image just installed.
m. *LED C46* performs miscellaneous post-install procedures.
n. LED C56 executes BOS installs customization.
o. LED C46 finishes and reboots the system

- After **pssp_script** script is invoked:

a. *u20* creates log directory (enter function `create_directories`).
b. *u21* establishes working environment (enter function `setup_environment`).
   - *u03* gets the node.install_info file from the master.
   - *u04* expands the node.install_info file.
c. *u22* configures the node (enter function **configure_node**).
   - *u57* gets the node.config_info file from the master.
   - *u59* gets the cuat.sp template from the master.
d. *u23* Create/update /etc/ssp files (enter function **create_files**).
   - *u60* Create/update /etc/ssp files.
e. *u24* updates /etc/hosts file (enter function **update_etchosts**).
f. *u25* gets configuration files (enter function **get_files**).
   - *u61* gets /etc/SDR_dest_info from the boot/install server.
   - *u79* gets script.cust from the boot/install server.
   - *u50* gets tuning.cust from the boot/install server.
   - *u54* gets spfbcheck from the boot/install server.
   - *u56* gets **psspfb_script** from the boot/install server.
   - *u58* gets **psspfb_script** from the control workstation.
g. *u26* gets authentication files (enters the function **authent_stuff**).
   - *u67* gets /etc/krb.conf from the boot/install server.
   - *u68* gets /etc/krb.realms from the boot/install server.
   - *u69* gets **krb-srvtab** from the boot/install server.
h. *u27* updates the /etc/inittab file (enters the function `update_etcinittab`).
i. *u28* performs MP-specific functions (enters the function **upmp_work**).
   - *u52* Processor is MP.
   - *u51* Processor is UP.
   - *u55* Fatal error in bosboot.
j. *u29* installs prerequisite filesets (enters the function **install_prereqs**).
k. *u30* installs ssp.clients (enters the function `install_ssp_clients`).
   - *u80* mounts lppsource and installs **ssp.clients**.
l. *u31* installs **ssp.basic** (enters the function `install_ssp_basic`).
   - *u81* installs **ssp.basic**.
m. *u32* installs **ssp.ha** (enters the function **install_ssp_ha**).
   - *u53* installs **ssp.ha**.
n. *u33* installs **ssp.sysctl** (enters the function `install_ssp_sysctl`).
   - *u82* installs **ssp.sysctl**.
o. *u34* installs **ssp.pman** (enters the function `install_ssp_pman`).

- *u41* configures switch (enters the function **config_switch**).

p. *u35* installs **ssp.css** (enters the function **install_ssp_css**).
   - *u84* installs **ssp.css**.

q. *u36* installs **ssp.jm** (enters the function **install_ssp_jm**).
   - *u85* installs **ssp.jm**.

r. *u37* deletes the **master .rhosts** entry (enters the function delete_master_rhosts).

s. *u38* creates a new dump logical volume (enters the function **create_dump_lv**).
   - *u86* creates a new dump logical volume.

t. *u39* runs the customer's **tuning.cust** (enters the function **run_tuning_cust**).

u. *u40* runs the customer's **script.cust** (enters the function **run_script_cust**).
   - *u87* runs the customer's **script.cust** script file.
   - *u42* runs the **psspfb_script** (enters the function run_psspfb_script).

Problem with 231 LED

When the node broadcasts a bootp request, it locates the remote boot image, and it is held in /etc/bootptab, which contains the IP addresses and the location of the boot image. The boot image in /tftpboot is simply a link to the correct type of boot image for the node. This is LED231. The following message is found in *AIX V4.3 Messages Guide and Reference,* SC23-4129:

```
Display Value 231
Explanation
Progress indicator. Attempting a Normal-mode system restart from Ethernet
specified by selection from ROM menus.
System Action
The system retries.
User Action
If the system halts with this value displayed, record SRN 101-231 in item 4
on the Problem Summary Form.Report the problem to your hardware service
organization, and then stop. You have completed these procedures.
```

To resolve this, try the following:

1. Try the manual node conditioning procedure and test network connectivity
2. Check the /etc/inetd.conf and look for bootps.
3. Check the /etc/bootptab file for an entry of the problem node. Note that in multiple frame configurations if you do not define the boot/install server in the Volume_Group class, it defaults to the first node in that frame.
4. Check for the boot/install server with the **splstdata -b** command.
5. Rerun the **spbootins** command with **setup_server**.

Problem with 611 LED

At this stage of the netboot process, all the files and directories are NFS mounted in order to perform the installation, migration, or customization. The following message is found in *AIX V4.3 Messages Guide and Reference,* SC23-4129:

```
Display Value 611
Explanation
Remote mount of the NFS file system failed.
User Action
Verify that the server is correctly exporting the client file systems.
Verify that the client.info file contains valid entries for exported file
systems and server.
```

To resolve this problem, try:

1. Check whether the NIM client machine has the exported directories listed, with the command:

   ```
   # lsnim -Fl <client> | grep exported
   ```

2. Compare with the output of the **exportfs** command.
3. Verify that the directory /spdata/sys1/install/<aix_version>/spot/spot_<aix_version>/usr/sys/inst.images is not a linked directory.
4. Check, with the following command, if the image file is linked to the correct boot image file:

   ```
   # ls -l /tftpboot/sp3n06.msc.itso.ibm.com
   ```

5. If you cannot find the cause of the problem, clean up the NIM setup and exported directory and do as follows:

   a. Remove entries from /etc/exports with:

      ```
      /export/nim/scripts/*
      /spdata/*
      ```

   b. Remove NFS-related files in /etc:

      ```
      # rm /etc/state
      # rm /etc/sm/* /etc/sm.bak/*
      ```

   c. Unconfigure and reconfigure NIM:

      ```
      # nim -o unconfig master
      # installp -u bos.sysmgt.nim.master
      ```

   d. Set the node or nodes back to **install** and run **setup_server**. This will also reinstall NIM:

      ```
      # spbootins -r install -l <node#>
      ```

   e. Refresh the newly created exports list:

      ```
      # exportfs -ua
      ```

```
# exportfs -a
```

f. Refresh NFS:

```
# stopsrc -g nfs
# stopsrc -g portmap
# startsrc -g portmap
# startsrc -g nfs
```

Problems with C45 LED

When you install the node, sometimes installation hangs at LED C45. The following message is found in *AIX V4.3 Messages Guide and Reference,* SC23-4129:

```
Explanation
Cannot configure the console.
System Action
The cfgcon command has failed.
User Action
Ensure that the media is readable, that the display type is supported, and
that the media contains device support for the display type.
```

If this happens, try the following:

1. Verify which fileset contains the **cfgcon** command by entering:

   ```
   # lslpp -w | grep cfgcon
   ```

   which returns:

   ```
   /usr/lib/methods/cfgcon bos.rte.console File
   ```

2. With the following command, verify if this fileset is in the SPOT:

   ```
   # nim -o lslpp -a filesets=bos.rte.console spot_aix432
   ```

3. Check if any device fileset is missing from SPOT.
4. If there is, install an additional fileset on the SPOT and re-create the boot image files.

Problems with C48 LED

When you migrate a node, the process hang at LED C48. The following message is found in *AIX V4.3 Messages Guide and Reference,* SC23-4129:

```
Display Value c48
Explanation
Prompting you for input.
System Action
BosMenus is being run.
User Action
If this LED persists, you must provide responses at the console.
```

To resolve the problem:

1. With the following command, check NIM information:

   ```
   # lsnim -l <node_name>
   ```

2. Open tty:

   ```
   # s1term -w frame_number node_number
   ```

3. If the node cannot read the image.data file, do as follows:

   a. Check if the bos fileset exists in lppsource:

      ```
      # nim -o lslpp -a filesets=bos lppsource_aix432
      ```

   b. Check if the image.data file exists:

      ```
      # dd if=/spdata/sys1/install/aix432/lppsource/bos bs=1k count=128 |
      restore -Tvqf ./image.data
      ```

   c. Then, check the file permission on image.data.

Problems with node installation from mksysb

When you have a problem installing from a mksysb image from its boot/install server:

- Verify that the boot/install server is available:

  a. Check with the clients' boot/install server and its hostname by issuing:

     ```
     # splstdata -b
     ```

  b. **telnet** to the boot/install server if not the CWS.

  c. Look at the /etc/bootptab to make sure the node you are installing is listed in this file. If the node is not listed in this file, you should follow the NIM debugging procedure shown on page 171 of the *PSSP Diagnosis Guide,* GA22-7350.

  d. If the node is listed in this file, continue to the next step.

- Open a write console to check for console messages.

  a. At the control workstation, open a write console for the node with the install problem by issuing:

     ```
     # spmon -o node<node_number>
     ```

     or

     ```
     # s1term -w frame_number node_number
     ```

  b. Check any error message from the console that might help determine the cause of the problem. Also, look for NIM messages that might suggest that the installation is proceeding. An example of a NIM progress message is:

     ```
     / step_number of total_steps complete
     ```

which tells how many installation steps have completed. This message is accompanied by an LED code of u54.

- Check to see if the image is available and the permissions are appropriate by issuing:

  ```
  # /usr/lpp/ssp/bin/splstdata -b
  ```

  The `next_install_image` field lists the name of the image to be installed. If the field for this node is set to default, the default image specified by the **`install_image`** attribute of the SP object will be installed. The images are found in the /spdata/sys1/install/images directory. You can check the images and their permissions by issuing:

  ```
  # ls -l /spdata/sys1/install/images
  ```

  This should return:

  ```
  total 857840
  -rw-r--r-- 1 root sys 130083840 Jan 14 11:15 bos.obj.ssp.4.3
  ```

  The important things to check are that the images directory has execute (**x**) permissions by all, and that the image is readable (**r**) by all.

  The **`setup_server`** script tries to clean up obsolete images on install servers. If it finds an image in the /spdata/sys1/install/images directory that is not needed by an install client, it deletes the image. However, **`setup_server`** deletes images on the control workstation only if the site environment variable REMOVE_IMAGES is true.

- Review the NIM configuration and perform NIM diagnostics for this Node.

15.3 Diagnosing SDR problems

This section shows a few common problems related to SDR and its recovery actions.

15.3.1 Problems with connection to server

Sometimes, when you change system or network and issue SDR command, such as **`splstdata -b`** on the node, you get the error message: `failing to connect to server`. If so, try the following:

1. Type **`spget_syspar`** on the node showing the failing SDR commands.
2. If the **`spget_syspar`** command fails, check the /etc/SDR_dest_info file on the same node. It should have at least two records in it. These records are the primary and the default records. They should look like this:

   ```
   # cat SDR_dest_info
   default:192.168.3.130
   ```

```
primary:192.168.3.130
nameofdefault:sp3en0
nameofprimary:sp3en0
```

If this file is missing or does not have these two records, the node may not be properly installed, or the file has been altered or corrupted. You can edit the file that contains the two records above or copy the file from a working node in the same system partition.

3. If the **spget_syspar** command is successful, check to make sure that the address is also the address of a valid system partition. If it is, try to ping that address. If the ping fails, contact your system administrator to investigate a network problem.
4. If the value returned by the **spget_syspar** command is not the same as the address in the primary record of the /etc/SDR_dest_info file, the SP_NAME environment variable is directing SDR requests to a different address. Make sure that this address is a valid system partition.
5. If the value of the SP_NAME environment variable is a hostname, try setting it to the equivalent dotted decimal IP address.
6. Check for the existence of the SDR server process (**sdrd**) on the CWS with:

   ```
   # ps -ef | grep sdrd
   ```

   If the process is not running, do the following:

   – Check the **sdrd** entry in the file /etc/inittab on the control workstation. It should read:

     ```
     sdrd:2:once:/usr/bin/startsrc -g sdr
     ```

   – Check the SDR server logs in /var/adm/SPlogs/sdr/sdrdlog. **<server_ip>.pid**, where **pid** is a process ID.

   – Issue `/usr/bin/startsrc -g sdr` to start the SDR daemon.

15.3.2 Problem with class corrupted or non-existent

If an SDR command ends with `RC=102` (internal data format inconsistency) or `026` (class does not exist), first make sure the class name is spelled correctly and the case is correct (see the table of classes and attributes in “The System Data Repository” appendix in *PSSP Administration Guide,* SA22-7348. Then follow the steps in “SDR Shadow Files” in the System Data Repository appendix in *PSSP Administration Guide,* SA22-7348.

Then, check if the /var file system is full. If this is the case, either define more space for /var or remove unnecessary files.

15.4 Diagnosing user access-related problems

As you have seen from the previous chapter, AMD is changed to AIX automount starting with PSSP 2.3. Thus, we briefly discuss general AMD checklists (for PSSP 2.2 or earlier) and extend the discussion to user access and AIX Automount problems.

15.4.1 Problems with AMD

- Check if the AMD daemon is running. If not, restart it with:
  `/etc/amd/amd_start`
- Make sure that the user's home directories are exported. If not, update /etc/exports and run the **`exportfs -a`** command.
- Check the /etc/amd/amd-maps/amd.u AMD map file for the existence of an user ID if you have problems with logging on to the system. An entry should look like this:

```
netinst type:=link;fs:=/home
.....
efri host==sp3en0;type:=link;fs:=/home/sp3en0 \
host!=sp3en0;type:=nfs;rhost:=sp3en0;rfs:=/home/sp3en0
```

- If there is no entry for the user ID you would like to use, add it to this file. Make sure that the updates are distributed after the change by issuing:

```
# dsh -w <nodelist> supper update user.admin sup.admin power_system
```

  Check whether the network connection is still working.

- Get the information about the AMD mounts by issuing the **`/etc/amd/amq`** command. If the output of **`amq`** looks as follows:

```
amq: localhost: RPC: Program not registered
```

  your problem could be:

  - The AMD daemon is not running.
  - The portmap daemon is not running.
  - The AMD daemon is waiting for a response from the NFS server that is not responding.

  Make sure that the portmap daemon is running and that your NFS server is responding. If the portmap daemon is inoperative, start it with the **`startsrc -s portmap`** command.

  If you have an NFS server problem, check the amd.log file located in the /var/adm/SPlogs/amd directory.

  Stop AMD by issuing **`kill -15 <amd_pid>`**, solve your NFS problems, and start AMD again with **`/etc/amd/amd_start`**.

- If you have user access problems, do the following:
  - Verify that the login and rlogin options for your user are set to `true`.
  - Check the user path or .rhosts on the node. If you have problems executing **rsh** to the node, check the user path to see if the user is supposed to be a Kerberos principal.
- If you have problems executing an SP user administrative command, you may get an error message similar to the following:

  ```
  0027-153 The user administration function is already in use.
  ```

  In this case, the most probable cause is that another user administrative command is running, and there is a lock in effect for the command to let it finish. If no other administrative command is running, check the /usr/lpp/ssp/config/admin directory for the existence of a .userlock file. If there is one, remove it and try to execute your command again.

15.4.2 Problems with user access or automount

This section shows a few examples about the problems logging into SP system or accessing user's home directories.

Problems with logging in an SP node by a user

Check the /etc/security/passwd file. If a user is having problems logging in to nodes in the SP System, check the login and rlogin attributes for the user in the /etc/security/passwd file on the SP node.

Check the Login Control facility to see whether the user's access to the node has been blocked. The system administrator should verify that the user is allowed access. The system administrator may have blocked interactive access so that parallel jobs could run on a node.

Problems with accessing users' directories

When you have a problem accessing a user's directory, verify that the automount daemon is running.

To check whether the automount daemon is running or not, issue:

```
# ps -ef | grep automount
```

for AIX 4.3.0 or earlier systems, and

```
# lssrc -g autofs
```

for AIX 4.3.1 or later systems.

Note: On AIX 4.3.1 and later systems, the AutoFS function replaces the automount function of AIX 4.3.0 and earlier systems. All automount functions are compatible with AutoFS. With AutoFS, file systems are mounted directly to the target directory instead of using an intermediate mount point and symbolic links

If automount is not running, check with the **mount** command to see if any automount points are still in use. If you see an entry similar to the following one, there is still an active automount mount point. For AIX 4.3.0 or earlier systems:

```
# mount
sp3n05.msc.itso.ibm.com (pid23450@/u) /u afs Dec 07 15:41 ro,noacl,ignore
```

For AIX 4.3.1 and later systems:

```
# mount
/etc/auto/maps/auto.u /u autofs Dec 07 11:16 ignore
```

If the **mount** command does not show any active mounts for automount, issue the following command to start the automounter:

```
# /etc/auto/startauto
```

If this command succeeds, issue the previous **ps** or **lssrc** command again to verify that the automount daemon is actually running. If so, verify that the user directories can be accessed or not.

Note that the automount daemon should be started automatically during boot. Check to see if your SP system is configured for automounter support by issuing:

```
# splsdata -e | grep amd_config
```

If the result is true, you have automounter support configured for the SP in your Site Environment options.

If the **startauto** command was successful, but the automount daemon is still not running, check to see if the SP automounter function has been replaced by issuing:

```
# ls -l /etc/auto/*.cust
```

If the result of this command contains an entry similar to:

```
-rwx ----- 1 root system 0 Dec 12 13:20 startauto.cust
```

the SP function to start the automounter has been replaced. View this file to determine which automounter was started and follow local procedures for diagnosing problems for that automounter.

If the result of the **ls** command does not show any executable user customization script, check both the automounter log file /var/adm/SPlogs/auto/auto.log and the daemon log file /var/adm/SPlogs/SPdaemon.log for error messages.

If the **startauto** command fails, find the reported error messages in the *PSSP: Messages Reference*, GA22-7352, and follow the recommended actions. Check the automounter log file /var/adm/SPlogs/auto/auto.log for additional messages. Also, check the daemon log file /var/adm/SPlogs/SPdaemon.log for messages that may have been written by the automounter daemon itself.

If automounter is running, but the user cannot access user files, the problem may be that automount is waiting for a response from an NFS server that is not responding or that there is a problem with a map file. Check the /var/adm/SPlogs/SPdaemon.log for information relating to NFS servers not responding.

If the problem does not appear to be related to an NFS failure, you will need to check your automount maps. Look at the /etc/auto/maps/auto.u map file to see if an entry for the user exists in this file.

Another possible problem is that the server is exporting the file system to an interface that is not the interface from which the client is requesting the mount. This problem can be found by attempting to mount the file system manually on the system where the failure is occurring.

Stopping and restarting automount

If you have determined that you need to stop and restart the automount daemon, the cleanest and safest way is to reboot the system. However, if you cannot reboot the system, use the following steps:

For AIX 4.3.0 or earlier systems:

1. Determine whether any users are already working in directories mounted by the automount daemon. Issue:

   ```
   # mount
   ```

2. Stop the automount daemon:

   ```
   # kill -15 process_id
   ```

   where `process_id` is the process number listed by the previous mount command.

Note : It is important that you *do not* stop the daemon with the kill -kill or kill -9. This will prevent the automount daemon from cleaning up its mounts and releasing its hold on the file systems. It may cause file system hangs and force you to reboot your system to recover those file systems

3. Start the automount daemon:

   ```
   # /etc/auto/startauto
   ```

   You can verify that the daemon is running by issuing the previous **mount** or **ps** commands.

For AIX 4.3.1 or later systems:

1. Determine whether any users are already working on the directories mounted by the autmountd daemon with the command: `mount`
2. Stop the automountd daemon with this command:

   ```
   # stopsrc -g autofs
   ```

3. Restart the autmounter:

   ```
   # /etc/auto/startauto
   ```

   You can verify that the daemon is running by issuing the previous **lssrc** command.

15.5 Diagnosing file collection problems

In this section, we summerize common checklists for file collection problems and explain how you can resolve them.

15.5.1 Common checklists

The following check lists give you an idea of what to do when you get error messages related to the file collection problems:

- Check the TCP/IP configuration because file collection uses the Ethernet network (en0). Check the en0 adapter status or routes if you have boot/install server exists and test it with the **ping** command from client to server. Also, check the hostname resolution with **nslookup** if DNS is setup.
- Check if the file collection is resident or not by issuing the **supper status** command. The output from the command looks like this:

  ```
  # /var/sysman/supper status

  Collection         Resident  Access Point Filesystem          Size
  ```

```
=====================================================================
 node.root Yes    /                          -                  -
 power_system Yes    /share/power/system -
 sup.admin Yes    /var/sysman                -                  -
 user.admin Yes    /                          -                  -
=====================================================================
```

If the update of the file collection failed, and this file collection is not resident on the node, install it by issuing the command:

```
# supper install <file collection>
```

- Check if the file collection server daemon is running on the CWS and boot/install server:

  On the CWS:

```
[sp3en0:/]# ps -ef | grep sup
    root  10502   5422   0   Dec 03      -  0:00
/var/sysman/etc/supfilesrv -p /var/sysman/sup/supfilesrv.pid

# dsh -w sp3n01 ps -ef | grep sup

sp3n01:     root  6640 10066   0 10:44:21      -  0:00
/var/sysman/etc/supfilesrv -p /var/sysman/sup/supfilesrv.pid
```

- Use **dsh /var/sysman/supper where** on the CWS to see which machine is each node's supper server, as follows:

```
[sp3en0:/]# dsh -w sp3n01,sp3n05 /var/sysman/supper where
sp3n01: supper:  Collection node.root would be updated from server
sp3en0.msc.itso.ibm.com.
sp3n01: supper:  Collection power_system would be updated from server
sp3en0.msc.itso.ibm.com.
sp3n01: supper:  Collection sup.admin would be updated from server
sp3en0.msc.itso.ibm.com.
sp3n01: supper:  Collection user.admin would be updated from server
sp3en0.msc.itso.ibm.com.
sp3n05: supper:  Collection node.root would be updated from server
sp3n01en1.msc.itso.ibm.com.
sp3n05: supper:  Collection power_system would be updated from server
sp3n01en1.msc.itso.ibm.com.
sp3n05: supper:  Collection sup.admin would be updated from server
sp3n01en1.msc.itso.ibm.com.
sp3n05: supper:  Collection user.admin would be updated from server
sp3n01en1.msc.itso.ibm.com.
```

- Check the server has the supman user ID created.
- Check the /etc/services file on the server machine as follows:

```
[sp3en0:]# grep sup /etc/services
```

```
supdup          95/tcp
supfilesrv      8431/tcp
```

- Check whether the supfilesrv daemon is defined and that it has a correct port.
- Check the log files located in the /var/sysman/logs directory.
- Check the log files located in the /var/adm/SPlogs/filec directory.

15.6 Diagnosing Kerberos problems

In this section, we summarize the common checklist of Kerberos problems. Then, we describe possible causes and the action needed to be taken to resolve them. In addition, we briefly describe the difference between PSSP v2 and PSSP v3.

15.6.1 Common checklists

Before we start the Kerberos problem determination, we recommand checking the following list:

- Check that the hostname resolution is OK or not whether you are using DNS or the local host file. Remember the encrypted Kerberos service key is created with hostname.
- Check your Kerberos ticket by issuing the **klist** or **k4list** command. If a ticket is expired, destroy it with the **kdestroy** or **k4destroy** command and reissue it with the command **kinit** or **k4init** as follows:

  ```
  # k4init root.admin
  ```

  Then, type the Kerberos password twice.
- Check the /.klogin file.
- Check the PATH variable whether Kerberos commands are in the environment PATH.
- Check your file systems by using the **df -k** command. Remember that **/var** contains a Kerberos database and **/tmp** contains a ticket.
- Check the date on the authentication server and clients. (Kerberos can handle only a five minute difference.)
- Check if the Kerberos daemons are running on the control workstation.
- Check /etc/krb.realms on the client nodes.
- Check if you have to recreate /etc/krb-srvtab on the node.
- Check /etc/krb-srvtab on the authentication server.

15.6.2 Problems with a user's principal identity

An example of a bad Kerberos name format generates the following error message:

```
sp3en0 # k4init
Kerbero Initialization
Kerberos name: root.admin
k4list: 2502-003 Bad Kerberos name format
```

The probable causes are a bad Kerberos name format, a Kerberos principal does not exist, an incorrect Kerberos password, or a corrupted Kerberos database. Recovery action is to repeat the command with the correct syntax. An example is:

```
# k4init root.admin
```

Another example is a missing root.admin principal in the /.klogin file on the control workstation as follows:

```
sp3n05 # dsh -w sp3en0 date
sp3en0:krshd:Kerberos Authentication Failed:User
root.admin@MSC.ITSO.IBM.COM is not authorized to login to account root.
sp3en0: spk4rsh: 0041-004 Kerberos rcmd failed: rcmd protocol failure.
```

Check the /.klogin file if it has entry for the user principal. If all the information is correct, but the Kerberos command fails, suspect a database corruption.

15.6.3 Problems with a service's principal identity

When a /etc/krb-srvtab file is corrupted on an node, and the remote command service (**rcmd**) fails to work from the control workstation, we have the following error message:

```
sp3en0 # dsh -w sp3n05 date
sp3n05:krshd:Kerberos Authentication Failed.
sp3n05: spk4rsh: 0041-004 Kerberos rcmd failed: rcmd protocol failure.
```

The probable causes for this problem are the krb-srvtab file does not exist on the node or on the control workstation, the krb-srvtab has the wrong key version, or krb-srvtab file is corrupted. Analyze the error messages to confirm the service's principal identity problem. Make sure the /.klogin file, /etc/krb.realms, and /etc/krb-conf files are consistent with those of the Kerberos authentication server.

15.6.4 Problems with authenticated services

When hardmon is having problems due to a Kerberos error, we have the following message:

```
sp3en0 # spmon -d
Opening connection to server
0026-706 Cannot obtain service ticket for hardmon.sp3en0
Kerberos error code is 8, Kerberos error message is:
2504-008 Kerberos principal unknown
```

The probable causes are that the ticket has expired, a valid ticket does not exists, the host name resolution is not correct, or the ACL files do not have correct entries. Destroy the ticket using **k4destroy** and issue a new ticket by issuing `k4init root.admin` if the user is root. Then, check the hostname resolution, ACL files, and the Kerberos database.

15.6.5 Problems with Kerberos database corruption

The database can be corrupted for many reasons, and messages also vary based on the nature of the corruption. Here, we provide an example of messages received because of Kerberos database corruption:

```
sp3en0 # k4init root.admin
Kerberos Initialization for "root.admin"
k4init: 2504-010 Kerberos principal has null key
```

Rebuild the Kerberos database as follows:

1. Ensure the following directories are included in your PATH:
   - /usr/lpp/ssp/kerberos/etc
   - /usr/lpp/ssp/kerberos/bin
   - /usr/lpp/ssp/bin
2. On the CWS, login as root and execute the following commands:

   ```
   # /usr/lpp/ssp/kerberos/bin/kdestroy
   ```

   The **kdestroy** command destroys the user's authentication tickets that are located in /tmp/tkt$uid.
3. Destroy the Kerberos authentication database, which is located in /var/kerberos/*:

   ```
   # /usr/lpp/ssp/kerberos/etc/kdb_destroy
   ```
4. Remove the following files:
   - krb-srvtab: Contains the keys for services on the nodes
   - krb.conf: Contains the SP authentication configuration
   - krb.realms: Specifies the translations from host names to authentication realms:

     ```
     # rm /etc/krb*
     ```

5. Remove the .klogin file that contains a list of principals that are authorized to invoke processes as the root user with the SP-authenticated remote commands **rsh** and **rcp**:

```
# rm /.klogin
```

6. Remove the Kerberos Master key cache file:

```
# rm /.k
```

7. Insure that the authentication database files are completely removed:

```
# rm /var/kerberos/database/*
```

8. Change the /etc/inittab entries for Kerberos:

```
# chitab "kadm:2:off:/usr/lpp/ssp/kerberos/etc/kadmind -n"
# chitab "kerb:2:off:/usr/lpp/ssp/kerberos/etc/kerberos"
```

9. Refresh the /etc/inittab file:

```
# telinit q
```

10. Stop the daemons:

```
# stopsrc -s hardmon
# stopsrc -s splogd
```

11. Configure SP authentication services:

```
# /usr/lpp/ssp/bin/setup_authent
```

    This command will add the necessary remote command (RCMD) principals for the nodes to the Kerberos database based on what is defined in the SDR for those nodes.

12. Set the node's bootp response to `customize` and run **setup_server**:

```
# sbootins -r customize -l <nodelist>
```

13. Reboot the nodes.

    After the node reboots, verify that the bootp response toggled back to disk.

14. Start the **hardmon** and **splogd** on the CWS:

```
# startsrc -s hardmon
# startsrc -s splogd
```

After step 12 and step 13 are done, the /etc/krb-srvtab files are distributed onto the nodes. However, if you cannot reboot the system, do as follows:

1. After running the command:

```
# spbootins -r customize -l <nodelist>
```

2. On the CWS, change the directory to the /tftpboot and verify that there is a <node_name>-new-srvtab file for each node

3. FTP in binary mode to each node's respective /tftpboot/<node-name>-new-srvtab file from the CWS to the nodes and rename the file to /etc/krb-srvtab.
4. Set the nodes back to disk on the CWS:

```
# spbootins -r disk -l <nodelist>
```

15.6.6 Problems with decoding authenticator

When you change the host name and do not follow the procedure correctly, sometimes /etc/krb-srvtab file produces an error, and you may see the following message:

```
kshd:0041-005 kerberos rsh or rcp failed:
2504-031 Kerberos error: can't decode authenticator
```

Recreate the /etc/krb-srvtab file from the boot/install server, and propagate it to the node. If you can reboot the node, simply set boot_response to `customize`, and reboot the node. Otherwise, do as follows:

On the control workstation, run **`spbootins`** by setting boot_response to: `customize`

```
# spbootins -r customize -l <node_list>
```

Then, on the control workstation, change the directory to /tftpboot and verify the <node_name>-new-srvtab file. FTP this file to the node's /etc, and rename the file to krb-srvtab. Then set the node back to **`disk`** as follows:

```
# spbootins -r disk -l <node_list>
```

15.6.7 Problems with the Kerberos daemon

Here is an example of messages when the Kerberos daemons are inactive because of missing krb.realms files on the control workstation. This message is an excerpt of admin_server.syslog file:

```
03-Dec-98 17:47:52 Shutting down admin server
03-Dec-98 17:48:15 kadmind:
2503-001 Could not get local realm.
```

Check all the Kerberos file exists on the authentication server that is usually the control workstation. Check the contents of the file to make sure the files are not corrupted. Check the log /var/adm/SPlogs/kerberos for messages related to Kerberos daemons.

15.7 Diagnosing system connectivity problems

This section shows a few examples related to network problems.

15.7.1 Problems with network commands

If you can not access the node using **rsh**, **telnet**, **rlogin**, or **ping**, you can access the node using the tty. This can be done by using the Hardware Perspectives, selecting the node, and performing an open tty action on it. It can also be done by issuing the `s1term -w frame number slot number` command, where frame number is the frame number of the node, and the slot number is the slot number of the node.

Using either method, you can login to the node and check the hostname, network interfaces, network routes, and hostname resolution to determine why the node is not responding.

15.7.2 Problems with accessing the node

If you can not access the node using **telnet** or **rlogin**, but can access the node using **ping**, then this is a probable software error. Initiate a dump, record all relevant information, and contact the IBM Support Center.

15.7.3 Topology-related problems

If the **ping** and **telnet** commands are successful, but *hostresponds* still shows the node not responding, there may be something wrong with the Topology Services (hats) subsystem. Perform these steps:

1. Examine the en0 (Ethernet adapter) and css0 (switch adapter) addresses on all nodes to see if they match the addresses in /var/ha/run/hats. partition_name/machines.lst.
2. Verify that the netmask and broadcast addresses are consistent across all nodes. Use the **ifconfig en0** and **ifconfig css0** commands.
3. Check the hats log file on the failing node with the command:

```
# cd /var/ha/log
# ls -lt | grep hats
-rw-rw-rw-   1 root     system     31474 Dec 07 09:26 hats.04.104612.sp3en0
-rwxr-xr-x   1 root     system        40 Dec 04 10:46 hats.sp3en0
-rw-rw-rw-   1 root     system     12713 Dec 04 10:36 hats.04.103622.sp3en0
-rw-rw-rw-   1 root     system    319749 Dec 04 10:36 hats.03.141426.sp3en0
-rw-rw-rw-   1 root     system    580300 Dec 04 03:13
hats.03.141426.sp3en0.bak
```

4. Check the hats log file for the Group Leader node. Group Leader nodes are those that host the adapter whose address is listed below the line **Group ID** in the output of the **lssrc -ls hats** command.
5. Delete and add the hats subsystem with the following command on the CWS:

   ```
   # syspar_ctrl -c hats.sp3en0
   ```

   Then:

   ```
   # syspar_ctrl -A hats.sp3en0
   ```

   or, on the nodes:

   ```
   # syspar_ctrl -c hats
   ```

   Then:

   ```
   # syspar_ctrl -A hats
   ```

15.8 Diagnosing 604 high node problems

This section provides information on:

- 604 high node characteristics, including:
  - Addressing power and fan failures in the 604 high node
  - Rebooting the 604 high node after a system failure
- Error conditions and performance considerations
- Using SystemGuard and BUMP programs

15.8.1 604 high node characteristics

The 604 high node operation is different from other nodes in several areas:

- A power feature is available that adds a redundant internal power supply to the node. In this configuration, the node will continue to run in the event of a power supply failure. Error notification for a power supply failure is done through the AIX Error Log on the node.
- The cooling system on the node also has redundancy. In the event that one of the cooling fans fails, the node will continue to run. Error notification for a power supply failure is done through the AIX Error Log on the node.
- If a hardware related crash occurs on the node, SystemGuard will re-IPL the node using the long IPL option. During long IPL, some CPU or memory resources may be deconfigured by SystemGuard to allow the re-IPL to continue.

15.8.2 Error conditions and performance considerations

You need to be aware of the following conditions that pertain to the unique operation of this node:

- An error notification object should be set up on the node for the label EPOW_SUS. The EPOW_SUS label is used on AIX Error Log entries that may pertain to the loss of redundant power supplies or fans.
- If the node is experiencing performance degradation, you should use the **lscfg** command to verify that none of the CPU resources have been deconfigured by SystemGuard if it may have re-IPLed the node using the long IPL option.

15.8.3 Using SystemGuard and BUMP programs

SystemGuard is a collection of firmware programs that run on the bringup microprocessor (BUMP). SystemGuard and BUMP provide service processor capability. They enable the operator to manage power supplies, check system hardware status, update various configuration parameters, investigate problems, and perform tests.

The BUMP controls the system when the power is off or the AIX operating system is stopped. The BUMP releases control of the system to AIX after it is loaded. If AIX stops or is shut down, the BUMP again controls the system.

To activate SystemGuard, the key mode switch must be in the SERVICE position during the standby or initialization phases. The standby phase is any time the system power is off. The initialization phase is the time when the system is being initialized. The PSSP software utilizes SystemGuard IPL flags, such as the FAST IPL default, when the netboot process starts.

15.8.4 Problems with physical power-off

If the 604 high node was physically powered off from the front panel power switch and not powered back on using the front panel switch, try as follow:

1. Using **spmon**, set the key to **service** mode.
2. Open a tty console with `spmon -o node<node_number>`.
3. Type at the prompt > `sbb`
4. On the BUMP processor menu, choose option **5**:

```
STAND-BY MENU :  rev 17.03
0 Display Configuration
1 Set Flags
2 Set Unit Number
```

```
3 Set Configuration
4 SSbus Maintenance
5 I2C Maintenance
Select(x:exit): 5
```

5. Select option **08** (I2C Maintenance):

```
I2C  Maintenance
00 rd OP status                 05 wr LCD
01 rd UNIT status               06 rd i/o port SP
02 rd EEPROM                    07 fan speed
03 margins                      08 powering
04 on/off OP LEDs
Select(x:exit):  08
```

6. Select option **02** and option **0**:

```
powering
00 broadcast ON
01 broadcast OFF
02 unit      ON
03 unit      OFF
Select(x:exit): 02
Unit (0-7): 0
```

7. At this point, the power LED should indicate on (does not blink), but the node will not power up.
8. Physically click the power switch (off and then on) on the node. The node should now boot in SERVICE mode.
9. After the node boots successfully, using **spmon -k normal <node_number>** to set the node key position to NORMAL on CWS, power off the node logically (not physically), and then power the node on.

15.9 Diagnosing switch problems

In this section, we discuss typical problems related to the SP Switch that you should understand to prepare for your exam. If your system partition has an SP Switch failure with the following symptoms, perform the appropriate recovery action described.

15.9.1 Problems with Estart failure

The **Estart** problems are caused by many different reasons. In this section, we discuss the following typical symptoms.

Symptom 1: System cannot find Estart command

Software installation and verification is done using the **CSS_test** script from either the SMIT panel or from the command line.

Run **CSS_test** from the command line. You can optionally select the following options:

-q To suppress messages.
-l To designate an alternate log file.

Note that if **CSS_test** is executed following a successful **Estart**, additional verification of the system will be done to determine if each node in the system or system partition can be pinged. If you are using system partitions, **CSS_test** runs in the active partition only.

Then review the default log file, which is located at /var/adm/SPlogs/css/CSS_test.log to determine the results.

Additional items to consider while trying to run **CSS_test** are as follows:

- Each node should have access to the /usr/lpp/ssp directory.
- /etc/inittab on each node should contain an entry for **rc.switch**.

For complete information on **CSS_test**, see page 56 in *PSSP Command and Technical Reference (2 Volumes),* SA22-7351.

Symptom 2: Primary node is not reachable

If the node you are attempting to verify is the primary node, start with Step 1. If it is a secondary node, start with Step 2.

1. Determine which node is the primary by issuing the **Eprimary** command on the CWS:

   ```
   Eprimary
   ```

   returns:

   ```
   1 - primary
   1 - oncoming primary
   15 - primary backup
   15 - oncoming primary backup
   ```

   If the command returns an oncoming primary value of none, reexecute the **Eprimary** command specifying the node you would like to have as the primary node. Following the execution of the **Eprimary** command (to change the oncoming primary), an **Estart** is required to make the oncoming primary node the primary.

If the command returns a primary value of none, an **Estart** is required to make the oncoming primary node the primary.

The primary node on the SP Switch system can move to another node if a primary node takeover is initiated by the backup. To determine if this has happened, look at the values of the primary and the oncoming primary backup. If they are the same value, then a takeover has occurred.

2. Ensure that the node is accessible from the control workstation. This can be accomplished by using **dsh** to issue the **date** command on the node as follows:

```
# /usr/lpp/ssp/rcmd/bin/dsh -w <problem hostname> date
TUE Oct 22 10:24:28 EDT 1997
```

   If the current date and time are not returned, check the Kerberos or remote command problem.

3. Verify that the switch adapter (css0) is configured and is ready for operation on the node. This can be done by interrogating the `adapter_config_status` attribute in the **switch_responds** object of the SDR:

```
# SDRGetObjects switch_responds node_number==<problem node number>
```

   returns:

```
node_number switch_responds autojoin isolated adapter-config_status
1 0 0 0 css_ready
```

   If the `adapter_config_status` object is anything other than **css_ready**, see P223 of *RS/6000 SP: PSSP 2.2 Survival Guide*, SG24-4928.

   Note: To obtain the value to use for problem node number, issue an SDR query of the node_number attribute of the Node object as follows:

```
# SDRGetObjects Node reliable_hostname==<problem hostname> node_number
```

   returns:

```
node_number
1
```

4. Verify that the fault_service_Worm_RTG_SP daemon is running on the node. This can be accomplished by using **dsh** to issue a **ps** command to the problem node as follows:

```
 # dsh -w <problem_hostname> ps -e | grep Worm
18422  -0:00 fault_service_Worm_RTG
```

   If the fault_service_Worm_RTG_SP daemon is running, SP Switch node verification is complete.

   If the fault_service_Worm_RTG_SP daemon is not running, try to restart it with: `/usr/lpp/ssp/css/rc.switch`

Symptom 3: Estart command times out or fails

Refer to the following list of steps to diagnose **Estart** failures:

1. Log in to the primary node.
2. View the bottom of the /var/adm/SPlogs/css/fs_daemon_print.file.
3. Use the failure listed to index from the Table 19 on the P133 of the *PSSP Diagnosis Guide*, **GA22-7350**.

If the message from the /var/adm/SPlogs/css/fs_daemon_print.file is not clear, we suggest to do the following before contacting IBM Software support:

- Check SDR with **SDR_test**.
- Run `SDRGetObjects switch_responds` to read the SDR switch_responds class and look for the values of the adapter_config_status attribute.
- Run `Etopology -read <file_name>`. Compare the output of the topology file with the actual cabling and make sure all the entries are correct.
- Make sure the Worm daemon is up and running on all the nodes. Check the worm.trace file on the primary node for Worm initialization failure.
- Make sure the Kerberos authentication is correct for all the nodes.
- Run **Eclock -d**, and bring the Worm up on all nodes executing the `/usr/lpp/ssp/css/rc.switch` script.
- Change the primary node to a different node using the **Eprimary** command. In changing the primary node, it is better to select a node attached to a different switch chip from the original primary or even a different switch board.
- Check if all the nodes are fenced or not. Use the `SDRChangeAttrValues` command as follows to unfence the primary and oncoming primary. Note that the command `SDRChangeAttrValues` is dangerous if you are not using it properly. It is recommended to archive SDR before using this command.

```
# SDRChangeAttrValues switch_responds node_number==<primary node_num>
isolated=0
```

- Now try **Estart**. If it fails, contact IBM Software support.

Symptom 4: Some nodes or links not initialized

When evaluating device and link problems on the system, first examine the out.top file in the /var/adm/SPlogs/css directory of the primary node. This file looks like a switch topology file except for the additional comments on lines where either the device or link is not operational.

These additional comments are appended to the file by the fault_service daemon to reflect the current device and link status of the system. If there are no comments on any of the lines, or the only comments are for wrap plugs where

they actually exist, you should consider all devices and links to be operational. If this is not the case, however, the following information should help to resolve the problem.

The following is an example of a failing entry in the out.top file:

```
s 14 2 tb3 9 0 E01-S17-BH-J32 to E01-N10 -4 R: device has been removed from
network-faulty (link has been removed from network or miswired-faulty)
```

This example means the following:

- Switch chip 14, port 2 is connected to switch node number 9.
- The switch is located in frame E01 slot 17.
- Its bulkhead connection to the node is jack 32.
- The node is also in frame E01, and its node number is 10.
- The -4R refers to the device status of the right side device (tb0 9), which has the more severe device status of the two devices listed. The device status of the node is `device has been removed from the network - faulty`.
- The link status is `link has been removed from the network or miswired -faulty`.

For a detail list of possible device status for SP switch, refer to P119-120 of the *PSSP Diagnosis Guide,* GA22-7350.

15.9.2 Problem with pinging to SP Switch adapter

If the SP node fails to communicate over the switch, but its switch_responds is on and **`ping`** or **`CSS_test`** commands fail. Check the following:

To isolate an adapter or switch error for the SP Switch, first view the AIX error log. For switch related errors, log in to the primary node; for adapter problems, log in to the suspect node. Once you are logged in, enter the following:

```
# errpt | more
ERROR_ID TIMESTAMP T CL Res Name ERROR_Description
34FFBE83 0604140393T T H Worm Switch Fault-detected by switch chip
C3189234 0604135793 T H Worm Switch Fault-not isolated
```

The Resource Name (Res Name) in the error log should give you an indication of how the failure was detected. For details, refer to Table 17 and Table 18 on pp.121-132 of *PSSP Diagnosis Guide,* GA22-7350.

15.9.3 Problems with Eunfence

The **`Eunfence`** command first distributes the topology file to the nodes before they can be unfenced. But, if the command fails to distribute the topology file, it

puts an entry in the dist_topology.log file on the primary node in the /var/adm/SPlogs/css directory.

The **Eufence** command fails to distribute the topology file if the Kerberos authentication is not correct.

The **Eunfence** command will time out if the Worm daemon is not running on the node. So, before running the **Eunfence** command, make sure the Worm daemon is up and running on the node. To start the Worm daemon on the node, it is required that you run the `/usr/lpp/spp/css/rc.switch` script.

If the problem persists after having correct Kerberos authentication, and the Worm daemon is running, the next step is to reboot the node. Then, try the **Eunfence** command again.

If neither of the previous steps resolve the problem, you can run diagnostics to isolate a hardware problem on the node.

The last resort, if all fails, would be to issue an **Eclock** command. This is completely disruptive to the entire switch environment; so, it should only be issued if no one is using the switch. An **Estart** must be run after **Eclock** completes.

15.9.4 Problems with fencing primary nodes

If the oncoming primary node becomes fenced from the switch use the following procedure to **Eunfence** it prior to issuing **Estart**:

- If the switch is up and operational with another primary node in control of the switch, then issue **Eunfence** on the oncoming primary, and issue **Estart** to make it the active primary node.

```
[sp3en0:/]# Eunfence 1
All node(s) successfully unfenced.

[sp3en0:/]# Estart
Switch initialization started on sp3n01
Initialized 14 node(s).
Switch initialization completed.
```

- If the switch is operational, and **Estart** is failing because the oncoming primary's switch port is fenced, you must first change the oncoming primary to another node on the switch and **Estart**. Once the switch is operational, you can then **Eunfence** the old oncoming primary node. If you also want to make it the active primary, then issue an **Eprimary** command to make it the oncoming primary node and **Estart** the switch once again.

```
[sp3en0:/]# Eprimary 5
```

```
Eprimary: Defaulting oncoming primary backup node to
sp3n15.msc.itso.ibm.com

[sp3en0:/]# Estart
Estart: Oncoming primary != primary, Estart directed to oncoming primary
Estart:  0028-061 Estart is being issued to the primary node:
sp3n05.msc.itso.ibm.com.
Switch initialization started on sp3n05.msc.itso.ibm.com.
Initialized 12 node(s).
Switch initialization completed.

[sp3en0:/]Eunfence 1
All node(s) successfully unfenced.

[sp3en0:/]# Eprimary 1
Eprimary: Defaulting oncoming primary backup node to
sp3n15.msc.itso.ibm.com

[sp3en0:/]# Estart
Estart: Oncoming primary != primary, Estart directed to oncoming primary
Estart:  0028-061 Estart is being issued to the primary node:
sp3n01.msc.itso.ibm.com.
Switch initialization started on sp3n01.msc.itso.ibm.com.
Initialized 13 node(s).
Switch initialization completed.
```

- If the oncoming primary's switch port is fenced, and the switch has not been started, you can not check that the node is fenced or not with the **Efence** command. The only way you can see which nodes are fenced is through the SDR. To check whether the oncoming primary fenced or not, issue:

  ```
  # SDRGetObjects switch_responds
  ```

  If you see the oncoming primary node is *isolated*, the only way you can change the SDR is through SDRChangeAttrValues command. Before using this command, do not forget to archive SDR.

  ```
  # SDRChangeAttrValues switch_responds node_number==<oncoming primary
  node_number> isolated=0
  # SDRGetObjects switch_responds node_number==<oncoming primary
  node_number>
  ```

  Then, issue the command **Estart**.

15.10 Impact of host name/IP changes on an SP system

In the distributed standalone RS/6000 environment, you simply update /etc/hosts file or DNS map file and reconfigure the adapters when you need to change the host name or IP address. However, in an SP environment, the task involved is

not simple, and it affects the entire SP system. The IP address and host names are located in the System Data Repository (SDR) using objects and attributes. The IP address and host names are also kept in system-related files that are located on SP nodes and the CWS.

This section describes the SDR classes and system files when you change either the primary Ethernet IP address and host name for the SP nodes or the CWS. We suggests that you avoid making any host name or IP address changes if possible. The tasks are tedious and in some cases require rerunning the SP installation steps. For detail procedures, refer the Appendix H in the *PSSP Administration Guide*, SA22-7348. These IP address and host name procedures support SP nodes at PSSP levels PSSP 3.1 (AIX 4.3), PSSP 2.4 (AIX 4.2 and 4.3), PSSP 2.2 (AIX 4.1-4.2), and PSSP 2.3 (AIX 4.2 or 4.3) systems. The PSSP 3.1 release supports both SP node coexistence and system partitioning.

Consider the following PSSP components when changing the IP address and hostnames:

- Network Installation Manager (NIM)
- System partitioning
- IBM Virtual Shared Disk
- High Availability Control Workstation (HACWS)
- RS/6000 Cluster Technology (RSCT) Services
- Problem management subsystem
- Performance monitor services
- Extension nodes
- Distributive Computing Environment (DCE)

15.10.1 SDR objects with host names and IP addresses

The following SDR objects reference the host name and IP address in the SP system for PSSP systems:

- Adapter - Specifies the IP addresses used with the switch css0 adapter, or the Ethernet, FDDI, or token ring adapters.
- Frame - Specifies the Monitor and Control Nodes MACN and HACWS.
- backup_MACN - Attributes on the control workstation that work with host names.
- JM_domain_info - Works with the host names for Resource Manager domains.

- JM_Server_Nodes - Works with the host names for Resource Manager server nodes.
- Node - Works with the initial or reliable host names and uses the IP address for SP nodes and boot servers. The nodes are organized by system partitions.
- Pool - Works with host names for Resource Manager pools.
- SP - Works with control workstation IP addresses and host names. Uses the host name when working with Network Time Protocol (NTP) printing, user management, and accounting services.
- SP_ports - Works with the host name used with hardmon and the control workstation.
- Switch_partition - Works with the host name for primary and backup nodes used to support the css SP switch.
- Syspar - Works with the IP address and SP_NAME with system partitions.
- Syspar_map - Provides the host name and IP address on the CWS for system partitions.
- pmandConfig - Captures the SP node host name data working on problem management.
- SPDM - Works with the host name for Performance Monitor status data.
- SPDM_NODES - Works with the host name for SP nodes and organized by system partition.
- DependentNode - Works with the host name for the dependent extension node.
- DependentAdapter - Works with the IP address for the dependent extension node adapter.

15.10.2 System files with IP addresses and host names

The following files contain the IP address or host name that exists on the SP nodes and the control workstation. We recommend that you look through these files when completing the procedures for changing host names and IP addresses for your SP system. The following files are available for PSSP systems:

- /.rhosts - Contains host names used exclusively with rcmd services.
- /.klogin - Contains host names used with authentication rcmd services.
- /etc/hosts - Contains IP addresses and host names used with the SP system.
- /etc/resolv.conf - Contains the IP address for Domain Name Service (DNS) (Optional).

- /var/yp/ NIS - References the host name and IP address with the Network Information Service (NIS).
- /etc/krb5.conf - Works with the host name for DCE.
- /etc/krb.conf - Works with the host name for the authentication server.
- /etc/krb.realms - Works with the host name of the SP nodes and authentication realm.
- /etc/krb-srvtab - Provides the authentication service key using host name.
- /etc/SDR_dest_info - Specifies the IP address of the control workstation and the SDR.
- /etc/ssp/cw_name - Specifies the IP address of control workstation host name on SP nodes that work with node installation and customization.
- /etc/ssp/server_name - Specifies the IP address and host name of the SP boot/install servers on SP nodes working with node customization.
- /etc/ssp/server_hostname - Specifies the IP address and host name of the SP install servers on SP nodes working with node installation.
- /etc/ssp/reliable_hostname - Specifies the IP address and host name of the SP node working with node installation and customization.
- /etc/ntp.conf - Works with the IP address of the NTP server (Optional).
- /etc/filesystems - Can contain the IP address or host name of NFS systems (mainly used on /usr client systems).
- /tftpboot/ host.config_info - Contains the IP address and host name for each SP node. It is found on the CWS and boot servers.
- /tftpboot/ host.intstall_info - Contains the IP address and host name for each SP node. It is found on the CWS and boot servers.
- /tftpboot/ host-new-srvtab - Provides authentication service keys using host name. It is found on the CWS and boot servers.
- /etc/rc.net - Contains the alias IP addresses used with system partitions.
- /etc/niminfo - Works with the NIM configuration for NIM master information.
- /etc/sysctl.acl - Uses host name that works with Sysctl ACL support.
- /etc/logmgt.acl - Uses host name that works with Error Log Mgt ACL support.
- /spdata/sys1/spmon/hmacls - Uses short host name that works with hardmon authentication services.
- /etc/jmd_config. SP_NAME - Works with host names for Resource Management on the CWS for all defined SP_NAME syspars.
- /usr/lpp/csd/vsdfiles/VSD_ipaddr - Contains the SP node IBM Virtual Shared Disk adapter IP address.

- /spdata/sys1/ha/cfg/em.<SP_NAMEcdb>.<Data> - Uses Syspar host name that works with configuration files for Event Management services.
- /var/ha/run/ Availability Services - Uses Syspar host name that contains the run files for the Availability Services.
- /var/ha/log/ Availability Services - Uses Syspar host name that contains the log files for the Availability Services.
- /var/adm/SPlogs/pman/ data - Uses Syspar host name that contains the log files for the Problem Management subsystem.
- /etc/services - Specifies short host name based on SP_NAME partition that work with Availability Services port numbers.
- /etc/auto/maps/auto.u - Contains host names of the file servers providing NFS mounts to Automount.
- /etc/amd/amd-maps/amd.u - Contains host names of the file servers providing NFS mounts to AMD.

15.11 Related documentation

The following documents are recommended for understanding the topics in this chapter and detail its recovery procedures.

SP manuals

This chapter introduces a summary of general problem diagnosis to prepare for the exam. Therefore, you should read Part 2 of *PSSP Diagnosis Guide,* GA22-7350, for a full description. In addition, you may read chapters 4, 5, 8, 12, and 14 of *PSSP Administration Guide,* SA22-7348, to get the basic concepts of each topic we discuss here.

SP redbooks

There is no problem determination redbook for PSSP 2.4. You can use *RS/6000 SP: PSSP 2.2 Survival Guide,* SG24-4928, for PSSP 2.2. This redbook discusses details on node installation and SP switch problems.

15.12 Sample questions

This section provides a series of questions to aid you in preparing for the certification exam. The answers to these questions are in Appendix A, “Answers to sample questions” on page 521.

1. During PSSP 2.4 installation, the **`setup_server`** script returns the following error:

```
mknimres: 0016-395 Could not get size of
/spdata/sys1/install/pssplpp/7[1]/pssp.installp on control workstation
```

You could correct the error by issuing:

a. `mv ssp.usr.2.4.0.0 /spdata/sys1/install/pssplpp/ssp.installp`

b. `mv ssp.usr.2.4.0.0 /spdata/sys1/install/pssplpp/pssp.installp`

c. `mv ssp.usr.2.4.0.0 /spdata/sys1/install/pssplpp/pssp-2.4/ssp.installp`

d. mv ssp.usr.2.4.0.0 /spdata/sys1/install/pssplpp/PSSP-2.4/pssp. installp

2. Select one problem determination/problem source identification methodology statement to resolve this situation:

   You discover you are unable to log in to one of the nodes with any ID (even root) over any network interface OR the node's TTY console. You begin recovery by booting the node into maintenance, getting a root shell prompt, and...

   a. 1) Run the **`df`** command, which shows 100 percent of the node's critical filesystems are used. Clear up this condition.

      2) Realize that Supper may have updated the /etc/passwd file to a 0 length file. Correct /etc/passwd.

      3) Reboot to Normal mode.

      4) Run **`supper update`** on the node.

      5) Now all IDs can log in to the node.

   b. 1) Check permissions of the /etc/passwd file to see if they are correct.

      2) Check that /etc/hosts file-all host lines show three duplicate entries. Edit out these duplicate entries.

      3) Reboot to Normal mode.

      4) Now all IDs can log in to the node.

   c. 1) Check name resolution and TCPIP (ping,telnet) functions to/from the nodes. No problems.

      2) On CWS: Check if hardmon is running. It is not; so, restart it.

      3) Correcting hardmon allows login of all IDs to the node.

   d. 1) Check if Kerberos commands work. They do.

      2) TCPIP (telnet, ping). Does not work.

      3) Fix TCPIP access with:

      ```
      # /usr/lpp/ssp/rcmd/bin/rsh /usr/lpp/ssp/rcmd/ \
      bin/rcp spcw1:/etc/passwd /etc/passwd
      # /usr/lpp/ssp/rcmd/bin/rsh /usr/lpp/ssp/rcmd/ \
      ```

```
bin/rcp spcw1:/etc/hosts /etc/hosts
```

4) Now all users can log in to the node.

3. Apart from a client node being unable to obtain new tickets, the loss of the CWS will not stop normal operation of the SP complex:
   a. True
   b. False
4. If a supper update returned the message `Could not connect to server`, the cause would most likely be:
   a. The supfilesrv daemon is not running and should be restarted.
   b. The SDR_dest_info file is missing and should be recreated.
   c. The root file system on the node is full.
   d. There is a duplicate IP address on the SP Ethernet.
5. If a user running a Kerberized **rsh** command receives a message including the text `Couldn't decode authenticator`, would the most probable solution be (more than one answer is correct):
   a. Remove the .rhosts file.
   b. Check that the time is correct and reset it if not.
   c. Generate a fresh krb-srvtab file for the problem server.
6. After having renamed the ssp.usr fileset to the appropriate name, you receive an error message from setup_server that says `the fileset indicated could not be found`. You should check that:
   a. The ssp.usr fileset is present.
   b. The table of contents for the /spdata/sys1/install/images directory.
   c. The .toc file for the pssplpp subdirectory mentioned is up to date.
   d. The correct file permissions on the /usr spot are set to **744**.

A

Answers to sample questions

This appendix contains the answers and a brief explanation to the sample questions included in every chapter.

A.1 Hardware validation and software configuration

Answers to questions in 2.17, “Sample questions” on page 96, are as follows:

Question 1 - The answer is B. Although primary backup nodes are recommended for high availability, it is not a requirement for switch functionality or for the SP Switch router node. In the event of a failure in the primary node, the backup node can take over the primary duties so that new switch faults can continue being processed. For more information on this, refer to 2.4, “Dependent nodes” on page 33.

Question 2 - The answer is B. The two switch technologies (SP Switch and SP Switch2) are not compatible. Both Switch technologies are still supported in PSSP 3.5 with AIX 5L 5.1 or 5.2, but only one can be used.

Question 3 - The answer is A. PSSP 3.5 requires AIX 5L 5.1 or later.

Question 4 - The answer is A. The new PCI thin nodes (both PowerPC and POWER3 versions) have two PCI slots available for additional adapters. The Ethernet and SCSI adapters are integrated. The switch adapter uses a special MX (mezzanine bus) adapter (MX2 for the POWER3 based nodes). For more information, refer to 2.3.1, “Internal nodes” on page 19.

Question 5 - The answer is A. The SP-attached server M/T 7040 p690 is controlled by the HMC, so only the connection between HMC and CWS is used for the control. The connections between HMC and p690 server include an RS-232 connection to the I/O Book of the server and the ethernet connection. No RS-232 connection between HMC and CWS, or p690 and CWS, exists.

Question 6 - The answer is B. The CWS acts as a boot/install server for other servers in the RS/6000 SP system. In addition, the control workstation can be set up as an authentication server using Kerberos. As an alternative, the control workstation can be set up as a Kerberos secondary server with a backup database to perform ticket-granting service.

Question 7 - The answer is B. For more information, refer to section 2.5.2, “Control workstation minimum hardware requirements” on page 39.

Question 8 - The answer is B. The M/T 7040 needs just a supported ethernet adapter installed in each LPAR. It is not required to be en0. M/T 7017 and M/T 7026 use an ethernet adapter installed in a specific slot, but it has to be en0. M/T 7039 and M/T 7028 use the integrated ethernet port but also do not require en0. Refer to 2.11, “Network connectivity adapters” on page 68.

Question 9 - The answer is C. Only four more LPARs can be configured when 12 are configured already. The limit for a p690/p690+ server is 16 LPARs when

installed in a Cluster 1600 environment. Refer to 2.16.2, “The slot numbering rule” on page 85.

Question 10 - The answer is B. A short frame support only a single SP Switch-8 board. For more information, refer to “SP Switch-8 short frame configurations” on page 78.

A.2 RS/6000 SP networking

Answers to questions in 3.8, “Sample questions” on page 119, are as follows:

Question 1 - The answer is D. Hardware control is done through the serial connection (RS-232) between the control workstation and each frame.

Question 2 - The answer is B. The reliable hostname is the name associated to the en0 interface on every node. The initial hostname is the hostname of the node. The reliable hostname is used by the PSSP components in order to access the node. The initial hostname can be set to a different interface (for example, the css0 interface) if applications need it.

Question 3 - The answer is B. If the /etc/resolv.conf file exist, AIX will follow a predefined order with DNS in the first place. The default order can be altered by creating the /etc/netsvc.conf file.

Question 4 - The answer is D. In a single segment network, the control workstation is the default route and default boot/install server for all the nodes. When multiple segments are used, the default route for nodes will not necessarily be the control workstation. The boot/install server (BIS) is selected based on network topology; however, for a node to install properly, it needs access to the control workstation even when it is being installed from a BIS other than the control workstation. In summary, every node needs a default route, a route to the control workstation, and a boot/install server in its own segment.

Question 5 - The answer is C. A netmask of 255.255.255.224 provides 30 discrete addresses per subnet.

Question 6 - The answer is C. Ethernet, Fiber Distributed Data Interface (FDDI), and token-ring are configured by PSSP. Other network adapters must be configured manually.

Question 7 - The answer is C. The default order in resolving host names is: BIND/DNS, NIS, and local /etc/hosts file. The default order can be overwritten by creating a configuration file, called /etc/netsvc.conf, and specifying the desire order.

Question 8 - The answer is D. There are four basic daemons that NIS uses: ypserv, ypbind, yppasswd, and ypupdated. NIS was initially called yellow pages; hence, the prefix *yp* is used for the daemons.

Question 9 - The answer is C. An NIS server is a machine that provides the system files to be read by other machines on the network. There are two types of servers: Master and Slave. Both keep a copy of the files to be shared over the network. A master server is the machine where a file may be updated. A slave server only maintains a copy of the files to be served. A slave server has three purposes: To balance the load if the master server is busy, to back up the master server, and to enable NIS requests if there are different networks in the NIS domain.

A.3 I/O devices and file systems

Answers to questions in 4.6, “Sample questions” on page 156, are as follows:

Question 1 - The answer is C. Nodes are independent machines. Any peripheral device attached to a node and can be shared with other nodes in the same way as stand-alone machines can share resources on a network. The SP Switch provides a very high bandwidth that makes it an excellent communication network for massive parallel processing.

Question 2 - The answer is C. Only Microchannel nodes support external SSA booting. The reason is that no PCI SSA adapters have been tested to certified external booting support. This is true by the time of this writing, but it may change by the time you read this. Refer to 4.3.4, “Booting from external disks” on page 141 for details.

Question 3 - The answer is A. PSSP 3.1 supports multiple rootvg definitions per node. Before you can use an alternate rootvg volume group, you need to install the alternate rootvg in an alternate set of disks. To activate it, you have to modified the boot list on that node. PSSP provides a command to modify the boot list remotely; it is `spbootlist`. Refer to “spbootlist” on page 140 for details.

Question 4 - The answer is B. The boot/install server is a NFS server for home directories. You can set a node to be a NFS server for home directories, but this does not depend on that node being a boot/install server. The control workstation is always a NFS server even in cases where all nodes are being installed from boot/install servers other than the control workstation. The control workstation always NFS exports the lppsource resources to all nodes.

Question 5 - The answer is A. The command **`spmirrorvg`** enables mirroring on a set of nodes given by the option `-l node_list`. You can force the extension of

the Volume Group by using the `-f` option (available values are: yes or no). Refer to “spmirrorvg” on page 136.

Question 6 - The answer is A. The command `splstdata` can now display information about Volume_Groups. Refer to 4.3.2.7, “Changes to `splstdata` in PSSP 3.1 or later” on page 123.

Question 7 - The answer is B. It is not recommended to use NFS in large production environments that require fast, secure, and easy to manage global file systems. On the other hand, NFS administration is fairly easy, and small environments with low security requirements will probably choose NFS as their global file system.

Question 8 - The answer is C. DFS is a distributed application that manages file system data. It is an application of DCE that uses almost all of the DCE services to provide a secure, highly available, scalable, and manageable distributed file system. DFS data is organized in three levels: Files and directories, filesets, and aggregates. Refer to “What is the Distributed File System?” on page 151.

Question 9 - The answer is C. The following are the other default values: the default install_disk is hdisk0, quorum is true, mirroring is off, copies are set to 1, there are no bootable alternate root Volume Groups, and all other attributes of the Volume_Groups are initialized according to the same rules as the Node object. Refer to “Volume_Group default values” on page 131.

Question 10 - The answer is B. Refer to “spmkvgobj” on page 132.

A.4 Cluster 1600 how-tos

Answers to questions in 5.9, “Sample questions” on page 211, are as follows:

Question 1 - The answer is D. Each SP-attached server must be connected to the control workstation through two RS-232 serial links and an Ethernet connection. One of the RS-232 lines connects the control workstation with the front panel of the SP-attached server and uses a System and Manufacturing Interface protocol (SAMI). The other line goes to the back of the CEC unit and attaches to the first integrated RS-232 port in the SP-attached server. This line serves as the s1term emulator. Remember that login must be enabled in that first integrated port (S1) in order to s1term to work. Refer to 5.2.3, “External node attachment” on page 165 for details.

Question 2 - The answer is D. SP-attached servers cannot be installed between switched frames and expansion frames. Although SP-attached servers can be placed anywhere in the SP complex because they do not follow the rules of standard SP frames, this restriction comes from the expansion frame itself. All

expansion frames for frame n must be numbered n+1, n+2, and n+3. Refer to 2.15, “Cluster 1600 configuration rules” on page 74 for details.

Question 3 - The answer is B. SP-attached servers do not have frame or node supervisor cards, which limits the capabilities of the hardmon daemon to monitor or control these external nodes. Most of the basic hardware control is provided by the SAMI interface, however most of the monitoring capabilities are provided by an internal sensor connected to the node supervisor cards. So, the lack of node supervisor cards on SP-attached servers limits those monitoring capabilities.

Question 4 - The answer is B. The s70d daemon is started and controlled by the hardmon daemon. Each time the hardmon daemon detects a SAMI frame (a SP-attached server seen as a frame), it starts a new s70d process. The hardmon daemon will keep a socket connection with this s70d. The s70d will translate the commands coming from the hardmon daemon into SAMI commands. Refer to 5.5.2, “Hardmon” on page 188 for details.

Question 5 - The answer is D. The SP-attached server cannot be the first frame in the SP system. So, the first frame in the SP system must be an SP frame containing at least one node. This is necessary for the SDR_config code, which needs to determine whether the frame is with or without a switch.

Question 6 - The answer is D. The SP-attached server M/T 7026 p660 needs an internal attachment adapter (CSP card) that has an RS-232 connection to the CWS. The protocol that is used is called CSP. The SAMI protocol is for MT 7017 server, the HMC protocol is for HMC-controlled servers like M/T 7040, 7039, 7038, and 7028; and for these servers there is no serial connection between the HMC and the CWS.

Question 7 - The answer is A. Your SP system must be operating with a minimum of PSSP 3.4 and AIX 5L 5.1 before you can use the SP-attached server Model 7028 p630 in this environment.

Question 8 - The answer is A. Each LPAR occupies one slot position. A maximum of 16 LPARs is allowed for one system. This is due to the 16 nodes per frame limitation in SP Systems.

Question 9 - The answer is B. For the S70 server, only the 10Mbps BNC or the 10Mbps AUI Ethernet adapters are supported for SP-LAN communication BNC adapters provides the BNC cables, but the AUI ethernet adapter does not provide the twisted pair cables.

Question 10 - The answer is D. Refer to , “System management” on page 197 for details.

A.5 SP security

Answers to questions in 6.14, “Sample questions” on page 245, are as follows:

Question 1 - The answer is C. The `rc.sp` script runs every time a node boots. This script checks the Syspar class in the SDR and resets the authentication mechanism based on the attributes in that class. Using the `chauthent` command directly on a node will cause the node to be in an inconsistent state with the rest of the system, and the change will be lost by the time of the next boot. It is recommended not to change the authentication setting directly on the node but through the use of PSSP command and SDR settings.

Question 2 - The answer is D. The /etc/krb-srvtab files contain the private password for the Kerberos services on a node. This is a binary file, and its content can be viewed by using the `klist -srvtab` command. By default the hardmon and the remote command (`rcmd`) principals maintain their private passwords in this file. Refer to 6.9, “Server key” on page 233 for details.

Question 3 - The answer is A. Although the SP Perspectives uses services that are Kerberos clients, the interface itself is not a Kerberos client. Event Perspectives requires you to have a valid Kerberos principal to generate automatic actions upon receiving event notifications (this facility is provided by the problem management subsystem). The Hardware Perspective requires you to have a valid Kerberos principal in order to access the hardware control monitoring facilities that are provided by the hardmon daemon. The VSD Perspective requires you to have a valid Kerberos principal to access the VSD functionality because the VSD subsystems uses sysctl for control and monitoring of the virtual shared disk, nodes, and servers.

Question 4 - The answers are A and D. Two service names are used by the Kerberos-authenticated applications in an SP system: `hardmon` used by the system monitor daemon on the control workstation by logging daemons, and `rcmd` used by sysctl.

Question 5 - The answer is B. One of the procedures to add a Kerberos Principal is to use the `mkkp` command. This command is non-interactive and does not provide the capability to set the principal’s initial password. The password must, therefore, be set by using the `kadmin` command and its subcommand, `cwp`. Refer to 6.8.1, “Add a Kerberos principal” on page 230 for more details.

Question 6 - The answer is D. On the SP, there are three different sets of services that use Kerberos authentication: the hardware control subsystem, the remote execution commands, and the sysctl facility.

Question 7 - The answer is C. PSSP support the use of an existing AFS server to provide Kerberos Version 4 services to the SP. Usage of AFS on SP systems is optional.

Question 8 - The answer is B. The three kerberos daemons are: Kerberos, kadmind, and kpropd.

Question 9 - The answer is D. The `kstash` command kills and restarts the kadmin daemon, and recreates the /.k file to store the new master key in it.

A.6 User and data management

Answers to questions in 7.8, “Sample questions” on page 268, are as follows:

Question 1 - The answer is C. If you are using the SP User Management facilities, File Collection will automatically replace the /etc/passwd and the /etc/security/passwd files every other hour. This makes it possible to have global SP users by having a common set of user files across nodes. The `passwd` command gets replaced by a PSSP command that will prompt the user to change its password on the control workstation, which is the password server by default.

Question 2 - The answer is C. SP users are global AIX users managed by the SP User Management facility (SPUM). All the user definitions are common across nodes. The SPUM provides mechanisms to NFS mount a home directory on any node and to provide the same environment to users no matter where they log in to. Refer to 7.3, “SP User data management” on page 251 for details.

Question 3 - The answer is D. The `spac_cntrl` command is used to set access control to node. This command must be executed on every node where you want to restrict user access, for example, to run batch jobs without users sniffing around. Refer to 7.3.6, “Access control” on page 253 for details.

Question 4 - The answer is B. All the user related configuration files are managed by the user.admin file collection. This collection is defined by default and it activated when you selected the SPUM as your user management facility. Refer to “user.admin collection” on page 257 for details.

Question 5 - The answer is D. PSSP is shipped with four predefined file collections: sup.admin, user.admin, power_system, and node.root.

Question 6 - The answer is B. The `supper` command is used to report information about file collections. It has a set of subcommands to perform file and directory management that includes verification of information and the checking of results when a procedure is being performed.

Question 7 - The answer is C. NIS allows a system administrator to maintain system configuration files in one place. These files only need to be changed once then propagated to the other nodes. Refer to 7.4, “Configuring NIS” on page 253 for details.

Question 8 - The answer is D. The default hierarchy of updates for file collections is in the following sequence: CWS/BIS/Nodes. However, the default hierarchy can be changed. Refer to 7.5.9, “Modifying the file collection hierarchy” on page 265 for details.

Question 9 - The answer is C. Make sure you are working with the master files. Refer to “Adding and deleting files in a file collection” on page 264 for details.

Question 10 - The answer is B. AIX Automounter is a tool that can make the RS/6000 SP system appear as only one machine to both the end users and the applications by means of a global repository of storage. It manages mounting activities using standard NFS facilities. It mounts remote systems when they are used and automatically dismounts them when they are no longer needed.

A.7 Configuring the control workstation

Answers to questions in 8.9, “Sample questions” on page 297, are as follows:

Question 1 - The answer is A. The initialization of the Kerberos is controlled by the `setup_authent` command. The **`install_cw`** script will not initialize the authentication services. Refer to 8.3.4, “install_cw” on page 279 for details.

Question 2 - The answer is B. Depending on the authentication method for the AIX remote commands you have selected you can set the authentication for SP trusted Services. Refer to table Table 8-6 on page 295 for details.

Question 3 - The answer is D. After you run the **`install_cw`** command for configuring the SDR on the CWS, the very next step is to verify whether the configuration was successful. The `SDR_test` command does the verification of the same.

Question 4- The answer is D. The /etc/rc.net file is the recommended location for setting any static routing information. In the case where the CWS and all nodes’ en0 adapters are not on the same Ethernet segment, the /etc/rc.net file of the CWS can be modified to include a routing statement.

Question 5 - The answer is A. The monitoring and control of the HMC Controlled server hardware from the CWS requires an ethernet connection between the CWS and each HMC in the Cluster 1600 system. Refer to Table 8-2 on page 285 for a complete list of different protocols used for Cluster 1600 servers.

Question 6 - The answer is C. The **`setup-authent`** command has no arguments. It configures the Kerberos authentication services for the Cluster 1600 system. This command can be used to configure the CWS in a number of ways as listed in the options. It cannot configure the SDR.

Question 7 - The answer is B. In the case that HMC servers are integrated in the Cluster 1600 environment an Ethernet connection is the only connectivity required between the CWS and the server configured on the en0 adapter of the server. Refer to Figure 2-1 on page 9 for a complete understanding of the connectivity.

Question 8- The answer is A. The installp images (lpp) of the AIX 5l 5.1 and latter must be stored in directories named /spdata/sys1/install/<name>/lppsource/installp/ppc. The filesets in the RPM format must be placed in /spdata/sys1/install/<name>/lppsourceRPMS/ppc. You can set <name> to the name your prefer. However, it is recommended to use a name identifying the version of the AIX lpps stored in this directory. The names generally used are aix51, aix52, and so on.

Question 9- The answer is D. All the filesets must be installed on the CWS before starting to configure the HMC.

A.8 Frames and nodes installation

Answers to questions in 9.5, “Sample questions” on page 343, are as follows:

Question 1 - The answers are A and C. The initial hostname is the real host name of a node, while the reliable hostname is the hostname associated to the en0 interface on that node. Most of the PSSP components will use the reliable hostname for accessing PSSP resources on that node. The initial hostname can be set to a faster network interface (such as the SP Switch) if applications use the node’s hostname for accessing resources.

Question 2- The answer is D. A boot/install server is defined when nodes have their install server field pointing to a particular node. By default, the control workstation is the boot/install server to all nodes, but in a multi-frame environment, PSSP will choose the first node in each frame to be the boot/install server for the nodes in that frame. The `spbootins` command will run the `setup_server` script remotely in any boot/install server node.

Question 3- The answer is D. The `spadaptrs` command is used to configure additional adapters into the SRD. It executes on the CWS only using the command line interface or the equivalent functions accessible from the SMIT Additional Adapter Database Information window (smitty add_adapt_dialog).

Question 4- The answer is C. The **syspar_crtl** command controls the system partition sensitive subsystems on the CWS and on the SP nodes. This command will start the daemons: hats, hags, haem, hr, pman, emon, spconfigd, emcond, and spdmd (optional). Since the daemons need to execute on all machines of the SP system for the subsystem to run successfully, **syspar_crtl -A** must also be executed on each node when it is up.

Question 5- The answer is C. The customization of the boot/install server (setup_server command) creates several files in the /tftpboot directory. Refer to 9.3.2, "/tftpboot" on page 335 for details.

Question 6- The answer is B. The **splstdata** command displays configuration information stored in the SDR. This command executes in the CWS or any SP node when using the command line interface. Refer to 9.2.14, "Verify all node information" on page 323 for more details about the `splstdata` command.

Question 7- The answer is D. The **setup_server** command configures the machine where it is executed (CWS or SP node) as a boot/install server. This command has no argument. It executes on the CWS and any additional boot/install servers. Refer to 9.2.19, "Configuring the CWS as boot/install server" on page 328 for details.

Question 8- The answer is A and D. Both these commands are used to customize the sample configuration file in the /etc/SP directory and store it in SDR. Refer to 9.2.18, "Setting the switch" on page 325.

Question 9- The answer is D. There are 5 ways of specifying the disk or disks to use for installation. For the actual format to be used, refer to "Selecting an installation disk" on page 319.

A.9 Verification commands and methods

Answers to questions in 10.8, "Sample questions" on page 366, are as follows:

Question 1 - The answer is D. The **SDR_test** script checks the SDR and reports any errors found. It will contact the SDR daemon and will try to create and remove classes and attributes. If this test is successful, then the SDR directory structure and the daemons are set up correctly. Refer to "Checking the SDR initialization: SDR_test" on page 349 for details.

Question 2 - The answer is D. The **spmon -d** command will contact the frame supervisor card only if the **-G** flag is used. If this flag is not used, the command will only report node information. Refer to "Monitoring hardware activity: spmon -d" on page 355 for details.

Question 3 - The answer is B. The hardmon daemon is not a partition-sensitive daemon. There is only one daemon running on the control workstation at any time, even though there may be more than one partition configured. The daemon uses the RS-232 lines and the Ethernet (for HMC) to contact the frames every five seconds by default.

Question 4 - The answer is C. The worm is started by the **rc.switch** script, which is started at node boot time.

Question 5 - The answer is B. The **SYSMAN_test** command is a very powerful test tool. It checks a large number of SP system management components. The command is executed on the CWS, but it does not restrict its checking to components of the CWS. If nodes are up and running, it will also perform several tests on them.

Question 6 - The answers are C and D. The **lsauthpar** command lists the remote command authentication methods that are active for the system partition. The **lsauthpts** command lists the trusted services authentication methods that are active for the system partition.

A.10 Cluster 1600 supported products

Answers to questions in Section 11.10, “Sample questions” on page 390, are as follows:

Question 1 - The answer is A. To run jobs on any machine in the LoadLeveler cluster, users need the same UID (the system ID number for a user) and the same GID (the system ID number for a group) for every machine in the cluster. If you do not have a user ID on a machine, your jobs will not run on that machine. Also, many commands, such as **llq**, will not work correctly if a user does not have an ID on the central manager machine.

Question 2 - The answer is C. The High Availability Cluster Multiprocessing Control Workstation (HACWS) requires two control workstations to be physically connected to any frame. A Y-cable is used to connect the single connector on the frame supervisor card to each control workstation.

Question 3 - The answer is D. If the primary CWS fails, the backup CWS can assume all functions with the following exceptions: updating passwords (if SP User Management is in use), adding or changing SP users, changing Kerberos keys (the backup CWS is typically configured as a secondary authentication server), adding nodes to the system, and changing site environment information.

Question 4 - The answer is B. The **lsvsd** command, when used with the **-l** flag, will list all the configured virtual shared disks on a node. To display all the virtual

shared disks configured in all nodes, you can use the **dsh** command to run the **lsvsd** command on all nodes.

Question 5 - The answer is D. In order to get the virtual shared disk working properly, you have to install the VSD software on all the nodes where you want VSD access (client and server), then you need to grant authorization to the Kerberos principal you will use to configure the virtual shared disks on the nodes. After you grant authorization, you may designate which node will be configured to access the virtual shared disks you define.

After doing this, you can start creating the virtual shared disks. Remember that when you create virtual shared disks, you have to make them ready to become active. By default, a virtual shared disk is put into a stopped mode after it is created; so, you have to use the `preparevsd` command to put it into a suspended state that can be made active by using the `resumevsd` command afterwards. Refer to *PSSP Managing Shared Disks,* SA22-7349 for details.

Question 6 - The answer is B. In GPFS, there is no concept of a GPFS server or client node. A GPFS node is whatever node has the GPFS code configured and up and running. GPFS nodes are always, at least, VSD client nodes, but they may also be VSD server nodes.

Question 7 - The answer is A. It is possible to change the configuration of GPFS for performance tuning purposes. The mmchconfig command is capable of changing the following attributes: pagepool, mallocsize, maxFiles To Cache, dataStructureDump.

Question 8 - The answer is D. GPFS automatically stripes data across VSDs to increase performance and balance disk I/O. There are three possible striping algorithms that you can choose for GPFS to implement: Round Robin, Balanced Random, and Random. A striping algorithm can be set when a GPFS FS is created or can be modified as an FS parameter later.

Question 9 - The answer is B. GPFS requires RVSD even though your installation does not have twin-tailed disks or SSA loops for multi-host disk connection.

A.11 Problem management tools

Answers to questions in 12.8, “Sample questions” on page 415, are as follows:

Question 1 - The answer is D. The `log_event` script uses the AIX **alog** command to write to a wraparound file. The size of the wraparound file is limited to 64 K. The **alog** command must be used to read the file. Refer to the AIX **alog** man page for more information on this command.

Question 2 - The answer is D. Access to the problem management subsystem is controlled by the /etc/sysctl.pman.acl configuration file. All users who want to use the problem management facility must have a valid Kerberos principal listed in this file before attempting to define monitors. Refer to 12.5.1, “Authorization” on page 404 for details.

Question 3 - The answer is C. The **haemqvar** command is a new command in PSSP 3.1 that allows you to display information regarding resource variables. Before this command was created, the only way you could get information for resource variables (such as syntax and usage information) was through the SP Perspectives graphical interface, in particular, through the Event Perspective.

Question 4 - The answer is D. Trace facility is available through AIX. However, it comes in an optional fileset called bos.sysmgt.trace. You need to install this optional component if you want to activate the trace daemon and generate trace reports.

Question 5 - The answer is B. All the PSSP log files are located in the /var/adm/SPlogs directory. All the RSCT log files are located in the /var/ha/log directory. Make sure you have enough free space for holding all the logged information.

Question 6 - The answer is D. Event Management gathers information on system resources using Resource Monitors (RMs). Refer to 12.4, “Event Management” on page 399 for details.

Question 7 - The answer is A. The Problem Management subsystem (PMAN) is a facility used for problem determination, problem notification, and problem solving. The PMAN subsystem consists of three components: `pmand`, `pmanrmd`, and `sp_configd`.

Question 8 - The answer is A. The following steps are required to create a condition: Decide what you want to monitor, identify the resource variable, define the expression, and create the condition. Refer to 12.6.1, “Defining conditions” on page 409 for details.

A.12 RS/6000 SP software maintenance

Answers to questions in 13.7, “Sample questions” on page 433, are as follows:

Question 1 - The answer is B. The **spsvrmgr** command can be used to check the supervisor microcode levels on frames and nodes. Use the **-G** flag to get all frame supervisor cards checked.

Question 2 - The answer is A. Every time a new PTF is applied, the supervisor microcode on frame and node supervisor cards should be checked.

Question 3 - The answer is B. Refer to 13.5.2, “Supported migration paths” on page 427 for details.

Question 4 - The answer is C.

Question 5 - The answer is B. To restore an image of the CWS, do the following: Execute the normal procedure to restore any RS/6000 workstation, issue the **/usr/lpp/ssp/bin/install_cw** command, and verify your CWS.

A.13 RS/6000 SP reconfiguration and update

Answers to questions in 14.9, “Sample questions” on page 476, are as follows:

Question 1 - The answers are C and D. When changes are made to IP addresses of adapters defined in the SDR, as is the case of the SP Switch adapter, the information should be updated into the SDR, and the node(s) affected should be customized.

Question 2 - The answer is A. New tall frames, announced in 1998, have higher power requirements. You should confirm that your current installation can handle this higher power demand.

Question 3 - The answer is D. If you set up the boot/install server, and it is acting as a gateway to the CWS, `ipforwarding` must be enabled. To turn it on issue: **/usr/sbin/no -o ipforwarding=1**.

Question 4 - The answer is D. There is only one partition in the SP system. Refer to 14.7, “Replacing to PCI-based 332 MHz SMP node” on page 471 for details.

Question 5 - The answer is B. If you need to update the microcode of the frame supervisor of frame 2, enter: **spsvrmgr -G -u 2:0**.

A.14 Problem diagnosis

Answers to questions in 15.12, “Sample questions” on page 518, are as follows:

Question 1 - The answer is D. When you download the PSSP installation tape into the control workstation or a boot/install server, the image is named ssp.usr.2.4.0.0 (for PSSP 3.1, it is called ssp.usr.3.1.0.0), but the **setup_server** script expects to find a file image called pssp.installp located in the main

directory for the version you are installing (in this case, it is /spdata/sys1/install/pssplpp/PSSP-2.4). If this file (pssp.installp) is not present in that directory, the **setup_server** script will fail with this error.

Question 2 - The answer is A. If for some reason the /etc/passwd file gets erased or emptied, as happened here, you will not be able to log on to this node until the file gets restored. To do that, you have start the node in maintenance mode and restore the /etc/passwd file before attempting to log on to that node again. Make sure you supper update the files if you keep a single copy of the /etc/passwd file for your system.

Question 3 - The answer is *True*. Although the control workstation plays a key role in the RS/6000 SP, it is not essential for having the nodes up and running. The most critical factor on the control workstation dependency is the fact that the SDR is located there, and by default, the control workstation is also the authentication server.

Question 4 - The answer is A. The supfilesrv daemon runs on all the file collection servers. If the daemon is not running, clients will prompt this error message when trying to contact the server.

Question 5 - The answers are B and C. Most cases when the error message refers to authenticator decoding problems, they are related to either the time difference between the client and the server machine because a time stamp is used to encode and decode messages in Kerberos; so, if the time difference between the client and server is more than five minutes, Kerberos will fail with this error. The other common case is when the /etc/krb-srvtab file is corrupted or out-of-date. This will also cause Kerberos to fail.

Question 6 - The answer is C. When installing PSSP, the **installp** command will check the .toc file. This file is not generated automatically when you move files around in the directory. Always use the **inutoc** command to update the table of contents of a directory before using the **installp** command.

B

NIS

There are four basic daemons that NIS uses: ypserv, ypbind, yppasswd, and ypupdated. NIS was initially called yellow pages; hence, the prefix yp is used for the daemons. They work in the following way:

- All machines within the NIS domain run the ypbind daemon. This daemon directs the machine's request for a file to the NIS servers. On clients and slave servers, the ypbind daemon points the machines to the master server. On the master server, its ypbind points back to itself.
- ypserv runs on both the master and the slave servers. It is this daemon that responds to the request for file information by the clients.
- yppasswd and ypupdated run only on the master server. The yppasswd daemon makes it possible for users to change their login passwords anywhere on the network. When NIS is configured, the **/bin/passwd** command is linked to the **/usr/bin/yppasswd** command on the nodes. The **yppasswd** command sends any password changes over the network to the yppasswd daemon on the master server. The master server changes the appropriate files and propagates this change to the slave servers using the ypupdated daemon.

Note: NIS serves files in the form of maps. There is a map for each of the files that it serves. Information from the file is stored in the map, and it is the map that is used to respond to client requests.

By default, the following files are served by NIS:

- /etc/ethers
- /etc/group
- /etc/hosts
- /etc/netgroup
- /etc/networks
- /etc/passwd
- /etc/protocols
- /etc/publickey
- /etc/rpc
- /etc/security/group
- /etc/security/passwd
- /etc/services

Tip : By serving the /etc/hosts file, NIS has an added capability for handling name resolution in a network. Refer to NIS and NFS publications by O'Reilly and Associates for detailed information.

To configure NIS, there are four steps, all of which can be done through SMIT. For all four steps, first run **`smit nfs`** and select **Network Information Service (NIS)** to access the NIS panels, then:

- Choose **Change NIS Domain Name of this Host** to define the NIS domain. Figure B-1 on page 539 shows what this SMIT panel looks like. In this example, SPDomain has been chosen as the NIS domain name.

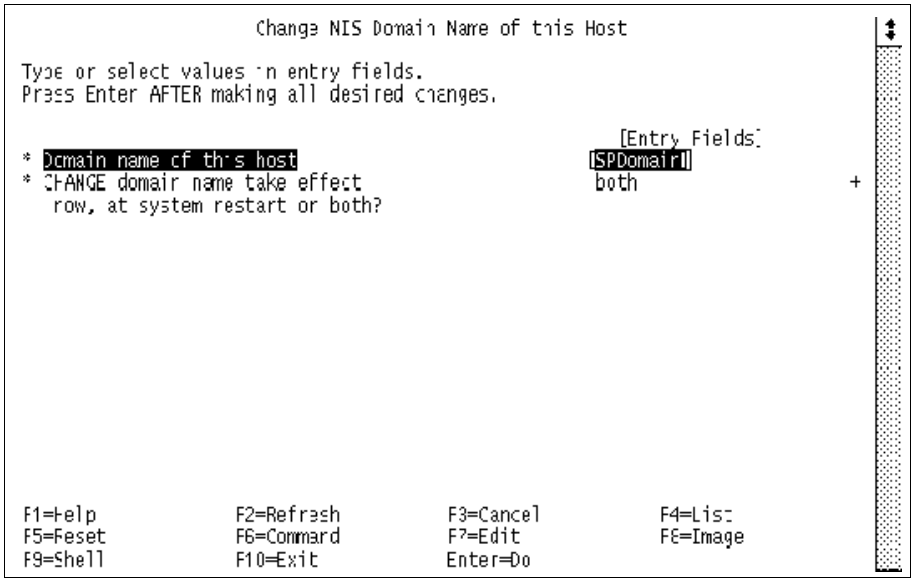


Figure B-1 SMIT panel for setting an NIS domain name

- On the machine that is to be the NIS master (for example, the control workstation), select **Configure/Modify NIS** → **Configure this Host as a NIS Master Server**. Figure B-2 on page 540 shows the SMIT panel. Fill in the fields as required. Be sure to start the yppasswd and ypupdated daemons. When the SMIT panel is executed, all four daemons (ypbind, ypserv, yppasswd, and ypupdated) are started on the master server. This SMIT panel also updates the NIS entries in the local /etc/rc.nfs files.

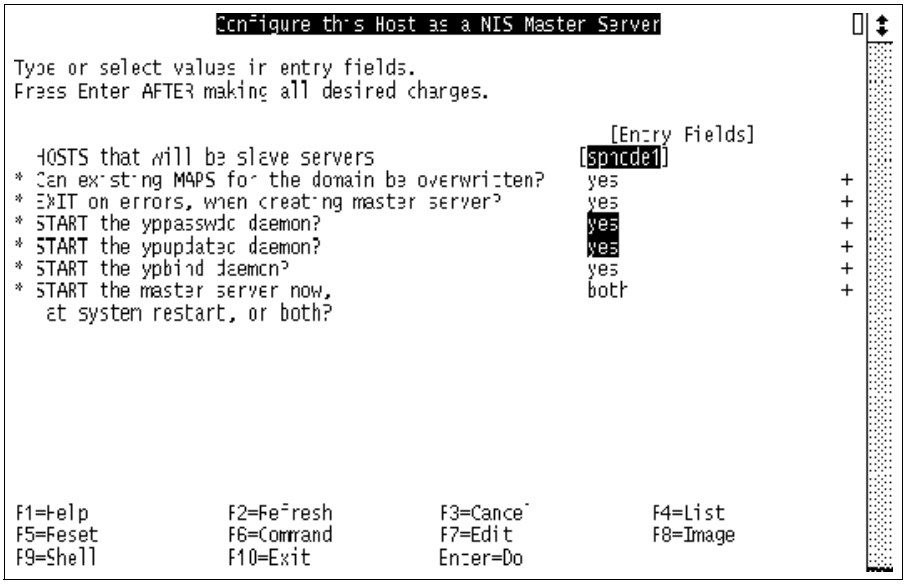


Figure B-2 SMIT panel for configuring a master server

- On the machines set aside to be slave servers, go to the NIS SMIT panels and select **Configure this Host as a NIS Slave Server**. Figure B-3 on page 541 shows the SMIT panel for configuring a slave server. This step starts the ypserv and ypbind daemons on the slave servers and updates the NIS entries in the local /etc/rc.nfs files.

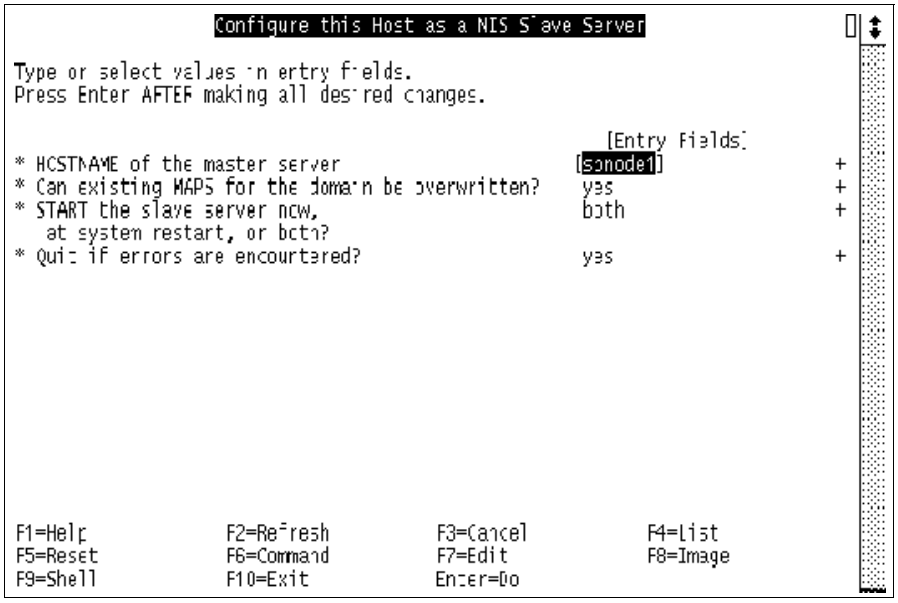


Figure B-3 SMIT panel for configuring a slave server

- On each node that is to be a NIS client, go into the NIS panels and select **Configure this Host as a NIS Client**. This step starts the ypbind daemon and updates the NIS entries in the local /etc/rc.nfs files.

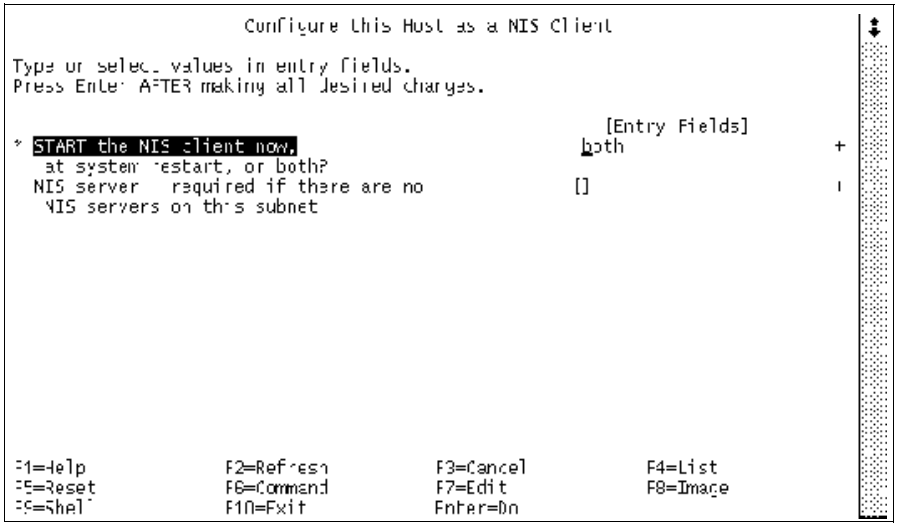


Figure B-4 SMIT panel for configuring an NIS client

Once configured, when there are changes to any of the files served by NIS, their corresponding maps on the master are rebuilt and either pushed to the slave servers or pulled by the slave servers from the master server. These are done through the SMIT panel or the **make** command. To access the SMIT panel, select **Manage NIS Maps** in the NIS panel. Figure B-5 on page 542 shows this SMIT panel.

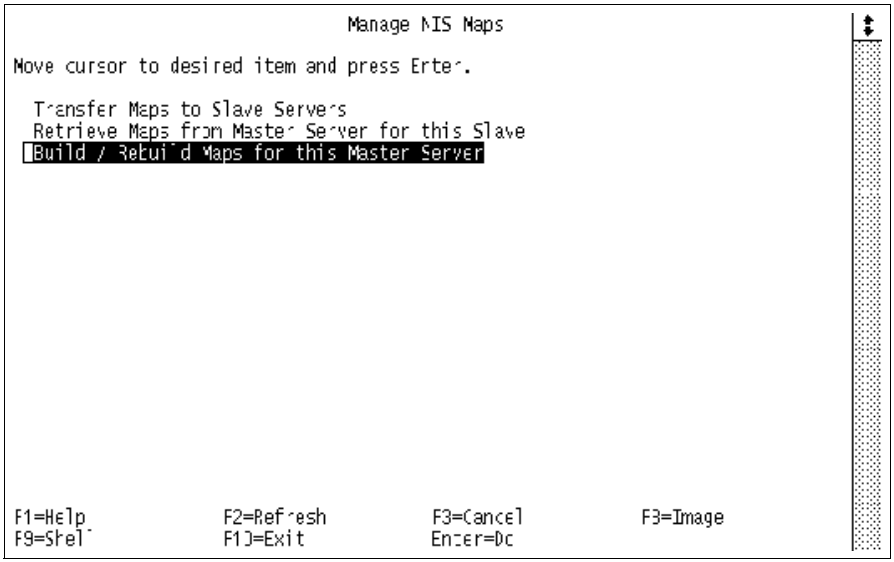


Figure B-5 SMIT panel for managing NIS maps

Select **Build/Rebuild Maps for this Master Server**, and then either have the system rebuild all the maps with the *all* option or specify the maps that you want to rebuild. After that, return back to the SMIT panel as shown n Figure B-5 on page 542, and select either **Transfer Maps to Slave Servers** (from the master server) or **Retrieve Maps from Master Server for this Slave** (from a slave server).

B.1 Setting up NIS

You can use SMIT to set up NIS, manage it, and control the NIS daemons. In your planning, you must decide whether you will have slave servers and whether you will allow users to change their passwords anywhere in the network.

B.1.1 Configuring a master server

Configuring a master server can be done by entering the **smit mkmaster** command.

By default, the NIS master server maintains the following files that should contain the information needed to serve the client systems:

- /etc/ethers
- /etc/group
- /etc/hosts
- /etc/netgroup
- /etc/networks
- /etc/passwd
- /etc/protocols
- /etc/publickey
- /etc/rpc
- /etc/security/group
- /etc/security/passwd
- /etc/services

Any changes to these files must be propagated to clients and slave servers using SMIT:

Select: **Manage NIS Maps**

Select: **Build / Rebuild Maps for this Master Server**

Either specify a particular NIS map by entering the name representing the file name, or leave the default value of *all*, then press Enter. You can also do this manually by changing to the directory /etc/yp and entering the command **make all** or **make <map-name>**. This propagates the maps to all NIS clients and transfers all maps to the slave servers.

B.1.2 Configuring a slave server

A slave server is the same as the master server except that it is a read-only server. Therefore, it cannot update any NIS maps. Making a slaver server implies that all NIS maps will be physically present on the node configured as the slave server. As with a master server, the NIS map files on a slave server can be found in /etc/yp/<domainname>.

You may configure a slave server with the **smit mkslave** command.

Configuring a slaver server starts the ypbind daemon that searches for a server in the network running **ypserv**. Shortly afterwards, the ypserv daemon of the slave server itself will start.

In many situations, the slave server must also be able to receive and serve login requests. If this is the case, the slave server must also be configured as an NIS client.

B.1.3 Configuring an NIS client

An NIS client retrieves its information from the first server it contacts. The process responsible for establishing the contact with a server is ypbind.

You may also configure a Client Server using SMIT by entering **smit mkclient** on every node or use `edit the appropriate entries` in the script.cust file. This can be done at installation time or later through changing the file and then doing a customized boot on the appropriate node.

B.1.4 Change NIS password

You may change an NIS user password with the **passwd** or **yppasswd** commands.

B.2 Related documentation

The following book is recommended reading to provide a broader view on implementing NIS in a Cluster 1600 managed by PSSP environment:

Managing NFS and NIS; O'Reilly

AFS as a Cluster 1600 Kerberos-based security system

PSSP supports the use of an existing AFS server to provide Kerberos Version 4 services for the Cluster 1600 configuration. It does not include the AFS server itself. Before installing PSSP on the control workstation, an AFS server must be configured and accessible. The **setup_authent** script, which initializes the authentication environment that comes with PSSP, supports AFS as the underlying Kerberos server. This is mainly contained in its **setup_afs_server** subcommand.

PSSP Installation and Migration Guide, GA22-7347 explains the steps that are required to initially set up the security on a Cluster 1600 configuration using an AFS server, and *PSSP Administration Guide,* SA22-7348 describes the differences in the management commands of PSSP Kerberos and AFS Kerberos.

AFS uses a different set of protocols, utilities, daemons, and interfaces for principal database administration.

Usage of AFS on systems is optional.

C.1 Setup to use AFS authentication server

- When running the **setup_authent** command, make sure to answer yes to the question on whether you want to set up authentication services to use AFS servers.
- The control workstation (CWS) may be an AFS server or an AFS client.
- The AFS files ThisCell and CellServDB should be in /usr/vice/etc, or a symbolic link should be created.
- The **kas** command should be in /usr/afsws/etc, or a symbolic link created.
- AFS must be defined with an administrative attribute.
- Run **setup_authent** providing the name and password of the AFS administrator.
- Issue the **k4list** command to check for a ticket for the administration account.

C.2 AFS commands and daemons

AFS provides its own set of commands and daemons. The AFS daemon is afsd, which is used to connect AFS clients and server.

Table C-1 describes some commands that can be used for managing AFS.

Table C-1 Commands for managing AFS

Commands	Description
kas	For adding, listing, deleting, and changing the AFS principal's attributes. **kas** has corresponding subcommands, as follows: **examine** for displaying Principal's information **create** for adding Principals and setting passwords **setfields** for adding an authentication administrator and for changing Principal passwords and attributes **delete** for deleting Principals
kinit	For obtaining authentication credentials
klog.krb (AFS command)	For obtaining authentication credentials
klist or **k4list**	For displaying authentication credentials

Commands	Description
token.krb (AFS command)	For displaying authentication credentials
kdestroy	For deleting authentication credentials, which involves removing tickets from the Kerberos ticket cache file
klog.krb	The user interface to get Kerberos tickets and AFS tokens
unlog	For deleting authentication credentials, which involves removing tokens held by AFS cache manager
kpasswd	For changing passwords
pts	This is the AFS protection services administration interface. It has the following subcommands: **adduser** for adding a user to a group **chown** for changing ownership of a group **creategroup** for creating a new group **delete** for deleting a user or group from the database **examine** for examining an entry **listowned** for listing groups owned by an entry **membership** for listing membership of a user or group **removeusers** for removing a user from a group **setfields** for setting fields for an entry

C.3 Related documentation

The following books are recommended reading to provide a broader view on implementing AFS in a Cluster 1600 managed by a PSSP environment.

SP manuals

PSSP Administration Guide, SA22-7348 covers “Relationship of PSSP Kerberos V4 and AFS.”

RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment, GA22-7281 covers “Using AFS authentication.”

PSSP Installation and Migration Guide, GA22-7347 covers “Initializing to use AFS authentication.”

Non-IBM publications

Managing AFS: The Andrew File System; Richard Campbell

Abbreviations and acronyms

ACL	Access Control Lists
AFS	Andrew File System
AIX	Advanced Interactive Executive
AMG	Adapter Membership Group
ANS	Abstract Notation Syntax
API	Application Programming Interface
ARP	Address Resolution Protocol
BIS	Boot/Install Server
BOS	Basic Overseer Server
BPC	Bulk Power Controller
BSD	Berkeley Software Distribution
BUMP	Bring-Up Microprocessor
CDS	Cell Directory Service
CEC	Central Electronic Complex
CLIOS/S	Client Input Output Socket
CP	Crown Prince
CSMA/CD	Carrier Sense, Multiple Access/Collision Detect
CSS	Communication Subsystem
CVSD	Concurrent Virtual Shared Disk
CWS	Control Workstation
DB	Database
DCE	Distributed Computing Environment
DFS	Distributed File System
DMA	Direct Memory Access
DNS	Domain Name Service
EM	Event Management
EMAPI	Event Management Application Programming Interface
EMCDB	Event Management Configuration Database
EMD	Event Manager Daemon
EPROM	Erasable Programmable Read-Only Memory
ERP	Enterprise Resource Planning
FCS	Fiber Channel Standard
FDDI	Fiber Distributed Data Interface
FIFO	First-In First-Out
FLDB	Fileset Location Database
FS	File System
GB	Gigabytes
GL	Group Leader
GPFS	General Parallel File System
GS	Group Services
GSAPI	Group Services Application Programming Interface
GUI	Graphical User Interface
GVG	Global Volume Group
HACMP	High Availability Cluster Multiprocessing
HACMP/ES	High Availability Cluster Multiprocessing Enhanced Scalability
HACWS	High Availability Control Workstation
HB	Heart Beat
HRD	Host Respond Daemon
HSD	Hashed Shared Disk
HSSI	High Speed Serial Interface

IBM	International Business Machines Corporation
IP	Internet Protocol
ISB	Intermediate Switch Boad
ISC	Intermediate Switch Chip
ITSO	International Technical Support Organization
JFS	Journaled File System
LAN	Local Area Network
LCD	Liquid Crystal Display
LED	Light Emitter Diode
LFS	Local File System
LP	Logical Partition
LRU	Last Recently Used
LSC	Link Switch Chip
LV	Logical Volume
LVM	Logical Volume Manager
MAC	Media Access Control
MACN	Monitor and Control Node
MB	Megabytes
MCA	Micro Channel® Architecture
MI	Manufacturing Interface
MIB	Management Information Database
MIMD	Multiple Instruction Stream, Multiple Data Stream
MPI	Message Passing Interface
MPL	Message Passing Library
MPP	Massive Parallel Processing
NFS	Network File System
NIM	Network Installation Management
NIS	Network Information System
NSB	Node Switch Board
NSC	Node Switch Chip
NVRAM	Non-volatile Memory
ODM	Object Data Management
OID	Object ID
OLTP	Online Transaction Processing
OSF	Open Software Foundation
PAIDE	Performance Aide for AIX
PDB	Power Distribution Bus
PE	Parallel Environment
PID	Process ID
PIOFS	Parallel I/O File System
PMAN	Problem Management
PP	Physical Partition
PSSP	Parallel System Support Programs
PTC	Prepare to Commit
PTX	Performance Toolbox for AIX
PV	Physical Volume
RAM	Random Access Memory
RCP	Remote Copy Protocol
RM	Resource Monitor
RMAPI	Resource Monitor Application Programming Interface
RPC	Remote Procedure Calls
RPQ	Request for Product Quotation
RSCT	Reliable Scalable Cluster Technology
RVSD	Recoverable Virtual Shared Disk
SAMI	Service and Manufacturing Interface
SBS	Structured Byte Strings
SCSI	Small Computer Systems Interface
SDR	System Data Repository
SMP	Symmetric Multiprocessor
SNMP	System Performance Measurement Interface
SPOT	Shared Product Object Tree

SPUM	SP User Management
SRC	System Resource Controller
SSA	Serial Storage Architecture
SUP	Software Update Protocol
TGT	Ticket-Granting Ticket
TLC	Tape Library Connection
TP	Twisted Pair
TS	Topology Services
UTP	Unshielded Twisted Pair
VLDB	Volume Location Database
VSD	Virtual Shared Disk
VSS	Versatile Storage Server™

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 556. Note that some of the documents referenced here may be available in softcopy only.

- *IBM pSeries 670 and pSeries 690 System Handbook,* SG24-7040
- *IBM (e)server Cluster 1600 Managed by PSSP 3.5: What's New,* SG24-6617
- *Configuring p690 in an IBM (e)server Cluster 1600,* REDP0187

Other publications

These publications are also relevant as further information sources:

- *RS/6000 and IBM @server pSeries PCI Adapter Placement Reference,* SA38-0538
- *Hardware Management Console for pSeries Installation and Operations Guide,* SA38-0590
- *Hardware Management Console for pSeries Maintenance Guide,* SA38-0603
- *AIX 5L Version 5.2 Commands Reference, Volume 4, N-R,* SC23-4118
- *AIX V4.3 Messages Guide and Reference,* SC23-4129
- *AIX Version 4.3 System Management Guide: Communications and Networks,* SC23-4127
- *AIX 5L V 5.2 Installation Guide and Reference,* SC23-4389
- *AIX 5L V5.2 Performance Tools Guide and Reference,* SC23-4859
- *AIX 5L V 5.2 Security Guide,* SC23-4860-01
- *Inside the RS/6000 SP,* SG24-5145
- *IBM 9077 SP Switch Router: Get Connected to the SP Switch,* SG24-5157
- *RS/6000 SP Software Maintenance,* SG24-5160

- *PSSP 3.1 Announcement,* SG24-5332
- *RS/6000 SP: PSSP 3 Survival Guide,* SG24-5344
- *RS/6000 SP Cluster: The Path to Universal Clustering,* SG24-5374
- *Exploiting RS/6000 SP Security: Keeping It Safe,* SG24-5521
- *RS/6000 SP Systems Handbook,* SG24-5596
- *AIX 4.3 Elements of Security Effective and Efficient Implementation,* SG24-5962
- *Aditional AIX Security Tools on IBM pSeries, RS/6000, and SP/Cluster,* SG24-5971
- *Managing IBM (e)server Cluster 1600 - Power Recipes for PSSP 3.4,* SG24-6603
- *IBM Cluster 1600 and PSSP 3.4 Cluster Enhancements,* SG24-6604
- *A Practical Guide for Resource Monitoring and Control (RMC),* SG24-6615
- *IBM (e)server Cluster 1600 Managed by PSSP 3.5: What's New,* SG24-6617
- *Performance and Tuning Considerations for the p690 in a Cluster 1600,* SG24-6841
- *Configuring Highly Available Clusters Using HACMP 4.5,* SG24-6845
- *An Introduction to CSM 1.3 for AIX 5L,* SG24-6859
- *An Introduction to Security in a CSM 1.3 for AIX 5L Environment,* SG24-6873
- *CSM for the PSSP System Administrator,* SG24-6953
- *IBM pSeries 670 and pSeries 690 System Handbook,* SG24-7040
- *RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment,* GA22-7280
- *RS/6000 SP Planning, Volume 2: Control Workstation and Software Environment,* GA22-7281
- *332 MHz Thin and Wide Node Service,* GA22-7330
- *PSSP Installation and Migration Guide,* GA22-7347
- *PSSP Administration Guide,* SA22-7348
- *PSSP Managing Shared Disks,* SA22-7349
- *PSSP Diagnosis Guide,* GA22-7350
- *PSSP Command and Technical Reference (2 Volumes),* SA22-7351
- *PSSP Messages Reference,* GA22-7352
- *CSM for Linux: Software Planning and Installation Guide,* SA22-7853

- *CSM for Linux: Hardware Control Guide,* SA22-7856
- *IBM (e)server Cluster 1600: Planning, Installation, and Service,* GA22-7863
- *CSM for Linux: Administration Guide,* SA22-7873
- *PSSP Implementing a Firewalled RS/6000 SP System,* GA22-7874
- *RSCT for AIX 5L: Group Services Programming Guide and Reference,* SA22-7888
- *RSCT for AIX 5L: Guide and Reference,* SA22-7889
- *RSCT for AIX 5L: Technical Reference,* SA22-7890
- *RSCT for AIX 5L: Messages,* GA22-7891
- *RSCT for Linux: Guide and Reference,* SA22-7892
- *RSCT for Linux: Technical Reference,* SA22-7893
- *RSCT for Linux: Messages,* GA22-7894
- *CSM for AIX 5L: Administration Guide,* SA22-7918
- *CSM for AIX 5L: Software Planning and Installation Guide,* SA22-7919
- *CSM for AIX 5L: Hardware Control Guide,* SA22-7920
- *CSM for AIX 5L: Command and Technical Reference,* SA22-7934
- *HACMP ES V4.5: Concepts and Facilities Guide,* SC23-4276
- *HACMP ES V4.5: Enhanced Scalability Installation and Administration Guide,* SC23-4306
- *HAGEO V2.4: Planning and Administration Guide,* SC23-1886
- *HAGEO V2.4: Concepts and Facilities,* SC23-1922
- *GEORM V2.4: Planning and Administration Guide,* SC23-4308
- *GEORM V2.4: Concepts and Facilities,* SC23-4307
- *GPFS V2.1: Concepts, Planning, and Installation,* GA22-7899
- *GPFS V2.1: Administration and Programming Reference,* SA22-7900
- *LoadLeveler V3.1: Using and Administering,* SA22-7881
- *IBM PE for AIX 5L V3.2: Installation,* GA22-7418
- *IBM PE for AIX 5L V3.2: Hitchhiker's Guide,* SA22-7424
- *IBM ESSL for AIX 5L V4.1: ESSL Products General Information,* GA22-7903
- *IBM ESSL for AIX 5L V4.1: ESSL Guide and Reference,* SA22-7904
- *IBM DCE for AIX: Introduction to DCE* and *IBM DCE for AIX: DFS Administration Guide and Reference* can be found at:

  `http://www.ibm.com/software/network/dce/library/publications/`

- *IBM AIX: Kernel and Subsystems Technical Reference, Volume 2 can be found at:*

  http://publibn.boulder.ibm.com/doc_link/en_US/a_doc_lib/libs/ktechrf2/About.htm

Online resources

These Web sites and URLs are also relevant as further information sources:

- pSeries whitepapers

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/power4.html
- IBM technical support home page

  http://techsupport.services.ibm.com/server/mdownload2/download.html
- HMC corrective services Web site

  http://techsupport.services.ibm.com/server/hmc/corrsrv.html
- pSeries library Web site

  http://www.ibm.com/server/eserver/pseries/library/sp_books/pssp.html
- Storage subsystems Web site:

  http://www.storage.ibm.com/
- pSeries storage subsystems Web site

  http://www.storage.ibm.com/products_pseries.html

How to get IBM Redbooks

You can search for, view, or download Redbooks, Redpapers, Hints and Tips, draft publications and Additional materials, as well as order hardcopy Redbooks or CD-ROMs, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

D

E

F

M

N

P

R

S

IBM @server Certification Study Guide: Cluster 1600 Managed by PSSP

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

IBM @server Certification Study Guide: Cluster 1600 Managed by PSSP

Redbooks

Handy deskside reference for Cluster 1600 systems

Detailed, step-by-step exercises

Sample certification test questions and answers

This IBM Redbook is designed as a study guide for professionals wishing to prepare for the certification exam to achieve IBM Certified Specialist - Cluster 1600 managed by PSSP.

The Cluster 1600 managed by PSSP certification validates the skills required to install and configure PSSP system software and to perform the administrative and diagnostic activities needed to support multiple users in an SP environment. The certification is applicable to specialists who implement and/or support Cluster 1600 managed by PSSP systems.

This redbook helps Cluster 1600 specialists seeking a comprehensive and task-oriented guide for developing the knowledge and skills required for certification. It is designed to provide a combination of theory and practical experience needed for a general understanding of the subject matter. It also gives sample questions that will help in the evaluation of personal progress and provides familiarity with the types of questions that will be encountered in the exam.

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

SG24-7013-00 ISBN 0738499056