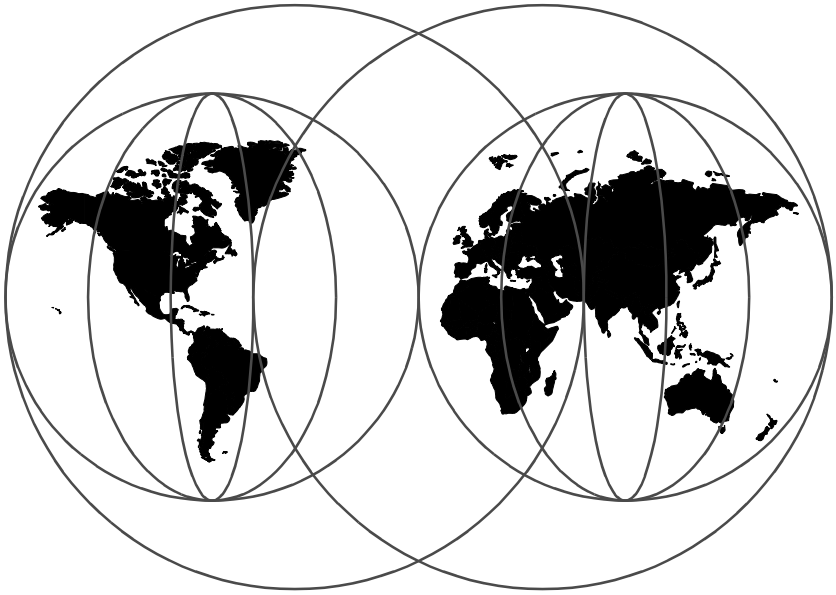
IBM Certification Study Guide: RS/6000 SP

Marcelo R. Barrios, Bruno Blanchard, Kyung C. Lee
Olivia P. Liu, Ipong Hadi Trisna



International Technical Support Organization

http://www.redbooks.ibm.com

SG24-5348-00

SG24-5348-00

International Technical Support Organization

IBM Certification Study Guide: RS/6000 SP

May 1999

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix B, “Special Notices” on page 467.

First Edition (May 1999)

This edition applies to PSSP Version 3, Release 1 (5765-D51) and PSSP Version 2, Release 4 (5765-529) for use with the AIX Version 4, Release 3 Operating System.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. HYJ Mail Station P099
522 South Road
Poughkeepsie, New York 12601-5400

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Figures

Tables

Preface

The AIX & 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 are available. They include:

- IBM Certified AIX User
- IBM Certified Specialist - RS/6000 Solution Sales
- IBM Certified Specialist - AIX System Administration
- IBM Certified Specialist - AIX System Support
- IBM Certified Specialist - RS/6000 SP
- 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 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 much more meaningful than the typical test that only captures classroom knowledge. These experienced Subject Matter Experts ensure 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.

This redbook is designed as a study guide for professionals wishing to prepare for the certification exam to achieve IBM Certified Specialist - RS/6000 SP.

The RS/6000 SP specialist certification validates the skills required to install and configure RS/6000 Scalable POWERparallel (SP) system software and to perform the administrative and diagnostic activities needed to support multiple users in an SP environment. The certification is applicable to specialist who implement and/or support RS/6000 SP systems.

This redbook helps RS/6000 SP 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 will not replace the practical experience you should have, but is an effective tool, 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 if you just want to validate your RS/6000 SP skills, this book is for you.

For additional information about certification and instructions on how to register for an exam, call IBM at 1-800-8322 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, Poughkeepsie Center.

Marcelo R. Barrios is a project leader at the International Technical Support Organization, Poughkeepsie Center. He has been with IBM for five years working in different areas related to RS/6000. Currently, he focuses on RS/6000 SP technology by writing redbooks and teaching IBM classes worldwide.

Bruno Blanchard is an IT specialist working at the EMEA Technical Support at IBM France. He holds a Engineer degree from Ecole Centrale de Paris, and a Master of Science from Oregon State University. He has been with IBM since 1983 as a system engineer for VM on 43xx, 308x, 309x and for AIX on PS/2, RT, RS/6000, and SP system. His areas of expertise also include network management of IBM 2220, 2219, and 2225.

Kyung C. Lee is a Senior IT Specialist working for RS/6000 National Practice based on Chicago at IBM US. He has worked on UNIX environments for 11 years and has extensive experience on SP environments since 1995. He holds BS and MS degree in Electrical Engineering and Computer Science from University of Illinois. His areas of expertise include AIX, RS/6000 SP, DB2, DB2/PE, and Oracle.

Olivia P. Liu is an Advisory IT Specialist working for RS/6000 Techline at IBM Australia. She has over ten years of experience in the Information Technology

field in the UNIX/AIX and COBOL environments. She holds a B.A. degree from the University of Western Ontario, Canada and an M.B.A. (Technology Management) degree from APESMA/Deakin University in Australia. She is currently doing her Doctoral studies in Business Administration involving research in Marketing and MIS. Her areas of expertise include UNIX, AIX, COBOL, RS/6000, marketing, and sales. She has written extensively on technology management.

Ipong Hadi Trisna has been working for IBM Indonesia since 1995 as a System Services Rep. for RS/6000 Division to support SP System and HACMP. He holds an Engineer degree from Indonesia Institute of Technology Jakarta, Indonesia majoring in Electronic and Computer Engineering. His expertise area is hands on in the field for RS/6000 especially SP System and HACMP.

We wish to thank the following people for their invaluable contributions to this project:

Becky Gonzalez
IBM Austin

John Owczarzak, Elizabeth Barnes
Editing Team, International Technical Support Organization, Austin Center

Comments Welcome

Your comments are important to us!

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

- Fax the evaluation form found in “ITSO Redbook Evaluation” on page 491 to the fax number shown on the form.
- Use the online evaluation form found at: `http://www.redbooks.ibm.com/`
- Send us a note at the following address: `redbook@us.ibm.com`

Chapter 1. Introduction

This guide is not a replacement for the SP product documentation or existing ITSO redbooks or to the value of real experience installing and configuring RS/6000 SP environments.

RS/6000 SP knowledge only 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 referred documentation listed in each chapter.

The RS/6000 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 the customers 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 as well as a similar scenario where all the chapters in the redbook refer to. This scenario is described in 1.2, “The Test Scenario” on page 2.

1.1 Book Organization

This guide intends to present you with all domains in the scope of the RS/6000 SP Certification exam. The structure of the book follows the normal flow that a standard RS/6000 SP installation may have.

Part 1, “System Planning” on page 5, contains chapters dedicated to the initial planning as well as to the initial setup of a standard RS/6000 SP environment. It also includes concepts and examples about SP security and user management.

Part 2, “Installation and Configuration” on page 233, contains chapters describing the actual implementation of the different 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 295, contains chapters for the planning and configuration of additional products that are present in most of the RS/6000 SP 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.

Part 4, “On-going Support” on page 367, 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* - It 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. First is to check your progress with the topics covered in the chapter. Second is to become familiar with the type of questions you may encounter in the exam.

There are many ways to perform the same action in an SP environment: Command line, SMIT or SMITTY, spmon -g (PSSP 2.4 or below), IBM SP Perspectives, and so on. The certification exam is not restricted to one of these methods. You are supposed to know each one, in particular, the syntax of the most useful commands.

1.2 The Test Scenario

As a way to present you with a similar situation to the one you may encounter in the SP Certification exam, we have included a test scenario that we will use in all sections of this study guide. The scenario is depicted in Figure 1 on page 3.

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

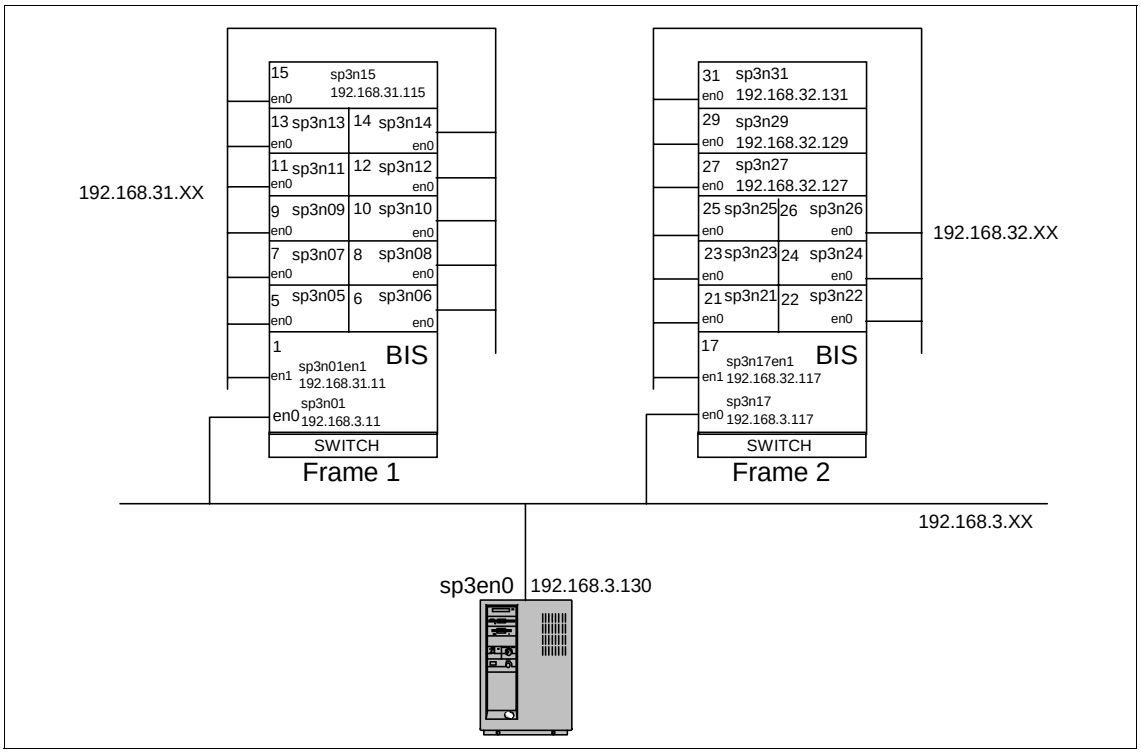


Figure 1. Study Guide Test Environment

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 have decided to include multiple segments in order to introduce an environment where networking and specially routing has to be considered. Details about networking can be found in Chapter 3, “RS/6000 SP Networking” on page 73.

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 nodes in that frame.

The frame numbering has been 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

Chapter 2. Validate Hardware and Software Configuration

This chapter discusses the hardware components of the RS/6000 SP, such as node types, control workstation, frames, and switches. It also provides some additional information on disk, memory, and software requirements.

2.1 Key Concepts You Should Study

The topics covered in this section provides a good preparation toward the RS/6000 SP certification exam. Before taking the exam, make sure you understand the following key concepts:

- What hardware components 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 RS/6000 SP are:

- The frame with its integral power supply.
- Processor nodes.
- Optional dependent nodes that serve a specific function, such as high-speed network connections.
- Optional SP Switch and Switch-8 to expand your system.
- Control workstation (a high-availability option is also available).
- 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) making the RS/6000 SP system accessible from any network-attached workstation.

Figure 2 on page 8 shows a sample of RS/6000 SP components.

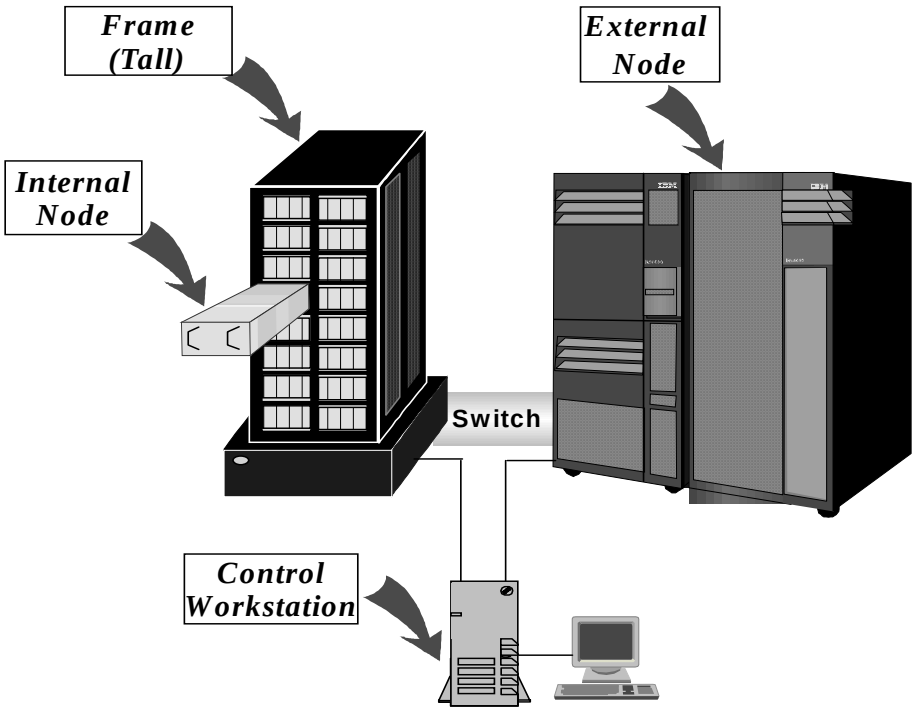


Figure 2. Sample RS/6000 SP with External Node

2.3 Frames

The building block of RS/6000 SP is the *frame*. There are two sizes: The tall frame (75.8" high) and the short frame (49" 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

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 units 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.

2.3.1 Tall Frames

The tall model frame (model 550) and the tall expansion frame (feature code #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).

2.3.2 Short Frames

The short model frame (model 500) and the short expansion frame (feature code #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 3 illustrates short frame components from the front view.

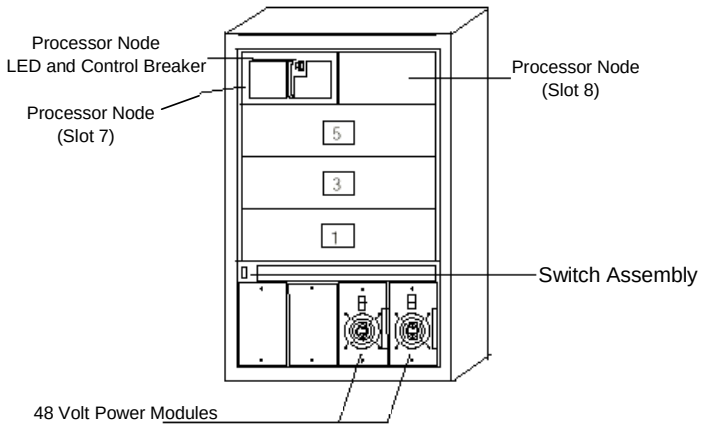


Figure 3. Front View of Short Components

2.3.3 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 (feature code #2031) contains four ISBs. An SP Switch frame with four ISBs will support 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 4 shows an SP Switch frame with eight ISBs.

Note

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

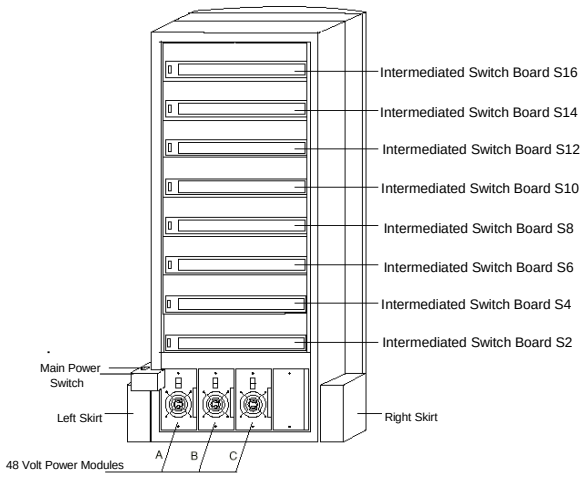


Figure 4. SP Switch Frame with Eight Intermediated Switch Boards (ISB)

2.3.4 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 on short frames (feature code #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 two power supplies and a third, optional one, for N+1 support.

Figure 5 on page 12 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 *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment,* GA22-7280.

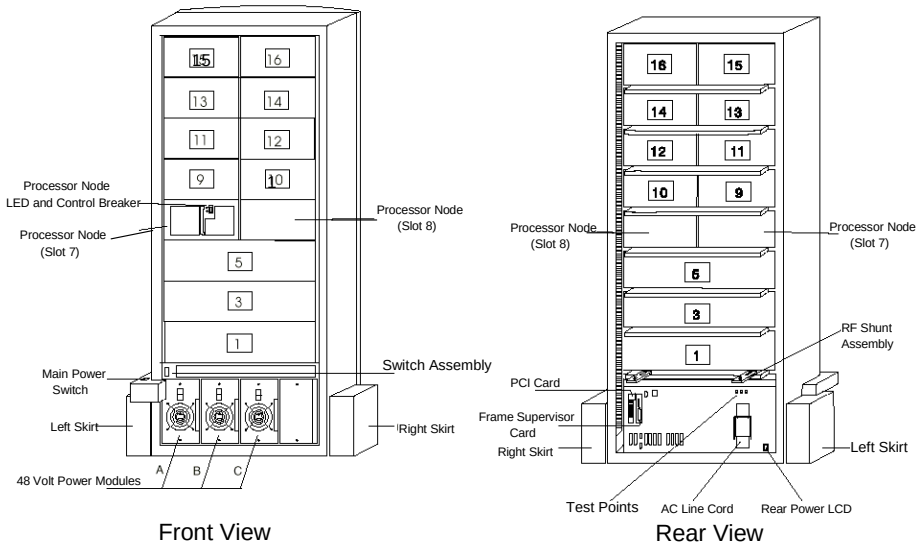


Figure 5. Front and Rear Views of Tall Frame Components

2.3.5 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 6 on page 13.

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)
- SAMI cable

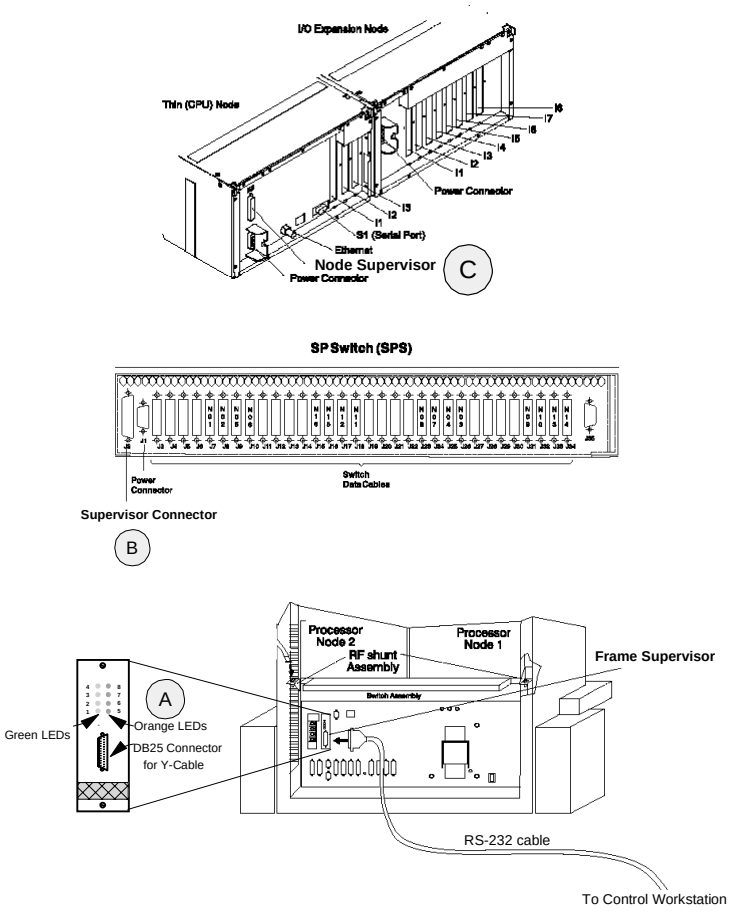


Figure 6. Frame Supervisor Attachment

There is a cable that 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.

2.4 Standard Nodes

The basic RS/6000 SP building block is the server node or standard node. Each node is a complete server system, comprising processor(s), memory, internal disk drive, expansion slots, and its own copy of the AIX operating system. The basic technology is shared with standard RS/6000 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 systems.

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

2.4.1 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 14 internal node types excluding some special *on request* node types. There are five most current nodes: 160 MHz Thin P2SC node, 332 MHz SMP Thin node, 332 MHz SMP Wide node, POWER3 SMP Thin node, and POWER3 SMP Wide node. Only the 160 MHz Thin P2SC node utilizes Micro Channel Architecture (MCA) bus architecture while the others use PCI bus architecture.

160 MHz Thin P2SC Nodes

This node is based on the POWER2 Super Chip (P2SC) implementation of the POWER architecture. Each node contains a 160 MHz P2SC processor combining IBM RISC microprocessor technology and the IBM implementation of the UNIX operating system, AIX. The standard memory in each node is 64 MB expandable to 1 GB maximum. The minimum internal disk storage in each node is 4.5 GB expandable to 18.2 GB. Each node has two disk bays, four Micro Channel slots, and integrated SCSI-2 Fast/Wide and Ethernet (10 Mbps) adapters. This node is equivalent to the RS/6000 stand-alone model 7012-397.

332 MHz SMP Thin Nodes

This node is the first PCI architecture bus node of the RS/6000 SP. Each node has two or four PowerPC 604e processors running at a 332 MHz clock

cycle, two memory slots with 256 MB, expandable to 3 GB, of memory and integrated Ethernet (10 Mbps) and SCSI-2 Fast/Wide I/O adapters to maximize the number of slots available for application use. This Thin node has two internal disk bays with a maximum of 18.2 GB (mirror) and two PCI I/O expansion slots (32-bit). The 332 MHz SMP Thin node can be upgraded to the 332 MHz SMP Wide node.

332 MHz SMP Wide Nodes

The 332 MHz SMP Wide node is a 332 MHz SMP Thin node combined with additional disk bays and PCI expansion slots. This wide node has four internal disk bays with a maximum of 36.4 GB (mirror) and ten PCI I/O expansion slots (three 64-bit, seven 32-bit). Both 332 MHz SMP Thin and Wide nodes are based on the same technology as the RS/6000 model H50 and have been known as the *Silver* nodes. Figure 7 shows a 332 MHz SMP node component diagram.

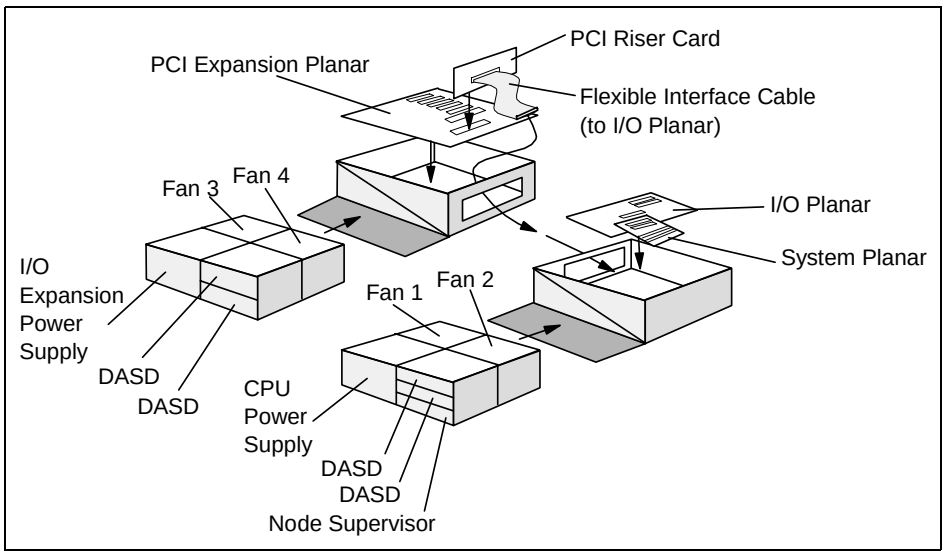


Figure 7. 332 MHz SMP Node Component Diagram

POWER3 SMP Thin Nodes

This node is the first 64-bit internal processor node of the RS/6000 SP. Each node has a one- or two-way (within two processor cards) configuration utilizing a 64-bit 200 MHz POWER3 processor with a 4 MB Level 2 (L2) cache per processor. The standard ECC SDRAM memory in each node is 256 MB expandable up to 4 GB (within two card slots). This new node is shipped with

disk pairs as a standard feature to encourage the use of mirroring to significantly improve system availability. This Thin node has two internal disk bays for pairs of 4.5 GB, 9.1 GB and 18.2 GB Ultra SCSI disk capacity. Each node has two 32-bit PCI slots and integrated 10/100 Ethernet and Ultra SCSI adapters. The POWER3 SMP Thin node can be upgraded to the POWER3 SMP Wide node.

POWER3 SMP Wide Nodes

The POWER3 SMP Wide node is a POWER3 SMP Thin node combined with additional disk bays and PCI expansion slots. This Wide node has four internal disk bays for pairs of 4.5 GB, 9.1 GB, and 18.2 GB Ultra SCSI disk capacity. Each node has ten PCI slots (two 32-bit, eight 64-bit). Both POWER3 SMP Thin and Wide nodes are equivalent to the RS/6000 43P model 260. A diagram of the POWER3 SMP node is shown in Figure 8. Notice that it uses docking connectors (position A) instead of flex cables as in the 332 MHz node.

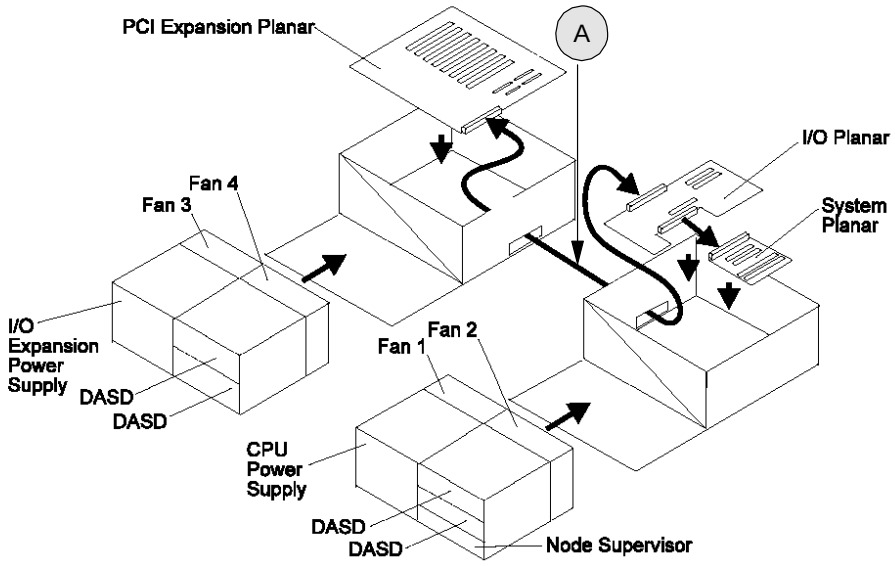


Figure 8. POWER3 SMP Node Component Diagram

The minimum software requirements for POWER3 SMP Thin and Wide nodes are the AIX Version 4.3.2 and PSSP Version 3.1.

Table 1 shows a comparison of current nodes.

Table 1. Current Nodes Comparison

Node Type	**160 MHz Thin**	**332 MHz SMP Thin**	**332 MHz SMP Wide**	**POWER3 SMP Thin**	**POWER3 SMP Wide**
Processor	160 MHz P2SC	332 MHz 2- or 4- way PowerPC 604e		200 MHz 1- or 2- way POWER3	
L1 Cache (Instr./Data) per processor	32 KB/ 128 KB	32 KB / 32 KB		32 KB / 64 KB	
L2 Cache (per processor)	-	256 KB		4 MB	
Std. Memory	64 MB	256 MB		256 MB	
Max. Memory	1 GB	3 GB		4 GB	
Memory Slots	4	2		2	
Disk Bays	2	2	4	2	4
Min. Int. Disk	4.5 GB	None Required		None Required	
Max. Int. Disk	18.2 GB	36.4 GB or 18.2 GB (Mirror)	72.8 GB or 36.4 GB (Mirror)	36.4 GB or 18.2 GB (Mirror)	72.8 GB or 36.4 GB (Mirror)
Expansion Slots	4 MCA	2 PCI (32-bit)	10 PCI (3 64-bit, 7 32-bit)	2 PCI (32-bit)	10 PCI (8 64-bit, 2 32-bit)
Adapters	Integrated SCSI-2 F/W and Ethernet (10 Mbps)	Integrated SCSI-2 F/W and Ethernet (10 Mbps)		Integrated Ultra SCSI and Ethernet (10/100 Mbps)	

2.4.1.1 332 MHz SMP Node System Architecture

The 332 MHz SMP Thin and Wide nodes provide two- or four- way symmetric multiprocessing utilizing PowerPC technology and extend the RS/6000 PCI I/O technology to the SP system. With their outstanding integer performance, these nodes are ideal for users who need mission-critical commercial computing solutions. The 332 MHz SMP node system structure is shown in Figure 9 on page 18.

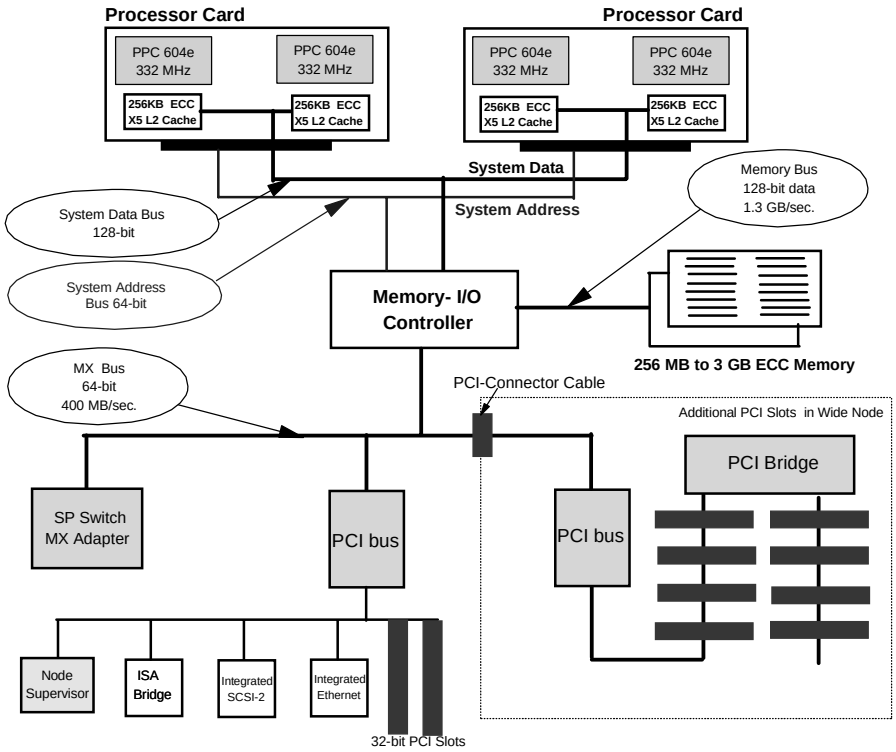


Figure 9. 332 MHz SMP Node System Architecture

Processor and Level 2 Cache Controller

The 332 MHz SMP node contains two- or four-way 332 MHz PowerPC 604e processors each with its own 256 KB Level 2 cache. The X5 Level 2 cache controller incorporates several technological advancements in design providing greater performance over traditional cache designs. The cache controller implements an eight-way, dual-directory, set-associative cache using SDRAM. When instructions or data are stored in a cache, they are grouped into sets of eight 64-byte lines. The X5 maintains an index to each of the eight sets. It also keeps track of the tags used internally to identify each cache line. Dual tag directories allow simultaneous processor requests and system bus snoops, thus, reducing resource contention and speeding up access.

System Bus

The SMP system bus is optimized for high performance and multiprocessing applications. It has a separate 64-bit address bus and 128-bit data bus. These buses operate independently in the true split transaction mode and are aggressively pipelined. For example, new requests may be issued before previous requests are completed. There is no sequential ordering requirement. Each operation is tagged with an 8-bit tag, which allows a maximum of up to 256 transactions to be in progress in the system at any one time.

System Memory

The 332 MHz SMP node supports 256 MB to 3 GB of 10-nanosecond SDRAM. System memory is controlled by the memory-I/O chip, which is capable of providing a sustained memory bandwidth of over 1.3 GB per second. The memory controller supports up to two memory cards with up to eight increments of SDRAM on each card.

I/O Subsystem

The memory-I/O controller implements a 64-bit, multiplexed address and data bus for attaching several PCI I/O buses and the SP Switch MX adapter. This bus runs concurrent with, and independent from, the system and memory buses. The peak bandwidth of this bus is 400 MB per second. Two 32-bit PCI slots are in the thin node, and three additional 64-bit PCI slots and five 32-bit PCI slots are in the wide node.

2.4.1.2 POWER3 SMP Node System Architecture

The POWER3 SMP node has excellent performance for compute-intensive analysis applications. The heart of this node is the POWER3 microprocessor based on IBM PowerPC architecture and RS/6000 Platform architecture. It provides a high bandwidth interface to a fast Level 2 (L2) cache and a separate high bandwidth interface to memory and other system functions. The POWER3 microprocessor implements the 64-bit PowerPC architecture and is fully compatible with existing 32-bit applications.

The POWER3 SMP node system structure is shown in Figure 10 on page 20.

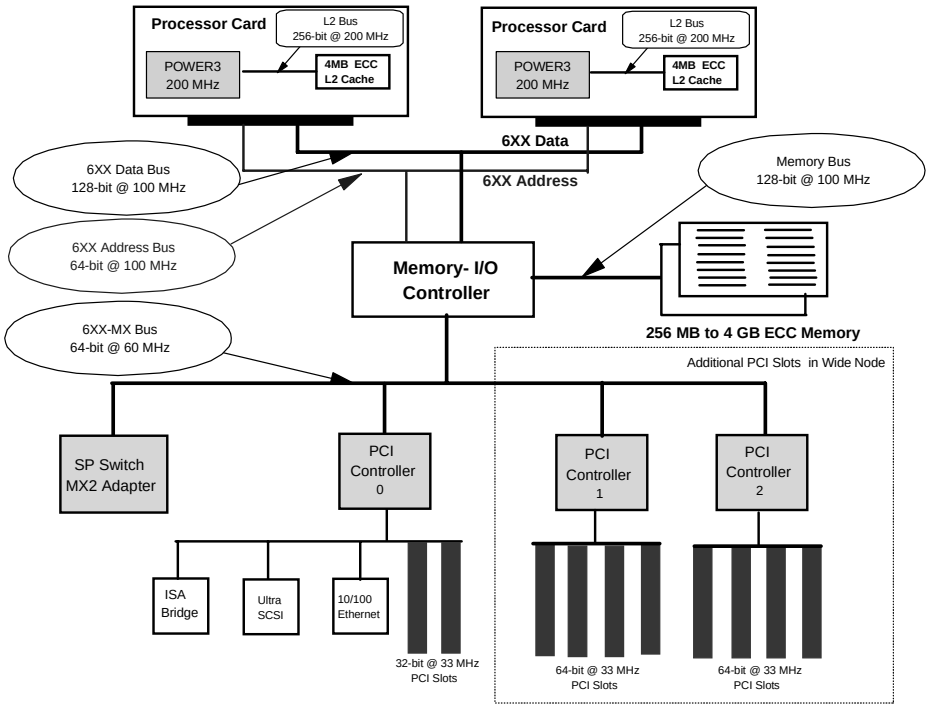


Figure 10. POWER3 SMP Node System Architecture

POWER3 Microprocessor

The POWER3 is a single chip implemented with 0.25 micron CMOS technology. It operates at a 200 MHz clock cycle. The POWER3 design contains a superscalar core that is comprised of eight execution units and allows concurrent operation of fixed-point, load/store, branch, and floating-point instructions. The processor can perform up to four floating-point operations per clock cycle. There is a 32 KB instruction and a 64 KB data Level 1 cache integrated within a single chip. Both instruction and data cache are parity protected. There is a 256-bit external interface to the 4 MB Level 2 cache, which operates at 200 MHz and is ECC protected.

System Bus

The system bus, referred to as the 6XX bus, connects up to two POWER3 processors to the memory-I/O controller chip set. It provides 40 bits of real address and a separate 128-bit data bus. The address, data, and tag buses

are fully parity protected. The 6XX bus runs at a 100 MHz clock rate, and peak data throughput is 1.6 GB/second.

System Memory

The POWER3 SMP node supports 256 MB to 4 GB of 10 nanosecond SDRAM. System memory is controlled by the memory-I/O chip set through the memory bus. The memory bus consists of a 128-bit data bus and operates a at 100 MHz clock cycle. It is separated from the system bus (6XX bus), which allows for concurrent operations on these two buses. For example, cache-to-cache transfers can occur while a Direct Memory Access (DMA) operation is proceeding to an I/O device. There are two memory card slots each supporting up to 16 128 MB memory DIMMs. Memory DIMMs must be used in pairs, and at least one memory card with a minimum of 256 MB memory is required to be operational. System memory is protected with a Single Error Correction, Double Error Detection ECC code.

I/O Subsystem

The Memory-I/O controller chip set implements a 64-bit plus parity, multiplexed address, and data bus (6XX-MX bus) for attaching three PCI controller chips and the SP Switch MX2 adapter. The 6XX-MX bus runs at 60 MHz clock cycle, the peak bandwidth of the 6XX-MX bus is 480 MB/second. The three PCI controller attached to the 6XX-MX bus provides the interface for ten PCI slots. Two 32-bit PCI slots are in the thin node, and eight additional 64-bit PCI slots are in the wide node. One of the PCI controller chips (controller chip 0) provides support for integrated Ultra2 SCSI and 10Base2, 10/100BaseT Ethernet functions. The Ultra2 SCSI interface supports up to four internal disks. An ISA bridge chip is also attached to PCI controller chip 0 for supporting two serial ports and other internally used functions in the POWER3 SMP node.

Service Processor

The service processor function is integrated on the I/O planner board. This service processor performs system initialization, system error recovery, and diagnostic functions that give the POWER3 SMP node a high level of availability. The service processor is designed to save the state of the system to 128 KB of nonvolatile memory (NVRAM) to support subsequent diagnostic and recovery actions taken by other system firmware and the AIX operating system.

2.4.2 External Nodes

A external node is a kind of processor node that cannot be housed in the frame due to its large size. The current supported external nodes are RS/6000 model S70 and RS/6000 model S70 Advanced (S7A). Both are large enterprise server class utilizing 64-bit symmetric multiprocessor (SMP) system that supports 32- and 64-bit applications concurrently. The bus architecture in these servers is PCI architecture. The differences between both models are the base processor (PowerPC RS64 125 MHz and PowerPC RS64 II 262 MHz, respectively), standard memory, and high availability I/O drawer on S70 Advanced Server. The external node is known as a SP-Attached server.

These servers excel in capacity and scalability in On-line Transaction Processing (OLTP), Server Consolidation, Supply Chain Management, and Enterprise Resource Planning (ERP), such as SAP, which single large database servers are required.

2.4.2.1 SP-Attached Servers

The RS/6000 Enterprise Server Model S70 and Model S7A 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 rack. Up to three more I/O racks can be added to a system. Figure 11 on page 23 shows the RS/6000 model S70 and S7A scalability.

The Central Electronics Complex contains:

- Either 64-bit 125 MHz PowerPC RS64 processors (S70) or 262 MHz PowerPC RS64 II processors (S7A).
- Optional 4-way processor cards (the same processor) that scale configuration to eight-way or 12-way SMP processing.
- 4 MB ECC L2 cache memory per 125 MHz processor or 8 MB per 262 MHz processor.
- Standard 512 MB ECC SDRAM memory (S70) or 1 GB ECC SDRAM memory (S7A) expands to 32 GB.
- A high-speed multi-path switch, memory controller and two high-speed memory ports with a total collective memory bandwidth of up to 5.33 GB/sec. (S70) and 5.6 GB/sec. (S7A).

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 4.5GB SCSI -2 Fast/Wide disk drive (S70) or 9.1 GB UltraSCSI disk drive (S7A).
- A 32X (Max) CD-ROM.
- A 1.44 MB 3.5-inch diskette drive.
- A service processor.
- Fourteen PCI slots (nine 32-bit and five 64-bit). Eleven slots are available.
- Three media bays (Two available) for S70.
- Two media bays (One available) for S7A.
- Twelve hot-swapped disk drive bays (Eleven available).

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 (S70) and Two (S7A) available media bays.
- Twelve available hot-swapped disk drive bays.

When all four I/O drawers are installed, the S70 contains twelve media bays (Eight media bays for S7A), forty-eight hot-swapped disk drive bays, and fifty-six PCI slots per system.

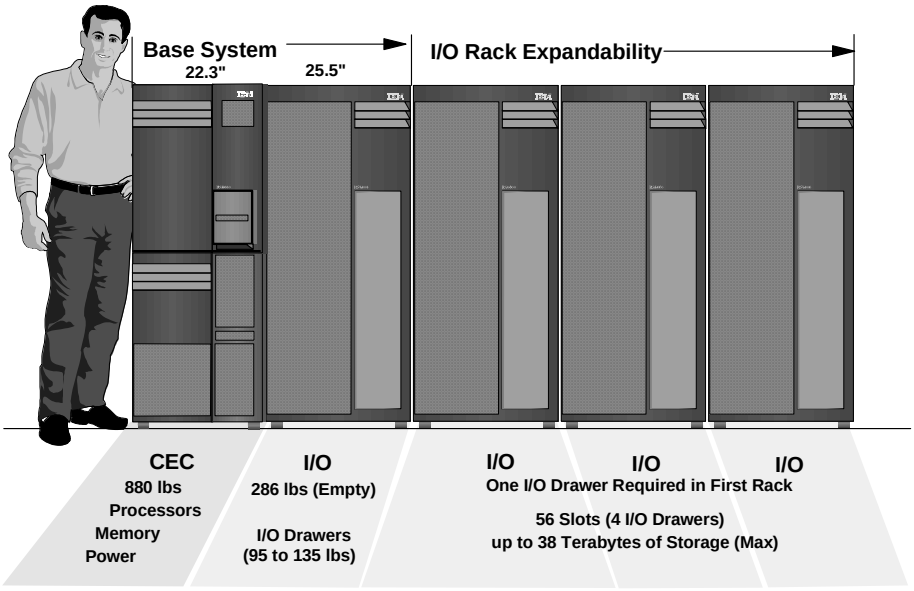


Figure 11. RS/6000 S70/S7A System Scalability

2.4.2.2 SP-Attached Server Attachment

It is important to note that the size of the S70 and S7A prohibit it from being physically mounted in the SP frame. Since the SP-attached server is mounted in its own rack and is directly attached to the control workstation using two RS-232 cable 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.

The SP-Attached server requires a minimum of four connections with the SP system in order to establish a functional and safe network. If your SP system is configured with an SP Switch, there will be five required connections as shown in Figure 12 on page 25.

Three connections are required with the control workstation.

1. An Ethernet connection to the SP-LAN for system administration purposes
2. A RS-232 cable connecting the control workstation to the SAMI port of SP-attached server (front panel)
3. A second RS-232 cable connecting the control workstation to the serial port of SP-attached server (S1 port)

The fourth connection is a 10 m frame-to-frame electrical ground cable.

The fifth connection is required if the SP system is switch configured.

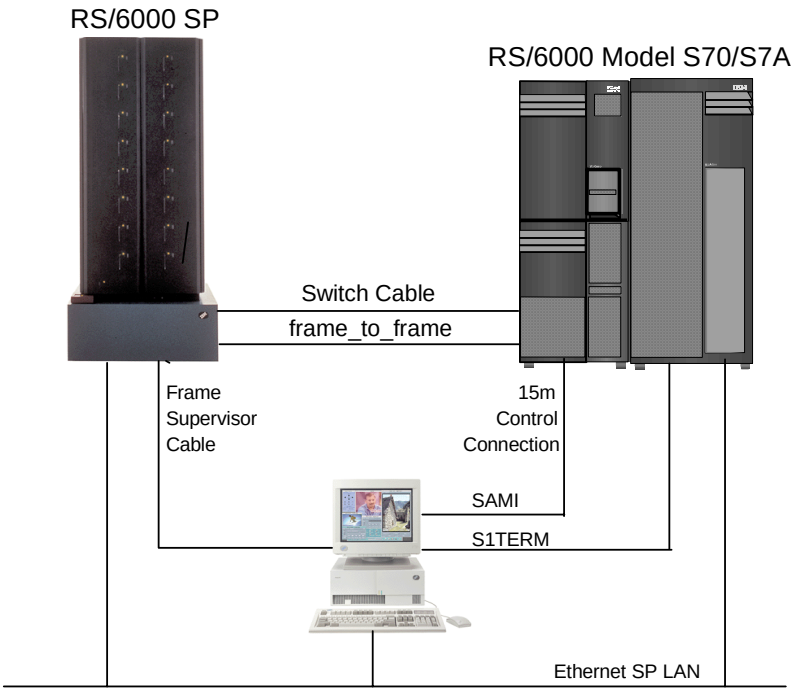


Figure 12. The SP-Attached Server Connection

2.5 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:

1. A physical dependent node - The hardware device requiring SP processor node support.
2. 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.
3. 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.

4. A cable - To connect the dependent node adapter with the logical dependent node. It connects the extension node to the SP system.

2.5.1 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 HIPPI backbones or other communications subsystems, such as ATM or 10/100 Ethernet (see Figure 13 on page 27).

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 (feature code #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 RS/6000 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

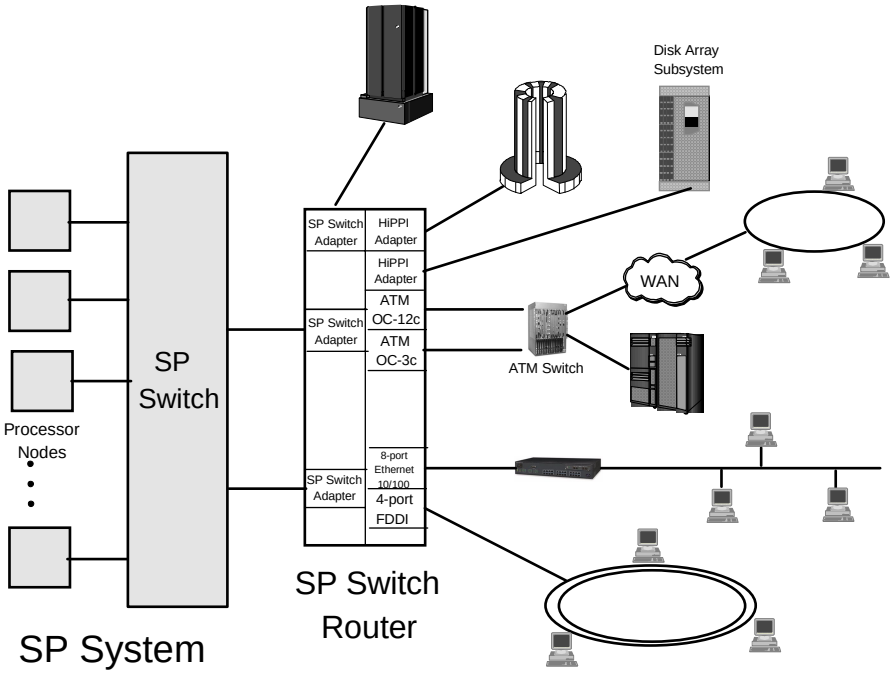


Figure 13. SP Switch Router

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 14.

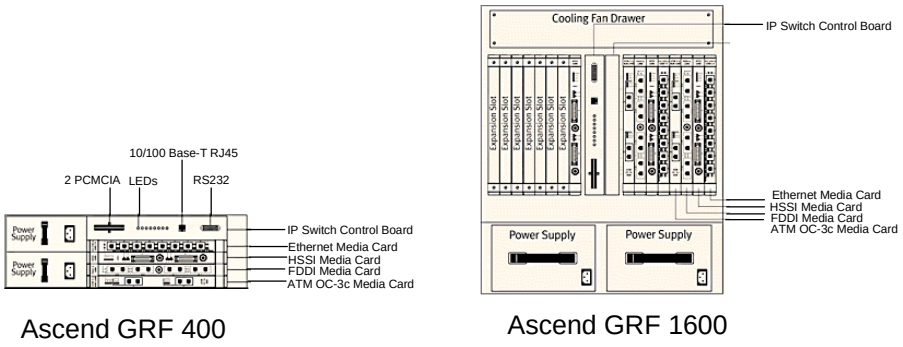


Figure 14. GRF Model 400 and 1600

2.5.2 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.

2.6 Control Workstation

The RS/6000 SP system requires an RS/6000 workstation. The control workstation serves as a point of control for managing, monitoring, 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. It also connects to each external node or SP-attached server with two 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 in the RS/6000 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.

2.6.1 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.

MCA-based Control Workstations:

- RS/6000 7012 Models 37T, 370, 375, 380, 39H, 390, 397, G30, and G40
- RS/6000 7013 Models 570, 58H, 580, 59H, 590, 591, 595, J30, J40, and J50 (see note 1)
- RS/6000 7015 Models 97B, 970, 98B, 980, 990, R30, R40, and R50 (see note 1)
- RS/6000 7030 Models 3AT, 3BT, and 3CT

Note:

1. Requires a 7010 Model 150 X-Station and display. Other models and manufacturers that meet or exceed this model can be used. An ASCII terminal is required as the console.

PCI-based Control Workstations:

- RS/6000 7024 Models E20 and E30 (see note 1)
- RS/6000 7025 Model F30 (see notes 1 and 2)
- RS/6000 7025 Models F40 and F50 (see note 3)
- RS/6000 7026 Models H10 and H50 (see note 3)
- RS/6000 7043 43P Models 140 and 240 (see notes 3, 4, and 5)

Notes:

1. Supported by PSSP 2.2 and later
2. On systems introduced since PSSP 2.4, either the 8-port (feature code #2493) or 128-port (feature code #2944) PCI bus asynchronous adapter should be used for frame controller connections. IBM strongly suggests you use the support processor option (feature code #1001). If you use this option, the frames must be connected to a serial port on an asynchronous adapter and not to the serial port on the control workstation planar board.

3. The native RS232 ports on the system planar can not be used as tty ports for the hardware controller interface. The 8-port asynchronous adapter EIA-232/ RS-422, PCI bus (feature code #2943), or the 128-port Asynchronous Controller (feature code #2944) are the only RS232 adapters that are supported. These adapters require AIX 4.2.1 or AIX 4.3 on the control workstation.
4. The 7043 can only be used on SP systems with up to four frames. This limitation applies to the number of frames and not the number of nodes. This number includes expansion frames.
5. The 7043-43P is NOT supported as a control workstation whenever an S70/S7A is attached to the SP. The limitation is due to the load that the extra daemons place on the control workstation.

2.6.2 Control Workstation Minimum Hardware Requirements

The minimum hardware requirements for the control workstation are:

- At least 128MB of main memory. For SP systems with more than 80 nodes, 256MB is required, and 512MB of memory is recommended.
- 4 GB of disk storage plus 1GB for each AIX release and modification level in your SP system. Double the number of physical disks if you plan on using rootvg mirroring.
- Physically installed with the RS232 cable to each SP frame.
- Physically installed with two RS232 cables to each external node SP-attached server, such as an RS/6000 Enterprise Server Model S70 or S70 Advanced.
- Equipped with the following I/O devices and adapters:
  - A 3.5 inch diskette drive.
  - Four or eight millimeter (or equivalent) tape drive.
  - SCSI CD-ROM drive.
  - One RS232 port for each SP frame.
  - Keyboard and mouse.
  - 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.
  - SP Ethernet adapters for connection to the SP Ethernet (see 3.3.1, “SP Ethernet” on page 85 for details).

2.6.3 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 SP frame supervisor card to each control workstation. This HACWS configuration provides automated detection, notification, and recovery of control workstation failures. Figure 15 shows the logical view of the HACWS attachment.

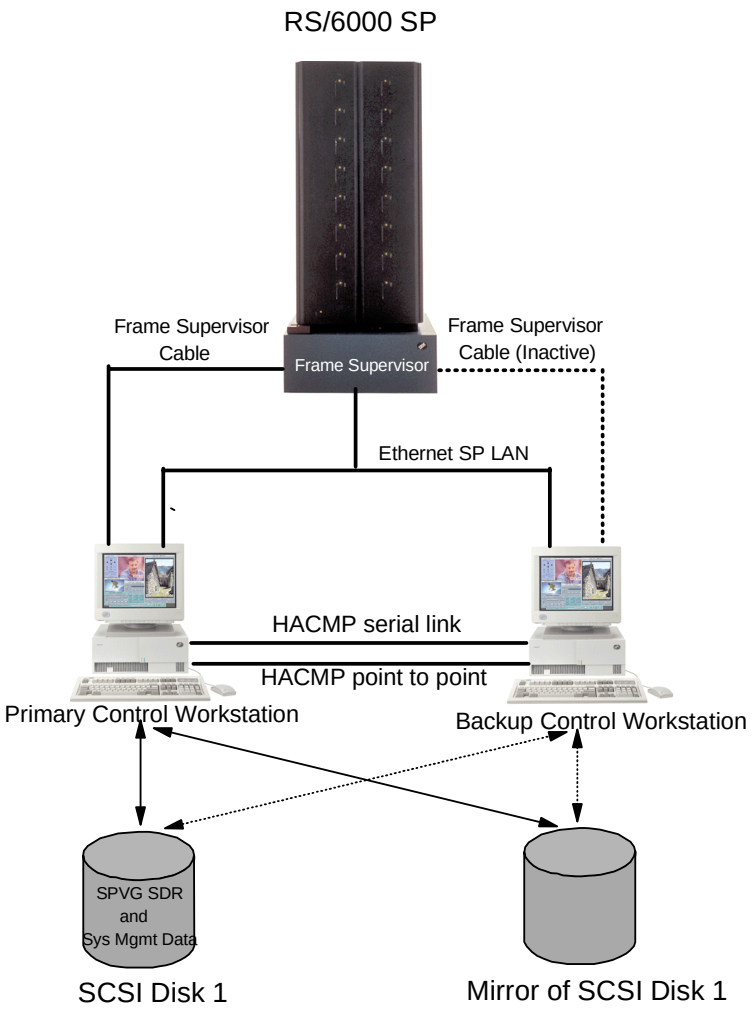


Figure 15. High Availability Control Workstation (HACWS) Attachment

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 at 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 a an SP authentication server, the backup control workstation must be a secondary authentication server.
- The S70 and S70 Advanced SP-attached servers are directly attached to the control workstation through two RS232 serial connections. There is no dual RS232 hardware support for these connections like 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. 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.

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

Figure 16 shows an example of a single frame with a boot/install server configured on node 1.

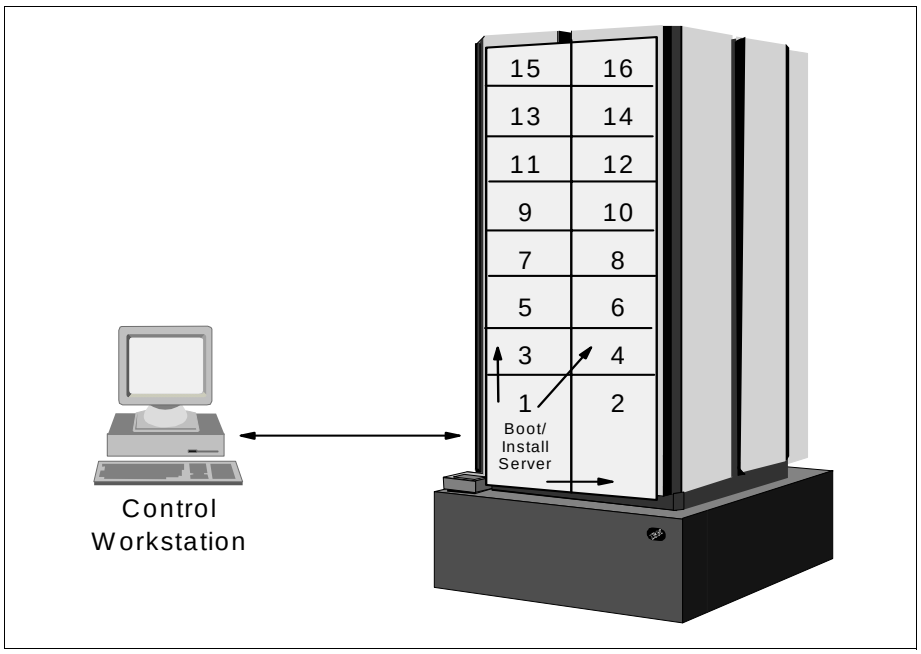


Figure 16. Boot/Install Servers

2.8 SP Switch Communication Network

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

2.8.1 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.

2.8.1.1 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 on Figure 17 on page 38

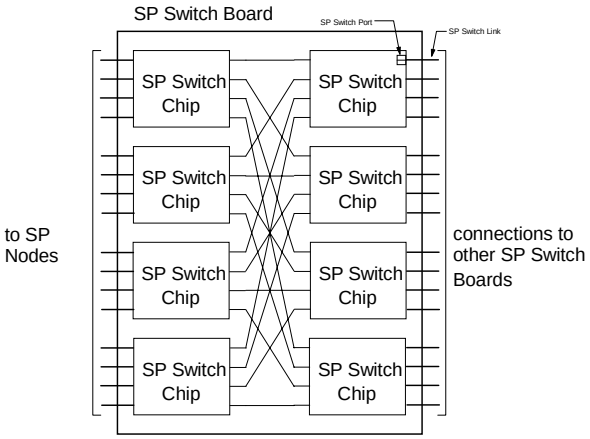


Figure 17. SP Switch Board

2.8.1.2 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.

2.8.1.3 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.

The relationship between the SP Switch Chip Link and the SP Switch Chip Port is shown in Figure 18 on page 39.

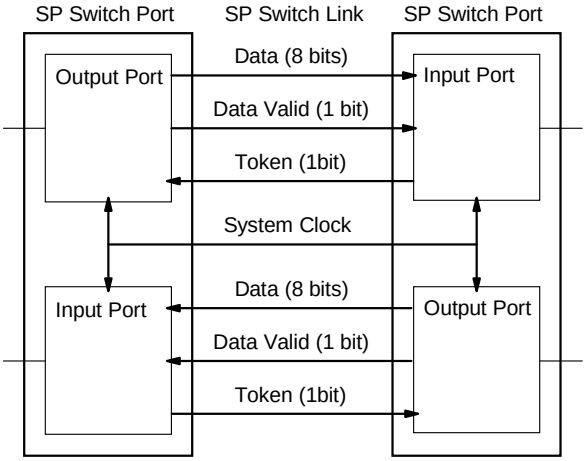


Figure 18. Relationship Between Switch Chip Link and Switch Chip Port

2.8.1.4 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 19 on page 40.

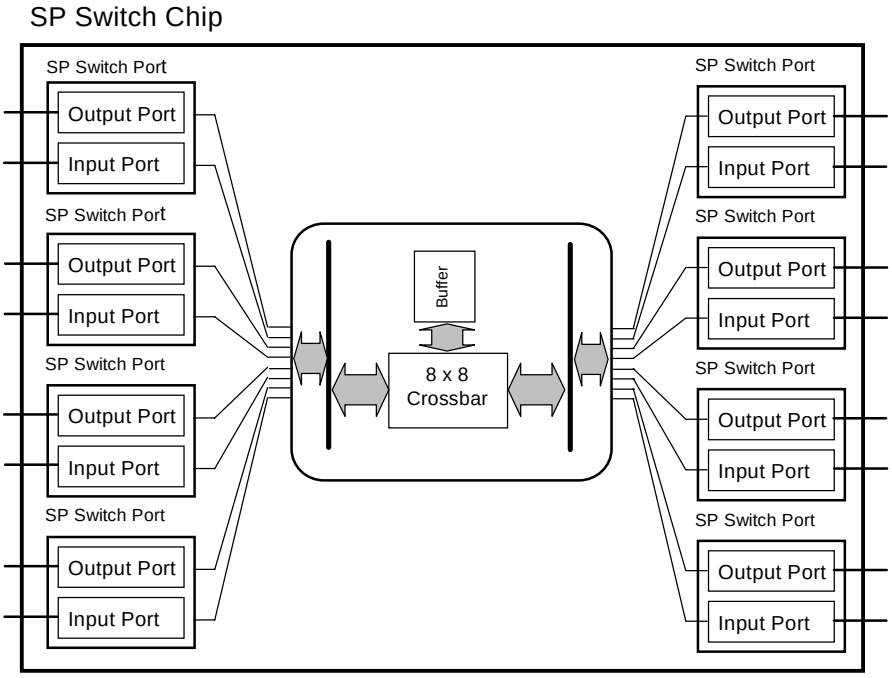


Figure 19. SP Switch Chip Diagram

2.8.1.5 SP Switch 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. The SP Switch Adapter is installed in an SP node.

Nodes based on RS/6000s that use the MCA bus obviously use the MCA-based switch adapter (#4020). The same adapter is used in uniprocessor thin, wide, and SMP high nodes.

New nodes based on PCI bus architecture (332 MHz SMP Thin and Wide Nodes, the 200 MHz POWER3 SMP Thin and Wide Nodes) must use the newer MX-based switch adapters (#4022 and #4023, respectively) since the switch adapters are installed on the MX bus in the node. The so-called mezzanine or MX bus allows the SP Switch adapter to be connected directly onto the processor bus providing faster performance than adapters installed on the I/O bus. The newer (POWER3) nodes use an improved adapter based on a faster mezzanine (MX2) bus.

External nodes, such as the 7017-S70 and 7017-S7A, 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 adapter (#8396).

Figure 20 shows a schematic diagram of the SP Switch adapter.

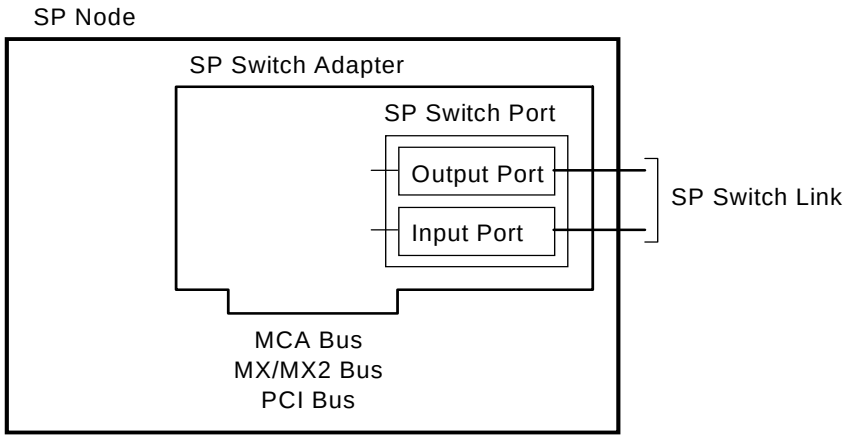


Figure 20. SP Switch Adapter

2.8.1.6 SP Switch System

The SP Switch system in a single frame of an SP and is illustrated in Figure 21 on page 42. 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.

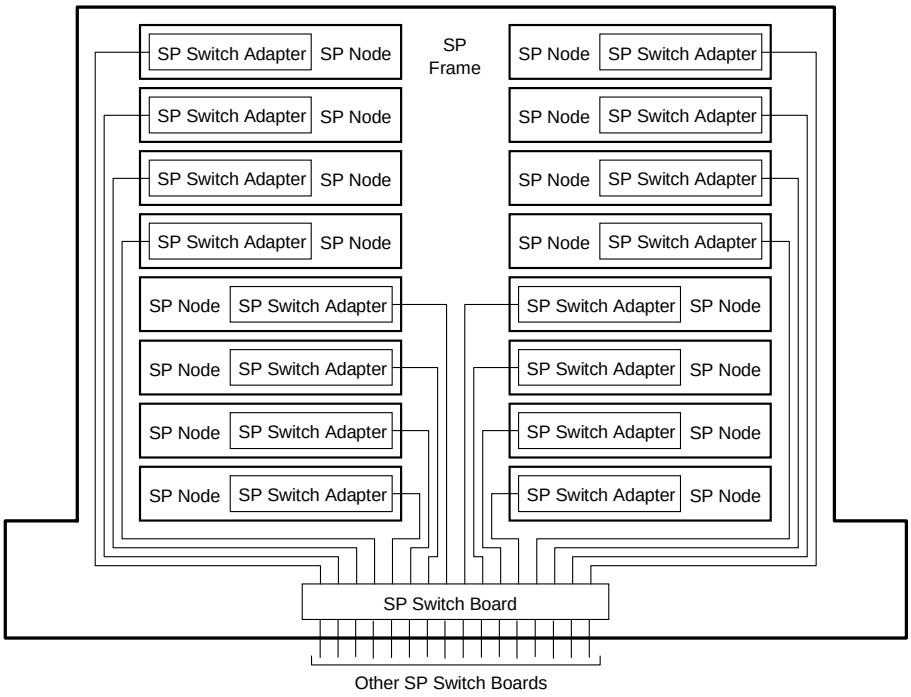


Figure 21. SP Switch System

2.8.2 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 22 on page 43. This SP Switch Board that connects nodes is called a node switch board (NSB). This figure 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.

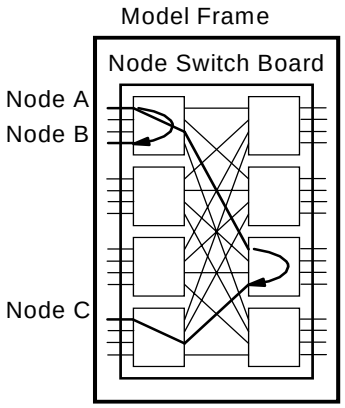


Figure 22. 16-Node SP System

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 23 on page 44. 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.

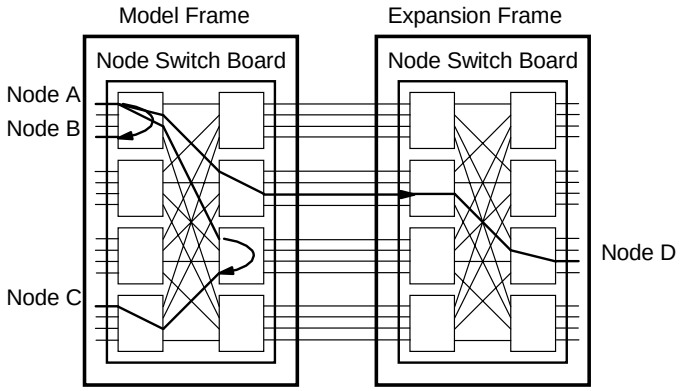


Figure 23. 32-node SP System

If we now consider an SP system made up of three frames of thin nodes (48 nodes in total, see Figure 24), 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.

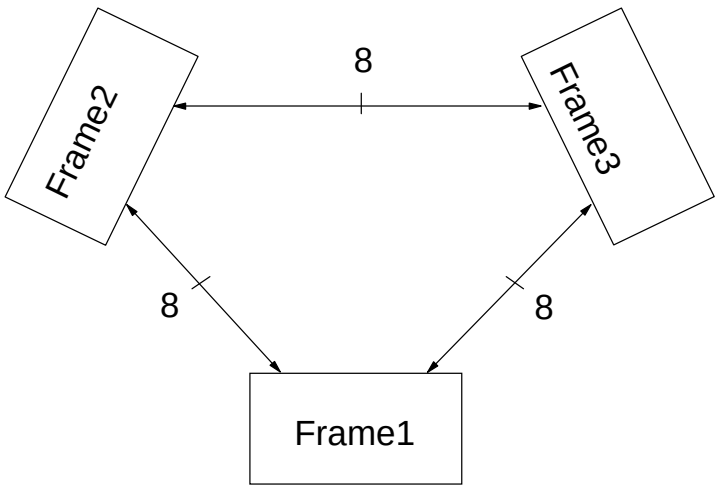


Figure 24. SP 48-Way System Interconnection

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 25. 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.

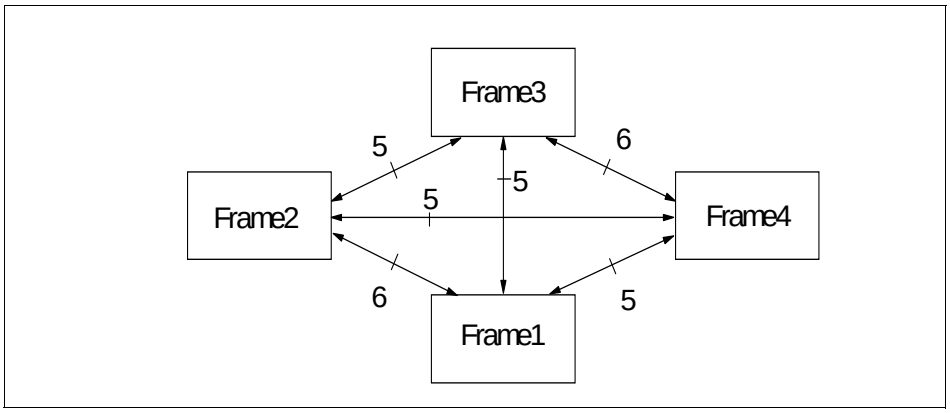


Figure 25. 64-Way System Interconnection

If we extend this 4-frame SP complex by adding another frame, the connections between frames are reduced again (see Figure 26 on page 46); 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.

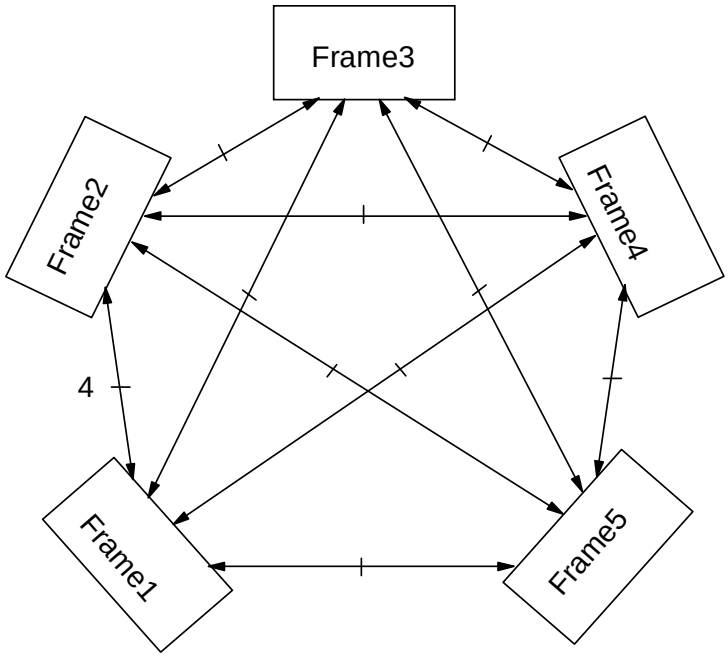


Figure 26. SP 80-Way System Interconnection

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 27 on page 47. 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, increasing network redundancy, and allowing the network bandwidth to scale linearly.

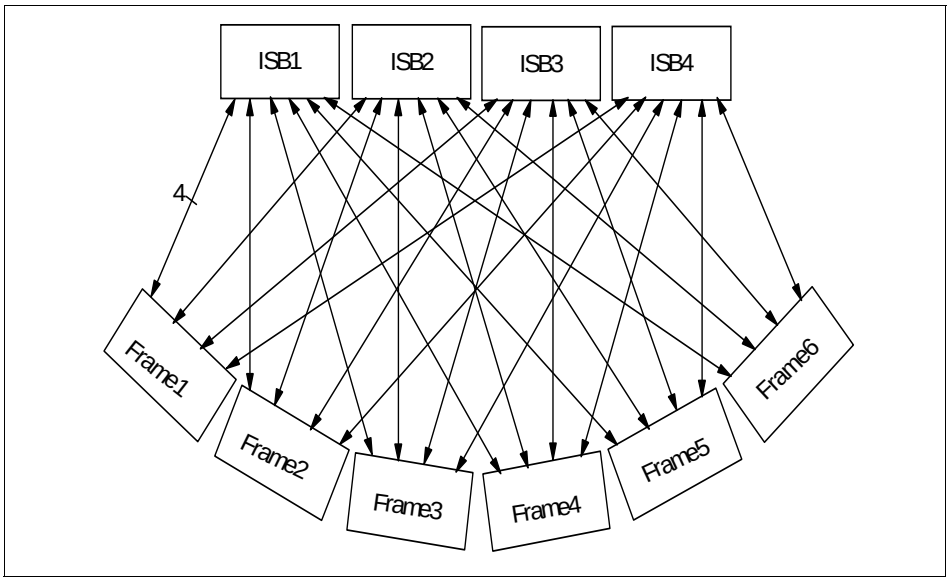


Figure 27. SP 96-way System Interconnection

2.8.3 SP Switch Network Products

Since the original RS/6000 SP product was made available in 1993, there have been two evolutionary cycles in switch technology. The original switch, known as the High Performance Switch (HiPS, feature code #4010), was last supported in Parallel System Support Programs (PSSP) Version 2.4. The latest version of PSSP software (Version 3.1) does not provide support for the HiPS switch. Switch adapters and switches (both 16-port and 8-port) based on the old HiPS technology are no longer available.

2.8.3.1 SP Switch

The operation of the SP Switch (feature code #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 multi-switch 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.

2.8.3.2 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, feature code #4008) provides switch functions for an 8-node SP system. The

SP Switch-8 is compatible with high nodes. The SP Switch-8 is the only low-cost switch available for new systems.

The 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

If a switch is configured in an SP system, an appropriate switch adapter is required to connect each RS/6000 SP node to the switch subsystem. Table 2 on page 48 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 *IBM 9077 SP Switch Router: Get Connected to the SP Switch*, SG24-5157.

Table 2. Supported Switch Adapters

SP Node Type	Supported SP Switch Adapter
160 Mhz Thin, 135 Mhz Wide, or 200 Mhz High	#4020 SP Switch Adapter
332 Mhz SMP Thin or Wide Node	#4022 SP Switch MX Adapter
200 Mhz POWER3 SMP Thin or Wide	#4023 SP Switch MX2 Adapter
External, S70 or S7A	#8396 SP System Attachment Adapter
External, SP Switch Router	#4021 SP Switch Router Adapter

The 332Mhz and 200Mhz 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 28 on page 49.

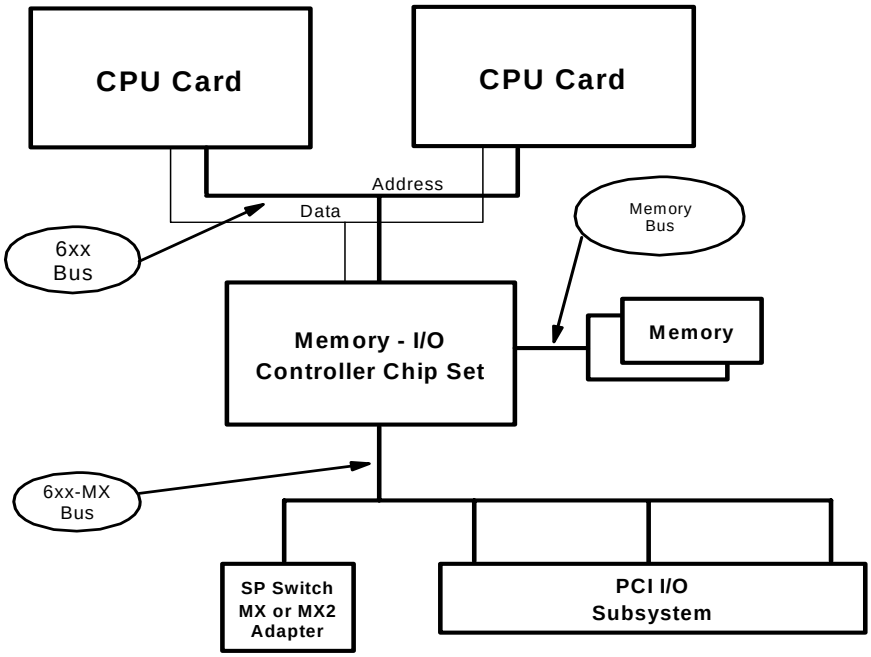


Figure 28. Internal Bus Architecture for PCI-based SMP Nodes

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.

2.9 Peripheral Devices

The attachment of peripheral devices, such as disk subsytems, tape drives, CD-ROMs, and printers, is very straightforward on the SP. There are no SP-specific peripherals; since the SP uses mainstream RS/6000 node technology, it simply inherits the array of peripheral devices available to the RS/6000 family. The SP's 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 peripheral: 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 subsytem 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.

2.10 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.

2.11 Space Requirements

Your 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 graphical user interface. In this case, the Resource Center, PTPE, IBM Virtual Shared Disk.

You can find more information on this space requirements in 7.6.2, “AIX Automounter” on page 227.

2.12 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

The version of PSSP that will run on each type of node is shown in Table 3 on page 51. 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.

Table 3. Minimum Level of PSSP and AIX That Is Allowed on Each Node

	Uni. Thin	332 SMP Thin & Wide	Uni. Wide	SMP High	SP Attached Server	POWER3 SMP Thin & Wide
Min. PSSP Level	2.2	2.4	2.2	2.2	3.1	3.1 plus IX85457
Minimum AIX Level	4.1.5	4.2.1	4.1.5	4.1.5	4.3.2	4.3.2

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 task and helps increase productivity by letting administrators view, monitor, and control system operations.

The PSSP Product ordered for the SP System (9076) but entitled for use across the entire SP Complex. PSSP V3.1 has been enhanced to allow attachment of an S70 or S70 advanced server. Here, the feature is ordered times the number of nodes.

Software Requirements for PSSP Version 2.2 on AIX 4.1:

- AIX Version 4.1.5 (5765-C34 or 5765-393) and
- C for AIX, Version 3.1 (5765-423) or later or

- C Set ++ for AIX, Version 3.1 (5765-421) or later and
- Performance Toolbox for AIX, Agent Component, Version 2.2 (5765-654)

Software Requirements for PSSP Version 2.3 on AIX 4.2:

- AIX Version 4.2 (5765-C34 or 5765-655) or later and
- C for AIX, Version 3.1 (5765-421) or later and
- C Set ++ for AIX, Version 3.1 (5765-421) or later and
- Performance Toolbox for AIX, Agent Component, Version 2.2 (5765-654)

Software Requirements for PSSP Version 2.4 on AIX 4.2.1:

- AIX Version 4.2.1, or later (5765-C34)
- C for AIX, Version 3.1.4.7 (5765-423) or later or
- C Set ++ for AIX, Version 3.1 (5765-421) or later and
- Performance Toolbox for AIX, Agent Component, Version 2.2 (5765-654)

Software Requirements for PSSP Version 2.4 on AIX 4.3.1:

- AIX Version 4.3.1 or later (5765-C34)
- C for AIX, Version 4.3 or later or
- C Set and C Set ++ for AIX, Version 3.6 or later
- Performance Toolbox for AIX, Agent component, Version 2.2 (5765-654)

Software Requirements for PSSP Version 3.1 on AIX 4.3.2, or later:

- AIX Version 4.3.2 or later (5765-C34)
- C for AIX, Version 4.3 or later or
- C Set and C Set ++ for AIX, Version 3.6 or later
- Performance Toolbox for AIX, Manager Component, Version 2.2 (5765-654)

AIX 4.3 and its Relation to PSSP:

- 32-bit or 64-bit application coexistence and concurrent execution for those who plan to implement 64-bit technology in SP system.
- An Internet- and intranet-ready operating environment.
- On-line HTML-based publications.
- Support for multiple authentication methods within the AIX remote commands.

- The Network Installation Management (NIM) component of AIX supports Distributed Computing Environment (DCE) authentication for remote commands.
- Supports PSSP 3.1 for AIX 4.3.2.

2.13 System Partitioning

In a switch 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.

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 29 on page 54.

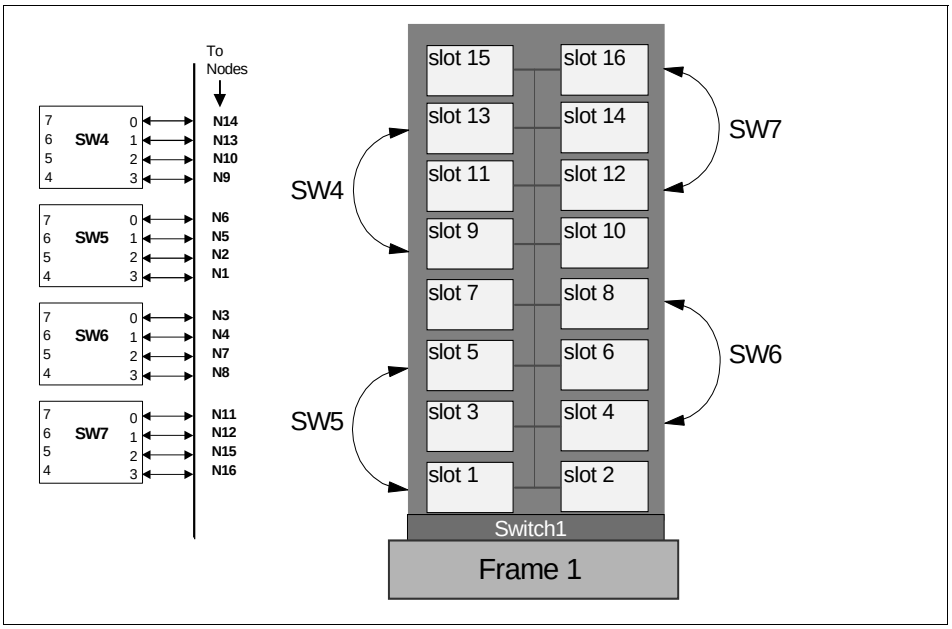


Figure 29. System Partitioning

2.14 Configuration Rules

The RS/6000 SP system has extremely wide scalability. For standard configuration, the RS/6000 SP system mounts up to 128 processor nodes. This section provides you with information on how you can expand your SP system and what kind of configuration fits your requirement. Also, in this section, we provide a set of rules and sample configurations to facilitate you to the design of more complex SP configurations. You may use these configuration rules as a check list when you configure your SP system.

This section uses the following current node, frame, switch, and switch adapter types to configure SP systems.

Nodes

- 160 MHz Thin node (feature code #2022).
- 332 MHz SMP Thin node (feature code #2050).
- 332 MHz SMP Wide node (feature code #2051).
- POWER3 SMP Thin node (feature code #2052).
- POWER3 SMP Wide node (feature code #2053).

- 200 MHz SMP High node (feature code #2009). This node is withdrawn from marketing.
- RS/6000 Server Attached node (feature code #9122).

Frames

- Short model frame (model 500)
- Tall model frame (model 550)
- Short expansion frame (feature code #1500)
- Tall expansion frame (feature code #1550)
- SP Switch frame (feature code #2031)
- RS/6000 server frame (feature code #9123)

Switches

- SP Switch-8 (8-port switch, feature code #4008)
- SP Switch (16-port switch, feature code #4011)

Switch Adapter

- SP Switch adapter (feature code #4020)
- SP Switch MX adapter (feature code #4022)
- SP Switch MX2 adapter (feature code #4023)
- SP System attachment adapter (feature code #8396)

The RS/6000 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 frame 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 of both. 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.

2.14.1 Short Frame Configurations

Short frames can be developed into two kind of configurations: Nonswitched and switched configurations. The supported switch for short frame configurations is the 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. Additional 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.

2.14.1.1 Nonswitched 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 30 on page 57.

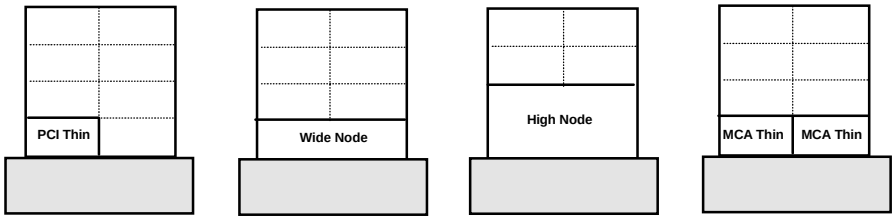


Figure 30. Minimum Nonswitched Short Frame Configurations

The short model frame must be completely full before the short expansion frame can mount nodes as shown in Figure 31

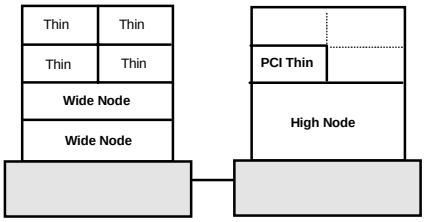


Figure 31. Example of Nonswitched Short Frame Configuration

2.14.1.2 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 the SP Switch-8 and additional nonswitched short expansion frames. Each node requires supported SP Switch adapters. Nodes in the nonswitched short expansion frames share unused switch ports in the short model frame. Figure 32 on page 58 shows the example of maximum SP Switch-8 short frame configurations.

Configuration Rule 4

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

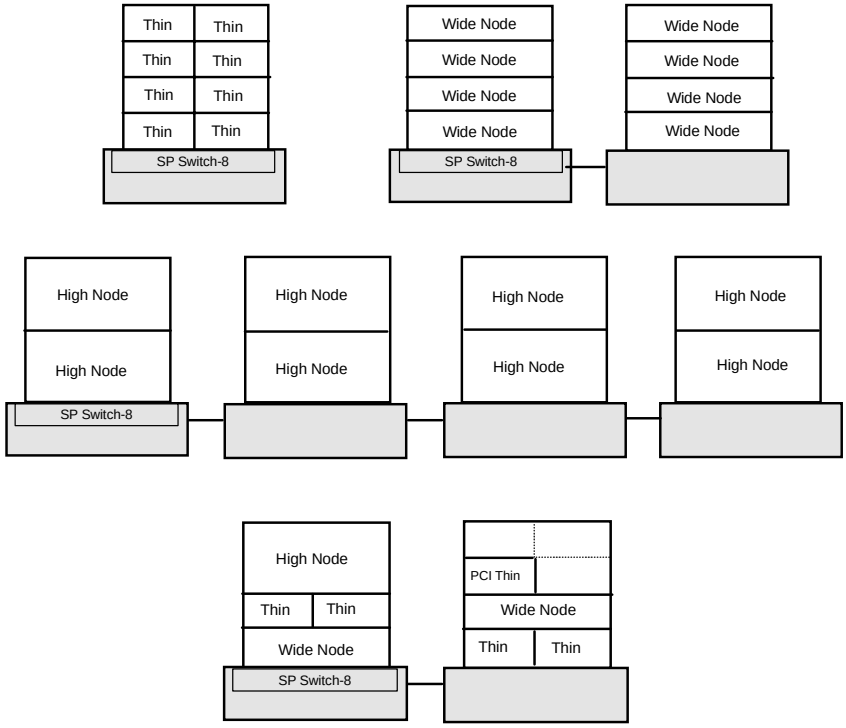


Figure 32. Maximum SP Switch-8 Short Frame Configurations

2.14.2 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. Nonswitched configuration
2. SP Switch-8 configuration
3. Single stage SP Switch configuration
4. Two stage SP Switch configuration

Configuration Rule 5

Tall frames support SP-Attached servers.

2.14.2.1 Nonswitched Tall Frame Configurations

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, or one High node, or one pair of MCA thin nodes. 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.

2.14.2.2 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 nonswitched 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 33 on page 59 shows example of SP Switch-8 tall frame configurations.

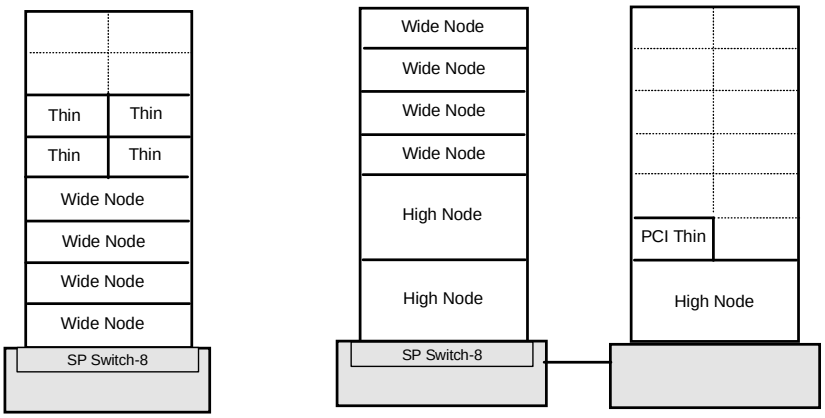


Figure 33. Example of SP Switch-8 Tall Frame Configurations

2.14.2.3 Single Stage SP Switch Configurations

This probably is 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 nonswitched 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 Configurations

If your SP system has no more than 16 nodes, a single SP Switch is enough. In this circumstance, nonswitched expansion may be added depending on the number of nodes and node locations (see 2.15.4, “The Switch Port Numbering Rule” on page 66 and Figure 39 on page 68).

Figure 34 on page 61 shows an example of single stage SP Switch configuration with no more than 16 nodes.

In configuration (a), 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 switched expansion frame has Thin nodes on the right side, it cannot support nonswitched expansion frames.

In configuration (b), six Wide nodes and two PCI Thin nodes are mounted in a tall model frame equipped with an SP Switch. There also is a High node, two Wide nodes, and four PCI Thin nodes mounted in a nonswitched 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 a 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 (c), there are eight Wide nodes mounted in a tall model frame equipped with an SP Switch and four High nodes mounted in a nonswitched expansion frame (frame 2). The second nonswitched 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 you try to describe the configuration (d). If you want to add two POWER3 Thin nodes, what would be the locations?

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

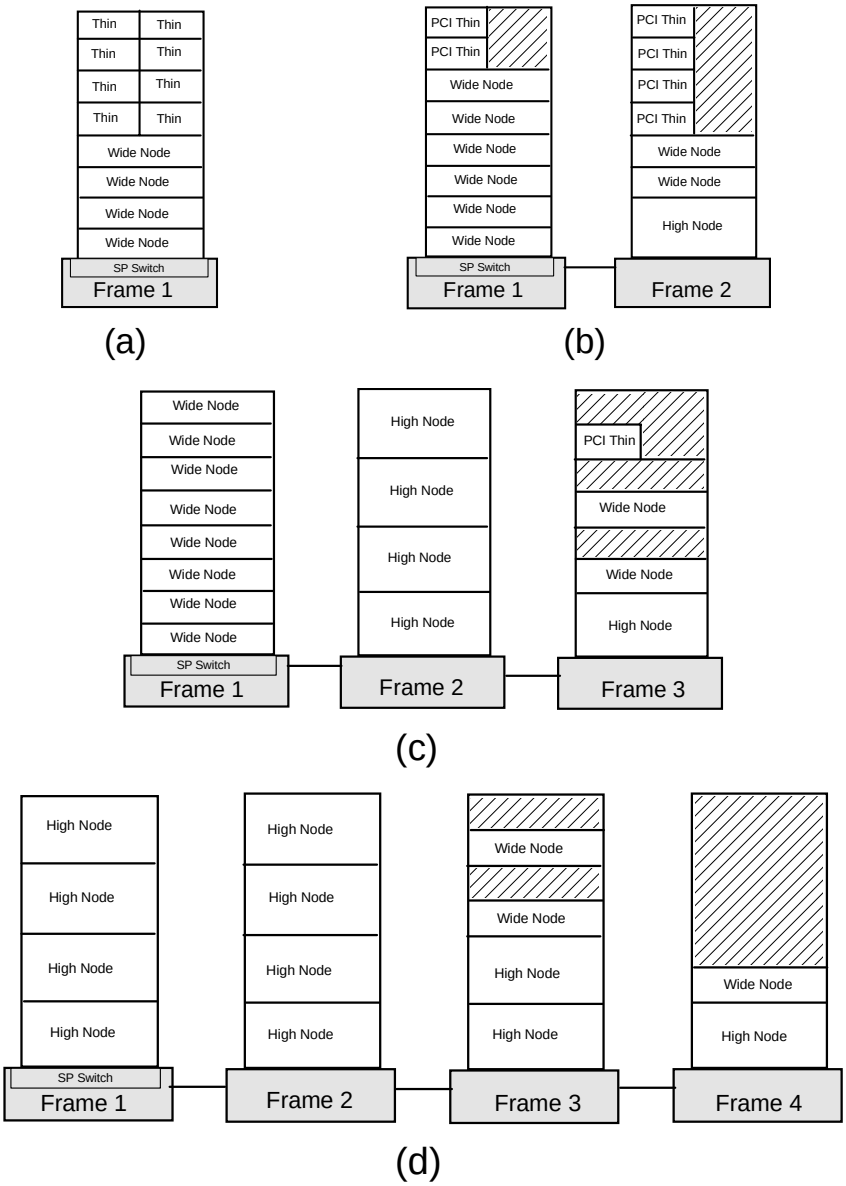


Figure 34. Example of Single SP-Switch Configurations

Single Stage with Multiple SP-Switches Configurations

If your SP system has 17 to 80 nodes, switched expansion frames are required. You can add switched expansion frames and nonswitched expansion frames. Nodes in the nonswitched expansion frame share unused switch ports that may exist in the model frame and in the switched expansion frames. Figure 35 shows an example of a Single Stage SP Switch with both switched and nonswitched 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.

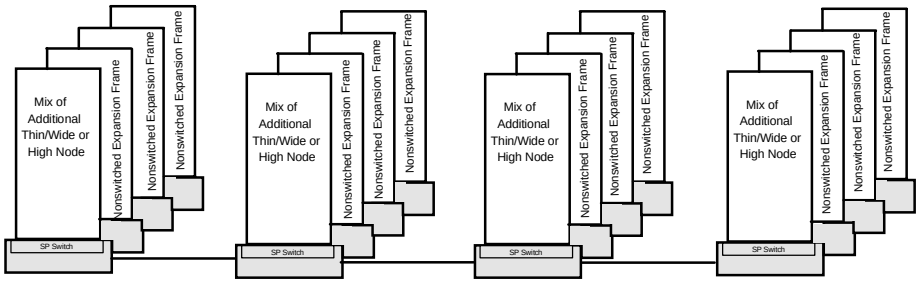


Figure 35. Example of a Multiple SP-Switches Configuration

2.14.2.4 Two Stage SP Switch Configurations

This configuration requires an SP Switch frame that forms the second switching layer. A minimum of 24 processor nodes are 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 nonswitched expansion frames as well. Nodes in the nonswitched expansion frame share unused switch ports that may exist in the model frame.

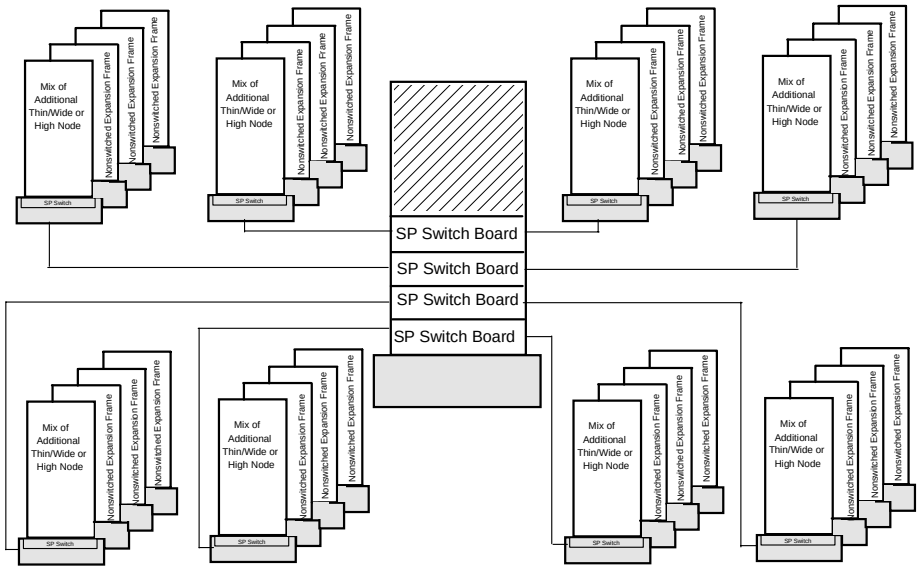


Figure 36. Example of Two Stage SP Switch Configurations

2.15 Numbering Rules

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

2.15.1 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 and nodes. This is also the order used to determine which frames share a switch.

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.

2.15.2 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. That slot 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 37 on page 65 shows slot numbering for tall frames and short frames.

An SP-Attached server is managed by the PSSP components as 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 is the only node in its frame. It occupies slot number 1 but uses the full 16 slot numbers. Therefore, 16 is added to the node number of the SP-Attached server to get the node number of the next node.
- 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 nonswitched expansion frame using that switch.

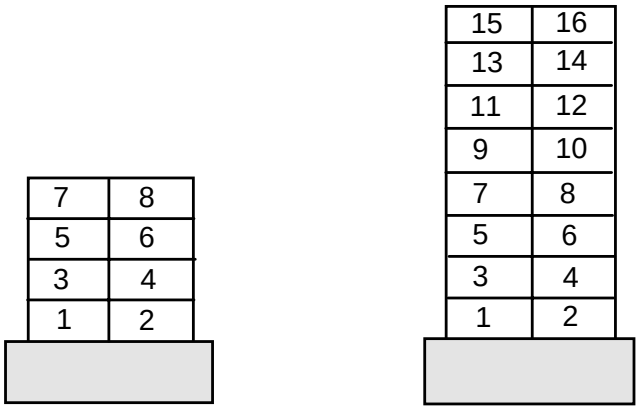


Figure 37. Slot Numbering for Short Frames and Tall Frames

2.15.3 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 by the following formula:

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

where *slot_number* is the lowest slot number occupied by the node. Each type (size) of node occupies a consecutive sequence of slots. For each node, there is an integer n such that a thin node occupies slot *n*, a wide node occupies slots *n, n+1*, and a high node occupies *n*, *n+1*, *n+2*, *n+3*. For wide and high nodes, n must be odd.

Node numbers are assigned independent of whether the frame is fully populated. Figure 38 on page 66 demonstrates node numbering. Frame 4 represents an SP-Attached server in a position where it does not interrupt the switched frame and companion nonswitched 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.

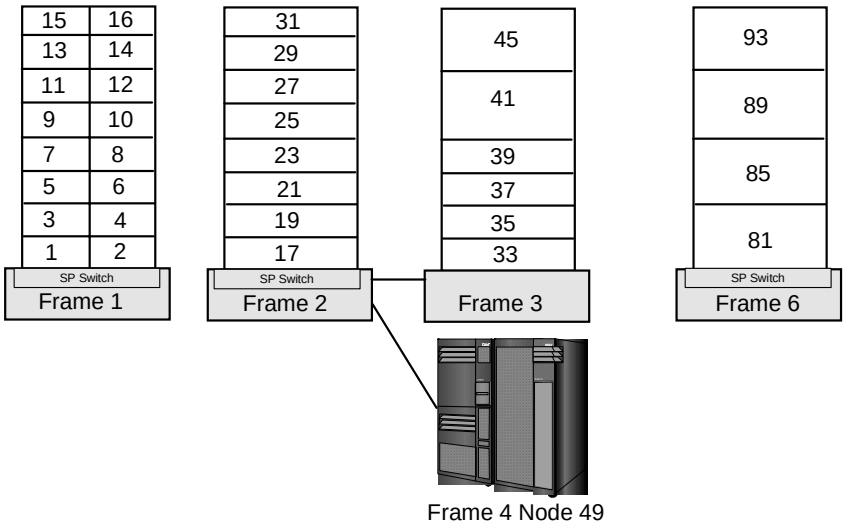


Figure 38. Node Numbering for an SP System

2.15.4 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 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 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 39 on page 68 shows the frame and switch configurations that are supported and the switch port number assignments in each node. Let us describe more details on 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 nonswitched 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 nonswitched expansion frames. These expansion frames are allowed to 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 nonswitched expansion frames. Each of these nonswitched 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 is allowed to be mount in a drawer. Therefore, it is allowed to mount in nonswitched 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).

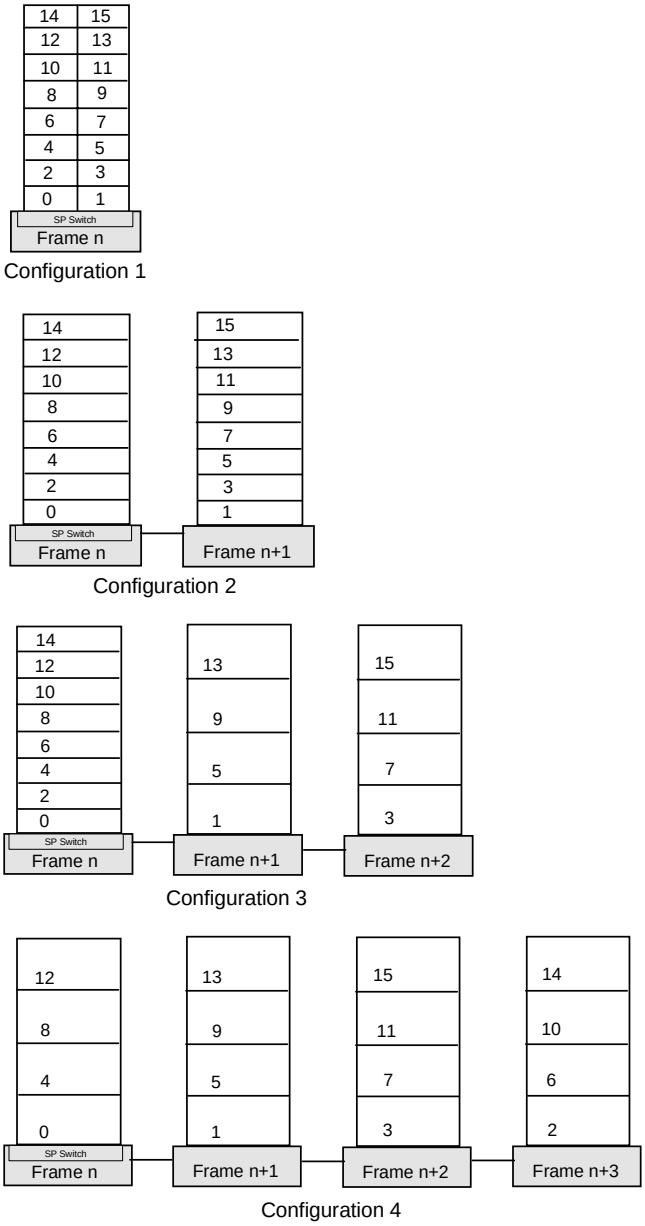


Figure 39. Switch Port Numbering for an SP Switch

Switch Port Numbering for an SP Switch-8

An SP system with SP switch-8 contains only switch port numbers zero through seven. 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 40 shows sample switch port numbers for a system with a short frame and an SP Switch-8.

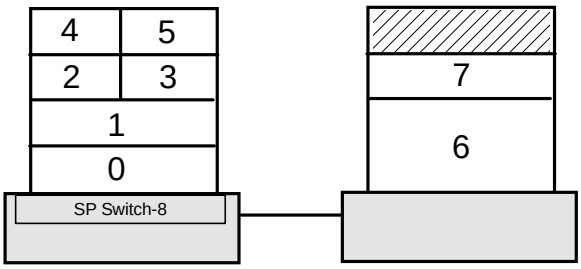


Figure 40. Example of Switch Port Numbering for an SP Switch-8

2.16 Related Documentation

These documents will 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 book *IBM 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 ensures that you have the correct 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. The redbook 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.

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 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 Switch over a High Performance Switch?

   A. Fault isolation

   B. Compatible with older HiPS Switches

   C. Improved bandwidth

   D. Higher availability

3. Which is a minimum prerequisite for PSSP Version 3 release 1?

   A. AIX Version 4.3.2

   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.

Chapter 3. RS/6000 SP Networking

This chapter covers the networking issues on a RS/6000 SP system. It discusses the different name resolution mechanisms you have available on the SP as well as Ethernet segmentation and routing. Network topology and the impact on the RS/6000 SP subsystems are also discussed.

3.1 Key Concepts You Should Study

The concepts explained in this section will give you a good preparation for the networking related questions in RS/6000 SP certification exam. In order to maximize your chances, you should become familiar with:

- How to create specific hostnames, TCP/IP address, Netmask value, and default routes.
- How to determine the name resolution mechanism, such as host table, DNS, or NIS, that better fit your needs.
- How to determine the Ethernet topology, segmentation, and routing in the SP System.

3.2 Name, Address, and Network Integration Planning

You must assign IP addresses and host names for each network connection on each node and on the control workstation in your SP system. Because you probably want to attach the SP system to your site networks, you need to plan how to do this. You need to decide what routers and gateways you will use, what default and network routes you need on your nodes, and how you will establish these default and network routes.

You need to ensure that all of the addresses you assign are unique within your site network and within any outside networks to which you are attached, such as the Internet. Also, you need to plan how names and addresses will be resolved on your systems (that is, using DNS name servers, NIS maps, /etc/hosts files, or some other method).

3.2.1 Set Host name

Independent of any of the network adapters, each machine has a hostname. Usually the hostname is the name given to one of the network adapters in the machine. We need to set the hostname on the control workstation.

A sample of `smit hostname` is shown in Figure 41 on page 74.

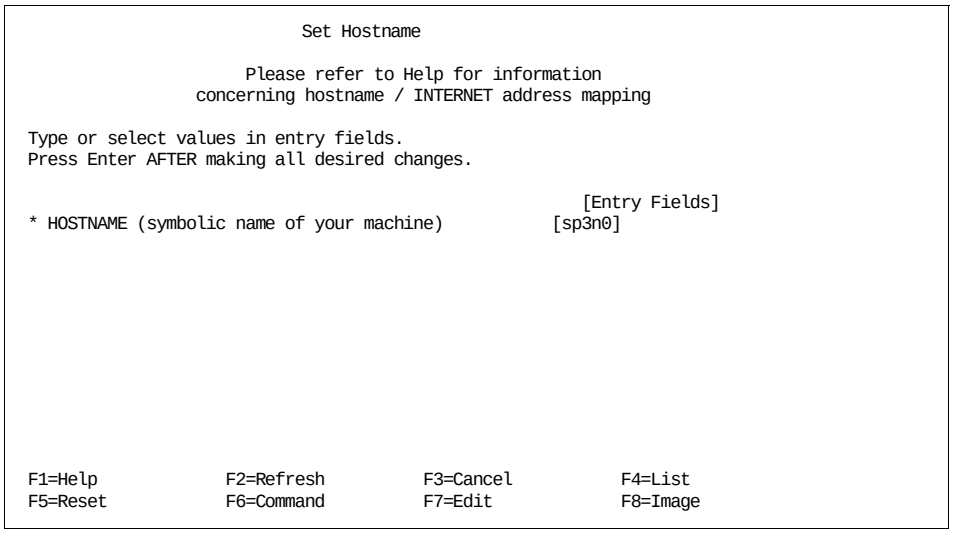


Figure 41. Set the Hostname on the Control Workstation

To set the hostname in control workstation, issue the `smit` fast path:

```
smit hostname
```

In SP systems, this is known as the Initial Hostname.

3.2.2 Set IP Address and Netmask

You will need at least one Ethernet subnet for your system. You will need an IP address per node and control workstation.

Each network adapter needs to have a specific IP address. To set an IP address to an adapter on the control workstation, enter the `smit` fastpath:

```
smit mktcpip
```

You select the network adapter you want to configure and fill in the IP address and netmask assigned for this adapter. Please be sure that you have the correct combination of IP address and netmask. The netmask can be defined based on the IP address class. A sample of `smit mktcpip` is shown in Figure 42 on page 75.

```
                        Minimum Configuration & Startup

 To Delete existing configuration data, please use Further Configuration menus

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

                                                      [Entry Fields]
* HOSTNAME                                           [sp3n0]
* Internet ADDRESS (dotted decimal)                  [192.168.3.130]
  Network MASK (dotted decimal)                      [255.255.255.0]
* Network INTERFACE                                   en0
  NAMESERVER
           Internet ADDRESS (dotted decimal)         []
           DOMAIN Name                               []
  Default GATEWAY Address                            [9.12.0.1]
  (dotted decimal or symbolic name)
  Your CABLE Type                                     bnc                     +
  START Now                                           no                      +


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 42. Set IP Address and Netmask on the Control Workstation

The en0 adapter (first Ethernet adapter) on the nodes needs to be configured with an IP address and a name. This name is know as the Reliable Hostname. The control workstation and several subsystems (such as Kerberos) will use this Reliable Hostname and the en0 adapter for communication.

3.2.3 Set Routes

If you have different subnet in your network, it is very important that you give a specific route from your SP system to this subnet. By defining a route, you basically show this node's adapter and 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.

Routing is very important in RS/6000 SP environments. PSSP supports multiple subnets, but all the nodes need to be able to access those subnets if nodes in the same partition reside there. Every node must have access to the control workstation even when it is being installed from a boot/install server other than the control workstation.

Before configuring boot/install servers for other subnets, make sure the control workstation has routes defined to reach each one 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><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
```

A sample of `smit mkroute` is shown in Figure 43.

```
                    Add Static Route

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

                                                    [Entry Fields]
  Destination TYPE                                   net                    +
* DESTINATION Address                               [192.168.15.0]
  (dotted decimal or symbolic name)
* Default GATEWAY Address                           [9.12.0.130]
  (dotted decimal or symbolic name)
* METRIC (number of hops to destination gateway)    [1]                     #
  Network MASK (hexadecimal or dotted decimal)      [255.255.255.0]









F1=Help            F2=Refresh         F3=Cancel          F4=List
F5=Reset           F6=Command         F7=Edit            F8=Image
F9=Shell           F10=Exit           Enter=Do
```

Figure 43. Adding a Route Using SMIT mkroute

3.2.4 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 to be not 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 sample, 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`.

3.2.5 NIS

NIS' main purpose is to centralize administration of files, such as /etc/passwd, within a network environment.

NIS separates a network into three components: Domain, server(s), and clients.

A 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 not all the machines on the network.

A 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:

1. To balance the load if the master server is busy.
2. To back up the master server.
3. 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 goes through a router.

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

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` 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. Please refer to the "NIS and NFS" publication 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 44 on page 80 shows what this SMIT panel looks like. In this example, SPDomain has been chosen as the NIS domain name.

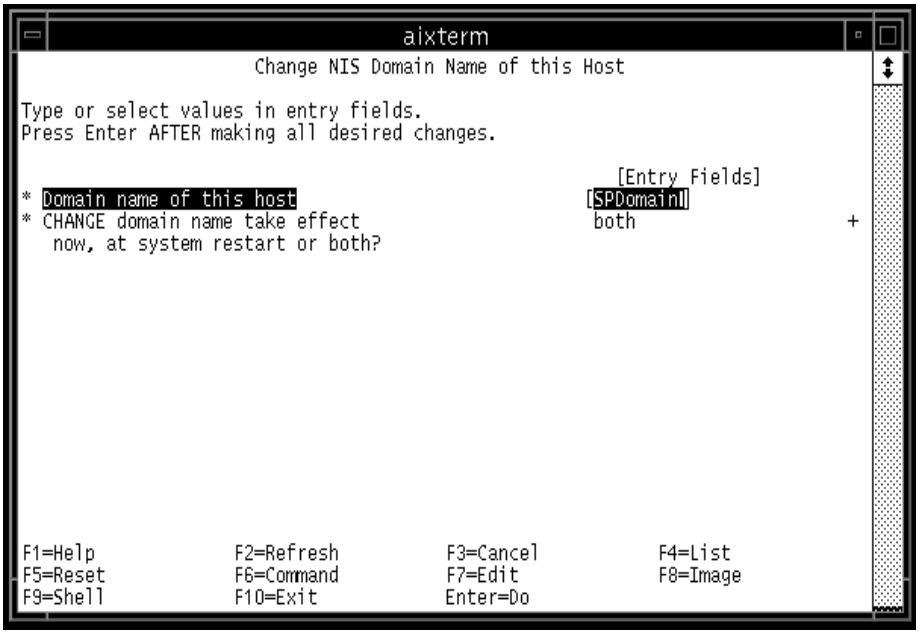


Figure 44. SMIT Panel for Setting a NIS Domain Name

- On the machine that is to be the NIS master (for example, the control workstation), select **Configure/Modify NIS** and then **Configure this Host as a NIS Master Server**. Figure 45 on page 81 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 file.

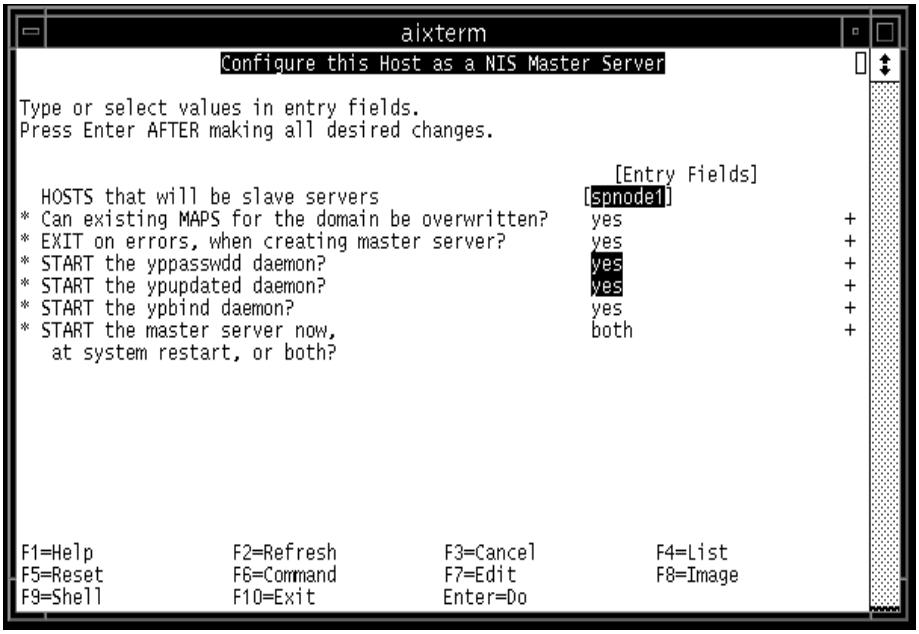


Figure 45. SMIT Panel for Configuring a Master Server

- 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 46 on page 82 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 file(s).

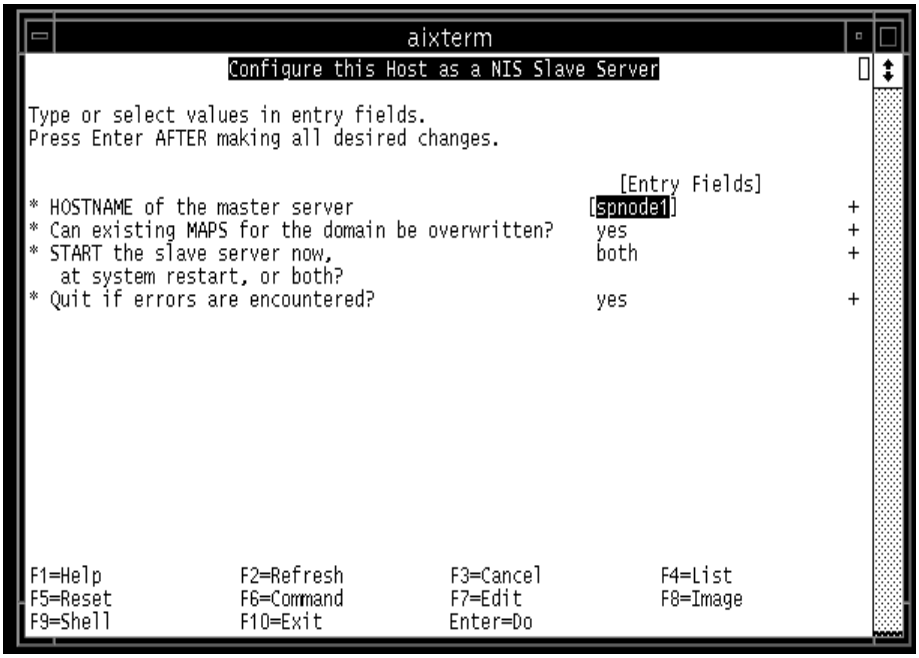


Figure 46. SMIT Panel for Configuring a Slave Server

- 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 file(s).

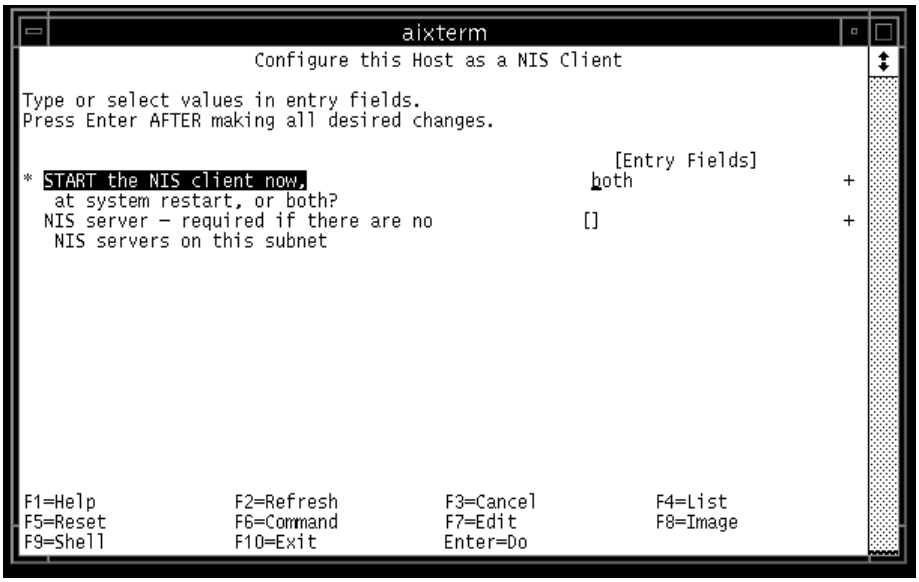


Figure 47. SMIT Panel for Configuring a NIS Client

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 command `make`. To access the SMIT panel, select **Manage NIS Maps** within the NIS panel. Figure 48 on page 84 shows this SMIT panel.

Figure 48. 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 option `all` or specify the maps that you want to rebuild. After that, return back to the SMIT panel in Figure 48 on page 84 and either **Transfer Maps to Slave Servers** (from the master server) or **Retrieve Maps from Master Server** to this Slave (from a slave server).

3.2.6 DNS

DNS (Domain Name Server) 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 as and convert it to 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 SP system has 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 similar way 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 as "/" is 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 subdomain of the com domain. The domain itso.ibm.com is subdomain of the ibm.com domain, and the .com domain.

You can set up your SP 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 copy of the /etc/hosts file, this becomes difficult to maintain for even 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 an SP 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.

3.3 The SP Networks

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

3.3.1 SP Ethernet

The SP requires an Ethernet connection between the control workstation and all nodes, which is used for network installation of the nodes and for system management. This section describes the setup of that administrative Ethernet, which is often called the SP LAN.

3.3.1.1 Frame and Node Cabling

SP frames include coaxial Ethernet cabling for the SP LAN also known as *thin-wire* Ethernet or 10BASE-2. All nodes in a 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, Multiple Access/Collision Detect (CSMA/CD). 10BASE-2 only supports half duplex (HDX). There is a hard limit of 30 stations on a single 10BASE-2 segment, and the total cable length must not exceed 185 meters. However, it is not advisable to connect more than 16 to 24 nodes to a single segment. Normally, there is one segment per frame, and one end of the coaxial cable is terminated in the frame. Depending on the network topology, the other end connects the frame to either the control workstation or to a boot/install server in that segment and is terminated there. In the latter case, the boot/install server and CWS are connected through an additional Ethernet segment; so, the boot/install server needs two Ethernet adapters.

It is also possible to use customer-provided Unshielded Twisted Pair (UTP) cabling of category 3, 4, or 5. An UTP cable can be directly connected to the RJ-45 Twisted Pair (TP) connector of the Ethernet adapter if one is available or through a transceiver/media converter to either the AUI or BNC connector. Twisted Pair connections are always point-to-point connections. So, all nodes have to be connected to a customer-provided repeater or Ethernet switch, which is normally located outside the SP frame and is also connected to the control workstation. Consequently, using UTP involves much more cabling. On the other hand, fault isolation will be much easier with UTP than with thin-wire Ethernet, and there are more opportunities for performance improvements. Twisted Pair connections at 10 Mbps are called 10BASE-T, those operating at 100 Mbps are called 100BASE-TX.

In order to use Twisted Pair in full duplex mode, there must be a native RJ-45 TP connector at the node (no transceiver), and an Ethernet switch, like the IBM 8274, must be used. A repeater always works in half duplex mode and will send all IP packets to all ports (such as in the 10BASE-2 LAN environment). We, therefore, recommend to always use an Ethernet switch with native UTP connections.

The POWER3 SMP nodes (made available in 1999) have an integrated 10/100 Mbps Ethernet adapter. They still may be connected and installed at 10 Mbps using 10BASE-T or 10BASE-2 and a transceiver. However, to fully utilize the adapter at 100 Mbps, category 5 UTP wiring to a 100 Mbps repeater or Ethernet switch is required (100BASE-TX). As mentioned above, we recommend the use of an Ethernet switch since this allows to utilize the

full duplex mode and avoids collisions. The control workstation also needs a fast connection to this Ethernet switch.

3.3.1.2 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. We strongly recommend to provide additional network connectivity (through the SP Switch or additional Ethernet, Token Ring, FDDI, or ATM networks) if the applications on the SP perform significant communication among nodes. 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 these additional networks.

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

Shared 10BASE-2 Network

In relatively small SP configurations, like single frame systems, the control workstation and nodes typically share a single thin-wire Ethernet. Figure 49 shows this setup.

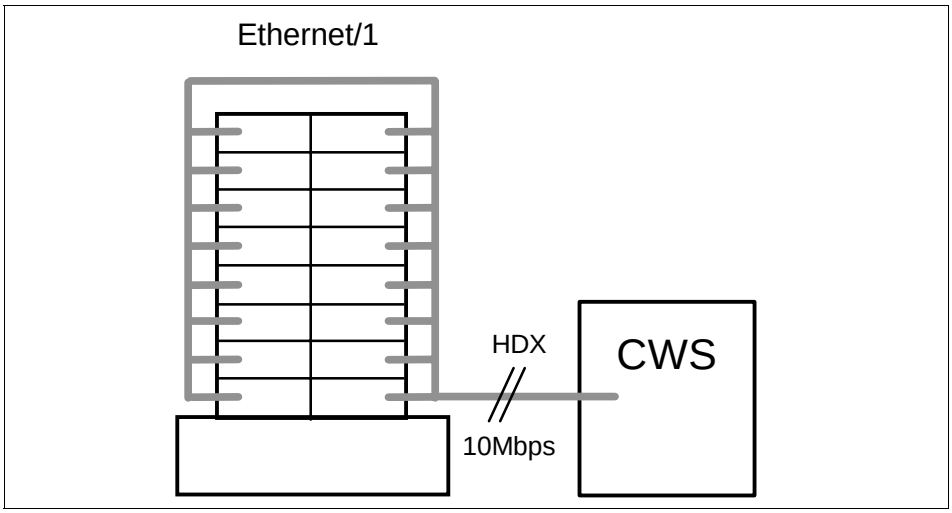


Figure 49. Shared 10BASE-2 SP Network

This configuration is characterized by the following properties:

- No routing is required since the CWS and all nodes share one subnet.

- Consequently, the whole SP LAN is a single *broadcast domain* as well as a single *collision domain*.
- The CWS acts as boot/install server for all nodes.
- Performance is limited to one 10 Mbps HDX connection at a time.
- Only six to eight network installs of SP nodes from the CWS NIM server can be performed simultaneously.

Even if this performance limitation is accepted, 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.

Segmented 10BASE-2 Network

A widely used approach to overcome the limitations of a single shared Ethernet is segmentation. The control workstation is equipped with additional Ethernet adapters, and each one is connected to a different shared 10BASE-2 Ethernet subnet. This is shown in Figure 50.

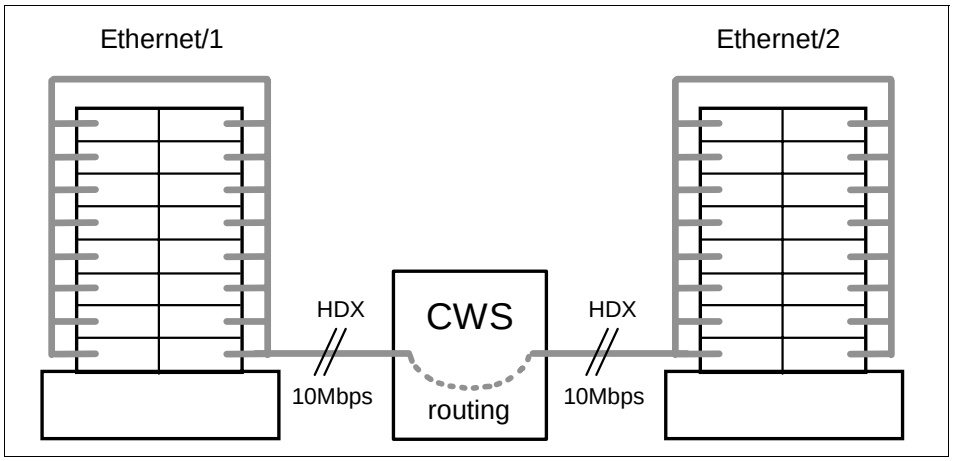


Figure 50. Segmented 10BASE-2 SP Network with Two Subnets

For a configuration with *N* separate subnets (and consequently *N* Ethernet cards in the CWS), the following holds:

- Nodes in one subnet need static routes to the *(N-1)* other subnets through the CWS, and routing (or IP forwarding) must be enabled on the CWS.
- The SP LAN is split into *N* broadcast domains.

- The CWS acts as boot/install server for all nodes since it is a member of all *N* subnets.
- Aggregate performance is limited to a maximum of *N* times 10 Mbps HDX. However, this is only achievable if the CWS communicates with one node in each of the subnets simultaneously.
- Only six to eight network installs per subnet should be performed simultaneously increasing the maximum to *6N to 8N* simultaneous installs.

This approach is limited primarily by the number of available adapter slots in the control workstation but also by the ability of the CWS to simultaneously handle the traffic among these subnets or to serve *6N to 8N* simultaneous network installations. In practice, more than four subnets should not be used.

Segmented 10BASE-2 Networks with Boot/Install Servers

For very large systems, where the above model of segmentation would require more 10 Mbps Ethernet adapters in the control workstation than possible, a more complex network setup can be deployed that uses additional boot/install servers. This is shown in Figure 51 on page 90. The CWS is directly connected to only one Ethernet segment, which is attached to the 10 Mbps Ethernet Adapter en0 of a set of *N* boot/install server (BIS) nodes typically the first node in each frame. We call this Ethernet subnet the Install Ethernet since it is the network through which the CWS installs the boot/install server nodes. The remaining nodes are grouped into *N* additional Ethernet segments (typically one per frame), which are not directly connected to the CWS. Instead, each of these subnets is connected to one of the boot/install servers through a second 10 Mbps Ethernet adapter in the boot/install servers.

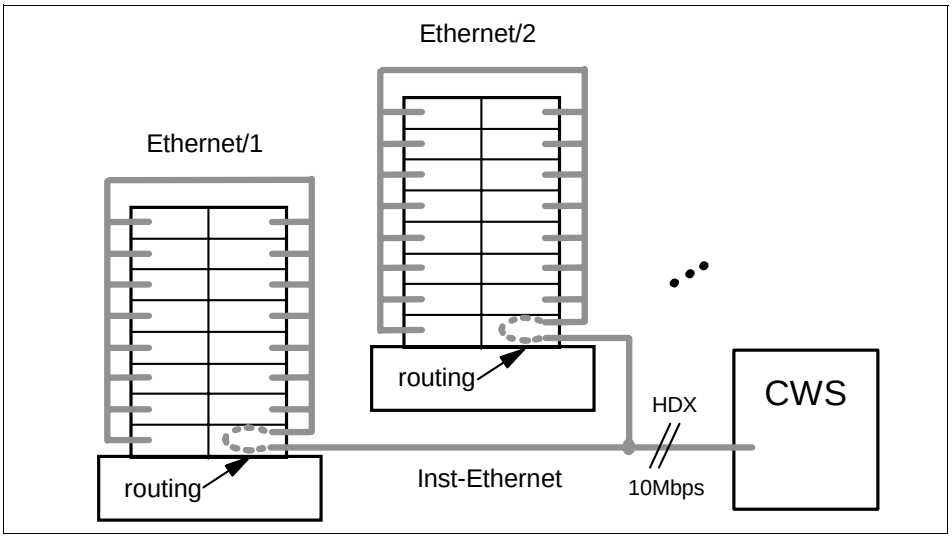


Figure 51. Segmented SP Network with Boot/Install Server Hierarchy

With such a network configuration:

- Routing is complicated:
  - Non-BIS nodes in a segment have routes to all other segments through their BIS node.
  - BIS nodes have routes to the *(N-1)* other nodes' segments through the BIS nodes attached to that segment.
  - The CWS has routes to the *N* nodes' segments through the BIS nodes in these segments.
- The SP LAN is split into *(N+1)* broadcast domains.
- The boot/install servers are installed from the NIM server on the CWS. After this, all non-BIS nodes are installed by the boot/install servers. Note that some NIM resources, such as the LPPSOURCE, are only served by the CWS.
- The maximum bandwidth in the Install Ethernet segment (including the CWS) is 10 Mbps HDX.
- Only six to eight BIS nodes can be installed simultaneously from the CWS in a first installation phase. In a second phase, each BIS node can install six to eight nodes in its segment simultaneously.

Apart from the complex setup, this configuration suffers from several problems. Communication between regular nodes in different subnets requires routing through two boot/install server nodes. All this traffic, and all communications with the CWS (routed through one BIS node), have to compete for bandwidth on the single shared 10 Mbps half duplex Install Ethernet.

The situation can be improved by adding a dedicated router. Connecting all the nodes' segments to this router removes the routing traffic from the BIS nodes, and using a fast uplink connection to the CWS provides an alternative, high bandwidth path to the CWS. The BIS nodes in each segment are still required because the network installation process requires that the NIM server and the client are in the same broadcast domain. Figure 52 shows such a configuration. Nodes in the frames now have a route to the control workstation and the other frames' networks through the router, which off-loads network traffic from the BIS nodes.

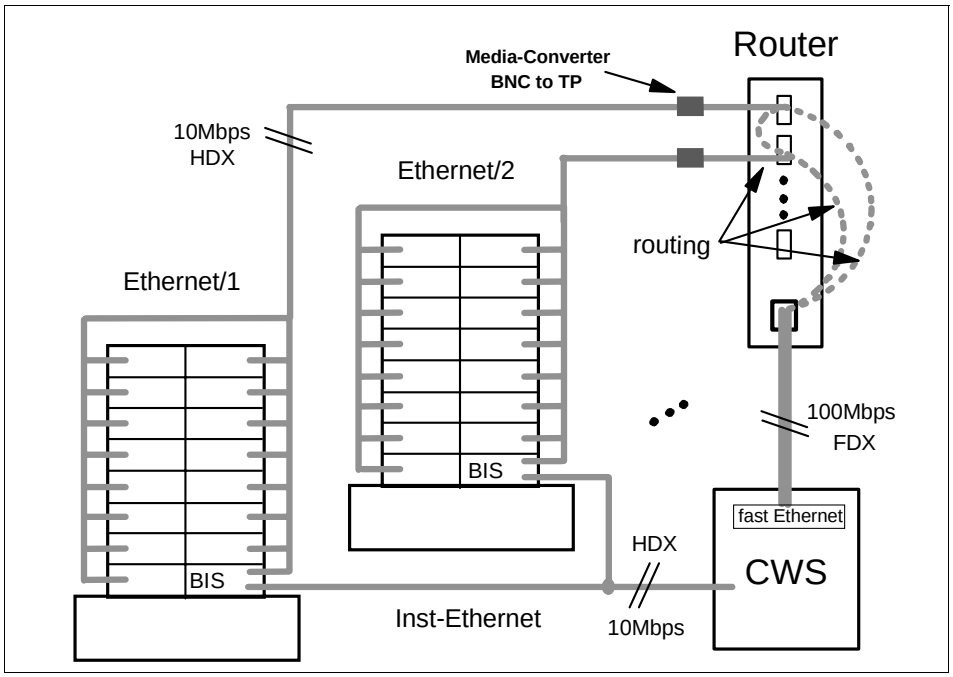


Figure 52. Boot/Install Server Hierarchy with Additional Router

Even when a router is added, the solution presented in the following section is normally preferable to a segmented network with boot/install servers both from a performance and from a management/complexity viewpoint.

Switched 10BASE-2 Network

An emerging technology to overcome performance limitations in shared or segmented Ethernet networks is Ethernet Switching, which is sometimes called micro-segmentation. An SP example is shown in Figure 53.

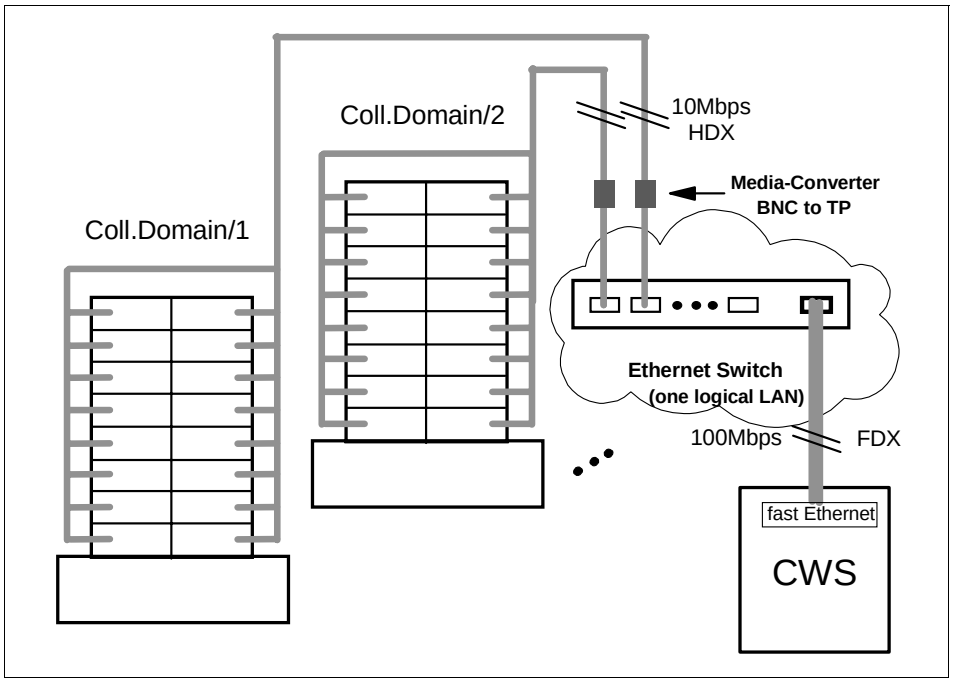


Figure 53. Switched 10BASE-2 SP Network with Fast Uplink

This configuration has the following properties:

- No routing is required. All Ethernet segments are transparently combined to one big LAN by the Ethernet switch.
- Of course, node-to-node connections within a single Ethernet segment still have to share that 10-BASE-2 medium in half duplex mode. But many communications between different ports can be switched simultaneously by the Ethernet switch. The uplink to the control workstation can be operated in a 100 Mbps full duplex mode.
- The control workstation can act as the boot/install server for all nodes since the Ethernet switch combines the CWS and nodes into one big network (or broadcast domain).

This setup eliminates the routing overhead for communications between nodes or a node and the control workstation. With a 100 Mbps, full duplex

Ethernet uplink to the CWS, there should also be no bottleneck in the connection to the CWS, at least if the number of 10BASE-2 segments is not much larger than ten.

Considering only the network topology, the control workstation 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 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 SP system, it may, therefore, 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 SP system, we advocate 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 control workstation.
3. Install the non-BIS nodes from their respective BIS nodes. This provides the desired scalability for the installation of a whole, large SP system.
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 whole SP system) 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 control workstation. 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 53 on page 92 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 use broadcast messages when the group leader sends PROCLAIM messages to attract new members.

Note: ARP cache tuning

Be aware that for SP systems with very large networks (and/or routes to many external networks), the default AIX settings for the ARP cache size might not be adequate. The Address Resolution Protocol (ARP) is used to translate IP addresses to Media Access Control (MAC) addresses and vice versa. Insufficient APR 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 control workstation. This can be accomplished either through multiple uplinks between one Ethernet switch and the CWS or through multiple switches that each have a single uplink to the control workstation.

Shared or Switched 100BASE-TX Network

With the introduction of the POWER3 SMP nodes in 1999, it has become possible to operate nodes on the SP LAN at 100 Mbps including network installation. This requires UTP cabling as outlined in 3.3.1.1, "Frame and Node Cabling" on page 86.

One possible configuration would be to use a repeater capable of sustaining 100 Mbps and a fast Ethernet adapter in the control workstation. This would boost the available bandwidth up to 100 Mbps, but it would be shared among all stations, and connections are only half duplex. Although the bandwidth would be higher by a factor of ten compared to a 10BASE-2 SP Ethernet, we recommend to use an Ethernet switch that supports full duplex connections at 100 Mbps instead of a repeater. Many node-to-node and node-to-CWS connections can be processed by the Ethernet switch simultaneously rather than the shared access through a repeater. This configuration is shown in Figure 54 on page 95. 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 control workstation not the network bandwidth.

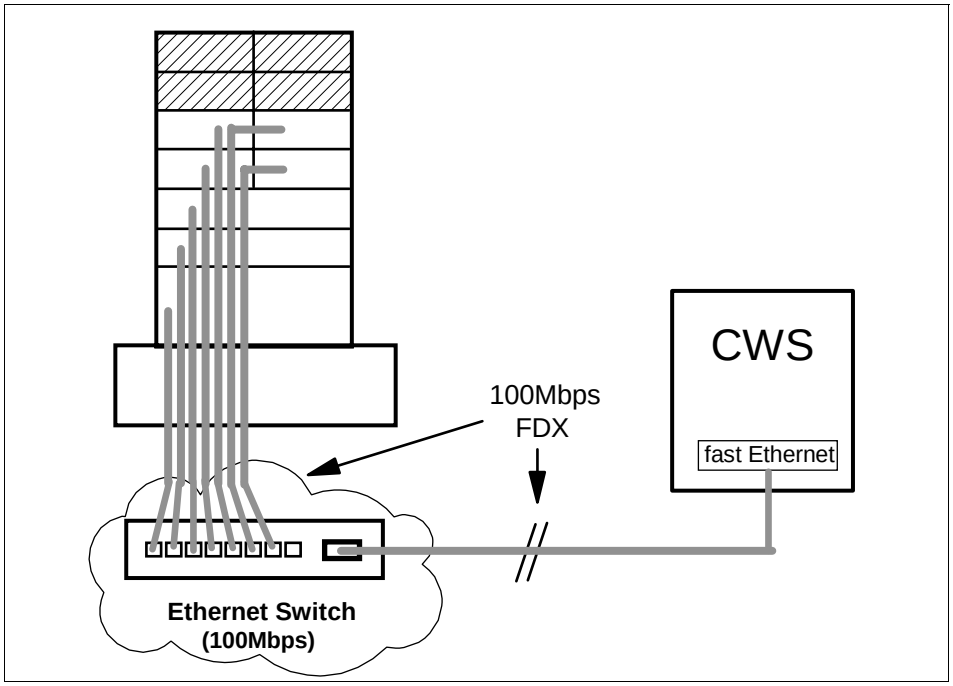


Figure 54. Simple 100BASE-TX SP Network

For larger SP configurations, the cabling required to establish point-to-point connections from all nodes to the Ethernet Switch can be impressive. An IBM 8274 Nways LAN RouteSwitch could be used to provide the required switching capacities. Models with 3, 5, or 9 switching modules are available.

Heterogeneous 10/100 Mbps Network

In many cases, an existing SP system will be upgraded by new nodes that have fast Ethernet connections, but older or less lightly loaded 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 55 on page 96.

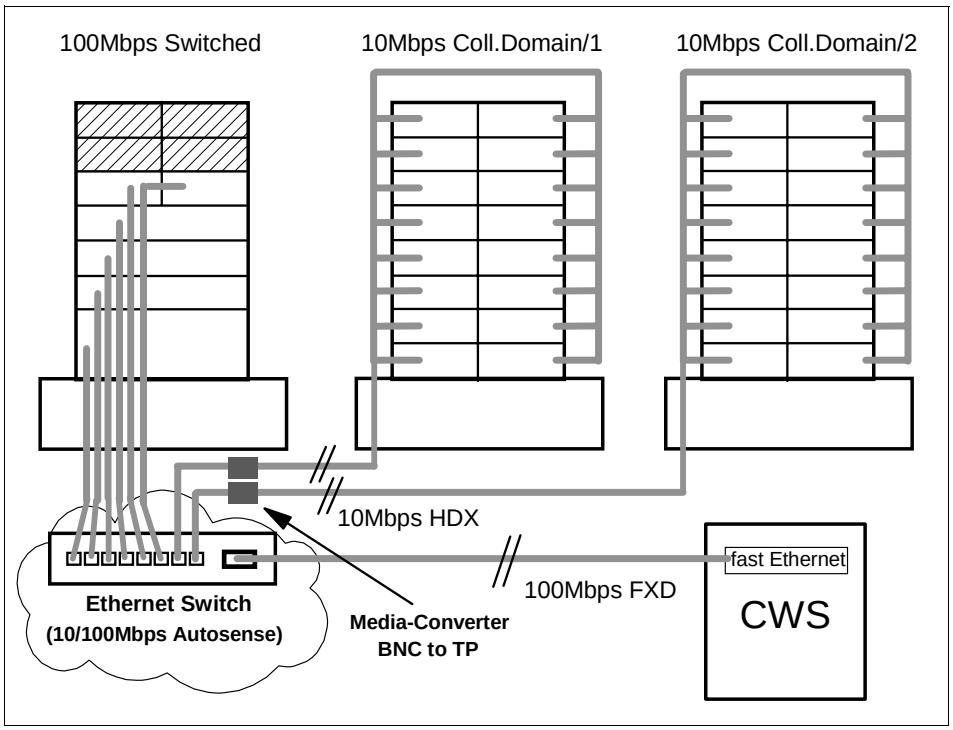


Figure 55. Heterogeneous 10/100 Mbps SP Network

In this configuration, again an Ethernet Switch, such as the IBM 8274, is used to provide a single LAN and connects to the control workstation at 100 Mbps FDX. One frame has new nodes with 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 as in the previous example. Two frames with older nodes and 10BASE-2 cabling are connected to ports of the same Ethernet Switch using media converters as in the configuration shown in Figure 53 on page 92. Ideally, a switching module with autosensing ports is used, which automatically detects the communication speed.

3.3.2 Additional LANs

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

3.3.3 IP over the Switch

If your SP has a switch, and you want to use IP for communications over the switch, each node needs to have an IP address and name assigned for its switch interface, the css0 adapter. If hosts outside the SP switch network need to communicate over the switch using IP with nodes in the SP system, those hosts must have a route to the switch network through one of the SP nodes or through the SP Switch router.

If you are not enabling ARP on the switch, specify the switch network subnet mask and the starting node's IP address. After the first address is selected, subsequent node addresses are based on the switch port number assigned. Unlike all other network interfaces, which can have sets of nodes divided into several different subnets, the switch IP network must be one contiguous subnet that includes all the nodes in the system partition.

If you want to assign your switch IP addresses as you do your other adapters, you must enable ARP for the css0 adapter. If you enable ARP for the css0 adapter, you can use whatever IP addresses you wish, and those IP addresses do not have to be in the same subnet for the whole system.

3.3.4 Subnetting Considerations

All but the simplest SP system configurations will likely include several subnets. Thoughtful use of netmasks in planning your networks can economize on the use of network addresses.

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

- 192.168.3.1-31 to the control workstation to node 1 subnet
- 192.168.3.33-63 to the frame 1 subnet

- 192.168.3.65-96 to frame 2

For example, if we used 255.255.255.0 as our netmask, then 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 56 on page 98.

Consider the example of a multi-frame SP that has a CWS with separate Ethernet connections to the node in the first slot in each frame. Each first node has a network that connects to every other node in that frame.

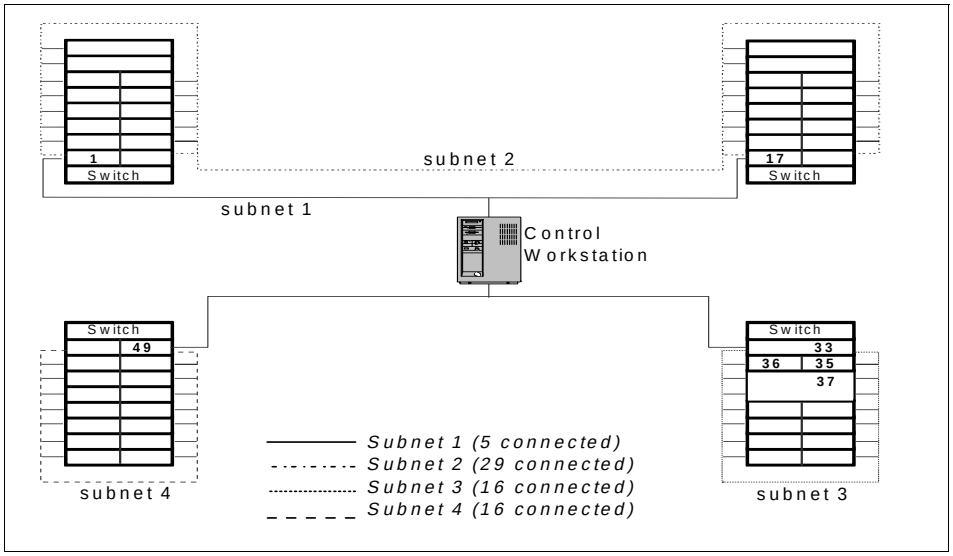


Figure 56. SP Ethernet Subnetting Example

3.4 Routing Considerations

When planning routers, especially router nodes, in your system, several factors can help determine the number of routers needed and their placement in the SP configuration. The number of routers you need can very 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. Refer to 2.5.1, “SP Switch Router” on page 26 for details.

3.5 Related Documentation

These following documentations will help you understand the concepts and examples covered in this guide. Refer to the documentations mentioned in this chapter 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 book is essential to understand the planning and requirements of SP system networking.

IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment, GA22-7280, Chapter 15. This chapter will help you to understand the SP-attached Server.

RS/6000 SP Overview, Planning and Installation Course AU91. This course material is easy to follow and will help you to understand your networking configuration.

SP Redbooks

Inside the RS/6000 SP, SG24-5145. This book will help you to understand how the RS/6000 SP is affected by the network.

Others

IBM Certification Study Guide: AIX V.4.3 System Support, SG24-5139. This book helps to understand some part of the SP system that relates closely to networking design.

TCP/IP, SNA, HACMP, and Multiple Systems, SG24-4653. This redbook contains in-depth discussion on protocols and will help you to strengthen your knowledge in this area.

3.6 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 requires an Ethernet connection between the control workstation and all nodes. Which of the following tasks do NOT use the SP Ethernet?

   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 to the en0 interface is know as:

   A. Initial hostname

   B. Reliable hostname

   C. Hostname

   D. Primary name

3. What is the default order for resolving host names if /etc/resolv.conf is present?

   A. /etc/hosts - DNS - NIS

   B. DNS - NIS - /etc/hosts

   C. DNS - NIS - /etc/hosts

   D. NIS - /etc/hosts - DNS

4. In a possible scenario with a segmented 10Base-2 network, the control workstation 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 the above

Chapter 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 RS/6000 SP environments. It also covers a discussion about 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.
- Basic understanding of 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 consider 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 satisfies your needs.

Disk options offer the following trade-offs 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 4 I/Os: A read and a write to two separate disks. 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.

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 if 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 on page 102 has more information on disk storage choices.

Table 4. Disk Storage Subsystems

Disk Storage	Description
2100	The Versatile Storage Server (VSS) offers the ability to share disks with up to 64 hosts through Ultra SCSI connections. The hosts can be RS/6000, NT, AS/400, and other UNIX platforms. The VSS has a protected storage capacity of up to 2 TB. It can be connected through multiple Ultra SCSI busses (up to 16) for increased throughput and has up to 6 GB of read cache. Internally, SSA disks are configured in RAID 5 arrays with fast write cache availability. The 7133 is an integral part of VSS. Your existing 7133 SSA disks can be placed under control of the VSS. They can remain in their current racks, or they can be placed in the VSS enclosures. Disks are configured into 6+P+S or 7+P RAID 5 arrays with at least one hot spare per loop and typically one 7133 drawer per SSA loop. These RAID 5 arrays are then divided into LUNs (logical units) with valid LUN sizes of 0.5, 1, 2, 4, 8, 12, 16, 20, 24, 28, and 32 GB. Each LUN is an hdisk in the RS/6000.
7027	The 7027 High Capacity Storage Drawer provides up to a maximum of 67.5GB of disk storage plus three tape or CD-ROM bays all in a single rack drawer. Supporting SCSI-2 Fast/Wide single-ended and SCSI-2 Fast/Wide differential, the 7207 can attach to Micro Channel-based RS/6000 systems. Offering hot-swap disk and remote power-on capabilities, it offers exceptional performance in storage expansion and growth.

Disk Storage	Description
7131	The tower has five hot swappable slots for 4.5, or 9.1 GB disk drives for a maximum 45.5 GB capacity. Two towers can provide a low cost mirrored solution.
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 support 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.
7137	The 7137 subsystem supports both RAID 0 and RAID 5 modes. It can hold from 4 to 33GB of data (29GB 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

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-hosts connections, or if 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`

Figure 57 on page 104 shows external devices configuration that can be connected to an SP system.

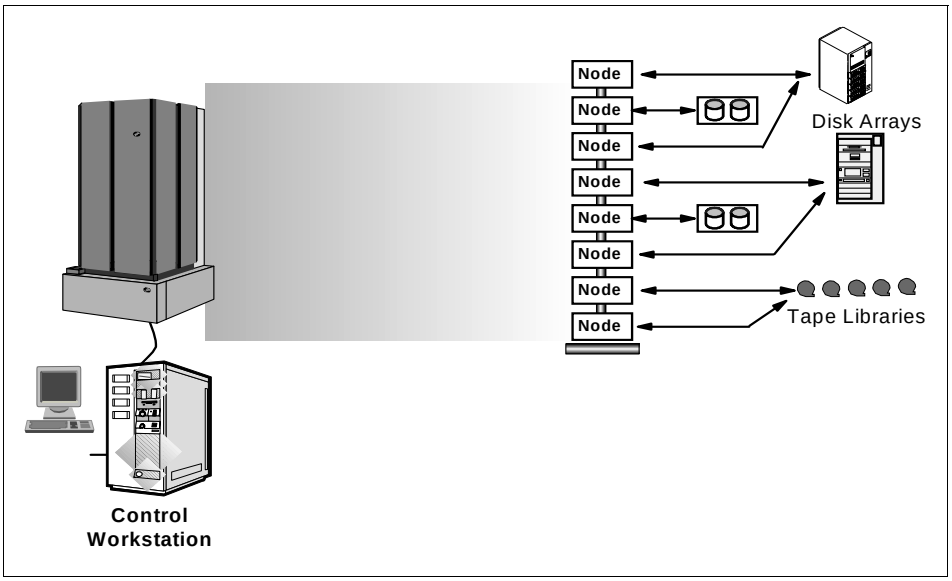


Figure 57. External Devices

4.2.2 Internal I/O Adapters

There are two types of internal I/O adapters you could use in your configuration depending on the types of nodes you have: PCI adapters and MCA adapters.

4.2.2.1 PCI Adapters

This section provides information on RS/6000 SP system PCI adapters. The following features are installed in the SP nodes and are used to connect the SP system with external networks. Network connections through SP nodes are typically slower than network connections through an SP Switch Router. Also, network connections through a node may not have the availability of those through an SP Switch Router.

Table 5 has more information on available PCI adapters.

Table 5. Available PCI Adapter Features

Feature Code	PCI Adapter Name
2741	FDDI SK-NET LP SAS
2742	FDDI SK-NET LP DAS

Feature Code	PCI Adapter Name
2743	FDDI SK-NET UP SAS
2751	S/390 ESCON Channel Adapter
2920	Token-Ring Auto Landstream
2943	EIA 232/RS-422 8-port Asynchronous Adapter
2944	WAN RS232 128 port
2962	2-port Multiprotocol X.25 Adapter
2963	ATM 155 TURBOWAYS UTP Adapter
2968	Ethernet 10/100 MB
2985	Ethernet 10 MB BNC
1987	Ethernet 10 MB AUI
2988	ATM 155 MMF
6206	Ultra SCSI Single Ended
6207	Ultra SCSI Differential
6208	SCSI-2 F/W Single-Ended
6209	SCSI-2 F/W Differential
6215	SSA RAID 5

4.2.2.2 MCA Adapters

This section provides information on RS/6000 SP system MCA adapters.

Table 6 has more information on available PCI adapters.

Table 6. Available MCA Adapter Features

Feature Code	Adapter Description
2402	IBM Network Terminal Accelerator - 256 Session
2403	IBM Network Terminal Accelerator - 2048 Session
2410	SCSI-2 High Performance External I/O Controller
2412	Enhanced SCSI-2 Differential Fast/Wide Adapter/A
2415	SCSI-2 Fast/wide Adapter/A

Feature Code	Adapter Description
2416	SCSI-2 DIfferential Fast/Wide Adapter/A
2420	SCSI-2 Differential High Performance External I/O Controller
2700	4-port multiprotocol Communication Controller
2723	FDDI Dual-Ring Attachment
2724	FDDI Single-Ring Attachment
2735	High Performance Parallel Interface - HIPPI
2754	S/390 ESCON Channel Emulator Adapter
2755	Block Multiplexer Channel Adapter - BMCA
2756	ESCON Control Unit Adapter
2930	8-port Async Adapter-EIA-232
2940	8-port Async Adapter-EIA-422A
2960	X-25 Interface Co-Processor/2
2970	Token Ring High Performance Network Adapter
2972	Auto Token Ring LANstreamer MC 32 Adapter
2980	Ethernet High Performance LAN Adapter
2984	TURBOWAYS 100 ATM Adapter
2989	TURBOWAYS 100 ATM Adapter
2992	High-Performance Ethernet LAN Adapter (AUI/10baseT)
2993	High-Performance Ethernet LAN Adapter (BNC)
2994	10/100 Ethernet Twisted Pair MC Adapter
4224	Ethernet 10BaseT Transceiver
6212	9333 High Performance Subsystem Adapter
6214	SSA 4-port Adapter
6216	Enhanced SSA 4-port Adapter
6217	SSA 4-port RAID Adapter
6219	Micro Channel SSA Multi-initiator/RAID EL Adapter

Feature Code	Adapter Description
6305	Digital Trunk Dual Adapter
7006	Real-time Interface Co-Processor Portmaster Adapter/A
8128	128-port Async Controller

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, a new SDR class called Volume_Group has been 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.

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 command `spbootlist`.

Another enhancement in PSSP 3.1 is the possibility of mirroring the Root Volume Group directly from the CWS. Mirroring is writing simultaneous copies of the operating systems logical volumes to provide redundancy. Either two or three copies (one or two mirrors) are allowed in AIX.

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

Prior to PSSP 3.1, on 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 3.1or 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 4.3.2. To display this information, a new option (-v) has been 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 lists its 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.

4.3.1.1 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.

4.3.1.2 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 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.
- 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 version (previous to PSSP 3.1) are described.

4.3.2.1 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 58 on page 110.

```
                        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                       [bos.obj.mksysb.aix432.090898]
  LPP Source Name                                  [aix432]
  PSSP Code Version                                 PSSP-3.1 +
  Set Quorum on the Node                                                    +


F1=Help                 F2=Refresh              F3=Cancel               F4=List
F5=Reset                F6=Command              F7=Edit                 F8=Image
F9=Shell                F10=Exit                Enter=Do
```

Figure 58. 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 the 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: 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:

- Issue the command:

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

- 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//0123456789AB00D
  ```

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

  ```
  ssar//0123456789AB00D:ssar//0123456789FG00D
  ```

Important

The ssar identifier must have a length of 21 characters.

Installation on external SSA disks is supported in PSSP 3.1 or later.

4.3.2.2 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 59.

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                      []
  LPP Source Name                                 []
  PSSP Code Version                                PSSP-3.1 +



F1=Help                 F2=Refresh              F3=Cancel               F4=List
F5=Reset                F6=Command              F7=Edit                 F8=Image
F9=Shell                F10=Exit                Enter=Do
```

Figure 59. 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
```

4.3.2.3 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 60 for details.

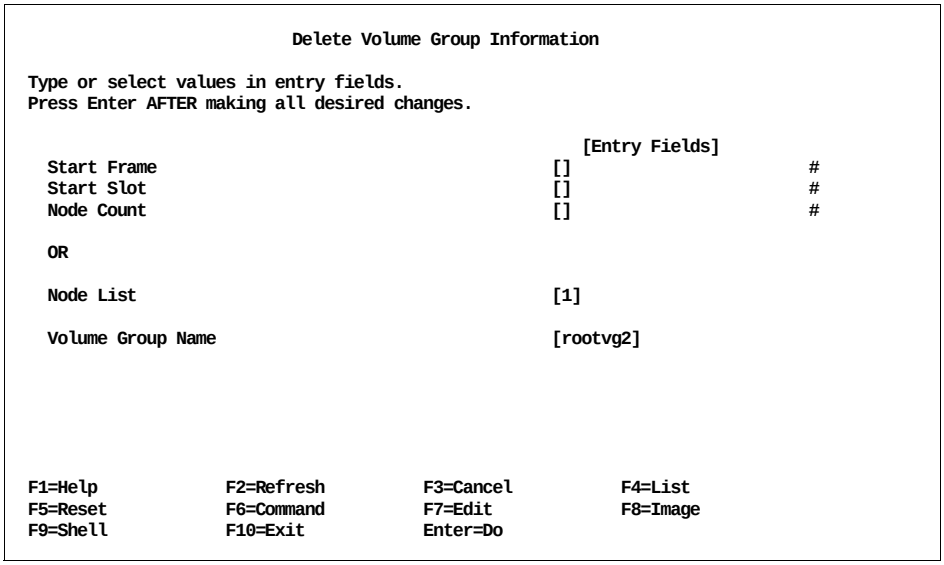


Figure 60. New SMIT Panel to Delete a Volume Group

The following is an example built by the SMIT panel used in Figure 60:

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

4.3.2.4 Changes to spbootins in PSSP 3.1 or later

spbootins is the command to set various node attributes in the SDR (code_version, lppsource_name, and so forth).

By using 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 new spbootins is as follows:

spbootins

-r <install|diag|maintenance|migrate|disk|customize>

-l <node_list>

-c <selected_vg>

-s <yes|no>

spbootins no longer have the following flags:

-h <install_disk>

-n <boot_server>

-v <lppsource_name>

-i <install_image_name>

-p <PSSP_level>

-u <usr_server_id>

-g <usr_gateway_id>

-a <interface name>

Note

-u, -g, and -a flags were dropped because PSSP 3.1 no longer supports /usr servers.

Figure 61 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 +


F1=Help              F2=Refresh           F3=Cancel            F4=List
F5=Reset             F6=Command           F7=Edit              F8=Image
F9=Shell                F10=Exit                Enter=Do
```

Figure 61. 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.

4.3.2.5 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:

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

Figure 62 on page 115 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                     +




F1=Help            F2=Refresh         F3=Cancel          F4=List
F5=Reset           F6=Command         F7=Edit            F8=Image
F9=Shell           F10=Exit           Enter=Do
```

Figure 62. New SMIT Panel to Initiate the spmirrorvg Command

The following is an example built by the SMIT panel in Figure 62:

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

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

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

4.3.2.6 spunmirrorvg

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

```
-l node_list
```

Figure 63 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

  Node List                                       []



F1=Help              F2=Refresh           F3=Cancel            F4=List
F5=Reset             F6=Command           F7=Edit              F8=Image
F9=Shell             F10=Exit             Enter=Do
```

Figure 63. New SMIT Panel to Initiate the spunmirrorvg Command

The following is the example built by the SMIT panel in Figure 63:

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

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

4.3.2.7 Changes to splstdata in PSSP 3.1 or Later

`splstdata` can now display information about Volume_Groups using the new option:

```
-v
```

Figure 64 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

        ssar//0004AC50532100D:ssar//0004AC50616A00D
    1 jmbvg           0          true      1         PSSP-3.1 aix432

        default                  Fri_Sep_25_11:50:47_EDT_1998 hdisk0

        ssar//0004AC5150BA00D
```

Figure 64. Example of splstdata -v

4.3.2.8 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.

Section 4.3, “Multiple rootvg Support” on page 107 gives 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 58 on page 110 and Figure 61 on page 114.

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.

4.3.3.1 How to Activate a New rootvg

Several rootvg's 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 65 on page 119. 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 by issuing the command:

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

Then reboot the node.

The following example gives the steps to follow 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
   ```

2. Note: it is not necessary to run `setup_server`.
3. Verify:

   ```
   splstdata -b
   ```

4. Change the node bootlist:

   ```
   spbootlist -l 1
   ```

5. Verify:

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

6. Reboot the node:

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

Important

The key switch must be in the normal position.

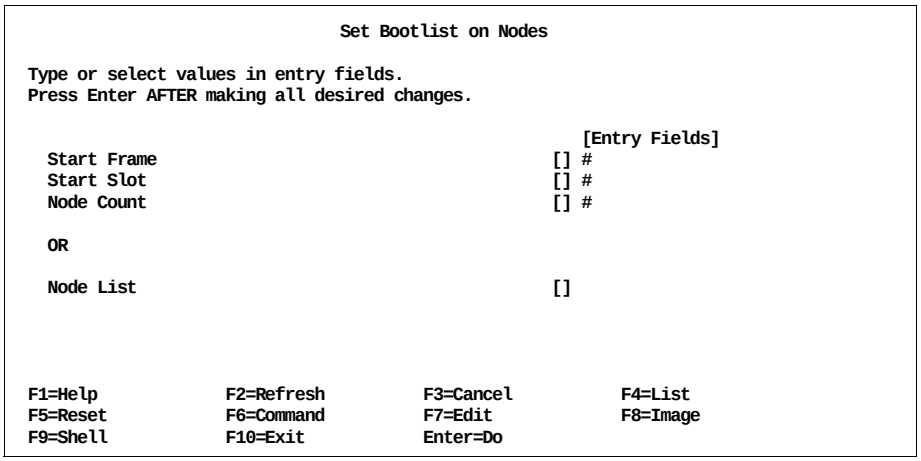


Figure 65. SMIT Panel for the spbootlist Command

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.

4.3.4.1 SSA Disk Requirements

Figure 66 and Figure 67 on page 120 show the SSA disk connections to a node.

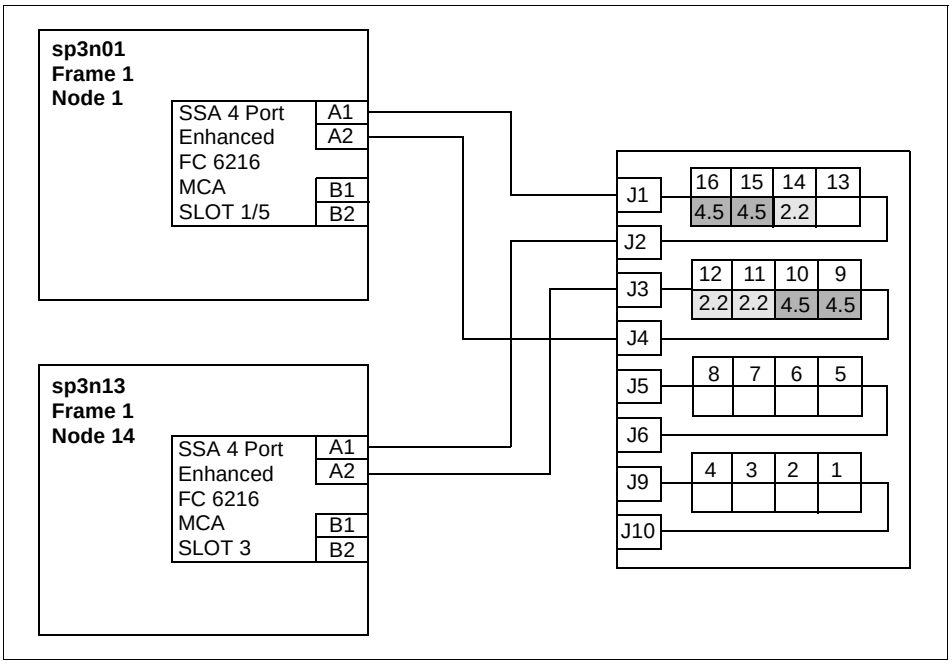


Figure 66. Cabling SSA Disks to RS/6000 SP Nodes

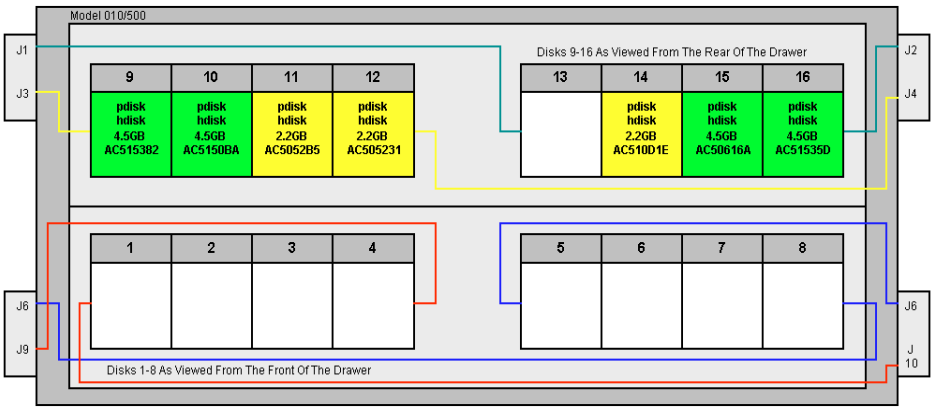


Figure 67. Connections on the SSA Disks

Not all node types can support an SSA boot. Table 7 shows the node types that support an SSA boot.

Table 7. Supported Adapters for Nodes with Full SSA Boot

Node Feature Code	Node Type	Feature Code Numbers of Supported SSA Adapters
2005	77 MHz Wide	#6214 SSA 4-Port Adapter #6216 Enhanced SSA 4-Port Adapter #6217 SSA RAID Adapter #6219 Enhanced SSA RAID Adapter
2006	604 High	Same as above
2007	120 MHz Thin	Same as above
2008	135 MHz Wide	Same as above
2009	604e High	Same as above
2022	160 MHz Thin	Same as above

The SP-supported external SSA disk subsystems are:

7133 IBM Serial Storage Architecture Disk Subsystems Models 010, 020, 500, and 600.

4.3.4.2 SCSI Disk Requirements

Some nodes can now be booted from an external SCSI-2 Fast/Wide disk 7027-HSD storage device. Not all nodes can support an SCSI boot. Table 8 lists the nodes and the adapters for external disk booting.

Table 8. Supported Adapters for Nodes with SCSI Boot

Node Feature Code	Node Type	Feature Code Numbers of Supported SCSI Adapters
2002	66 MHz Thin	#2412, #2416
2003	66 MHz Wide	Same as above
2004	66 MHz Thin 2	Same as above
2005	77 MHz Wide	Same as above
2006	604 High	Same as above
2007	120 MHz Thin	Same as above
2008	135 MHz Wide	Same as above
2009	604e High	Same as above

Node Feature Code	Node Type	Feature Code Numbers of Supported SCSI Adapters
2022	160 MHz Thin	Same as above
2050	332 MHz SMP Thin	#6207, #6209
2051	332 MHz SMP Wide	#6207, #6209

The SP-supported external SCSI disk subsystems are:

7027-HSD IBM High Capacity Drawer with an SP SCSI-DE/FW adapter for Micro Channel machines or SP Ultra-SCSI adapter for PCI machines.

4.3.4.3 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 59 on page 112) to get straight there.

Figure 68 on page 123 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 4.3.2.1, “spmkvgobj” on page 109 for more information on how to enter the information).

```
                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]
  Number of Copies of Volume Group                    1                       +
  Set Quorum on the Node                                                      +
  Boot/Install Server Node                           []                        #
  Network Install Image Name                         []
[MORE...2]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 68. 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 69 on page 124. 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 is a new command in PSSP 3.1 that recognizes the new 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
                PSSP-3.1            rootvg
```

Figure 69. Output of the splstdata -b Command

4.3.4.4 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 70 on page 125 for node 1's bosinst.data file (node 1 was used as an example node in Figure 68 on page 123 and Figure 70 on page 125). NIM puts in the new 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//0004AC50532100D

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

Figure 70. 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 12.4, "General Parallel File Systems" on page 323.

4.4.1 Network File System (NFS)

Sun Microsystem's Network File System (NFS) is a widely used global file system, which is 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 71 on page 127.

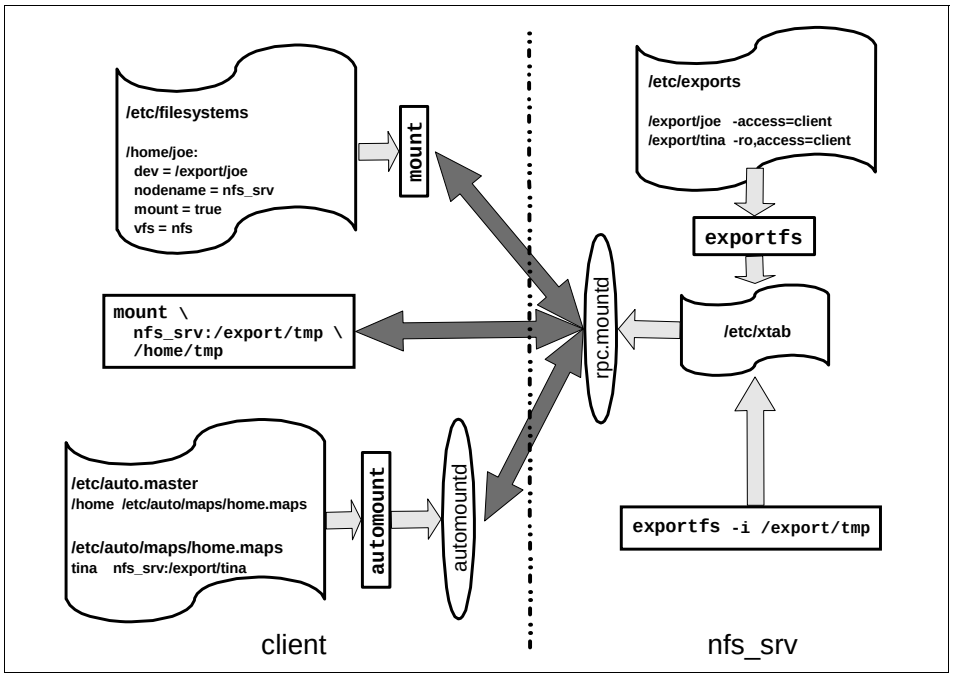


Figure 71. Conceptual Overview of NFS Mounting Process

The PSSP software uses NFS for network installation 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 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 to compare NFS to other global file systems.

Performance: NFS Version 3 contains several improvements over NFS Version 2. The most important change probably is that NFS Version 3 no longer limits the buffer size to 8 kB 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 needs 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 filesystem 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 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.

4.4.2.1 What Is the Distributed File System?

The 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 the 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 a 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.

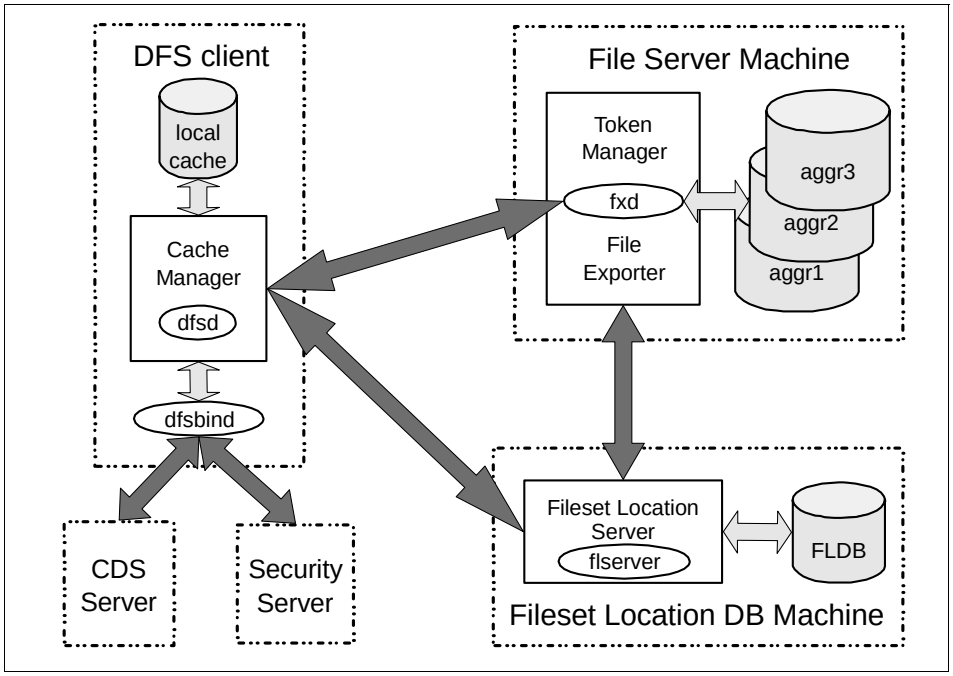


Figure 72. Basic DFS Components

Figure 72 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: DFS Administration Guide and Reference*.

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 126.

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 LFS 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.

4.4.2.2 Differences of 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 different concepts and examples covered in this chapter. We recommend you to 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 you detailed explanations on I/O devices.

IBM 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.

4.6 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. If you attach a tape drive to one of your nodes, which of the following statements is 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 is 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 is 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.

Chapter 5. SP-Attached Server Support

PSSP 3.1 provides support for the RS/6000 Enterprise Server Models S70 and S7A known as SP-attached servers. These are high-end RS/6000 PCI-based and are the first 64-bit SMP architecture nodes that attach independently to the SP as they are simply too large to physically reside in an SP frame.

The main section in this chapter is subdivided into the following five sections:

1. The system attachment of the SP-attached server to the SP is discussed in “Hardware Attachment” on page 135.
2. Installation and configuration of an SP-attached server are discussed in “Installation and Configuration” on page 146.
3. The PSSP support to SP-attached server is discussed in “PSSP Support” on page 152.
4. User interface panels and commands are discussed in “User Interfaces” on page 161.
5. Different attachment scenarios to the SP are discussed in “Attachment Scenarios” on page 166.

5.1 Key Concepts You Should Study

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 are the software levels required to attach an RS/6000 Enterprise Server (S70/S7A)?
- The difference between an SP-attached server and a dependent node.
- What are node, frame, and switch numbering rules when attaching an RS/6000 Enterprise 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 Brief RS/6000 Enterprise Server Overview

The RS/6000 Enterprise Server Model S70 (7017) is a 64-bit symmetric multiprocessing (SMP) system that supports 32- and 64-bit applications concurrently.

Until now, all nodes in an SP environment resided within the slot location of an SP frame. However, the SP-attached server is physically too large to reside in an SP frame slot location as it is packaged in two side-by-side rack units as shown in Figure 73 on page 137.

The first unit is a 22w x 41d x 62h-inch (56w x 104d x 157h-cm) Central Electronics Complex (CEC). The CEC system rack contains:

- A minimum of one processor card and a maximum of three processor cards with a 4-, 8-, or 12-way PowerPC processor configuration. The system can contain up to a maximum of 12 processors sharing common system memory.
- Each processor card has four 64-bit processors operating at 125 Mhz or 262 Mhz.
- A 4 MB ECC L2 cache memory per 125 Mhz processor and an 8 MB per 262 Mhz processor.
- System memory is controlled through a multiport controller that supports up to 20 memory slots. All the system memory is contained in the system rack up to a maximum of 16 GB.
- An operator panel that consists of the display unit, scroll up and down push-button, an Enter button, and two indicator LEDs. The power on/off button is also located on the operator panel. In addition, it contains a port that can be used through an RS-232 cable to communicate to the S70. The operator panel is used for selecting boot options and initiating system dumps as well as for service functions and diagnostic support of the entire system.
- Reliability from redundant fans, hot-swappable disk drives, power supplies and fans, and a built-in service processor.

The second unit is a standard I/O rack similar in size to the CEC. Each I/O rack accommodates up to two I/O drawers with a maximum of four drawers per system. Up to three more I/O racks can be added to a system. The base I/O drawer contains:

- Up to 14 PCI slots per drawer.
- Drawer zero reserves slots two and eight for support of system media.
- Service processor and hot-pluggable DASD.

- Drawers one through three are reserved for supported PCI adapters.
- One fully configured system of four I/O drawers and up to 56 PCI slots.
- Support for SCSI/SSA six-packs, looped SSA, and SIO.

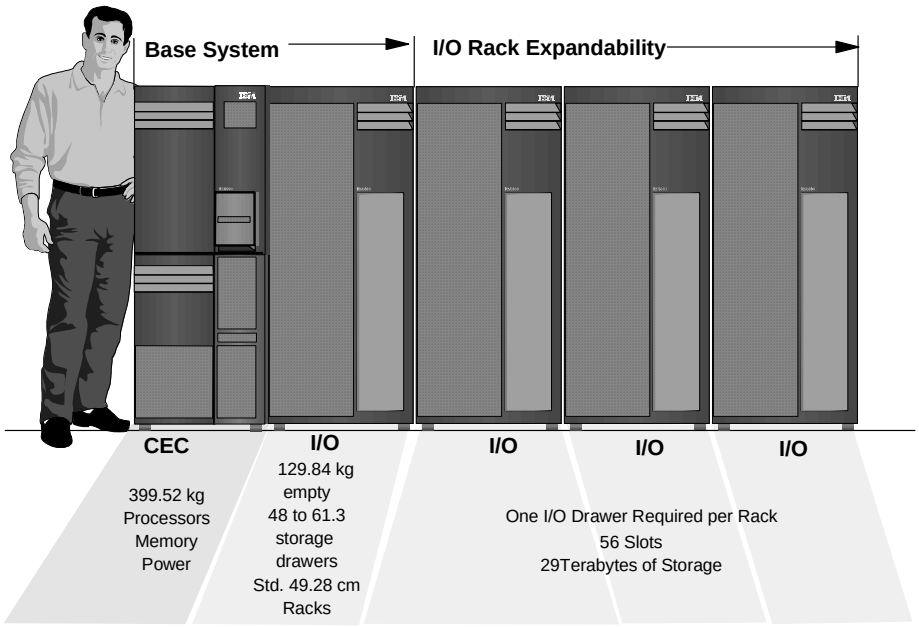


Figure 73. The S70 Components

Since the CEC and I/O racks are so large, the SP-attached server must be attached to the SP system externally.

5.2.2 SP-Attached Server Attachment

This section describes the attachment of the SP-attached server to the SP highlighting the potential areas of concern that must be met before installation. The physical attachment is subdivided and described in three connections.

- Connections between the CWS and the SP-attached server are described in “Control Workstation Connections” on page 141.
- Connections between the SP Frame and the SP-attached Server are described in “SP Frame Connections” on page 142.
- 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 143.

These connections are illustrated in Figure 74 on page 138.

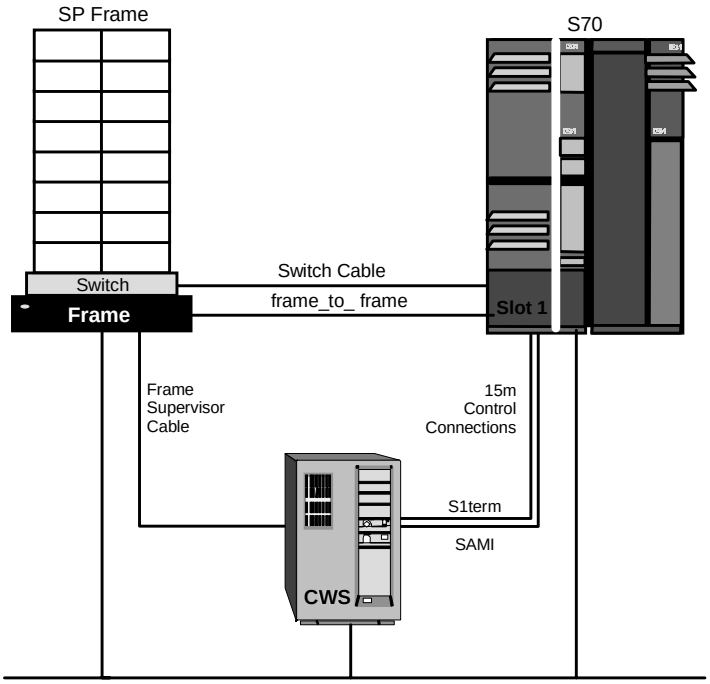


Figure 74. The S70 Attachment to the SP

The following diagram outlines the two RS-232 connections to the S70 machine.

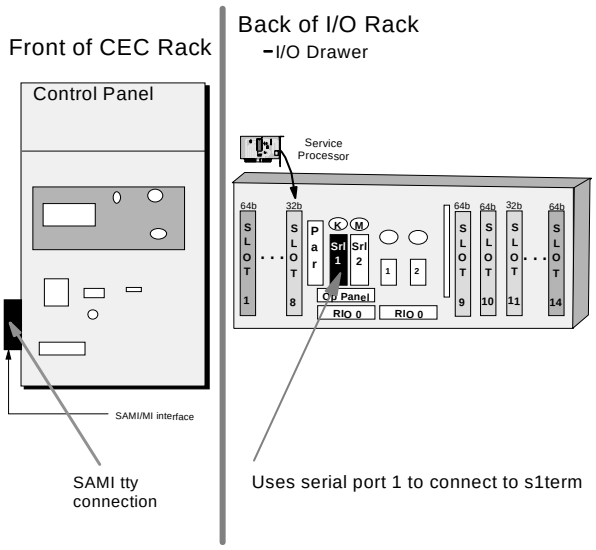


Figure 75. RS-232 Connections to the S70

It is important to note that the size of the S70 prohibits it from being physically mounted in the SP frame. Since the SP-attached server is mounted in its own rack and is directly attached to the CWS using RS-232, the SP system must view the SP-attached server as a frame. The SP-attached server 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 as 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 5.6, "Attachment Scenarios" on page 166.

Once the frame number has been assigned, the server's node numbers, which are based on the frame number, are automatically generated. The following system defaults are used:

- The SP-attached server is viewed as a single frame containing a single node.
- The 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 76; for further information on the frame numbering issue, refer to Figure 92 on page 166:

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

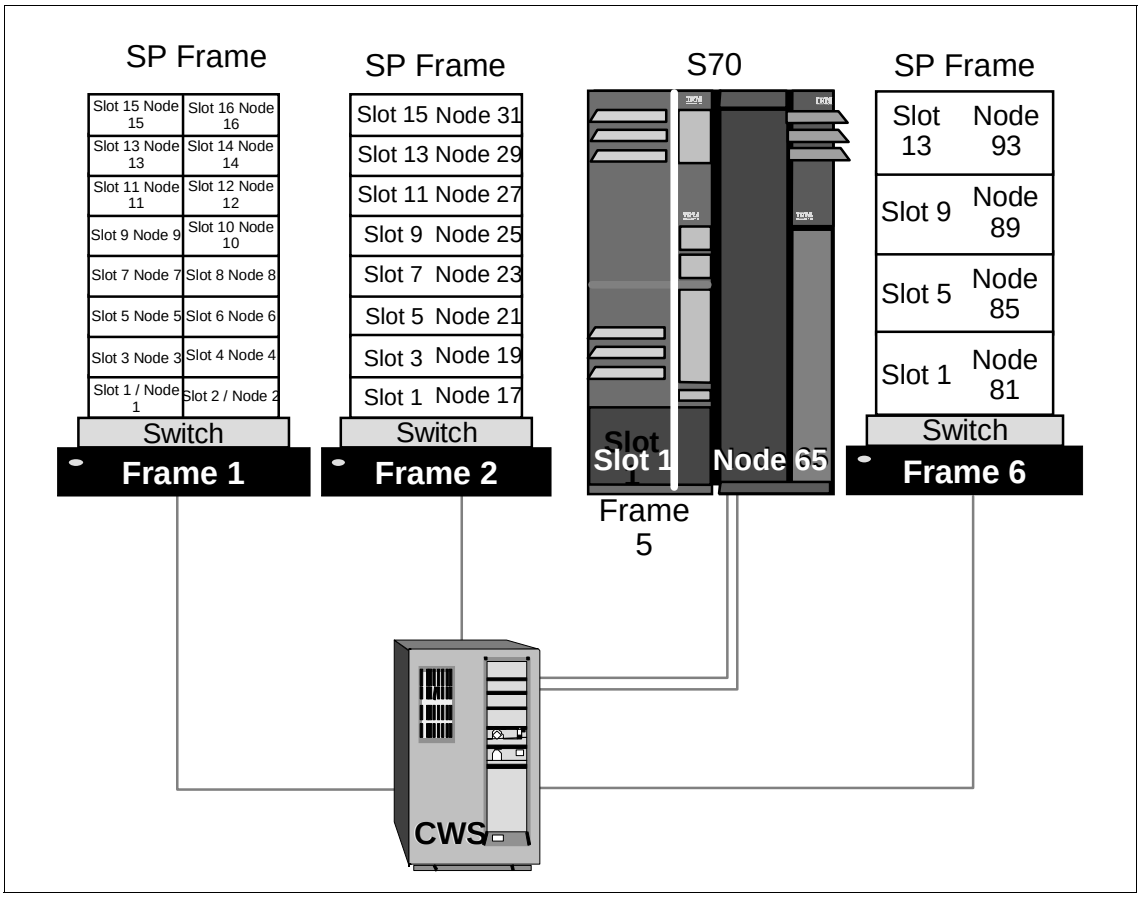


Figure 76. Node Numbering

5.2.2.1 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.

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 know 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.

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 SP attachment. The first serial port on the S70/S7A *must* be used for S1TERM connectivity.
- Floor placement planning to account for the effective usable length of RS-232 cable.

  The CWS-to-S70 connection cables are 15 meters in length, but only 11.5 meters is effective. So, the S70 must be placed at a distance where the RS-232 cable to the CWS is usable.
- In a HACWS environment, there will be no S70 control from the backup CWS. In the case where a failover occurs to the backup CWS, hardmon and s1term support of the S70 is not available until fail back to the primary CWS. The node will still be operational with switch communications and SP Ethernet support.

5.2.2.2 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* as the 49inch 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.
- Maximum of eight SP-attached servers 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 *Site and Hardware Planning Information*, SA38-0508.

5.2.2.3 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 TB3PCI adapter, known as the RS/6000 SP system attachment adapter, of the SP-attached server connects to the 16-port SP switch through a 10 meter switch cable.

  This TB3PCI adapter is used in those systems that are connected to the switch board using a PCI adapter, and it has the following characteristics:

  - It is driven by a 99 Mhz 603e PowerPC processor.
  - It has a sustained bandwidth of 85 MByte/sec.
  - It has components familiar to the SP environment.
  - Its device driver is derived from TB3MX.
  - It is supported *only* in the S70 server family.

Switch Considerations

In connecting the SP-attached server to the SP Switch, it is important to note the following:

- The High Performance switch (HiPS) cannot be used with an SP-attached server since this switch is not supported in PSSP 3.1.
- The S70/S7A servers will be the first, and currently the only, nodes attached to the switch using an RS/6000 SP Attachment adapter.

- Only *one* RS/6000 SP Attachment adapter is allowed per SP-attached server.
- The RS/6000 SP Attachment adapter that is placed in the SP-attached server requires:
  - One valid, unused switch port on the SP switch, corresponding to a legitimate node slot in your SP configuration.
  - The SP attachment adapter reserves three media slots in the I/O tower of the S70 server and has the following placement restrictions:
    - Must be installed in slot 10 of the SP-attached server's I/O tower
    - Slot 9 must be left open to ensure that the adapter has sufficient bandwidth.
    - Slot 11 must be left open to provide clearance for the switch adapter's heat sinks.

  These restrictions are illustrated in Figure 77 on page 144.

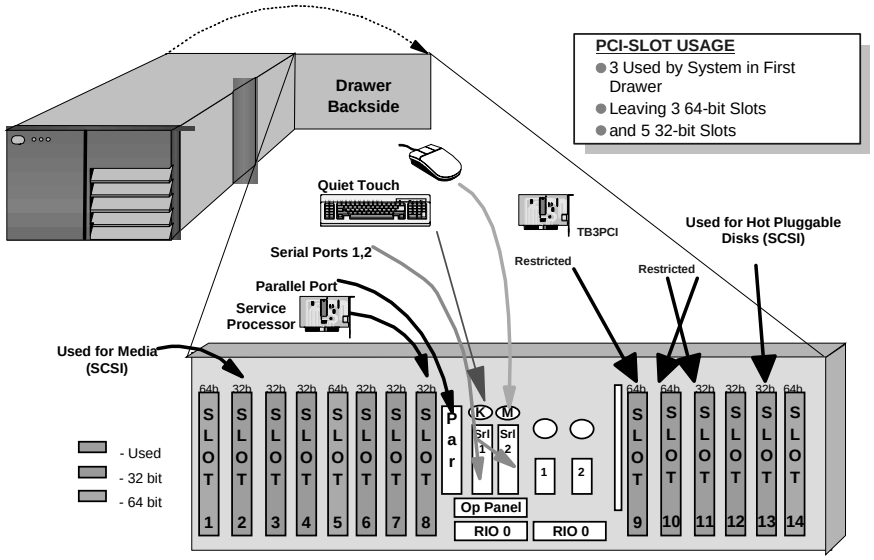


Figure 77. S70 Switch Adapter Attachment Slot

- Floor placement planning to account for the effective usable switch cable.

  The SP switch-to S70 connection cable is 10 meters in length but only 6.5 meters is effective. So, the S70 switch adapter located in slot 10

must be within 6.5 meters of the SP switch as illustrated in Figure 78 on page 145.

.

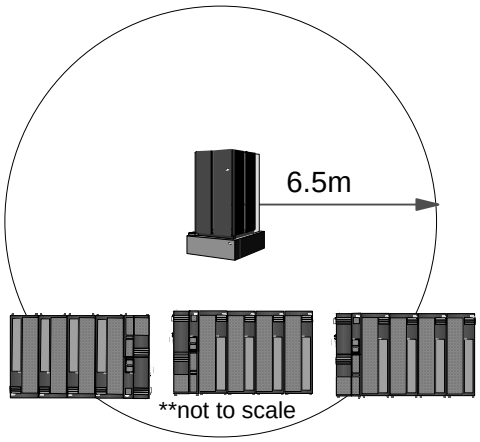


Figure 78. S70 Floor Placement

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 (S70). However, not all currently supported SP-attached server adapters are supported in the SP switch-attached server environment. If the S70 possesses adapters that are not currently supported in the SP environment, they *must* be removed from the SP-attached server.

  The following is a list of supported adapters:

  - F/C 2741 FDDI SK-NET LP SAS
  - F/C 2742 FDDI SK-NET LP DAS
  - F/C 2743 FDDI SK-NET UP SAS
  - F/C 2751 S/390 ESCON Channel Adapter
  - F/C 2920 Token Ring Auto Lanstream
  - F/C 2943 EIA 232/RS-422 8-port Asynchronous Adapter
  - F/C 2944 WAN RS-232 128-port
  - F/C 2962 2-port Multiprotocol X.25 Adapter
  - F/C 2963 ATM 155 TURBOWAYS UTP
  - F/C 2968 Ethernet 10/100 MB
  - F/C 2985 Ethernet 10 MB BNC

- F/C 2987 Ethernet 10 MB AUI
- F/C 2988 ATM 155 MMF
- F/C 6206 Ultra SCSI SE
- F/C 6207 Ultra SCSI DE
- F/C 6208 SCSI-2 F/W SE
- F/C 6209 SCSI-2 F/W DE
- F/C 6215 SSA RAID 5

- SP-attached server Ethernet required as en0:

  For the S70 server, only the 10Mbps BNC or the 10Mbps AUI Ethernet adapters are supported for SP-LAN communication in accordance with the existing SP-LAN configuration. Note that the BNC adapters provides the BNC cables, but the AUI ethernet adapter does *not* provide the twisted pair cables.

  The SP-LAN adapter must be configured as the en0 adapter of the SP-attached server (that is, the lowest numbered Ethernet bus slot in the first I/O tower).

- Minimum code requirements:

  The CWS and SP-attached server must be running AIX 4.3.2 and PSSP 3.1 at the minimum. Hence, an existing S70 may require an AIX upgrade before installation of PSSP 3.1 to achieve SP-attachment.

Note

Each SP-attached server S70 must have a PSSP 3.1 licence separately chargeable against each S70's serial number.

5.3 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 5.2.2, "SP-Attached Server Attachment" on page 137.

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

- 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 SP-attached server:

  The rules for assigning the frame number of the SP-attached server are detailed in section 5.2.2, “SP-Attached Server Attachment” on page 137.

  The SP-attached server must be defined to PSSP, using the `spframe` command, and using the new options that are available for SP-attached server for this command:

  ```
  /usr/lpp/ssp/bin/spframe -p {hardware protocol}
  -n {starting_switch_port}
  [-r {yes|no}]   [-s {s1tty}]
  start_frame frame_count starting_tty_port
  ```

  Alternatively, you can use the `smitty nonsp_frame_dialog` menu as shown in Figure 79.

```
                    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                        +




F1=Help            F2=Refresh         F3=Cancel          F4=List
Esc+5=Reset        Esc+6=Command      Esc+7=Edit         Esc+8=Image
Esc+9=Shell        Esc+0=Exit         Enter=Do
```

Figure 79. 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.

- Hardware Ethernet address collection:

  The MAC address of the SP-Attached server is retrieved by `sphrdwrad` in just the same way as a normal SP node and placed in the SDR.

  Now that the SP-attached server is configured as a 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 is 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 `spdelfram` 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.

  1. 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 166.

  2. Software levels.

     If your SP-attached server is not at AIX 4.3.2, upgrade to that level.

     Ensure that the PSSP code_version is set to PSSP-3.1.

  3. 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
     ```

  4. 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
```

5. 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.

6. 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 hostname must be defined. For example, the control workstation's hostname 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
   ```

7. 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.

8. Verify perfagent.

   Ensure that perfagent.tools 2.2.32.x are installed on the SP-attached server.

9. 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
   ```

10. 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
```

11. 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
    ```

12. Run pssp_script.

    Run the pssp_script by issuing:

    ```
    /usr/lpp/ssp/install/bin/pssp_script
    ```

13. Reboot.

    Perform a reboot of the SP-attached server.

5.3.1 Pre-Installation Checklist

Using the SP configurator, the following hardware and software components for the SP-attached server should be ordered.

1. Feature 9122 Node Attachment.

   The feature provides the following:

   - 15 meters RS-232 cable between S70 and CWS (S1TERM).
   - 15 meters RS-232 cable between S70 and CWS (SAMI).
   - This feature includes the frame-to-frame electrical ground cable.

2. Feature 9123 Frame Attachment.

   This feature keeps track of how many frames are in your SP system to avoid exceeding the limit.

3. Feature 5700/1/2 for SP-Attached Server PSSP.

   PSSP 3.1 is a separately charged software license for each SP-attached server.

   AIX 4.3.2 is included with the SP-attached server and preloaded at the factory and, therefore, does not need to be ordered separately.

   This feature must be ordered for a non-switched system as well.

4. 9222 Node Attachment Ethernet BNC Boot Feature.

   Includes BNC cable for SP Ethernet Communications.

5. 9223 Node Attachment Ethernet Twister pair Boot Feature.

This feature does not provide twisted pair cables.

6. The following features are optional and are only required if the SP-attached server should be attached to the switch. In a switchless system, this feature is not necessary.
   - Feature 8396 RS/6000 SP System Attachment Adapter
   - Feature 9310, 10 meter SP switch cable

5.4 PSSP Support

This section describes the PSSP software support to 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.

5.4.1 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.

- 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 new 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 method is SAMI (Service and Manufacturing Interface), which is the protocol used to communicate across the serial connection to the SP-attached server service processor.

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 Figure 80.

frame_number	tty	frame_type	MAC	b_MACN	slots	f_in_config	snn_index	switch_config	hardware_protocol	s1_tty
1	/dev/tty0	switch	spcw	" "	16	1	0	0	sp	" "
2	/dev/tty2	" "	spcw	" "	1	" "	" "	" "	SAMI	/dev/tty1

Figure 80. Example of a Frame Class with an SP-Attached Server

- **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 section 5.2.2, “SP-Attached Server Attachment” on page 137.

  Likewise, the same algorithm is used to compute the node number of a 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 outlined in Figure 81 on page 154.

Node Class	Nodes is an SP	attached S70 Node
Node Number	1-16	17
Slot Number	1-16	1(always)
Switch_node_number	0-15	1
Switch_chip_port	0-15	any port used from 0-15
Switch_chip	4-7	any chip used from 4-7
Switch_number	1	1
Boot_device	en0	en0
Description	112_MHZ_SMP_High 66_MHZ_PWR2_Thin 66_MHZ_PWR2_Wide	7017-S70
Platform	rs6k	chrp
hardware_control_type	161 high, 97 thin, 81 wide, ...,etc.	10 (S70/S7A)

Figure 81. Entries of the Node Class for SP Nodes and SP-Attached Server

The platform attribute has a value of Common Hardware Reference Platform (chrp) for the SP-attached server.

The hardware_control_type key value is used to access the NodeControl class. A value of 10 suggests an SP-attached server.

- **Syspar_map Class**

  The Syspar_map class contains one entry for each switch port in potential switch port assuming each frame would contain a switch.

  As 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 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, the *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, the switch_node_number will be assigned by the administrator using the `spframe` command.

  An example of the syspar_map class is shown in Figure 82 on page 155.

syspar_name	syspar_addr	node_number	switch_node_number	used	node_type
k48s	9.114.11.48	1	0	1	standard
k48s	9.114.11.48	17	1	1	standard
k48s	9.114.11.48	3	2	1	standard
k48s	9.114.11.48	16	15	1	standard

Figure 82. Example of the Syspar_map Class with SP-Attached Server

The `SDR_config` command has been modified to accommodate these new SDR attribute values and now to handle the assignment of switch_port_numbers 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 that is loaded during installation and is not be 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 illustrated in Figure 83.

.

NodeControl Class

Type	Capabilities	Slots_used	Platform_type	Processor_type
65	Power,reset,tty,KeySwitch,LED,NetworkBoot	1	rs6k	UP
161	Power,reset,tty,KeySwitch,LCD,NetworkBoot	4	rs6k	MP
33	Power,reset,tty,KeySwitch,LED,NetworkBoot	1	rs6k	UP
10	Power,tty,LCD,NetworkBoot	1	chrp	MP
177	Power,reset,tty,LCD,NetworkBoot	1	chrp	MP
115	Power,reset,tty,KeySwitch,LED,NetworkBoot	2	rs6k	UP

Figure 83. Example of the NodeControl Class with the SP-Attached Server

The key link between the Node class and the NodeControl class is the node type, which is a new attribute stored in the SDR Node object. The

SP-attached server has a node type value of 10 with hardware capabilities of power on/off, tty, LCD, and network boot as outlined in Figure 84.

Node Class	Nodes is an SP	attached S70 Node
Node Number	1-16	17
Slot Number	1-16	1(always)
Switch_node_number	0-15	1(always)
Switch_chip_port	0-15	any port used from 0-15
Switch_chip	4-7	any chip used from 4-7
Switch_number	1	1
Boot_device	en0	en0
Description	112_MHZ_SMP_High 66_MHZ_PWR2_Thin 66_MHZ_PWR2_Wide	7017-S70
Platform	rs6k	chrp
hardware_control_type	161 high, 97 thin, 81 wide ...,etc.	10 (S70/S7A)

NodeControl Class

Type	Capabilities	Slots_used	Platform_type	Processor_type
65	Power,reset,tty,KeySwitch,LED,NetworkBoot	1	rs6k	UP
161	Power,reset,tty,KeySwitch,LCD,NetworkBoot	4	rs6k	MP
33	Power,reset,tty,KeySwitch,LED,NetworkBoot	1	rs6k	UP
10	Power,tty,LCD,NetworkBoot	1	chrp	MP
177	Power,reset,tty,LCD,NetworkBoot	1	chrp	MP
115	Power,reset,tty,KeySwitch,LED,NetworkBoot	2	rs6k	UP

Figure 84. The Relationship between Node and Node-Control Class

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.

5.4.2 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 new SP-attached server does not have a frame or node supervisor card that can communicate with the hardmon daemon. Therefore, a new mechanism to control and monitor SP-attached servers is provided in PSSP3.1.

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.

Discover the SP-Attached Server

For hardmon to discover the hardware, it must first identify the hardware and its capabilities. Today, for each frame configured in the SDR's 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.

With PSSP 3.1, two new fields have been added to the SDR's frame class: *hardware_protocol* and *s1_tty*. They 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.

Currently, the only two supported values for the *hardware_protocol* field are SP and SAMI. However, these values are extensible for new hardware protocol drivers that will emerge as more externally connected hardware is supported.

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 hardware since it is the SAMI software protocol that allows the communication, both sending messages and receiving packet data to the S70/S7A's Service Processor.

Once hardmon recognizes the existence of one or more S70/S7As 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 to normal SP frames. This is illustrated in Figure 85 on page 158.

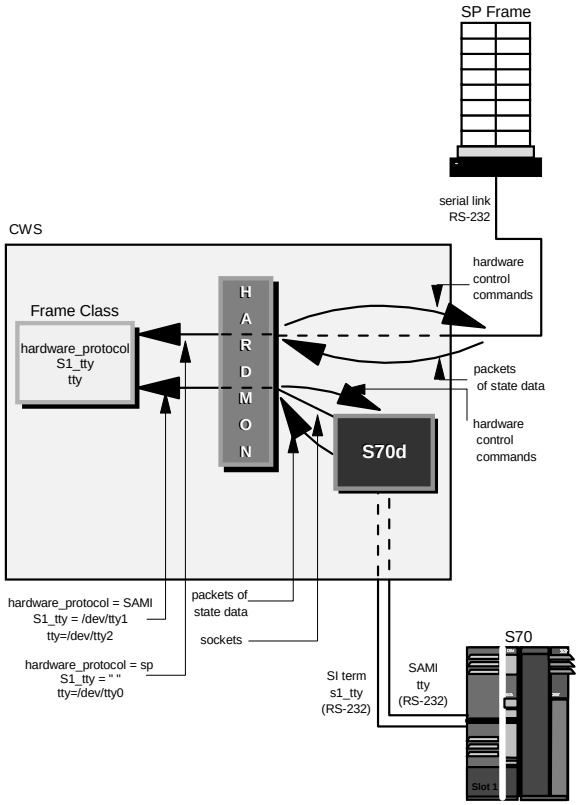


Figure 85. Hardmon Flow of Control

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 SAMI/MI operator panel port, and `/dev/tty1` 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.

Dataflow

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 frames' 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 node.

The node supervisor interface polls state data from the S70/S7A 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 software protocol and sends the command through to the S70/S7A 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 86 on page 160.

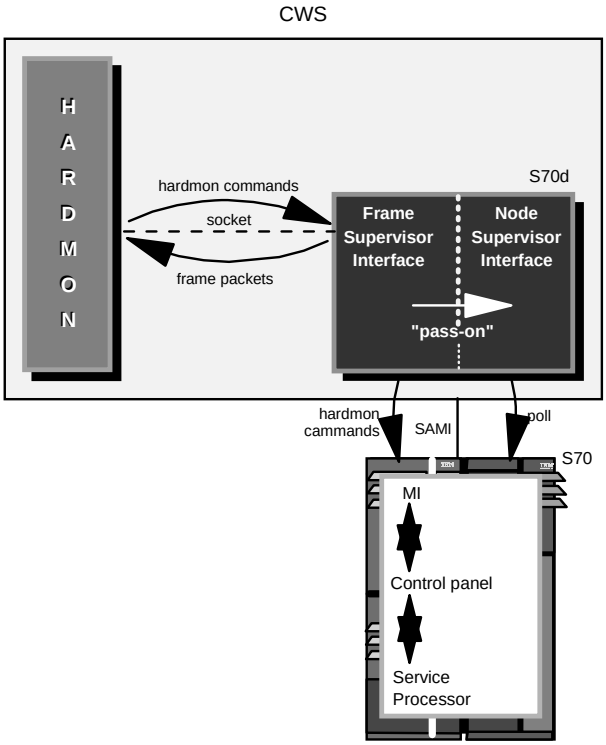


Figure 86. S70 Daemon Internal Flow

The S70 daemon uses 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 will indicate the hardware capabilities for monitoring. This relationship is illustrated in Figure 84 on page 156.

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 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 as are all the nodes in the affected system partition.

5.5 User Interfaces

This section highlights the changes in the different user interface panels and commands that have been made to represent the SP-attached server to the user.

5.5.1 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.

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 Perspective representation of the SP-attached server is shown in Figure 87 on page 162.

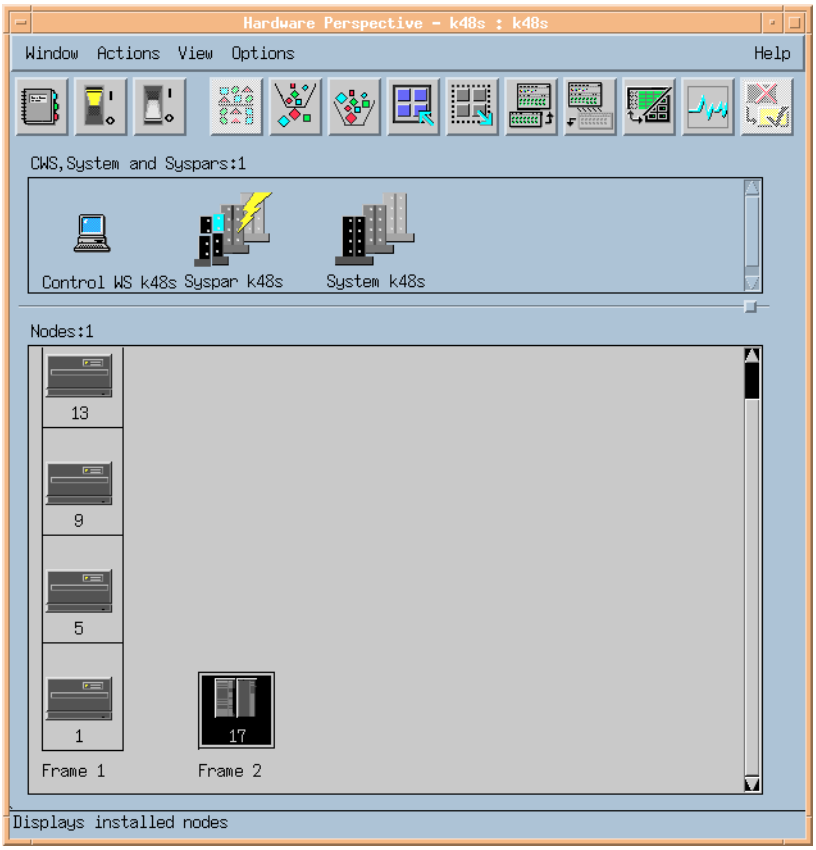


Figure 87. Example of Perspectives with SP-Attached Server

The monitored resource variables are handled the same as for standard SP nodes. Operations, status, frame, and node information are handled the same 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 off 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.

5.5.1.1 System Management

The various system management commands that display new SDR attributes for SP-attached servers are:

- `spmon`

  Figure 89 on page 165 outlines the `spmon -d -G` output in an SP system that consists of an SP Frame and an SP-attached server.

```
1.  Checking server process
    Process 11454 has accumulated 9 minutes and 27 seconds.
    Check ok

2.  Opening connection to server
    Connection opened
    Check ok

3.  Querying frame(s)
    2 frame(s)
    Check ok

4.  Checking frames

        Controller   Slot 17  Switch   Switch     Power supplies
Frame   Responds     Switch   Power    Clocking   A   B   C   D
-------------------------------------------------------------
  1       yes          no      N/A       N/A      on  N/A N/A N/A
  2       yes          no      N/A       N/A      N/A N/A N/A N/A

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  no      normal   no   LCDs are blank   no
  5       5    high    on   yes  no      normal   no   LCDs are blank   no
  9       9    high    on   yes  no      normal   no   LCDs are blank   no
 13      13    high    on   yes  no      normal   no   LCDs 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    extrn   on    no  no      normal no no no
                                                                  LCD2 is blank
```

Figure 88. The Output of the spmon Command

- splstdata

  Figure 89 on page 165 is the output of `splstdata -n`. It shows two frames. Figure 90 on page 165 shows the output from `splstdata -f` where the S70 is shown as a second frame. Figure 90 on page 165 shows the hardware description of each node in the SP system.

- 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 2, with one node of node_number 17 occupying one slot.

.

```
List Node Configuration Information

node# frame# slot# slots  initial_hostname  reliable_hostname  dcehostname
      default_route   processor_type processors_installed description
-------------------------------------------------------------------------
    1      1     1     4  c60n01.ppd.pok.i   c60n01.ppd.pok.i  ""
           9.114.88.94            MP                   4 112_MHz_SMP_High
    5      1     5     4  c60n05.ppd.pok.i   c60n05.ppd.pok.i  ""
           9.114.88.94            MP                   4 75_MHz_SMP_High
    9      1     9     4  c60n09.ppd.pok.i   c60n09.ppd.pok.i  ""
           9.114.88.94            MP                   4 75_MHz_SMP_High
   13      1    13     4  c60n13.ppd.pok.i   c60n13.ppd.pok.i  ""
           9.114.88.94            MP                   4 112_MHz_SMP_High
   17      2     1     1  c60tp1n02.ppd.po   c60tp1n02.ppd.po  ""
           9.114.88.1             MP                   1 ""
```

Figure 89. splstdata -n Output

Figure 90 is the output of `splstdata -f`, which shows two frames:

```
 List Frame Database Information

frame#             tty           s1_tty      frame_type  hardware_protocol
-------------------------------------------------------------------------
    1        /dev/tty0              ""          switch               SP
    2        /dev/tty1       /dev/tty2             ""             SAMI
```

Figure 90. splstdata -f Output

Figure 91 on page 166 is the output of `spgetdesc -u -a`, which shows the hardware description obtained from the Node class.

```
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.
```

Figure 91. spgetdesc -u -a Output

5.6 Attachment Scenarios

The following sections describe the different 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-attach 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 RS/6000 SP Attachment adapter.

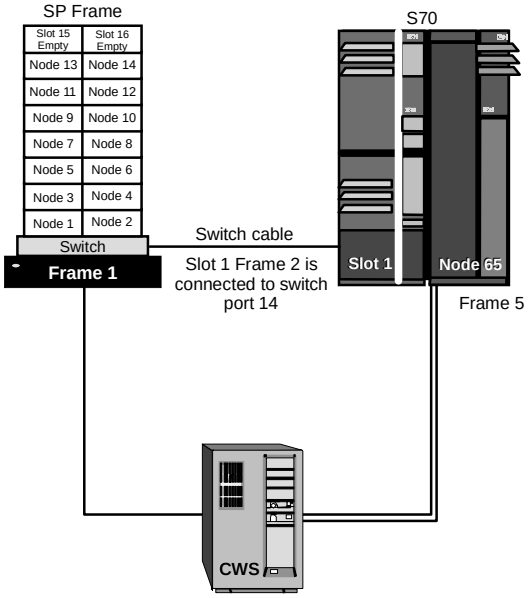


Figure 92. Scenario 1: SP-Attached Server and One SP Frame

Scenario 2: SP-Attached Server to a Two-Frame SP System

This scenario shows a two-frame system with four high nodes in each frame. This configuration will use eight switch ports and leave 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 five to allow for the two possible non-switch frames.

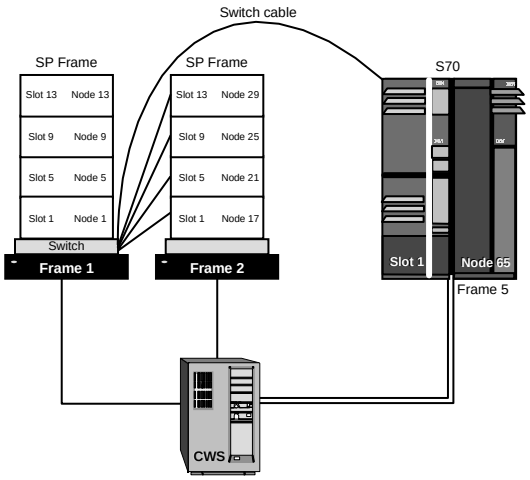


Figure 93. Scenario 2: SP-Attached Server to Two SP Frames

Note that the switch cable from frame one connects to the S70; for example, in this case, slot one frame five connects to switch port three of switch chip five.

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 as 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.

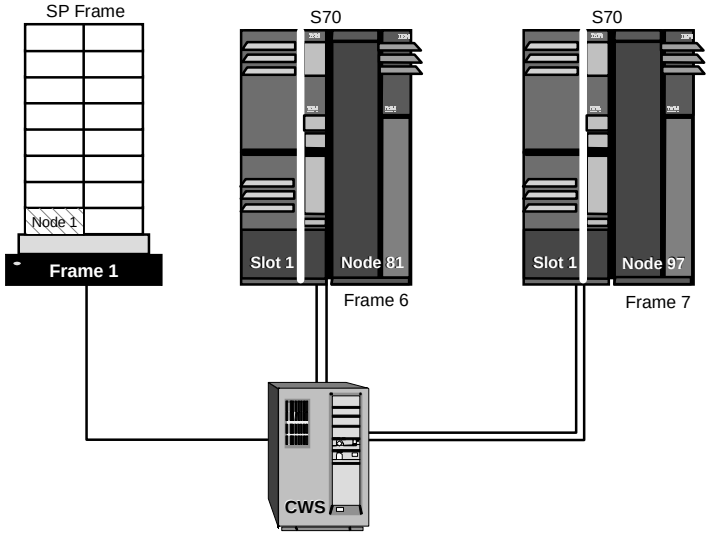


Figure 94. Scenario 3: SP Frame and Multiple SP-Attached Servers

Scenario 4: Non-Contiguous SP-Attached Server Configuration
Frame one and three of the SP system are switch-configured. Frame two is a non-switched expansion frame attached to frame one. In this configuration, the SP-attached server could be given frame number four, but that would forbid any future attachment of nonswitched expansion frames to frame one'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 nonswitched expansion frames.

Frame three is another switch-configured frame, and the SP-attached server has previously been assigned frame number 10 for future scalability purposes.

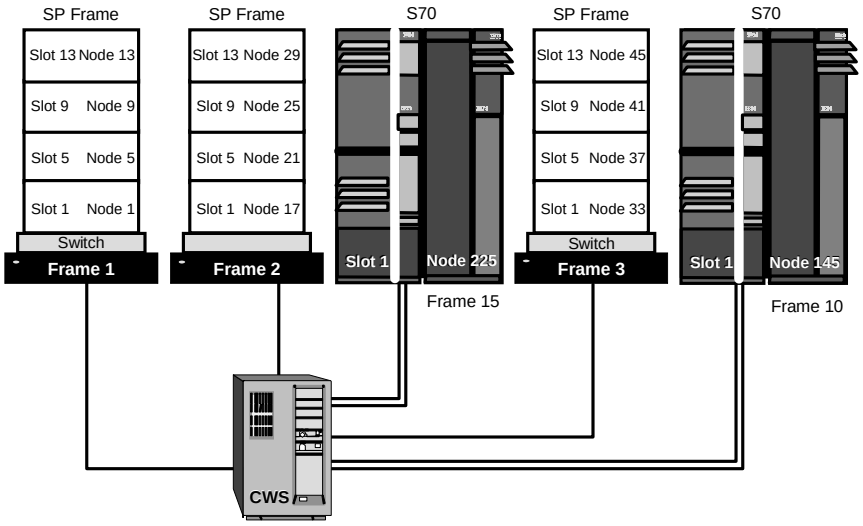


Figure 95. Scenario 4: Non-Contiguous SP-Attached Server

For more information see: *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* GA22-7281*.*

5.7 Related Documentation

These documents will help you to 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 *IBM 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.

5.8 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. There must be three connections between the control workstation 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 is 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 is 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 control workstation.

   C. The control workstation should have enough spare serial ports to connect the SP-attached server. Additional 16-port adapter 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 restrict installation of SP-attached servers to one at the 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 is false?

   A. The s70d is partition-sensitive; so, it will be one s70d daemon per SP-attached server per partition running on the control workstation.

   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 control workstation.

Chapter 6. SP Security

This chapter covers security facilities available to the SP system environment.

Three common security concepts on systems are defined in terms of identification, authentication, and authorization.

Emphasis is placed on Kerberos, which is a security service included in PSSP. A definition on Kerberos, how Kerberos authentication service works, how to set up Kerberos, and how to manage Kerberos are discussed.

Two other Kerberos-based security systems that may be used on the SP system are also discussed. These are AFS authentication and sysctl authorization services on the SP.

6.1 Key Concepts You Should Study

The key concepts of security on the SP system are listed below in order of importance.

- Concepts of using Kerberos for authentication services on the SP system. These include client/server activities, principals, realms, and tickets.
- Procedures in managing Kerberos that covers adding and deleting principals and authentication administrators.
- Concepts in AFS authentication management and its usage of a different set of protocols, utilities, daemons, and interfaces to manage the principal database.
- Concepts of sysctl as an SP 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.

Recommended reading can be found in 6.15, “Related Documentation” on page 202.

6.2 Security-Related Concepts

There are three security-related concepts. The following is a brief description of the three concepts and how they may be applied to the SP system environment.

1. Identification: This is a process by which one entity tells another who it is, that is, its identity. In the SP system environment, identification simply means a process that presents client identity credentials.
2. 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 SP system environment, authentication simply means a service requester's name and encrypted password are checked with the usage of available system utilities.
3. 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.

In an SP system, there are at least two levels of security: AIX and PSSP.

Kerberos, which comes in bundled with PSSP, has been entrusted to perform authentication on SP environments.

6.3 AIX Security

AIX provides the basic security elements to control user access to files, directories, and networks. Details of AIX security can be obtained in the redbook named *Elements of Security: AIX 4.1*, GG24-4433.

However, a machine may be programmed to send information across the network impersonating another machine (which means assuming the identity of another machine or another user). One way to protect the machines and users from being impersonated is to authenticate the packets when travelling within the network. Kerberos, which is included in PSSP, can provide such authentication services to the SP system environment.

6.3.1 Secure Remote Execution Commands

In AIX 4.3.1, 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:

`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.

The main security concern with this authentication for the above commands is the fact that passwords are sent in plain text over the network. They can be easily captured by any root user on a machine that is on the network(s) through which the connection is established.

`rcp`**,** `rlogin`**, and** `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, which are accepted for incoming connections. This is on a per-user basis and also works for the root user.

Here, the 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 the system. Of course, if a password is requested rather than using $HOME/.rhosts or /etc/hosts.equiv files, this is also normally sent in plain text.

With AIX v4.3.1, all these commands except `rexec` also support Kerberos Version 5 authentication. The base AIX operating system does not include Kerberos. It is recommended that DCE for AIX Version 2.2 is used to provide Kerberos authentication. Note that previous versions of DCE did not make the Kerberos services available externally. However, DCE for AIX Version 2.2, which is based on OSF DCE Version 1.2.2, provides the complete Kerberos functionality as specified in RFC 1510, *The Kerberos Network Authentication Service (V5)*.

For backward compatibility with PSSP 3.1 (which still requires Kerberos Version 4 for its own commands), the AIX r-commands `rcp` and `rsh` also support Kerberos Version 4 authentication. See 6.5, “How Kerberos Works” on page 179 for details on Kerberos.

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 a new library, libauthm.a. Three options are available: `chauthent -std` enables standard AIX authentication; `chauthent -k5` and `-k4` enable Version 5 or 4 Kerberos authentication. More than one method can be specified, and authenticated applications/commands will 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

On the SP, the `chauthent` command should not be used directly. The authentication methods for SP nodes and the control workstation 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.

If Kerberos Version 5 is activated as an authentication method, the `telnet` connection is secured by an optional part of the telnet protocol specified in RFC 1416, *Telnet Authentication Option*. Through this mechanism, clients and servers can negotiate the authentication method. The `ftp` command uses another mechanism: Here the authentication between client and server takes place through a protocol specified in RFC 1508, *Generic Security Service API*. These extensions are useful but have no direct relation to SP system administration. An impediment to widespread use of these facilities is that they rely on all clients being known to the Kerberos database including the clients’ secret passwords.

The kerberized rsh and rcp commands are of particular importance for the SP as they replace the corresponding Kerberos Version 4 authenticated r-commands which have been part of PSSP Versions 1 and 2. Only the PSSP versions of rsh, rcp, and kshd have been removed from PSSP v3.1. It still includes and uses the Kerberos Version 4 server. This Kerberos server can be used to authenticate the AIX r-commands. A full description of the operation of the rsh command in the SP environment can be found in 6.12.2, “Remote Execution Commands” on page 192 including all three possible authentication methods.

6.4 Defining Kerberos

Kerberos can be used to prevent machine impersonation by means of authenticating the packets in a two-party communication.

Kerberos is a service for authenticating users in a network environment. It consists of a set of distributed software with encrypted exchanges of information to allow a user access to servers. It also provides for cryptographic checks to make sure that data passing between workstations and servers is not corrupted either by accident or by tampering.

6.4.1 AFS and Sysctl Are Kerberos-Based Security Systems

Both AFS and sysctl are Kerbereos-based security system that may be used on the SP system.

AFS is a distributed file system. Since AFS includes Kerberos Version 4, SP systems may use an AFS Kerberos authentication server instead of SP servers. However, AFS uses a different set of protocols, utilities, daemons, and interfaces for principal database administration.

Sysctl is an SP Kerberos-based client/server system designed to run commands remotely and in a parallel fashion with a high degree of authentication.

Further descriptions of the AFS and Sysctl security systems are included in the later parts of this chapter.

6.4.2 Main Reasons for Using Kerberos on the SP

The main reasons are:

- To prevent unauthorized access to the system.
- Prevents non-encrypted passwords from being passed on the network.

- Provides security on remote commands, such as `rsh`, `rcp`, `dsh`, and `sysctl`. Description of these commands are in the following table.

Table 9. Some Kerberos Authenticated Commands

Commands	Description
`spmon`	Controls and monitors SP system activity through the hardware monitor, `hardmon`. Replaced by Perspectives on PSSP V3.1.
`rsh`	`rsh` is the remote shell command. On PSSP V3.1 this command is no longer SP provided. The `/usr/lpp/ssp/rcmd/bin/rsh` command is linked to the Berkeley command `/usr/bin/rsh` (which uses.rhosts file).
`rcp`	`rcp` is remote copying of files between local and remote hosts. On PSSP V3.1, this command is no longer SP provided. The `/usr/lpp/ssp/rcmd/bin/rcp` command is now linked to the Berkeley command `/usr/bin/rcp`.
dsh	Can be issued to groups of SP nodes at the same time. For example, `dsh -w sp3n05 sp3n06`. dsh is not interactive. Therefore, `telnet`, `rcp`, `rsh`, and so forth, may be used.
`sysctl`	Uses the SP authentication service. When the client issues the `sysctl` command, a Kerberos ticket will be sent to the server to validate the identity of the client.

6.4.3 Kerberos Terms

The following table consists of basic Kerberos Terms.

Table 10. 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

Basic Kerberos Terms	Description
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 covered 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 and 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.

6.5 How Kerberos Works

Kerberos authenticates information exchange over a network, and there are three daemons that deal with the Kerberos services. 6.5.2, "Kerberos Authentication Process" on page 180 illustrates the Kerberos authentication process.

6.5.1 Kerberos Daemons

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, kprop, which runs on the primary server.

6.5.2 Kerberos Authentication Process

Three entities are involved in the Kerberos authentication process: The client, server, and the authentication database server. The following is an example of authentication:

1. The client (Host A) issues the `kinit` command that requests for a ticket-granting ticket (TGT) to perform the `rcp` command on the destination host (Host B).

   For example: issue command lines `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 information is valid, then it will issue 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 will then be sent to the client (Host A).

6.6 Kerberos Paths, Directories, and Files

The location of Kerberos directories and files are in the following paths.

For PSSP 2.4:

```
PATH=/usr/lpp/ssp/rcmd/bin:$PATH:/usr/lpp/ssp/bin:/usr/lpp/ssp/kerberos/bi
n:/usr/lpp/ssp/kerberos/etc
```

For PSSP V3.1:

```
PATH=$PATH:/usr/lpp/ssp/bin:/usr/lpp/ssp/kerberos/bin:/usr/lpp/ssp/kerbero
s/etc
MANPATH=$MANPATH:/usr/lpp/ssp/man:/etc
```

Table 11 displays the Kerberos directories and files on the Primary Authentication Server, which is usually the control workstation (CWS).

Table 11. 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 /.k file to store the new master key in the /.k file.
$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

Directories and Files	Description
/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.
/var/kerberos/database/*	This directory includes the authentication database created by setup_authent. Files residing in this directory 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/kerboros.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>

6.7 Authentication Services Procedures

This section gives an overview of 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. Perform login to SP Kerberos authentication services.
   - Use the command `k4init <principal>` 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.

6.8 Kerberos Passwords and Master Key

Initial setup of passwords on the primary authenticator server:

During the installation stage, the `setup_authent` command is entered to configure the SP authentication services on the control workstation (CWS) and other RS/6000 workstations connected to the SP system. The `setup_authent` command gives and 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

Change a principal's password:

Enter the `kpasswd` command to change a Kerberos principal's password. For example, to change the password of current user, use the `kpasswd` command.

Change Kerberos master password:

1. Login to Kerberos as 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 by entering 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
```

6.9 Kerberos Principals

Kerberos principals are either users who use authentication services to run the Kerberos-authenticated applications supplied with the SP system or the individual instances of the servers that run on SP nodes, the control workstation, and on other IBM RS/6000 workstations that have network connections to the SP system.

- User Principals for SP System Management

  An implementation of the SP 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 SP 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.

  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:

  1. `hardmon` used by the System Monitor daemon on the control workstation by logging daemons.
  2. `rcmd` used by sysctl.

  The hardmon daemon runs only on the control workstation. The SP logging daemon, splogd, can run on other IBM RS/6000 workstations. Therefore, for each (short) network interface name on these workstations, a service principal is created with the name hardmon and the network name as the instance. The remote commands can be run from, or to, any IBM RS/6000 host on which the SP system authenticated client services (ssp.clients) are installed. Therefore, for each (short) network interface name on all SP nodes, the control workstation, and other client systems, a

service principal is created with the name rcmd and the network name as the instance.

6.9.1 Add a Kerberos Principal

It is desirable to allow users to perform certain system tasks. Such users must be set up as Kerberos principals. These users 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
     ```

6.9.2 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 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 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. Instead of not finding the specified principal, the command finds it already exists and prompts for changes to all its attributes starting with the password followed by the expiration date and 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).

6.9.3 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 through 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 to

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`

6.10 Server Key

The server keys are located in the /etc/krb-srvtab file on the control workstation (CWS) and all the nodes. The file is used to unlock (decrypt) tickets coming in from clients authentication.

- On the CWS hardmon and rcmd, service principals are in the file.
- On 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.
- `setup_server` script creates server key files for nodes and stores them in /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 local key file /etc/krb-srvtab, use:

  ```
  ksrvutil list

  k4list -srvtab

  krsvutil list -f /tftpboot/sp31n1-new-srvtab
  ```

6.10.1 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.

6.11 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.

6.11.1 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 prompt for a secondary server. (Note: It will also prompt you to login as the same administrative principal name as that was 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.

6.11.2 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.

6.12 SP Services That Utilize Kerberos

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. This section describes the authentication of theses services and the different means they use to authorize clients that have been successfully authenticated.

6.12.1 Hardware Control Subsystem

The SP hardware control subsystem is implemented through the hardmon and splogd daemons, which run on the control workstation and interface with the SP hardware through the serial lines. To secure access to the hardware, Kerberos authentication is used, and authorization is controlled through hardmon-specific Access Control Lists (ACLs). PSSP v3.1 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 a 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 control workstation including IP aliases in case of

multiple partitions. The secret keys of these hardmon principals are stored in the /etc/krb-srvtab file of the control workstation. The `k4list -srvtab` command shows the existence of these service keys.

```
# k4list -srvtab
Server key file:   /etc/krb-srvtab
Service         Instance        Realm      Key Version
-------------------------------------------------------
hardmon         sp4cw0          MSC.ITSO.IBM.COM 1
rcmd            sp4cw0          MSC.ITSO.IBM.COM 1
hardmon         sp4en0          MSC.ITSO.IBM.COM 1
rcmd            sp4en0          MSC.ITSO.IBM.COM 1
```

The above client commands performs a Kerberos Version 4 authentication: They require that the user who invokes them has signed on to Kerberos by 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, which is read by hardmon when it starts up. Since hardmon runs only on the control workstation, this authorization file also only exists on the control workstation.

```
> cat /spdata/sys1/spmon/hmacls
sp4en0 root.admin a
sp4en0 hardmon.sp4en0 a
1 root.admin vsm
1 hardmon.sp4en0 vsm
2 root.admin vsm
2 hardmon.sp4en0 vsm
```

Each line in that 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 control workstation 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:

- m (Monitor) - monitor hardware status
- v (Virtual Front Operator Panel) - control/change hardware status
- s (S1) - access to node's console through the serial port (s1term)

- a (Administrative) - use hardmon administrative commands

Note, that for the control workstation, 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 control workstation 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 re-read the hmacls file.

Care must been taken if any of the hardware monitoring/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, for example:

```
hmmon: 0026-614 You do not have authorization to access the Hardware Monitor.
```

In other cases, no, or 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.

6.12.2 Remote Execution Commands

In releases prior to PSSP v3.1, PSSP provided its own remote execution commands and the corresponding krshd daemon that were Kerberos Version 4 authenticated. These were located in /usr/lpp/ssp/rcmd/. All the SP management commands used the PSSP version of `rsh` and `rcp`, and AIX provided the original r-commands in /usr/bin/. This is shown in Figure 96 on page 193.

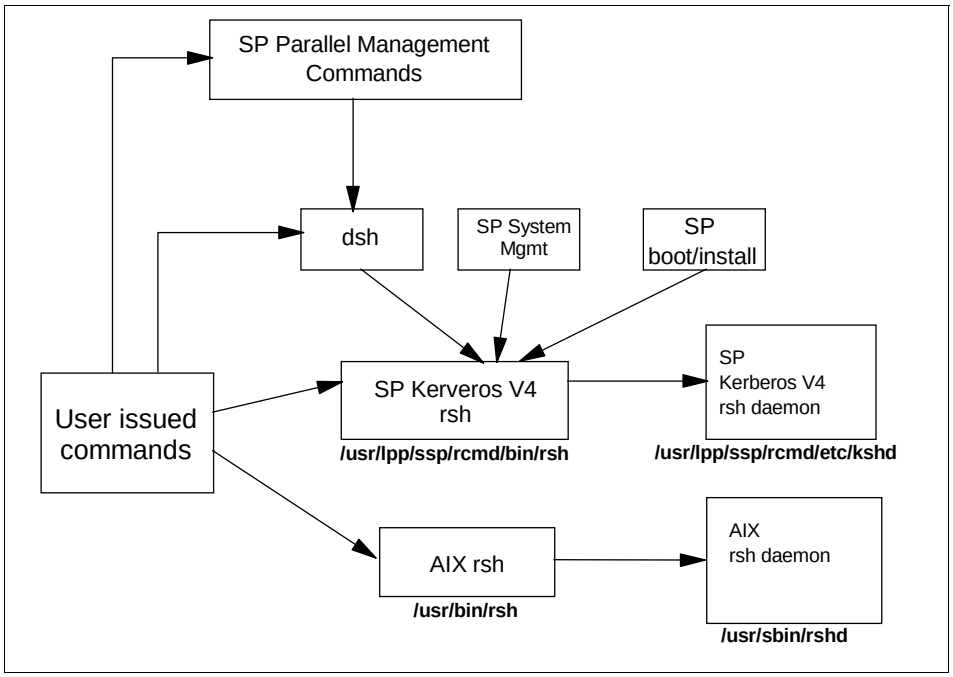


Figure 96. Remote Shell Structure before PSSP 3.1

In PSSP v3.1, the authenticated r-commands in the base AIX 4.3.2 operating system are used instead. 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 simply removed but have been replaced by links to the corresponding AIX commands. This new calling structure is shown in Figure 97 on page 194.

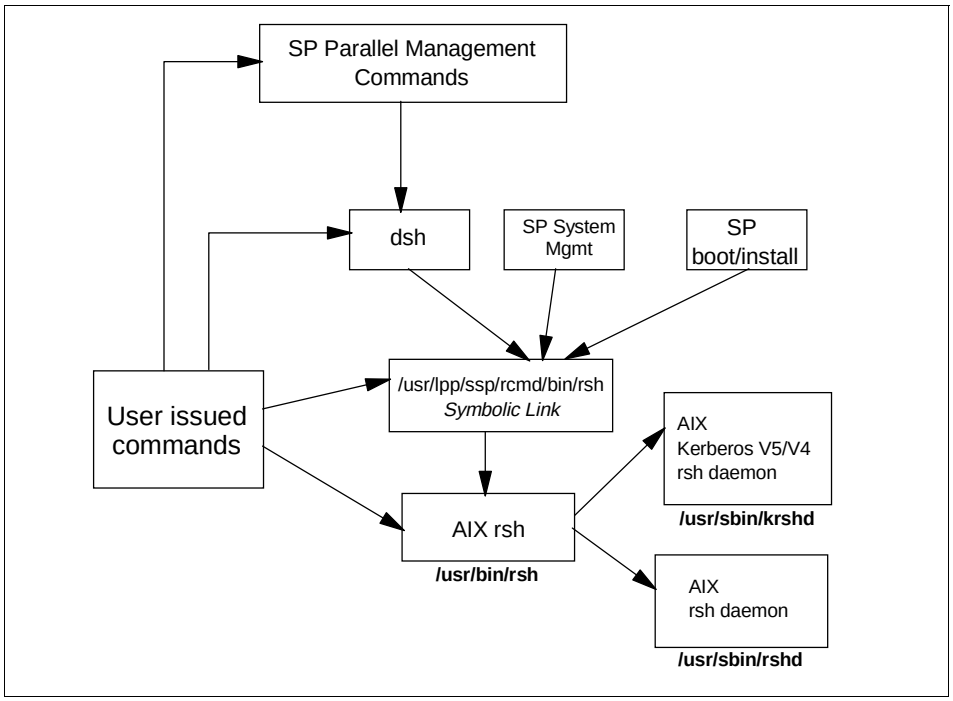


Figure 97. Remote Shell Structure in PSSP 3.1

For the user, this switchover from PSSP-provided authenticated r-commands to the AIX-provided authenticated r-commands should be transparent. In the remainder of this section, 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.

6.12.2.1 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.

Note: K5MUTE

Authentication methods are set on a system level not on a user level. This means that, for example, on an SP where Kerberos Version 4 and Standard AIX is set, a user's `rsh` command will produce a Kerberos authentication failure if that user has no Kerberos credentials (which is normally the case unless the user is an SP system administrator). After that failure, the `rsh` attempts to use the standard AIX methods. The delay caused by attempting both methods can not 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 though.

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, will be invoked by the AIX `rsh` command. It will acquire 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 are not forwardable.

These requests are then processed by the rshd and krshd daemons.

6.12.2.2 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 the *AIX Version 4.3 Commands Reference Volumes*, SC23-4119. Notably, the rshd daemon:

- 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 if the client host is listed in /etc/hosts.equiv.
- If the pervious check is negative, rshd checks if 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 will honor requests even though `lsauthent` still reports that Standard AIX authentication is disabled.

6.12.2.3 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 if 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 completely reverse to 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 both the kvalid_user() man page and the *AIX Version 4.3 System Managment Guide: Communications and Networks*, SC23-4127.

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 that 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.

Note: 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 so 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 so 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. For this reason, Kerberos Version 4 authentication is only available on SP systems not on normal RS/6000 machines.

Note: rcmdtgt

PSPP 3.1 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.

6.12.2.4 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 control workstation to update status information about the installation process that is stored on the control workstation. 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 control workstation and removes this entry after network installation.

6.13 AFS as an SP Kerberos-Based Security System

PSSP supports the use of an existing AFS server to provide Kerberos Version 4 services to the SP. 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 SP's authentication environment, supports AFS as the underlying Kerberos server. This is mainly contained in its `setup_afs_server` sub-command.

PSSP: Installation and Migration Guide, GA22-7347 explains the steps which are required to initially set up SP security 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 SP systems is optional.

6.13.1 Set Up to Use AFS Authentication Server

- When running the `setup_authent` command, ensure 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.
- AFS files ThisCell and CellServDB should be in /usr/vice/etc, or a symbolic link created.
- `kas` command located 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.

6.13.2 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 12, contains some commands that may be used for managing AFS.

Table 12. Some Commands for Managing AFS

Commands	Description
kas	For adding, listing, deleting, and changing the AFS principal's attributes. kas has corresponding *subcommands*, which are 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.
token.krb (AFS commands)	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).

6.14 Sysctl Is an SP 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 SP systems is optional.

6.14.1 Sysctl Components

The server daemon for sysctl server is *sysctld*. The sysctl server also contains built-in commands, configuration files, access control lists (ACL), and client programs.

Figure 98 shows the sysctl architecture.

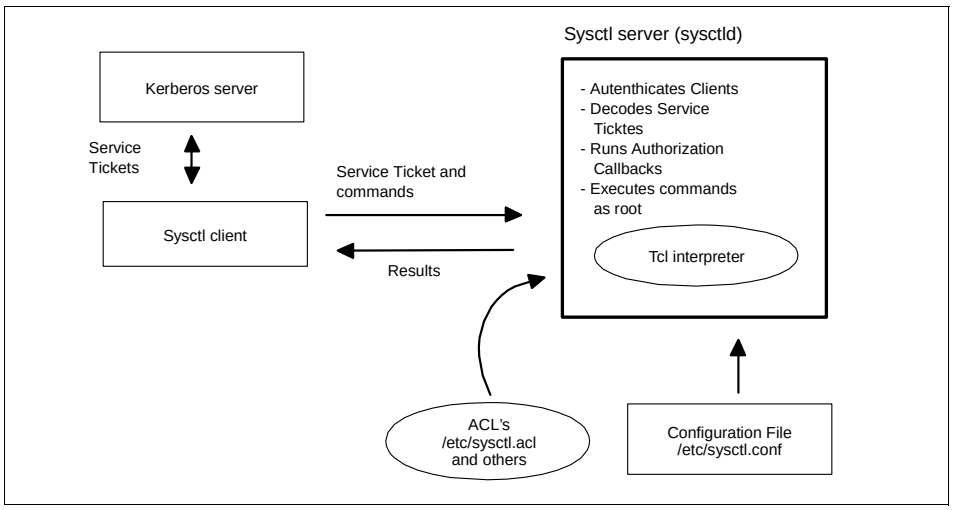


Figure 98. Sysctl Architecture

6.14.2 Sysctl Process

The following is the sysctl process.

1. The sysctl client code gets its authentication information from SP 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.
- Decode service ticket.
- Performs an authorization callback.
- Executes commands as root.

6.14.3 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:
    1. The /etc/sysctl.conf file that 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.
    2. 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. Tcl-based set of commands can be separated in the following three classes.
    1. Base `Tcl` commands: These are the basic Tcl interpreter commands. They are also defined in the /etc/sysctl.conf file.
    2. 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.
    3. User-written scripts: These are programmer written applications that use the base `Tcl` commands and built-in `sysctl` commands.

6.15 Related Documentation

The following documentation list consists of books that provide further explanation on the key concepts discussed in this chapter.

SP Manuals

PSSP: Administration Guide, SA22-7348. Two Chapters are on security on the SP system. Chapter 12 concentrates on security features of the SP system that includes conceptual information regarding Kerberos Version 4. Chapter 13 is on sysctl, which covers its relationship to the SP system.

SP Redbooks

Inside the RS/6000 SP, SG24-5145. Section 4.6 of Chapter 4 is on SP Security. It gives a good overview of Kerberos, AFS, and sysctl.

RS/6000 Scalable POWERparallel Systems, SG24-4542. Part 5 is solely on Kerberos and contains Chapters 12-19. They give details on Kerberos, which covers Kerberos secure authentication, Kerberos authentication protocols, installing Kerberos primary and secondary servers, how to implement Kerberos on the SP, and description of a list of Kerberos files.

Study Guides

IBM Global Services, RS/6000 SP, System Administration: Course Code AU96. (Unit 1 is on managing Kerberos authentication in the SP environment.) This book covers what Kerberos is used for on the SP, how to manage Kerberos principal authentication, how to keep Kerberos secure, and considerations on authentication server backup and recovery. Unit 5 is on working with the sysctl security system. Appendix C covers an overview of AFS authentication.

6.16 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. On the SP, the AIX command `chauthent` should not be used directly because:

   A. It is not supported on RS/6000 SP environments.

   B. 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. PSSP requires Kerberos v4 because some components still use this Kerberos level. These components are:

   A. hardmon and the nodecond scripts.

B. The partition-sensitive daemons and file collection.

C. hardmon and NIM.

D. hardmon and sysctl.

3. 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.

4. 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).

Chapter 7. User and Data Management

This chapter covers user management that consists of adding, changing, and deleting users on the SP system and how to control user login access.

Data and user management using the file collections facility is also discussed. File collection provides the ability to have a single point of update and control of file images that will then be replicated across nodes.

AMD and AIX Automounters are also discussed. These allows users local access to any files and directories no matter which node they are logged in to.

7.1 Key Concepts You Should Study

The key concepts on user and data management are listed below in order of importance.

- Considerations for administering SP users and SP user access control and procedures to perform them.
- File collections and how it works in data management in the SP system.
- How to work with and manage file collections and procedures to build and install file collections.
- The concepts of AMD and AIX Automounter and how they manage mounting and unmounting activities using NFS facilities.
- AMD to AIX Automounter migration and the main differences between the two.

7.2 Issues on Administering Users on the SP System

Table 13 consists of the issues and solutions on user and data management. You need to consider them when installing an SP system.

Table 13. Issues and Solutions when Installing an SP 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

Issues	Solutions
WIthin a single user space, how to restrict access to individual node?	Login Control
Where should user's home directories reside?	Control Workstation (CWS) Nodes Other Network System
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 SP system

7.3 SP User Data Management

The following three options may be used to manage the user data on the SP:

- SP User Management (SPUM).
- Network Information System (NIS).
- Manage each user individually over each machine on the network.

The first two are more commonly used and are discussed in this chapter.

7.3.1 SP User Management (SPUM)

The following information is covered by this chapter:

- How to set up SP User management?
- How to add/change/delete/list SP users?
- How to change SP user passwords?
- SP user login and access control.

7.3.2 Setup SP User Management

1. Enter `smit site_env_dialog`. The output is shown in Figure 99 on page 207.

```
                        Site Environment Information

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

                                                      [Entry Fields]
  Default Network Install Image                      [bos.obj.ssp.432]
  Remove Install Image after Installs                 false                  +

  NTP Installation                                    consensus              +
  NTP Server Hostname(s)                             [""]
  NTP Version                                         3                      +

  Automounter Configuration                           true                   +

  Print Management Configuration                      false                  +
  Print system secure mode login name                [""]

  User Administration Interface                       true                   +
  Password File Server Hostname                      [sp3en0]
  Password File                                      [/etc/passwd]
  Home Directory Server Hostname                     [sp3en0]
  Home Directory Path                                [/home/sp3en0]

  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                [aix432]




F1=Help            F2=Refresh         F3=Cancel          F4=List
F5=Reset           F6=Command         F7=Edit            F8=Image
F9=Shell           F10=Exit           Enter=Do
```

Figure 99. Setup SP User Management

2. Activate SPUM by setting the following fields to `true`.
   - Automounter Configuration
   - User Administration Interface
   - File Collection Management

7.3.3 Add/Change/Delete/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 *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*, GA22-7347 contains detailed instructions for entering site environment information.

The following are the steps in adding an SP User by entering `smit spmkuser`:

- Check /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 is stored in the /usr/lpp/ssp/config/admin/newpass.log file.

Figure 100 shows the output screen for changing the characteristics of an SP user. All value fields can be changed except the `name` field. Nodes will pull the SPUM file collection from the CWS and update its configuration.

```
                *Change/Show Characteristics of a User

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

                                                      [Entry Fields]
* User NAME                                           spuser1
  User ID                                            [218]                    #
  LOGIN user?                                                                 +
  PRIMARY group                                      [1]                      +
  Secondary GROUPS                                   [staff]                  +
  HOME directory                                     [sp3en0:/home/sp5en0/sp>
  Initial PROGRAM                                    [/bin/ksh]                /
  User INFORMATION                                   []









F1=Help            F2=Refresh         F3=Cancel          F4=List
F5=Reset           F6=Command         F7=Edit            F8=Image
F9=Shell           F10=Exit           Enter=Do
```

Figure 100. Changing the Characteristics of an SP User

Figure 101 shows the output screen for removing an SP user with the `smit sprmuser` command. Both authentication information and home directory may be removed. When deleting a user, the entry for that user in the newpass.log file doesn't get removed.

```
                          Remove a User

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

                                                      [Entry Fields]
  Remove AUTHENTICATION information?                  No                      +
  Remove HOME directory?                              No                      +

* User NAME                                           spuser1
  User ID                                             218
  PRIMARY group                                       1
  Secondary GROUPS                                    staff
  HOME directory                                      /u/spuser1 on sp3en0(/>
  Initial PROGRAM                                     /bin/ksh
  User INFORMATION




F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 101. Removing an SP User

The following example shows how to list SP users with the `spluser` command:

```
spluser spuser1
```

The output will appear as the following:

```
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 control workstation (CWS).
- Use the `passwd` command to change the password.
- /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 control workstation (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`, it 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 in 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 an SP is a machine containing multiple RS/6000 nodes, you do not want to maintain an SP 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 the user only needs to maintain one password. When the user's home directory is maintained on one machine and made available through NFS, the user's environment is also easier to maintain.

From an SP point of view, an NIS solution removes the SP restriction of changing user's passwords on the control workstation. When you would 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 SP from any given node.

7.4.1 Setting UP NIS

You can using 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.

7.4.1.1 Configuring a Master Server

This 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 to 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 filename 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.

7.4.1.2 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 a lot of 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.

7.4.1.3 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.

7.4.1.4 Change NIS password

You may change an NIS user password with the `passwd` or `yppasswd` commands.

7.5 File Collections

The SP system also has another tool that ensures that system configuration files look the same throughout your SP network. 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 SP 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 SP 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 for File Collections.

7.5.1.1 Terms Used when Defining File Collections

- *Resident*: A file collection that is installed in its true location and able to be served to other systems.
- *Available*: A file collection that is not installed in its true location but able to be served to other systems

7.5.1.2 Unique Features on File Collections

The following are the unique features on 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.

- `supper` command interprets the Master Files:

  You handle files in a collection with special procedures and the `supper` commands rather than with the standard AIX file commands. The `supper` commands interpret the Master Files and use 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, 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 list 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 SP 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 served out to the nodes. Secondary file collections are served 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/installl server, and all of the files that have been defined within that file collection will be installed on that boot/installl 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 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.

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 SP or when partitioning the system into groups.

7.5.3 Predefined File Collections

On the SP, 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
```

7.5.3.1 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.

7.5.3.2 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

7.5.3.3 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.

7.5.3.4 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 102 on page 218 shows the structure of /var/sysman/sup directory, which consists of the master files for a file collection.

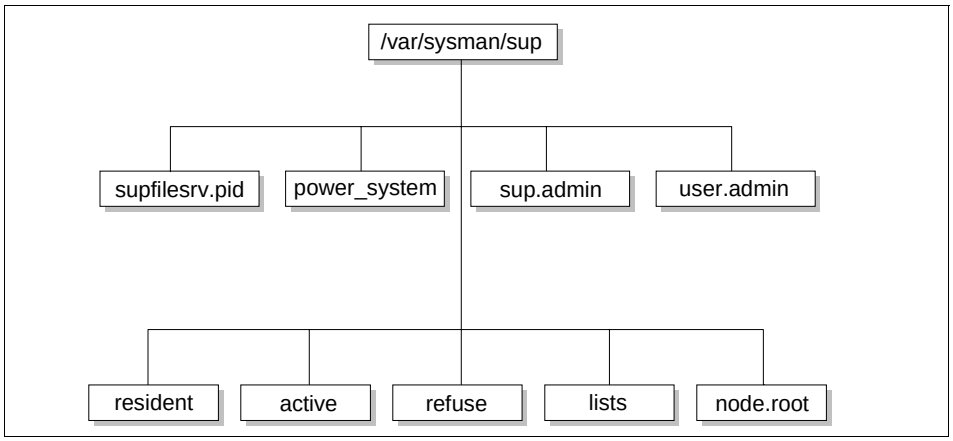


Figure 102. /var/sysman/sup Files and Directories

The following provides an explanation on these files and directories.

.active: Identifies the active volume group. It is not found on the control workstation.

.resident: Lists each file collection in the SP 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 103 on page 219. It shows the structure of the /var/sysman/sup/sup.admin file collection.

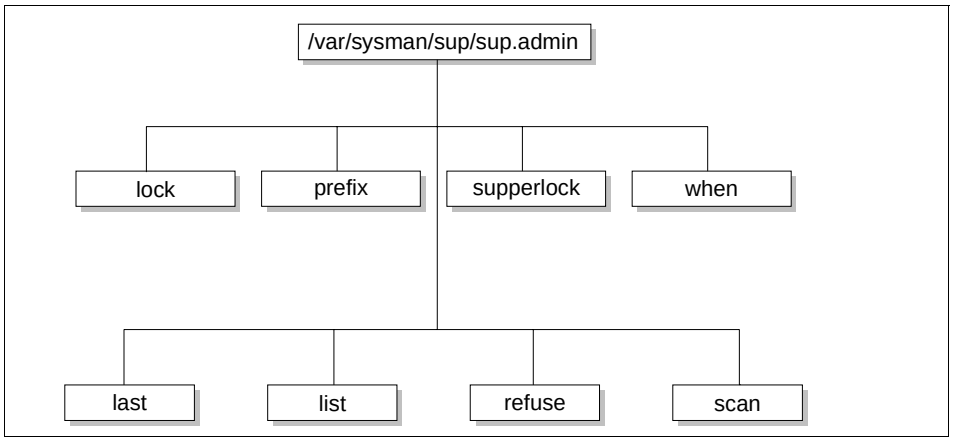


Figure 103. sup.admin Master Files

The following provides an explanation on these files.

last: Consists of a list of files and directories that have been updated.

list: Consists of a list of files that is 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: Set 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 volume.

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/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 command `/var/sysman/super update user.admin` on each node. This can also be performed remotely through 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.
- 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

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 explanation on these activities are as follows.

7.5.8.1 Reporting File Collection Information

The supper command is used to report information about file collections. It has a set of subcommands to perform files and directories management that includes verification of information and the checking of results when a procedure is being performed.

Table 14 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 14. Brief Description of Supper Subcommands

Supper Subcommands	Runs on	Reports on
`where`	Node Boot/Install Server	Current boot 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 will search the directory, apply the criteria in the master files, and add the new file.
`install`	CWS	To install a collection.

7.5.8.2 Verifying File Collections Using Scan

By running the `supper scan` command, a scan file will be 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.

7.5.8.3 Updating Files In A File Collection

Make sure changes are made on the files on 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 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.

7.5.8.4 Adding and Deleting Files in a File Collection

Prior to performing addition or deletion of files in a file collection, the following must be considered:

- 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 for the start pointing of the tree for file collection.
- If the file not found in tree structure, copy the file to it.
- If the entry is needed in the list file, add entry to list file.
- If there is no scan file on the master, run the `supper scan` command.
- Run the `supper update` command on the nodes.

7.5.8.5 Refusing 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 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 offline on Frame 2 with the `supper offline` command. This will eliminate the logical path from the CWS to the boolt/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 a file collection 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 files you wish to collect. For example, it has been decided that program files (which are graphic tools) in /usr/local directory are to be included on all nodes.
2. Create the file collection directory. In this case, create /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 lists file in the /var/sysman/sup directory. For example,

   ln -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 file .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 `supper install` command on each boot/install server or node which 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 pre-defined file collections that come with the SP 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 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 following is the cross reference on a summary of common file collection problems and solutions:

Section 16.5, “Diagnosing File Collection Problems” on page 432

7.6 SP User Files and Directories Management

Berkeley Automounter (also known as AMD) and AIX Automounter have been used for SP user files and directories management.

AMD has been used by PSSP 1.2, 2.1 and 2.2. AIX Automounter has been used by PSSP 2.3 onwards.

An NFS automounting system and the SPUM Interface environment allow users local access to any files and directories no matter which node they are logged on to.

7.6.1 Berkeley Automounter, AMD

AMD is used for automatic and transparent mounting and unmounting of NFS file systems and is a simple and effective way for managing NFS file systems and directories.

The AMD daemon runs on CWS, boot/install servers, and all nodes on the SP system. It monitors specified directory mount points, and when a file I/O operation is requested to that mount point, it performs the RPC call to complete the NFS mount to the server specified in the automount map files.

Any mount point directories that do not already exist on the client will be created. After a period of inactivity, two minutes by default, the automount daemon will attempt to unmount any mounted directories under its control. The mounted directories can come from SP boot/install servers or any workstation or server on the network.

AMD is not an IBM product, and its information can be found in the compressed file named /usr/ssp/lpp/public/amd_up102.tar.Z.

There are two types of file maps. These are indirect maps and direct maps.

- Indirect maps are useful for commonly-used, higher-level directories, such as /home.
- Direct maps are useful when directories cannot be dedicated for automount such as /usr.

AMD can be enabled by entering the command `smit enter_data` and selecting `true` for AMD configuration. The system will then run the `amd_config` Perl script. This script is located in the /usr/lpp/ssp/install/bin directory that adds the `amd_start` script to `/etc/rc.sp`.

7.6.2 AIX Automounter

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.

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 SP, 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 control workstation (CWS) to enable changes on a system-wide basis.

The 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_conifg variable is set to `true` so that the automountd (which is the automounter daemon) will start.
4. Ensure usermgmt_config variable is set so that the maps for the user's 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

7.6.3 AMD to AIX Automounter Migration

As of PSSP 2.3, use of the public domain BSD automounter, the AMD daemon, was replaced with native AIX automounter support, which is available as part of NFS in the Network Support Facilities of the AIX Base Operating System (BOS) Runtime. The AIX automount daemon is shipped with AIX 4.3.0 and older systems. In AIX 4.3.1, this daemon was replaced with the AutoFS implementation. AMD uses map files to define the automounter control. These map files are not compatible with the AIX automounter and must be converted.

7.6.3.1 Migration Considerations

If your current installation has SP automounter support configured (the amd_config site environment variable is true) when migrating to PSSP 3.1 from PSSP 2.2, the system configuration process (services_config) will create a new /etc/auto directory structure and default automount configuration files.

If SP User Management services is also configured (the usermgmt_config site environment variable is true), your existing /etc/amd/amd-maps/amd.u map file will be used to automatically create a new /etc/auto/maps/auto.u map file.

The `mkautomap` command is a migration command used to generate an Automount map file from the AMD map file AMD_map created by a previous

SP release. Only AMD map file entries created by a previous SP release will be recognized. If the AMD map file was modified by the customer, results may be unpredictable. If an AMD map entry cannot be properly interpreted, a message will be written to standard error, and that entry will be ignored. Processing will continue with the next map entry. All recognized entries will be interpreted and equivalent.

The `mkautomap` command migrates /etc/amd/amd-maps/amd.u file and add /u filesystem to /etc/auto.master. It also modifies the syslog configuration so that errors are directed to /var/adm/SPlogs/SPdaemon.log file.

7.6.4 Diagnosing AMD and Automount Problems

The following are cross references on AMD and Automount problem diagnosis:

Section 16.4.1, “Problems with AMD” on page 428.

Section 16.4.2, “Problems with User Access or Automount” on page 429.

7.6.5 Coexistence of the AMD and AIX Automounters

A system may consist of newer nodes running PSSP Version 2.3 as well as older nodes running PSSP Versions 2.2 or earlier. Nodes running PSSP 2.3 use AIX Automounter, and nodes running prior versions use AMD.

The SP will configure and run the native AIX automounter on the newer nodes containing PSSP 2.3 and later releases and the BSD AMD daemon on the older nodes containing PSSP 2.2.

If the SP User Management services have also been configured (usermgmt_config site environment variable is also true), the control workstation will create and maintain both the automount map file /etc/auto/maps/auto.u and the AMD map file /etc/amd/amd-maps/amd.u. The `spmkuser`, `spchuser`, and `sprmuser` commands (and their SMIT equivalents) will process user home directory entries in both map files.

If filecoll_config variable has been set to `true` under Site Environment Information during installation, then the SP is configured to manage file collections. In this case, AIX automounter map files will be automatically distributed to all the nodes by means of `supper`.

7.7 Related Documentation

The following books are recommended readings to provide a broader view on user and data management.

SP Manuals

PSSP: Administration Guide, SA22-7348. Chapter 4 describes file collections thoroughly that cover the concepts, how to create file collections, how it works, and so forth. Chapter 5 gives detailed description on managing user accounts that covers how to set up SP users and how to change, delete, and list them.

SP Redbooks

Inside the RS/6000 SP, SG24-5145. Chapter 4, Section 4.8 contains description on file collection that cover the definition, file collection building, installation, organization, maintenance, and so forth. Section 4.9 covers managing AIX Automounter and its difference from BSD Automounter.

Technical Presentation on PSSP Version 2.3, SG24-2080. Chapter 5 contains detailed descriptions on AIX Automounter that covers distribution of files, how to create map files, migration from AMD (prior to PSSP V2.3) to AIX Automounter, and so forth.

Study Guides

IBM Global Services, RS/6000 SP, System Administration: Course Code AU96. Unit 2 describes the managing of user accounts, which covers considerations for setting up and administering users in a distributed system and setting up login control. Unit 3 describes the managing of user directories that cover automounting of NFS file systems, usage, and setting up of the AMD Automounter. Unit 4 covers data management that covers file collection concepts, how to work with and manage file collections, build and install them, the difference between using NIS and file collections, and so forth.

7.8 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. Why passwords cannot be changed directly on the node if SP Users Management is being used?

   A. Because there is not `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 a SP user?

   A. An AIX user is able to access a local resources in a node, while a 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 SP.

   D. SP user can Telnet to the control workstation, while AIX user 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

Part 2. Installation and Configuration

Chapter 8. Configuring the Control Workstation

This chapter addresses various topics related to the initial configuration of the CWS: Preparation of the environment, copy of the AIX and PSSP lpp from the distribution media to the CWS disks, initialization of Kerberos services and of 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.

8.1 Key Concepts You Should Study

Before taking the RS/6000 SP certification exam, you should understand the following CWS configuration concepts:

- PSSP product packaging: Required and optional lpps and filesets.
- Connectivity between the CWS, SP frames, non-SP frames, and SP nodes.
- Storage requirements and directory organization for PSSP software.
- AIX system configuration files related to the SP system.
- CWS configuration commands.
- Setup of Kerberos authentication services.

8.2 Summary of CWS Configuration

This section presents a summary of the initial configuration of the CWS. It corresponds to:

- Steps 1 to 21 of the PSSP 2.4 *PSSP: Installation and Migration Guide*, GC23-3898, Chapter 2.

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 several steps:

1. You need to update the AIX system environment: You have to 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. You must 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 SP frames and nodes. You cannot start configuring the SP system on the CWS until the physical installation is complete. 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 SP nodes 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. Please refer to Chapter 3, “RS/6000 SP Networking” on page 73 for more detail on the TCP/IP configuration step.
4. You must allocate disk space on the CWS for storing the PSSP software, and restore it from the distribution media.
5. You have to install the PSSP on the CWS using the `installp` command.
6. You must configure authentication services on the CWS either by using the Kerberos implementation distributed with PSSP or by using another Kerberos implementation.
7. Finally, you have to initialize the SDR database that will be used to store all your SP system configuration information.

The tasks described in steps one through four can be executed in any order. Steps 5, 6, and 7 must be performed in this order after all other steps.

The following sections describes in more detail the commands, files, and concepts related to these seven steps.

8.3 Key Commands

The commands described in this chapter are to be used only on the CWS and not on the SP nodes.

8.3.1 setup_authent

This command has no argument. It configures the Kerberos authentication services for the SP system. The command, `setup_authent`, 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, on an

AFS based Kerberos service. If you choose the SP provided Kerberos services, `setup_authent` will initialize the primary authentication server on the CWS.

8.3.2 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 initial customization of PSSP onto the CWS (setup of PSSP SMIT panels, initialization of the SDR, and so on), configures the default partition, and starts the SP daemons necessary for the following steps of the SP installation.

8.4 Key Files

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`

8.4.1 .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
```

8.4.2 /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 look like:

```
srcmstr:2:respawn:/usr/sbin/srcmstr # System Resource Controller
```

/etc/inittab is also used to define which PSSP daemons are to be started at boot time. It is updated automatically during the PSSP installation with the entries `sdrd`, `sp`, `fsd`,`hardmon`, `sysctld`, `st_swnum`, `spmgr`, `kerb`, `kadm`, `aplogd`, `swtlog`, `swtadmd`, `hats`, `hags`, `haem`, `hr`, `pman`, and `sp_configd`.

8.4.3 /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 will not check or modify these entries.

8.4.4 /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 adding the following lines:

```
#  additions for tuning of SP-PSSP system
no -o thewall=16384
no -o sb_max=163840
no -o ipforwarding=1
no -o tcp_sendspace=65536
no -o tcp_recvspace=65536
no -o udp_sendspace=32768
no -o udp_recvspace=65536
no -o tcp_mssdflt=1448
```

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' 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.

For example, in our environment, we would add:

```
/usr/sbin/route add -net 192.168.31.0 192.168.3.11
```

8.4.5 /etc/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).

8.5 Environment Requirements

Before starting the installation of the PSSP software onto the CWS, you must prepare the hardware and software environment and pay attention to some rules and constraints.

8.5.1 Connectivity

During normal operation, the TCP/IP connectivity needed for user applications between the CWS and the SP nodes 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 SP nodes from the CWS, there *must* exist an Ethernet network connecting all SP nodes 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) SP node(s) with multiple Ethernet adapters, Ethernet hubs, or Ethernet routers.

Furthermore, the monitoring and control of the SP frames and nodes hardware from the CWS requires a serial connection between the CWS and *each* frame in the SP system. If there are many frames, there may not be enough build-in serial adapters on the CWS and additional serial adapter cards may need to be installed in the CWS.

In the case that SP-Attached Servers (RS/6000 S70 or S7A) 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 server configured on the en0 adapter of the server.

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, and the SP nodes through the serial links, and between the CWS and the nodes through the Ethernet network, must be set up before starting the PSSP installation.

8.5.2 Disk Space and File System Organization

Before starting installation of the PSSP software, plan the allocation of disk space dedicated to the storage of the product code as well as to the archiving of the AIX images (mksysb) loaded on each SP node. This disk space is organized in a directory structure that must comply with naming conventions.

8.5.2.1 /spdata Size and Disk Allocation

Most of the PSSP-related disk storage is allocated within the /spdata directory. (A small amount of storage is also needed in /usr and /var.) The exact size that must be available on the CWS and on each boot/install server can be computed using the formulas in Chapter 3 of *Planning Volume 2, Control Workstations and Software Environment*, GA22-728. The exact size depends on the installation of optional filesets and the number of mksysb images to be kept on the CWS. However, some rules of thumb can be used for a rough evaluation of the size. For a simple configuration, the same system image installed on all nodes and AIX with a reasonable number of options, 1.8 GB will be necessary. For a system with *n* images (archiving backup images, or using different images for different nodes), the size will be in the order of (1100+(n * 700)) MB. Note that the same image can be used for uniprocessor or multiprocessor nodes.

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 91 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 will 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 will change when the systems installed on the SP nodes 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.

8.5.2.2 /spdata Directory Structure and Naming Convention

You must manually create the /spdata directory before the beginning of the PSSP installation with a minimum substructure consisting of the following directories:

```
/spdata/sys1/install/
/spdata/sys1/install/<source_name>
/spdata/sys1/install/<source_name>/lppsource
/spdata/sys1/install/aix431
/spdata/sys1/install/aix431/lppsource
/spdata/sys1/install/aix432
/spdata/sys1/install/aix432/lppsource
/spdata/sys1/install/images
/spdata/sys1/install/pssp
/spdata/sys1/install/pssplpp
/spdata/sys1/install/pssplpp/PSSP-2.2
/spdata/sys1/install/pssplpp/PSSP-2.3
/spdata/sys1/install/pssplpp/PSSP-2.4
/spdata/sys1/install/pssplpp/PSSP-3.1
```

Figure 104. /spdata Initial Structure

The installable images (lpp) of the AIX systems must be stored in directories named `/spdata/sys1/install/<source_name>/lppsource`.You can set `<source_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 aix421, aix431, and so on.

Except for `<source_name>`, the name of all directories listed in Figure 104 *must* be left unchanged.

In Section 8.5.2.1, “/spdata Size and Disk Allocation” on page 240, we have 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 (may be only once at

installation time). Practically, this implies that you create on the spstdvg volume group one file systems for holding each directory:

/spdata/sys1/install/<source_name>

/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.

8.6 LPP Filesets

An SP 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 SP system.

8.6.1 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.

The minimum set of AIX components to be copied to the /spdata/sys1/install/<source_name>/lppsource directory is shown is Table 15.

Table 15. Minimum AIX LPP Requirements

bos	bos.diag.*	bos.mp.*	bos.net.*
bos.powermgt.*	bos.sysmgt.*	bos.terminfo.*	bos.up.*
bos.64bit	devices.*	xlC.rte.*	X11.apps.*
X11.base.*	X11.compat.*	X11.Dt.*	X11.fnt.*
X11.loc.*	X11.motif.*	X11.msg.*	X11.vsm.*

For installation on AIX releases earlier or equal to 4.2, you also need to install the filesets bos.info.*

In addition, the right level of perfagent must be installed on the CWS and copied to each /spdata/sys1/install/<source_name>/lppsource directory, according to Table 16.

Table 16. Perfagent Filesets

AIX level	PSSP level	Required Filesets
AIX 4.1.5	PSSP 2.2	perfagent.server 2.1.5.x
AIX 4.2.1	PSSP 2.2	perfagent.server 2.2.1.x where x>=2
AIX 4.2.1	PSSP 2.3	perfagent.server 2.2.1.x where x>=2
AIX 4.3.1	PSSP 2.3	perfagent.server 2.2.31.x
AIX 4.3.1	PSSP 2.4	perfagent.server 2.31.x
AIX 4.3.2	PSSP 2.3	perfagent.tools and perfagent.server 2.2.32.x
AIX 4.3.2	PSSP 2.4	perfagent.tools and perfagent.server 2.2.31.x
AIX 4.3.2	PSSP 3.1	perfagent.tools 2.2.32.x

8.6.2 PSSP Filesets

The installation of the PSSP software on the CWS disks is a two-step process.

1. All PSSP software is first restored from the distribution media into the /spdata/sys1/install/pssplpp/PSSP-x.x directory using the `bffcreate` command.

   The ssp.user file must then be renamed into pssp.installp, and the table of contents must be regenerated (execute the `inutoc` command). The exact filename of ssp.user depends on the PSSP version:

   PSSP 2.4: ssp.usr.2.4.0.0

   PSSP 3.1: ssp.usr.3.1.0.0

2. Part of the code of the /spdata/sys1/install/pssplpp/PSSP-x.x directory is then installed (`installp`) on the CWS. You must first choose which of the PSSP filesets you need to install.

The filesets to be installed depend on the version of PSSP.

8.6.2.1 PSSP 2.4 Filesets

For PSSP 2.4, Table 17, Table 18, Table 19, and Table 20 list the filesets that are, respectively, required when using a Switch, when using a Switch Router, or optional.

Table 17. PSSP 2.4 Required Filesets

Fileset Description	Fileset Name
SP System Support package	ssp.basic
Authentication Client Commands	ssp.clients
System Monitor and Perspectives	ssp.gui
Sysctl	ssp.sysctl
Availability subsystems	ssp.ha
Topology services	ssp.topscvs

Table 18. PSSP 2.4 Required Filesets (with an SP Switch)

Fileset Description	Fileset Name
Communication subsystem	ssp.css
Communication subsystem topology	ssp.top

Table 19. PSSP 2.4 Required Filesets (with an SP Switch Router)

Fileset Description	Fileset Name
Extension nodes SNMP manager	ssp.spmgr

Table 20. PSSP 2.4 Optional Packages

Fileset Description	Fileset Name
System management tools (NTP, file collection, and so on).	ssp.sysman
Resource manager	ssp.jm
Public tools	ssp.public
PSSP documentation	ssp.docs

Fileset Description	Fileset Name
Authentication Server: This package is compulsory if you wish to use the CWS as the master Kerberos authentication server for the PSSP provided authentication method.	ssp.authent
System Partitioning Aid	ssp.top.gui
Job Switch Resource Table Services	ssp.st
Perl	ssp.perlpkg
Problem management	ssp.pman
High Availability Control Workstation	ssp.hacws
Performance Monitor	ssp.ptpegui
Virtual Shared Disk supports	ssp.csd.vsd, sp.csd.cmi, ssp.csd.hsd, ssp.csd.sysctl, ssp.csd.gui
Minimal AIX mksysb images: This package is only needed if the user does not provide his own mksysb for the nodes.	spimg
Supervisor Microcode	ssp.ucode

8.6.2.2 PSSP 3.1 Filesets

For PSSP 3.1, Table 21, Table 22, Table 23, and Table 24 describe the filesets that are required to support a Switch, to support a Switch Router, or optional.

Table 21. PSSP3 3.1 Required Filesets

Fileset Description	Fileset Name
Cluster Technology	rsct.basic.hacmp, rsct.basic.rte, rsct.basic.sp, rsct.clients.hacmp, rsct.clients.rte, rsct.clients.sp
SP System Support package	ssp.basic
Authentication Client Commands	ssp.clients

Fileset Description	Fileset Name
Compatibility package	ssp.ha_topcvcs.compat
Perl	ssp.perlpkg
Sysctl	ssp.sysctl
System management tools (NTP, file collection, and so on).	ssp.sysman

Table 22. PSSP 3.1 Required Filesets (with an SP Switch)

Fileset Description	Fileset Name
Communication subsystem	ssp.css
Communication subsystem topology	ssp.top

Table 23. PSSP 3.1 Required Filesets (with an SP Switch Router)

Fileset Description	Fileset Name
Extension nodes SNMP manager;	ssp.spmgr

Table 24. PSSP 3.1 Optional Filesets

Fileset Description	Fileset Name
Performance Toolbox Parallel Extensions	ptpe.docs, ptpe.program
Minimal AIX mksysb images: This package is only needed if you do not provide your own mksysb for the nodes.	spimg
Authentication Server: This package is compulsory if you wish to use the CWS as the master authentication server for the PSSP provided authentication method.	ssp.authent
PSSP Documentation	ssp.docs, ssp.resctr.rte
Perspectives	ssp.gui

Fileset Description	Fileset Name
High Availability Control Workstation	ssp.hacws
Problem Management	ssp.pman
Performance Monitor	ssp.ptpegui
Public Tools	ssp.public
Job Switch Resource Table Services	ssp.st
TEC Event Adapter	ssp.tecad
System Partitioning Aid	ssp.top.gui
Supervisor Microcode	ssp.ucode
Virtual Shared Disk support	vsd.cmi, vsd.hsd, ssp.vsdgui, vsd.sysctl, vsd.vsddd
Recoverable Virtual Shared Disks	vsd.rvsd.hc, vsd.rvsd.rvsdd, vsd.rvsd.scripts

8.7 Related Documentation

For complete reference and ordering information for the documents listed in this section, see Appendix C, “Related Publications” on page 471.

SP Manuals

You can refer to two sets of documents related to either Version 2.4 or Version 3.1 of PSSP:

RS/6000 SP Planning Volume 2, Control Workstation and Software Environment, GA22-7281. Chapters 2, 3, and 5 provide detailed information about the SP connectivity, storage requirements, and site information.

PSSP: Installation and Migration Guide, GC23-3898. 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, GC23-3900, for PSSP 2.4 or *PSSP: Command and Technical Reference*, SA22-7351 for PSSP 3.1 contains a complete description of each CWS installation command listed in 8.3, “Key Commands” on page 236.

SP Redbooks

RS/6000 SP: PSSP 2.2 Survival Guide, SG24-4928. Chapter 2 describes the logical flow of steps that make the installation process.

Inside the RS/6000 SP, SG24-5145. Chapter 5 contains a high-level description of the installation process.

8.8 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 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 SP installation. Which of the following is not done by the install_cw script?

   A. Creates RS/6000 SP SMIT panels

   B. Initializes the SDR

   C. Creates and starts the hardmon daemon

   D. Creates and starts the partition-sensitive daemons

2. Which of the following is not a pre-requisite for PSSP 3.1?

   A. AIX 4.3.2

   B. perfagent.server.2.2.32.X

   C. perfagent.tools.2.2.32.X

   D. xlC.rte

3. Which filesets are a pre-requisite for PSSP 3.1?

   A. rsct.basic and rsct.clients

   B. rsct.basic.sp and rsct.clients.sp

   C. ssp.ha and ssp.clients

   D. ssp.topsvc and ssp.hacws

Chapter 9. Frames and Nodes Installation

In Chapter 8, “Configuring the Control Workstation” on page 235, we presented the initial configuration of the CWS. This chapter addresses all of the other steps of the installation of an SP system from the configuration of the PSSP software on the CWS through the installation of AIX and PSSP on SP nodes up to the first boot of nodes and switches.

9.1 Key Concepts You Should Study

Before taking the RS/6000 SP certification exam, you should understand the following frame, nodes, and switch installation concepts:

- Structure of the SDR configuration information: Site information, frame information, and node information
- Contents of the predefined subdirectories of /spdata
- Files used for SDR configuration and SP frames, nodes, and switches installation
- NIM concepts applied to the SP environment
- Setup of boot/install servers (primary and secondary)
- Network-installation concepts
- Automatic and manual node conditioning
- SP system customization
- SP 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 an SP system configuration and installation.

To help you understand the use of each command, they are presented in association to the installation step in which they are performed and in the order in that 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 an SP system. In this case, we also provide information that is not related to the installation step but that you may need at a later stage in the life of your SP system.

Finally, this section is not intended to replace the SP manuals referenced in 9.4, “Related Documentation” on page 275. You should refer to the these manuals to get a 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 be initialized (the last command executed on the CWS was `install_cw`). We are now at the beginning of the SP system customization and installation.

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 command `spsitenv` defines all site wide configuration information (name of the default installable image, NTP, and so on) and which of the optional PSSP features will be used (SP User Management, SP File Collection Management, SP Accounting).

Because of the number of parameters you must provide in the `spsitenv` command, we recommend that you use the SMIT interface rather than the command.

In our environment, the site configuration is defined as shown on Figure 105 on page 251.

```
                    Site Environment Information

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

                                                    [Entry Fields]
  Default Network Install Image                    [bos.obj.ssp.432]
  Remove Install Image after Installs               false                +

  NTP Installation                                  consensus            +
  NTP Server Hostname(s)                           [""]
  NTP Version                                       3                    +

  Automounter Configuration                         true                 +

  Print Management Configuration                    false                +
  Print system secure mode login name              [""]

  User Administration Interface                     true                 +
  Password File Server Hostname                    [sp3en0]
  Password File                                    [/etc/passwd]
  Home Directory Server Hostname                   [sp3en0]
  Home Directory Path                              [/home/sp3en0]

  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              [aix432]



F1=Help           F2=Refresh        F3=Cancel         F4=List
F5=Reset          F6=Command        F7=Edit           F8=Image
F9=Shell          F10=Exit          Enter=Do
```

Figure 105. Site Environment Information

9.2.2 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. This command is used to enter frame configuration information into the SDR: association between the frame number and the tty port on the CWS to which the frame is attached through a serial link.

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 3.1, this command also defines the hardware protocol used on the serial link (SP for SP nodes, SAMI for SP-Attached Servers) and the switch port to which a non-SP frame is attached.

This command must be performed during the first installation of an SP 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 in between the frames installed originally in the system.

In our environment, we define the first frame using:

```
spframe -r yes 1 1 /dev/tty0
```

The second frame will be defined later in Chapter 15, “RS/6000 SP Reconfiguration and Update” on page 385.

9.2.3 Check the Level of Supervisor Microcode

Once the frames have been configured, and before starting to configure the nodes, we recommend to check that the 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, which must have been installed on the CWS to perform this operation.

The `spsvrmgr` command manages the supervisor code on the SP frames. It executes on the CWS only. It can be called from the command line or from SMIT. Each of the command line option is equivalent to a one of the functions accessible from the SMIT RS/6000 SP Supervisor Manager window.

It 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 to use the SMIT panels to perform these operations. However, two commands can be used for system wide checking and update:

- `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 this stage at the beginning of the SP system installation rather than later when the system is in production.

9.2.4 Check the Previous Installation Steps

Since the complete SP system installation is a complex process involving more than 50 steps, it is a good idea to perform some checking at several points of the process to insure that already executed steps were successful. Chapter 10, “Verification Commands and Methods” on page 279 is addressing the various aspect of SP system checking. We will, therefore, not mention the actions to perform at each checkpoint in this chapter but only the most useful command: `splstdata`

This command executes on the CWS or any SP node when using the command line interface. It can only be used on the CWS when called from one of the SMIT windows accessible from the List Database Information window (`smitty list_data`).

This command displays configuration information stored in the SDR about the frames and nodes.

This command has many options. During the SP installation, the most useful ones are:

- `-n` for node general configuration
- `-b` for node boot/installation configuration
- `-a` for node adapters configuration
- `-f` for frame information

At this point in the installation, you can use `splstdata -f` and `splstdata -n` to verify that the frames have been correctly configured in the SDR, and that the execution of the `spframe` command has correctly discovered the nodes in each frame.

9.2.5 Define the Nodes Ethernet Information

Once the frame information has been configured, and the microcode level is up to date, you have to define in the SDR the IP addresses of the en0 adapters of each of the SP nodes as well as the type and the default route for this Ethernet adapter.

This task is performed by the `spethernt` 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 `spethernt` command can define adapters individually, by group, or by ranges.

In our environment, we need to use the command three times to define all adapters since the en0 adapter of node 1, the en0 adapters of nodes 5 to 9, and the en0 adapters of nodes a10 to 15 are in three different ranges of the sequential Ethernet IP address and because the default route is different for node 1 than for other nodes:

```
spethernt -s yes -t bnc 1 1 192.168.3.11 192.168.3.130
spethernt -s yes -t bnc 5 5 192.168.31.15 192.168.31.11
spethernt -s yes -t bnc 10 6 192.168.31.110 192.168.31.11
```

Figure 106 shows which adapters are defined by each of these commands.

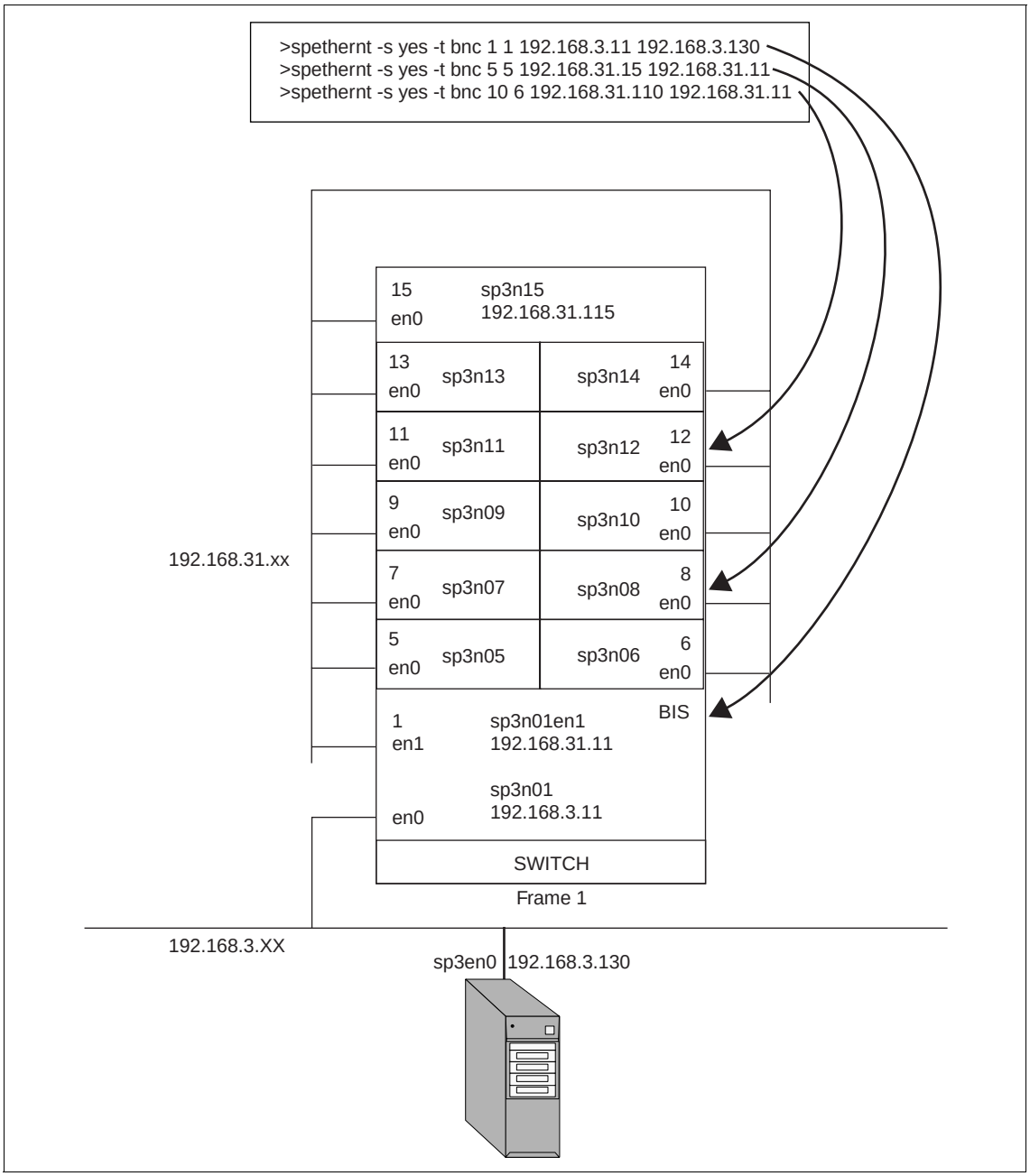


Figure 106. Definition of Additional Adapters

9.2.6 Discover or Configure the Ethernet Hardware Address

Once the nodes en0 IP address 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 a file /etc/bootptab.info (for more details, see 9.3.1, "/etc/bootptab.info" on page 267) to speed-up the `sphrdwrad` command. For each node in the argument list of the command, `sphrdwrad` will look if it finds its hardware address in the /etc/bootptab.info. If it cannot find it, it will then query the node through the hardware connection to the frame (serial link). In the latter case, the node will be powered-down and powered-up.

Note

Do not use the `sphrdwrad` command on a running node since it will be powered off.

In our environment, we can use either command to discover the en0 adapter hardware addresses:

```
sphrdwrad 1 1 rest
```

or

```
sphrdwrad 1 1 12
```

9.2.7 Configure Additional Adapters for Nodes

In addition to the en0 adapters, SP 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`).

The `spethernt` command configures the en0 adapter, and the `spadaptrs` command is its counterpart for the other adapters. Similar to the `spethernt`

command, you can configure with `spadaptrs` the IP address of individual adapters or range of adapters; 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 (ATM, ESCON) can not 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 272).

For the switch adapters, two options `-a` and `-n` allow 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 node 1 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
```

9.2.8 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.

In our environment, we only want to change the format of the name and use the short names but keep the en0 adapter name as host name:

```
sphostnam -f short 1 1 12
```

9.2.9 Create Authorization Files

This step is only executed in PSSP 3.1. There is no equivalent step in PSSP 2.4.

You now have to create the appropriate authorizations files for use of root remote commands, such as `rcp`, `rsh`, and so on, on the CWS. Possible methods are Kerberos4 and AIX standard authentication.

On the CWS, you can use either `spsetauth` or the SMIT Select Authorization Methods for Root access to Remote Commands window (`smitty spauth_rcmd`).

In our environment, we configured both Kerberos 4 and AIX standard methods for the default partition (`sp3en0`):

```
spsetauth -d -p sp3en0 k4 std
```

9.2.10 Enable Selected Authentication Methods

This step is only executed in PSSP 3.1. There is no equivalent step in PSSP 2.4.

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`).

In our environment, we wish to use both Kerberos 4 and AIX standard methods in the default partition:

```
chauthpar k4 std
```

9.2.11 Start System Partition-Sensitive Subsystems

After the SDR has been loaded with the frame and nodes information, IP addresses, symbolic names, and routes, you have to add and start all subsystems in the default partition.

The `syspar_ctrl` command controls the system partition sensitive subsystems on the CWS and on the SP nodes. At this point, it is used only on the CWS since the nodes are still not installed:

```
syspar_ctrl -A
```

This command will start the daemons: hats, hags, haem, hr, pman, emon, spconfigd, emcond, and spdmd (optional).

Execution of this command on the CWS only starts the daemons on the CWS and not on any SP node. 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.

9.2.12 Set Up Nodes to Be Installed

This step is different in PSSP 2.4 and PSSP 3.1.

In PSSP 2.4, the `spbootins` completely perform this task, while in PSSP 3.1 the task is split between the `spchvgobj` and the `spbootins` command. Sections 9.2.13, “spchvgobj” on page 259 and 9.2.14, “spbootins” on page 260 describe the functions performed by the commands in each case.

9.2.13 spchvgobj

The `spchvgobj` command executes on CWS only.

It is equivalent to the SMIT Change Volume Group Information window (`smitty changevg_dialog`).

This command is only available in PSSP 3.1. It has been added as part of the new PSSP feature, which has the possibility of having several bootable volume groups. The boot/install configuration, that was, up to PSSP 2.4, specific of a node, is now specific to a volume group.

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 as the boot/install server(s):

- 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 `spchvgobj` command or its SMIT equivalent to specify for a set of SP nodes, and for a bootable volume group, the names of disk(s), 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 lpp source directory, and the level of PSSP to be install on the nodes.

In our environment, since at the time of the installation of the first frame we already plan for adding a new frame, we want to force node 5 to 15 to point to node 1 as the boot/install server. We can therefore use:

```
spchvgobj -n 1 1 5 11
```

where `-n 1 indicates that node 1 is the server, and 1 5 11` indicates that 11 nodes starting from frame 1 node 5 will point to this server.

9.2.14 spbootins

The `spbootins` command executes on CWS only.

It is equivalent to the SMIT Boot/Install Server Information window (`smitty server_dialog`).

This command behaves differently in PSSP 2.4 and PSSP 3.1.

9.2.14.1 spbootins in PSSP 2.4

In PSSP 2.4, the boot/install server configuration is associated to each node. The use of `spbootins` during installation is optional. You can use it to change the default boot/install configuration for a set of nodes, the names of disk(s) where to install the AIX image, 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 and the level of PSSP to be install on the nodes, and how they will perform their next boot (from a server or from their disk).

If our environment was running under PSSP 2.4 instead of PSSP 3.1, then we would have used

```
spbootins -n 1 1 5 11
```

to perform the same customization in the example given in 9.2.13, “spchvgobj” on page 259.

9.2.14.2 spbootins in PSSP 3.1

As mentioned in 9.2.13, “spchvgobj” on page 259, most of the boot/install server configuration in PSSP 3.1 is associated with a volume group. The `spbootins` command is used to define in the SDR for 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).

In our environment, we use

```
spbootins -s no -r install 1 1 12
```

to specify that, at their next reboot, all nodes in frame one are to load the AIX image from their respective boot/install server and to ask not to run setup_server.

9.2.15 Configure the CWS as Boot/Install Server

The SDR now contains all required information to create a boot/install server on the CWS.

The command `setup_server` 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 on **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`.

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 will contain 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 performs various tasks, such as configuring PSSP or setting the NIM environment. Here is a simplified control flow sequence of the `setup_server` command (if present, the text in italic is the name of the wrapper actually performing the action associated with the step):

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: `kinit`
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 filesystem 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.

9.2.16 Set the Switch Topology

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 Topology File Annotator (`smitty annotator`) and Store a Topology File windows (`smitty etopology_store`).

In our environment, since we have one Node Switch Board and no Intermediate Switch Board, we use:

```
Eannotator -F /etc/SP/expected.top.1nsb.0isb.0 -f /etc/SP/expected.top/annotated -O no
Etopology /etc/SP/expected.top/annotated
```

9.2.17 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`).

Assuming that, in our environment, we wish to use node 5 instead of the default value, node 1, as the primary node, we use:

```
Eprimary -init 5
```

9.2.18 Set the Clock Source for All Switches

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.

In our environment, and since there is only one switch, we let `Eclock` automatically make the decision:

```
Eclock -d
```

9.2.19 Network Boot the Boot/Install Server Nodes

After configuring the switch, we are finally ready to install the SP 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.

In normal cases, the installation of a node requires that you open two shell windows on your CWS display. One will be used to monitor the execution of the installation using the `s1term` command, while the other one is used to initiate the installation using the `nodecond` command. These two commands execute in parallel. We present them sequentially in 9.2.20, “s1term” on page 263 and 9.2.21, “nodecond” on page 264.

9.2.20 s1term

The `s1term` command executes on CWS only.

The `s1term` command opens a connection to the SP node serial port. Since the node console is, by default, associated to this port, `s1term` provides a remote console access to the SP node from the CWS through the serial link.

It is, therefore, a very useful command to take control of a node when the IP connection through the Ethernet network is not available.

By default, `s1term` provides a read-only connection. If you wish to enter commands on the node, you need to establish a read-write connection by using the `-w` option.

During installation, the nodecond command needs write access to the node on the serial link. The write access cannot be shared by several clients. You must, therefore, only open a read-only connection to monitor the node installation and see all messages displayed on the node console.

In our environment, we first boot the node 1, which is a boot/install server. We, therefore, open a connection on this node:

```
s1term 1 1
```

After the boot/install servers have successfully been installed, you can start the installation of the other nodes. To monitor this installation, you can open a parallel one `s1term` session to each of these nodes

9.2.21 nodecond

The nodecond command executes on CWS.

This is equivalent to clicking on **Run setup_server Command** in SMIT Enter database Information window (`smitty enter_data`).

In parallel with the s1term window, you can now initiate the boot and system installation on the target node. This phase is called node conditioning and it is executed by the nodecond command. It is executed on the CWS 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.

In our environment, we first need to install node 1 and start the command in the background:

```
nodecond 1 1 &
```

After all boot/install server has successfully been installed, you can condition the remaining nodes of your SP system.

In our environment, we would perform:

```
for i in 5 6 7 8 9 10
do
   nodecond 1 $i &
done
```

wait for completion since they all boot from node 1

```
for i in 11 12 13 14 15
do
   nodecond 1 $i &
done
```

It is also possible to perform the installation using the Perspectives graphical user interface. In the Hardware Perspective window, you select the node you want to install, and then you need only to click on the **Network Boot...** item of the Action menu.

Manual Node Conditioning

In some cases, you may want to perform the node installation manually rather than automatically using nodecond. Manual node conditioning is a more complex task consisting of several steps. Since it highly depends on the hardware type of the node, these steps differ for each category of nodes.

For a Thin or Wide node (non-SMP), using either spmon -g (PSSP 2.4) or Hardware Perspective (PSSP 3.1) and a `s1term -w` window the steps are:

1. Power off the node (if it is powered on).
2. Put the key in Secure mode.
3. Power on the node.
4. When the led gets to 200, put the key in Service mode.
5. Reset the node.
6. When the led reaches 260 or 262, the Main menu is displayed. Select option 1 **Select Boot Device**.
7. In the Select Boot (startup) Device menu, select your network adapter to boot from.
8. You will be prompted to enter IP addresses for the client (the node to be installed), the server (the boot/install server for the node being installed), and a gateway (which you may leave empty).
9. Return to the main menu.
10. Select option **3** to send a test transmission ping between the client and the server.
11. Return to the main menu and select option **4** to start the system boot.
12. At this point, make sure the key is in Normal mode before the installation finishes using the command:

    `spmon -key normal <node>`

For a High Node, the steps are:

1. Power the node off using `spmon`.
2. Set the key service by invoking `spmon`.
3. Open a tty connection using a read/write `s1term`.
4. Press **Enter** to get the BUMP prompt and type `sbb`.
5. Choose option 1 **Set Flags** from the STAND-BY MENU.
6. Verify the if Fast IPL is enabled. If it is disabled, enable it.
7. Select **x** until the > prompt reappears.
8. Power on the node by using `spmon`.
9. Wait for the Maintenance Menu to appear.
10. Select option 6 **System Boot**.
11. In the System Boot Menu, select option 1 **Boot from network**.
12. Wait for Main Menu of the boot install and select option 4 **Exit Main Menu and Start System (BOOT)**.

For a CHRP Node (SMP Thin and Wide), perform the following steps:

1. Power the node off by invoking `spmon`.
2. Open a tty connection using a read/write `s1term`.
3. Power on the node using `spmon`.
4. Wait for a panel to appear as shown in Figure 107.

```
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000
RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000 RS/6000


memory      keyboard     network      scsi
```

Figure 107. Boot Screen

5. Type `1` after `scsi` appears on the screen.
6. When SMS Main Menu screen appears, select option 3 **Utilities**.

7. When the Utilities Menu appears, choose option 4 **Remote Initial Program Load Setup**.
8. In the Network Parameters Menu, choose option 2 **Adapter Parameters**.
9. In the Adapter Parameters Menu, select the appropriate adapter parameters.
10. Select **X** to exit menus until you get the SMS Main Menu.
11. Select **2** to go to the Multiboot Menu.
12. When the Multi Menu appears, select option 4 **Select Boot Devices**.
13. When the Boot Device Menu appears, select option 3 **Configure the first boot device**.
14. In the Configure Boot Device Menu, select the appropriate network interface.
15. Select **X** until the SMS Main Menu appear.
16. Select **X** to start the booting process.

9.2.22 Check the System

At this point, we recommend that you check the SP system using the `SYSMAN_test` command (see 10.3.1.7, “Checking Sysman Components: SYSMAN_test” on page 281).

9.2.23 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 on **Start Switch** in the SMIT Perform Switch Operation Menu.

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 SP 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 with a comma with no blanks.

The MAC address is separated from the node identifier by a blank. It is formatted in hexadecimal with no . or :. 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 Figure 108.

```
> 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
>
```

Figure 108. Example of /etc/bootptab.info

`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 SP client 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 (`setup_server` command) 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 109 on page 269.

```
[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 sp3n01.msc.itso.ibm.com.install_info
-rw-r--r--   1 root     system   3928595 Dec 15 15:33 spot_aix432.rs6k.mp.ent
-rw-r--r--   1 root     sys         2250 Nov 13 13:59 tuning.cust
[sp3en0:/]#
```

Figure 109. Contents of the CWS /tftpboot Directory

We will now describe in more detail the role of each of these files.

9.3.2.1 <spot_name>.<archi>.<kernel_type>.<network>

Files with this format of name are bootable images. The naming convention is:

- **<spot_name>** 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>** is the machine architecture that can load this bootable image. It is one of `rs6k`, `rspc`, or `chrp`.
- **<kernel_type>** 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>** 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 will contain a symbolic link to the appropriate bootable image. You can see an example of this in Figure 109 on page 269, where this file is called spot_aix432.rs6k.mp.ent.

9.3.2.2 <hostname>-new-srvtab

These files are created by the `create_krb_files` 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 the passwords for the rcmd principals of the SP node. Each SP node retrieves its <hostname>-new-srvtab file from the server and stores it in its /etc directory as krb-srvtab.

9.3.2.3 <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 like psspfb_script.

9.3.2.4 <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.

9.3.2.5 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).

9.3.2.6 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.

9.3.2.7 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/<aix_level>/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/<aix_level>/lppsource directory. The recommended rule is to set the relative pathname <aix_level> to a name significantly indicating the level of AIX: aix414, aix421, aix432. 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. Refer to the 8.6.1, “PSSP Prerequisites” on page 242 for the minimal sets of AIX and perfagent lpp to install in this directory. Starting with AIX release 4.3.2, perfagent.tools is part of AIX and not PAIDE as it used to be for previous AIX releases.

If the SP system contains several boot/install server, 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 whatever 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-2.2, PSSP-2.3, PSSP-2.4 or PSSP-3.1.

During the first step of the PSSP software installation on the CWS (refer to 8.6.2, “PSSP Filesets” on page 243), the PSSP source images must be installed using `bffcreate` into these directories.

If the SP 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 nodes 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 268).

- 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 SP 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, please refer to the *AIX 4.3 Network Installation Management Guide and Reference*, SC23-4113.

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 imagedata 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 Appendix C, “Related Publications” on page 471.

SP Manuals

The reader can refer to two sets of documents related to either version 2.4 or version 3.1 of PSSP.

PSSP:Administration Guide, GC23-3897 for PSSP 2.4 and *PSSP: Administration Guide*, SA22-7348 for PSSP 3.1. Chapters 7, 8, 12, 14, and 15 provide detailed information about the services that may be configured in the SP system: Time Server, Automounter, Security, Switch and System partitions.

PSSP: Installation and Migration Guide, GC23-3898, for PSSP 2.4. In Chapter 2, steps 22 to 59 detail the complete installation process. Appendix C, D, F, and G describe the customization of nodes, the wrappers associated with the `setup_server` command, and the procedure to solve port contention issues.

PSSP: Command and Technical Reference, GC23-3900, for PSSP 2.4 and *PSSP: Command and Technical Reference*, SA22-7351 for PSSP 3.1 contain a complete description of each command listed in 9.2, “Installation Steps and Associated Key Commands” on page 249

SP Redbooks

RS/6000 SP: PSSP 2.2 Survival Guide, SG24-4928. Chapter 2 contains practical tips and hints about specific aspects of the installation process.

Inside the RS/6000 SP, SG24-5145. Chapter 4 presents the concepts underlying the SP system software and provides help in the planning the installation of this software. Section 5.2 describes the high-level design of the installation process.

Others

We recommend the use of either *AIX 4.2 Network Installation Management Guide and Reference*, SC23-1926 or *AIX 4.3 Network Installation Management Guide and Reference*, SC23-4113 for getting any detailed information about NIM.

9.5 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. In an SP 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. Once the frames have been configured, and before starting configuring nodes, it is recommended to check that the frame supervisor microcode is at the latest level supported by the PSSP level being installed. The command that checks the supervisor microcode level is:

   A. `spchksvr - G -r status all`

   B. `spsvrmgr -G -u all`

   C. `spsvrmgr -G -r status all`

   D. It is done through SP Perspectives.

3. How 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 node to install.

4. How does the `nodecond` script access the node to start the network booting process?

   A. Through the RS-232 line from the control workstation.

   B. Through TCP/IP from the control workstation.

C. Through the RS-232 line from the boot/install server node.

D. Through the Ethernet network from the control workstation.

Chapter 10. Verification Commands and Methods

This chapter presents some of the commands and methods available to the SP administrator to check that the SP system has been correctly configured, initialized, and started.

10.1 Key Concepts You Should Study

Before taking the RS/6000 SP certification exam, you should understand the following concepts related to verifying and checking an SP system:

- The `splstdata` command.
- The various components of an SP system and the different verifications 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 SP System Checking

Several options are available to the SP user or administrator who wish 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 279 presents the commands that are available for checking various aspects of an SP system. Section 10.4, “Graphical User interface” on page 288 give a few hints about the use of `spmon -g` and Perspectives. Section 10.5, “Key Daemons” on page 290 focuses on the daemons that are important to monitor in an SP system. Section 10.6, “SP Specific Logs” on page 292 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 SP user. We present in this section the most widely used AIX and PSSP commands for this purpose.

10.3.1 Verify Installation of Software

During the CWS and SP system installation, your first verification task consist in checking that the AIX and PSSP software have been successfully installed and that the basic components are configured correctly before you start entering configuration data specific to your environment.

10.3.1.1 Checking Software Levels: lslpp

The `lslpp` command is the standard AIX command to check that 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, using the tables of Section 8.6.1, “PSSP Prerequisites” on page 242:

```
lslpp -La bos* devices* perf* X11* xlC* ssp* rsct* | more
```

You should also verify that you have installed all PSSP filesets corresponding to your SP hardware configuration and to the options you wish to use (VSD, RVSD, and so on).

10.3.1.2 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. This command can also be used later, during operation of the SP system, if you suspect problems with the SDR.

10.3.1.3 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 that spmon is correctly installed using the `spmon_itest` command.

10.3.1.4 Checking the System Monitor Configuration: spmon_ctest

After the SP hardware has been discovered by the CWS (`spframe`), you can check that the System Monitor has been correctly configured with the information about the SP frames and nodes hardware using the `spmon_ctest` command. This command also checks that the hardmon daemon is running, that the serial RS232 links to the frames and nodes are properly connected, that the CWS can access the frames and nodes hardware through these connections, and that the hardware information has been stored in the SDR.

10.3.1.5 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.

For example, to check all PSSP related filesets, you can use:

```
lppdiff -Ga ssp* rsct*
```

10.3.1.6 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 as shown in Figure 110.

```
[sp3en0:/usr/lpp/ssp]#  splst_versions -tG
1 PSSP-3.1
5 PSSP-3.1
6 PSSP-3.1
7 PSSP-3.1
8 PSSP-3.1
9 PSSP-3.1
10 PSSP-3.1
11 PSSP-3.1
12 PSSP-3.1
13 PSSP-3.1
14 PSSP-3.1
15 PSSP-3.1
[sp3en0:/usr/lpp/ssp]#
```

Figure 110. PSSP Versions Installed on Each Node

10.3.1.7 Checking Sysman Components: SYSMAN_test

The `SYSMAN_test` command is a very powerful test tool. It checks a large number of SP 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. However, its use is not limited to the installation phase of the life of your SP system. It can provide valuable information during normal operation of an SP system.

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 will also perform several tests on them. 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 to redirect the output to a file to prevent flooding the screen with messages that display too fast and to then use a file browser or editor to look at the results of the command. An alternative is to look at file /var/adm/SPogs/SYSMAN_test.log, but this file does not contain all the information provided by the verbose option.

10.3.1.8 ssp.css: Switch Code CSS_test

The last command we will 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 this command 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 SP nodes.

10.3.2 Verify System Partitions

Two commands are particularly useful for checking the SP system partitions.

10.3.2.1 Listing Existing Partition: splst_syspars

The first of these commands, splst_syspar, only lists the existing partitions in the SP 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]#
```

10.3.2.2 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 SP system. It is only to be used when the system has more partitions than the initial default partition.

10.3.3 Checking Subsystems

These are some useful commands for checking the different PSSP subsystems.

10.3.3.1 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 SP environment, it is especially used to obtain information about the status of the system partition-sensitive subsystems. To check if these subsystem are running on the CWS, you can use the `lssrc` with the `-a` option to get the status of all AIX subsystem, and then filter (`grep`) the result on the partition name. In our environment, the result is listed in Figure 111.

```
[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 111. Listing Status of System Partition-Sensitive Subsystems on the CWS

You can also use `lssrc` on SP nodes or to get detailed information about a particular subsystem. Figure 112 on page 284 shows a long listing of the status of the Topology Services subsystem on one of the SP 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 112. Listing Topology Services Information on Node sp3n06

10.3.3.2 syspar_ctrl -E

The `syspar_ctrl` command is the PSSP command providing control of the system partition-sensitive subsystems. In 9.2.11, “Start System Partition-Sensitive Subsystems” on page 258, we have seen that the `-A` option of this command adds and starts the subsystems.

The `syspar_ctrl -E` command displays (Examine) 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`.

10.3.4 Monitoring Hardware Status

This monitoring is done through the RS-232 line that connects the control workstation and each frame. From the control workstation the hardmon daemon uses a designated tty to connect to each frame supervisor card.

10.3.4.1 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 (serial links) between the CWS and the SP nodes. However, it also checks that the hardmon and sdr daemons are running, that it can communicate with the frame, and that the System Monitor has been correctly configured.

We recommend to use this command each time a new frame or node has been added to an SP system, after using the `spframe` command, to check that the new nodes have been correctly discovered by PSSP and that they have been taken into account in the SDR.

10.3.4.2 Monitoring Hardware Activity: spmon -d

The `spmon` command is a monitoring and control command. It can retrieve and display information about the hardware component of the SP system as well as act on them. We present here only a few options.

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. Figure 113 shows the result of this command on our CWS.

```
[sp3en0:/]# spmon -d -G
1.  Checking server process
    Process 16262 has accumulated 42 minutes and 14 seconds.
    Check ok

2.  Opening connection to server
    Connection opened
    Check ok

3.  Querying frame(s)
    1 frame(s)
    Check ok

4.  Checking frames

        Controller   Slot 17  Switch   Switch     Power supplies
Frame   Responds     Switch   Power    Clocking   A   B   C   D
-------------------------------------------------------------
  1       yes         yes      on         0       on  on  on  on

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
[sp3en0:/]#
```

Figure 113. spmon -d -G

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 288).

10.3.5 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 for one node the current value of the LED display.
- 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 windows stays open until you terminate the spled process.

10.3.6 Extracting SDR Contents

The SDR is the main repository for holding 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 in this section two of these commands. We strongly encourage you to refer to the *PSSP: Command and Technical Reference*, SA22-7351 and to read about these two commands as well as about all commands whose names start with SDR.

10.3.6.1 SDRGetObjects

The SDRGetObjects command extracts information about all objects in a class. For example, you can list the reliable hostname of all SP nodes:

```
[sp3en0:/]# SDRGetObjects Node reliable_hostname
reliable_hostname
sp3n01.msc.itso.ibm.com
sp3n05.msc.itso.ibm.com
sp3n06.msc.itso.ibm.com
sp3n07.msc.itso.ibm.com
sp3n08.msc.itso.ibm.com
sp3n09.msc.itso.ibm.com
sp3n10.msc.itso.ibm.com
sp3n11.msc.itso.ibm.com
sp3n12.msc.itso.ibm.com
sp3n13.msc.itso.ibm.com
sp3n14.msc.itso.ibm.com
sp3n15.msc.itso.ibm.com
[sp3en0:/]#
```

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
```

10.3.6.2 splstdata

The SDRGetObjects command is very powerful and is often used in SP 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 Section 9.2.4, “Check the Previous Installation Steps” on page 253. You must also know how to use:

splstdata -v	to display volume group information (PSSP 3.1 only)
splstdata -s	to access switch information
splstdata -h	to extract hardware configuration information
splstdata -i	to display node IP configuration
splstdata -e	to display site environment information

10.3.7 Checking IP Connectivity: ping/telnet/rlogin

The availability of IP communication between the CWS and the SP nodes is critical for the successful operation of the SP system. However, PSSP does not provide any tool to check the TCP/IP network since there is nothing specific to the SP in this area. Common TCP/IP commands can be used in the SP environment: ping, telnet, rlogin, traceroute, netstat, arp, and so on. These commands will return information for all IP connection 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.

10.3.8 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 114 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 three corresponds to commands we did not mention in this section: `st_verify`, `jm_install_verify`, and `jm_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
  Resource Manager Installation
  Resource Manager Configuration



F1=Help             F2=Refresh          F3=Cancel           F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 114. SMIT Verification Window

10.4 Graphical User interface

PSSP provides an alternative to the use of command line interface or SMIT panels for monitoring a system. You can use two graphical interfaces for that purpose. These interfaces are started by the commands: `spmon -g` and `perspectives`.

The `spmon -g` command is available in all versions of PSSP up to release 2.4. Although Perspectives has been avaliable since PSSP 2.2, all graphical tools used for management of an SP system are now accessible through the Perspectives Launch Pad and have been greatly enhanced in PSSP 3.1. In

PSSP 3.1, The `spmon -g` functionalities have been replaced with the SP Hardware Perspectives tool.

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 an SP system can be accessed through this interface. We, therefore, recommend that you refer to *SP Perspectives: A New View of Your SP*, SG24-5180, for further information about this tool. Another good source of information is the Perspectives on-line help available from the Perspectives Launch Pad.

The Perspective 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 have presented in previous sections:

- Monitoring of `hostsResponds`, `switchResponds`, `nodePowerLEDs`
- SMIT `SP_verify`
- `syspar_ctrl -E`

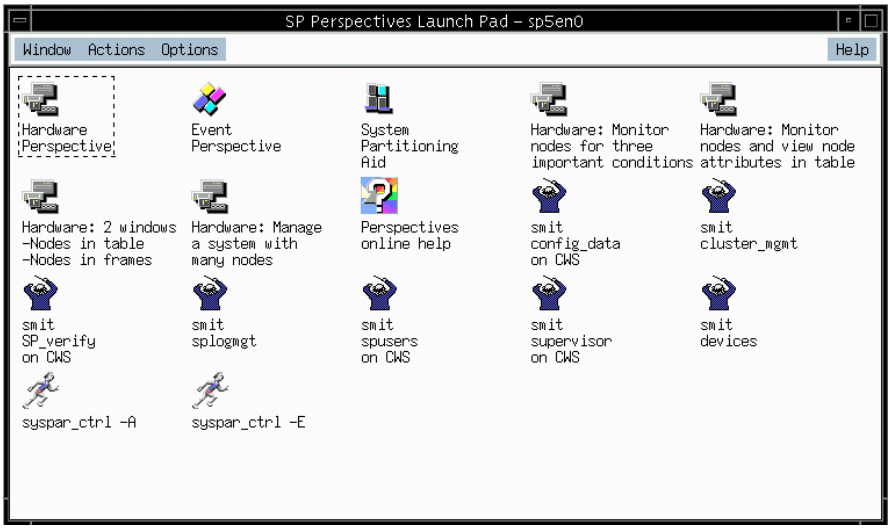


Figure 115. Perspectives Launch Pad

If you decide to perform most of your SP monitoring through the Perspectives tools, we recommend that you add your favorite tools to the Launch Pad.

10.5 Key Daemons

The management of an SP 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 they are running, and how they interact.

The SP related daemons are listed in Table 25 on page 290

Table 25. SP Daemons

Hardware monitoring	**hardmon, S70d**
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 provide below a very brief description of some of these daemons.

10.5.1 Sdrd

The sdrd daemon runs on the CWS. It serves all request from any client application to manipulate SDR information. It is managed using the AIX SRC commands. There is an entry for sdrd in the /etc/inittab, and sdrd is started at CWS boot time. This daemon must be running before any SP management action can be performed.

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 manages the serial port of the CWS that are connected to the SP frame. It controls all frames and node hardware through an SP specific protocol for communicating over the serial links. It also manages the S70d, which performs the hardware monitoring of non-SP frames over serial links. There is an entry for hardmon in the /etc/inittab, and it is started at CWS boot time.

No management of the SP 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 SP nodes in an SP system equipped with a switch. The worm 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*, GC23-3897 for PSSP 2.4 and *PSSP: Administration Guide*, SA22-7348 for PSSP 3.1. 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 282).

These subsystems are closely related. The Topology Services provides information about the SP systems to the Group Services, and 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 clients' applications of the Topology Services, Group Services, and Event Management subsystems.

We recommend that you refer to Chapter 22, 23, and 24 of the *PSSP:Administration Guide*, GC23-3897 for PSSP 2.4 and *PSSP: Administration Guide*, SA22-7348 for PSSP 3.1 for more details about the Topology Services, Group Services, and Event Management.

10.6 SP Specific Logs

Since SP systems are complex, the amount of data that an SP administrator may need to look at to manage such systems is far beyond what can be reasonably be gathered in one file or displayed in one screen.

The various components of PSSP therefore store information about their processing in several different logs. PSSP generates information in about 30 log files. A complete list of all these logs can be found on page 77, Chapter 4 "Error Logging Overview", of the *IBM Parallel System Support Programs for AIX: Diagnosis Guide*, GA22-7350.

Most of the SP related logs can be found in /var/adm/SPlogs on the CWS and on the SP nodes. A few other logs are stored in /var/adm/ras and /var/tmp/SPlogs.

You generally only look at logs for problem determination. For the purpose of this chapter (verifying the PSSP installation and operation), we will only mention the /var/adm/SPlogs/sysman directory. On each SP node, this directory contains the trace of the AIX and PSSP installation, their configuration, and the execution of the customization scripts described in Section 9.3.2, "/tftpboot" on page 268. 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 indication 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 Appendix C, "Related Publications" on page 471.

SP Manuals

The reader can refer to the related document for Version 2.4 of PSSP:

Installation and Migration Guide, GC23-3898 for PSSP 2.4 . The installation of an SP system is a long process involving several steps (up to 50 or more depending on the complexity of the system). Therefore, several verifications can be performed during installation to ensure that the already executed steps have been completed correctly. Chapter 2 of this guide documents the use of these verifications methods during the SP installation.

PSSP: Command and Technical Reference, GC23-3900, for PSSP 2.4 and *PSSP: Command and Technical Reference*, SA22-7351 for PSSP 3.1 contain a complete description of each command listed in 10.3, “Key Commands” on page 279.

Chapter 19 of *PSSP: Diagnosis and Messages*, GC23-3899 for PSSP 2.4 and Chapter 24 of *IBM Parallel System Support Programs for AIX: Diagnosis Guide*, GA22-7350 for PSSP 3.1 describe, in detail, the verification of System Management installation using the `SYSMAN_test` command.

The *Administration Guide*, GC23-3897 for PSSP 2.4 and *PSSP: Administration Guide*, SA22-7348 for PSSP 3.1. Chapter 14 describes the Switch related daemons, while Chapters 22, 23, and 24 provide you with detailed information about the partition-sensitive subsystems and their daemons.

SP Redbooks

RS/6000 SP Monitoring: Keeping It Alive, SG24-4873. Chapter 5 provides you with a detailed description of the Perspectives graphical user interface.

SP Perspectives: A New View of Your SP, 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 help aid you in preparation for the certification exam. The answers to these questions can be found in Appendix A.

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 setup 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 3.1?

   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 control workstation only. Which of the following statements is false?

   A. It uses the RS-232 lines to contact the frame supervisor cards.

   B. It is a partition-sensitive daemon.

   C. It requires read/write access to each tty connected to frames.

   D. It logs information in /var/adm/SPlogs/hardmon directory.

Part 3. Application Enablement

Chapter 11. Understanding Additional SP-Related Products

In addition to PSSP, several products are used in RS/6000 SP environment to provide workload management, connectivity, higher availability, and so on. This chapter provides an overview of some of these products.

11.1 Key Concepts You Should Know

Although most of these products are not essential for any SP installation, they are commonly found in customer environments. In preparation for the SP Certification exam, you should understand how the following products work and what solutions they provide:

- LoadLeveler
- Performance Toolbox Parallel Extension (PTPE)
- High Availability Control Workstation (HACWS)
- NetTAPE
- Client Input Output Socket (CLIO/S)

11.2 Understanding LoadLeveler

LoadLeveler is a software 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 standalone RS/6000 may be part of a cluster although LoadLeveler is most often run in the RS/6000 SP environment. A sample LoadLeveler cluster is shown in Figure 116.

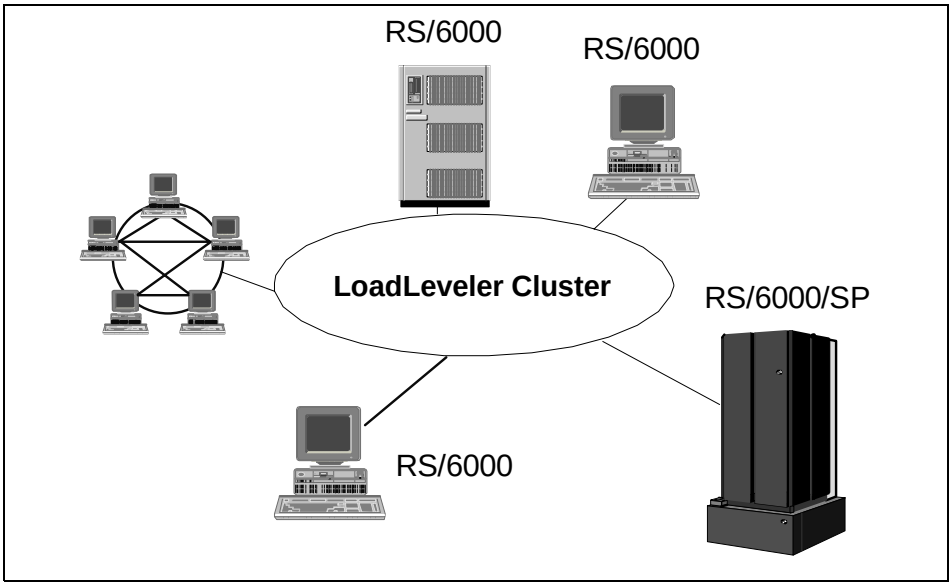


Figure 116. Example LoadLeveler Configuration

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 in to 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.

By way of example, Figure 117 on page 299 schematically illustrates a series of job steps. In this figure, data is read from storage in job step one.

Depending on the exit status of this operation, the job is either terminated or continues on to job step two. Again, LoadLeveler examines the exit status of job step two and either proceeds on to job step three, which, in this example, prints the data that the user requires or terminates.

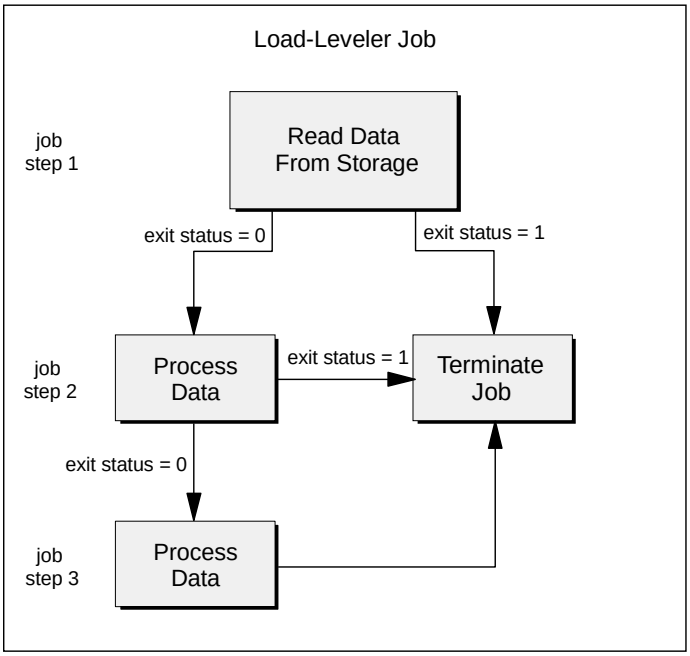


Figure 117. A LoadLeveler Job

11.2.1 A Breakdown of How It Works

There are three important functional machine types in LoadLeveler.

Scheduling machine. When a job is submitted to LoadLeveler, it gets placed in a queue that is managed by the scheduling machine. The latter then asks the central manager to find a machine that can process the job.

Central manager machine. This machine evaluates the resources required by the job that were specified in the job command file and selects a machine that is capable of running it. The central manager is also called the negotiator.

Executing machines. Machines that are assigned and run jobs.

Figure 118 on page 300 shows how these machine types fit together and the order in which they communicate.

1. A job has been submitted to LoadLeveler.
2. The scheduling machine contacts the central manager to inform it that a job has been submitted and to find out if there is a machine available that matches the job's requirements.
3. The central manager checks to determine if a machine exists that is capable of running the job. Once a machine is found, the central manager informs the scheduling machine which machine is available.
4. The scheduling machine contacts the executing machine and sends it the job information and executable program. The executing machine sends job status information to the scheduling machine and notifies it when the job has completed.

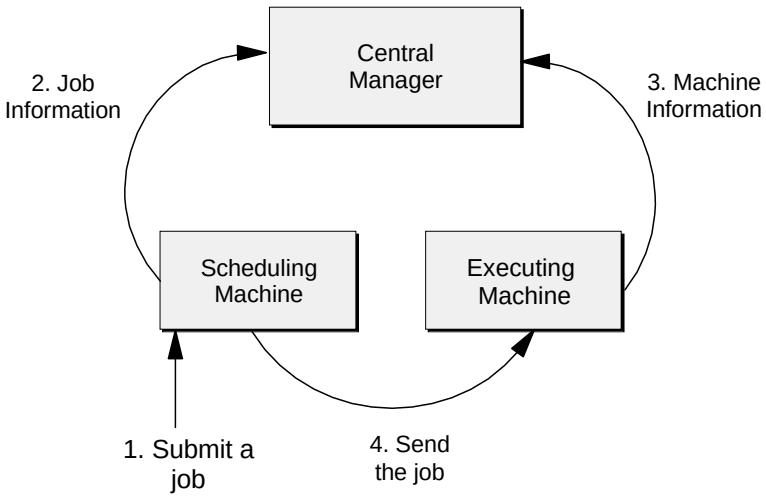


Figure 118. LoadLeveler Job Flow

In addition, there is another type of machine known as a submit-only machine. As its name indicates, this type of machine can only submit jobs although it is also able to query and cancel them.

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 will run before those with a low value**.**

The system administrator can specify several different 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 will be the job that has the higher user priority.

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 in the day, could run both. LoadLeveler can be configured to also take job class in to 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 run on a particular pool of machines. In this way, the right machines can be allocated the right jobs.

11.3 Understanding PTPE

The performance Toolbox is a performance analysis tool for standalone RS/6000 machines. PTPE is a parallel extension to this tools that enables performance monitoring and analysis on large SP systems. In addition to the capabilities of PTX/6000, PTPE provides:

- **Collection of SP-specific data**. PTPE provides ptpertm, an additional data supplier that complements the data xmservd collects. The

SP-specific performance data is currently implemented for:
- SP Switch
- LoadLeveler
- VSD

- **SP runtime monitoring**. The system administrator should have a global view of SP performance behavior. With reference to Figure 119 on page 302, similar nodes of the first tier, or Collectors, can be grouped and their performance data summarized by their respective Data Manager node in the second tier. This way, large SP systems can be easily monitored from a single presentation application by viewing node groups instead of individual nodes. The Data Managers are administered by the Central Coordinator in the third tier. The Central Coordinator aggregates the Data Managers' summary data to provide a total performance overview of the SP. Of course, the base PTX/6000 monitoring functions can be used to focus on any particular performance aspect of an individual node.

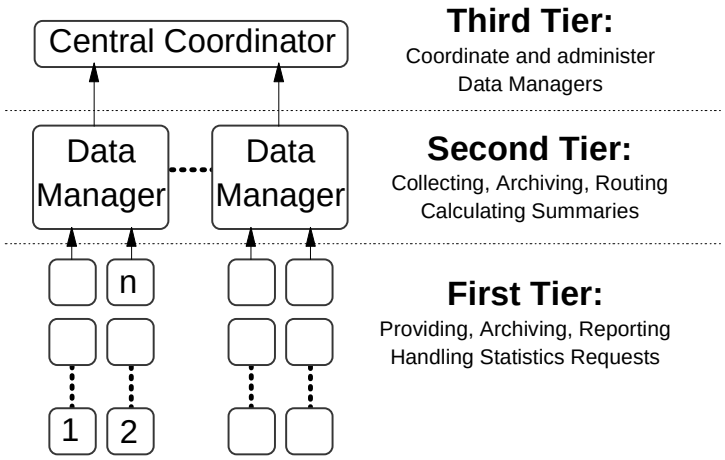


Figure 119. PTPE Monitoring Hierarchy

- **Data Analysis and Data Relationship Analysis**. PTPE provides an API to allow analysis applications to sift through all requested data. The data archive created by PTPE exists on every node and is completely accessible through the API. In base PTX/6000, performance data is analyzed with the azizo utility, which is restricted to simple derivatives, such as maximum, minimum, and average. With the PTPE API, programs of any statistical complexity can be written to find important trends or relationships. Also, data captured for azizo use is far more limited with base PTX/6000.

In PSSP 3.1 and later, PTPE is included in PSSP at no extra charge. For levels prior to PSSP 3.1, PTPE is a separately orderable and a priced feature of the PSSP LPP.

11.4 Understanding HACWS

HACWS is an optional collection of components that implement a backup CWS for an SP. The backup CWS takes over when the primary control workstation requires upgrade service or fails. The HACWS components are:

- A second RS/6000 machine supported for CWS use.
- 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, physically attached to each CWS, to allow access to data in the /spdata file system.

An HACWS cluster is depicted in Figure 120 on page 304.

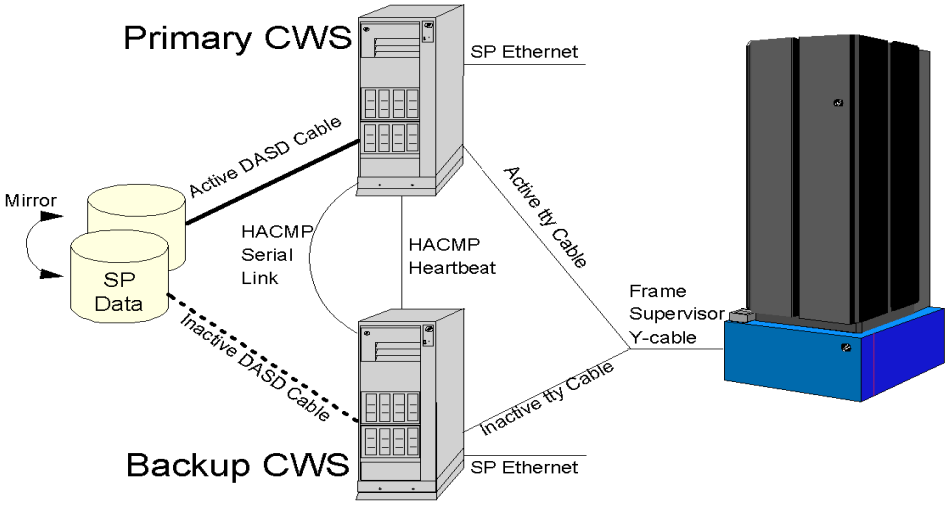


Figure 120. HACWS Cluster

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

11.5 Understanding NetTAPE

NetTAPE lets you manage a group of tape devices from a single workstation or multiple workstations using either a Motif/X-Window System-based graphical user interface or a set of commands.

NetTAPE can:

- Consolidate Control of Distributed Tape Operations

  NetTAPE provides a single system image of all of the network's tape devices. Tape device allocation, mount queue management, and tape device monitoring functions are performed using a graphical user interface.

- Customize Operator Views of Tape Operations

  NetTAPE allows you to assign each tape device to an operator domain. Using the NetTAPE GUI, operators limit the display of tape devices to those in their own domain. They see only the devices for which they are responsible.

- Use Tape Device Pools to Process Mount Requests More Efficiently

  NetTAPE lets you create pools of tape devices organized by device type. With device pools, mount requests for a certain type of device can be satisfied by any device in the pool. As a result, mount requests can be processed more quickly and efficiently.

- Support for Advanced Tape Devices and Features

  NetTAPE Tape Library Connection (NetTAPE TLC) supports advanced tape devices, such as the IBM 3494, 3495, and 3575 Tape Library Dataservers, and StorageTek Tape Libraries. It also supports the automatic cartridge loading functions of several types of tape devices. NetTAPE lets installations take advantage of the large capacity and automatic features of these tape devices in an AIX environment.

  With the use of ADSTAR Distributed Storage Manager (ADSM) device drivers, a myriad of SCSI-attached autochangers and libraries, from small 16 GB Autochangers to 14.4 TB libraries, are also supported.

- Coexistence with ADSM and CLIO/S

  NetTAPE works with ADSM for AIX Version 2.1 and can coexist on the same network with IBM's CLIO/S. The IBM 3494, 3495, and 3575 Tape Library Dataservers, StorageTek Tape Libraries, and SCSI-attached libraries and autochangers can be managed by NetTAPE and shared with ADSM allowing you to make better use of tape resources.

  Starting in Version 1, Release 2, NetTAPE TLC supports remote devices and esoteric device pools for ADSM. This eliminates the requirement that devices accessed by ADSM be physically located on the same node as the ADSM server.

11.6 Understanding CLIO/S

IBM Client Input Output/Sockets (CLIO/S) is a set of commands and application programming interfaces (API) that can be used for high-speed communication and for accessing tape devices on a network of AIX workstations and MVS mainframes. CLIO/S makes it easier to distribute work and data across a network of mainframes, workstations, and RS/6000 SP

systems. CLIO/S also provides an API to tape drives anywhere in your network. CLIO/S can be used to:

- Quickly move data between your MVS/ESA system and your workstation (or SP). For example, you can store large volumes of seismic data on tape and manage it using a mainframe acting as a data server to multiple workstations. This solution retains tape management as the responsibility of a single mainframe system while permitting seismic processing capacity to increase by distributing the work.
- Transfer very large files. For example, you can use applications on AIX to update customer files during the day, then use CLIO/S for fast backups to take advantage of MVS as a file server with extensive data management capabilities. Using CLIO/S for frequent file copying can mean shorter interruptions to your ongoing applications.
- Transfer files using familiar workstation commands. The CLIO/S CLFTP subcommands are similar to those of TCP/IP's ftp command; so, there's no need for users to learn a new interface. Users can even access tape data on MVS with the CLFTP subcommands.
- Access a tape drive on MVS from your workstation as though it were a local tape drive. For example, you can store data on MVS controlled tape drives and access it using CLIO/S connections to the compute servers.
- Start servers on other workstations and mainframes in your network to create a parallel processing environment. For example, CLIO/S can be used to schedule work on several workstations running in parallel. It also provides high data transfer rates and low processor utilization permitting very high parallel efficiency.
- Use AIX named pipes and BatchPipes/MVS. For example, you can access data on MVS (either on DASD or on tape) with an AIX named pipe. Or, an MVS program can use an MVS BatchPipe to send its output to AIX where another program using an AIX named pipe can do further processing.

11.7 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.

SP Manuals

Product manuals are very helpful for installing, configuring, and managing these products. If you are interested in installing and configuring these

products, you can consult the product manuals listed in Appendix C, “Related Publications” on page 471.

SP Redbooks

There are many redbooks that cover each one of these products in great detail. However, since the idea is to get an understanding only, we recommend the book *Inside the RS/6000 SP*, SG24-5145. This book covers most of the products in more detail than they appear here; so this redbook may be useful if you want to explore the product in greater depth.

11.8 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. 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.

Chapter 12. Application Specific Resources

Once PSSP has been configured and installed, you may need to install and configure additional products before you may start using your applications. These products, although some of them are not part of PSSP, are usually installed and configured in RS/6000 SP environments.

This chapter provides the basic concepts and setup procedures for understanding, installing, and configuring additional RS/6000 SP products.

12.1 Key Concepts You Should Study

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

- How the IBM Virtual Shared Disk works and what solution it provides.
- What are the filesets that are part of the VSD packaging and where they should be installed.
- How you create and configure VSD nodes and disks.
- How you manage VSD nodes and disks.
- How the IBM Recoverable Virtual Shared Disk works and what solution it provides.
- What are the hardware prerequisites for installing and configuring RVSD.
- What are the filesets that are part of the RVSD packaging and where they should be installed.
- How you set up and manage a RVSD environment.
- What is a General Parallel File System (GPFS).
- What are the hardware prerequisites for installing and configuring GPFS.
- How you configure and manage GPFS.
- What is NetTape.

12.2 IBM Virtual Shared Disks

The IBM Virtual Shared Disks (VSD) allows data stored in logical devices (logical volumes) to be access transparently from remote nodes. VSD is a thin layer of software that runs between the logical device and the Logical Volume Manager (LVM) as shown in Figure 121 on page 310.

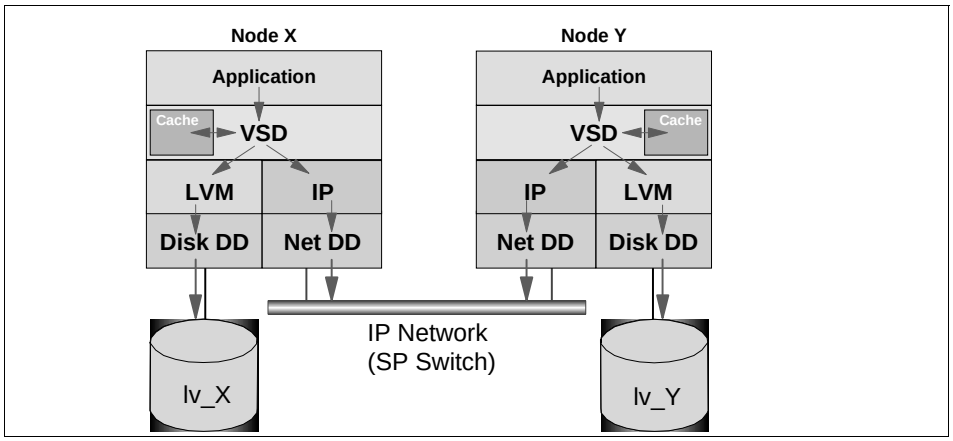


Figure 121. VSD Architecture

In Figure 121, there are two logical devices (lv_X and lv_Y). Each one is owned by Node X and Node Y respectively. When applications in Node X need to access lv_X, they will go trough the logical volume manager as usual for local access. However, when they need to access lv_Y, which is remote, the VSD layer will take the requirement and ship it through a TCP/IP network (in this case the SP Switch) to the disk server for lv_Y. For the application on Node X, both accesses were of the same kind (both access special devices named /dev/lv_X and /dev/lv_Y, respectively).

The nodes that manage physical disks are called *VSD Server*, and those that only access VSD disks are called *VSD Clients*. A VSD Server can be a VSD client.

In order to use VSD, it is necessary to install the VSD filesets on all the nodes that are going to be using or managing VSD disks. The VSD filesets have changed from PSSP 2.4 to PSSP 3.1, as shown in Table 26.

Table 26. VSD Filesets

PSSP 2.4 or below	PSSP 3.1	Description
ssp.csd.cmi	vsd.cmi	VSD SMIT Panels
ssp.csd.vsd	vsd.vsdd	VSD Device Driver
ssp.csd.hsd	vsd.hsd	VSD Hash Shared Disk
ssp.csd.sysctl	vsd.sysctl	VSD Sysctl Commands
ssp.csd.gui	ssp.vsdgui	VSD Perspective

PSSP 2.4 or below	PSSP 3.1	Description
ssp.csd.loc.ma_RP.gui	ssp.vsdgui.msg.ma_RP	VSD Perspective Locale Information
ssp.csd.msg.ma_RP.gui	ssp.vsdgui.msg.ma_RP	VSD Perspective Messages

12.2.1 Installing IBM Virtual Shared Disk

Before you install VSD into your nodes and control workstation, make sure that you are using the right level of AIX and PSSP. The VSD components have some prerequisites in terms of AIX and PSSP level as described in 8.6.1, “PSSP Prerequisites” on page 242.

The filesets involved are:

For the IBM Virtual Shared Disk component:

- vsd.vsdd
- vsd.sysctl
- vsd.cmi

For the Hashed Shared Disk component:

- vsd.hsd

If you are working with PSSP level older than PSSP 3.1, make the conversion to the correspondent fileset according to Table 26.

Note

The IBM Virtual Shared Disk Perspective component is in ssp.vsdgui. The PostScript file for VSD manual and the man pages for the related commands are contained in ssp.docs. They are in the ssp install image which should be installed on the control workstation.

The filesets to be installed are as follows:

On the control workstation:

- vsd.vsdd
- vsd.sysctl
- vsd.cmi
- ssp.vsdgui (if you want to use the VSD Perspective)

On the VSD client and server nodes:

- vsd.vsdd
- vsd.sysctl

If you are going to use HSD, then on HSD server and client nodes:

- vsd.hsd

12.2.2 Establishing Authorization

The IBM Virtual Shared Disk component uses `sysctl` for configuration and management. Your Kerberos principal has to be listed in the VSD ACL file in order to execute any VSD configuration command. The file /etc/sysctl.vsd.acl is shown in Figure 122 on page 312.

```
#acl#

# These are the users that can issue sysctl_vsdXXX command on this node
# Name must have a Kerberos name format which defines user@realm
# Please check your security administrator to fill in correct realm name
# you may find realm name from /etc/krb.conf

# _PRINCIPAL root@PPD.POK.IBM.COM
_PRINCIPAL root.admin@MSC.ITSO.IBM.COM
# _PRINCIPAL rcmd@PPD.POK.IBM.COM
# _PRINCIPAL userid@PPD.POK.IBM.COM
```

Figure 122. The sysctl.vsd.acl File

This file should be copied to all the nodes where VSD has been installed. Once copied, check that you have authorization to the VSD nodes.

To check your sysctl authorization, first run the `klist` command to look at your ticket and then run the sysctl `whoami` command and compare both:

```
[sp3en0:/]# klist
Ticket file:    /tmp/tkt0
Principal:      root.admin@MSC.ITSO.IBM.COM

  Issued           Expires          Principal
Dec  4 14:34:30  Jan  3 14:34:30  krbtgt.MSC.ITSO.IBM.COM@MSC.ITSO.IBM.COM
Dec  4 14:43:22  Jan  3 14:43:22  rcmd.sp3en0@MSC.ITSO.IBM.COM
Dec  4 14:43:43  Jan  3 14:43:43  hardmon.sp3en0@MSC.ITSO.IBM.COM
Dec  4 14:56:04  Jan  3 14:56:04  rcmd.sp3n01@MSC.ITSO.IBM.COM
Dec  4 14:56:04  Jan  3 14:56:04  rcmd.sp3n05@MSC.ITSO.IBM.COM
```

```
Dec  4 14:56:04  Jan  3 14:56:04  rcmd.sp3n06@MSC.ITSO.IBM.COM
Dec  4 14:56:04  Jan  3 14:56:04  rcmd.sp3n08@MSC.ITSO.IBM.COM
Dec  4 14:56:04  Jan  3 14:56:04  rcmd.sp3n07@MSC.ITSO.IBM.COM
Dec  4 14:56:04  Jan  3 14:56:04  rcmd.sp3n09@MSC.ITSO.IBM.COM
Dec  4 14:56:04  Jan  3 14:56:04  rcmd.sp3n10@MSC.ITSO.IBM.COM
Dec  4 14:56:05  Jan  3 14:56:05  rcmd.sp3n11@MSC.ITSO.IBM.COM
Dec  4 14:56:05  Jan  3 14:56:05  rcmd.sp3n13@MSC.ITSO.IBM.COM
Dec  4 14:56:05  Jan  3 14:56:05  rcmd.sp3n12@MSC.ITSO.IBM.COM
Dec  4 14:56:05  Jan  3 14:56:05  rcmd.sp3n14@MSC.ITSO.IBM.COM
Dec  4 14:56:05  Jan  3 14:56:05  rcmd.sp3n15@MSC.ITSO.IBM.COM
[sp3en0:/]# sysctl whoami
root.admin@MSC.ITSO.IBM.COM
```

To check that you can run VSD multinode commands, use the following command:

```
[sp3en0:/]# vsdsklst -n 1,15
>> sp3n01.msc.itso.ibm.com
Node Number:1; Node Name:sp3n01.msc.itso.ibm.com
    Volume group:rootvg; Partition Size:4; Total:537; Free:233
        Physical Disk:hdisk0; Total:537; Free:233
    Not allocated physical disks:
        Physical disk:hdisk1; Total:2.2
<<
>> sp3n15.msc.itso.ibm.com
Node Number:15; Node Name:sp3n15.msc.itso.ibm.com
    Volume group:rootvg; Partition Size:4; Total:958; Free:665
        Physical Disk:hdisk0; Total:479; Free:311
        Physical Disk:hdisk3; Total:479; Free:354
    Not allocated physical disks:
        Physical disk:hdisk1; Total:2.0
        Physical disk:hdisk2; Total:2.0
<<
```

This command lists information about physical and logical volume manager states as seen by the IBM Virtual Shared Disk software.

In this case, VSD have been installed and configured in node 1 and node 15.

12.2.3 Configuring

At this point in the installation, you are required to define and enter disk parameters for the VSD nodes into the System Data Repository (SDR).

This can be done through the vsdnode command or the IBM Virtual Shared Disk Perspective graphical interface (spvsd command). The syntax for the vsdnode command is as follows:

```
Usage: vsdnode node_number ... adapter_name init_cache_buffer_count
max_cache_buffer_count vsd_request_count rw_request_count
min_buddy_buffer_size max_buddy_buffer_size max_buddy_buffers
VSD_maxIPmsgsz
```

For example, to define and configure nodes 1 and 15, we should run the following command:

```
vsdnode 1 15 css0 256 256 256 48 4096 262144 2 61440
```

Or we may use the VSD Perspective as shown in Figure 123.

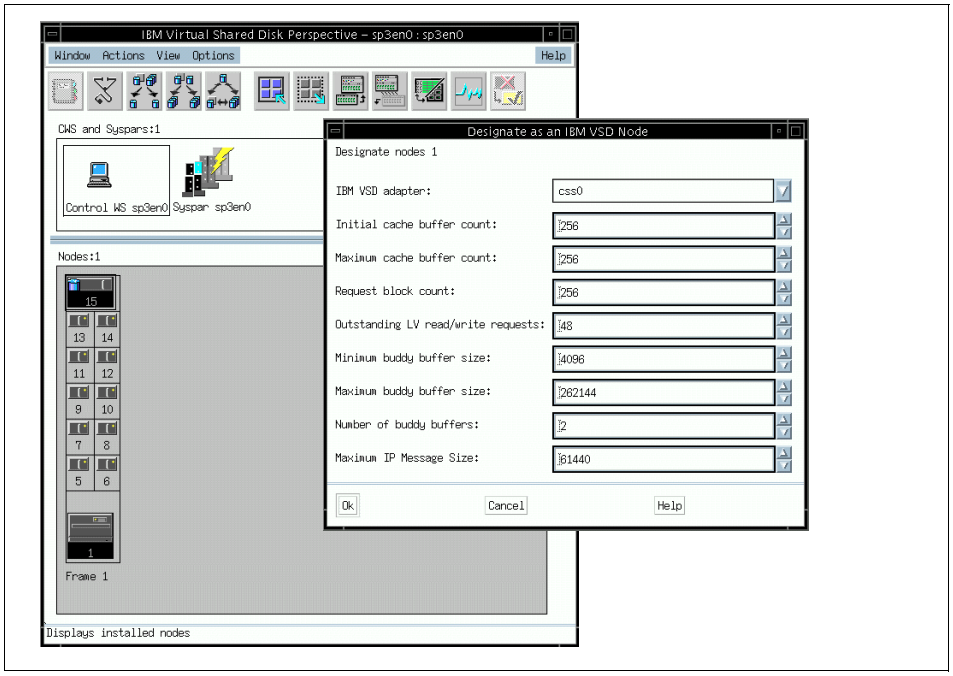


Figure 123. IBM Virtual Shared Disk Perspective

Once the nodes have been designated, we can start creating VSD disks on the designated nodes. To create a VSD disk, you have to decide first which volume group your are going to use. It can be rootvg or a global volume group you have previously created.

Volume groups used for virtual shared disks must be given a global name that is unique across system partitions.

This tasks is always done, but you do not have to always perform it. The **Create...** actions and the comparable `createvsd` and `createhsd` commands do

this for you. You only have to do this explicitly if you need to use the **Define...** action or the defvsd command to create your virtual shared disks because you already have logical volumes.

You can use the **Run Command...** action and run the vsdvg command to define global volume groups.

12.2.4 Creating Virtual Shared Disks

Remember, your procedure is based on whether or not you already have logical volumes. The **Create...** actions and commands take care of logical volumes and global volume groups for you. If you already have them, you must do the *define* steps instead.

If you are using VSD to create the logical volumes and define the global volume groups for you, then it is a good idea to check old rollback files. Refer to *IBM Parallel System Support Programs for AIX: Managing Shared Disks*, SA22-7349 for details on how to check old rollback files.

You can create a virtual shared disks with the graphical user interface action or a line command (on both primary and secondary nodes if you have the IBM Recoverable Virtual Shared Disk component running). You must first have used the IBM Virtual Shared Disk Perspective or the vsdnode command to set up information in the SDR about each node involved in this virtual shared disk configuration.

To create virtual shared disk using the IBM Virtual Shared Disk Perspective, launch the graphical interface using the spvsd command. Figure 124 on page 316 shows the initial start-up window.

In the main window select the **View->Add Pane** as shown Figure 125 on page 317. Once the new pane has been added, you can create virtual shared disks by selecting the **Create...** option from the Action menu.

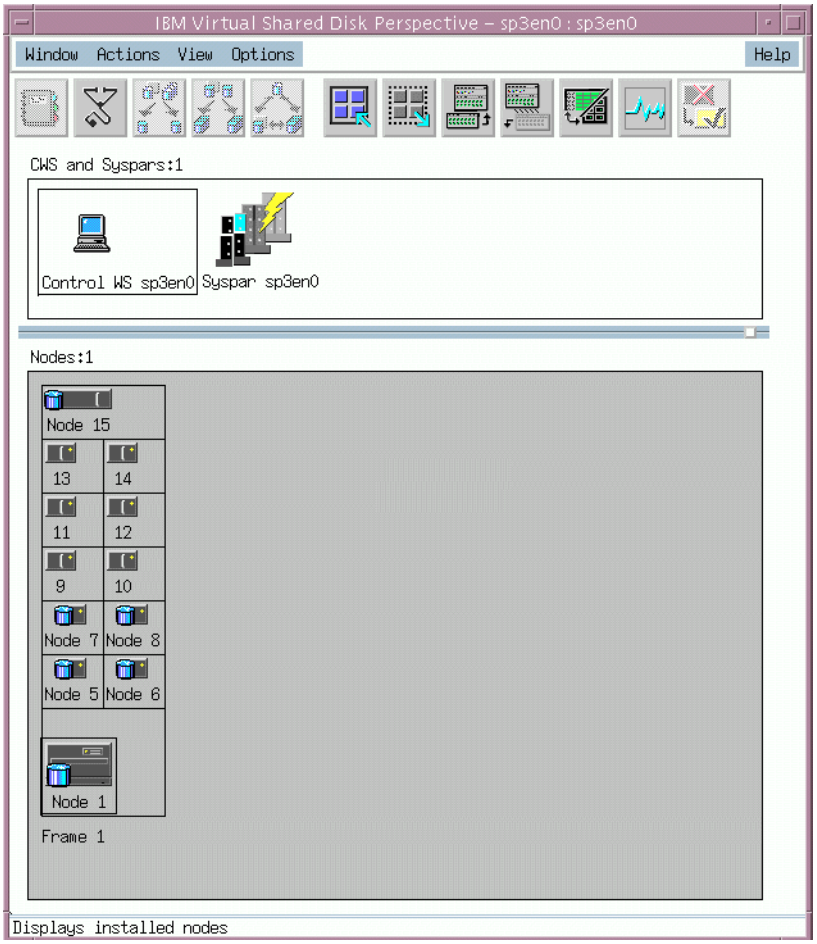


Figure 124. IBM Virtual Shared Disk Perspective (spvsd)

When creating virtual shared disks, you have to enter the pertinent information in the dialog box or as arguments to the `createvsd` command.

The information you need to enter is:

- The number of IBM VSDs per node.
- The IBM VSD name prefix.
- The logical volume name prefix.
- The volume group name.
- The IBM VSD size (MB).

- The mirroring count.
- The physical partition size (MB).
- Select the nodes that have been designated as virtual shared disk nodes, the primary node, and the backup node.
- Select the physical disks that the virtual shared disk is to span.

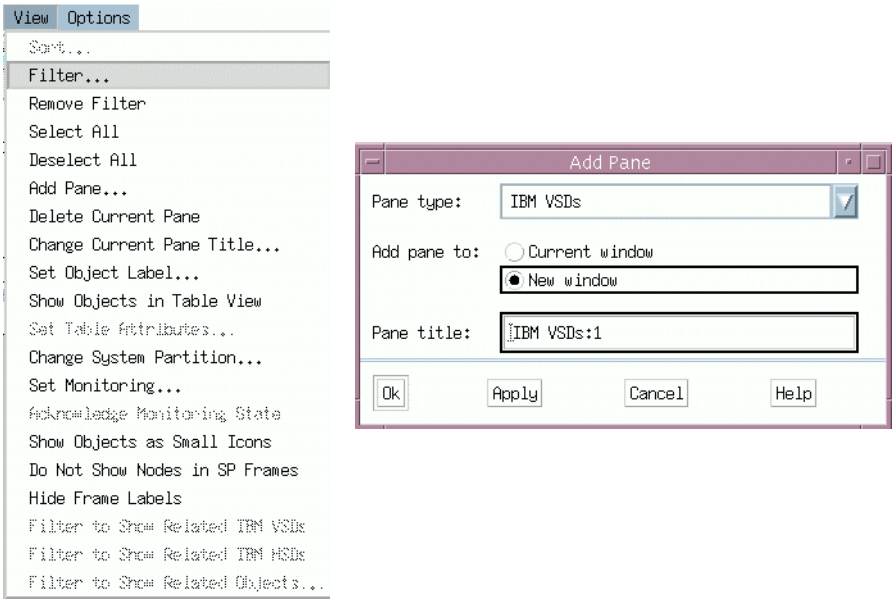


Figure 125. Adding a VSD Pane

The window for creating virtual shared disks is shown in Figure 126 on page 318.

If you prefer the command line interface instead, you can use the `createvsd` command as follows:

```
createvsd -n 1,15 -s 4 -g ITSOVG -v ITSOVSD
```

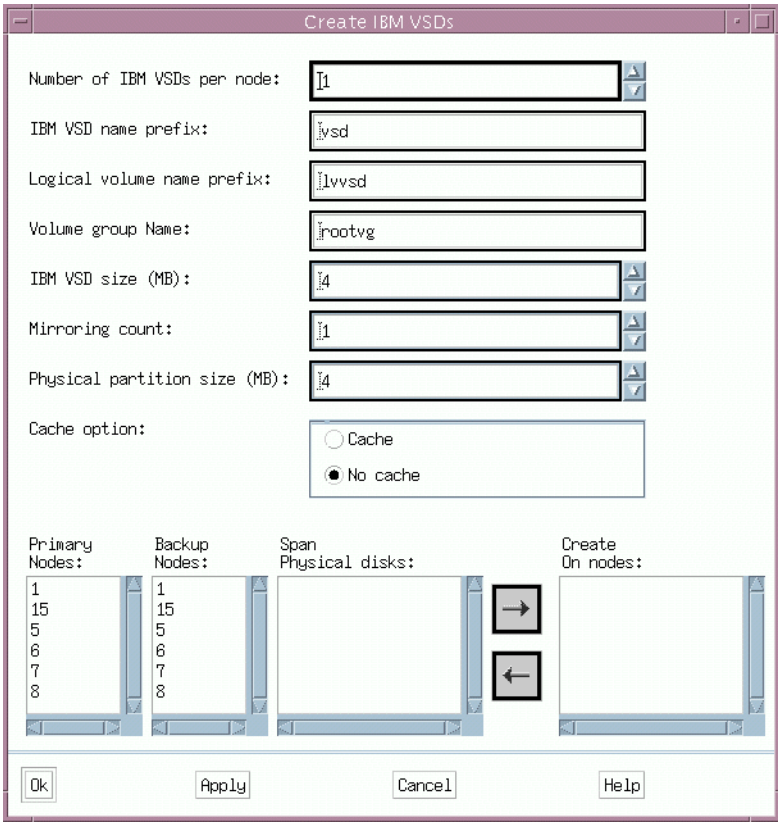


Figure 126. Creating Virtual Shared Disks

This creates the following virtual shared disk definitions:

- ITSOVSDn1 on node 1. The local volume group name on node 1 is ITSOVG. The global volume group name is ITSOVGn1. The logical volume is lvITSOVSDn1.
- ITSOVSDn15 on node 15. The local volume group name on node 15 is ITSOVG. The global volume group name is ITSOVGn15. The logical volume is lvITSOVSDn15.

This can be seen from the two nodes we have just configured and that created the virtual shared disks:

```
[sp3en0:/usr/lpp/vsd]# dsh -w sp3n01,sp3n15 lsvg
sp3n01: rootvg
sp3n01: ITSOVG
sp3n15: rootvg
sp3n15: ITSOVG
[sp3en0:/usr/lpp/vsd]# dsh -w sp3n01,sp3n15 lsvg -l ITSOVG
sp3n01: ITSOVG:
sp3n01: LV NAME             TYPE       LPs   PPs   PVs  LV STATE      MOUNT POINT
sp3n01: lvITSOVSD1n1        jfs        1     1     1    closed/syncd  N/A
sp3n15: ITSOVG:
sp3n15: LV NAME             TYPE       LPs   PPs   PVs  LV STATE      MOUNT POINT
sp3n15: lvITSOVSD2n15       jfs        1     1     1    closed/syncd  N/A
```

No secondary nodes are defined. The space allocated to a virtual shared disk is spread across all the physical disks (hdisks) within its local volume group on each node (1 and 15).

To assign each disk in the previous example a secondary node (with the IBM Recoverable Virtual Shared Disk component running), type:

```
createvsd -n 1/5/,15/6/ -s 4 -g ITSOVG -v ITSOVSD
```

This creates the following virtual shared disk definitions:

- ITSOVSDn1 on node 1 with a twin-tailed connection to node 5. The local volume group name on node 1 is ITSOVG. The global volume group name is ITSOVGn1. The logical volume is lvITSOVSD1n1.
- ITSOVSDn15 on node 15 with a twin-tailed connection to node 6. The local volume group name on node 15 is ITSOVG. The global volume group name is ITSOVGn15. The logical volume is lvITSOVSD2n15.

After you have created your virtual shared disks, you must configure them on all nodes that need to read from and write to them.

If you want recoverability, you should also have installed the IBM Recoverable Virtual Shared Disk software on each virtual shared disk node. In this case, you can use the actions from the Nodes pane **Control IBM RVSD subsystem...**, which will automatically configure and activate all the virtual shared disks as soon as quorum is met and activates recoverability on all the virtual shared disk nodes after you set the state to Initial Reset. If you prefer to use the command `ha_vsd reset`, you must run it on each virtual shared disk node.

To configure all the virtual shared disks, you can use the IBM Virtual Shared Disk Perspective (spvsd) graphical interface, or you can use the command `cfgvsd`. From the graphical interface, select the nodes you want to configure and then select **Configure IBM VSDs...** from the Actions menu. Figure 127 on page 320 shows the graphical window for configuring the virtual shared disks we previously defined.

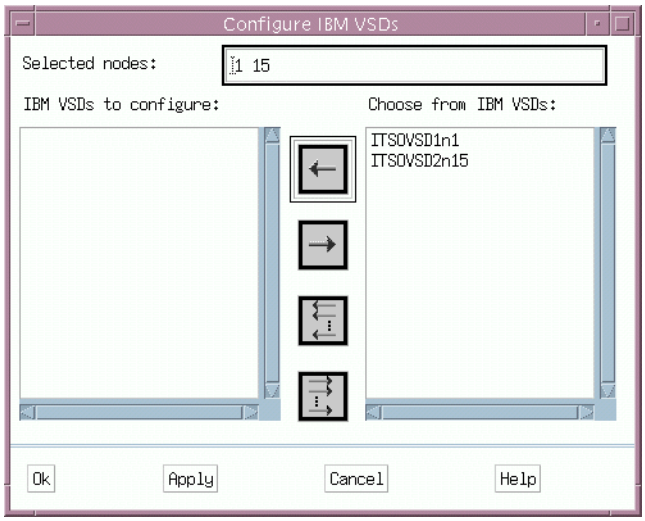


Figure 127. Configuring Virtual Shared Disks

12.2.5 Changing States of Virtual Shared Disks

After your virtual shared disks are configured, they are put into a *stopped* state. Before they can be of any good, you have to start them. If you are using the IBM Recoverable Virtual Shared Disk component, then this step is done automatically.

To check the status of your virtual shared disks, you may use the `lsvsd` command as follows:

```
# lsvsd -l
minor  state server lv_major lv_minor vsd-name                    option
size(MB)
 1      ACT    1      34       1      ITSOVSD1n1                  nocache         4
 2      ACT   15       0       0      ITSOVSD2n15                 nocache         4
```

The state column represents the state of the virtual shared disk.

Before you start your virtual shared disks, you have put the virtual shared disks you just configured into a suspended state. To do this, you use the `preparevsd` command. Once the virtual shared disks are in a suspended state, you can use the `resumevsd` command to make them active.

Figure 128 on page 321 shows all the possible states of a virtual shared disk and the transitions between states.

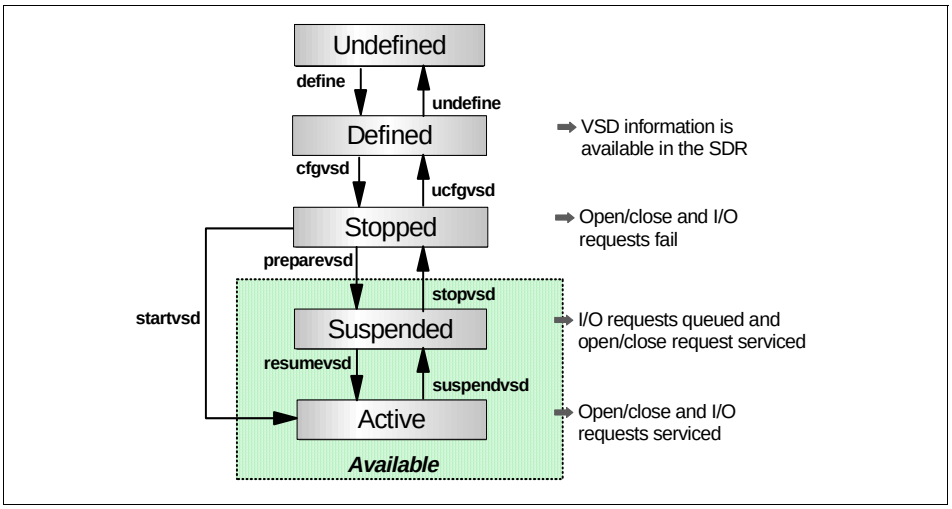


Figure 128. Virtual Shared Disk States and Associated Commands

12.3 IBM Recoverable Virtual Shared Disks

Recoverable Virtual Shared Disk (RVSD) adds availability to VSD. RVSD allows you to twin-tail disks, that is, physically connect the same group of disks to two or more nodes and provide transparent failover of VSDs among the nodes.

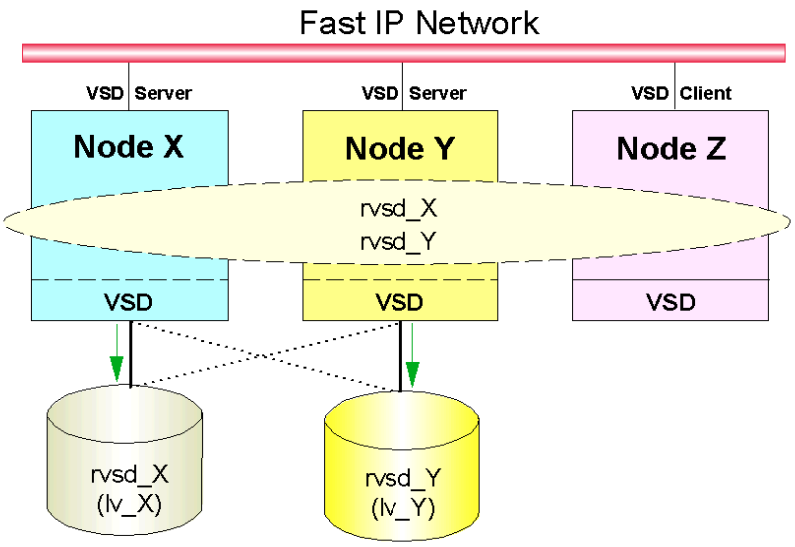


Figure 129. RVSD Function

With reference to Figure 129, Nodes X, Y, and Z form a group of nodes using VSD. RVSD is installed on Nodes X and Y to protect VSDs rvsd_X and rvsd_Y. Nodes X and Y physically connect to each other's disk subsystem where the VSDs reside. Node X is the primary server for rvsd_X and the secondary server for rvsd_Y and vice versa for Node Y. Should Node X fail, RVSD will automatically failover rvsd_X to Node Y. Node Y will take ownership of the disks, vary-on the volume group containing rvsd_X, and make the VSD available. Node Y serves both rvsd_X and rvsd_Y. Any I/O operation that was in progress and new I/O operations against rvsd_X are suspended until failover is complete. When Node X is repaired and rebooted, RVSD switches the rvsd_X back to its primary Node X.

RVSD subsystems are shown in Figure 130 on page 323. The rvsd daemon controls recovery. It invokes the recovery scripts whenever there is a change in the group membership. When a failure occurs, the rvsd daemon notifies all surviving providers in the RVSD node group so they can begin recovery. Communication adapter failures are treated the same as node failures.

The hc daemon is also called the Connection Manager. It supports the development of recoverable applications. The hc daemon maintains a membership list of the nodes that are currently running hc daemons and an incarnation number that is changed every time the membership list changes. The hc daemon shadows the rvsd daemon recording the same changes in state and management of VSD that rvsd records. The difference is that hc

only records these changes after rvsd processes them to assure that RVSD recovery activities begin and complete before the recovery of hc client applications takes place. This serialization helps ensure data integrity.

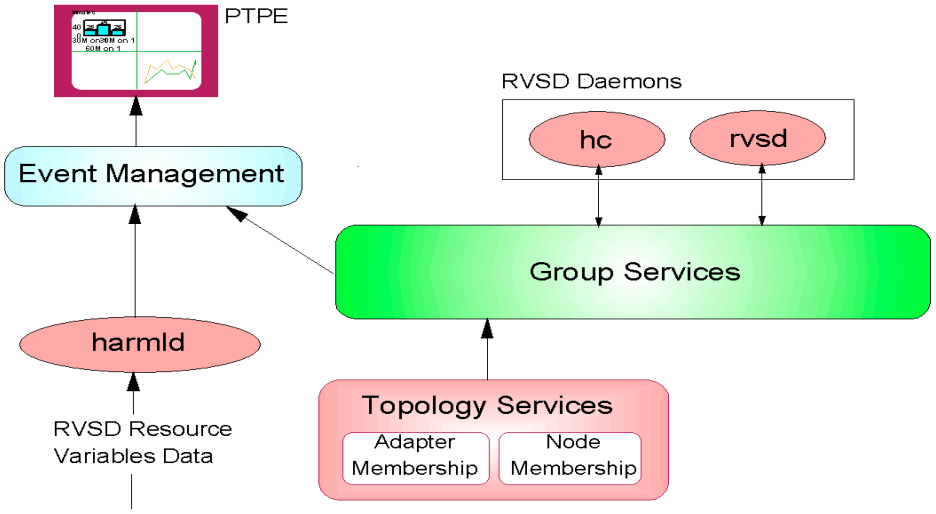


Figure 130. RVSD Subsystems and HAI

12.4 General Parallel File Systems

GPFS provides a standard, robust file system for serial and parallel applications on the SP. From a user's view, it resembles NFS, but unlike NFS, the GPFS file system can span multiple disks on *multiple nodes*. GPFS exploits VSD technology and the Kerberos-based security features of the SP and, thus, is only supported on SP systems.

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.

GPFS is implemented as kernel extensions, a multi-threaded daemon, and a number of commands. The kernel extensions are needed to implement the virtual file system layer that presents a GPFS file system to applications as a local file system. In the first version of GPFS (GPFS v1.1), the locking

mechanism, also called token management, was implemented as a kernel extension. In the second version currently available (GPFS v1.2), the token management facility has been moved to user space as part of the GPFS daemon (mmfsd).

The multi-threaded daemon provides specific functions within GPFS. Basically, the daemon provides data and metadata management (such as disk space allocation, data access, and disk I/O operations). It also provides security and quotas management.

The GPFS daemon runs on every node participating in the GPFS domain and may take on different *personalities*. Since GPFS is not the client-server type of file system, as NFS or AFS may be seen, it uses the concept of VSD servers, which are nodes physically connected to disks. Each node running GPFS (including VSD servers) will use the virtual shared disk extensions to access the data disks.

GPFS works within a system partition, and the node in this partition running GPFS will be able to access any defined GPFS file system. In order to access the file systems created in GPFS, nodes need to mount them like any other file system. To mount the file systems, nodes have two options:

- Nodes running GPFS

  For these nodes, mounting a GPFS file system is the same as mounting any local (JFS) file system. The mounting has no syntax difference with the local mounting done with JFS. At creation time, GPFS file systems can be set to be mounted automatically when the nodes start up.

- Nodes not running GPFS

  For these nodes, GPFS file system can be made available through NFS. Nodes running GPFS, and after mounting the file systems, can NFS export them. The same applies to any NFS-capable machine.

12.4.1 Requirements

GPFS environment is specific to AIX on the RS/6000 SP. Various software requirements must be installed and configured correctly before you can create a GPFS file system.

12.4.1.1 Hardware Requirements

GPFS runs only on the RS/6000 SP, and the switch must be installed and configured. Altough GPFS does not required twin-tailed or SSA loops of disks, it is recommended to install such configurations in order to provide higher data availability at the hardware level.

12.4.1.2 Software Requirements

There are two versions of GPFS available at the time this redbook is being written. GPFS v1.1 requires PSSP 2.4, which requires AIX 4.2.1 or AIX 4.3. GPFS v1.2 requires PSSP 3.1, which, in turn, requires AIX 4.3.2.

GPFS also requires the IBM Virtual Shared Disk and the IBM Recoverable Virtual Shared Disk products, which level are defined by the level of PSSP installed. So, if PSSP 2.4 is installed, VSD and RVSD Version 2.1.1 are required. If PSSP 3.1 is used, then VSD and RVSD 3.1 are required.

GPFS requires RVSD even though your installation does not have twin-tailed disks or SSA loops for multi-host disk connection.

12.4.2 Configuring GPFS

Chapter 2 of *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278 is devoted to a series of steps in planning GPFS. It is recommended that this section be read and understood prior to installing and using GPFS.

GPFS tasks cannot be done on the CWS; they must be performed on one of the GPFS nodes.

There are three areas of consideration when GPFS is being setup: The nodes using GPFS, the VSDs to be used, and the FS to be created. Each area is now examined. A sample FS setup consisting of four nodes is provided. Nodes 12, 13, and 14 are GPFS nodes, while node 15 is the VSD server node.

Warning

Do not attempt to start the mmfsd daemon prior to configuring GPFS. Starting the mmfsd daemon without configuring GPFS causes dummy kernel extensions to be loaded, and you will be unable to create a FS. If this occurs, configure GPFS and then reboot the node(s).

Carry out the following procedures to configure GPFS, then start the mmfsd daemon to continue creating the FS.

Nodes

The first step in setting up GPFS is to define which nodes are GPFS nodes. The second step is to specify the parameters for each node.

There are three areas where nodes can be specified for GPFS operations: Node count, node list, and node preferences.

The Node Count is an estimate of the maximum number of nodes that will mount the FS and is entered into the system only *when the GPFS FS is created*. It is recommended to overestimate this number. This number is used in the creation of GPFS data structures that are essential for achieving the maximum degree of parallelism in file system operations. Although a larger estimate consumes a bit more memory, insufficient allocation of GPFS data structures can limit a node's ability to process certain parallel requests efficiently, such as the allotment of disk space to a file. If it is not possible to estimate the number of nodes, apply the default value of 32. A larger number may be specified if more nodes are expected to be added. However, it is important to avoid wildly overestimating since this can affect buffer operations. This value cannot be changed later. The FS must be destroyed and recreated.

A node list is a file that specifies to GPFS the actual nodes to be included in the GPFS domain. This file may have any file name. However, when GPFS configures the nodes, it copies the file to each GPFS node as /etc/cluster.nodes. The GPFS nodes are listed one per line in this file, and the switch interface is to be specified because this is the interface over which GPFS runs.

Figure 131 on page 327 is an example of a node list file. The file name in this example is /var/mmfs/etc/nodes.list.

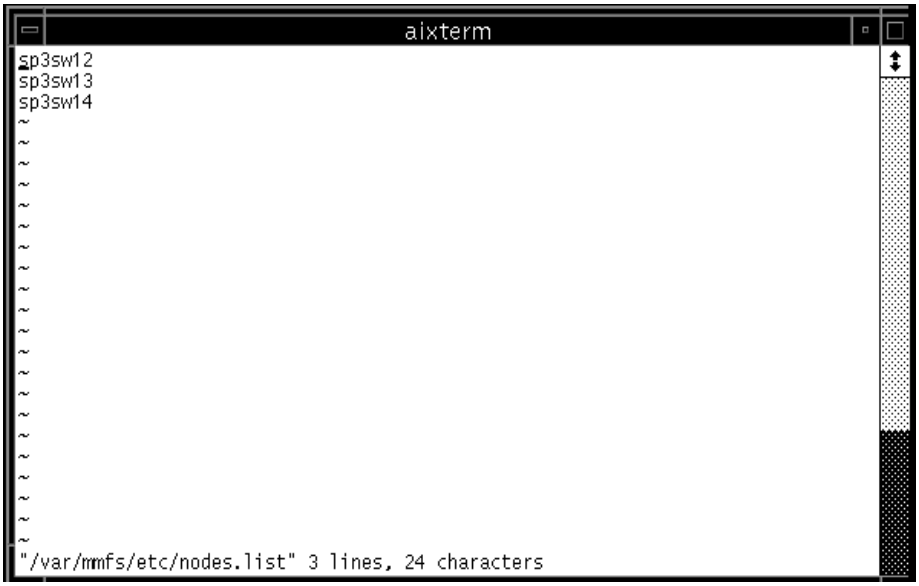


Figure 131. Sample Node List File

Having chosen the nodes that form the GPFS domain, there is the option to choose which of these nodes are to be considered for the personality of stripe group manager. There are only three nodes in the GPFS domain in this example; so, this step is unnecessary. However, if there are a large number of nodes in the GPFS domain, it may be desirable to restrict the role of stripe group manager to a small number of nodes. This way, if something happens and a new stripe group manager has to be chosen, GPFS can do so from a smaller set of the nodes (the default is every GPFS node). To carry this out, follow the format for creating a node list to create the file /var/mmfs/etc/cluster.preferences (this file name must be followed).

To configure GPFS, you can use SMIT panels or the `mmconfig` command. The `mmconfig` command is further described in *General Parallel File System by AIX: Installation and Administration Guide*, SA22-7278. The SMIT panel maybe accessed by typing `smit gpfs` and then selecting the **Create Startup Configuration** option. Figure 132 on page 328 shows the SMIT panel used to configure GPFS (this is being run on node 12 in our example). This step needs to be run on only one node in a GPFS domain.

Figure 132. SMIT Panel for Configuring GPFS

It is possible to configure GPFS to automatically start on all nodes whenever they come up. Simply specify yes to the autoload option in the SMIT panel or the -A flag in the mmconfig command. This eliminates the need to manually start GPFS when nodes are rebooted.

The pagepool and mallocsize options specify the size of the cache on each node dedicated for GPFS operations. mallocsize sets an area dedicated for holding GPFS control structures data, while pagepool is the actual size of the cache on each node. In this instance, pagepool is specified to the default size of 4 M while mallocsize is specified to be the default of 2 M where M stands for megabytes and must be included in the field. The maximum values per node are 512 MB for pagepool and 128 MB for mallocsize.

The priority field refers to the scheduling priority for the mmfsd daemon. The concept of priority is beyond the scope of this redbook. Please refer to AIX documentations for more information.

Notice the file /usr/lpp/mmfs/samples/mmfs.cfg.sample. This file contains the default values to be used to configure GPFS if none are specified either through the fields in the SMIT panel or in another file. The use of another file to set GPFS options may appeal to more experienced users or those who want to configure multiple GPFS domains with the same parameters. Simply copy this file (/usr/lpp/mmfs/samples/mmfs.cfg.sample) to a different file, make the changes to according to your specifications, propagate it out to the nodes, and configure it using SMIT or the `mmconfig` command.

Further information, including details regarding the values to set for `pagepool` and `mallocsize`, is available in the manual *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278.

Once GPFS has been configured, mmfsd has to be started on the GPFS nodes before a FS can be created. Here are the steps to do so:

1. Set the WCOLL environment variable to target all GPFS nodes for the `dsh` command. *PSSP: Administration Guide*, SA22-7348, *PSSP: Command and Technical Reference*, SA22-7351, and *IBM RS/6000 SP Management, Easy, Lean, and Mean*, GG24-2563 all contain information on the WCOLL environment variable.
2. Designate each of the nodes in the GPFS domain as an IBM VSD node.
3. Ensure that the rvsd and hc daemons are active on the GPFS nodes.

Note

It is necessary to have set up at least one VSD. The rvsd and hc do not start unless they detect the presence of one VSD defined for the GPFS nodes. This VSD may or may not be used in the GPFS FS; the choice is up to you.

4. Start the mmfsd daemon by running it on one GPFS node:

   ```
   dsh startsrc -s mmfs
   ```

   The mmfsd starts on all the nodes specified in the /etc/cluster.nodes file. If the startup is successful, the file /var/adm/ras/mmfs.log* looks like Figure 133 on page 330.

*Figure 133. Sample Output of /var/adm/ras/mmfs.log**

VSDs

Before the FS can be created, the underlying VSDs must be setup. The nodes with the VSDs configured may be strictly VSD server nodes, or they can also be GPFS nodes. The application needs to be studied, and a decision needs to be made as to whether the VSD server nodes are included in the GPFS domain.

A decision also needs to be made regarding the level of redundancy used to guard against failures. Should the VSDs be mirrored? Should they run with a RAID subsystem on top? Should RVSD be used in case of node failures? Again, this depends on the application, but it can also depend on your comfort and preferences with dealing with risk.

In addition to these options, GPFS provides two further recovery strategies at the VSD (disk) level.

GPFS organizes disks into a number of failure groups. A failure group is simply a set of disks that share a common point of failure. A common point of failure is defined as that which, if it goes down, causes the set of disks to

become simultaneously unavailable. For example, if a VSD spans two physical disks within one node, the two disks can be considered a failure group because if the node goes down, both disks become unavailable.

Recall that there are two types of data that GPFS handles: Metadata and the data itself. GPFS can decide what is stored into each VSD: Metadata only, data only, or data and metadata. It is possible to separate metadata and data to ensure that data corruption does not affect the metadata and vice versa. Further, it can impact performance. This is best seen if RAID is involved. RAID devices are not suited for handling metadata because metadata is small in size and can be handled using small I/O block sizes. RAID is most effective at handling large I/O block sizes. Metadata can, therefore, be stored in a non-RAID environment, such as mirrored disks, while the data can be stored in a RAID disk. This protects both data and metadata from each other and maximizes the performance given that RAID is chosen.

Once the redundancy strategy has been adopted, there are two choices to creating VSDs: Have GPFS do it for you or manually create them. Either way, this is done through the use of a Disk Descriptor file. This file can be manually set up or done through the use of SMIT panels. If using SMIT, run `smit gpfs` and then select the **Prepare Disk Descriptor File** option. Figure 134 on page 332 shows the SMIT panel for our example.

In this case, the VSD vsd1n15 has already being created on node 15 (sp3n15). *Do not* specify a name for the server node because the system has all of the information in needs from the configuration files in the SDR. In addition, the VSD(s) must be in the Active state on the VSD server node and all the GPFS nodes prior to the GPFS FS creation.

If the VSDs have not been created, specify the name of the disk (such as hdisk3) in the disk name field instead of vsd1n15 and specify the server where this hdisk is connected. GPFS then creates the necessary VSDs to create the FS.

The failure group number may be system generated or user specified. In this case, a number of 1 is specified. If no number is specified, the system provides a default number that is equal to the VSD server node number + 4000.

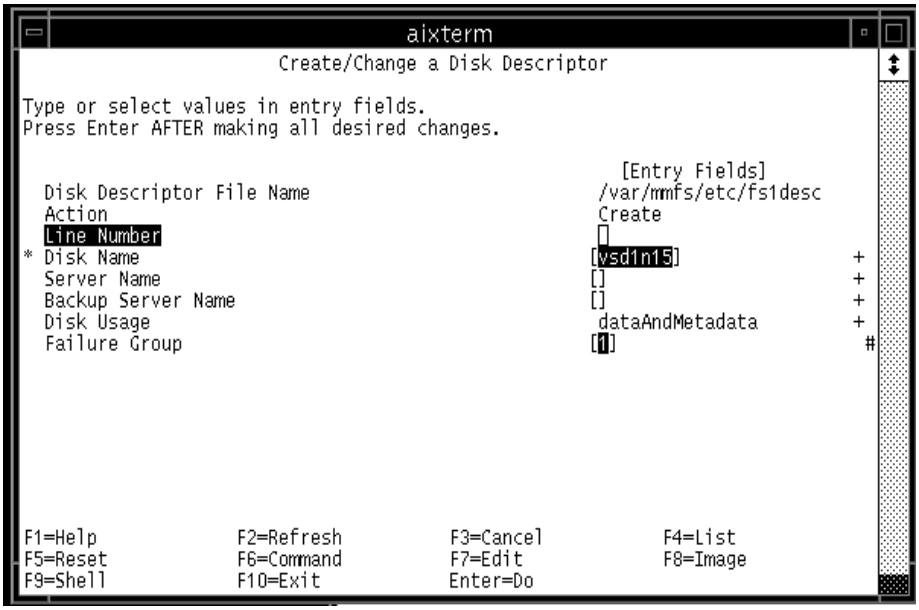


Figure 134. SMIT Panel for Creating Disk Descriptor File

File System

There are two ways to create a GPFS FS: Using SMIT panels or the `mmcrfs` command. Figure 135 on page 333 shows the SMIT panel to be used. This is accessed by running `smit gpfs` and then selecting the **Create File System** option. Details on `mmcrfs` can be found in *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278.

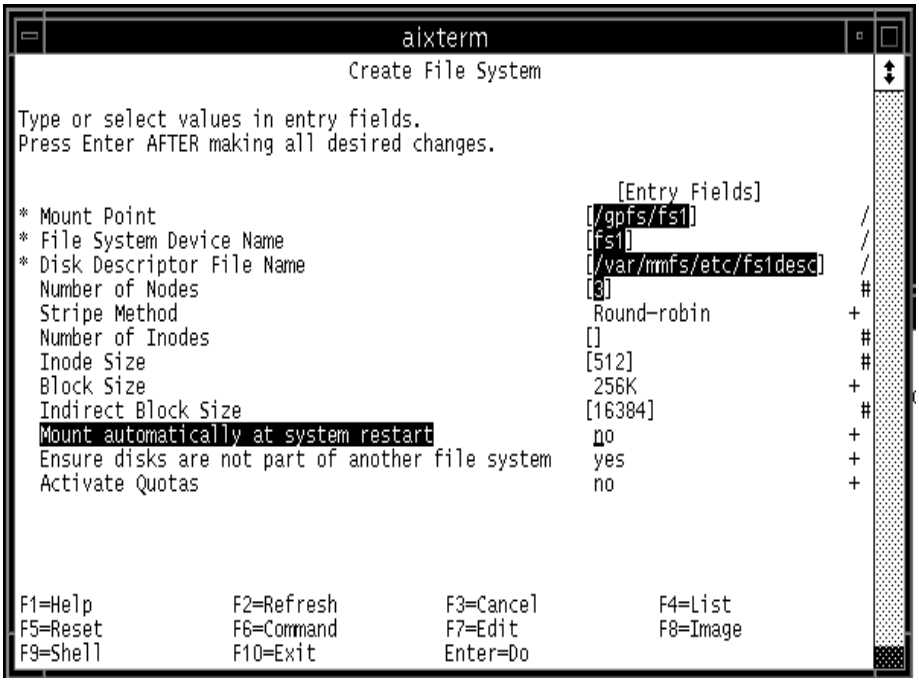


Figure 135. SMIT Panel for Creating a GPFS FS

Before creating the FS, several decisions have to be made:

- Decide how to structure the data in the FS.

  There are three factors to consider for structuring the data in the FS: Block size, i-node size, and indirect block size.

  GPFS offers a choice of three block sizes for I/O to and from the FS: 16 KB, 64 KB, or 256 KB. Familiarity with the applications running on your system will help you determine which block size to use. If the application handles large amounts of data in a single read/write operation, then a large block size may prove more suitable. If the size of the files handled by the application is small, a smaller block size may be more suitable. The default is 256 KB.

  GPFS further divides each block of I/Os into 32 sub-blocks. If the block size is the largest amount of data that can be accessed in a single I/O operation, the sub-block is the smallest unit of disk space that can be allocated to a file. For a block size of 256 KB, GPFS reads as much as 256

KB of data in a single I/O operation, and a small file can occupy as little as 8 KB of disk space.

Files smaller than one block size are stored in fragments, which are made up of one or more sub blocks. Large files are, therefore, often stored in a number of full blocks plus one or more fragments to hold the data at the end of the file.

The i-node is also known as the file index. It is the internal structure that describes an individual file to AIX holding such information as file size and the time of the last modification to the file. In addition, an i-node points to the location of the file on the hard disk. If the file is small, the i-node stores the addresses of all the disk blocks containing the file data. If the file is large, i-nodes point to indirect blocks that point to the disk blocks storing the file data (indirect blocks are set aside to specifically only hold data block addresses).

The default size of an i-node is 512 bytes. This number can increase to 4 KB depending on the size of the files the application uses.

An indirect block can be as small as a single sub-block or as large as a full block (up to an absolute maximum of 32 KB). The only additional requirement is that the value of an indirect block is a multiple of the size of a sub-block.

It is also possible to specify the number of i-nodes to limit the maximum number of files that can be created in the FS. In older versions of GPFS, the maximum number of i-nodes is set at GPFS FS creation time and cannot be changed after. At GPFS v1.2, it is now possible to set a limit at FS creation time, and if it proves necessary, change this upper limit. The upper limit is changed by the `mmchfs` command and the exact syntax can be found in *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278.

- Decide the striping method.

  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 may be set when a GPFS FS is first created or can be modified as a FS parameter later on.

  The three algorithms are now detailed:

  - Round Robin

    This is the default option the system chooses. Data blocks are written to one VSD at a time until all the VSDs in the FS have received a data

block. The next round of writes will then write a block to each VSD *in exactly the same order*.

This method yields the best write performance. There is, however, a penalty when a disk is added or removed from the FS. When a disk is added or removed from the FS, a re-striping occurs. The round Robin method takes the longest amount of time among the three algorithms to handle this re-striping.

- Balanced Random

  This method is similar to round Robin. When data blocks are written, one block is written to each VSD. When all the VSDs have received one block of data, the round begins. However, in balanced Random, the order in the second round is not the same as the first round. Subsequent rounds are similarly written to all VSDs but in an order different than that of the previous round.

- Random

  As its name implies, there is no set algorithm for handling writes. Each data block is written to a VSD according to a random function. If data replication is required, GPFS does ensure that both copies of the date are not written to the same disk.

- Decide whether to use GPFS Quotas or not.

  GPFS quotas define the amount of space in the FS that a user or a group of users is allowed to use. There are three parameters with which quotas operate: Soft limit, hard limit, and grace period.

  The hard limit is the maximum disk space and files that a user or group can accumulate. A soft limit are the levels below which a user or group can safely operate. A grace period is only used for soft limits and define a period of time in which a user or group can exceed the soft limit.

  The usage and limits data are stored in the files quota.user and quota.group files that reside in the root directories of GPFS FSs.

  In a quota-enabled configuration, one node is automatically nominated as the quota manager whenever GPFS is started. The quota manager allocates disk blocks to the other nodes writing to the FS and compares the allocated space to the quota limits at regular intervals. In order to reduce the need for frequent space requests from nodes writing to the FS, the quota manager allocates more disk blocks than requested.

  Quotas can be turned on by switching the Activate Quotas entry as shown in Figure 135 on page 333 to `Yes` or by specifying the `-Q` yes flag for the `mmcrfs` command.

Quotas are further discussed in *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278.

- Decide whether to replicate the files or not.

  At the FS level, GPFS provides an option to have additional copies of data and metadata be stored on the VSDs. This is above and beyond disk mirroring. Therefore, with both replication and mirroring turned on, it is possible to have a minimum of four copies of data being written.

  It is possible to replicate metadata, data, or both. The parameters for this are Max Meta Data Replicas and Max Data Replicas, which control the maximum factors of replication of metadata and data (respectively), and Default Meta Data Replica and Default Data Replicas, the actual factors of replication. Acceptable values are 1 or 2. 1 is the default and means no replication (only one copy), and 2 means replication is turned on (two copies). The Default values must be less than or equal to the Max values. In other words, the Max values grant permission for replication, while the Default values turn the replication on or off.

  Replication can be set at FS creation time and *cannot* be set through SMIT panels. The only way to turn on replication is with the command `mmcrfs` and the flags `-M` for Max Metadata Replicas, `-m` for Default Metadata Replicas, `-R` for Max Data Replicas, and `-r` for Default Data Replicas. Using the same example in Figure 135 on page 333, we can create a FS with both metadata and data replication turned on:

  ```
  mmcrfs /gpfs/fs1 fs1 -F /var/mmfs/etc/fs1desc -A yes -B 256K -i 512 -I
  16K -M 2 -m 2 -n 3 -R 2 -r 2 -v yes
  ```

  More information on these flags and the `mmcrfs` command can be found in *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278.

Once a GPFS FS has been setup, it can be mounted or unmounted on the GPFS nodes using the AIX `mount` and `umount` commands. Or, you can use the SMIT panel by running `smit fs` and then selecting **Mount File System**. Figure 136 on page 337 shows the SMIT panel for mounting a FS:

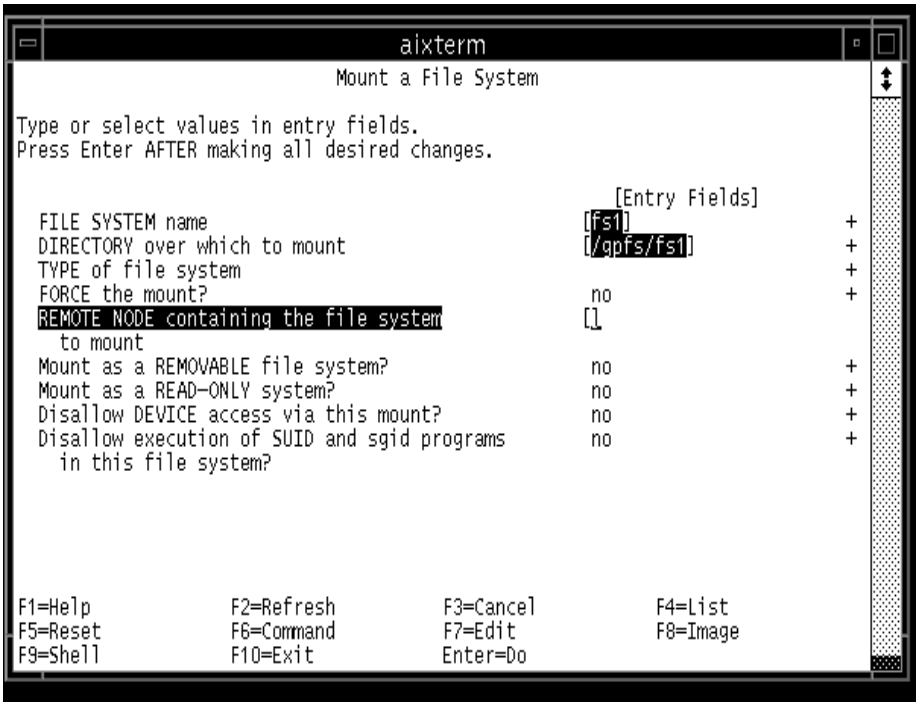


Figure 136. SMIT Panel for Mounting a File System

12.4.3 Managing GPFS

Once a GPFS FS has been set up, there are a number of tasks that can be performed to manage it. Some of the tasks and the commands to execute them are included here for reference. Note that SMIT panels are available as well to execute the commands. The commands and the SMIT panels are further described in the manual *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278.

Changing the GPFS Configuration

It is possible to change the configuration of GPFS for performance tuning purposes. The command to do so is `mmchconfig`, and it is capable of changing the following attributes:

1. pagepool
2. data Structure Dump
3. mallocsize

4. maxFiles To Cache
5. priority
6. autoload

Changes to pagepool may take effect immediately if the **-i** option is chosen; otherwise, the changes will take effect the next time GPFS is started. Changes to data Structure Dump, mallocsize, maxFiles To Cache, and priority require a re-start of GPFS. Changes to autoload require a re-boot of the nodes where these are affected.

For example, to immediately change the size of pagepool to 60 MB, run:

```
mmchconfig pagepool=60M -i
```

It is also possible to add and delete nodes from a GPFS configuration. The commands to do so are `mmaddnode` and `mmdelnode`. Be careful when adding or subtracting nodes from a GPFS configuration. GPFS uses quorum to determine if a GPFS FS stays mounted or not. It is easy to break the quorum requirement when adding or deleting nodes. Adding or deleting nodes automatically configures them for GPFS usage! Newly added nodes are considered GPFS nodes in a down state and are not recognized until a restart of GPFS. By maintaining quorum, you ensure that you can schedule a good time to refresh GPFS on the nodes.

For example, consider a GPFS configuration of four nodes. The quorum is three. With all four nodes running, we can add or delete one node, and the quorum requirement is still satisfied. We can add up to three nodes into the GPFS group as long as all four current nodes stay up. If we try to add four nodes, the GPFS group consists then of eight nodes with a quorum requirement of five. However, at that point, GPFS can only see four nodes up (configured) and exits on all the current nodes.

Deleting a FS

The command to do so is `mmdelfs`. Before deleting a GPFS FS, it must be unmounted from all GPFS nodes.

For example, if we want to delete `fs1`, which we have created in Figure 135 on page 333, we can run:

`umount fs1` on all GPFS nodes, then:

```
mmdelfs fs1
```

Checking and Repairing a FS

If a FS cannot be mounted, or if messages are received saying that a file cannot be read, it is possible to have GPFS check and repair any repairable damages to the FS. The FS has to be in the unmounted state for GPFS to check it.

The command is `mmfsck`. This command checks for and repairs the following file inconsistencies:

- Blocks marked allocated that do not belong to any file. The blocks are marked free.
- Files for which an i-node is allocated, but no directory entry exists. `mmfsck` either creates a directory entry for the file in the /lost+found directory, or it destroys the file.
- Directory entries pointing to an i-node that is not allocated. `mmfsck` removes the entries.
- Ill-formed directory entries. They are removed.
- Incorrect link counts on files and directories. They are updated with the accurate counts.
- Cycles in the directory structure. Any detected cycles are broken. If the cycle is a disconnected one, the new top level directory is moved to the /lost+found directory.

FS Attributes

FS attributes can be listed with the `mmlsfs` command. If no flags are specified, all attributes are listed. For example, to list all the attributes of `fs1`, run:

```
mmlsfs fs1
```

To change FS attributes, use the `mmchfs` command. There are eight attributes that can be changed:

1. Automatic mount of FS at GPFS startup
2. Maximum number of files
3. Default Metadata Replication
4. Quota Enforcement
5. Default Data Replication
6. Stripe Method
7. Mount point
8. Migrate FS

For example, to change the FS to permit data replication, run

```
mmchfs -r 2
```

Querying and Changing File Replication Attributes

The command `mmlsattr` shows the replication factors for one or more files. If it is necessary to change this, use the `mmchattr` command.

For example, to list the replication factors for a file /gpfs/fs1/test.file, run:

```
mmlsattr /gpfs/fs1/test.file
```

If the value turns out to be `1` for data replication, and you want to change this to `2`, run:

```
mmchattr -r 2 /gpfs/fs1/test.file
```

Re striping a GPFS FS

If disks have been added to a GPFS, you may want to re-stripe the FS data across all the disks for system performance. This is particularly useful if the FS is seldom updated, for the data has not had a chance to propagate out to the new disk(s). To do this, run `mmrestripefs`.

There are three options with this command, and any one of the three must be chosen. The `-b` flag stands for rebalancing. This is used when you simply want to re-stripe the files across the disks in the FS. The `-m` flag stands for migration. This option moves all critical data from any suspended disk in the FS. Critical data is all data that would be lost if the currently suspended disk(s) are removed. The `-r` flag stands for replication. This migrates all data from a suspended disk and restores all replicated files in the FS according to their replication factor.

For example, a disk has been added to `fs1`, and you are ready to re-stripe the data onto this new disk, run:

```
mmrestripefs fs1 -b
```

Query FS Space

The AIX command `df` shows the amount of free space left in a FS. This can also be run on a GPFS FS. However, if information regarding how balanced the GPFS FS is, the command to use is `mmdf`. This command is run against a specific GPFS FS and shows the VSDs that make up this FS and the amount of free space within each VSD.

For example, to check on the GPFS FS `fs1` and the amount of free space within each VSD that houses it, run:

```
mmdf fs1
```

12.4.4 Migration and Coexistence

The improvements in GPFS v1.2 have made it necessary that all nodes in a GPFS domain be at the same level of GPFS code. That is, in a GPFS domain, you cannot run both GPFS v1.1 and v1.2.

It is, however, possible to run multiple levels of GPFS codes provided that each level is in its own group within one system partition.

There are two possible scenarios to migrate to GPFS v1.2 from previous versions: Full and staged. As its name implies, a full migration means that all the GPFS nodes within a system are installed with GPFS v1.2. A staged migration means that certain nodes are selected to form a GPFS group with GPFS v1.2 installed. Once you are convinced that this test group is safe, you may migrate the rest of your system.

Migration and coexistence are further described in both *PSSP 3.1 Announcement*, SG24-5332, and *General Parallel File System for AIX: Installation and Migration Guide*, SA22-7278.

12.5 Related Documentation

Some extra documentation will help you better understand the different concepts and examples covered in this chapter. We recommend you take a look at some of these books in order to maximized your chances of success in the SP Certification exam.

SP Manuals

For the IBM Virtual Shared Disk (VSD) and the IBM Recoverable Virtual Shared Disk (RVSD), the manual *IBM Parallel System Support Programs for AIX: Managing Shared Disks*, SA22-7349 is an excellent guide on installing and configuring the virtual disk technology especially Chapters 1 to 6.

For the General Parallel Filesystem for AIX (GPFS), this manual will help you: *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278.

SP Redbooks

Redbooks are always good references. There are a couple of redbooks that you may want to take a look at. *Inside the RS/6000 SP*, SG24-5145 gives you

a broad coverage of the different components in the RS/6000 SP. For VSD/RVSD and GPFS, we recommend you read Chapter 4, especially 4.7 "Parrallel I/O". Another redbook that covers in much more detail this technology is *GPFS: A Parallel File System*, SG24-5165. This book will provide you with practical information about installing, configuring, and managing R/VSD and GPFS. We recommend you read Chapters 1 and 2.

12.6 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. 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`

2. You 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.

3. What is the definition of a GPFS node?

   A. It is the server node that provides token management.

   2. It is the node that has GPFS up and running.

   3. It is the node that provides the data disks for the file system.

   4. It is the server node that has GPFS and VSD up and running.

4. How do you start GPFS?

   A. `startsrc -s mmfs`

   B. `startsrc -s gpfs`

   C. `dsh startsrc -s mmfs`

   D. `dsh startsrc -s gpfs`

Chapter 13. Problem Management Tools

This chapter provides an overview and several examples for problem management by using the tools available within the RS/6000 SP. By problem management, we understand problem notification, log consolidation, and automatic recovery.

This chapter covers this by first giving an explanation about the technology used by all the problem management tools available on the RS/6000 SP. It then describes two ways of using these tools and setting up monitors for critical components, such as memory, file system space, and daemons. This first method is using the command line interface through the Problem Management subsystem (PMAN), and the second method is using the user graphical interface (SP Event Perspective).

13.1 Key Concepts You Should Study

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

- What is a resource monitor?
- What is and where the configuration data for the Event Management subsystem is 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.

13.2 AIX Service Aids

Basically, every node (and the control workstation) is an AIX machine. This means that all the problem determination tools available for standard RS/6000 machines are also available for SP nodes and control workstations.

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 within AIX and, therefore, 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 workstation.

13.2.1 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 possible 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 SP where errors could be caused by components on different machines. We will 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* facility or through the *errsave* facility if they are running at kernel level. In any case, the errdemon daemon adds the entries in the error log in 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 *IBM Parallel System Support Programs for AIX: Diagnostic Guide*, GA22-7350, Section "Using the AIX Error Log Notification Facility" on page 72, 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 within your 128 nodes 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.

13.2.2 Trace Facility

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" of the *AIX Problem Solving Guide and Reference*, SC23-4123.

13.2.3 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 AIX v3, 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.

If there is not enough space in /var/adm/ras for copying the dump, the system will ask you what to do with this dump (throw it away, copy it to tape, and so on). This is changed for SP nodes since they usually do not have people staring at the console because there is no console (at least not a physical console). Similar to machines running AIX v3, 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 use for anything else. Although your nodes are running AIX v4, so 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 who took the machine out of order. 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 keeps 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 the *Problem Solving Guide and Reference*, SC23-4123.

13.3 PSSP Service Aids

PSSP provides several tools for problem determination. Therefore, in this sometimes complex environment, you are not alone. The facilities that PSSP provides range from log files being present on every node and the control workstation to SP Perspectives that utilize the RS/6000 Cluster Technology.

13.3.1 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 control workstation (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. The *IBM Parallel System Support Programs for AIX: 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 of the cases, they 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. However, this information may be obsolete for the next release of PSSP. 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 SP 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 the *RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281 for details on disk space requirement.

13.4 Event Management

Event Management (EM) provides an application 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 137 illustrates EM's functional design.

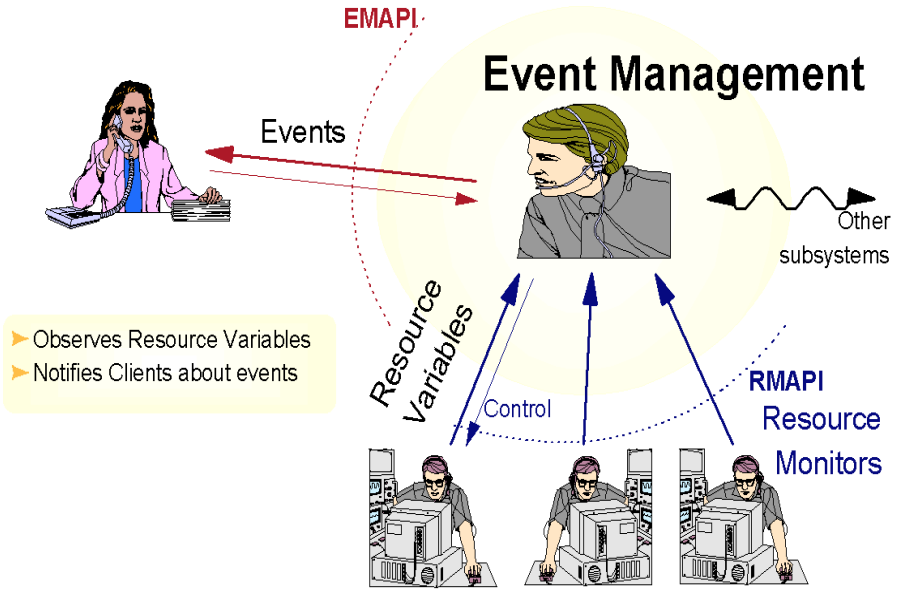


Figure 137. EM Design

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. Resource Monitors (RMs) are integral to EM, but how do RMs get their data? Data-gathering mechanisms would vary according to platform (for example, sampling CPU data in an AIX environment is implemented completely different than in a Windows NT environment). The SP-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 lives to serve applications, such as EM. As EM must communicate reliably among its daemons, it uses the Reliable Messaging information built from TS. This is shown in Figure 138.

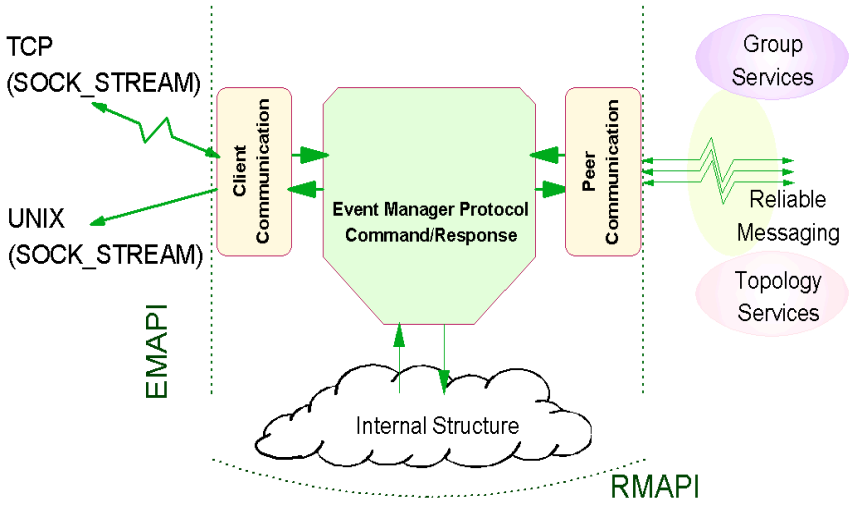


Figure 138. EM Client and Peer Communication

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 communications 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.

13.4.1 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.

13.4.2 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 through 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 139.

```
[sp3en0:/]# SDRGetObjects Syspar haem_cdb_version
haem_cdb_version
913591595,334861568,0
```

Figure 139. EMCDB Version Stored in the Syspar Class

To check the version number of the run-time version, you can use the following command:

`lssrc -ls haem.`*domain_name* from the CWS

or

`lssrc -ls haem` from a node

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 run-time 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 if the peer group has been dissolved, use the following command:

For PSSP 2.2/2.3/2.4:

`/usr/lpp/ssp/bin/hagsgr -s hags.`*domain_name* `| grep ha_em_peers`

For PSSP 3.1:

`/usr/sbin/rsct/bin/hagsgr -s hags.`*domain_name* `| grep ha_em_peers`

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.

13.5 Problem Management

The Problem Management subsystem (PMAN) is a facility, present on systems running PSSP v2.2 or later, for problem determination, problem notification, and problem solving. It uses the RSCT infrastructure for monitoring conditions on behalf of authorized users, and then it generate 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 control workstation 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 additional 16 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 354 for an example about how to use this facility.
- **sp_configd** - Through this daemon, PMAN can send Simple Network Management Protocol (SNMP) traps to SNMP managers to report pre-defined conditions.

13.5.1 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 it is located at /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 the *PSSP: Command and Technical Reference,* SA22-7351, on page 350 for a complete definition of this command. Chapter 25 "Using the Problem Management Subsystem" in the *PSSP :Administration Guide*, SA22-7348 contains several examples and a complete explanation about 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 about 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 will use one of these variables to monitor an 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; 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 by using the `pmanrmloadSDR` command.

PSSP provides a template for this configuration file. It is located in the /spdata/sys1/pman directory on the control workstation. 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 sixteen 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) will update 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 special syntax for this variable is 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) by executing the following command:

```
[sp5en0:/]# haemqvar "" IBM.PSSP.pm.User_state1 "*"|more
```

This command gives you a very good explanation along with examples on 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 control workstation) about the intruder.

The script `/usr/local/bin/SayItLoud.pl` will display 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 "------ Pres 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 second (a minute at most), we receive the warning window shown in following warning message at the control workstation:

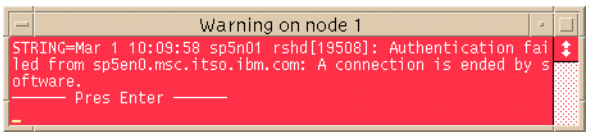


Figure 140. User-defined Resource Variables - Warning Window Example

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 an 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.

13.6 Event Perspectives

The SP Perspectives is a set of applications, each of which has a graphical interface (GUI), that enables 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.

13.6.1 Defining Conditions

The procedure for creating monitors is very straightforward. A condition needs to be defined prior 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:

Step 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.

Step 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 find out the right variable. In releases previous to PSSP 3.1 the only way to get some information on resource variables is through the help facility on SP Perspectives. However, in PSSP 3.1 there is a new command that will help you find the right variable, and it will provide you with information about how to use it. Let's use this new command called `haemqvar`.

We can use this command 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 within 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 that represents exactly what we want to monitor.

Step 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 141 on page 360.

```
[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
    has free space. The "%nodesused" resource variable can be used
    to monitor the percent of file nodes which are in use.

    Example expression:

    To receive a notification that the file system mounted on /tmp on any
    node is more than 90% full, and also receive a notfication when the
    percentage has subsequently dropped below 80%, one could register
    for the following event using the HA_EM_CMD_REG2 command:

        Resource variable:   IBM.PSSP.aixos.FS.%totused
        Resource ID:         VG=rootvg;LV=hd3;NodeNum=*
        Expression:          X > 90
        Re-arm expression:   X < 80

e....lines not displayed....
```

Figure 141. 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. Therefore, our expression would be: `X>90`

We could use a rearm expression in our condition. A rearm expression is optional, and 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 will send us a notification, and then it will continue monitoring the file system; but now it will send us a notification when the space used falls below 60 percent.

Step 4 Create the condition. To create the condition, let us move the focus to the conditions pane on Event Perspective and then select **Actions->Create...**, as shown in Figure 142.

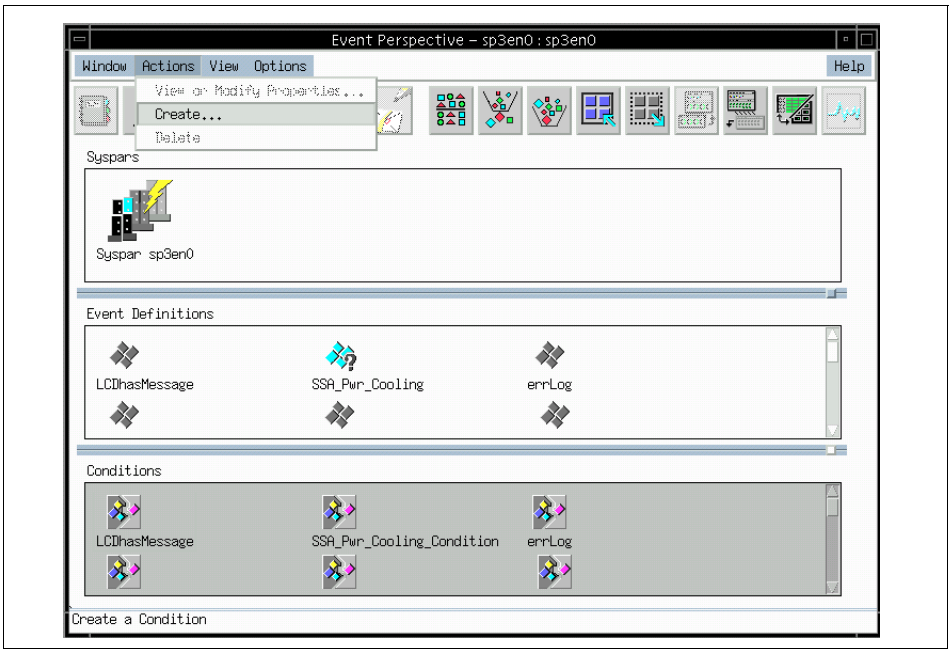


Figure 142. Create Condition Option from Event Perspectives

Once you click on the **Actions->Create...** option, you will be presented with the Create Condition pane as shown in Figure 143 on page 362.

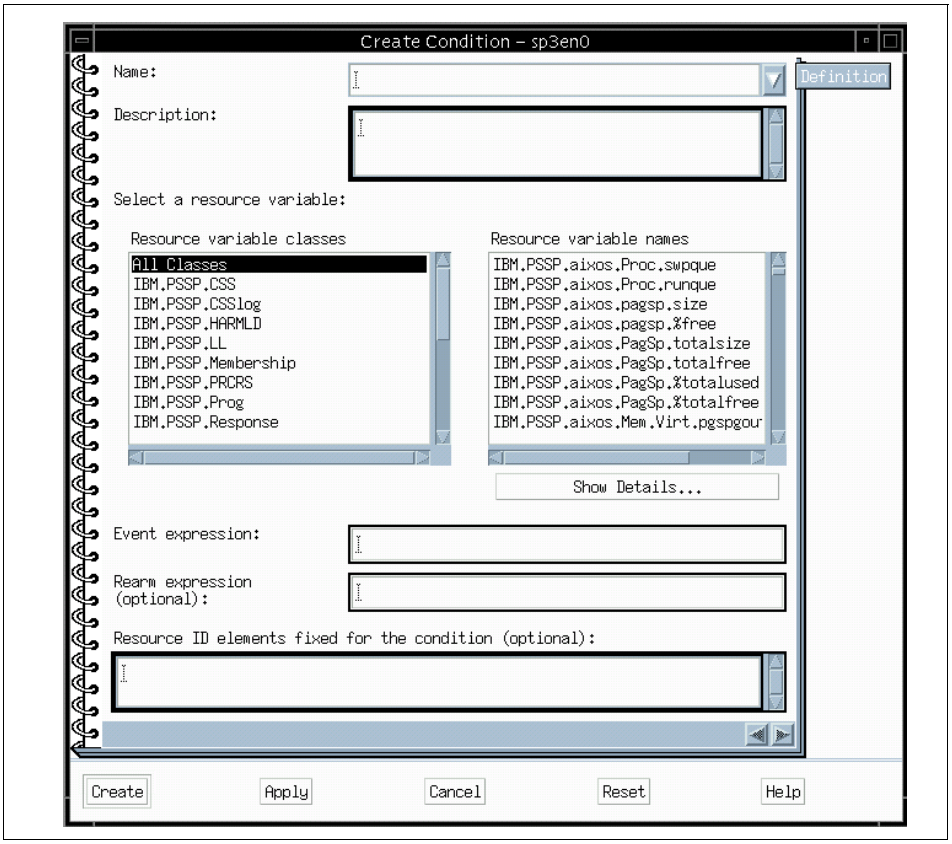


Figure 143. Create Condition Pane

As you can see in the Create Condition pane, 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 144 on page 363.

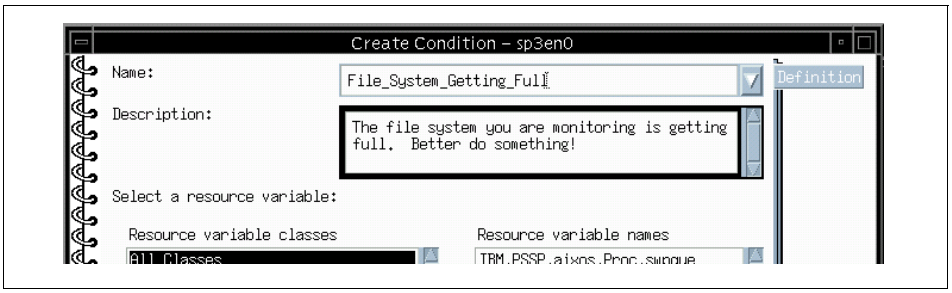


Figure 144. Defining Name and Description of a Condition

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 expression we defined in the previous step. This is shown in Figure 145.

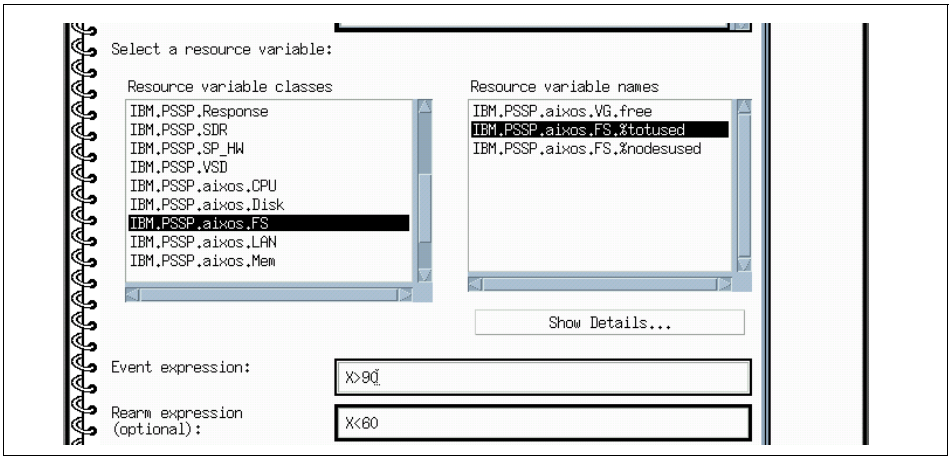


Figure 145. Selecting Resource Variable and Defining Expression

If you click on **Show Details...**, it will present you the same output we got through 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) and the other is the logical volume name (LV). By using the last input box, we could have fixed one or the two resource IDs to a specific file system; so, this condition could 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 pane as shown in Figure 146.

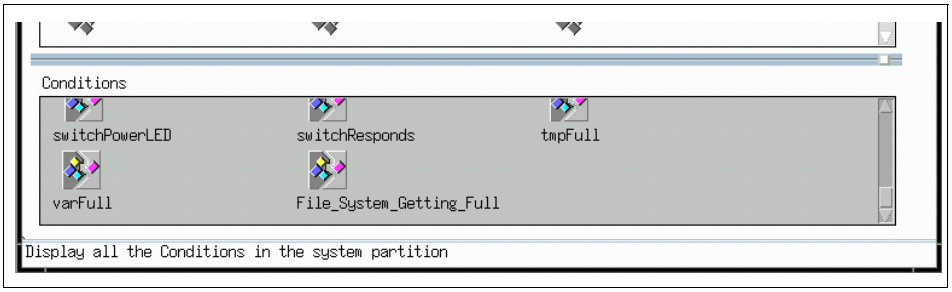


Figure 146. Conditions Pane - New Condition

13.7 Related Documentation

This documentation will help you getting more detailed information in the different topics covered in the chapter. Also, remember that good hands-on experience may reduce the amount of preparation for the SP Certification exam.

SP Manuals

The only SP manual that can help you with this is the *PSSP: Administration Guide, SA22-7348 for PSSP 3.1* and the *PSSP: Administration Guide*, GC23-3897 for PSSP 2.4. In both books, there is a section dedicated to availability and problem management as well as SP Perspectives. We recommend you to read at least Chapters 24 and 25 of the PSSP 3.1 guide and Chapters 23 and 24 of the PSSP 2.4 guide.

SP Redbooks

There are several books that cover the topics in this Chapter. However, 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 about the concepts involved. The other redbook is *Inside the RS/6000 SP*, SG24-5145. This redbook contains an excellent description of the Event Management and Problem Management subsystems. Finally, the redbook *RS/6000 SP PSSP 2.2 Technical Presentation*, SG24-4868, contains detailed information on these topics.

For a PSSP 3.1 update, we recommend Chapter 6 of *PSSP 3.1 Announcement*, SG24-5332.

13.8 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 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>`

Part 4. On-going Support

Chapter 14. RS/6000 SP Software Maintenance

This chapter discusses how to maintain backup images for the CWS and SP 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.

14.1 Key Concepts You Should Study

This section gives you key concepts for the preparation for the certification exam for maintaining the software of the RS/6000 SP. You should understand:

- How to create and manage backup images for CWS and SP nodes.
- How to restore CWS or SP nodes and what are the necessary procedures after restoring.
- How to apply the PTFs and what are the required tasks for AIX and PSSP on CWS and nodes.
- What are the influences of the PTFs you applied on your SP system.
- The concept of software migration and coexistence in supported environments.
- What are the changes made between PSSP v2 and PSSP v3.

14.2 Backup of the Control Workstation and SP Node Images

Maintaining a good copy of backup images is as important as initial implementation of your SP system. Here we discuss how to maintain the CWS backup image and how to efficiently create SP node images with a scenario we set up in our environment.

14.2.1 Backup of the Control Workstation

The backup of the CWS is the same as the strategy you use for standalone RS/6000 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`.

14.2.2 Backup of SP Node Images

In scientific or parallel computing environments, we may only need one copy of node images across the SP 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 SP node. Therefore, you need to understand your environment and set up the SP node backup strategy.

In general, it is recommended to 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 ADSM, for backup.

Also, remember that the node image you create is a file and is not bootable so that you should follow the network boot process, as discussed in Chapter 9, "Frames and Nodes Installation" on page 249, to restore it.

There are many ways you can set up an SP node backup depending upon your environment. Here, we introduce the way we set it up in our environment.

14.2.3 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 need 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 run `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 147 on page 371.

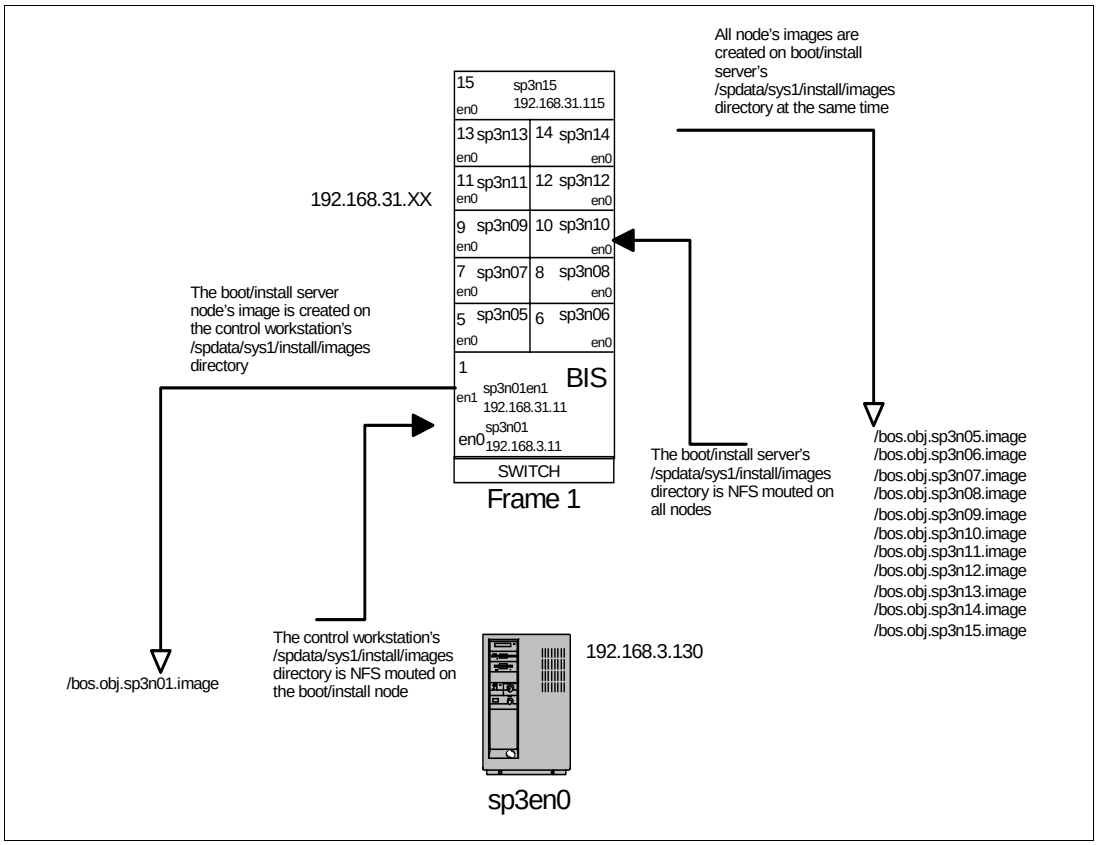


Figure 147. Mechanism of SP Node Backup in Boot/Install Server Environment

Of course, you can write scripts to automate this process. Due to the nature of this book, we only introduce the mechanism of the node backup strategy.

14.3 Restoring from mksysb Image

In the following sections, we discuss the recovery of the CWS and nodes when a system has crashed.

14.3.1 Restoring the Control Workstation

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 RS/6000 workstation 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 RS/6000 workstation.
2. Issue the `/usr/lpp/ssp/bin/install_cw` command.

   When 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 8.3.2, “install_cw” on page 237.
3. Verify your CWS.

14.3.2 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 in PSSP 2.4 or earlier, you do the following:

On the command line, you execute the following command:

```
# spbootins -r install -i <mksysb image name> -l <node list>
```

PSSP 3.1 has some modifications to the `spbootins` command; you do not have the same flags you had in PSSP 2.4 or earlier. If you try to change the environment using SMIT in PSSP 3.1 with the procedure just described, you will get a response similar to the following:

```
spbootins: 0016-601 An option was used that is no longer supported by
this command.
Use the "spchvgobj" command.
spbootins: Syntax:
spbootins [ -c selected_vg ]
[ -r {install | customize | disk | maintenance | diag | migrate }][ -s
yes | no ]{start_frame start_slot node_count | -l <node_list>}
spbootins: Syntax:
spchvgobj < -r < volume_group >
[ -h pv_list ]
[ -i install_image ]
[ -p code_version ]
[ -v lppsource_name ]
[ -n boot_server ]
[ -c 1 | 2 | 3 ]
[ -q true | false ]
{start_frame start_slot node_count | -l <node_list>}
```

The reason for the error is that the option `-i`, used to change the name of the image of installation, is no longer supported in PSSP 3.1. The new command `spchvgobj` should be used to change this field. This change is needed to support the new possibility of having multiple rootvg volume groups.

To change the environment in PSSP 3.1, you run the following command:

```
# 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 specific configuration of it. After the node is installed, `pssp_script` customizes it with the correct information.

14.4 Applying 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.

14.4.1 On the Control Workstation

This section briefly describes how to apply AIX and PSSP PTFs on the CWS.

14.4.1.1 Applying AIX PTFs

The steps for applying AIX PTFs are as follows:

1. Create 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/aix432/lppsource.
4. Create a new .toc file by executing the commands:

```
# cd /spdata/sys1/install/aix432/lppsource
# 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_aix432`

   Network Install Operation to perform: `update_all`

If the status of the install is OK, then you are done with the update of the AIX PTFs on the CWS. If the status of the install is that it has failed, then review the output for the cause of the failure and resolve the problem.

14.4.1.2 Applying PSSP PTFs

The steps for applying PSSP PTFs are as follows:

1. Create a 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-3.1 for PSSP 3.1.
3. Create a new .toc file by issuing the following commands:

   ```
   # cd /spdata/sys1/install/pssplpp/PSSP-3.1
   # 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 see 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.

14.4.2 To the Node

There are many ways you can install PTFs on the nodes. If you have a server consolidation environment and have 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.

Note, that 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.

14.4.2.1 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 for installing the PTFs.

1. Log in as root and mount the lppsource directory of the CWS in `sp3n01` by issuing the command:

   ```
   # mount sp3en0:/spdata/sys1/install/aix432/lppsource /mnt
   ```

2. Apply the PTFs using the command:

   ```
   # smitty update_all
   ```

   INPUT device for directory / software: `/mnt`

   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/aix432/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. Before rebooting a node, if you have a switch, you may need to fence it using the command:

```
# Efence -autojoin sp3n01
```

14.4.2.2 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.

For installing PSSP PTFs, 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-3.1 /mnt
```

When updating the ssp.css fileset of PSSP, you must reboot the nodes for the Kernel extensions to take effect.

It is recommended to make another backup image after you have applied the PTFs.

14.5 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 3.1 when you migrate your system to 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.

14.5.1 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 4.2 to 4.3, it is a migration, while if you change the AIX level from 4.3.1 to 4.3.2, it is an update. However, all PSSP level changes are updates.

14.5.2 Supported Migration Paths

In PSSP 3.1, the only supported paths are those shown in Table 27. If your current system, CWS, or any node is running at a PSSP or AIX level not listed in the `From` column of Table 27, you must update to one of the listed combinations before you can migrate to PSSP 3.1. Refer the manual *IBM Parallel System Support Install & Migration Guide Version 3 Release 1*, GA22-7347 for detail migration procedure.

Table 27. Supported Migration Paths to 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 to 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 the document *IBM RS/6000 SP Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.

14.5.3 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: To preserve 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).
- RS/6000 SP 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 3.1 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.

14.5.4 Overview of CWS PSSP Update

This section briefly describes what is new in PSSP 3.1 for updating the CWS. For further information refer, to *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*, GA22-7347.

We describe the main steps in the installation process but with the migration goal in mind. We 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 run-time 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 and 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.

   Issue the `lssrc -a` command to verify that the daemons are no longer running on the CWS.

   ```
   # 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
   ```

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.

14.5.5 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 the document *IBM Parallel System Support Install & Migration Guide Version 3 Release 1*, GA22-7347, on page 128.

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`, `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 Install of Nodes

   This is the node installation that we discussed in Chapter 9, “Frames and Nodes Installation” on page 249.

3. Update to a new level of PSSP and update 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, you must set the `bootp_response` parameter to `customize` with the new PSSP 3.1 commands (`spchvgobj`, `spbootins`) for the 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 PSSP 3.1 node to the new PSSP 3.1 level.

14.5.6 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 28 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 28. 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.

14.6 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 Chapter 6 of *PSSP:Installation and Migration Guide (Version 2 Release 4)*, GC23-3898, and *Installation and Migration Guide (Version 3 Release 1)*, GA22-7347. For details on how to boot from the mksysb tape, read the *AIX Version 4.3 Installation Guide*, SC23-4111.

SP Redbooks

RS/6000 SP Software Maintenance, SG24-5160. This redbook provides everything you need for software maintenance. It is strongly recommended to read this for real production work. For the sections of backup and PTFs, you

may refer to Chapter 7 and Chapter 8. For the section on software migration, you may read Chapter 2 of *PSSP 3.1 Announcement*, SG24-5332.

14.7 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. 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's 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.

Chapter 15. RS/6000 SP Reconfiguration and Update

Most commercial environments start with a small number of nodes and expand their environment as time goes by or new technology becomes available. In Chapters 7 and 8, we discussed the key commands and files used for initial implementation based on our environment. In this chapter, we go through the procedures used to reconfigure an SP system, such as a adding frame, 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.

15.1 Key Concepts You Should Study

This section gives you the key concepts you have to understand when you prepare the certification exam about reconfiguration and migration of RS/6000 SP. You should understand:

- 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 multi-frame 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.

15.2 Environment

This section describes the environment for our RS/6000 SP system. From the initial RS/6000 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 148 on page 386.

In the Figure 148, 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 of the rest of 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 the *Installation and Migration Guide(Version 3 R 1)*, GA22-7347,. even though the physical installation was done at the same time.

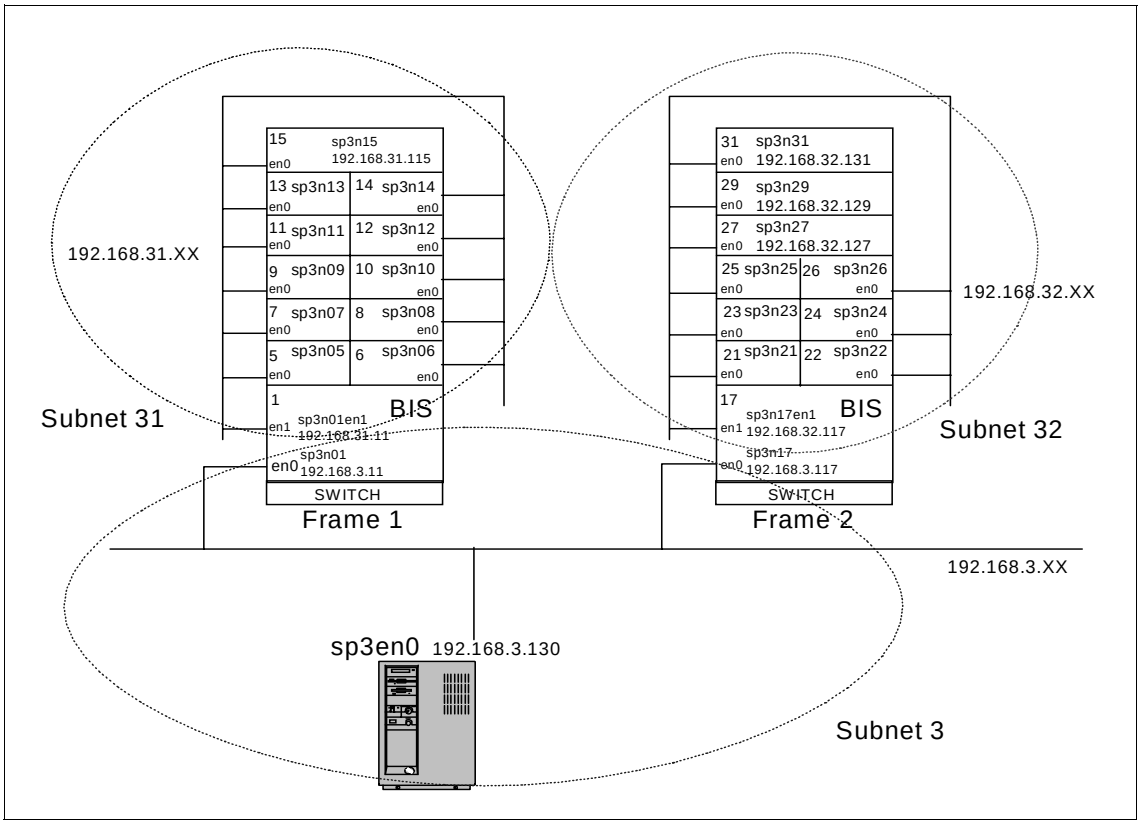


Figure 148. Environment after Adding a Second Switched Frame and Nodes

15.3 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 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. Everytime 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. Unpartition 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 15.2, “Environment” on page 385, 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 to Frame 2) and indicate that Frame 1 is connected to /dev/tty0 and Frame 2 to /dev/tty1 and reinitializes 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 or S70 advanced server attached to your SP.

The S70 and S70 Advanced Server 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 and S70 Advanced Server for hardware controls. The s1 tty port value defines the connection to the serial port on the S70 and S70 Advanced Server for serial terminal (s1term) support. A switch port value is required for each S70 or S70 Advanced Server 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 2 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
```

And 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 like this:

```
[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 will be 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.

15.4 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 148 on page 386. Assume that all nodes were installed when the frame was installed. Thus, the following steps are the continuation of 14.1, “Key Concepts You Should Study” on page 369. After we enter all nodes information into 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 address 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 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` adapters 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 *Installation and Migration Guide(Version 3 R 1)*, 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:

```
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 running 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 like:

```
                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 such:

   ```
   [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.
The output looks as such:

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 have checked sp3n17 node points to the CWS as boot/install server, and all the rest of nodes point to sp3n17 as boot/install server, which is the default in a multi-frame environment. However, if you need to select the different node to be boot/install server, you can use `-n` option of the `spchvgobj` command.

To change these 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 like:

```
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 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 command `splstdata` 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` 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 /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 and S70 Advanced Server 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. The settings in the /tftpboot/tuning.cust file are maintained across a boot of the 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 the 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 function that 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 in a more *normal* environment where most all services have been fully configured.

15. Additional switch configuration (optional)

    If you have added a frame with a switch, perform:

    1. 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.

    2. 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 overridden 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.

3. 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
   ```

4. Set the switch clock source for all switches with the command: `Eclock`

   For our environment, select **/etc/SP/Eclock.top.2nsb.0isb.0** as an topology file and enter:

   ```
   [sp3en0:/]# Eclock -f /etc/SP/Eclock.top.2nsb.0isb.0
   ```

   To verify the switch configuration information, enter:

```
[sp3en0:/]# splstdata -s
                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.

    1. To monitor installation progress by opening the node's read-only console, issue:

       ```
       [sp3en0:/]# s1term 2 1
       ```

    2. 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 shows 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&
    ```

    Then, finished the rest of nodes. 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.

15.5 Adding Existing S70 to SP System

If you want to preserve the environment of your existing S70 or S7A server, perform the following steps to add as 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 control workstation 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 on your SP-attached server.
6. Mount the `pssplpp` directory: 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 pssp_script: Run the pssp_script by issuing:

   ```
   # /usr/lpp/ssp/install/bin/pssp_script
   ```

10. Reboot: Perform a reboot:

    ```
    # shutdown -Fr
    ```

15.6 Adding a Switch

This section was already summarized as part of the previous section. However, here we introduce the following two cases when you just add the SP switch only:

- Adding a switch to a switchless system
- Adding a switch to a system with existing switches

15.6.1 Adding a Switch to a Switchless System

1. Redefine the system to a single partition.

   Refer to the *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 command: `installp`

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.

   To do this, 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 the step 15 in 15.4, “Adding a Node” on page 390.

10. Refresh system partition-sensitive subsystems with the 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.

15.6.2 Adding 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 390.

6. Refresh system partition-sensitive subsystems with the 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.

15.7 Replacing to PCI-Based 332 MHz SMP Node

This scenario of migration is summarized only for preparing for the exam and will not provide full information for conducting actual migration. However, this section will provide enough information to understand the migration process.

15.7.1 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.

15.7.2 Software Requisites

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. By and large, 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).

15.7.2.1 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.

15.7.2.2 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.

15.7.2.3 Migration and Coexistence Considerations

Table 29 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 29. 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

15.7.3 Control Workstation Requirements

The CWS has a 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).

15.7.3.1 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 the *PSSP: Installation and Migration Guide*, GC23-3898, for documentation of these levels. AIX must be at a supported level before PSSP can be installed.

15.7.3.2 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.

15.7.3.3 NIM Configuration

The NIM configuration on the CWS will also need to be updated to current levels. Please 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.

15.7.4 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.

15.7.4.1 Phase I: Preparation on the CWS and Existing Nodes

1. Plan any necessary client and server verification testing.
2. Plan any external device verification (tape libraries, and so on).
3. Capture all required node documentation.
4. Capture all non-rootvg VG information.
5. 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
```

6. 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
```

7. Create system backup (rootvg) for the control workstation.

```
# mksysb -i /dev/rmt0
```

8. Copy required AIX filesets including PCI device filesets to the /spdata/sys1/install/aix432/lppsource directory.
9. Copy required AIX fixes including PCI device fixes to the /spdata/sys1/install/aix432/lppsource directory.
10. Copy PSSP to the /spdata/sys1/install/pssplpp/PSSP-3.1 directory.
11. Copy latest PSSP fixes to /spdata/sys1/install/pssplpp/PSSP-3.1 directory.
12. Copy coexistence fixes to /spdata/sys1/install/pssplpp/PSSP-3.1 directory if needed.

13. Create /spdata volume group backup.

```
# savevg -i /dev/rmt0 spdatavg
```

15.7.4.2 Phase II: Perform on the Existing Nodes

1. Perform the preparation steps.
2. 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.
3. 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.
4. Verify operation of the upgraded CWS with the nodes. Perform a SDRArchive.
5. Upgrade PSSP and AIX (if needed) on the nodes that will be replaced by 332 MHz SMP nodes. Install the latest PSSP fixes.
6. Verify operation of the nodes and back up the nodes after successful verification. Archive the SDR through the `SDRArchive` command.
7. Shutdown the original SP nodes which are being replaced.
8. 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.
9. Bring in and physically install the new nodes. You will move all external node connections from the original nodes to the new nodes.

15.7.4.3 Phase III: Rebuild SDR and Install New 332 MHz SMP Nodes

1. Rebuild SDR with all required node information on the CWS.
2. 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).
3. Netboot all nodes being sure to select the correct AIX & PSSP levels.
4. Verify AIX and PSSP base code levels on nodes.
5. Verify AIX and PSSP fix levels on nodes and upgrade if necessary.

6. Verify node operation (`/usr/lpp/ssp/install/bin/node_number, netstat -in`).
7. You will need the node documentation acquired during the preparation step.
8. Perform any necessary external device configuration (tape drives, external volume groups, and so on).
9. Perform any necessary client and server verification testing.
10. Perform any external device verification (tape libraries, and so on).
11. Create a full system backup for nodes.
12. Create a system backup (rootvg) for the CWS.
13. Create a /spdata volume group backup.

15.8 Related Documentation

We assume that you already have experience with the key commands and files from Chapter 7 and Chapter 8. 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*, GA22-7281 give you a good description of what you need. For details about reconfiguration of you SP system, you can refer to Chapter 5 of the following two manuals: *PSSP: Installation and Migration Guide (Version 2 R 4)*, GC23-3898, and *Installation and Migration Guide(Version 3 R 1)*, GA22-7347.

Other Sources

Migrating to the RS/6000 SP 332 MHz SMP Node, IBM Intranet:
`http://dscrs6k.aix.dfw.ibm.com/`

15.9 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. 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.

Chapter 16. 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.

16.1 Key Concepts You Should Study

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 from common problems.
- The mechanism of SP user management with automount and file collection and the technique 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.

16.2 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.

16.2.1 Diagnosing setup_server Problems

The problems with `setup_server` are complicated and a 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 the 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 CWS.

16.2.1.1 Problems with the SDR

The most common problem with 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.

16.2.1.2 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 can not 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.

16.2.1.3 Problems with Conflicting NIM Cstate and SDR

Before we discuss this problem, it is helpful to understand NIM client definition. Table 30 shows information on this.

Table 30. 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.	

boot_response	Cstate	Allocations
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 command `lsnim -l <client_name>` 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.

16.2.1.4 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 recreate the SPOT.

16.2.1.5 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.

16.2.1.6 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.

Login 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.

16.2.1.7 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 with the command: `df -k`
2. Check the valid lppsource resource is available with 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 the `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`.

16.2.2 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.

16.2.2.1 Common Checklists

When you have a problem with network booting, you should check the following lists:

- 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, “nodecond” on page 264.
- Check if there is any Kerbero error.
- Check if there is any SDR error.

16.2.2.2 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.
  1. *LED 231* sends a bootp broadcast packet through the network.
  2. *LED 260* reaches the limit for not receiving a reply packet.
  3. Attempts to retrieve the boot image file.
  4. *LED 299* received a valid boot image.
  5. Continued to read the boot record from the boot image and create the RAM file system.
  6. Invokes: `/etc/init(/usr/lib/boot/ssh)`
  7. Invokes: `/sbin/rc.boot`
- After `rc.boot` is executed:
  1. Cannot execute `bootinfo_<platform>`, hang at LED C10.
  2. Remove unnecessary files from RAM file system.
  3. Read IPL Control Block.
  4. Can not determine the type of boot, hang at LED C06.
  5. *LED 600* executes `cfgmgr -fv`. Set IP resolution by /etc/hosts.
  6. *LED 606* configures lo0,en0. If error, hang at LED 607.
  7. *LED 608* retrieves `niminfo (/tftpboot/<reliable_hostname>)` file through tftp. If error, hang at LED 609.
  8. Create /etc/hosts and Configure IP route. If error, hang at LED 613.
  9. *LED 610* performs NFS mount of the SPOT file system. If error, hang at LED 611.
  10. *LED 612* executes the `rc.bos_inst` script.
  11. Change Mstate attribute of the NIM client object to: `in the processing of booting`
  12. *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.
  13. *LED 622* links the configuration methods and executes `cfgmgr -vf` for the first and second phase.
  14. Exit `/etc/rc.boot` for the first phase and start the second phase.
  15. Set /etc/hosts for IP resolution and reload niminfo file.
  16. Execute `rc.bos_inst` again.
  17. Delete the `rc.boot` file.
  18. Define the IP parameters.

19. Copy ODM objects for pre-test diagnostics.
20. Clear the information attribute of the NIM client object.
21. Invoke the `bi_main` script.

- After the `bi_main` script is invoked:

1. Invoke the initialization function and change the NIM Cstate attribute to `Base Operation System Installation is being performed`.
2. *LED C40* retrieves bosinst.data, image.data and preserve.list files and creates a file with the description of all the disks.
3. *LED C42* changes the NIM information attribute to `extract_diskette_data` and verify the existence of image.data.
4. Change the NIM information attribute to `setting_console` and set the console from the bosinst.data file. If error, hang at LED C45.
5. Change the NIM information attribute to `initialization`.
6. *LED C44* checks for available disks on the system.
7. *LED C46* validates target disk information.
8. *LED C48* executes the BOSMenus process.
9. *LED C46* initializes the log for bi_main script and sets the minimum values for LVs and file systems.
10. Prepare for restoring the operating system.
11. *LED C54* restores the base operating system.
12. *LED C52* changes the environment from RAM to the image just installed.
13. *LED C46* performs miscellaneous post-install procedures.
14. LED C56 executes BOS installs customization.
15. LED C46 finishes and reboots the system

- After `pssp_script` script is invoked:

1. *u20* creates log directory (enter function `create_directories`).
2. *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.
3. *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.
4. *u23* Create/update /etc/ssp files (enter function `create_files`).
   - *u60* Create/update /etc/ssp files.
5. *u24* updates /etc/hosts file (enter function `update_etchosts`).
6. *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.

7. *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.
8. *u27* updates the /etc/inittab file (enters the function `update_etcinittab`).
9. *u28* performs MP-specific functions (enters the function `upmp_work`).
   - *u52* Processor is MP.
   - *u51* Processor is UP.
   - *u55* Fatal error in bosboot.
10. *u29* installs prerequisite filesets (enters the function `install_prereqs`).
11. *u30* installs ssp.clients (enters the function `install_ssp_clients`).
    - *u80* mounts lppsource and installs `ssp.clients`.
12. *u31* installs `ssp.basic` (enters the function `install_ssp_basic`).
    - *u81* installs `ssp.basic`.
13. *u32* installs `ssp.ha` (enters the function `install_ssp_ha`).
    - *u53* installs `ssp.ha`.
14. *u33* installs `ssp.sysctl` (enters the function `install_ssp_sysctl`).
    - *u82* installs `ssp.sysctl`.
15. *u34* installs `ssp.pman` (enters the function `install_ssp_pman`).
    - *u41* configures switch (enters the function `config_switch`).
16. *u35* installs `ssp.css` (enters the function `install_ssp_css`).
    - *u84* installs `ssp.css`.
17. *u36* installs `ssp.jm` (enters the function `install_ssp_jm`).
    - *u85* installs `ssp.jm`.
18. *u37* deletes the `master .rhosts` entry (enters the function `delete_master_rhosts`).
19. *u38* creates a new dump logical volume (enters the function `create_dump_lv`).
    - *u86* creates a new dump logical volume.
20. *u39* runs the customer's `tuning.cust` (enters the function `run_tuning_cust`).
21. *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`).

16.2.2.3 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 the *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 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 command `splstdata -b`.
5. Rerun the `spbootins` command with `setup_server`.

16.2.2.4 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 the *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, with the following command, if the NIM client machine has the exported directories listed:

```
# 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 can not find the cause of the problem, clean up the NIM setup and exported directory and do as follows:
   1. Remove entries from /etc/exports with:

      ```
      /export/nim/scripts/*
      /spdata/*
      ```

   2. Remove NFS-related files in /etc:

      ```
      # rm /etc/state
      # rm /etc/sm/* /etc/sm.bak/*
      ```

   3. Unconfigure and reconfigure NIM:

      ```
      # nim -o unconfig master
      # installp -u bos.sysmgt.nim.master
      ```

   4. Set the node or nodes back to `install` and run `setup_server`. This will also reinstall NIM:

      ```
      # spbootins -r install -l <node#>
      ```

   5. Refresh the newly created exports list:

      ```
      # exportfs -ua
      # exportfs -a
      ```

   6. Refresh NFS:

      ```
      # stopsrc -g nfs
      # stopsrc -g portmap
      # startsrc -g portmap
      # startsrc -g nfs
      ```

16.2.2.5 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 recreate the boot image files.

16.2.2.6 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:

   1. Check if the bos fileset exists in lppsource:

      ```
      # nim -o lslpp -a filesets=bos lppsource_aix432
      ```

   2. Check if the image.data file exists:

```
# dd if=/spdata/sys1/install/aix432/lppsource/bos bs=1k count=128
| restore -Tvqf ./image.data
```

3. Then, check the file permission on image.data.

16.2.2.7 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:
  1. Check with the clients' boot/install server and its hostname by issuing:

     ```
     # splstdata -b
     ```

  2. `telnet` to the boot/install server if not the CWS.
  3. 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 *IBM PSSP for AIX: Diagnosis Guide (Version 3 R 1)*, GA22-7350.
  4. If the node is listed in this file, continue to the next step.
- Open a write console to check for console messages.
  1. 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
     ```

  2. 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.

16.3 Diagnosing SDR Problems

This section shows a few common problems related to SDR and its recovery actions.

16.3.1 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:

```
# 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.

16.3.2 Problem with Class Corrupted or Nonexistent

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 *IBM PSSP for AIX: Administration Guide (Version 3 R 1)*, SA22-7348). Then, follow the steps in “SDR Shadow Files” in the System Data Repository appendix in the *IBM PSSP for AIX: Administration Guide (Version 3 R 1)*, 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.

16.4 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.

16.4.1 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.

16.4.2 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.

16.4.2.1 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.

16.4.2.2 Problems with Accessing User's 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 autmounter:

```
# /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 *PSSP: Messages Reference* 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.

16.5 Diagnosing File Collection Problems

In this section, we summerize common checklists for file collection problems and explain how you can resolve them.

16.5.1 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:

```
# /var/sysman/supper status

CollectionResident  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.

16.6 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.

16.6.1 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 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.

16.6.2 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.

16.6.3 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 or the krb-srvtab has the wrong key version or krb-srvtab file is corrupted. Analyze the error messages to confirm services'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.

16.6.4 Problems with Authenticated Services

When hardmon is having problems due to 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, host name resolution is not correct, or 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 hosname resolution, ACL files, and the Kerberos database.

16.6.5 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`, `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 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>
   ```

16.6.6 Problems with Decoding Authenticator

When you change the hostname 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 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>
```

16.6.7 Problems with the Kerberos Daemon

Here is an example of messages when 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.

16.7 Diagnosing System Connectivity Problems

This section shows a few examples related to network problems.

16.7.1 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.

16.7.2 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.

16.7.3 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
   ```

16.8 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

16.8.1 604 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.

16.8.2 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.

16.8.3 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.

16.8.4 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 the 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.

16.9 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 following symptoms, perform the appropriate recovery action described.

16.9.1 Problems with Estart Failure

The `Estart` problems are caused by many different reasons. In this section, we discuss the following typical symptoms.

16.9.1.1 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,*SA22-7351.

16.9.1.2 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 with: /usr/lpp/ssp/css/rc.switch

16.9.1.3 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 *IBM PSSP for AIX: Diagnosis Guide (Version 3 R 1)*, 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 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.

16.9.1.4 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 detail list of possible device status for SP switch, refer to P119-120 of the *IBM PSSP for AIX: Diagnosis Guide (Version 3 R 1)*, GA22-7350

16.9.2 Problem with Pinging to SP Switch Adapter

If the SP node fails to communicate over the switch, but its switch_responds is on `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 in P121-132 from *IBM PSSP for AIX: Diagnosis Guide (Version 3 R 1)*, GA22-7350.

16.9.3 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 else 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.

16.9.4 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

16.10 Impact of Hostname/IP Changes on 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 hostname 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 IBM RS/6000 SP: 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)

16.10.1 SDR Objects with Hostnames 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.

16.10.2 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.

16.11 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 *IBM PSSP for AIX: Diagnosis Guide (Version 3 R 1)*, GA22-7350, for full description. In addition, you may read Chapters 4, 5, 8, 12, and 14 of *IBM for AIX PSSP: Administration Guide*, SA22-7348, to get the basic concepts of each topic we discuss here.

SP Redbooks

There is no problem determination redbook available for PSSP 2.4. You can use *RS/6000 SP: PSSP 2.2 Survival Guide*, SG24-4928, for PSSP 2.2 This redbook discusses extreme details on node installation and SP switch problems.

16.12 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. 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 login 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 by:

      ```
      # /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:

   - True

- 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 filesystem 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`.

Appendix 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 70, 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 2.5, “Dependent Nodes” on page 25.

Question 2 - The answer is B. The two switch technologies (SP Switch and HiPS) are not compatible. PSSP 2.4 is the last PSSP level that support the HiPS switch. PSSP 3.1 or later does not support the older switch.

Question 3 - The answer is A. PSSP 3.1 requires AIX 4.3.2 or later. The Performance Toolbox manager extension (perfagent.server fileset) is no longer a prerequisite in PSSP 3.1. Refer to 2.12, “Software Requirements” on page 51 for details.

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 2.4.1, “Internal Nodes” on page 14.

A.2 RS/6000 SP Networking

Answers to questions in 3.6, “Sample Questions” on page 100, 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.

A.3 I/O Devices and File Systems

Answers to questions in 4.6, “Sample Questions” on page 134, 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 119 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 4.3.2.8, “spbootlist” on page 117 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.

A.4 SP-Attached Server Support

Answers to questions in 5.8, “Sample Questions” on page 170, 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 5.2.2, “SP-Attached Server Attachment” on page 137 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 2.14, “Configuration Rules” on page 54 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 5.4.2, “Hardmon” on page 156 for details.

A.5 SP Security

Answers to questions in 6.16, “Sample Questions” on page 203, 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. One of the reasons why PSSP 3.1 still requires Kerberos v4, although it supports, through AIX, Kerberos v5, is the fact that the hardmon daemon and the sysctl facility still require Kerberos v4 for authentication. Refer to 6.12, “SP Services That Utilize Kerberos” on page 190 for details.

Question 3 - 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 6.10, “Server Key” on page 188 for details.

Question 4 - 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.

A.6 User and Data Management

Answers to questions in 7.8, “Sample Questions” on page 230, 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 206 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 210 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 7.5.3.2, “user.admin Collection” on page 216 for details.

A.7 Configuring the Control Workstation

Answers to questions in 8.8, “Sample Questions” on page 248, are as follows:

Question 1 - The answer is D. The partition-sensitive daemons are controlled by the `syspar_ctrl` command. The `install_cw` script will not create or start those daemons. Refer to 8.3.2, “install_cw” on page 237 for details.

Question 2 - The answer is B. In the release prior to PSSP 3.1, the System Performance Measurement Interface (SPMI) library was required by some PSSP components. This library was packaged as part of the Performance Toolbox Aide (PAIDE) package companion of the Performance Toolbox for AIX. In AIX 4.3.2, which is a prerequisite for PSSP 3.1, the SPMI library is shipped in the perfagent.tools fileset and not in the perfagent.server component as in previous releases. Although most of the PSSP components will not use the SPMI library, the aixos resource monitor will need it in order to provide resource variables to Event Management. In summary, the perfagent.tools components is a pre-requisite for PSSP 3.2 running on AIX 4.3.2.

Question 3 - The answer is A. The RS/6000 Cluster Technology (RSCT) is a prerequisite for the PSSP. It was packaged as ssp.ha in releases prior to PSSP 3.1. These filesets provide the program and the configuration files for the three key components within the RS/6000 SP (Topology Services, Group Services, and Event Management).

A.8 Frames and Nodes Installation

Answers to questions in 9.5, “Sample Questions” on page 276, 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 C. The `spsvrmgr` command is used for checking frame and node supervisor microcode levels. The `-G` flag will contact all the frame supervisor cards in the system. Refer to 9.2.3, “Check the Level of Supervisor Microcode” on page 252 for details.

Question 3 - 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 4 - The answer is A. The `nodecond` script runs on the control workstation, and it accesses each node’s console through a read/write s1term. The s1term uses the RS-232 line to the frame for opening the console of a node in that frame. The Ethernet network is not used until the node starts network booting after the `nodecond` script has selected all the necessary options in the network boot menu.

A.9 Verification Commands and Methods

Answers to questions in 10.8, “Sample Questions” on page 293, 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 10.3.1.2, “Checking the SDR Initialization: SDR_test” on page 280 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 `spmon -d`

command will only report node information. Refer to 10.3.4.2, “Monitoring Hardware Activity: spmon -d” on page 285 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 to contact the frame supervisor cards every five seconds by default.

A.10 Understanding Additional SP-Related Products

Answers to questions in 11.8, “Sample Questions” on page 307, 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.

A.11 Application Specific Resources

Answers to questions in 12.6, “Sample Questions” on page 342, are as follows:

Question 1 - 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 may use the `dsh` command to run the `lsvsd` command on all nodes.

Question 2 - 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 them into a suspended state that can be made active by using the resumevsd command afterwards. Refer to 12.2.5, “Changing States of Virtual Shared Disks” on page 320 for details.

Question 3 - The answer is B. In GPFS, there is no concept of a GPFS server or client node. A GPFS node is whatever node that 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 4 - The answer is A. The GPFS subsystem is a system resource controlled subsystem called mmfs. The name comes from the multimedia AIX product (video streamer) that was developed in San Jose, California. GPFS shares this common past; so, that is why the mmfs name for the multimedia file system.

A.12 Problem Management Tools

Answers to questions in 13.8, “Sample Questions” on page 365, 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 13.5.1, “Authorization” on page 353 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.

A.13 RS/6000 SP Software Maintenance

Answers to questions in 14.7, “Sample Questions” on page 383, 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. The `-G` flag has to be used in order 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.

A.14 RS/6000 SP Reconfiguration and Update

Answers to questions in 15.9, “Sample Questions” on page 410, 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.

A.15 Problem Diagnosis

Answers to questions in 16.12, “Sample Questions” on page 453, 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. 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.

Appendix B. Special Notices

This publication is intended to help IBM Customers, Business Partners, IBM System Engineers, and other RS/6000 SP specialists who are involved in Parallel System Support Programs (PSSP) projects including the education of RS/6000 SP professionals responsible for installing, configuring, and administering PSSP. The information in this publication is not intended as the specification of any programming interfaces that are provided by Parallel System Support Programs. See the PUBLICATIONS section of the IBM Programming Announcement for PSSP for more information about what publications are considered to be product documentation.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter 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 the IBM Director of Licensing, IBM Corporation, 500 Columbus Avenue, Thornwood, NY 10594 USA.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The information about non-IBM ("vendor") products in this manual has been supplied by the vendor and IBM assumes no responsibility for its accuracy or completeness. The use of this information or the implementation of any of these techniques is a customer

responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

Any performance data contained in this document was determined in a controlled environment, and therefore, the results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.

Reference to PTF numbers that have not been released through the normal distribution process does not imply general availability. The purpose of including these reference numbers is to alert IBM customers to specific information relative to the implementation of the PTF when it becomes available to each customer according to the normal IBM PTF distribution process.

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

IBM ®	AIX
BookManager	Global Network
ESCON	HACMP/6000
LoadLeveler	OS/390
POWERparallel	RS/6000
S/390	SP
System/390	TURBOWAYS
VM/ESA	

The following terms are trademarks of other companies:

C-bus is a trademark of Corollary, Inc.

Java and HotJava are trademarks of Sun Microsystems, Incorporated.

Microsoft, Windows, Windows NT, and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation.

PC Direct is a trademark of Ziff Communications Company and is used by IBM Corporation under license.

Pentium, MMX, ProShare, LANDesk, and ActionMedia are trademarks or registered trademarks of Intel Corporation in the U.S. and other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

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

Appendix C. Related Publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

C.1 International Technical Support Organization Publications

For information on ordering these ITSO publications see “How to Get ITSO Redbooks” on page 475.

- *PSSP 3.1 Announcement,* SG24-5332
- *Inside the RS/6000 SP*, SG24-5145
- *RS/6000 SP: PSSP 2.2 Survival Guide*, SG24-4928
- *GPFS: A Parallel File System*, SG24-5165
- *IBM 9077 SP Switch Router: Get Connected to the SP Switch*, SG24-5157
- *A Holistic Approach to AIX V4.1 Migration*, SG24-4653
- *RS/6000 Scalable POWERparallel Systems*, SG24-4542
- *Technical Presentation on PSSP Version 2.3*, SG24-2080
- *RS/6000 SP Monitoring: Keeping It Alive*, SG24-4873
- *SP Perspectives: A New View of Your SP*, SG24-5180
- *RS/6000 SP PSSP 2.2 Technical Presentation*, SG24-4868
- *Elements of Security: AIX 4.1*, GG24-4433
- *IBM RS/6000 SP Management, Easy, Lean, and Mean*, GG24-2563

C.2 Redbooks on CD-ROMs

Redbooks are also available on the following CD-ROMs: **Order a subscription** and receive updates 2-4 times a year.

CD-ROM Title	Collection Kit Number
System/390 Redbooks Collection	SK2T-2177
Networking and Systems Management Redbooks Collection	SK2T-6022
Transaction Processing and Data Management Redbook	SK2T-8038
Lotus Redbooks Collection	SK2T-8039
Tivoli Redbooks Collection	SK2T-8044
AS/400 Redbooks Collection	SK2T-2849

CD-ROM Title	Collection Kit Number
RS/6000 Redbooks Collection (HTML, BkMgr)	SK2T-8040
RS/6000 Redbooks Collection (PostScript)	SK2T-8041
RS/6000 Redbooks Collection (PDF Format)	SK2T-8043
Application Development Redbooks Collection	SK2T-8037

C.3 Other Publications

These publications are also relevant as further information sources:

- *PSSP: Installation and Migration Guide*, GA22-7347
- *IBM Parallel System Support Programs for AIX: Diagnosis Guide*, GA22-7350
- *PSSP: Command and Technical Reference, Volume 1* and Volume 2, SA22-7351
- *IBM Parallel System Support Programs for AIX: Managing Shared Disks*, SA22-7349
- *IBM RS/6000 SP Planning Volume 1, Hardware and Physical Environment*, GA22-7280
- *IBM RS/6000 SP Planning Volume 2, Control Workstation and Software Environment*, GA22-7281
- *332 MHz Thin and WIde Node Service*, GA22-7330
- *AIX Version 4.3 System Management Guide: Communications and Networks*, SC23-4127
- *Site and Hardware Planning Information*, SA38-0508
- *AIX Version 4.3 Commands Reference Volumes*, SC23-4119
- *PSSP: Administration Guide*, SA22-7348
- *PSSP: Installation and Migration Guide*, GC23-3898
- *PSSP: Command and Technical Reference*, GC23-3900
- *PSSP: Administration Guide*, GC23-3897
- *AIX Problem Solving GUide and Reference*, SC23-4123
- *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7349
- *PSSP: Diagnosis and Messages*, GC23-3899

- *AIX 4.3 Network Installation Management Guide and Reference*, SC23-4113
- *General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278
- *AIX V4.3 Messages Guide and Reference*, SC23-4129
- *AIX 4.3 Network Installation Management Guide and Reference*, SC23-4113

How to Get ITSO Redbooks

This section explains how both customers and IBM employees can find out about ITSO redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** http://www.redbooks.ibm.com/

  Search for, view, download or order hardcopy/CD-ROM redbooks from the redbooks web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this redbooks site.

  Redpieces are redbooks in progress; not all redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

  Send orders via e-mail including information from the redbooks fax order form to:

	e-mail address
In United States	usib6fpl@ibmmail.com
Outside North America	Contact information is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl/

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl/

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl/

This information was current at the time of publication, but is continually subject to change. The latest information for customer may be found at http://www.redbooks.ibm.com/ and for IBM employees at http://w3.itso.ibm.com/.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and redbooks by accessing the IBM Intranet Web site at http://w3.itso.ibm.com/ and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may also view redbook. residency, and workshop announcements at http://inews.ibm.com/.

IBM Redbook Fax Order Form

Please send me the following:

Title	Order Number	Quantity

First name	Last name	
Company		
Address		
City	Postal code	Country
Telephone number	Telefax number	VAT number

☐ Invoice to customer number _____

☐ Credit card number _____

Credit card expiration date	Card issued to	Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

List of Abbreviations

ACL	Access Control Lists
ADSM	ADSTAR Distributed Storage Manager
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
BSD	Berkeley Software Distribution
BUMP	Bring-Up Microprocessor
CDS	Cell Directory Service
CEC	Central Electronics Complex
CLIO/S	Client Input Output Socket
CP	Crown Prince
CPU	Central Processing Unit
CSMA/CD	Carrier Sense, Multiple Access/Collision Detect
CSS	Communication Subsystem
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
FDDI	Fiber Distributed Data Interface
FIFO	First-In First-Out
FLDB	Fileset Location Database
FS	File System
GB	Gigabytes
GL	Group Leader
GPFS	General Purposes File System
GS	Group Services
GSAPI	Group Services Application Programming Interface
GUI	Graphical 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
HIPS	High Performance Switch
HRD	Host Respond Daemon
HSD	Hashed Shared Disk
IBM	International Business Machines Corporation
IP	Internet Protocol
ISB	Intermediate Switch Board
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 Nodes
MB	Megabytes
MCA	Micro Channel Architecture
MI	Manufacturing Interface
MIB	Management Information Base
MPI	Message Passing Interface
MPL	Message Passing Library
MPP	Massive Parallel Processors
NFS	Network File System
NIM	Network Installation Management
NIS	Network Information System
NSB	Node Switch Board
NSC	Node Switch Chip
NVRAM	Nonvolatile Memory
OID	Object ID
ODM	Object Data Management
OLTP	On-Line Transaction Processing
OSF	Open Software Foundation
P2SC	POWER2 Super Chip
PAIDE	Performance Aide for AIX
PE	Parallel Environment
PID	Process ID
PMAN	Problem Management
PP	Physical Partition
PSSP	Parallel System Support Programs
PTC	Prepare to Commit
PTPE	Performance Toolbox Parallel Extensions
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	RS/6000 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	Simple Network Management Protocol
SPMI	System Performance Measurement Interface
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

Index

Symbols

Numerics

A

B

C

D

E

F

L

M

N

P

R

S

T

U

V

ITSO Redbook Evaluation

IBM Certification Study Guide RS/6000 SP
SG24-5348-00

Your feedback is very important to help us maintain the quality of ITSO redbooks. **Please complete this questionnaire and return it using one of the following methods:**

- Use the online evaluation form found at http://www.redbooks.ibm.com
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Which of the following best describes you?
_ **Customer** _ **Business Partner** _ **Solution Developer** _ **IBM employee**
_ **None of the above**

Please rate your overall satisfaction with this book using the scale:
(1 = very good, 2 = good, 3 = average, 4 = poor, 5 = very poor)

Overall Satisfaction _____

Please answer the following questions:

Was this redbook published in time for your needs? Yes___ No___

If no, please explain:

_____

_____

_____

_____

What other redbooks would you like to see published?

_____

_____

_____

Comments/Suggestions: (THANK YOU FOR YOUR FEEDBACK!)

_____

_____

_____

_____

SG24-5348-00
Printed in the U.S.A.

IBM Certification Study Guide RS/6000 SP

SG24-5348-00