RS/6000 SP

Planning Volume 2, Control Workstation and Software Environment

GA22-7281-07

RS/6000 SP

Planning Volume 2, Control Workstation and Software Environment

GA22-7281-07

Note!
Before using this information and the product it supports, read the information in “Notices” on page 301.

Eighth Edition (October 2002)

This edition applies to version 3 release 5 of the IBM Parallel System Support Programs for AIX (PSSP) licensed program (number 5765-D51) and to all subsequent releases and modifications until otherwise indicated in new editions. This edition replaces GA22-7281-06. Significant changes or additions to the text and illustrations are indicated by a vertical line (|) to the left of the change.

IBM welcomes your comments. A form for readers' comments may be provided at the back of this publication, or you may address your comments to:

International Business Machines Corporation
Department 55JA, Mail Station P384
2455 South Road
Poughkeepsie, NY 12601-5400
United States of America

FAX (United States and Canada): 1+845+432-9405
FAX (Other Countries):
Your international Access Code +1+845+432+9405

IBMLink (United States customers only): IBMUSM10(MHVRCFS)
Internet e-mail: mhvrcfs@us.ibm.com

If you want a reply, be sure to include your name, address, telephone number, or FAX number.

Make sure to include the following in your comment or note:

- Title and order number of this book
- Page number or topic related to your comment

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

Contents

Figures

Tables

About this book

This book helps you plan for the installation of new software or migration of existing software to the newest levels on new, changed, or existing IBM systems where PSSP 3.5 is to be the primary system management software, optionally working cooperatively with related IBM licensed programs. The applicable systems are:

- The IBM RS/6000 SP system
- The IBM @server Cluster 1600 system managed by PSSP

RS/6000 SP systems are specific types of Cluster 1600 systems that are being managed by the PSSP software product. As you read this book, keep this in mind:

- ***A Cluster 1600 system that has an SP building block, and is managed by PSSP, is an SP system.***
- Except where otherwise explicitly stated, statements in this book about function on the SP system or nodes apply to all nodes running the PSSP software, regardless of the physical system configuration in which they participate.

This book tells you what your control workstation and IBM system software options are, what to consider for integrating the system into your existing computing network, what decisions you need to make, and what information to prepare and record to facilitate the successful installation or migration of software on your system. A hardware overview is included and hardware considerations are discussed throughout the book, but only as they relate to the successful operation of the control workstation and software environment.

Do not attempt to install, expand, or migrate your system without first reading this book. Read this book and complete the worksheets in Appendix C, “SP system planning worksheets” as you plan your control workstation and software environment.

Your software choices can lead to hardware requirements. Work closely with your hardware planners. The book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* provides information to help you understand hardware requirements and scaling limits, plan your physical environment, and prepare your site for hardware installation.

For a list of related books and information about accessing online information, see “Bibliography” on page 313.

This book applies to PSSP version 3 release 5. To find out which version of PSSP is running on your control workstation (node 0), run the following command:

```
splst_versions -t -n0
```

In response, the system displays something similar to:

```
0 PSSP-3.5
```

Since this is a planning book, you might want to know what you have running on the nodes so that you can plan for your next install or upgrade. To find out which version of PSSP is running on the nodes on your system, run the following command from your control workstation:

```
splst_versions -t -G
```

The response indicates the version of PSSP that is running on each node. The system displays something similar to:

```
1 PSSP-3.5
2 PSSP-3.4
7 PSSP-3.2
8 PSSP-3.5
```

Save your old manuals!
If you are running mixed levels of PSSP, take care to keep and use the relevant books for every version of PSSP that you intend to continue running.

Who should use this book

This book is intended for experienced computer professionals responsible for planning and preparing for the network, control, and system installation of a new Cluster 1600 system managed by PSSP (SP system) or for the expansion or migration of an existing system.

This book assumes that you have a working knowledge of AIX or UNIX and experience with network systems. In addition, you should already know what the basic SP and AIX features are, and have a basic understanding of computer systems, networks, and applications.

Typographic conventions

This book uses the following typographic conventions:

Typographic	Usage
Bold	**Bold** words or characters represent system elements that you must use literally, such as commands, flags, and path names.
Italic	• *Italic* words or characters represent variable values that you must supply. • *Italics* are also used for book titles and for general emphasis in text.
`Constant width`	Examples and information that the system displays appear in `constant width` typeface. All references to the hypothetical customer, `Corporation ABC`, and any choices made by `Corporation ABC` are in this font.
[ ]	Brackets enclose optional items in format and syntax descriptions.
{ }	Braces enclose a list from which you must choose an item in format and syntax descriptions.
|	A vertical bar separates items in a list of choices. (In other words, it means “or.”)
< >	Angle brackets (less-than and greater-than) enclose the name of a key on the keyboard. For example, **<Enter>** refers to the key on your terminal or workstation that is labeled with the word Enter.
...	An ellipsis indicates that you can repeat the preceding item one or more times.
<Ctrl-*x*>	The notation **<Ctrl-*x*>** indicates a control character sequence. For example, **<Ctrl-c>** means that you hold down the control key while pressing **<c>**.
\	The continuation character is used in coding examples in this book for formatting purposes.

Software level notation

The following demonstrates the meaning of the notation used in this book:

PSSP 3.2 or later	PSSP version 3 release 2 and any modification and fix level, or later version, release, modification, and fix level.
AIX 5L 5.1	AIX 5L version 5 release 1 and any modification and fix level
AIX 4.3.3	Only AIX version 4 release 3 modification level 3, and any fix level
AIX 4.3.3 (or later)	AIX version 4 release 3 modification 3, and any fix level, or later version, release, modification, and fix level.

Part 1. Planning your system

Chapter 1. Introduction to system planning

IBM Cluster 1600 systems managed by PSSP are comprised of hardware and software building blocks. They also involve a continually changing set of human requirements. In order to get the highest level of performance out of your system, you need to plan for all of the internal and external activities. System planning produces the solid foundation you need for managing your system as it evolves over time. Some of the basic areas you have to plan for include:

- Network design
- The physical equipment and its operational software
- Operational environments
- System security and authentication
- System partitions
- Migration and coexistence on existing systems
- Defining user accounts
- Backup procedures

This chapter introduces the hardware and software building blocks of a Cluster 1600 system managed by PSSP that you might want to integrate into your operational environment, describes hardware in a general sense, and gets your project team started on the planning tasks. “Question 3: Which related IBM licensed programs do you need?” on page 25 describes the software building blocks in a general sense. Other chapters provide more details and planning information.

If you already have an SP system and want to move to the newest level of AIX and PSSP software, you need to plan your migration, possibly by taking advantage of coexistence or system partitioning.

You can choose to contract with IBM to plan and install your system. Contact your IBM representative if you want help with these tasks.

Note!

Be sure to read “About this book” on page xv for important statements about the systems to which discussions in this book apply.

The topics addressed are:

- “Planning services”.
- “System overview” on page 4.
- “Hardware overview” on page 5.
- “Software overview” on page 12.
- “What’s new in AIX and PSSP?” on page 13.
- “Planning issues” on page 16.
- “Using this and other books for planning” on page 16.

Planning services

This optional IBM service provides a specialist on site to assist you with planning your implementation. Activities offered with this service include:

- Planning for integrating a Cluster 1600 system managed by PSSP into your network.
- Defining name service requirements.
- Defining volume group and file system.
- Planning for migration.
- Defining accounting practices and policies.
- Defining security policies.

For further details, call 1-800-CALLAIX.

System overview

An IBM @server Cluster 1600 system managed by PSSP (SP system) brings together ***cluster hardware building blocks*** and ***cluster software building blocks*** into a unified cluster with its own unique identification number. A single system number simplifies cluster administration and makes it easier for customers to track, manage, and upgrade their clusters over time.

The hardware building blocks include a choice of cluster interconnect technologies, either IBM or industry standard, and cluster-enabled IBM pSeries and IBM RS/6000 servers:

- RS/6000 SP nodes
- S70
- S7A
- S80
- IBM @server pSeries 680
- H80
- M80
- IBM @server pSeries 660 Model 6H1
- IBM @server pSeries 660 Model 6M1
- IBM @server pSeries 660 Model 6H0
- IBM @server pSeries 690
- IBM @server pSeries 670

See “Hardware overview” on page 5 to learn about the overall system hardware. See “Question 8: Which and how many nodes do you need?” on page 43 to learn more about specific hardware building blocks.

The IBM Cluster 1600 software building blocks, cluster management software, and parallel computing tools, include:

- Parallel System Support Programs for AIX (PSSP) 3.5

  Offers a central point-of-management control. Built upon the system management tools and commands of the AIX operating system, PSSP allows system administrators and operators to manage clustered nodes easily and cost effectively. PSSP allows all local and remote administrative functions to be performed from a centralized control workstation.
- General Parallel File System (GPFS) 1.6

  Available for both AIX and Linux Clusters, GPFS provides a cluster-wide file system allowing users shared access to files spanning multiple disk drives. GPFS is based on a shared disk model, providing lower overhead access to disks not directly attached to the application nodes, and using a distributed protocol to provide data coherence for access from any node.
- High Availability Cluster Multi-Processing for AIX (HACMP) 4.5

HACMP allows continuous access to data and applications typically through component redundancy and failover in mission-critical environments. With HACMP, customers can scale up to 32 nodes and mix and match system sizes and performance levels, as well as network adapters and disk subsystems to satisfy specific application, network and disk performance needs.

- High Availability Geographic Cluster for AIX (HAGEO) and Geographic Remote Mirror for AIX (GeoRM) 2.3

  Geographic clustering software is designed to help ensure that data and business critical applications are continuously available, even when a natural disaster threatens the computer complex. HAGEO and GeoRM extend an HACMP cluster to include two physically separate data centers. Data entered at one site is sent across a point-to-point TCP/IP network and is mirrored at a second geographically distant location. If a disaster disables one site, the data is available within minutes at the other site.
- LoadLeveler for AIX 5L (LL) 3.1

  Used for dynamic workload scheduling, LoadLeveler is a distributed network-wide job management facility designed to dynamically schedule work on servers, such as the IBM @server pSeries and IBM RS/6000 systems.
- Parallel Environment for AIX 5L (PE) 3.2

  Used to develop, debug, analyze, tune and execute parallel processing applications.
- Engineering and Scientific Subroutine Library (ESSL) 3.3

  The ESSL libraries provide a variety of complex mathematical functions for many different scientific and engineering applications.
- Parallel ESSL 2.3

  Provides optimum performance for floating-point-intensive engineering and scientific workloads.

See “Software overview” on page 12 to learn more about the overall system software. See “Question 3: Which related IBM licensed programs do you need?” on page 25 to learn more about these specific licensed programs.

Hardware overview

A Cluster 1600 system managed by PSSP provides an overall view and a unique identification number for your cluster-ready hardware building blocks. In general terms, the hardware that can comprise a Cluster 1600 system where PSSP is the primary system management software are:

- Processor nodes
- Frames
- Switches
- Extension nodes
- A control workstation
- Network connectivity and I/O adapters
- SP Expansion I/O units

These components connect to each other by a PSSP administrative local area network (LAN). That network is often called the SP Ethernet admin LAN, the SP LAN, or the SP Ethernet. Remember, those names are also used in a Cluster 1600 system that is managed by PSSP even if it does not have an SP. The nodes can connect to your existing computer network through another LAN, making the Cluster 1600 system accessible from any network-attached workstation.

Hardware details are in other books.

Keep in mind that this is merely a high level overview explaining some physical features used as points of reference in later discussions regarding the software support. Each type of hardware has a set of requirements. Be sure to read the books relevant to your system for physical specifications, connectivity, requirements, and scaling limits:

- *IBM @server Cluster 1600 Hardware Planning, Service and Installation*
- *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*
- See also "Related hardware publications" on page 315 for where to find hardware publications about IBM @server pSeries servers.

The topics addressed in this section are:

Processor nodes

As used in this book, generally a ***node*** is where PSSP is running along with AIX and PSSP-related licensed programs. PSSP 3.5 can run on nodes in an SP frame or in any server supported as a hardware building block in a Cluster 1600 system. The supported servers can be connected to an SP frame, in which case they are called ***SP-attached servers***, or they can be in a Cluster 1600 system managed by PSSP with no SP frame, in which case they are called ***clustered servers***. Depending on the physical type, a server might have one node or might be physically partitioned into multiple nodes. These terms signify the system configuration in which nodes participate when running the PSSP software. Remember these terms because some supported features apply when nodes are in one system configuration and not in the other.

Remember, *clustered servers* are in a Cluster 1600 system configuration with no SP building block. That configuration is called a *clustered servers system*. With the PSSP 3.5 software on the control workstation and a supported level of PSSP on each node in a cluster of IBM @server pSeries or RS/6000 servers, a clustered servers system functions like an SP system with some variation due to configuration and hardware differences.

Both the SP and clustered servers system configurations are scalable over time to meet your changing computing requirements.

A Cluster 1600 system managed by PSSP can be comprised of:

- ***SP processor nodes***

  These are RS/6000 computers mounted in short or tall SP frames. They are of three types: thin nodes, wide nodes, and high nodes. These are sometimes called SP rack-mounted nodes. The frame spaces into which nodes fit are called drawers. A tall frame has eight drawers, while a short frame has four drawers. Each drawer is further divided into two slots. One slot can hold one thin node. A

single thin node in a drawer, one that is not paired with another thin node in the same drawer, must occupy the odd numbered slot. A wide node occupies one drawer (two slots) and a high node occupies two drawers (four slots). The SP system is scalable from one to 128 processor nodes that can be contained in multiple SP frames in standard configurations. The maximum number of high nodes supported ranges from 64 to 128 depending on which high nodes you have. Systems that can have from 129 to 512 nodes are available by special bid.

Figure 1 illustrates a basic SP suitable for parallel and serial batch technical computing in a departmental setting.

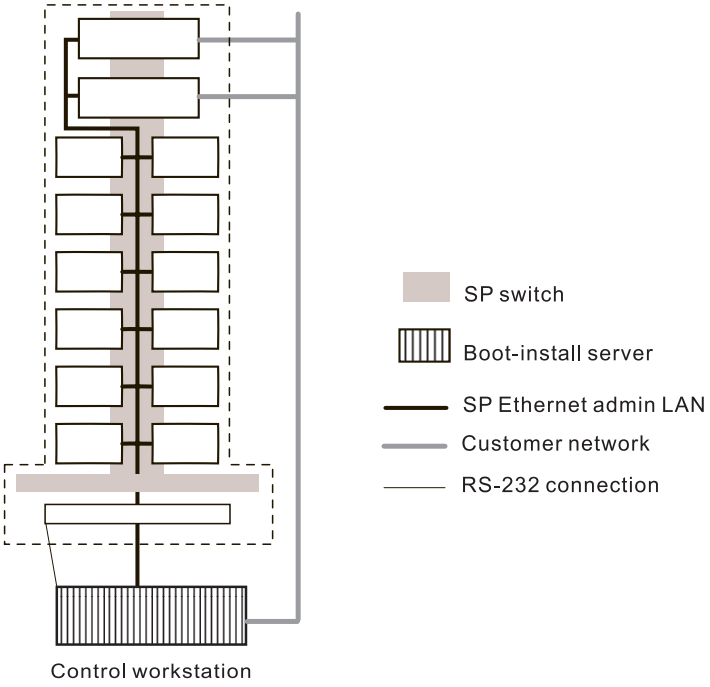


Figure 1. Basic SP configuration

- ***SP-attached servers***

  These are nodes that are not mounted in an SP frame. Generally, they are 24 inch or 19 inch rack-mounted nodes. Some are in physical units that might resemble an SP frame. They connect directly to the control workstation. They connect to the SP directly by the PSSP admin LAN. Some have limited hardware control and monitoring from the control workstation because they have no SP frame supervisor or SP node supervisor. Others do have hardware control and monitoring capabilities comparable to an SP frame. Except for the physical differences, after they are installed and running the PSSP software, they function just like SP processor nodes and interact with the other nodes in the system. Except for the IBM @server pSeries 690 and 670 running in LPAR mode, each physical server is managed by the PSSP software as a separate frame with one node.

  The p690 and p670 have physical components that can be assigned to logical partitions (LPARs) that can be multiple nodes in one frame. Those servers have features that are similar to an SP frame. Each server can have several LPARs managed by the PSSP software as nodes in one frame, however, there are constraints in a system with a switch configuration.

  The number of nodes in SP-attached servers counts toward the maximum number of nodes in the SP system. The number of SP-attached servers counts toward the maximum number of frames with nodes in the system.

Figure 2 illustrates a basic SP system that includes one SP-attached server.

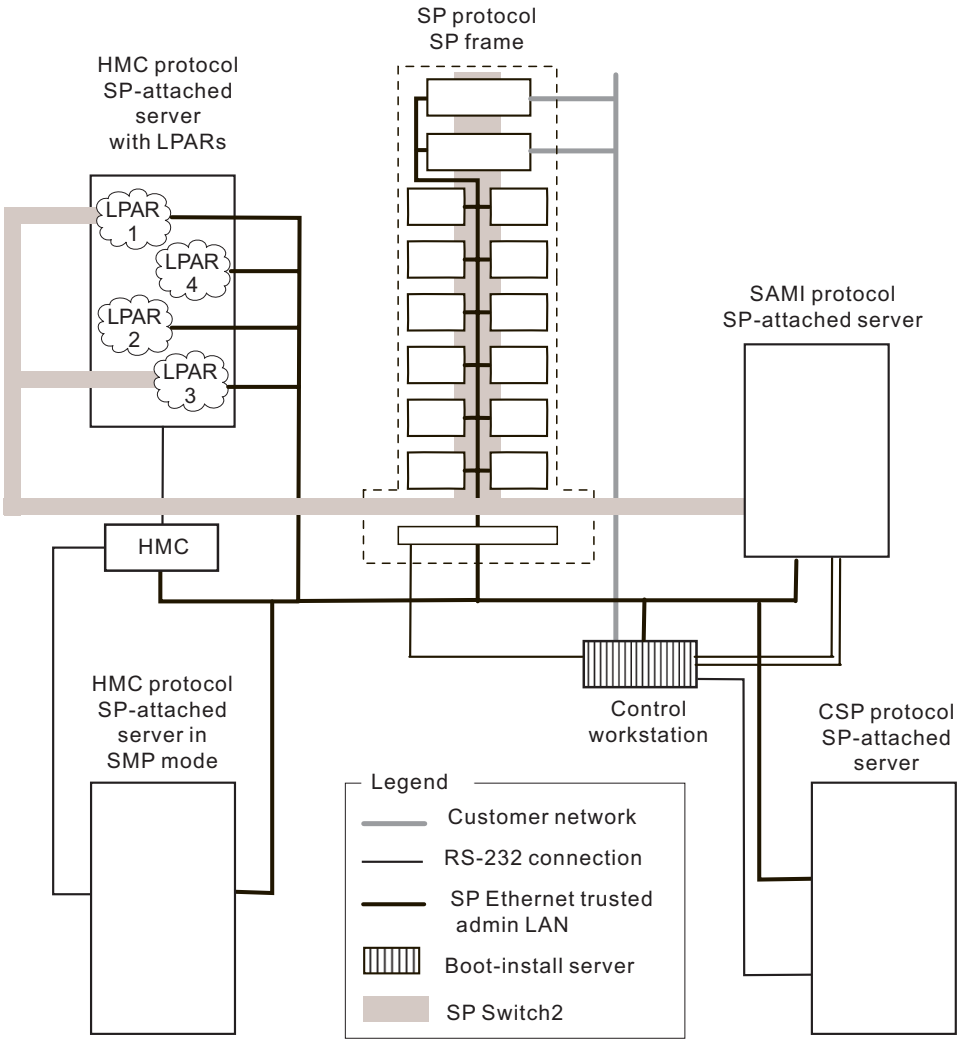


Figure 2. SP with an SP-attached server

A Cluster 1600 system managed by PSSP with no SP building block is comprised of ***clustered servers***. The same machine types supported as SP-attached servers can participate in a clustered servers system configuration. Figure 3 on page 9 illustrates a clustered servers system. When the control workstation and each of the servers are appropriately connected and running the PSSP software and there are no SP nodes or frames, the term *clustered servers system* applies as explained here and used in other PSSP publications.

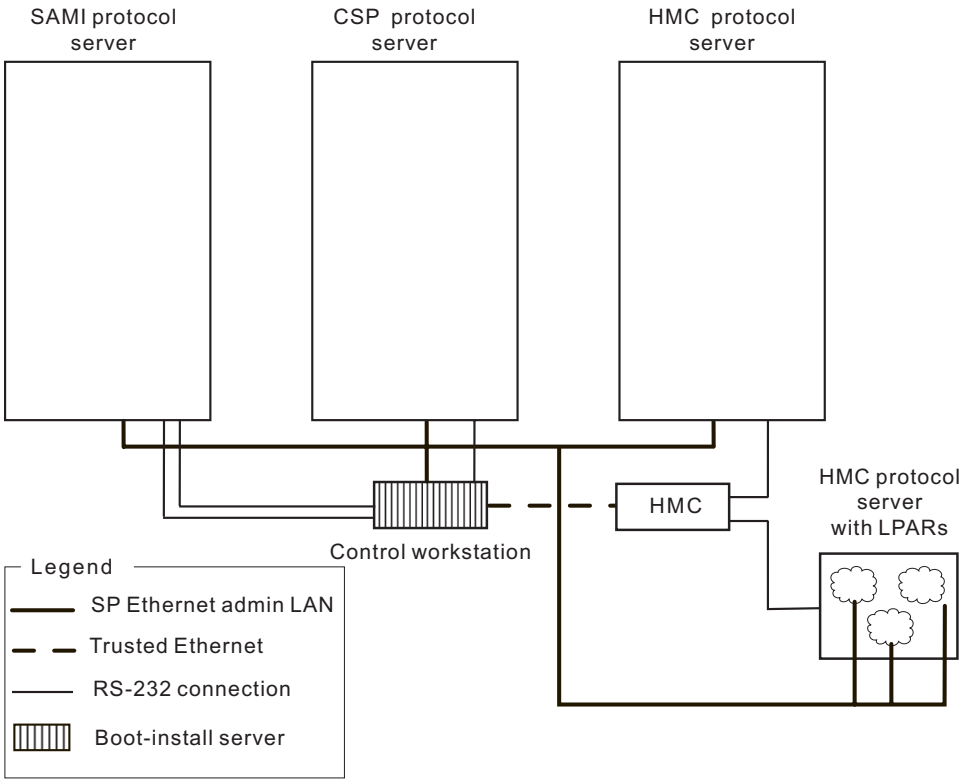


Figure 3. A clustered servers system

Keep in mind

Unless otherwise explicitly stated, the information in this book about nodes applies to all nodes that run PSSP, whether in an SP frame, in an SP-attached server, or in a clustered server or LPAR. Functionally they are all simply nodes in the system.

These processor nodes are all symmetric multiprocessor (SMP) computers with varying levels of function, capacity, and performance. Each processor node includes memory, internal direct access storage devices (DASD) that are optional in some nodes, optional connectivity to external networks and DASD, and a method for Ethernet connection. The type of node and optional equipment it contains can lead to other requirements.

Base your choice of processor nodes on the function and performance you require today and in the foreseeable future. Thin nodes are typically configured as compute nodes, while wide nodes are more often used as servers to provide high-bandwidth data access. High nodes are typically used for database operations and for applications with extensive use of floating point. SP-attached servers are particularly suitable in SP systems with large serial databases. If you do not require a full scale SP system, a system of clustered servers might be right for you. No rigid rule governs the logical configuration of a node. You can configure any physical node type for the logical functions that best serve your computing requirements.

Note: Remember, this is an overview. Do not make your choices before reading the planning information about the servers that are currently available from IBM as nodes on which you can run the PSSP 3.5 software. The maximum

number of servers supported depends on the types and configuration. See the information in ″Chapter 2, “Defining the system that fits your needs”″ under the heading “Question 8: Which and how many nodes do you need?” on page 43.

Frames

SP frames have spaces into which the nodes fit. These spaces are called drawers. A tall frame has eight drawers and a short frame has four drawers. Each drawer is further divided into two slots. One slot can hold one thin node or SP Expansion I/O Unit. A wide node occupies one drawer (two slots) and a high node occupies two drawers (four slots). An internal power system is included with each frame. Frames get equipped with the optional processor nodes and switches that you order.

SP processor nodes can be multiply mounted in a tall or short SP frame. The maximum number of SP frames with nodes supported in an SP system is 128. Frames with only switches or SP I/O Extension Units can be numbered from 129 to 250 inclusive in order to allow the maximum of 128 frames with nodes in a standard system. The maximum number of high nodes supported in a 128-frame SP system varies depending on which high nodes you have. The SP system supports up to 128 POWER3 SMP high nodes while older 604 series high nodes are limited to 64. If your system is fully populated with SP rack-mounted nodes, there is no room for SP-attached servers.

Servers, whether configured as SP-attached or clustered, are conceptually *self-framed* and apply 1 to 1 in the count of frames. The maximum number of frames and servers depends on the types and the system configuration that are supported in a clustered servers system.

Switches

Switches are used to connect processor nodes, providing the message passing network through which nodes communicate with a minimum of four disjoint paths between any pair of nodes. In any complete Cluster 1600 system managed by PSSP, you can use only one type of switch, either the SP Switch2 or the SP Switch.

SP-attached servers can be connected to a switch in a tall SP frame. SP-attached servers are not supported with short SP frames.

By definition, a clustered servers system is a Cluster 1600 system managed by PSSP with no SP frames. However, nodes in a clustered configuration are also supported with either switch. Though you are not required to have an SP frame or SP node in order to use the PSSP software, you do need the appropriate frame for any SP switch you decide to use. If you add an SP switch to what might otherwise be a clustered servers system, the system becomes technically an SP system and the nodes become SP-attached, because the node numbering and placement rules to be honored are those related to the SP Switch2 or the SP Switch, whichever you add.

Adapters are required to connect any processor node or extension node to the switch subsystem. See the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for which adapter is required for each supported node.

To consider whether you need a switch and which switch to choose, see “Choosing a switch” on page 21.

Extension nodes

Extension nodes are non-processor nodes that extend the capabilities of the SP system, but cannot be used in the same ways as SP processor nodes.

A specific type of extension node is a dependent node. A dependent node depends on SP processor nodes for certain functions, but much of the switch-related protocol that processor nodes use is implemented on the SP Switch.

A physical dependent node can support multiple dependent node adapters. If a dependent node contains more than one dependent node adapter, it can route data between SP system partitions. ***The only node of this type is the SP Switch Router. It is available only to enhance an SP system that uses the SP Switch.*** Data transmission is accomplished by linking the dependent node adapters in the SP Switch Router with valid switch ports on the SP Switch. If these SP Switches are located in different SP system partitions, data can be routed at high speed between the system partitions.

The SP Switch Router can be used to scale your SP system into larger systems through high speed external networks such as a FDDI backbone. It can also dramatically speed up TCP/IP, file transfers, remote procedure calls, and relational database functions.

Control workstation

The SP system uses an IBM RS/6000 or pSeries workstation with a suitable hardware configuration, the PSSP software, and other optional software as a central point of control for managing and maintaining the nodes and related hardware and software (see “Question 10: What do you need for your control workstation?” on page 70). An authorized system administrator can log in to the control workstation from any other workstation on the PSSP admin LAN to perform system management, monitoring, and control tasks.

The control workstation connects directly to each SP frame and server to provide hardware control functions. Each server connects directly to the control workstation. Depending on which machine types you choose to have as nodes in your system, the hardware control might be comparable to that of an SP frame or it might be minimal. Only some servers have features that are comparable to an SP frame and node supervisor.

The control workstation acts as a boot-install server for nodes in the system. For security, the control workstation can be set up as a Distributed Computing Environment (DCE) or Kerberos Version 4 (V4) authentication server. See Chapter 6, “Planning for security” on page 145 for more information.

The High Availability Control Workstation option enables you to have a primary and secondary control workstation for automatic failover and reintegration in the event that the primary control workstation is not available. See Chapter 4, “Planning for a high availability control workstation” on page 129 for more information.

Network connectivity and I/O adapters

Network connectivity is supplied by various adapters, some built in, some optional, that can provide connection to I/O devices, networks of workstations, and mainframe networks. Ethernet, FDDI, token-ring, HIPPI, SCSI, FCS, and ATM are some types of adapters that can be used.

The PSSP admin LAN is the network that connects all nodes to each other and to the control workstation. An Ethernet cable is provided with each frame to use in the wiring of this network. Additional optional adapters such as Ethernet, FDDI, and token-ring are automatically configured on each node. Other optional adapters are supported and can be individually configured on each node.

Note: See the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for information about required and optional adapters.

SP Expansion I/O Unit

An SP Expansion I/O Unit is designed to satisfy the needs of customers running applications with a greater demand for internal DASD, external DASD, and network connectivity than is available in the node alone. The unit expands the capacity of a node by providing eight PCI slots and up to four hard disks. These hard disks are considered internal DASD of the associated node.

The SP Expansion I/O Unit has the following characteristics and restrictions:

- It is available only for connecting to a POWER3 SMP ***high*** node.
- It connects to a POWER3 SMP high node directly by cable.
- Up to six SP Expansion I/O Units can connect to one POWER3 SMP high node.
- Each unit is about the size of a thin node and occupies a thin node slot in a tall SP frame (it is not supported in a short frame).
- It can be in any tall frame in the SP system, not necessarily in the same frame as the node to which it connects.
- It is irrelevant to SP system partitioning. Only the node to which it is connected is considered to be in a system partition.
- It does not support connection to a switch. The node to which it is connected can be connected to a switch.
- It does not support twin-tailed internal I/O.

Software overview

The Cluster 1600 system managed by PSSP software infrastructure includes:

- AIX 5L 5.1, the base operating system
- PSSP 3.5, a higher-level set of support programs and interfaces that enables you to take advantage of the powerful parallel processing features of a Cluster 1600 or SP system
- Other IBM system and PSSP-related licensed programs:
  - General Parallel File System (GPFS)
  - High Availability Cluster Multi-Processing (HACMP)
  - LoadLeveler
  - Parallel Environment
  - Engineering and Scientific Subroutine Library (ESSL)
  - Parallel ESSL

  See “Question 3: Which related IBM licensed programs do you need?” on page 25 to learn about these programs. See the “Bibliography” on page 313 for a listing of publications.

AIX

AIX is an integrated UNIX operating environment conforming to industry standards for open systems. It provides the basic operating system functions such as the AIXwindows user interface, extended real-time support, network installation management (NIM), advanced file system support, physical disk space management, and a platform for application development and execution. AIX capabilities that some PSSP components use for system management include the following:

- The Network Installation Management (NIM) environment provides the ability to install an AIX **mksysb** image over the network.
- The Logical Volume Manager (LVM) improves data management productivity, enables files to span multiple disk drives, and provides, with the disk mirroring and concurrent logical volume support, a high availability option for critical data.
- The System Management Interface Tool (SMIT) provides a single, consistent, and expandable interface to system management commands.
- The System Resource Controller (SRC) simplifies the management of subsystems and daemons.
- AIX device support lets you add and delete devices at any time without disrupting the system.
- The option to install either the 32-bit or 64-bit AIX kernel with PSSP 3.5 using AIX 5L Version 5.1. See “64-bit kernel” on page 15.

PSSP

PSSP is a comprehensive suite of applications to manage an SP system as a full-function parallel processing system. It provides a single point of control for administrative tasks and helps increase productivity by letting administrators view, monitor, and control system operation. Most functions are base components of PSSP while others are optional: they come with PSSP, but you choose whether to install and use them. The following are available as optional components:

- High Availability Control Workstation (HACWS) lets your control workstation be a single point of control while keeping it from becoming a single point of failure.
- The IBM Virtual Shared Disk management components:
  - IBM Virtual Shared Disk makes data on physical disks accessible from multiple nodes in IBM Virtual Shared Disks that you create; otherwise, the data is accessible only from the node connected to the disk.
  - IBM Recoverable Virtual Shared Disk makes IBM Virtual Shared Disks automatically recoverable on the event of a failure.
  - Hashed Shared Disk stripes your data across multiple IBM Virtual Shared Disks and multiple nodes.

See the book *PSSP: Administration Guide* for a summary of the PSSP components.

What’s new in AIX and PSSP?

New RS/6000 SP systems come with PSSP 3.5 and AIX 5L 5.1 on installation media. You can use these software building blocks for managing any Cluster 1600 system managed by PSSP even if it does not include the SP system hardware. The following sections briefly describe selected functions in AIX 5L as well as the latest modifications in PSSP 3.5. The AIX features included are only those of significant general interest or that relate particularly to the PSSP licensed program.

What's new in AIX 5L 5.1?

AIX 5L Version 5.1 represents the next generation of AIX. Fortified with open technologies from some of the world's top providers, AIX 5L builds on a solid heritage of supplying integrated, enterprise-class support for IBM @server pSeries and RS/6000 systems. It provides an industrial-strength UNIX operating system with increased levels of integration, flexibility, and performance for meeting the high demands of today's mission-critical applications. It offers an advanced operating system with a strong affinity with Linux and built-in capabilities designed to accommodate IBM POWER-based nodes, which include those supported as Cluster 1600 hardware building blocks.

These are some of the new or changed features:

- Support for the latest IBM @server pSeries systems, adapters, and I/O products.
- Network Authentication Service Version 1.2 improves performance of concurrent login and availability of Network Authentication Service servers is improved.
- Open Secure Shell (OpenSSH) 2.9.9, a set of client/server connectivity tools that is designed to encrypt network traffic to help eliminate network-level attacks is available with the Bonus Pack.

Additional technology is available with the Bonus Pack.

What's new in PSSP 3.5?

The PSSP 3.5 licensed program provides enhanced quality and support for the enablement of new nodes on which it can operate and the AIX 5L 5.1 operating system. Functional enhancements include new or expanded support in the following areas:

Hardware

Support has been added for the following hardware:

- After PSSP 3.4 became available, the IBM @server pSeries 670 was introduced. It is a high-end 4, 8, or 16-way Gigabit processor POWER4 system. The physical components can be assigned to separate logical partitions (LPARs) within one physical frame. You can have up to 16 LPARs. Each LPAR functions as an individual node within that frame and each is fully functional as a PSSP node in a Cluster 1600 or SP system. With PSSP 3.5, these nodes require AIX 5L 5.1. The p670 is supported in an SP-attached configuration with the SP Switch2, the SP Switch, or no switch, or in a clustered server configuration.
- The IBM 375/450 MHz POWER3 SMP Thin Node and Wide Node. These nodes have been available with 375 MHz processors in an SP frame. Now you can have these thin and wide nodes with 450 MHz processors.
- Additional node attachment and switch attachment adapters.

See "Hardware overview" on page 5 for introduction and references to planning information with respect to the software support of new nodes.

Switch

- With the control workstation running PSSP 3.4 or later software, you have optional switch connectivity. This means you can use the SP Switch2 with newer nodes and still keep older nodes in your system. Any node that is not supported on the SP Switch2 can remain in the system, but not connected to the SP Switch2.
- For both the SP Switch and SP Switch2, you can specify which nodes you want to exclude from serving as a switch primary or primary backup, and which nodes are available for that purpose.

SP-attached servers

SP system partitioning is supported by default in switchless systems with at least one SP node frame. In that case, the number of SP-attached servers you can have is limited by the number of available switch ports. But you might have no need for multiple SP system partitions while you might need more SP-attached servers than are accommodated by the available switch ports in your system. You can force a switchless SP system to be nonpartitionable. Then you can have more SP-attached servers because you can ignore the switch port numbering rules and assign sequential numbers.

64-bit kernel

PSSP 3.5 and the software stack it supports (this includes GPFS, MPI, LAPI, KLAPI, and IBM Virtual Shared Disks) now provide support for use of 64-bit kernels. This support is essential for running applications with large virtual memory requirements. 64-bit kernel support only exists on those control workstations and nodes that meet the following minimal requirements:

- Running PSSP 3.5 or later
- Running AIX 5L 5.1 or later
- Running the 64-bit kernel on hardware that AIX supports use of the 64-bit kernel on

Note that some applications may place additional restrictions on 64-bit kernel support. For example, some applications may require that all nodes that use that application must be using the 64-bit kernel if any one of them are.

File collections

The supman user ID is used by PSSP when distributing file collections. Not managing a supman AIX user ID password is a security risk. PSSP 3.4 and later releases provides new commands to initially set and routinely change the supman password. See the section ″Establish password for secure file collections,″ in the ″Installing and Configuring a new RS/6000 SP system″ chapter of *PSSP: Installation and Migration Guide*.

IBM Virtual Shared Disks

The IBM Virtual Shared Disk component of PSSP has performance improvements and improved flow control. With AIX trace hooks, you can better assess overall system problems.

Migration and coexistence

With PSSP 3.5, AIX 5L 5.1 must be on the control workstation. Support is provided for migrating to PSSP 3.5 running on AIX 5L 5.1. If you have any IBM @server pSeries 690 or 670 servers in your system, AIX 5L 5.1 must be on each p690 and p670 node along with the PSSP 3.5 or PSSP 3.4 software. PSSP 3.5 nodes can coexist with nodes that have PSSP 3.4 or PSSP 3.2.

Planning issues

You need to read all of this book. The planning steps you take depend on what you have now and which system configuration you want to have. Chapter 2, "Defining the system that fits your needs" on page 19 helps you define a system that meets your needs. The following list contains some of the major issues you need to consider when setting up your system:

- The type of computing your Cluster 1600 system managed by PSSP will perform.
- Possible operational benefits from using the Parallel Environment.
- Future expansion (scaling) plans for your system.
- The number of nodes you will need.
- The type of nodes you will need.
- The IBM @server pSeries or RS/6000 model you can use for a control workstation.
- High availability (system backup) requirements for data and hardware.
- The amount and type of data storage.
- External network connections for your Cluster 1600 system managed by PSSP.
- The language in which you want AIX and PSSP to operate.
- The security service to use for AIX remote command authentication.
- The security services to use for SP trusted services.
- Migration for system upgrades.
- The ability to use partitioning and coexistence as migration tools.
- Coexistence requirements and limitations.

Remember, as your planning begins to shape your system, you will need to work closely with your hardware planners. Many of the software decisions you make might create a related requirement in hardware, such as:

- Cables to connect frames, control workstations and extension nodes.
- The number and types of nodes affect power and cooling requirements.
- Data recovery and connections to external systems influence which adapters you need.

Using this and other books for planning

Work closely with your hardware planner because choices of hardware can lead to requirements on software, and choices of software can lead to requirements on hardware. This section briefly describes the content in Cluster 1600 and SP planning books and the PSSP books for your convenience. See "Bibliography" on page 313 for how to access the books.

To help you plan either a new Cluster 1600 system managed by PSSP or a changed SP system running the PSSP 3.5 and related software, books are available in the *IBM RS/6000 SP* planning library:

- Use the book *IBM @server Cluster 1600 Hardware Planning, Service and Installation* to plan and check that you have the correct physical configuration and environment for your Cluster 1600 system.
- Use the book *Planning Volume 1, Hardware and Physical Environment* to plan and check that you have the correct physical configuration and environment for your SP system.
- Use this book, *Planning Volume 2, Control Workstation and Software Environment*, to help you plan and make your decisions about which components

to install, which nodes to use for what purposes, and how to plan system expansions and software upgrades using migration, coexistence, and system partitioning. This book is organized in the following manner:

– Chapter 1 is a high level overview of the system, introducing the hardware, software, and available services.
– Chapters 2 through 9 address subjects you need to attend to when planning a new system or when planning to use a PSSP function or PSSP-related licensed program for the first time on a system you already have.
  - Chapter 2 leads you through a list of questions you need to address. As it does, you are introduced to the PSSP components and PSSP-related licensed programs that warrant your consideration during the planning phase, without yet burdening you with planning details. It also tells you more about some hardware options from which to choose. Make these choices with respect to what you need to support the software you are considering. When you are done with Chapter 2, you should have a good idea of which hardware and software is to comprise your system. You will also have prepared some of the information you need to do a software install or migration.
  - Chapter 3 guides you through what you need to consider and what rules to follow in deciding how to configure and layout your hardware, software, and network from an operational control workstation and software environment point of view. This chapter helps you prepare more of the information that you need to do a software install or migration.
  - Chapters 4 through 9 cover PSSP components and related licensed programs, some for a second time, this time giving you or directing you to planning details, requirements, and dependencies. Read each chapter to learn more about the respective subject before you make a final choice or proceed to installation or migration.
– Chapters 10 and 11 are useful for considering growth. They address activities that are primarily for customizing a system you already have:
  - Chapter 10 addresses what to consider if you want to expand your SP system.
  - Chapter 11 discusses what to consider before migrating to the newest PSSP and PSSP-related software.

The following books are available in the *PSSP* library:

- Use the book *Installation and Migration Guide* for new installations or migrations to the new PSSP software. Use this book for system changes as well.
- Use the book *PSSP: Managing Shared Disks* for planning, installing, and using the optional components of PSSP with which you can create, use, monitor, and manage IBM Virtual Shared Disks. The components are IBM Virtual Shared Disk, Hashed Shared Disk and Recoverable Virtual Shared Disk.
- Use the book *Administration Guide* for the day-to-day operation and management of your system.
- Use the book *Command and Technical Reference* for PSSP command descriptions and technical reference.
- Use the book *Diagnosis Guide* to help you diagnose problems.
- Use the book *Messages Guide* to find information about messages; what might have caused them and how to respond to them.

Chapter 2. Defining the system that fits your needs

This chapter helps you define a new Cluster 1600 system managed by PSSP (SP system) that meets your hardware and software computing needs. You will be asked to answer many questions about the type of system you want and you will be instructed to complete a set of worksheets as you progress through the questions.

Decision making is an iterative and recursive process. Therefore, you might find yourself modifying answers to questions you previously answered. Reviewing and modifying your plans is a necessary part of a thorough planning process. The output of this exercise should be a completed layout of your system hardware and software that will help you to prepare for your installation.

Contact your network administrator.
Connecting your Cluster 1600 system managed by PSSP to a network has important benefits because networking information is critical to the success of the system installation. Network planning is a complex but necessary part of a thorough planning process. It is important to consider networking information early in the process so do not delay contacting your network administrator.

As you plan your system you will make many decisions. The remainder of this chapter poses several questions for you to answer. Review these questions and become familiar with the type of information you need to gather throughout the planning process.

This chapter contains sample worksheets for a hypothetical corporation called the `ABC Corporation`. Review these sample worksheets and use them as a guide to see the decisions you need to make and how to complete them. Decisions made by the `ABC Corporation` are shown in `constant width` font to help distinguish them from the fixed text in the worksheet.

Your decisions will most likely be different from those of the `ABC Corporation`. This is natural since every company is different and the decisions you make must meet the needs of your organization.

As you go through these questions, fill in your copies of the worksheets in Appendix C, “SP system planning worksheets” on page 281 with the information about your system. First, make several copies of all the worksheets. You might change your mind as you go through the worksheets. You will need the original blank copies plus your completed worksheets in the future when you want to add to or otherwise change your system.

Your completed worksheets serve primarily to prepare you with much of the information you need to have during installation and configuration of your system frames, nodes, and switches, as well as for installation or migration and configuration of the PSSP software. The book *PSSP: Installation and Migration Guide* instructs you through that process.

Your completed worksheets can also serve to document your system hardware, software, and network configuration. You might consider some of the information to be confidential. Take care to handle completed worksheets according to the security policy established in your organization.

Question 1: Why do you need a Cluster 1600 system managed by PSSP?

Why do you need a Cluster 1600 system managed by PSSP? For LAN consolidation? For data mining? For engineering or scientific computing? See Figure 4 for some typical uses.

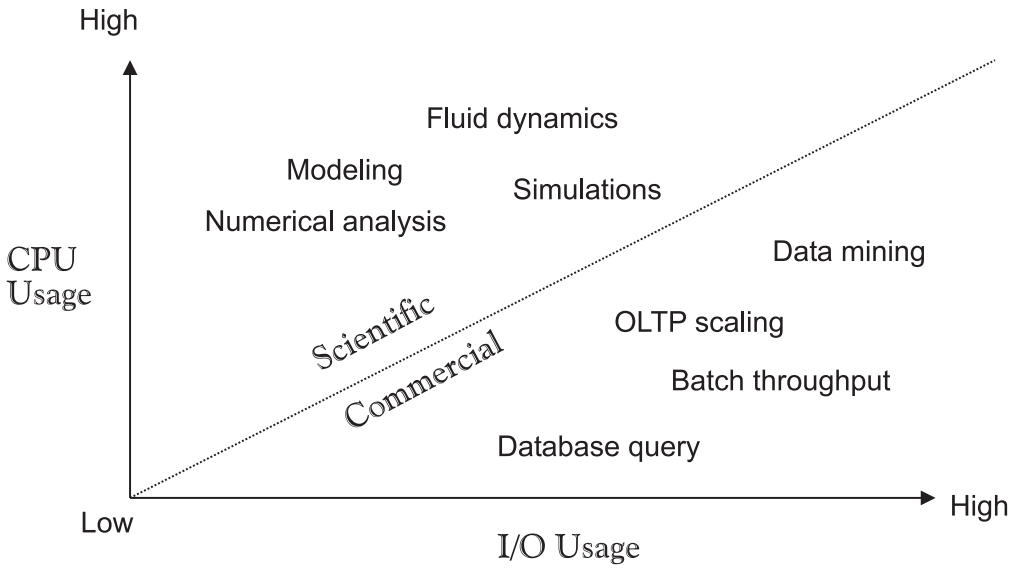


Figure 4. Typical uses of a Cluster 1600 system managed by PSSP (not to scale)

Which applications do you want to run on this system? Each new node is a powerful SMP that can run the AIX 5L 5.1 operating system and PSSP 3.5, and perhaps other Cluster 1600 software building blocks that you choose to install, with a single point of control for system management. Thousands of IBM RS/6000 and @server pSeries applications can run unchanged on a Cluster 1600 system managed by PSSP 3.5. Do you need the additional power of parallel computing? Do you need the high-performance communication within your system that an SP switch can provide?

Considering parallel computing

Along with this question, you need to decide whether you want to reap the benefits that parallel computing offers by running parallel applications. Parallel computing involves breaking a serial application into its logical parts and running those parts simultaneously. As a result, you can solve large, complex problems quickly.

Parallel applications can be broadly classified into two classes by considering whether the parallelism can be achieved through the use of a *middleware* software layer or whether the application developer needs to explicitly parallelize the problem by working with the source code and adding directives and code to achieve speedup.

Examples of a parallelized software layer are a parallel relational database such as DB2 Parallel Edition, or the Parallel Engineering and Scientific Subroutine Library (Parallel ESSL), which lets you execute an SQL statement, or call a matrix multiplication routine, and achieve problem speedup without having to specify how to achieve the parallelism.

An example of explicit parallelism is taking an existing serial Fortran program and adding calls to a message passing library to distribute the computations among the PSSP nodes. In this case, various parallel tools such as compilers, libraries, and debuggers are required.

Choosing a switch

Consider whether you can benefit from using a switch to interconnect the nodes.

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

- Interframe connectivity and communication
- Scalability up to 128 node connections, 512 by special bid, including switch board frames
- Constant bandwidth and latency between node pairs
- Support for Internet Protocol (IP) communication between nodes
- IP Address Resolution Protocol (ARP) support
- Support for dedicated or multi-user environments
- Error detection and retry
- High availability
- Fault isolation
- Concurrent maintenance for nodes
- Improved switch chip bandwidth

Your choices for a switch supported with PSSP 3.5 are the SP Switch2 and the SP Switch. Table 1 on page 22 describes the PSSP 3.5 support of these switches.

Table 1. The switches supported by PSSP 3.5

Switch feature	Description
SP Switch2	The SP Switch2 (feature code 4012) has 16 ports for node connections and 16 ports for switch to switch connections. The SP Switch2 can interconnect all the processor nodes that are currently available from IBM. Some older nodes that you might already have are also supported on this switch. See “Question 8: Which and how many nodes do you need?” on page 43 for information about them. There is optional switch connectivity. This means you can connect some nodes to the SP Switch2 and leave other nodes off the switch. For instance, the RS/6000 Enterprise Server S70 and S7A are not supported with the SP Switch2. If you have any, they can still be in your system running PSSP 3.5, but not connected to the SP Switch2. The SP Switch2 supports two different switch configurations so that nodes can communicate over one or two switch planes. You can have improved communication performance with two switch planes operational, and higher availability since one switch plane can continue operating even when you take a switch down for maintenance. A two-plane SP Switch2 system has two sets of switches and two switch adapters per node. The switch planes are disjoint – each is cabled exactly like a single plane, both with the same topology, and communication across the pair of planes is achieved by a data striping algorithm in the software. You can only use this option with nodes that can have two adapters – one for connecting to one switch plane and another for connecting to the other switch plane. You cannot install one adapter in some nodes and two adapters in others. The SP Switch2 supports multiple node switch boards in one frame. You can have up to eight SP Switch2 node switch boards in a single SP frame. This is advantageous for installations that require a large number of switch connections for SP-attached servers or clustered server configurations. This is a frame configured only with SP Switch2 switches. The SP Switch2 can be configured with one switch plane. With one adapter per node attached to the one switch plane network it is like the SP Switch. But it also has the added functional benefits, like relaxed node placement rules and optional connectivity. The PSSP software supports the removable and hot-pluggable interposer cards that provide the bulkhead connections to the switch cable, the concurrent replacement of any failed power supplies or fans, and replacement of the supervisor while the switch is operating. The SP Switch2 does not support SP system partitioning or the SP Switch Router.
SP Switch (16-port)	The SP Switch (feature code 4011) has 16 ports for node connections and 16 ports for switch to switch connections. It interconnects all the currently supported nodes and the SP Switch Router. When you use the SP Switch, all the nodes in the system must be connected to it.
SP Switch (8-port)	This is a lower-cost version of the SP Switch (feature code 4008) that offers 8 internal connections to provide enhanced functions for small systems with up to 8 total nodes. It can interconnect only thin and wide SP rack-mounted nodes, not high nodes. It does not support SP-attached servers or scaling to larger systems.

Switch incompatibility statement:
Switch types cannot be mixed in a Cluster 1600 system managed by PSSP – not even in separate SP system partitions. If you want to use the SP Switch2 but you still have some nodes that are not supported on the SP Switch2, you can leave them off the switch.

Hardware planning is described in Volume 1.
This book covers switch planning only in the context of system configuration. Adapters are used to connect a node to the switch subsystem. For physical planning regarding switch adapters, wiring, and cabling see the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*.

For the SP Switch there are frame, switch, and node placement and numbering rules and switch port numbering rules to be understood and honored for the supported switch capsule configurations. The SP Switch2 has some system configuration rules, but node placement rules have been relaxed. Nodes can be placed anywhere allowed by the physical constraints. The node switch port numbers are not predefined. You can connect a node to any available switch port and the numbers are generated when the switch is started. See “Understanding placement and numbering” on page 118.

Question 2: Do you want preloaded software or the default order?

You have the option of ordering your Cluster 1600 system managed by PSSP with default software that you must load or you can order one that IBM has preloaded with software to meet your organization’s specific needs. There are manufacturing-based services available from which you can choose when you decide to order an SP.

Default order service

By default, every Cluster 1600 system managed by PSSP with an SP hardware building block is delivered with the newest software level, including the latest Program Temporary Fixes (PTFs), of AIX 5L 5.1 and PSSP 3.5. You receive a 32-bit kernel and a 64-bit kernel mksysb image on installation media for the control workstation and nodes. You also receive a node image on tape for backup purposes. The mksysb for the control workstation might also have a LoadLeveler install image. These are delivered with your hardware.

The versions installed are determined by one of the following feature codes:

- FC #9535: AIX 5L 5.1 and PSSP 3.5
- FC #9534: AIX 5L 5.1 and PSSP 3.4
- FC #9435: AIX 4.3.3 and PSSP 3.4
- FC #9433: AIX 4.3.3 and PSSP 3.2

Note: If you do not specify a feature code, IBM provides AIX 5L 5.1 and PSSP 3.5. ***This book applies only to planning for PSSP 3.5 with AIX 5L 5.1.*** If you order an older version of PSSP, be sure to use the books that correlate to your order.

With the default order service, you do the loading and customizing work including network configuration and the installation of any additional licensed programs. You do the complete installation of **all** the mksysb images you choose to install for **both** the control workstation **and** the nodes.

Customizing services

Customizing services based on cost-recovery charges are available. The Customized Solutions Group has a team of professionals skilled in the following services:

- Pre-install review of the control workstation, frame, network, and node configurations.
- Design review focused on how to implement application suites on specific hardware configurations.
- High Availability Control Workstation (HACWS) customization.
- High Availability Cluster Multi-Processing (HACMP) customization and integration. This includes a complete integration from pre-sales consulting and design sessions through on-site integration.
- Oracle Parallel Query installation, customization, and integration.
- Lotus Notes server integration and implementation of Lotus Notes with HACMP.
- Implementation of DB/2 Parallel Edition.

What you get as a package

Backup tapes, installation media, and instructions that simplify installation are shipped with your system. They are in clear plastic inside the wooden shipping container. The contents and the instructions vary depending on your order:

- With the default order service, the nodes are not preloaded. You receive installation media that contains the PSSP and AIX 5L software with instructions for you to load the workstation and the nodes. You receive a 32-bit kernel and a 64-bit kernel mksysb image on installation media for the control workstation and nodes. Use one of the mksysb backups for the control workstation and nodes. You also receive a node image on tape for backup purposes.
- With customized service, you can have the nodes preloaded with the software and networking parameters you order or you can have onsite service. In any case, you receive a customized mksysb for the control workstation and the instructions to load the workstation. You also receive a node install image on tape for backup purposes.

If you plan to further modify or customize your nodes, also read the books *PSSP: Installation and Migration Guide*, *PSSP: Managing Shared Disks*, and *PSSP: Administration Guide*.

Contacting the Customized Solution Organization

If you have questions about these preload services, please contact the World Wide Customized Solution Organization. You can call or send a note. In any messages you send or leave on the phone include your name, phone number, user ID, and order number. A member of the Customized Solution Department will then work with you to help make your installation a success.

To contact the Customized Solution Organization:

- from the US, do any of the following:
  - call 1-800-426-4955 and select option 2

  - have your IBM representative send a note to JUST ASK/Poughkeepsie/IBM
- from EMEA, do any of the following:
  - call 33-4-6734-4679
  - have your IBM representative send a note to VM user ID SP2LOAD at MOPVMA
  - have your IBM representative request instruction files by typing on the command line of a VM session:

    ```
    REQUEST SP2INST FROM SP2INST AT MOPVMA
    ```
- from any other location, call your IBM representative.

Listing your applications

Though you've only just begun to consider the software, begin a list of the applications you want on your system. If you already know that you want the services of the Customized Solution Organization, contact them. Whether or not you use the custom service, use the worksheets in this book as you progress in the planning process.

List the applications you are considering on your copy of Worksheet 1, "Preliminary list of applications" in Table 60 on page 282. Indicate that you want parallel processing or that you need a switch for better application performance by marking the **Parallel** and **Need Switch** columns respectively. Use a "?" if you are not sure yet. If you think of additional applications, you can add them to this list at any time.

The hypothetical customer, the `ABC Corporation`, considered their application requirements and filled in Table 2.

Table 2. Preliminary list of applications for the ABC Corporation

Preliminary list of applications – Worksheet 1		
Application	**Parallel**	**Need switch?**
`DB2 Parallel Edition`	y	y
`DCE for AIX`		
`Customer-written application`	y	y

Save your list. You'll use it again in the planning process.

Question 3: Which related IBM licensed programs do you need?

There are two sets of IBM licensed programs from which you must know what to order. One set includes programs that are part of the PSSP environment and the other includes programs that are for the base operating system, like C and C++ compilers, to run on each node.

An IBM C or C++ compiler is required for the PSSP software to use. The compiler is necessary for service of the PSSP software. Also, without the compiler

preprocessor, dump diagnosis tools like **kdb** will not work effectively. You need at least one concurrent use license for the C or C++ compilers of VisualAge C++ Professional for AIX, Version 6.0 or later.

Note: PSSP does not support the incremental compiler and runtime libraries. It only supports the batch C and C++ compilers and runtime libraries that are included in this VisualAge package.

If you intend to do C application development work, you need to decide how many developers you want to support at a given time and acquire enough licenses for them. You might also want to develop 64-bit applications for AIX 5L 5.1 with:

- The compiler XL Fortran 7.1.0.2 or later
- The debugger TotalView 4.1 or later

PSSP is comprised of software components that are an integral part of your system. It is the software that makes a collection of RS/6000 and @server pSeries nodes into a Cluster 1600 system managed by PSSP (SP system) into a virtual SP system. PSSP helps a system administrator manage the SP system. It provides a single point of control for administrative tasks and helps increase productivity by letting administrators view, monitor, and control system operation.

Some components of PSSP are optional. You will receive all of the PSSP software but you can choose whether or not to install and use them. For instance, you might want to consider the following optional components when planning your SP system:

- High Availability Control Workstation (HACWS)

  See Chapter 4, “Planning for a high availability control workstation” on page 129.
- IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk

  See Chapter 5, “Planning for IBM Virtual Shared Disks” on page 141.

In addition, there are many licensed programs available to run on your SP which can add to the productivity of your enterprise. You can see a list on the Internet at the Web site:

`http://www.ibm.com/servers/aix/products/ibmsw/list/`

Some licensed programs in the SP software suite are particularly closely related to the PSSP software. Each of those programs is briefly described here. If you think one of the following programs might provide a service you want on your SP system, see Chapter 9, “Planning for PSSP-related licensed programs” on page 195 for planning information. If you decide to seriously consider ordering any of them for an existing system, be sure to carefully read Chapter 11, “Planning for migration” on page 213 for possible release level dependencies. At the end of this section, you will find a worksheet where you can list the licensed programs you want.

General Parallel File System

IBM General Parallel File System for AIX (GPFS) provides concurrent shared access to files spanning multiple disk drives located on multiple nodes. This provides file system service to parallel and serial applications on the SP system.

Using GPFS to store and retrieve your files can improve your system by:

- Allowing multiple processes or applications on all nodes in a GPFS nodeset simultaneous access to the same file using standard file system calls.
- Increasing aggregate bandwidth of your file system by spreading reads and writes across multiple disks.

- Balancing the load evenly across all disks to maximize their combined throughput.
- Supporting large amounts of data.
- Allowing concurrent reads and writes, which is important in parallel processing.
- Assuring data consistency by a sophisticated token management system.
- Simplifying administration through simple, multiple node file system commands that function across a GPFS nodeset.

If you might want to use GPFS, see “Planning for General Parallel File System (GPFS)” on page 198.

High Availability Cluster Multi-Processing

IBM’s tool for building UNIX-based mission-critical computing platforms is the IBM High Availability Cluster Multi-Processing for AIX (HACMP) software package. HACMP ensures that critical resources are available for processing. HACMP has several features which you can choose to use independently. You can run HACMP on the Cluster 1600 system managed by PSSP (SP system) with or without:

- The Enhanced Scalability feature (HACMP/ES)

  HACMP/ES builds on the Event Management and Group Services components of RSCT to scale HACMP function.
- IBM High Availability Geographic Cluster (HAGEO)

  HAGEO provides real-time mirroring of customer data between systems connected by local or point-to-point networks, bringing disaster recovery capability to a cluster of IBM @server pSeries or RS/6000 nodes placed in two widely separated geographic locations. HAGEO automatically responds to site and communication failures and provides for automatic site takeover. Tools are available for data synchronization after an outage, configuration management, capacity planning, performance monitoring, and problem determination.
- IBM Geographic Remote Mirror (GeoRM)

  The geographic remote mirroring capability is available alone, as a separate feature without the failure detection and automatic takeover capability of HAGEO. GeoRM allows customer data to be mirrored in real time between geographically dispersed locations using LANs or point-to-point networks, with no limitation on the distance between locations.

Typically, HACMP or HACMP/ES is run on the control workstation only if HACWS is being used. HACMP/ES is run on the nodes. If you might want to use this licensed program, see “Planning for High Availability Cluster Multi-Processing (HACMP)” on page 196.

LoadLeveler

LoadLeveler is an IBM software product that provides workload management of both interactive and batch processing on a Cluster 1600 system managed by PSSP (SP system), or pSeries and RS/6000 workstation. The LoadLeveler software lets you build, submit, and process both serial and parallel jobs. It allows multiple user space tasks per node, enabling applications that use the message passing interface (MPI) protocol to exploit parallel processing and realize significant performance improvements. Applications using the latest levels of the LAPI User Space API or Parallel Environment MPI and LoadLeveler software, can start up to four user space tasks per node with the SP Switch and up to 64 with the SP Switch2, depending on the switch configuration. These tasks can be from the same or from different parallel jobs. This allows parallel applications to exploit the symmetric

multiprocessors on nodes without restructuring or recompiling. A user space MPI or LAPI job can consist of up to 1024 tasks on SP Switch systems and up to 4096 tasks on SP Switch2 system (up to 2048 tasks for the MPI/IP library).

LoadLeveler is an integral piece of a total System Management solution. LoadLeveler can take advantage of PSSP features like event management, performance monitoring, and SP switch management. LoadLeveler can also interoperate with other schedulers to support batch job processing on other hardware platforms. These schedulers can include Network Queueing System (NQS) and the IBM Network Queuing System/MVS (NQS/MVS).

LoadLeveler can be included with your new Cluster 1600 system managed by PSSP (SP system) order. You choose whether to use it or not. If you might want to use this licensed program, see “Planning for LoadLeveler” on page 197.

Parallel Environment

The IBM Parallel Environment for AIX licensed program provides support for parallel application development on a Cluster 1600 system managed by PSSP (SP system), on a single RS/6000 processor, or a TCP/IP-networked cluster of IBM RS/6000 or @server pSeries processors. The Parallel Environment licensed program contains tools to support the development and analysis of parallel applications written in Fortran, C, or C++. It also provides a user-friendly runtime environment for their execution. Parallel Environment supports the Message Passing Library (MPL) subroutines, the Message Passing Interface (MPI) standard, Parallel ESSL, and the Communications Low-Level Applications Programming Interface (LAPI). LAPI is designed to provide optimal communication performance on the SP Switch2 or the SP Switch.

The main Parallel Environment components are:

- **Parallel Operating Environment**, which provides the ability to create and execute parallel application programs.
- **Message Passing and Collective Communications Application Programming Interface** and **Message Passing Subroutine Library**, which help application developers parallelize their code.
- **Parallel Utilities**:
  - PE Benchmarker Performance Collection Tool (ppe.prf)

    The Performance Collection Tool enables you to collect either MPI and user event data or hardware and operating system profiles for one or more application processes or *tasks*. This tool is built on dynamic instrumentation technology, the Dynamic Probe Class Library (DPCL). When you collect MPI and user event traces using the Performance Collection Tool, the collected information is saved as a standard AIX event trace file on each machine running instrumented processes. The UTE utilities enable you to convert one or more of these AIX trace files into UTE interval files. The UTE utilities, installed in the directory /usr/lpp/ppe.perf/bin as part of the ppe.perf fileset, consist of:
    - The convert utility, which converts AIX event trace records into UTE interval trace files.
    - The utemerge utility, which merges multiple UTE interval files into a single UTE interval file.
    - The utestats utility, which generates statistics tables from UTE interval files.

- The slogmerge utility, which converts and merges UTE interval files into a single SLOG file for analysis within Argonne National Laboratory's Jumpshot tool.

When you collect hardware and operating system profiles using the Performance Collection Tool, the collected profile information is saved as NetCDF (Network Common Data Form) files on each machine running instrumented processes. The Profile Visualization Tool can read NetCDF files and summarize the profile information in reports.

– PE Benchmarker Profile Visualization Tool (ppe.pvt)

  This is a post-mortem analysis tool designed to process profile data files generated by the Performance Collection Tool. The profile data includes function call count, wall clock timer, system resource usage, and hardware performance counter events. You can run this tool using the PVT graphical user interface or command. Results can be viewed, placed in report files, or saved to a summary file.

– Parallel Debugger (PDBX)

  Parallel Debugger maintains the same look and feel as the common DBX debugger available on AIX and UNIX operating systems. This debugger extends the DBX debugger commands and subcommands. Some have been modified for use on parallel programs. The PDBX debugger is a POE application with some modifications on the home node to provide a user interface.

  Two instances of PDBX do not interoperate. PDBX is run on the home node, it communicates with a partition manager daemon on each remote node which communicates with a set of dbx processes (one dbx per task) where each instance of dbx communicates with an instance of the program (a.out).

– Xprofiler

  Xprofiler is a stand-alone application to analyze application performance – a graphical front-end to *prof* and *gprof*.

DPCL is a C++ class library that enables tool developers to build highly scalable platform independent tools. It is based on a technology that allows dynamic instrumentation of applications with software patches (Probes) for the purpose of gathering data from the application during execution. It was specifically created to tune applications for optimal performance but is applicable to a variety of end user tool requirements including logic debugging, memory debugging, software diagnostics, test coverage, trace gathering, and profiling. DPCL is available only as open source. It is packaged with PE, but supported in open source.

If you might want to use the Parallel Environment, see "Planning for Parallel Environment" on page 195.

Parallel Engineering and Scientific Subroutine Library

The IBM Engineering and Scientific Subroutine Library for AIX (ESSL) family of products is a state-of-the-art collection of mathematical subroutines. Running on IBM @server pSeries and RS/6000 nodes, the ESSL family provides a wide range of high-performance mathematical functions for a variety of scientific and engineering applications:

- ESSL contains over 400 high-performance mathematical subroutines tuned for IBM UNIX hardware.
- Parallel ESSL contains over 100 high-performance mathematical subroutines specifically designed to exploit the full power of parallel processing.

Parallel ESSL subroutines make it easier for developers, especially those not proficient in advanced parallel processing techniques, to create or convert applications to take advantage of the parallel processors of a Cluster 1600 system managed by PSSP (SP system).

Parallel ESSL accelerates applications by substituting comparable math subroutines and in-line code with high performance, highly-tuned subroutines. They are tuned for optimal performance on a system with the SP Switch2 or the SP Switch (16-port or 8-port). Both new and current numerically intensive applications written in Fortran, C, or C++ can call Parallel ESSL subroutines. New applications can be designed to take advantage of complete Parallel ESSL capabilities. Existing applications can be enabled by replacing comparable routines and in-line code with calls to Parallel ESSL subroutines. Parallel ESSL supports the Single Program Multiple Data programming model and provides subroutines in six major areas of mathematical computations:

- Level 2 PBLAS
- Level 3 PBLAS
- Linear Algebraic Equations
- Eigensystem Analysis and Singular Value Analysis
- Fourier Transforms
- Random Number Generation

The design of Parallel ESSL centers on exploiting operational characteristics and the parallel processing architecture of a a Cluster 1600 system managed by PSSP (SP system). There are performance advantages when run on a system with the SP Switch2 or the SP Switch.

If you might want to use Parallel ESSL, see “Planning for Parallel ESSL” on page 196.

Selecting IBM licensed programs

Our hypothetical customer, the `ABC Corporation`, chose to order the programs marked in Table 3 on page 31 to run with PSSP 3.5. You can mark your copy of “IBM licensed programs to order – Worksheet 2” from Table 61 on page 283.

Table 3. IBM licensed programs to order for ABC Corporation

IBM licensed programs to order – Worksheet 2		
Order		**Level**
x	IBM VisualAge C++ Professional (batch C and C++)	6.0 or later
x	IBM DCE	3.2
x	IBM DCE Base Services (6693, 41L2819) and Servers (6688, 41L2813)	3.2
x	IBM Parallel System Support Programs (PSSP)	3.5
		3.4
		3.2
x	IBM General Parallel File System	2.1
		1.5
		1.4
	IBM High Availability Cluster Multi-Processing (HACMP features HAS, CRM, ES, ESCRM)	4.5
	IBM HACMP features HAGEO or GeoRM	2.4
x	IBM LoadLeveler	3.1
		2.2
x	IBM Parallel Environment	3.2
		3.1
x	IBM Engineering and Scientific Subroutine Library (ESSL)	3.3
		3.2
x	IBM Parallel ESSL	2.3
		2.2

Question 4: Which levels of AIX do you need?

New Cluster 1600 systems managed by PSSP (SP systems) come with AIX 5L 5.1 and PSSP 3.5 on installation media. Before you decide, you need to consider which existing and new hardware and software features you need and what they require. For instance, do you want PSSP services that support 64-bit addressing and themselves use the 64-bit kernel? If you are not planning an entirely new system but are adding to an existing one, you especially need to consider the current migration and coexistence support to best understand the requirements.

This section addresses the topics:

Considering AIX and PSSP in another language

The software is NLS-enabled and ready for *Internationalization*. This means that the software has been made translation-ready and can be translated to a language that is supported by AIX. It is not already translated. Contact your IBM representative if you choose to consider translating it. ***This discussion is to inform you of the possibilities.***

All of the PSSP software components and the following PSSP-related licensed programs have been enabled with National Language Support (NLS):

- High Availability Cluster Multi-Processing
- LoadLeveler

The SP software is configured by default to operate in the US English language locale. Before you decide to translate the software and run AIX and PSSP in another language, you need to consider the effects of using different language locales. If you decide to use another language locale, after you get the software translated, do the following:

- Set the AIX language locale.
- Set the SP administrative locale.

Effects of using different language locales

For AIX, the installation default language locale refers to the locale selected during AIX system installation as the system-wide locale. On an SP system, that AIX process still applies on each node. On an SP system, you can install AIX on the control workstation in any supported language, followed by network setup, the PSSP software installation, SDR configuration, activation of the setup server, then node install of AIX based on mksysb images provided by IBM or generated by you, followed by node install of PSSP based on mksysb images. The PSSP 3.2 software supports a heterogeneous (multilingual) system configuration. That means you can operate with the system locale on some nodes being different from the administrative locale on the control workstation. The SP will function correctly in such a case, but output from commands or processes that span multiple nodes might be received in mixed languages and might be unreadable. Therefore, the nodes and control workstation can run with different locales but with the following restrictions:

1. When installing, configuring, and customizing a multilingual SP system, any system-wide control and configuration data (data written to the SDR) must be written in the SP administrative locale or in the US English locale. If, for instance, a node is running in a locale different from the administrative locale, any system-wide control or configuration data written to the SDR on behalf of that node must be restricted to the portion of the locale code set that is common across all locales: the standard 7–bit ASCII character set.
2. If the SP administrative locale is a multibyte character set (MBCS) locale, system-wide control and configuration data can be stored in the SDR using multibyte character data (MBCD) specific to the SP administrative locale. However, IBM strongly suggests that in a heterogeneous environment all system-wide control and configuration data be stored using the standard 7–bit ASCII character set because it is available in all code sets and locales. That assures that data written to the SDR can be displayed and is readable from the SP administrative locale. It also reduces the probability of MBCD from several different locales being written into the SDR in the event that the SP administrative locale is changed.
3. If the SP administrative locale is an MBCS locale and system-wide control and configuration MBCD is stored into the SDR, the system will not be able to use the IBM domain name server (DNS) code. The general DNS standard for host names forbids any characters not in the set [A-Z, a-z, 0–9, -]. The upper and lower case English letters, the Arabic numerals, the hyphen, and for HACMP only the underscore character (_), are included in the set. You can use a different name resolution mechanism that is MBCS-enabled.
4. If the SP system is configured as a multilingual heterogeneous system, any distributed application or client-server application that spans several nodes

which might be running different locales can receive responses in the node's default installation locale or in the SP administrative locale. The response will be in the language in which the node is operating and might not be readable. The application can compensate by imposing a homogeneous restriction, ensuring that daemons are initiated in the same locale or in the SP administrative locale. It can compensate by providing an API to transport the locale information from the client to the server and daemons. If the serviceability and functioning of the application is not impacted, the application might handle different language responses from nodes.

5. Administration of several languages from a single point of control and ensuring that all information can be displayed and is readable across different code pages is available only when using Unicode. Unicode only works if the AIX system platform across the SP system is declared to be Unicode only.

Setting the AIX language locale

For AIX, the installation default language locale refers to the locale selected during AIX system installation as the system-wide locale. On an SP system, that AIX process still applies on each node.

This locale is defined in the **LANG** environment variable in the **/etc/environment** file. The **/usr/lib/nls/loc** directory contains the system-wide locale set during installation of AIX. These two directories are defined in the **/etc/environment** file under the environment variables **LOCPATH** and **NLSPATH**, respectively.

To change the locale, you can use the AIX **chlang** or **export LANG=***locale* command, where *locale* is the AIX-supported locale you choose and, preferably, the one to which you translated the PSSP and related licensed programs, and with which you want PSSP to operate.

Changes made to the NLS environment are not immediate. Changes made to the **/etc/environment** file requires rebooting the system. Changes made in the user's **.profile** file requires executing **.profile** or logging in to AIX again.

The AIX **locale** command lists all local environment variables with which you can set up a locale category preference. An application can impose the values of locale categories, but then it is not considered to be a code set independent application.

Setting the SP administrative locale

The purpose of the SP administrative locale includes the following:

- To determine the language to be used for data written to the SDR in a multilingual configuration.
- To use a default locale when installing nodes.
- To be used by applications that must always run in the same locale.
- To be used by client-server applications that must run in a single locale.

The default is for all nodes to operate in the administrative locale. In this locale, the SP system administrator experiences a consistent view of the SP system control and external information.

The administrative locale is set by the **install_cw** process during installation and can be changed at any time using the **spsitenv** command.

Considering new features

You should plan to order the basic operating system and levels that support the functions and hardware you need. You might need to stay at earlier levels of AIX to

support specific hardware or software you already have installed. On the other hand, you might need the newest software to handle your capacity and performance requirements.

Before deciding, consider the new features available in this new release. See "What's new in AIX and PSSP?" on page 13 for brief descriptions of new features.

Considering migration and coexistence

Migration addresses upgrading AIX, PSSP, and PSSP-related licensed programs on an existing SP system or a cluster of IBM @server pSeries or RS/6000 nodes from earlier supported levels to the new levels. Coexistence refers to the ability of a program to support multiple levels of AIX, PSSP, and itself in the same system partition. Coexistence is important in the ability to migrate one node at a time and is a key component of migration. However, some licensed programs support coexistence only to facilitate migration and function is not available until all nodes have been migrated to the new level.

A direct migration path is supported only from a control workstation with PSSP 3.2 or PSSP 3.4 to PSSP 3.5 and AIX 5L 5.1. The migration from PSSP 3.2 and AIX 4.3.3 must be done in one service window. A control workstation with earlier releases must first be migrated to PSSP 3.4, then to AIX 5L 5.1, then to PSSP 3.5. Nodes can migrate to PSSP 3.5 and AIX 5L 5.1 in one step. Table 4 lists the migration paths to PSSP 3.5 that are available.

Table 4. Migration paths to PSSP 3.5 for nodes

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

You need to migrate nodes with AIX 4.3.3 and older versions of PSSP first to PSSP 3.4 or 3.2 and later use one of the listed migration paths.

In general, coexistence is supported in the same SP system partition or a single default system partition (the entire Cluster 1600 system managed by PSSP) for nodes running any combination of:

- PSSP 3.5 and AIX 5L 5.1.
- PSSP 3.4 and AIX 5L 5.1 or AIX 4.3.3
- PSSP 3.2 and AIX 4.3.3

For exceptions and more details, see Chapter 11, "Planning for migration" on page 213.

Recording your decision for question 4

Now you should know which level of AIX you need. To run PSSP 3.5, you must have AIX 5L 5.1 running on the control workstation. If you need to run more than one level of AIX on nodes, you might require SP system partitions. Planning for SP system partitions is discussed in Chapter 7, "Planning SP system partitions" on page 171. Without multiple SP system partitions, you must rely on the coexistence support. Study Chapter 11, "Planning for migration" on page 213 before you decide.

The ABC Corporation's worksheet appears in Table 5.

Table 5. Operating system level selected by the ABC Corporation

Need to run	AIX	PSSP
x	AIX 5L 5.1	PSSP 3.5
	AIX 5L 5.1	PSSP 3.4
	AIX 4.3.3	PSSP 3.4
	AIX 4.3.3	PSSP 3.2

Question 5: What type of network connectivity do you need?

In order to answer this question, you need to consult with your network administrator to decide how this system will connect into your existing computing network. Also consult with your hardware planner regarding which adapters work with the node types you are considering (see the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*).

You might not have determined your system requirements quite enough yet to be able to fully answer all the questions you need to consider. But do start to think about them and know that you will have to come back to them and fill in all the network information when your physical layout plan is complete. Also, "Planning your network configuration" on page 110 will help you understand the SP network capabilities.

Your answers for question 5 are applied in the collection of worksheets for "Question 8: Which and how many nodes do you need?" on page 43.

Questions to consider

Review the following questions to determine the type of network connectivity your organization needs:

- Do you currently have a TCP/IP network?
- Do you intend to connect that network to the SP Ethernet admin LAN and the control workstation? Only IPv4 networks are supported on the SP system. Some PSSP components tolerate IPv6 aliases for IPv4 network addresses but not with DCE, HACMP, HACWS, or an SP switch. For information about the SP system tolerating IPv6 aliases for IPv4 network addresses, see the appendix on the subject in the book *PSSP: Administration Guide*.
- What type of physical Ethernet LAN do you require? The PSSP software supports twisted pair (TP), thin Ethernet (BNC), and thick Ethernet (DIX) for the SP Ethernet admin LAN.
- Do you have a TCP/IP address range? What is the address range and the subnet scheme? What subnet masks are employed? Will the address range be sufficient to cover the addresses of the SP nodes and control workstations? Remember to consider the future. For instance, you might start with a 10-node system but plan to grow to 100 nodes. Be sure to define an address range that is broad enough to accommodate your future needs.
- Do you have a domain name? If so, what is it? How are IP addresses resolved to names and vice versa? Do you have DNS, NIS, or **/etc/hosts**? How are your domain name servers configured?

- What is the topology of your TCP/IP network? Where do you intend to connect the IBM @server pSeries or RS/6000 workstations and the nodes into the network?

For the SP system and the control workstations, also consider the following:

- A Cluster 1600 system managed by PSSP with an SP switch has a minimum of two networks, the PSSP admin LAN that connects each node to the control workstation and the switch network. With an SP Switch2, you can have two planes in the switch network. The two networks must each be assigned unique TCP/IP network addresses. The switch subsystem provides an optional aggregate IP device abstraction for the SP Switch2 networks. This pseudo-device driver enables you to address the two SP Switch2 planes with one IP address. If you have a switch, you must plan for the switch network.
- In addition to the PSSP admin LAN and the switch network, additional communications adapters like ATM or FDDI are often installed in the nodes. In that case, separate TCP/IP network addresses need to be assigned. Have such networks been considered?
- What domain name will be assigned to your Cluster 1600 system managed by PSSP?
- What IP networks, addresses, subnet masks and default gateways will be assigned to the SP networks?
- Do you want a firewalled Cluster 1600 system managed by PSSP? See the document on this subject at the Internet address given in “Finding documentation on the World Wide Web” on page 313.
- Will nodes be configured as primary and secondary name servers?
- What SP-attached or clustered servers considerations are there? If you plan to use a server that is otherwise already in use and connected to an IPv6 network, you must remove it from the IPv6 network before using it with the PSSP software. For more information about servers, see page 45.

Considering the SP Switch router

To provide higher bandwidth communications, an SP Switch Router can be connected to the SP Switch by the SP Switch Router Adapter. The SP Switch Router Adapter provides a high performance, full duplex interface between the SP Switch and the SP Switch Router.

Note: The SP Switch Router is not supported on the SP Switch2.

When the SP Switch Router Adapter is installed in an SP Switch Router, it allows the SP Switch Router to be used as a networking gateway for the SP system. The SP Switch Router can be populated with additional adapters for standard network interfaces, including the types:

- Ethernet
- FDDI
- ATM
- SONET
- HIPPI
- HSSI

More than one SP Switch Router Adapter can be installed in an SP Switch Router. These SP Switch Routers can be connected to the same SP system, system partition, or to other SP systems. When multiple SP Switch Router Adapters are

installed and connected to more than one SP system or system partition, they can be used to provide a high-bandwidth link between SP systems or system partitions and to provide the SP systems or system partitions with a shared set of interfaces to external networks.

Each SP Switch Router Adapter requires one available node switch port on the SP Switch that meets the criteria for valid extension node ports as described in Chapter 3, “Defining the configuration that fits your needs” on page 89. A 10 meter SP Switch cable and a 10 meter ground strap are provided (other lengths are available) for connecting the SP Switch Router Adapter, located in the SP Switch Router chassis, to the SP Switch.

Question 6: What are your disk storage requirements?

Consult with your system administrator to answer this question. You need to understand your existing environment to be able to project your future disk requirements. In planning how much disk space you need, be aware of considerations that relate to internal and external disk storage.

Address the considerations in this section and record your answers. Later you will enter them in your copy of the worksheet “Hardware configuration by frame” in Table 64 on page 286.

Disk space for user home directories

You need to decide whether you will serve your user home directories from an existing server or from a new server.

Disk space for system programs

Installing AIX and some subset of PSSP and related programs consumes disk storage on each node. Use the tables in “Determining install space requirements” on page 98 to calculate the disk storage needed for AIX and PSSP and the related programs.

Think about the licensed programs and applications you plan to install. For instance, do you plan to install DCE security services? How much space do you need for **/usr**, **/**, and other file systems? For the **/spdata** file system, you will need extra space on the control workstation and boot-install servers if you maintain more than one level of AIX or PSSP on the system.

Will your applications be installed in one root volume group (rootvg) with the base AIX programs or will they be installed elsewhere? Will you want to use alternate root volume groups and mirrored root volume groups? Maybe you plan to boot from external disks. These considerations might help you decide whether to add additional internal or external disks. Adding additional disks gives you the flexibility to preserve the application installation in the event that a node requires a reinstallation or a service upgrade.

Disk space for databases

Will you install any databases on your system? How many? How large? Are they production or development databases? What is the high availability strategy and do you require multi-host attached disks or do you plan to use the concurrent disk environment of AIX? What is the data protection and disk failure recovery strategy? Do you require disk mirroring or RAID 1 or RAID 5?

For each database you need to determine:
- How much temporary space you need.
- How much space you need for logs.
- How much space you need for the database definition and data dictionary.
- How much space you need for rollback files.

If you plan to use multi-host attached disks or disk mirroring, you must also take into account what types and how many adapters you will need. This might later determine the node models you need because some nodes have fewer adapter slots than others.

If you plan to use DCE security services, see the book *IBM DCE Version 3.1 for AIX: Quick Beginnings*, which describes DCE and explains how to plan for, install, and configure the program.

Disk requirements for the IBM Virtual Shared Disk component of PSSP

IBM Virtual Shared Disk is a subsystem that allows applications running on multiple nodes within the same switch partition to access a raw logical volume as if it were local at each of the nodes. Volume groups defined to be used by the IBM Virtual Shared Disk subsystem ought to be used only to define IBM Virtual Shared Disks.

Another optional component, the IBM Recoverable Virtual Shared Disk subsystem, enhances IBM Virtual Shared Disk availability by supporting multi-host attached physical disks such that another node can take over the I/O service if the IBM Virtual Shared Disk node or related communication adapter fails. HACMP/ES is another licensed program which you can use to failover volume groups for applications that do not use IBM Virtual Shared Disks, including applications that use raw logical volumes or file systems.

Define all IBM Virtual Shared Disks on external disk storage drives. Data residing on an internal disk that is not multi-host attached to another disk will be lost if the node containing that internal disk fails. That type of data loss will occur whether or not the IBM Recoverable Virtual Shared Disk subsystem is in use.

File system requirements

Plan ahead of time for expected growth of all your file systems. Also, monitor your file system growth periodically and adjust your plans when necessary.

When the AIX 5L 5.1 32-bit kernel is enabled, the default file system is JFS. When the AIX 5L 5.1 64-bit kernel is enabled, the default file system is JFS2. It is possible to switch kernels but maintain the original file system. If a kernel switch is done, existing file systems remain the same, but all new ones are created as JFS for the 32-bit kernel and JFS2 for the 64-bit kernel. For more information, see the chapter ″Base Operating System (BOS)″, heading ″64 bit Kernel″, sub-heading ″Installation and Enablement″ in *AIX 5L for POWER Version 5.1 Release Notes*.

Boot-install requirements

The number of boot-install servers and the network layout of their Ethernet connections can affect the efficiency of your system. See “System topology considerations” on page 102 for suggested boot-install configurations for various system sizes.

Multiple boot requirements

Definable multiple boot images (alternate root volume groups) provide you with the fall back mechanism for SP systems or system partitions in case a problem is found in the system software, hardware, or application software. This requires one disk or set of disks, each holding a complete and different image of the operating system. If you want alternate boot system images, make sure you plan for enough disk space.

Note: The AIX **alt_disk_install** function is not related to the PSSP alternate root volume group support and is not supported with PSSP installation.

Mirrored root volume group requirements

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

When you install a node, you choose how many copies to make of the root volume group. AIX supports one (the original), two (the original plus a mirror), or three (the original plus two mirrors) copies of a volume group. IBM urges you to mirror the root volume group by choosing at least two copies. PSSP provides commands to facilitate mirroring on the SP system. The book *PSSP: Administration Guide* has information about mirroring a root volume group.

To mirror a root volume group, you need one disk or set of disks for each copy, depending on how many are needed to contain one complete system image. IBM, with a desire to provide a system having improved availability, delivers all new POWER3 SMP nodes with disk pairs as a standard feature. IBM urges you to use the extra disk for a mirror of the root volume group when you install your system.

Note: When you install a POWER3 SMP high node, the root volume group is automatically mirrored using the two internal disk drives by default. Each of these disks have the capacity to contain one complete system image. You must override the default setting during the install process if the node is configured without the two internal drives or if you choose to not use mirroring.

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. External disk can be used for booting the AIX operating system and for data.

Disk options offer the following trade-offs in price, performance, and availability:

- For availability, you can use the software mirroring function of AIX or a hardware RAID solution:
  - RAID 0: Disk spanning and striping for increased performance (No redundancy)
  - RAID 1: Data mirroring using fully redundant disk
  - RAID 3: Data Striping with dedicated parity disk
  - RAID 5: Data Striping with block interleaving of distributed parity across disk
  - RAID 10: RAID 1 mirroring of RAID 0

- For best performance when availability is needed, you can use mirroring, RAID 1, or RAID 10, but these require twice the disk.
- 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 in the RAID array. An N+P RAID 5 array, comprised of N+1 disks, offers N disks worth of storage, therefore it does not require twice as much disk.

  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 8 disks is a 7+P RAID 5 array, providing 7 disks worth of available protected storage. A hot spare provides no additional usable storage but provides a disk which quickly replaces a failed disk in the RAID 5 array. All disks in a RAID 5 array should be the same size, otherwise disk space will be wasted.
- For low cost when availability due to disk failure is not an issue, you can use what is known as JBOD (Just a Bunch of Disk).

After you choose a disk option, be sure to get enough disk drives to satisfy the I/O requirements of your applications. 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 and scalability requirements of the storage.
- The technology of the interconnect to the disk, like SCSI, SSA, or Fibre Channel.
- The availability protection strategy (mirroring, RAID 5), if any.
- The I/O rate you require for storage performance.
- Whether you will use external disk storage for boot-install devices as part of the root volume group. If you do, it should not be shared across nodes or have multiple paths within a node.
- Connectivity to the storage per node, such as single or multiple paths. For example, the SSA device driver and Subsystem Device Driver for IBM Enterprise Storage Server each provide multiple path capability to a disk.
- Whether you need multi-host connections to a disk:
  - The IBM Recoverable Virtual Shared Disk component of PSSP needs multi-host attached external disks available to be primary and backup IBM Virtual Shared Disk servers.
  - High Availability Cluster Multi-Processing (HACMP) supports two through eight-way sharing.
  - AIX clusters running GPFS which can support two through eight nodes sharing SSA disk or three through thirty-two nodes sharing Fibre Channel disks.

Note: For more information about Fibre Channel, see the book *Fibre Channel Planning and Integration Guide*. You can find information about the available storage subsystems on the Internet at the address:

http://www.storage.ibm.com

Completing the external disk storage worksheet

The following table shows how the `ABC Corporation` recorded their external disk storage needs. Record your external disk storage needs on your copy of Worksheet 3, Table 62 on page 283. You will also apply that information in your copy of Worksheet 4, “Major system hardware components” in Table 63 on page 284.

Table 6. ABC Corporations's external disk storage needs

External disk storage – Worksheet 3			
Disk subsystem	**Adapters (# - type)**	**Number of disks**	**Disk size**
7133 (SSA)	4 - FC 6230 PCI Advanced Serial RAID Plus	32	9.1GB

Question 7: What are your reliability and availability requirements?

What are your reliability and availability requirements? Who is going to use the SP? For some users reliability is not worth the cost. For others it is worth any cost and extremely important to keep the production system up and running. Two of the functions in PSSP that assist in reliability and availability are the High Availability Control Workstation and SP system partitions.

Systems that use a switch for connectivity also offer enhanced availability.

High Availability Control Workstation

One function providing enhanced reliability is the High Availability Control Workstation (HACWS). It supports a second control workstation that effectively eliminates the control workstation as a single point of failure. When the primary control workstation becomes unavailable, either through a planned event or a hardware or software failure, this SP high availability component detects the loss and shifts the workload off that component and to a backup control workstation.

To provide this extra reliability and eliminate the control workstation as a single point of failure, you need both extra hardware and software as summarized in Table 7.

Table 7. Requirements for the High Availability Control Workstation

Software	An additional AIX license
	The HACWS optional component of PSSP
	2 licenses for the HACMP program
Hardware	A second control workstation
	HACWS Connectivity Feature

Planning and using the HACWS component of PSSP will be simpler if you configure your backup control workstation identical to the primary control workstation. Some components must be identical, others can be similar. Wait until the last question about control workstation hardware and software to specify the components. For now, you need only decide if you need HACWS support. For more information and usage restrictions, see Chapter 4, "Planning for a high availability control workstation" on page 129.

SP system partitions

Partitioning your SP system can aid in system availability. This support lets you logically divide the SP system into non-overlapping groups of nodes called system partitions. You can then use one system partition, for example, to test new levels of AIX, PSSP, licensed programs, application programs, or other software on an SP system that is currently running a production workload in another system partition, without disrupting that workload. The partitioning solution assumes that you have nodes available for another system partition. A minimum system partition consists of at least two drawers (or four slots). If you do not partition your SP system, everything is considered to be in one SP system partition.

Note: System partitioning is not supported in systems with the SP Switch2, nor with clustered server systems.

You might not have to partition your system just for installation and testing. Coexistence support, which allows you to migrate one node of your system at a time, also promotes system availability. With coexistence, you are permitted to have multiple levels of PSSP operating within a single SP system partition. However, limits and restrictions do apply. There are issues particularly involving security support. See Chapter 11, “Planning for migration” on page 213 to help you decide whether coexistence alone is right for you.

A good use for system partitions is to create multiple production environments with the same non-interfering characteristics that benefit a testing partition. With system partitions the environments are sufficiently isolated so that the workload in one environment is not adversely affected by the workload in the other. They might be especially useful to isolate services which have critical implications to job performance, for example the switch. System partitions let you isolate switch traffic in one system partition from the switch traffic in other system partitions.

You might consider partitioning your SP system if you need a strong security policy for part of your SP system, but you do not want the overhead of enforcing that security policy on the entire system. Whatever your purpose for having SP system partitions, the same set of authentication methods for security must be enabled on all nodes within one SP system partition. Keep in mind that any system partition with a node running at a level earlier than PSSP 3.2 must have Kerberos V4 configured and enabled as an authentication method for AIX remote commands while the compatibility mode must be enabled for use by SP trusted services.

Remember, initially the system is a single partition. The number of system partitions you can define depends upon the size of your SP system. See Chapter 7, “Planning SP system partitions” on page 171 for more information and rules about system partitions. If you decide you want system partitions, study that chapter before completing your system plan. For now, you only need to decide if it is something you want and how many system partitions you think you’ll need. You can return and modify these answers if you find other information that affects your decision.

Table 8. Function checklist

Check	Function
	Do you want the redundancy of a High Availability Control Workstation?
0	How many system partitions do you want?
x	• Will some nodes run PSSP 3.5 and AIX 5L 5.1?
x	• Will some nodes run PSSP 3.4 and AIX 5L 5.1?
	• Will some nodes run PSSP 3.4 and AIX 4.3.3?

Table 8. Function checklist (continued)

Check	Function
	• Will you run PSSP 3.2 and AIX 4.3.3?
Note: Review coexistence limitations in Chapter 11, “Planning for migration” on page 213 to help you decide if you need to partition your system.	

Question 8: Which and how many nodes do you need?

Your answer to this question might be based on financial limits or it might be based on performance requirements. Keep in mind that any Cluster 1600 system managed by PSSP (SP system) is a ***scalable*** parallel processing system which means that you can add more nodes later. A Cluster 1600 system managed by PSSP is technically an SP system if there is at least one SP frame with nodes in it or it uses the SP Switch2 or SP Switch.

For new SP systems there are SP rack-mounted thin and wide processor nodes. If you already have an SP system you might also have high nodes. There are other processor nodes which can function like SP rack-mounted nodes, but are in servers that are not in an SP frame. Each server can be within an SP system, in which case it is called an ***SP-attached server***, or it can be part of a ***clustered servers*** system. These terms merely signify the system configuration in which a server participates when it runs the PSSP software.

A *clustered servers* system configuration has no SP frames. It is comprised of a control workstation running the PSSP software and connected to servers containing what PSSP recognizes as nodes, also running their own images of AIX, PSSP, and PSSP-related system management software. These configurations are differentiated in order to be able to associate connectivity, layout, and numbering rules with them later. ***Except where otherwise noted, statements in this book about nodes apply equally whether a node is in an SP frame, one of multiple LPARs in a server, or is a single node in an SP-attached server or clustered servers configuration. Functionally, it is simply a node in the system.***

There is also the SP Switch Router, not a processor node, but an extension node to handle high data transmission demands in an SP system that uses the SP Switch.

Your answers to the prior questions should have helped you determine the type of work for which you will be using the system. For example, if you determined that you want to use SP system partitions, it can affect the number of nodes you require. If you are planning a system with an SP switch, you need to understand and follow the support and configuration criteria associated with that switch, regardless of whether your system has an SP or a clustered servers configuration. The PSSP software, primarily the communication subsystem, works differently based on whether or not the system has a switch, which switch, and can there be multiple SP system partitions. Since the system is scalable, you can select fewer nodes now and add more later or select more now and scale down later.

Note: Some helpful hardware information is included here to help you select nodes and other major system hardware components and record your choices. For complete hardware information including dependencies, like which adapters go with which nodes and switches, see the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*.

To decide which and how many nodes you need, consider the overall capacity of each node keeping in mind your function, performance, network, communication, and data transmission requirements, together with the software that you have selected to satisfy your needs, and which nodes can handle those combined demands. There are also some limitations resulting from hardware differences.

Several worksheets originally for information about one SP frame at a time, with rows that relate to the nodes in slots within that frame, have been adjusted to accommodate servers as well. The p690 and p670 servers have characteristics similar to SP frames. Use one copy of such worksheets for each p690 or p670 server with multiple LPARs. Use a row for each LPAR node.

For servers in a one-to-one frame and node relationship, including unpartitioned p690 and p670 servers, use one copy of each worksheet to document up to 16 servers. Ignore the frame and switch number header fields and use a row for each server. Use the fixed row number in the worksheet as a sequence number and add the frame number you assign to each server.

Before you answer this question for yourself, address the topics:

- "Considering processor nodes"
- "Considering extension nodes" on page 52
- "Considering SP node frames" on page 53

Then proceed to the topics:

- "Completing the system hardware components worksheet" on page 53
- "Completing the node layout worksheets" on page 56
- "Completing the hardware configuration worksheet" on page 58
- "Completing network configuration worksheets" on page 60

Considering processor nodes

You choices include SP rack-mounted nodes and IBM @server pSeries servers.

A *clustered servers* system configuration can have up to 64 servers, depending on the hardware type, not exceeding a total of 64 PSSP nodes, also depending on the hardware type.

You can choose from the following SP rack-mounted nodes:

- Thin node

  The 375/450 MHz POWER3 SMP thin node can have two or four 64-bit 375 or 450 MHz processors, 256 MB to 8 GB of memory, two 32-bit PCI slots, a slot for the SP Switch MX2 adapter, integrated Ethernet, 4.5 GB to 18.2 GB of mirrored internal DASD, integrated Ultra SCSI, and an external RS-232 connection with active heartbeat used only by the HACMP application. This node is available as a single node. It does not have to be used in a pair. When it is a single node in a drawer, it must be in the odd-numbered slot.

  Thin nodes are supported in a short or tall SP frame in switchless, SP Switch2, or SP Switch configurations. In a tall frame they can connect to the SP Switch2 or the SP Switch. In a short frame they can connect to the SP Switch-8.

- Wide node

  Wide nodes greatly expand the I/O and network server functions of the SP system. Wide nodes occupy two slots in a frame and have more attachment options than thin nodes to increase DASD and network connectivity.

The 375/450 MHz POWER3 SMP wide node can have two or four 64-bit 375 or 450 MHz processors, 256 MB to 8 GB of memory, ten PCI slots (two 32-bit and eight 64-bit), a slot for the SP Switch MX2 adapter, integrated Ethernet, and 4.5 GB to 54.6 GB of mirrored internal DASD, integrated Ultra SCSI, and an external RS-232 connection with active heartbeat used only by the HACMP application.

Wide nodes are supported in a short or tall SP frame in switchless, SP Switch2, or SP Switch configurations. In a tall frame they can connect to the SP Switch2 or the SP Switch. In a short frame they can connect to the SP Switch-8.

You can order certain IBM @server pSeries servers as a Cluster 1600 system managed by PSSP 3.5. The following characteristics significant to PSSP software and configuration planning apply to any SP-attached or clustered server:

- It functions as at least one SP processor node, running the PSSP and PSSP-related licensed programs, but it is not physically in an SP frame.
- For a server that has function similar to the SP frame and node supervisor, there is SP-equivalent hardware control and monitoring. It is represented by PSSP as one frame with possibly multiple nodes. A model without a comparable frame and node supervisor has limited hardware control and monitoring from the control workstation. It is treated functionally by PSSP like one SP frame with one node.
- It connects directly to the PSSP admin LAN and to the control workstation.
- For optimal performance, you might have an SP switch in your system. In an SP Switch system configuration, each node must be connected to the SP Switch by a suitable adapter. In an SP Switch2 system configuration, you have optional connectivity. Nodes that are not supported on the SP Switch2 can remain in the system but not connected to the SP Switch2 It is not supported with an SP Switch-8.
- If the server on which you plan to run PSSP is already in use and connected to an IPv6 network, remove it from the IPv6 network before connecting it to the system. Some PSSP components tolerate IPv6 aliases for the IPv4 network addresses but not with DCE, HACMP, HACWS, or an SP switch. For information about the SP tolerating IPv6 aliases for IPv4 network addresses, see the appendix on the subject in the book *PSSP: Administration Guide*.
- These are the numbering considerations:
  - You will have to assign a frame number to the server.
  - Node numbers are automatically generated based on the frame number.
  - If you plan to use this server in a switchless SP system or with the SP Switch, you need to assign a valid switch port number for each node. However, in a switchless SP system where you will never use SP system partitioning, you can force the system to be non-partitionable and simply assign switch port numbers sequentially. See “Understanding partitionability choices” on page 97.
  - In clustered server systems you do not have to assign a switch port number, but you should if you anticipate expanding in the future by using an SP Switch or by adding a switchless SP system where you want to have multiple SP system partitions.
  - If you plan to use the server with the SP Switch2, switch port numbers are automatically generated.
- Neither SP-attached nor clustered servers are automatically supported by HACWS. If you already have HACWS running on your SP system, you might be able to continue using it if you understand and accept all the limitations, and you have the expertise necessary to support it. See “Limits and restrictions” on page 134.

The following considerations are uniquely significant to planning for running PSSP software in a clustered servers configuration, that is a configuration with no SP node frames:

- SP system partitioning is not supported in a clustered servers configuration.
- You might begin with a clustered servers configuration and later decide to add an SP frame containing the SP Switch2 or the SP Switch. When you use an SP switch, functions that depend on it are also available, like GPFS, IBM Virtual Shared Disks, and user space jobs. However, when you introduce an SP frame, even if just for a switch, the system then has an SP-attached server configuration – it is no longer a clustered servers configuration. These configurations are differentiated in order to discuss connectivity, layout, and numbering rules. The PSSP software, primarily the communication subsystem, works differently based on whether or not the system has a switch, which switch, and can there be multiple SP system partitions.
- If you add the SP Switch2, SP system partitioning does not apply, but the system must honor the rules associated with the SP Switch2.
- If it is likely that with future growth you might add the SP Switch, the system becomes partitionable and the clustered servers become SP-attached servers subject to the pertinent rules. They must honor all the placement and numbering rules associated with the SP Switch. In that case, plan your system in advance with appropriate frame numbers, and switch port numbers so that you can migrate to an upscaled SP system without having to totally reconfigure existing servers.
- If it is likely that with future growth you might add an SP node frame, the system becomes a switchless SP system, which is partitionable, and the clustered servers become SP-attached servers subject to the pertinent rules, but you have the following options:
  - If you will never use the SP Switch, you can force the system to be non-partitionable and ignore the switch port numbering rules for SP-attached servers. They can remain numbered sequentially.
  - If you might want to eventually use SP system partitioning, plan your system in advance with appropriate frame numbers, and switch port numbers so that you can migrate to an upscaled SP system without having to totally reconfigure existing servers.

Note: Be sure to read and understand the information regarding SP-attached servers and clustered servers in “Planning your network configuration” on page 110 and in “Understanding placement and numbering” on page 118.

Each of these servers have PCI-based 64-bit symmetric multiprocessors of varied capacities delivering performance, scalability, and reliability for today’s critical e-business applications. Each supports concurrent 32-bit and 64-bit application processing. You can use any of the following in a Cluster 1600 system managed by PSSP as SP-attached servers or in a clustered servers configuration:

- **IBM @server pSeries 690 and 670**

  The ***p690*** is an 8- to 32-way SMP server. It can be configured with 8, 16, or 24 1.0 GHz processors or with 16, 24, or 32 1.3 GHz processors. These physical resources can be configured into up to 16 logical partitions (LPAR). The ***p670*** is a 4- to 16-way SMP mid-range server utilizing the POWER4 dual processor on a chip. It can be configured with 4, 8, or 16 1.1 GHz processors. These physical resources can be configured into up to 16 logical partitions (LPAR). See “Scaling considerations for clusters” on page 51.

Each of these servers is represented by PSSP as a single frame. If it is not partitioned, it is represented as a single node in that frame. If the physical resources of the server are partitioned, each LPAR in that frame is a separate node running AIX 5L 5.1 with the correlating PSSP and PSSP-related software. Every LPAR in the server is recognized as a PSSP node – no provision is made in PSSP to ignore any LPAR as being a node.

An unpartitioned server has one LPAR and is seen by PSSP as one node. A partitioned server is seen by PSSP as one frame with as many nodes as there are LPARs. The number of these servers counts toward the total number of servers in one system. Additional constraints apply to p690 and p670 servers depending on the switch configuration. See “Scaling rules for clusters” on page 52.

Connectivity from the control workstation to the p690 or p670 server is through the Hardware Management Console (HMC) by a network connection to the PSSP admin LAN. The HMC is a processor that runs only the HMC software for installation and service support of these servers. The HMC has a serial connection to the Common Service Processor (CSP) that is integrated in the server. The HMC represents and manages the server through the Common Information Model (CIM) and provides hardware control capability and the function to logically partition the physical resources in the server. For a stand-alone server, the HMC is optional if the LPAR function is not needed. For a server that is SP-attached or are part of a system of clustered servers, the HMC is required whether or not the server is configured with LPARs. It is necessary for a single point of control from the control workstation. Redundant HMCs are not supported at this time.

These servers use the HMC hardware protocol. No RS-232 cable connecting to the control workstation is necessary.

Since each LPAR functions like an SP node, PSSP requires that each LPAR be configured with an Ethernet adapter connected to the PSSP admin LAN. Previously, PSSP required that the PSSP admin LAN connection be made explicitly through the en0 adapter from the node. For the PSSP admin LAN connection from this node, you can specify any Ethernet adapter that is currently supported for connection to the PSSP admin LAN. The adapter can be installed at any logical location on the device tree. You can specify an adapter to be configured by its physical location and the PSSP software will map the physical location to the logical device name when using AIX commands to define and configure the adapter on the node.

Each LPAR that is to use a switch needs to have a suitable adapter. If you plan to use the SP Switch2 with two switch planes, each LPAR needs two adapters, **css0** to connect to the first switch plane and **css1** to connect to the second switch plane. If you plan to use the SP Switch, you need to specify the switch port connections by using a PSSP configuration file. The switch primary and primary backup nodes can be assigned to LPARs on p690 or p670 servers. However, do not assign both the primary and primary backup nodes in the same physical server if possible. Since these LPARs share some common level of hardware and power controls, the entire switch would not be available if that server is not available for any reason.

PSSP does not provide an interface to logically partition these servers. You need to use the WebSM facility provided on the HMC console. However, you can display the HMC graphical user interface on the control workstation monitor. An interface to launch a remote WebSM session to the HMC console is provided by SP Perspectives.

An option is available through the HMC to logically partition the server such that the partitions contain processors, memory, and I/O that are in physical proximity.

Using this option results in more predictable computing performance for these LPARs. It might be an important consideration when the LPARs are used as compute nodes for large parallel applications or in systems where predictable node performance is required. See the book *Hardware Management Console Operations Guide* for additional information on configuring the LPARs.

- **IBM @server pSeries 660 Models 6M1, 6H0, 6H1**

  Each processor is represented as a single node in a single frame. It uses the Common Service Processor (CSP) to provide hardware control and monitoring similar to the SP frame and node supervisor. These servers use the CSP hardware protocol with one RS-232 cable to the control workstation. See "Scaling considerations for clusters" on page 51. The following are characteristics of these models:

  – The **p660 Model 6M1** is a 2, 4, 6, or 8-way SMP server with RS64 IV 668 MHz processors, 1 to 64GB of memory, 4 RIO ports, 8 EIA high, 8MB of L2 cache, and dual line cord support. The I/O drawer has 14 PCI slots with option of 2 boot DASD. You can have drawers with only SCSI DASD or with only SSA DASD.

  – The **p660 Model 6H0** is a 1-, 2- or 4-way SMP server that features IBM's innovative silicon-on-insulator (SOI) chip technology. You can have 450 MHz RS64 III or 600 MHz RS64 IV SMPs, up to 32GB of memory, and up to 28 PCI slots.

  – The **p660 Model 6H1** is a midrange, up to 6-way, 64-bit SMP server with superior performance, mainframe reliability and availability, and enhanced capacity over the RS/6000 Model H80. A fully configured system consists of one processor and two I/O drawers for a capacity of 28 PCI slots. It has up to 32GB of memory and 8MB of L2 cache.

- **IBM @server pSeries 680**

  The pSeries 680 is a 24 inch rack-mounted node. See "Scaling considerations for clusters" on page 51. Each is represented as a single node in a single frame. It is an up to 24-way, 64-bit SMP server with state-of-the-art copper silicon-on-insulator technology. A fully configured system has 24 600 MHz RS64 IV SMPs, 96GB of memory, 56 PCI adapter slots, 48 hot-swappable disk bays, 8 media bays, and 873.6GB of internal disk. This server uses the SAMI hardware protocol with two RS-232 cables to the control workstation.

Note: See information online that might help you decide which is best for your needs at address:

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

Table 9 on page 49 summarizes the nodes that you can currently order from IBM. They are all supported with the SP Switch2 or the SP Switch. The nodes that do not have to be in a drawer of an SP frame can be SP-attached or in a clustered system configuration where the control workstation and all nodes are running the PSSP software. See the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for hardware requirements.

Table 9. SP and other nodes currently available from IBM for use with PSSP 3.5.

Node (min nodes to SP drawer)	Processor	Processing	Min to max memory	Min to max internal disk space	Max switch planes	PSSP level at announce
p690 1-16 LPARs	8, 16, 24 or 32 1.0 or 1.3 GHz	64-bit SMP w AIX 5L 5.1	8 to 256GB	36.4GB to 9.3TB	2	3.4
p670 1-16 LPARs	4, 8 or 16 1.1 GHz	64-bit SMP w AIX 5L 5.1	4 to 128GB	36.4GB to 3.5TB	2	3.4
p660 6M1	2, 4, 6 or 8 750 MHz or 2 or 4 500 MHz	64-bit SMP w AIX 5L 5.1 or AIX 4.3.3	2 to 64GB	0 to 72.8GB	2	3.2
p660 6H1	1, 2, 4 or 6 750 MHz or 1, 2 or 4 600 MHz	64-bit SMP w AIX 5L 5.1 or AIX 4.3.3	512MB to 32GB	0 to 72.8GB	2	3.2
p660 6H0	1, 2 or 4 750 or 600 MHz	64-bit SMP w AIX 5L 5.1 or AIX 4.3.3	512MB to 32GB	0 to 72.8GB	2	3.2
p680	4, 6, 8, 12, 16 or 24 600 or 450 MHz	64-bit SMP w AIX 5L 5.1 or AIX 4.3.3	2 to 96GB	18.2 to 873.6GB	2	3.2
POWER3 Wide (1-1)	2 or 4 375 or 450 MHz	64-bit SMP w AIX 5L 5.1 or AIX 4.3.3	256MB to 16GB	0 to 109.2GB	1	3.1.1
POWER3 Thin (1-1/2)	2 or 4 375 or 450 MHz	64-bit SMP w AIX 5L 5.1 or AIX 4.3.3	256MB to 16GB	0 to 220.2GB	1	3.1.1

Note:

- w = with
- With p690 and p670 servers, the control workstation and the LPARs have to run AIX 5L 5.1.
- The p690 and p670 servers have no internal drives. They can have up to eight remote I/O (RIO) units, but must have at least one, each of which can have up to 16 disks of 36GB each. A RIO is equivalent to internal disk capacity.

You might already have SP rack-mounted nodes or other servers that are still supported and you can migrate to PSSP 3.5 with AIX 5L 5.1. You can still use any of the following nodes:

- The RS/6000 Enterprise Server H80, M80, and S80
- POWER3 SMP high nodes (375 or 222 MHz)
- POWER3 SMP thin and wide nodes (200 MHz)
- The RS/6000 Enterprise Server S70 or S7A
- The 332 MHz SMP thin and wide nodes

Not all nodes can be connected to the SP Switch2. Some can be in a system that has the SP Switch2 but cannot be connected to the switch.

Table 10 summarizes the nodes you might already have that can run PSSP 3.5 with AIX 5L 5.1. Nodes that do not have to be in a drawer of an SP frame might be attached to an SP system or might be in a cluster configuration where the control workstation and all nodes are running the PSSP software. If you plan to migrate an earlier level of PSSP to PSSP 3.4 on nodes you already have or you plan to run mixed levels of PSSP, be sure to carefully read Chapter 11, “Planning for migration” on page 213.

Table 10. Nodes you might already have that can run PSSP 3.5

Node (nodes to SP drawer)	Processor	Processing at announce	Min to max memory	Min to max internal disk space	Max switch planes	PSSP level at announce
M80	500 MHz	64-bit SMP w AIX 5L 5.1 or AIX 4.3.3	1 to 32GB	0 to 36.4GB	2	3.2
POWER3 High (1-2)	375 MHz	64-bit SMP w AIX 5L 5.1 or AIX 4.3.3	1 to 64GB	0 to 1.9TB	2	3.1.1
H80	450, 500, 600, or 668 MHz	64-bit SMP w AIX 4.3.3	512MB to 32GB	0 to 36.4GB	2	3.2
S80	450 MHz	64-bit SMP w AIX 4.3.3	512MB to 64GB	4.5GB to 218GB	2	3.1.1
POWER3 High (1-2)	222 MHz	64-bit SMP w AIX 4.3.3	1 to 16GB	9.1 to 18.2GB mirrored	2	3.1.1
POWER3 Wide (1-1)	200 MHz	64-bit SMP w AIX 4.3.3	256MB to 4GB	4.5 to 36.4GB mirrored	1	3.1
POWER3 Thin (1-1/2)	200 MHz	64-bit SMP w AIX 4.3.3	256MB to 4GB	4.5GB to 18.2GB mirrored	1	3.1
S7A	262 MHz	64-bit SMP w AIX 4.3.3	512MB to 32GB	4.5GB to 218GB	1 ¬sps2	3.1
S70	125 MHz	64-bit SMP w AIX 4.3.3	512MB to 16GB	4.5GB to 218GB	1 ¬sps2	3.1
Wide (1-1)	332 MHz	32-bit SMP w AIX 4.2.1 or 4.3.2	256MB to 3GB	4.5GB to 36.4GB	1	2.4
Thin (1-1/2)	332 MHz	32-bit SMP w AIX 4.2.1 or 4.3.2	256MB to 3GB	4.5GB to 18.2GB	1	2.4
Thin (1-1/2)	160 MHz	32-bit SMP w AIX 4.2.1 or 4.3.2	64MB to 1GB	4.5GB to 18.2GB	1¬sps2	2.3
High (1-2)	200MHz	32-bit SMP w AIX 4.2.1 or 4.3.2	256MB to 4GB	4.5GB to 18.2GB	1¬sps2	2.3
Wide (1-1)	135 MHz	32-bit SMP w AIX 4.2.1 or 4.3.2	256MB to 3GB	4.5GB to 36.4GB	1¬sps2	2.2
Thin (2-1)	120 MHz	32-bit SMP w AIX 4.2.1 or 4.3.2	64MB to 1GB	2GB to 18.2GB	1¬sps2	2.2

Table 10. Nodes you might already have that can run PSSP 3.5 (continued)

Node (nodes to SP drawer)	Processor	Processing at announce	Min to max memory	Min to max internal disk space	Max switch planes	PSSP level at announce
High (1-2)	112MHz	32-bit SMP w AIX 4.2.1 or 4.3.2	64MB to 2GB	2GB to 6.6GB	1¬sps2	2.2
Wide (1-1)	77 MHz	32-bit w AIX 4.2.1 or 4.3.2	64MB to 2GB	4.5GB to 18GB	1¬sps2	1.2
Wide (1-1)	66 MHz	32-bit w AIX 4.2.1 or 4.3.2	64MB to 2GB	4.5GB to 18GB	1¬sps2	1.2
Thin (2-1)	66 MHz	32-bit w AIX 4.2.1 or 4.3.2	64MB to 512MB	2GB to 9GB	1¬sps2	1.2

Note:

- w = with, ¬sps2 = cannot connect to SP Switch2
- Servers must run PSSP 3.4 or later to connect to the SP Switch2 or the SP Switch.
- SP nodes must run PSSP 3.4 or later to connect to the SP Switch2.
- The POWER3 high node can have up to six SP Expansion I/O Units to increase internal DASD and connectivity to external DASD and networks.

Scaling considerations for clusters

A Cluster 1600 managed by PSSP can consist of from two to 128 AIX operating system images. An operating system image, or logical node, can be:

- A 7017 server (S70, S7A, S80, p680)
- A 7026 server (H80, M80, p660, 6H0/6H1, 6M1)
- A 7040 server (p670, p690) running as a full system partition
- An LPAR of a 7040 server (p670, p690)
- A 9076 SP node

Logical nodes are limited in a cluster running PSSP as described in Table 11 on page 52 and “Scaling rules for clusters” on page 52. Any cluster system that exceeds any of the announced supported logical node limits is not supported.

Note: Refer to the *Read This First* document for the latest information on scaling limits.

Table 11. 7040 Cluster Limits

Maximum Values	SP Switch	SP Switch2	SP Switch2, total switched plus not switched, subject to other limits*	Switchless
p690/p670 servers per cluster	32	32	32	32
LPARs per p690 server	8	16	16	16
LPARs per p670 server	4	4	16	16
LPARs per Cluster	128	128	128	128
Additional Information				
Number of switch planes supported	1	1 or 2	1 or 2	0
Number of p690/p670 servers per HMC	8	8	8	8
Number of LPARs per HMC	32	32	32	32
Notes: 1. Hardware Management Console (note that redundant HMCs are not supported). 2. With the SP Switch, switched and non-switched logical nodes cannot be mixed.				

Scaling rules for clusters

A PSSP cluster must meet all of the following limits:

- No more than 128 logical nodes from the set {7017, 7026, 7040, 7040 LPAR, 9076}.
- No more than 64 servers from the set {7017, 7026, 7040}
- No more than 32 servers from the set {7017, 7040}
- No more than 16 servers from the set {7017}

For example:

- 32 p690s with 4 LPARs each or 16 p690s with 8 LPARs each
- 16 p690s with 4 LPARs each or 64 p660s
- 16 p690s with 4 LPARs each, 32 p660s, and 32 SP nodes
- 12 p690s with 4 LPARs each, 16 p680s, 16 p660s, and 48 SP nodes

Considering extension nodes

Configuring a system using extension nodes requires special planning with respect to processor nodes. The only unit currently in this category is a dependent node, the SP Switch Router. ***Extension nodes are only supported with the SP Switch, not with the SP Switch2.***

Regardless of which processor node types you use, if you plan to use the SP Switch and order an SP Switch Router, you have to reserve one node slot for each connection from the SP Switch Router to the SP Switch. This is necessary to have a switch port available for each SP Switch Router Adapter.

Each SP Switch Router Adapter in an SP Switch Router must be connected to a valid switch port on the SP Switch. To accommodate that requirement, each

dependent node logically occupies a slot in an SP frame and physically occupies the switch port corresponding to that slot. A processor node must not be assigned to the same slot, although a wide or high node can overlay the slot. For a discussion of valid extension node slots read ″Chapter 3, “Defining the configuration that fits your needs”″ beginning with “Considering the SP Switch Router” on page 115.

Considering SP node frames

There are four SP frame models for the SP system which you can populate with optional nodes and switches to create the SP system configuration of your choice. Your layout can range from a single-frame starter system to a highly parallel, large-scale system. The frame models for SP nodes currently available from IBM are listed in Table 12. Other frames with no nodes are available but not discussed here.

Table 12. The basic SP node frames

Frame Model	Description
500	Short base frame, power supply, additional equipment: • up to eight thin or wide nodes, one drawer required, one node required to become a functional SP. Must be replaced with tall base frame in order to scale up from eight nodes. • SP Switch-8 optional, nodes must be in sequence and not interspersed with empty drawers • no SP-attached servers or POWER3 high nodes.
550	Tall base frame, power supply, additional equipment: • up to 16 nodes, type is optional, one drawer required, one node required to become a functional SP • option to use one type of SP switch: – SP Switch2 – nodes can be in any sequence, interspersed with empty drawers, and placed at any slot that is physically suitable for the node within a switch capsule. – SP Switch with 16 ports – nodes can be in any sequence and interspersed with empty drawers but must honor the node placement rules within a switch capsule. – SP Switch with 8 ports – nodes must be in sequence and not interspersed with empty drawers. Limits frame to eight thin or wide nodes, no high nodes or other processor nodes are supported with this switch. Must be replaced with SP Switch2 or SP Switch with 16 ports in order to scale up. • scalable up to 128 nodes with SP Switch2 or SP Switch
1500	Short expansion frame, same support as short base frame but has no prerequisite of a node
1550	Tall expansion frame, same support as tall base frame but has no prerequisite of a node.
Note: If you plan to add any new node to a frame you already have, your frame might need a power upgrade. See the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for more information about frame options, especially relating to SP switches.	

Completing the system hardware components worksheet

Now it’s time to take all the information you have thought about and start to lay out your system requirements on detailed worksheets. These worksheets are an

invaluable tool for helping you plan your configuration and installation in detail. If you have not done it already, make copies of the worksheets in Appendix C, “SP system planning worksheets” on page 281. The worksheets in this chapter have been filled out for a hypothetical customer, the `ABC Corporation`. The major system hardware components selected for the `ABC Corporation` are in Table 13 on page 55.

Note: See the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for requirements of other hardware based on your choices of the major system hardware components. For instance, each node or type needs a specific adapter to attach to a specific switch.

Table 13. Major system hardware components

Major system hardware components – Worksheet 4		
Company name: ABC Corporation		
Customer number: 999999		**Date:** October 18, 2002
Customer contact: Jim Smith		**Phone:** 1-800-555-5678
IBM contact: Susann Burns		**Phone:** 1-800-555-6789
Complete the following by entering quantities to order...		
Frames	**Nodes**	**Attached Servers**
550 (tall): 1	375/450 MHz Thin: 6	p680:
1550 (tall):	375/450 MHz Wide: 2	p660 6H1:
500 (short):	375 MHz High: 1	p660 6M1:
1500 (short):		p660 6H0:
		p690 (number LPARs:)
		p670 (number LPARs:)
Switch subsystem components		
SP Switch2: 1	SP Switch2 Adapter: 1	SP Switch2 PCI-X Attachment Adapter:
	SP Switch2 MX2 Adapter: 8	SP Switch2 PCI Attachment Adapter:
SP Switch 16-port:	SP Switch MX Adapter:	SP Switch MX2 Adapter:
SP Switch 8-port:	SP Switch Adapter:	RS/6000 SP System Attachment Adapter:
SP Switch Router:	SP Switch Router Adapter:	
External storage units:	*Type*	*Quantity*
	7133	32
Network media cards:	*Type*	*Quantity*
Fill in after you place your order...		
Cluster model number: 9078-160	Purchase order number:	
Cluster serial number: 200200770	SP model number: 9076	
Control workstation:	SP serial number: 510077730	
Peripherals:		

Complete your copy of Worksheet 4, “Major system hardware components” in Table 63 on page 284 with the heading information, the number of frames, the number of each node type, the number of switches and adapters, and other components you want. If you selected a external disks in “Question 6: What are your disk storage requirements?” on page 37, copy the information from that table to your copy of Worksheet 4.

When you place your order for an SP system, fill in the SP model number, SP serial number, Cluster model number, Cluster serial number, and the purchase order number for easy reference.

Completing the node layout worksheets

These instructions explain one way to document your node layout. First draw a diagram of your system. Then add network information to that diagram. After that, write your network information into your copies of the worksheets. Fill in as many copies of Worksheet 5, Figure 58 on page 285, as you need. Use one copy for each SP frame or equivalent. An example network is shown in Figure 5 on page 57 and Figure 6 on page 58.

To complete the node layout worksheets, do the following:

1. For each frame, fill in the frame number and the switch number on the lines marked **Frame** and **Switch** at the bottom of the diagram.

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

2. Write the unique ID numbers, such as node number and expansion number. Indicate whether each node is a wide, thin, or high node using a unique identifier for each. For example, you might represent wide nodes with a *w*, thin nodes with a *t*, and high nodes with an *h*. Slot numbers are already present in each frame diagram. Wide nodes occupy two slots and use the odd slot number. High nodes occupy four slots and use the lowest odd slot number in the set. Cross out the even slot numbers in wide nodes and the three excess slots in high nodes except those to be used by an extension node or SP-attached server.

   If you are planning an SP Expansion I/O Unit, extension node, or SP-attached server, make up an indicator for each of them and mark the slots they use. An SP Expansion I/O Unit physically occupies a slot but leaves the switch port free. An extension node or SP-attached server only logically occupies the slot but uses the associated switch port in a system with the SP Switch. You can use the same slot for both an SP Expansion I/O Unit and either an extension node or an SP-attached server.

   Figure 5 on page 57 shows a single frame with numbered slots.

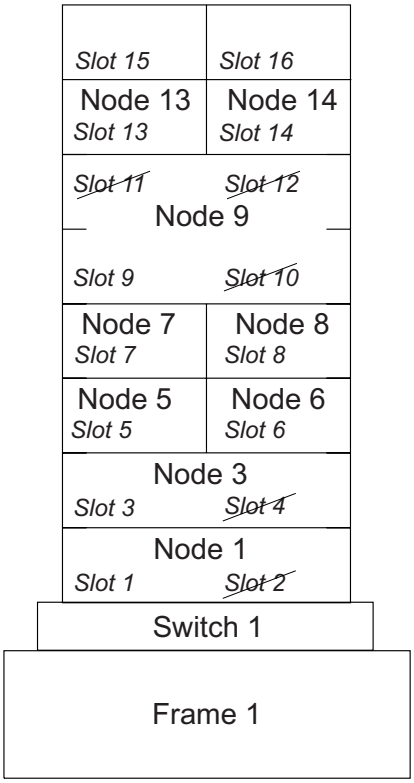


Figure 5. A node layout example

For instruction on node and switch port numbering, see “Understanding placement and numbering” on page 118.

3. Sketch your SP Ethernet admin LAN connections to each node and to the control workstation. Indicate specific adapter connections (for example, en0 and en1 connections). See “System topology considerations” on page 102 for SP Ethernet admin LAN tuning considerations.
4. Sketch additional network connections.

   At this point, your layout might look something like that in Figure 6 on page 58.

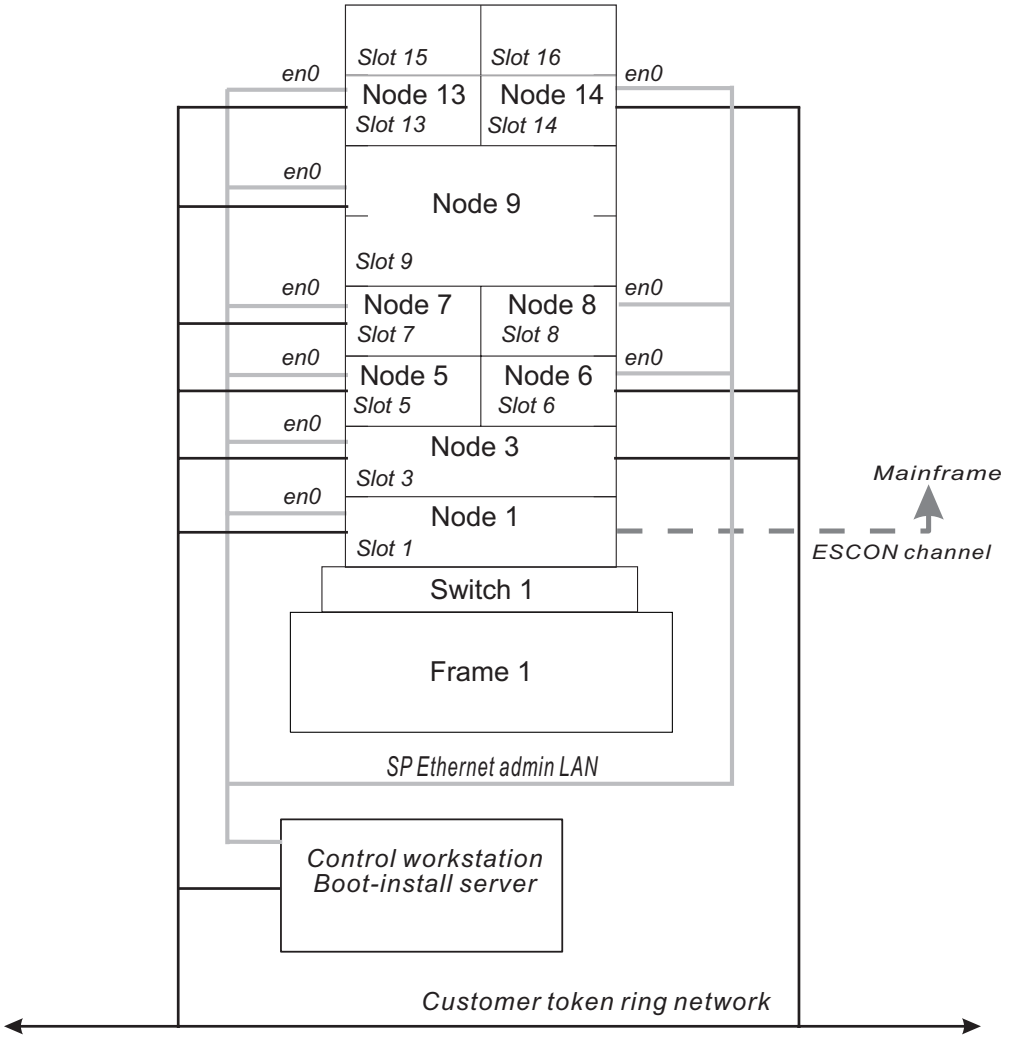


Figure 6. A node layout example with communications information

5. Sketch connections to any routers, gateways, or other networks.
6. Indicate network addresses, netmasks and host names for each subnet and node address on each node interface.

Completing the hardware configuration worksheet

You need to record the hardware configuration of your frames. At the same time you decide what types and how many nodes and other units you want, you also need to decide and keep track of how many processors, how much processor memory, and how much internal disk storage each processor node will have. Each of these values will affect the performance of your system, so choose carefully.

After you decide on that information, fill in your copy of Worksheet 7, “Hardware configuration by frame” in Table 64 on page 286. You need multiple copies of this worksheet depending on what is to comprise your system. Use one copy of this worksheet for each SP frame or equivalent like the p690 or p670 server.

For SP frames with nodes, this worksheet is intended for information about one frame at a time and the rows relate to the nodes and SP Expansion I/O Units in slots within that frame. Complete column three with node type for SP rack-mounted nodes. You might want to put the frame number there for SP-attached servers. To

be thorough, include any SP Expansion I/O Units. Record the frame and slot of the associated node, as demonstrated by the entries for slots 15 and 16 in Table 14 on page 60. Leave blank any fields or entries that do not apply.

For p690 and p670 servers, use one copy for each server, completing up to 16 rows in each copy with each row being a node based on an LPAR. See "Considering an HMC trusted network" on page 113 before you complete the worksheet.

For servers in a one-to-one node and frame relationship, including unpartitioned p690 and p670 servers, use one copy of this worksheet to document up to 16 servers, using a row for each server. Use more copies for more servers. Ignore the frame and switch number header fields and use each row as a frame. You might want to put the server model in column three.

The `ABC Corporation` planned an SP system and made the choices in Table 14 on page 60. Fields that do not apply to their system are left blank.

Table 14. ABC Corporations's choices for hardware configuration by frame

Hardware configuration by frame – Worksheet 7					
Frame number: 1		**Hardware protocol:** SP		**Switch number:** 1	
p690/p670 name:			**HMC hardware monitor user id:**		
HMC trusted network adapter names or IP addresses:					
Slot or **LPAR** or **frame**	**Node** or **Expansion number**	**Node type** or **Associated frame/slot**	**Number processors, memory**	**Internal disk**	**Additional adapters**
1	1	wide	4, 8GB	36.4GB mirrored	TokenRing(1), SSA(4), Ethernet(1), ESCON(1)
2	—				
3	3	wide	4, 8GB	36.4GB mirrored	TokenRing(1), SSA(4), Ethernet(1)
4	—				
5	5	thin	4, 8GB	18.2GB mirrored	TokenRing(1), SSA(2)
6	6	thin	4, 8GB	18.2GB mirrored	TokenRing(1), SSA(2)
7	7	thin	2, 3GB	18.2GB mirrored	TokenRing(1), Ultra SCSI(2)
8	8	thin	2, 3GB	18.2GB mirrored	
9	9	high	16, 32GB	18.2GB mirrored	TokenRing(1), SSA(8), Ethernet(4), FDDI(2), SCSI(4)
10	—				
11	—				
12	—				
13	13	thin	2, 3GB	18.2GB mirrored	TokenRing(1), Ultra SCSI(2)
14	14	thin	2, 3GB	18.2GB mirrored	TokenRing(1), Ultra SCSI(2)
15	15	1/9			
16	16	1/9			
Note: For p690 or p670 name, use the name you assigned to the server using the HMC user interface.					

Completing network configuration worksheets

Each adapter in each node, workstation, and router has an IP address. Each of these addresses has a separate name associated with it. The SP system uses only IPv4 addresses. Some PSSP components tolerate IPv6 aliases for IPv4 network addresses but not with DCE, HACMP, HACWS, or an SP switch. For information about the SP system tolerating IPv6 aliases for IPv4 network addresses, see the appendix on the subject in the book *PSSP: Administration Guide*.

Interface names

During installation and configuration, all addresses, including the router addresses, must be resolvable into names. Likewise, all names both long and short, must be resolvable into addresses. If your network administrator or support group provides name-to-address resolution through DNS, NIS, or some other means, they need to plan for the addition of all these names to their servers before the system arrives. You must specify these names during configuration to be set in the PSSP System Data Repository (SDR). Since AIX is case sensitive, the names must match exactly.

Host names

Independent of any of the network adapters, each node has a *host name*. Usually the host name of a node is the name given to one of the network adapters in the node.

The host name in the worksheet is referring to the name given to that adapter. You need to select which of these adapter host names is to be the one given to the node. Mark the column of the adapter that will be the host name. While completing these worksheets, keep the following criteria in mind:

- An application might require that the node host name be the name associated with the adapter over which its traffic will flow.
- If the host name of the control workstation is not set to the name of the SP Ethernet admin LAN adapter, the default route of the nodes must be an adapter that can be automatically configured. For a list of adapter types that can be automatically configured, see the **spadaptrs** command in the book *PSSP: Command and Technical Reference*.

Completing the PSSP admin LAN and additional node network configuration worksheets

Review your network topology and fill in your copies of worksheets 8 and 9 which start with Table 65 on page 287. Be sure to make extra copies before you complete them. If you have additional network adapters planned for some or all of your nodes, you need to plan their network information also. See the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for information about required and optional adapters.

For SP frames with nodes, these worksheets are intended for information about one frame at a time and the rows relate to the nodes in slots within that frame. For p690 and p670 servers, use one copy for each server, using multiple rows in each copy. Each row is a node based on an LPAR. For servers in a one-to-one node and frame relationship, including unpartitioned p690 and p670 servers, use one copy of this worksheet to document up to 16 servers, using a row for each server. Use two copies to document up to the maximum of 32 servers in one system. Ignore the frame and switch number header fields and use each row as a frame. Use the fixed row number in the worksheet for sequencing, and enter the frame number you assign to the server.

For all nodes except those on a p690 and p670 server, you must use the adapter name `en0`. For nodes on a p690 and p670 server, you can use `en0`, another Ethernet name, or the physical location, like `en2` or `U1.9-P1-I2/E1`, of any Ethernet adapter that is supported in the node for connection to the SP Ethernet admin LAN.

The `ABC Corporation` completed their network configuration worksheets starting with the “PSSP admin LAN configuration” in Table 15 on page 62. They chose to plan only the additional token ring and ESCON connections at this time. They also completed the worksheet for additional node network adapters shown in Table 16 on page 63.

Table 15. ABC Corporation's SP Ethernet admin LAN

PSSP admin LAN configuration – Worksheet 8				
Company name: ABC Corporation			**Date:** October 18, 2002	
Frame number: 1			**p690/p670 server name:**	
Associated node slot or **frame**	**Admin LAN netmask:** 255.255.255.192 *(must be en0 unless p690/p670)*			**Default route**
	Hostname	**Adapter name** or **physical location**	**IP Address**	
1	spnode01	en0	129.40.60.1	129.40.60.125
2	--	--		
3	spnode03	en0	129.40.60.3	129.40.60.125
4	--	--		
5	spnode05	en0	129.40.60.5	129.40.60.125
6	spnode06	en0	129.40.60.6	129.40.60.125
7	spnode07	en0	129.40.60.7	129.40.60.125
8	spnode08	en0	129.40.60.8	129.40.60.125
9	spnode09	en0	129.40.60.9	129.40.60.125
10	--	--		
11	--	--		
12	--	--		
13	spnode13	en0	129.40.60.13	129.40.60.125
14	spnode14	en0	129.40.60.14	129.40.60.125
15	--	--		
16	--	--		

Notes:

1. AIX is case sensitive. If name-to-address resolution is provided by DNS, NIS or some other means, the names in the SDR must match exactly. Otherwise, use lower case for the host name and addresses.
2. Wide nodes occupy two slots and use the *odd-numbered* slot.
3. High nodes occupy four slots (2 drawers) and use the lowest odd-numbered slot.
4. Use Ethernet adapter name or physical location for p690 and p670 nodes only. All other nodes must use en0.
5. For p690 or p670 server name, use the name you assigned to the server with the HMC user interface.

Table 16. ABC Corporation's additional adapters node network

Additional adapters node network configuration – Worksheet 9				
Company name: ABC Corporation			**Date:** October 18, 2002	
Frame number: 1			**p690/p670 server name:**	
Token ring speed: 16				
Associated node slot or **frame**	**Additional adapters netmask:** 255.255.255.192			**Default route**
	Adapter name or **physical location**	**Hostname**	**IP address**	
1	tr0	sptok01	129.40.61.1	129.40.60.125
2	--			
3	tr0	sptok03	129.40.61.3	129.40.60.125
4	--			
5	tr0	sptok05	129.40.61.5	129.40.60.125
6	tr0	sptok06	129.40.61.6	129.40.60.125
7	tr0	sptok07	129.40.61.7	129.40.60.125
8	--			
9	tr0	sptok09	129.40.61.9	129.40.60.125
10	--			
11	--			
12	--			
13	tr0	sptok13	129.40.61.13	129.40.60.125
14	tr0	sptok14	129.40.61.14	129.40.60.125
15	--			
16	--			

Notes:

1. AIX is case sensitive. If name-to-address resolution is provided by DNS, NIS or some other means, the names in the SDR must match exactly. Otherwise, use lower case for the host name and addresses.
2. Wide nodes occupy two frame slots and use the *odd-numbered* slot.
3. High nodes occupy four frame slots (2 drawers) and use the lowest odd-numbered slot.
4. Use adapter physical location for p690 and p670 nodes only. All other nodes must use the adapter name.
5. For p690 or p670 server name, use the name you assigned to the server with the HMC user interface.

Completing the switch configuration worksheet

The advantage of an SP switch is that it has its own subnet. You need to plan this switch network whenever you plan to use any of the following:

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

Do you plan to enable ARP over the switch? If not, you need to derive the switch IP addresses from the address of the first node plus the switch port number.

A two-plane SP Switch2 system has two sets of switches and two adapters per node. The switch planes are disjoint – each is cabled exactly like a single plane and communication across the pair of planes is achieved via software striping. In a

two-plane SP Switch2 configuration, the first SP frame has a switch to be configured as plane 0 and the first expansion frame has a second switch to be configured as plane 1. Every node that will be connected to the switch has two adapters. The first switch adapter in the node is to be configured as **css0** and connected to **plane 0**. The second switch adapter in the node is **css1** and connected to **plane 1**.

Although each node has access to both switch planes, there is no physical connection between the planes. The switch subsystem provides an aggregate or multi-link IP device abstraction for the SP Switch2 networks. This pseudo-device driver enables you to address the two SP Switch2 planes with one IP address. The purpose for this virtual-device interface, the third IP address to be configured as **ml0**, is to allow IP messages to be transmitted in a more economical manner called striping. The striping technique provides a capability to transmit consecutive IP data across two fully operational adapters. It takes advantage of the combined bandwidth of both adapters. For example, when an IP message is sent between nodes and both nodes have access to both available switch networks, consecutive datagrams are sent in a pattern like *adapter0*, *adapter1*, *adapter0*, *adapter1*, ... Using **ml0** can ensure that a single failure in the SP Switch2 subsystem does not cause a complete outage to a node or other subsystem that is dependent on a switch. If a fault occurs between a node and one of the two switch planes, a transparent failover condition occurs using the **ml0** interface in order to access the remaining functional switch plane. For example, if *adapter0* malfunctions the resulting data flow would be *adapter1*, *adapter1*, *adapter1*, *adapter1*, ... To use this feature, plan to use the **spaggip** command or the SP Configuration Database Management SMIT tool after you configure **css0** and **css1** for each node that is to be attached to the SP Switch2.

Make copies of the switch configuration worksheet in Table 67 on page 289 and complete it for your system. Before you start, see “Switch port numbering” on page 125 and “IP address assignment” on page 127 for additional guidance. If the hypothetical `ABC Corporation` were to use the SP Switch2 with two switch planes, their completed worksheet might look like that in Table 17 on page 65.

Table 17. ABC Corporation's choices for the switch configuration worksheet

Switch configuration – Worksheet 10							
Frame number: 1	**Switch number:** 1	**css0 netmask:** 255.255.255.192		**css1 netmask:** 255.255.255.192		**ml0 netmask:** 255.255.255.192	
Slot number	**Switch port number**	**css0 hostname**	**css0 IP address**	**css1 hostname**	**css1 IP address**	**ml0 hostname**	**ml0 IP address**
1		spsw01	129.40.62.1	spsw101	129.40.63.1	spml01	129.40.64.1
2							
3		spsw03	129.40.62.3	spsw103	129.40.63.3	spml03	129.40.64.3
4							
5		spsw05	129.40.62.5	spsw105	129.40.63.5	spml05	129.40.64.5
6		spsw06	129.40.62.6	spsw106	129.40.63.6	spml06	129.40.64.6
7		spsw07	129.40.62.7	spsw107	129.40.63.7	spml07	129.40.64.7
8		spsw08	129.40.62.8	spsw108	129.40.63.8	spml08	129.40.64.8
9		spsw09	129.40.62.9	spsw109	129.40.63.9	spml09	129.40.64.9
10							
11							
12							
13		spsw13	129.40.62.13	spsw113	129.40.63.13	spml013	129.40.64.13
14		spsw14	129.40.62.14	spsw114	129.40.63.14	spml014	129.40.64.14
15							
16							
Note: Switch port number is necessary only with the SP Switch. Use of **css1** and **ml0** are options with the SP Switch2 only.							

When you plan to use an SP-attached server, you need to fill in the switch worksheet to set a switch port number even in a switchless SP system. This is because of the limited hardware interface to SP-attached servers. The SP functions cannot always derive all the information it needs like it can for SP nodes. During the SP installation and configuration process of your frames and nodes you will be asked to supply that number along with other values you are preparing during this planning phase.

Keep in mind that switch node numbers are used for nodes and SP-attached servers in all types of SP and clustered servers systems, including those systems with no switch. It is important to recognize that the algorithms for assigning switch node numbers to nodes and SP-attached servers differ depending on the type, or lack of, switch in the system. See "Switch port numbering" on page 125 for a discussion on how switch node numbers are assigned for each system and switch type.

This knowledge becomes critical when you are adding an SP Switch to a switchless system that has SP-attached servers. The algorithm for assigning switch node numbers changes, and the switch node numbers that you use for SP-attached servers in your switchless system might not be valid in the same system with an SP Switch. When first assigning switch node numbers to your SP-attached servers in a switchless SP system, consider if you might ever add an SP Switch to the system in the future. If you might, try to choose switch node numbers for the SP-attached servers that will be valid in both the switchless and SP Switch environments.

For example, if each existing SP frame were to have an SP Switch added to it, a reasonable number would be one that is available because a wide or high node is overlaying the associated slot. Fill out Worksheet 10 in Table 67 on page 289 to determine your switch node number allocations.

Question 9: Defining your system images

After determining the quantity and the type of nodes you need, you can decide which system image you want installed on which nodes. The system image is the collection of SP software components that is stored at a node. You can have a different system image on every node, the same system image on every node, or any combination in between. As you make this decision, there are performance and system management implications to consider.

The most significant implication is that if all the node images are the same, the installation and backup or restore functions are simpler. As discussed in the disk storage question, whether you install your applications on each node or on one node greatly affects the amount of disk storage space required for each node. While local node copies of applications offer faster performance, they require separate upgrades and system backups.

If you decide to have system partitions, you need to decide how many partitions you want and which nodes go with which partition. To fully understand partitioning, read Chapter 7, “Planning SP system partitions” on page 171 before you make any decisions about system partitions.

You can also have one or more alternate root volume groups defined on any of the nodes. This allows you to easily switch between multiple system images on the node. The node can assume any one of several different *personalities*. Remember that the alternate root volume groups on a node cannot share a disk. You must have at least one disk for each root volume group.

The AIX mksysb image (SPIMG) that comes with PSSP does not have the AIX trusted computing base option enabled. If you want to have the trusted computing base option enabled, you must create and use your own customized mksysb. Customization to enable that option can only be done during your installation phase before you create the mksysb. See the information about working with the AIX image in the book *PSSP: Installation and Migration Guide*.

Specifying more than one system image

Worksheets 11 and 12 help you lay out each system image that you want to define for the SP nodes. The control workstation is defined in the next section. Make a copy of both worksheets for each image or alternate image that you plan to have. Use Worksheet 11 for AIX and its options, IBM licensed programs, and other programs you choose to have in your system image. Use Worksheet 12 to select the optional components of PSSP that you choose to install. The ssp image is included for informational purposes. It contains all the base components of PSSP that are not optional.

IBM provides one or more minimal system images (SPIMG) with the PSSP installation media. It might or might not contain all the parts of AIX that you want installed on each node. For example, it does not contain AIX windows support. The *Read This First* document that you receive with PSSP will give you the latest information on the minimal image file sets. Make certain you use the listing for the PSSP level on your system.

When you come to the question about where you want to install the rootvg, you are deciding on which internal disk drive the SPIMG should be placed. You might be planning to install external disks, but IBM suggests that the SPIMG be placed on an internal drive.

To specify system images, the `ABC Corporation` filled out Worksheet 11, Table 18 on page 68. To specify PSSP components, they filled out Worksheet 12, Table 19 on page 69.

Table 18. ABC Corporation's system images

Specifying system images – Worksheet 11	
System image name	SPIMG1
AIX level	5.1
Partition number	1
Install on node numbers	1, 3, 5, 6, 7, 8, 9, 13, 14
Specify *internal* **disks where you want to install rootvg** disk 1	
Check here if you want only the SPIMG minimal image of AIX ____	
IBM licensed programs	
	AIX
	PSSP
	GPFS
	IBM C and C++ compilers
	LoadLeveler
	Parallel Environment
	ESSL (required for Parallel ESSL)
	Parallel ESSL
Additional AIX software	
	IBM HTTP Server
	IBM AIX Developer Kit, Java 2 Technology Edition
Other applications	
	NFS
	Open Secure Shell

Table 19. File set list for PSSP 3.5

PSSP 3.5 file sets – Worksheet 12		
System image name `spimg1`		
	File set	**Description**
x	**Image of AIX: spimg**	
	bos.obj.ssp.510	File with mksysb image of minimal AIX 5L 5.1 system with 32-bit kernel and JFS file system
	bos.obj.ssp.510_64	File with mksysb image of minimal AIX 5L 5.1 system with 64-bit kernel and JFS2 file system
x	**PSSP image: ssp** Base components of PSSP	
	ssp.authent	SP Kerberos V4 Server
	ssp.basic	SP System Support Package
	ssp.cediag	SP CE Diagnostics
	ssp.clients	SP Client Programs
	ssp.css	SP Communication Subsystem Package
	ssp.docs	SP man pages, PDF files, and HTML files
	ssp.gui	SP System Monitor Graphical User Interface
	ssp.ha_topsvcs.compat	Compatibility for ssp.ha and ssp.topsvcs clients
	ssp.perlpkg	SP PERL distribution package
	ssp.pman	SP Problem Management
	ssp.public	Public Code compressed tarfiles
	ssp.spmgr	SP Extension Node SNMP Manager
	ssp.st	Switch Table API package
	ssp.sysctl	SP sysctl package
	ssp.sysman	Optional System Management programs
	ssp.tecad	SP HA TEC Event Adapter package
	ssp.top	SP Communication Subsystem Topology package
	ssp.top.gui	SP System Partitioning Aid
	ssp.ucode	SP Supervisor microcode package
	PSSP image: ssp.hacws Optional component of PSSP	
	ssp.hacws	SP High Availability Control Workstation
x	**PSSP image: vsd** Components for managing IBM Virtual Shared Disks	
	vsd.cmi	IBM Virtual Shared Disk Centralized Management Interface (SMIT)
	vsd.hsd	Hashed Shared Disk data striping device driver
	vsd.rvsd.hc	IBM Recoverable Virtual Shared Disk Connection Manager
	vsd.rvsd.rvsdd	IBM Recoverable Virtual Shared Disk daemon
	vsd.rvsd.scripts	IBM Recoverable Virtual Shared Disk recovery scripts
	vsd.sysctl	IBM Virtual Shared Disk sysctl commands
	vsd.vsdd	IBM Virtual Shared Disk device driver
x	**PSSP image: ssp.vsdgui** IBM Virtual Shared Disk Perspectives GUI	
x	**PSSP image: ssp.resctr** Resource Center with links to online publications and other information.	

Table 19. File set list for PSSP 3.5 (continued)

PSSP 3.5 file sets – Worksheet 12		
System image name `spimg1`		
	File set	**Description**
	ssp.resctr.rte	Cluster Resource Center
x	**PSSP image: ssp.en_US.*** US English ISO8859-1, ISO8859-15	
	ssp.msg.en_US.authent	SP Authentication Server Messages
	ssp.msg.en_US.basic	SP System Support Messages
	ssp.msg.en_US.cediag	SP CE Diagnostic Messages
	ssp.msg.en_US.clients	SP Authenticated Client Messages
	ssp.msg.en_US.pman	SP Problem Management Messages
	ssp.msg.en_US.spmgr	SP Extension Node Manager Messages
	ssp.msg.en_US.sysctl	SP Package Messages
	ssp.msg.en_US.sysman	Optional System Management Messages
	PSSP image: ssp.En_US.* US English IBM-850	
	ssp.msg.En_US.authent	SP Authentication Server Messages
	ssp.msg.En_US.basic	SP System Support Messages
	ssp.msg.En_US.cediag	SP CE Diagnostic Messages
	ssp.msg.En_US.clients	SP Authenticated Client Messages
	ssp.msg.En_US.pman	SP Problem Management Messages
	ssp.msg.En_US.spmgr	SP Extension Node Manager Messages
	ssp.msg.En_US.sysctl	SP Package Messages
	ssp.msg.En_US.sysman	Optional System Management Messages
Note: You can choose to install complete images or only selected file sets. Keep in mind that some optional components require others. See the respective planning and migration sections in this book for dependencies. For information on which PSSP file sets to install on the control workstation and which to install on a node, see the books *PSSP: Installation and Migration Guide* and *PSSP: Managing Shared Disks*.		

Question 10: What do you need for your control workstation?

Consider the control workstation as a server of the PSSP and PSSP-related system management applications for the PSSP nodes. The PSSP nodes are clients of the control workstation server applications. The control workstation serves as a single point of control for these server applications that provide configuration data, security, hardware monitoring, diagnostics, and optionally, job scheduling data and a time source.

As in all servers, reliability of the servers affects availability of the clients. In this case, availability of the system as a whole is affected. See “Eliminating the control workstation as a single point of failure” on page 132 for more details and what happens when a single control workstation configuration has a control workstation failure. Make availability of resources a key consideration for configuring your control workstation.

IBM offers multiple ways to configure the control workstation and each way enables a different level of reliability for the control workstation and the SP system:

- Single control workstation without using AIX fault tolerant functions.

  This configuration has no redundant or backup functions. Its advantage is a configuration that costs less and is less complex. Its disadvantage is that a single hardware or software component failure can affect the availability of the SP system.
- Single control workstation that utilizes AIX fault tolerant functions.

  This configuration has some redundant or backup functions but does not protect against all failures. Its disadvantage is that most software failures and base system hardware failures are not protected against. Its advantage is that, although it is a slightly more costly configuration than the single control workstation without using AIX fault tolerant functions, it is still less costly than an HACWS configuration.
- An HACWS (High Availability Control Workstation) configuration.

  This configuration provides the most reliability for the control workstation and the SP system. All hardware and software components are redundant, which allows recovery from any single failure. Its disadvantage is that it costs more than the previous two control workstation configurations. Its advantage is that the SP system is better suited for production environments with this feature enabled.

  For more information on planning for HACWS, see Chapter 4, “Planning for a high availability control workstation” on page 129.

Software requirements for control workstations

The control workstation and some of its software might *not* be part of the SP package and must be ordered separately. Make sure you have ordered them in time to arrive when the rest of your SP does. To coordinate delivery of the SP and control workstation, your IBM representative must link the SP and control workstation orders with a System Order Number.

The control workstation is delivered with CE diagnostics preloaded. Do not begin your loading of software on the control workstation or nodes until after the CE turns the system over to you.

Required software

The following software is required on the control workstation:

- AIX 5L 5.1 server (5765-C34)
- PSSP 3.5 (5765-D51)
- At least one concurrent user license of VisualAge C++ 6.0 or later

  Note: PSSP does not support the incremental compiler and runtime libraries. It only supports the batch C and C++ compilers and runtime libraries that are included in this VisualAge package.

  Compilers are necessary for IBM service of PSSP. Also, without the compiler, dump diagnosis tools like **kdb** cannot work effectively. At least one concurrent user license is required for the SP system. Concurrent licensing allows the one license to float across the SP nodes and the control workstation. You can order the license as part of the SP system. It is not required directly on the control workstation if a license server for AIX for C and C++ exists some where in the network and the SP is included in the license server’s cell.

Optional software

Some PSSP components and related licensed programs are optional but when you do choose to use them, some must be installed on the control workstation while others can optionally be installed on the control workstation. While the *PSSP:*

Installation and Migration book has the complete list of what can be installed on the control workstation, certain choices have implications that you need to consider during planning. They are identified here as they relate to your plans for the control workstation.

The following are optional but if you use them, you must install them on your control workstation:

- HACWS

  See “Requirements for HACWS configurations” on page 136.
- IBM Virtual Shared Disk and Recoverable Virtual Shared Disk

  See Chapter 5, “Planning for IBM Virtual Shared Disks” on page 141. Also see the book *PSSP: Managing Shared Disks* to plan your IBM Virtual Shared Disks and which software to install on the control workstation and on the nodes.
- AIX DCE

  If you plan to use AIX DCE authentication methods as part of security on your SP, you must order and install the AIX DCE licensed program. See Chapter 6, “Planning for security” on page 145.
- PSSP graphical user interfaces

  These include the SP Perspectives and other optional interfaces such as:
  - Hardware Perspective
  - Event Perspective
  - IBM Virtual Shared Disk Perspective
  - System Partitioning Aid
  - Cluster Resource Center
  - SP Perspectives Launch Pad
  - Perspectives online help

  In addition to installing the correct file sets, you must establish proper authorizations for actions that these interfaces might issue to operate on nodes.

Service Director for RS/6000 is provided with the SP system, but you need to consider it during this planning phase if you want to use it. If you decide to use it, consider whether you want to install it on the control workstation or somewhere else. Service Director is a set of IBM software applications that monitor the health of your system. Service Director analyzes AIX error logs and runs diagnostics against those error logs. You can define which systems have the Service Director clients and servers and the level of error log forwarding or network access. See “Service Director for RS/6000” on page 192 for requirements and planning.

Hardware requirements for control workstations

To help you plan your control workstation hardware, this section includes the following:

- The default control workstation hardware configurations.
- The supported control workstations.
- The control workstation minimum hardware requirements.
- The HACWS minimum hardware requirements.

Default control workstations

There are several default control workstation configurations suggested for new customers. These can be upgraded as your needs increase. Your marketing

representative might specify alternative configurations for different environments. The available default CWS configurations are:

- IBM RS/6000 44P Model 170 (Table 20)
- IBM @server pSeries 610 Model 6C1 (rack: small Table 21, medium Table 24 on page 76)
- IBM @server pSeries 610 Model 6E1 (tower: small Table 23 on page 75, medium Table 24 on page 76)
- IBM @server pSeries 620 Model 6F1 (small Table 25 on page 77, medium Table 26 on page 77, large Table 27 on page 78)
- IBM @server pSeries 660 Model 6H1 (small Table 25 on page 77, medium Table 26 on page 77)

Table 20. 44P-170 Default control workstation configuration

Program	Description
7044-170	• RS/6000 44P Model 170 • 1.44MB 3.5″ Diskette Drive • Integrated Ultra SCSI Adapter • Integrated External Ultra2 SCSI Port • Integrated Ethernet Adapter
2624	32x Speed CD-ROM
2830	POWER GXT130P Graphics Adapter (PCI)
2909	18.2 GB 1″ Ultra SCSI Hard Disk Drive
2943	8 Port Async Adapter, EIA-232/422 (PCI)
2968	10/100 Mbps Ethernet PCI Adapter
3628	P260 Color Monitor, Stealth Black
4110	256 MB (2x128MB) SDRAM DIMMs
4223	Ethernet 1 Base2 Transceiver
4349	333MHz POWER3-II Processor Card
5005	Preinstall
6041	Mouse - Ivory
6159	12GB/24GB 4mm Tape Drive
8700	Quiet Touch Keyboard, Stealth Black - English (US)
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)

Table 21. 7028-6E1 small rack control workstation configuration

Program	Description
7028-6C1	• pSeries 610 Model 6C1 • 1.44MB 3.5″ Diskette Drive • Integrated Internal Ultra3 SCSI Adapter • Integrated External Ultra3 SCSI Port • Integrated Ethernet Adapter (2)
2633	48x Speed IDE CD-ROM
2848	POWER GXT135P Graphics Accelerator (PCI)
2943	8 Port Async Adapter, EIA-232/422 (PCI)

Table 21. 7028-6E1 small rack control workstation configuration (continued)

Program	Description
3263	18.2 GB Ultra3 SCSI Hot-Plug (2)
3628	P260/P275 Color Monitor, Stealth Black, and cable
4120	512 MB (2x256MB) SDRAM DIMMs
4242	6' 15-pin D-shell to 15-pin D-shell Extender Cable
4246	IDE 2-Drop Connector Cable
4248	SCSI Connector Cable and Repeater Card
4249	SCSI 3-Drop Connector Cable
4962	10/100 Mbps Ethernet PCI Adapter II
5005	Preinstall
5300	1-Way Power3 - II 375MHz Processor Card, 4MB L2 Cache (2)
6158	20 GB/40 GB 4mm Tape Drive
6277	Redundant Power Supply, 250 Watt AC, Hot Swap (2)
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3-Button Mouse - Stealth Black
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)
9911	Rack Power cord - ALL IBM Racks, 4m
7014-T00	Enterprise Rack --36 EIA
0197	Content :7 28-6C1 (5 EIA)
0198	Content :991 -P33 (2 EIA)
6088	Front Door for 1.8m Rack,Black
6098	Side Panel for 1.8m or 2.0m Rack, Black (2)
9171	Power Distribution Unit, Side-Mount, 1 Phase
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)
9910-P33	Powerware 5125 3 VA,2 -24 V - Rackmount
6630	Rail Kit
9010	Factory Install for P33//B38
9851	Power Cord set (L6-3)

Table 22. 7028-6E1 medium rack control workstation configuration

Program	Description
7028-6C1	• pSeries 610 Model 6C1 • 1.44MB 3.5″ Diskette Drive • Integrated Internal Ultra3 SCSI Adapter • Integrated External Ultra3 SCSI Port • Integrated Ethernet Adapter (2)
2633	48x Speed IDE CD-ROM
2848	POWER GXT135P Graphics Accelerator (PCI)
2943	8 Port Async Adapter, EIA-232/422 (PCI)

Table 22. 7028-6E1 medium rack control workstation configuration (continued)

Program	Description
3264	36.4 GB Ultra3 SCSI Hot-Plug (2)
3628	P260/P275 Color Monitor, Stealth Black, and cable
4121	1024 MB (2x512MB) SDRAM DIMMs
4242	6' 15-pin D-shell to 15-pin D-shell Extender Cable
4246	IDE 2-Drop Connector Cable
4248	SCSI Connector Cable and Repeater Card
4249	SCSI 3-Drop Connector Cable
4962	10/100 Mbps Ethernet PCI Adapter II
5005	Preinstall
5300	1-Way Power3 - II 375MHz Processor Card, 4MB L2 Cache (2)
6158	20 GB/40 GB 4mm Tape Drive
6277	Redundant Power Supply, 250 Watt AC, Hot Swap (2)
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3-Button Mouse - Stealth Black
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)
9911	Rack Power cord - ALL IBM Racks, 4m
7014-T00	Enterprise Rack --36 EIA
0197	Content :7 28-6C1 (5 EIA)
0198	Content :991 -P33 (2 EIA)
6088	Front Door for 1.8m Rack,Black
6098	Side Panel for 1.8m or 2.0m Rack, Black (2)
9171	Power Distribution Unit, Side-Mount, 1 Phase
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)
9910-P33	Powerware 5125 3 VA,2 -24 V - Rackmount
6630	Rail Kit
9010	Factory Install for P33//B38
9851	Power Cord set (L6-3)

Table 23. 7028-6E1 small tower control workstation configuration

Program	Description
7028-6E1	• pSeries 610 Model 6E1 • 1.44MB 3.5″ Diskette Drive • Integrated Internal Ultra3 SCSI Adapter • Integrated External Ultra3 SCSI Port • Integrated Ethernet Adapter (2)
2633	48x Speed IDE CD-ROM
2848	POWER GXT135P Graphics Accelerator (PCI)
2943	8 Port Async Adapter, EIA-232/422 (PCI)

Table 23. 7028-6E1 small tower control workstation configuration (continued)

Program	Description
3263	18.2 GB Ultra3 SCSI Hot-Plug (2)
3628	P260/P275 Color Monitor, Stealth Black, and cable
4120	512 MB (2x256MB) SDRAM DIMMs
4246	IDE 2-Drop Connector Cable
4248	SCSI Connector Cable and Repeater Card
4249	SCSI 3-Drop Connector Cable
4962	10/100 Mbps Ethernet PCI Adapter II
5005	Preinstall
5300	1-Way Power3 - II 375MHz Processor Card, 4MB L2 Cache
6158	20 GB/40 GB 4mm Tape Drive
6277	Redundant Power Supply, 250 Watt AC, Hot Swap (2)
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3-Button Mouse - Stealth Black
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)

Table 24. 7028-6E1 medium tower control workstation configuration

Program	Description
7028-6E1	• pSeries 610 Model 6E1 • 1.44MB 3.5″ Diskette Drive • Integrated Internal Ultra3 SCSI Adapter • Integrated External Ultra3 SCSI Port • Integrated Ethernet Adapter (2)
2633	48x Speed IDE CD-ROM
2848	POWER GXT135P Graphics Accelerator (PCI)
2943	8 Port Async Adapter, EIA-232/422 (PCI)
3263	18.2 GB Ultra3 SCSI Hot-Plug (2)
3264	36.4 GB Ultra3 SCSI Hot-Plug (2)
3628	P260/P275 Color Monitor, Stealth Black, and cable
4120	512 MB (2x256MB) SDRAM DIMMs
4246	IDE 2-Drop Connector Cable
4248	SCSI Connector Cable and Repeater Card
4249	SCSI 3-Drop Connector Cable
4962	10/100 Mbps Ethernet PCI Adapter II
5005	Preinstall
5300	1-Way Power3 - II 375MHz Processor Card, 4MB L2 Cache (2)
6158	20 GB/40 GB 4mm Tape Drive
6277	Redundant Power Supply, 250 Watt AC, Hot Swap (2)
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3-Button Mouse - Stealth Black

Table 24. 7028-6E1 medium tower control workstation configuration (continued)

Program	Description
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)

Table 25. p620 Model 6F1 small control workstation configuration

Program	Description
7025-6F1	• pSeries 620 Model 6F1 • 1.44MB 3.5″ Diskette Drive • Integrated SCSI-2 F/W Adapter • Integrated Ultra2 SCSI Adapter 1 • Integrated Ethernet Adapter
2624	32x Speed CD-ROM
2830	POWER GXT130P Graphics Adapter (PCI)
2943	8 Port Async Adapter, EIA-232/422 (PCI)
2968	10/100 Mbps Ethernet PCI Adapter
3153	18.2 GB 10K RPM 1″ Ultra3 SCSI 16-bit Disk (2)
3628	P260 Color Monitor, Stealth Black
3752	Service Pack
4110	256 MB (2x128MB) SDRAM DIMMs
5005	Preinstall
5211	1-Way RS64 IV 600 Mhz Processor Card, 2MB L2 Cache
5663	SCSI Hot Swap
6159	12GB/24GB 4mm Tape Drive
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3-Button Mouse - Stealth Black
9300	Language - English (US)
9800	Power Cord - US/Canada (125V, 15A)

Table 26. p620 Model 6F1 medium control workstation configuration

Program	Description
7025-6F1	• pSeries 620 Model 6F1 • 1.44MB 3.5″ Diskette Drive • Integrated SCSI-2 F/W Adapter • Integrated Ultra2 SCSI Adapter 1 • Integrated Ethernet Adapter
2624	32x Speed CD-ROM
2830	POWER GXT130P Graphics Adapter (PCI)
2943	8-Port Async Adapter, EIA-232/422 (PCI)
2968	10/100 Mbps Ethernet PCI Adapter (3)
3109	SCSI external port to SCSI internal 6-pack cable assembly
3153	18.2 GB 10K RPM 1″ Ultra3 SCSI 16-bit Disk (4)
3628	P260 Color Monitor, Stealth Black

Table 26. p620 Model 6F1 medium control workstation configuration (continued)

Program	Description
3752	Service Pack
4075	Memory Board, 16-position
4119	512 MB (2x256MB) SDRAM DIMMs (2)
5005	Preinstall
5202	2-Way RS64 III 450 MHz Processor Card, 4MB L2 Cache
6159	12GB/24GB 4mm Tape Drive
6553	SCSI Hot Swap 6-Pack
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3-button mouse - Stealth Black
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)

Table 27. p620 Model 6F1 large control workstation configuration

Program	Description
7025-6F1	• pSeries 620 Model 6F1 • 1.44MB 3.5″ Diskette Drive • Integrated SCSI-2 F/W Adapter • Integrated Ultra2 SCSI Adapter 1 • Integrated Ethernet Adapter
2624	32x Speed CD-ROM
2830	POWER GXT130P Graphics Adapter (PCI)
2944	128-Port Async Controller (PCI)
2968	10/100 Mbps Ethernet PCI Adapter (3)
3109	SCSI external port to SCSI internal 6-pack cable assembly
3153	18.2 GB 10K RPM 1″ Ultra3 SCSI 16-bit Disk (6)
3628	P260 Color Monitor, Stealth Black
3752	Service Pack
4075	Memory Board, 16-position
4119	512MB (2x256MB) SDRAM DIMMs (2)
5005	Preinstall
5204	4-Way RS64 III 450 MHz Processor Card, 4MB L2 Cache
6158	12GB/24GB 4mm Tape Drive
6553	SCSI Hot Swap 6-Pack
8131	4.5mm Controller Cable
8133	RJ-45 to DB-25 Converter Cables (4 cables per set)
8137	Enhanced Async Node 16-Port EIA-232
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3-button mouse - Stealth Black
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)

Table 28. p660 Model 6H1 small control workstation configuration

Program	Description
7026-6H1	• pSeries 660 Model 6H1 • 1.44MB 3.5″ Diskette Drive • Integrated SCSI-2 F/W Adapter • Integrated Ultra2 SCSI Adapter 1 • Integrated Ethernet Adapter
2624	32x Speed CD-ROM
2830	POWER GXT130P Graphics Adapter (PCI)
2943	8 Port Async Adapter, EIA-232/422 (PCI)
2968	10/100 Mbps Ethernet PCI Adapter
3102	18.2 GB 10K RPM Ultra SCSI Disk Drive (2)
3142	Remote I/O Cable -3m (2)
3628	P260 Color Monitor, Stealth Black
4110	256 MB (2x128MB)SDRAM DIMMs
5005	Preinstall
5211	1-Way RS64 IV 600 Mhz Processor Card, 2MB L2 Cache
5992	System Control and Initialization Cable
6132	CEC to Primary I/O Drawer Power Control Cable, 3m
6159	12GB/24GB 4mm Tape Drive
6324	Primary I/O Drawer, 5 EIA
6540	IPL Disk Mounts, Cables, Terminator
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3-Button Mouse -Stealth Black
9172	AC Power Specify
9300	Language - English (US)
9800	Power Cord -US/Canada (125V,15A)
7014-T00	Enterprise Rack - 36 EIA
0176	Content: FC 6324 (5 EIA)
0188	Content: 7026-6H1 (5 EIA)
6088	Front Door for 1.8m Rack, Black
6098	Side Panel for 1.8 or 2.0m Rack, Black (2)
9171	Power Distribution Unit, Side-Mount, 1 Phase
9300	Language - English (US)
9800	Rack Power Cord - US/Canada
7014-T00	Enterprise Rack - 36 EIA
0183	Content: 9910-A30 (5 EIA)
6088	Front Door for 1.8m Rack, Black
6098	Side Panel for 1.8 or 2.0m Rack, Black (2)
9171	Power Distribution Unit, Side-Mount, 1 Phase
9300	Language - English (US)
9800	Rack Power Cord - US/Canada

Table 28. p660 Model 6H1 small control workstation configuration (continued)

Program	Description
9910-A30	APC 5000VA Smart-UPS Rack-Mount

Table 29. p660 Model 6H1 medium control workstation configuration

Program	Description
7026-6H1	• pSeries 660 Model 6H1 • 1.44MB 3.5″ Diskette Drive • Integrated SCSI-2 F/W Adapter • Integrated Ultra2 SCSI Adapter 1 • Integrated Ethernet Adapter
2624	32x Speed CD-ROM
2830	POWER GXT130P Graphics Adapter (PCI)
2943	8 Port Async Adapter, EIA-232/422 (PCI)
2968	10/100 Mbps Ethernet PCI Adapter (3)
3102	18.2 GB 10K RPM Ultra SCSI Disk Drive (2)
3142	Remote I/O Cable -3m (2)
3628	P260 Color Monitor, Stealth Black
4075	Memory Board, 16-position
4119	512 MB (2x256MB)SDRAM DIMMs
5005	Preinstall
5212	2-Way RS64 IV 600 Mhz Processor Card, 4MB L2 Cache
5992	System Control and Initialization Cable
6132	CEC to Primary I/O Drawer Power Control Cable, 3m
6159	12GB/24GB 4mm Tape Drive
6230	Advanced SerialRAID Plus Adapter
6324	Primary I/O Drawer, 5 EIA
6540	IPL Disk Mounts, Cables, Terminator
8700	Quiet Touch Keyboard, Stealth Black - English (US)
8741	3 button mouse - Stealth Black
9172	AC Power Specify
9300	Language - English (US)
9800	Power cord - US/Canada (125V, 15A)
7014-T00	Enterprise Rack - 36 EIA
0176	Content: FC 6324 (5 EIA)
0188	Content: 7026-6H1 (5 EIA)
6088	Front Door for 1.8m Rack, Black
6098	Side Panel for 1.8 or 2.0m Rack, Black (2)
9171	Power Distribution Unit, Side-Mount, 1 Phase
9300	Language - English (US)
9800	Rack Power Cord - US/Canada
7014-T00	Enterprise Rack - 36 EIA

Table 29. p660 Model 6H1 medium control workstation configuration (continued)

Program	Description
0183	Content: 9910-A30 (5 EIA)
6088	Front Door for 1.8m Rack, Black
6098	Side Panel for 1.8 or 2.0m Rack, Black (2)
9171	Power Distribution Unit, Side-Mount, 1 Phase
9300	Language - English (US)
9800	Rack Power Cord - US/Canada
7133-D40	Advanced SSA Disk Subsystem (Rack-Mounted)
0550	Hungary Manufacturing Ship Direct to Customer
8022	50/60Hz AC, 300 VDC Power Supplies
8031	Raven Black Drawer Cover
8518	One 10K/18.2GB Advanced Disk Drive Module (4)
8801	1m Advanced SSA Cable
9300	Language - English (US)
9910-A30	APC 5000VA Smart-UPS Rack-Mount

Supported control workstations

Running the PSSP software requires an IBM @server pSeries or RS/6000 workstation as a point-of-control for managing, monitoring, and maintaining the nodes in a system of clustered servers or in a system of SP frames and processor nodes. The control workstation you supply connects to each frame and to the SP Ethernet. Table 30 lists the control workstations you might already have and can still use to run PSSP 3.5 and it includes those you can currently order from IBM. Pay particular attention to the notes. Some indicate certain conditions for support. Only those so noted in the PCI category are currently available from IBM.

Table 30. Supported control workstations

MCA	• 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 (Note 1 on page 82) • RS/6000 7015 Models 97B, 970, 98B, 980, 990, R30, R40, and R50 (Notes 1 on page 82, 2 on page 82) • RS/6000 7030 Models 3AT, 3BT, 3CT
PCI	• IBM @server pSeries 610 Model 6C1 – 7028-6C1 rack (Note 7 on page 82) • IBM @server pSeries 610 Model 6E1 – 7028-6E1 tower (Note 7 on page 82) • IBM @server pSeries 620 Model 6F1 – 7025-6F1 (Note 7 on page 82) • IBM @server pSeries 660 Model 6H1 – 7026-6H1 (Note 7 on page 82) • RS/6000 7024 Models E20 and E30 • RS/6000 7025 Model F30 (Note 3 on page 82) • RS/6000 7025 Model F40 (Notes 4 on page 82, 5 on page 82) • RS/6000 7025 Model F50 and F80 (Notes 4 on page 82) • RS/6000 7026 Models H10 and H50 (Notes 4 on page 82, 5 on page 82) • RS/6000 7026 Model H80 (Notes 4 on page 82) • RS/6000 7043 Models 140, and 240 (Notes 4 on page 82, 6 on page 82) • RS/6000 7044 Model 170 – 44P-170 (Note 7 on page 82)

Table 30. Supported control workstations (continued)

Notes:

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.
2. Installed in either the 7015-99X or 7015-R00 rack.
3. On systems introduced since PSSP 2.4, either the 8-port (#2493) or 128-port (#2944) PCI bus asynchronous adapter should be used for frame controller connections. IBM strongly suggests you use the support processor option (#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.
4. The native RS-232 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 (#2943) or the 128-port Asynchronous Controller (#2944) are the only RS-232 adapters that are supported.
5. IBM strongly suggests you use the support processor option (#1001).
6. 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. The 7043 **cannot** be used for SP systems with SP-attached servers.
7. This is currently available from IBM.

See Chapter 4, “Planning for a high availability control workstation” on page 129 for more planning information about the control workstation.

Control workstation minimum hardware requirements

The minimum requirements for the control workstation are:

- At least 128MB of main memory. An extra 64MB of memory should be added for each additional system partition. For SP systems with more than 80 nodes, 256MB is required, 512MB of memory is suggested. For systems containing p670 and p690 servers, a minimum of 2GB is suggested.
- At least 9GB of disk storage. If the SP is going to use an HACWS configuration, you can configure 9GB of disk storage in the rootvg volume group and 9GB for the spdata in an external volume group.

  Because the control workstation is used as a Network Installation Manager (NIM) server, the number of unique file sets required for all the nodes in the SP system might be larger than a normal single system. You should plan to reserve 6GB of disk storage for the file sets, and 2GB for the operating system. This will allow adequate space for future maintenance, system mksysb images and growth. Keep in mind that if you have nodes at different levels of PSSP or AIX, each node requires its own directory for source which will take up extra space.

  A good rule of thumb to use for disk planning for a production system is 4GB for the rootvg to accommodate additional logging and **/tmp** space, plus 4GB for each AIX release and modification level and for licensed programs in **lppsource** files. Add more disk space for mksysb images for the nodes.

  If you plan on using rootvg mirroring, for one mirror double the number of physical disks you estimated so far. Triple the estimate for two mirrors.
- You might plan to use the control workstation to initially build your own customized mksysb AIX image. One reason might be that you want to enable the AIX trusted computing base option. If you plan to create a customized mksysb for any reason, you must have at least two physical disks in the control workstation, one for the alternate volume group (not rootvg) that the mksysb command will use for the output.

- Physically installed to within 12 meters of RS-232 cable to each SP frame or IBM @server pSeries or RS/6000 server. See the book *IBM @server Cluster 1600 Hardware Planning, Service and Installation* for instructions on how to connect to each of the supported servers.
- Equipped with the following I/O devices and adapters:
  - A 3.5 inch diskette drive
  - Four or eight millimeter (or equivalent) tape drive
  - A SCSI CD-ROM device
  - One RS-232 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.
  - An appropriate network adapter for your external communication network. The adapter does not have to be on the control workstation. If it is not on the control workstation, the SP Ethernet must extend to another host that is not part of the SP system. A backup control workstation does not satisfy this requirement. This additional connection is used to access the control workstation from the network when the SP nodes are down.
  - Ethernet adapters for connection to the PSSP admin LAN

    The number of Ethernet adapters required depends completely on the Ethernet topology you use on your SP system. The following types of Ethernet adapters can be used:

    - Ethernet adapters with thin BNC

      Each Ethernet adapter of this type can have only 30 network stations on a given Ethernet cable. The control workstation and any routers are included in the 30 stations.
    - Ethernet adapters with twisted pair (RJ45/AUI). A network hub or switch is required.
    - 10/100 Mbps Ethernet adapters. A network hub or switch is required.
  - If your SP has more than a total of two standard SP frames and non-LPAR-capable attached servers, or you are considering a control workstation that has the PCI type of bus, you need an asynchronous adapter card to provide ports for the SP frames.

HACWS minimum hardware requirements

If you plan to use HACWS, the following are required in addition to the previous requirements:

- You need two IBM @server pSeries or RS/6000 workstations that are supported by PSSP 3.5 and are also supported by the version of HACMP that can run with PSSP 3.5 and AIX 5L 5.1.

  Each of the workstations must have the same set of I/O required for control workstations as previously listed. They can be different models but the tty configuration **must** be exactly the same on each control workstation. The disks should be of the same type and configured the same way on both control workstations to allow the `hdisk`*x* numbers to be consistent between the two control workstations.
- You need external disk storage that is supported by HACMP and the control workstation being used.

Two external disk controllers and mirrored disks are strongly suggested but not required. If a single external disk controller is used the control workstation single point of failure has not been eliminated but moved to the disk subsystem.

- You need the HACWS connectivity feature #1245 on each SP frame.
- If target mode SCSI is not being used for the HACMP communication, you need an additional RS-232 connection for HACMP communication.

Hardware controller interface planning

Each SP frame uses an RS-232 line to connect to a serial port on the control workstation. Supported servers, whether SP-attached or in a cluster configuration, connect directly to the control workstation in one way or another depending on the hardware characteristics.

Be prepared to identify one of the following hardware protocol values during configuration for the nodes in a frame. Table 31 shows the values to specify for each type of frame or server connected to the control workstation.

Table 31. Hardware protocol values

Hardware protocol	Frame or server
SP	SP frame. *This is the default.*
SAMI	Servers connected by two RS-232 lines to the CWS: @server pSeries 680, RS/6000 S80, S7A, and S70
CSP	Servers with Common Service Processor (CSP) connected by one RS-232 line to the CWS: RS/6000 H80, M80, and @server pSeries 660 (6H0, 6H1, 6M1)
HMC	Servers with a Hardware Management Console (HMC) use a trusted network connection to the CWS: the IBM @server pSeries 690 and 670. A trusted network can be the SP Ethernet admin LAN or an HMC trusted network. See “Considering an HMC trusted network” on page 113.

HACWS does not automatically support SP-attached servers. However, with specialized experience and manual attention, you might be able to have it work for you in limited cases. If you want to use an SP-attached server in an HACWS configuration, see Chapter 4, “Planning for a high availability control workstation” on page 129, particularly “Limits and restrictions” on page 134, for other considerations.

Note: IBM strongly suggests that you use asynchronous adapter cards instead of the native RS-232 ports on the system units. Several IBM @server pSeries or RS/6000 systems do not support the use of the native serial ports for the frame controller connections.

On HACWS configurations one of these native serial ports can be used for HACMP communication.

The native serial ports can be used for remote service by Service Director for RS/6000.

See the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for additional hardware planning and how to connect nodes.

Completing the control workstation worksheets

You need to complete one set of worksheets for each control workstation you will configure. The `ABC Corporation` completed Worksheet 13 in Table 32 on page 86, and Worksheet 15, in Table 33 on page 87.

Complete your copies of Worksheet 13, ″Control workstation″ in Table 70 on page 293, Worksheet 14, ″Select a time zone″ in Table 71 on page 294, and Worksheet 15, “Control workstation connections” in Table 72 on page 295.

Table 32. ABC Corporations's Cluster 1600 managed by PSSP control workstation plan

Control workstation – Worksheet 13	
Control workstation image:	
Control Workstation Name	cws01
Model	7025-6F1 (p620 Model 6F1 medium)
Install rootvg on disk	dsk01
Disk Space	18.2GB x 4
Memory Size	512MB
Hardware options and adapters:	
Type	***Quantity***
ATM	
Ethernet	1
FDDI	
Token ring (speed 16Mbps)	1
Multiport Serial Adapters	
8 mm tape drive	1
CD-ROM	1
IBM licensed programs:	
AIX	
PSSP	
IBM C and C++ Compilers	
LoadLeveler	
Other applications:	
NFS	

Table 33. ABC Corporation's control workstation connections

Control workstation connections – Worksheet 15							
Company name: ABC Corporation				**Date:** October 18, 2002			
System name: spsystem1				**Control workstation name:** cws01			
Frame hardware control connections				**Control workstation network connections**			
Frame number	**Serial port for RS-232 control line**	**tty device**		**Adapter**	**Hostname**	**IP address**	**Netmask**
1	s1	tty0		en0	spcwsen0	129.40.60.125	255.255.255.192
				tr0	spcwstr0	129.40.60.1	255.255.255.192

Note:

Column 2 applies to nodes with CSP or SAMI hardware protocol. Record one serial port for nodes with CSP and two for nodes with SAMI.

Nodes with the HMC protocol require a network connection only. Columns 2 and 3 do not apply.

Chapter 3. Defining the configuration that fits your needs

This chapter provides the information you need to plan for configuring your system before it gets installed. The following sections discuss some of the system planning and configuration issues you need to consider and decisions you need to make. Planning and configuring your Cluster 1600 system managed by PSSP system software has an impact on the SP site plan options you select. Your choice of software applications and features drives your DASD usage. The physical layout is addressed to the extent of which layouts are supported and which rules apply that relate to the software environment.

Note: Some helpful hardware information is included only as it relates to what the software supports. For complete hardware requirements and dependencies regarding adapters and physical networks, see the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*.

With this chapter, you will complete the worksheets in Table 73 on page 296, Table 69 on page 291, and Figure 58 on page 285. You will use the completed worksheets to:

- Review your installation plan with your IBM hardware and software installation team.
- Help you configure your system during the installation.

Make copies of the worksheets before you begin.

Planning your site environment

You set your site environment configuration options during the install process by using SMIT or the **spsitenv** command on the control workstation. SMIT is the System Management Interface Tool supplied as part of the PSSP software package.

If you plan to use DCE security services, this is a good time to consider if you want control over which interfaces DCE uses for remote processing. You can set AIX *environment variables* to control that for each process or for all processes. Chapter 6, “Planning for security” has more information in “Considering to exclude network interfaces from DCE Remote Procedure Call binding” on page 157.

The installation and configuration scripts read what you enter and customize your system configuration according to your choices. The entries you make on your copy of Table 73 on page 296 are the entries you’ll make during installation of PSSP on the control workstation. See “Using the Site Environment Worksheet” on page 90.

Site environment choices you can make include the following:

- The name of the default network install image. (See “Understanding network install image choices” on page 90.)
- Your method of time service, the name of your time servers, and the version of NTP in use. (See “Understanding time service choices – Network Time Protocol (NTP)” on page 91.)
- Whether you want to have the PSSP services configure and manage the Automounter. (See “Understanding user directory mounting choices – AIX Automounter” on page 92.)
- Whether you want to use the SP User Account Management facility. (See “Understanding user account management choices” on page 93.)

- Whether you want to use PSSP file collections and where. (See “Understanding system file management choices – file collections” on page 94.)
- Whether you want to use SP Accounting. (See “Understanding accounting choices” on page 95.)
- Whether you use the default lppsource directory for AIX file sets. (If you change the directory name you must also change it in the Site Environment Data label. See “Understanding lppsource directory name choices” on page 95.)
- Whether you want to have restricted root access and a secure remote command process. (See “Understanding remote command choices” on page 96.)
- Whether you want to disable system partitioning on your switchless SP system. See“Understanding partitionability choices” on page 97

You can easily change the choices discussed in the following sections any time after the installation. If you are not sure about any of these options, you can use the defaults and change your selections later.

Using the Site Environment Worksheet

The following sections help you make decisions about your site environment. These sections are listed in the same order as the items in the **Site Environment Worksheet** on page 296. A brief description of the function of each area along with a discussion of the alternatives should give you enough information to make your choices and fill out the worksheet. More information about these and other system administration functions is in the book *PSSP: Administration Guide*.

Keep in mind, the defaults are designed to provide an operational system and they might be just right for you. You can change them later, if you find it necessary.

Understanding network install image choices

The *install_image* attribute lets you specify the name of the default network install image to be used for any PSSP node when the install image field is not set. The default shipped with the system is **bos.obj.ssp.510**. A second image, **bos.obj.ssp.510_64**, is available for use with nodes capable of running an AIX 64-bit kernel. Specify **bos.obj.ssp.510_64** as the default only if all your nodes are 64-bit capable.

Note: Some nodes and adapters are not supported by the 64-bit kernel. Refer to the following Web site to view the various AIX processors that support the 64-bit kernel:

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

then select ″AIX 5L release notes″ and select the HTML option for ″AIX 5L for POWER Version 5.1 Release Notes.″ Using the ″Contents,″ search for the 64-bit Kernel entry under the ″Base Operating System (BOS)″ heading.

If you configure one or more nodes of your SP system as boot-install servers, each will act as an intermediate repository for a network install image of the AIX operating system. This network install image is a single file that occupies significant space on the file system of the boot-install server on which it resides. It uses approximately 25 MB per lppsource level.

You can reclaim this disk space by setting the *remove_image* attribute to **true**, which deletes this network install image after all new installation processes

complete. Alternatively, you can retain the image to improve the speed of a successive install that uses this same image.

Note: This does not apply to the control workstation. The network install images are never automatically deleted from the control workstation.

Site Environment Worksheet entries

You can set two attributes for these options. *install_image* lets you set the name of the default image. *remove_image* specifies what to do with the image after all installations are complete.

Table 34. Network install image choices

	Worksheet entries to be filled in...	
To do this...	*install_image*	*remove_image*
Use **bos.obj.ssp.510** as the default	**bos.obj.ssp.510**	
Use **bos.obj.ssp.510_64** as the default	**bos.obj.ssp.510_64**	
Remove the install image after all installs have completed		**true**
Do not remove the install image		**false** (default)
Note: • Change default attribute values to suit your environment. • Blank entries imply that you make no substitutions for these values.		

Understanding time service choices – Network Time Protocol (NTP)

A Cluster 1600 system managed by PSSP requires that time be synchronized on the control workstation and PSSP nodes. Your options are the following:

- If you already have an established NTP server, you can use it.
- You can choose an NTP server from the Internet.
- You can use the NTP server that comes with AIX by default.
- You can choose not to use NTP at all, relying on another method at your site.

Notes:

1. If you choose not to use NTP, you must have another way to manage clock synchronization.
2. Do not use the control workstation or backup control workstation to be the time master server.

See the book *PSSP: Administration Guide* for managing the NTP server.

High Availability Control Workstation considerations

If you install the High Availability Control Workstation software with a second control workstation and you select **timemaster** as your site's existing NTP time server, both control workstations must use the site time server. If you use the Internet configuration, both control workstations need access to the Internet.

Site Environment Worksheet entries

There are three attributes to set for NTP. *ntp_version* defaults to **3** (the version shipped with AIX 4.2 or later). If your installation is using an earlier version of NTP, change this value to the number for the version in use. The other two attributes are described in Table 35 on page 92.

Table 35. Time service choices

	Worksheet entries to be filled in...	
To do this...	*ntp_config*	*ntp_server*
Use your site's existing NTP time server to synchronize the SP system clocks.	**timemaster**	*hostname* of your current NTP time server
Use an NTP time service from the Internet to synchronize the SP system clocks.	**internet**	*hostnames* of time servers on the Internet*
Run NTP locally on the SP to generate a consensus time.	**consensus** (default)	
Do not use NTP on the SP; instead, use some other method to synchronize system clocks.	**none**	
Note: • Change default attribute values to suit your environment. • Blank entries imply that you make no substitutions for these values. • * See the **README.public** file in the **/usr/lpp/ssp/public** directory for information on Internet time servers.		

Understanding user directory mounting choices – AIX Automounter

An automounter is an automatic file system that dynamically mounts users' home directories and other file systems when a user accesses the files and unmounts them after a specified period of inactivity. The automounter manages directories specifically defined in the automounter map files. Using an automounter will minimize system hangs and, through mapping, will also provide a method of sharing common file system mount information across many systems.

Automounter daemons run independently on the control workstation and on every node in the SP system. Since these daemons run independently, you will be able to simultaneously run different automounters, if you have different levels of PSSP on your system. Also, a system configuration variable gives you the option of turning off the automount daemons on all or none of the system partitions.

Automounter considerations

Booting the SP nodes invokes a similar process creating node directories and logs. Map files are downloaded from the control workstation to the nodes during node boot. Once it has been created, the user directory automounter map is updated automatically as users are added and deleted from the system provided you have configured SP User Management Services on the control workstation.

The AIX automounter uses NFS (Network File Systems) to mount or AIX to link directories. Nodes running PSSP use the AIX automounter by default. As an alternative to the AIX automounter, you can provide your own technique for directory access.

One method of directory access would be to leave the SP automounter support turned on and replace the default SP function with support you provide for using your own automounter. You would do this using a set of user customization scripts that would be recognized by the SP. Another method would be setting the configuration variable so that the automounter daemon is off for the entire system. You would then have to provide some other means for users to access their home

directories. Alternatively, since the use of an automounter is optional, you might choose to not use an automounter on your SP system.

See the chapter on managing Automount in the book *PSSP: Administration Guide*.

Site Environment Worksheet entries

Only one attribute applies to the Automount option.

Table 36. User directory mounting choices - system automounter support

	Worksheet entry to be filled in
To do this...	*amd_config*
Use **AIX Automounter** supplied with PSSP	**true** (default)
Use some other means of mounting user directories to the SP	**false**
Note: • Change default attribute values to suit your environment. • Blank entries imply that you make no substitutions for these values.	

Understanding user account management choices

The SP user account management facility is designed to fit in with your current computing environment. If you already have procedures in place for managing user accounts, you can configure the SP system to use them. Alternatively, you can use the set of commands and tools provided with the SP for this purpose. The SP uses a single **/etc/passwd** file replicated across all nodes in the SP system using the SP file collection facilities. If you are using Network Information Service (NIS), these commands will use the NIS function. A set of customer commands is provided to interface to the NIS function.

These options are offered to help you manage user accounts. They involve passwords and directory paths. Read the brief descriptions that follow and record your choices on the Site Environment Worksheet.

Password management

The *passwd_file* lets you specify the name of your password file.

The default name of the password file is **/etc/passwd**.

The *passwd_file_loc* attribute should contain the host name of the machine where you maintain your password file. This defaults to your control workstation. The value of the *passwd_file_loc* cannot be one of the nodes in the SP system.

Home directories

Specify a default location for user home directories in the *homedir_server* attribute. If you are using Amd, the user management commands will use this host name when building Amd maps. If you do not specify a default, the user management commands assume the host on which you enter the commands. You can override this value when adding or modifying a user account with the **spmkuser** and **spchuser** commands.

Use the *homedir_path* attribute to specify the path of user home directories. The default base path for user home directories is **/home/***localHostname*. Change this value if you wish to set a different path as the default for your site. You can also override the default path with the **home** attribute on the **spmkuser** and **spchuser** commands.

See the chapter on managing accounts in the book *PSSP: Administration Guide*.

Site Environment Worksheet entries

Five attributes apply to PSSP User Management, but four of them are used only if you set *usermgmt_config* to **true**.

Table 37. PSSP user account management choices

	Worksheet entries to be filled in...				
To do this...	*usermgmt_config*	*passwd_file_loc*	*passwd_file*	*homedir_server*	*homedir_path*
Do not use PSSP user account management	**false**				
Use PSSP user account management	**true** (default)	`ctl wkstn` (default)	`/etc/passwd` (default)	`ctl wkstn` (default)	`/home/` (*home directory name*)
Note: • Change default attribute values to suit your environment. • Blank entries imply that you make no substitutions for these values.					

Understanding system file management choices – file collections

The PSSP file collection component simplifies the task of maintaining duplicate files across the nodes of the SP system. File collections provide a single point of control for maintaining a consistent version of one or more files across the entire system. You can make changes to the files in one place and the system replicates the updates on the other copies.

The files that are required on the control workstation, the file servers and the SP nodes are grouped into file collections. A file collection consists of a directory of files which includes special master files that define and control the collection.

The file collection structure is created along with the initial installation and configuration of your SP system. You must decide which files to specify for replication.

See the chapter on managing file collections in the book *PSSP: Administration Guide*.

Site Environment Worksheet entries

The SP system gives you the option of using file collections or not using them. If you choose to use them you must specify a unique (unused) user ID for the file collection daemon along with a unique (unused) port through which to communicate.

Table 38. System file management choices

	Worksheet entries to be filled in...		
To do this...	*filecoll_config*	*supman_uid*	*supfilesrv_port*
Do not use PSSP file collections	**false**		
Use PSSP file collections	**true** (default)	**102** (default for **supman**)	**8431** (default port number)

Understanding accounting choices

The accounting utility lets you collect and report on individual and group use of the SP system. This accounting information can be used to bill users of the system resources or monitor selected aspects of the system's operation.

Because the level of hardware resources is probably not distributed evenly across your SP system, you might want to charge different rates for different nodes. SP accounting lets you define *classes* or groups of nodes for which accounting data is merged, providing a single report for the nodes in that class. In addition, you can suppress or disable the collection of accounting data. Individual nodes within a class can be enabled or disabled for accounting.

Site Environment Worksheet entries

The attributes in the following table apply to the PSSP accounting option, but are used only if you set *spacct_enable* to **true**. Use *spacct_actnode_thresh* to specify the minimum percentage of nodes for which accounting data must be present. Use *spacct_exclusive_enable* to specify whether, by default, separate accounting records are generated when a LoadLeveler job requests exclusive use of a node.

Use *acct_master* to specify which node is to act as the accounting master. The default value is **0** (the control workstation).

Table 39. Accounting choices

	Worksheet entries to be filled in...			
To do this...	*spacct_enable*	*spacct_actnode_thresh*	*spacct_exclusive_enable*	*acct_master*
Do not use PSSP accounting	**false** (default)			
Use PSSP accounting	**true**	**80**	**false** (default)	**0**
Note: • Change default attribute values to suit your environment. • Blank entries imply that you make no substitutions for these values.				

For information about this feature of PSSP and how to set up an accounting system, see the chapter on accounting in the books *PSSP: Administration Guide* and the relevant AIX version of *System Management Guide*.

Understanding lppsource directory name choices

The *cw_lppsource_name* attribute lets you specify the name of the directory to which the AIX and related file sets, collectively referred to as the lppsource, will be copied.

You must ensure that the AIX level of the licensed programs contained in the lppsource (indicated by the value given to *cw_lppsource_name*) matches the AIX level installed on your control workstation.

The attribute value makes up just one part of the directory name in the form:

```
/spdata/sys1/install/<cw_lppsource_name>/lppsource
```

where *cw_lppsource_name* is the new lppsource name for the control workstation (such as aix510 if that is what you choose to call the directory with the AIX 5L 5.1

licensed program source files). Keep in mind that the **setup_server** program looks for and uses this name later in the installation process. By default, it is set to `default`, so that if you use that as your directory name, you do not have to change the value of *cw_lppsource_name*. If you do not provide a name, the **setup_server** program assumes the value is `default`.

See the chapter on preparing the control workstation in the book *PSSP: Installation and Migration Guide*.

Site Environment Worksheet entries

Only one attribute applies to the lppsource directory name option.

Note: The lppsource name specified here might or might not be the same that is used by the nodes. Be sure to specify the appropriate lppsource name to be used by the nodes when installing your system.

Table 40. lppsource directory choices

	Worksheet entry to be filled in...
To do this...	*cw_lppsource_name*
Use **aix510** to uniquely identify the new lppsource directory	**aix510**
Use **default** as the default lppsource directory	
Note: • Change default attribute values to suit your environment. • Blank entries imply that you make no substitutions for these values.	

Understanding remote command choices

You have the option of running PSSP with an enhanced level of security called *restricted root access*. With the restricted root access option enabled, PSSP system management software does not internally issue **rsh** and **rcp** commands as a root user from a node. Any such actions can only be run from the control workstation or from nodes configured to authorize them. If you enable this option, PSSP does not automatically grant authorization for a root user to issue **rsh** and **rcp** commands from a node and some procedures might not work as documented. For example, to run HACMP an administrator must grant the authorizations for a root user to issue **rsh** and **rcp** commands that PSSP otherwise grants automatically.

In addition, you can choose to use a *secure remote command process* to replace the **rsh** and **rcp** commands issued by PSSP system management software running as root on the control workstation. You must acquire and install the secure remote command software on the control workstation before you can enable a secure remote command process to be used by the PSSP software. The secure remote command software must be running on the control workstation and root must have the ability to successfully use it to issue remote commands to the nodes without being prompted for passwords or passphrases.

See “Considering restricted root access” on page 146 and “Considering a secure remote command process” on page 151 for more explanation and limitations before you decide to use these options.

Site Environment Worksheet entries

You must enable restricted root use of remote commands in order to use a secure remote command process.

Table 41. Remote command choices

	Worksheet entries to be filled in...			
To do this...	*restrict_root_rcmd*	*rcmd_pgm*	*dsh_remote_cmd*	*remote_copy_cmd*
Do not restrict PSSP root use of remote commands	**false** (default)	**rsh** (default)		
Do not restrict PSSP root use of remote commands but use alternative executables	**false** (default)	**rsh** (default)	**/usr/local/bin/rsh** for example	**/usr/local/bin/rcp** for example
Restrict PSSP to use remote commands from the cws only	**true**	**rsh** (default)		
Restrict PSSP to use remote commands from the cws only and use alternative executables	**true**	**rsh** (default)	**/usr/local/bin/rsh** for example	**/usr/local/bin/rcp** for example
Restrict PSSP use of remote commands and use the default secure remote command process	**true**	**secrshell**		
Restrict PSSP use of remote commands and use an alternative secure remote command process	**true**	**secrshell**	**/usr/local/bin/ssh** for example	**/usr/local/bin/scp** for example
Note: • Change default attribute values to suit your environment. • Blank entries imply that you make no substitutions for the default values.				

Understanding partitionability choices

You might be in a position to expand your switchless SP system, but you find that you do not have enough available switch ports in your existing SP frames to add all the servers you want. If you have no need for multiple SP system partitions or the SP Switch now and in the future, you might want to disable the SP system partitioning function. Then the SP Switch port numbering rules do not apply and any servers you add are numbered sequentially.

When the PSSP software is installed, a single default system partition is automatically created. It consists of the control workstation and all the nodes connected to the system administrative LAN. The single default system partition is all you can have if your system uses the SP Switch2 or it is a switchless clustered

server system. SP system partitioning is supported by default in switchless systems with at least one SP node frame. However, you can force a switchless SP system to be nonpartitionable. You might want to disable SP system partitioning if you have no need for multiple SP system partitions and you want to ignore switch node number rules in order to have more SP-attached servers than available switch ports. You can set the `force_non_partitionable` attribute by using the SP site environment SMIT panel or the **spsitenv** command. See the **spsitenv** command in the book *PSSP: Command and Technical Reference*.

Site Environment Worksheet entries

Disable SP system partitioning in order to add more SP-attached servers than there are available switch port numbers to accommodate them.

Table 42. Partitioning choices

	Worksheet entries to be filled in...
To do this...	*force_non_partitionable*
Force a switchless SP system to be not partitionable.	**true**
Make a switchless SP system partitionable.	**false** (default)
Note: • Change default attribute values to suit your environment. • Blank entries imply that you make no substitutions for the default values.	

Determining install space requirements

Your SP install package includes standard installation images plus the optional images that you order. The install images require disk space to contain them on each control workstation and PSSP boot-install server.

AIX 5L 5.1 requires a minimum of 64 megabytes of physical memory and the initial paging space (/dev/hd6) has to be a minimum of 64 megabytes in size. Table 43 shows approximate amounts of space allocated during the base AIX install process. Use these as guidelines to help you estimate the space required in directories on the control workstation and boot-install servers.

Table 43. Approximate space allocated during base AIX install

Directory	**AIX 5L 5.1**	**AIX 4.3.3**
/ (root)	8MB	4MB
/usr	385MB	294MB
/var	4MB	4MB
/tmp	32MB	16MB
/opt	4MB	N/A

These directories actually need more space available than the amounts listed to accommodate other processes as well. The installation process involves additional file sets that require disk space on each control workstation and PSSP node on which you plan to install them. Find requirements in the AIX and other licensed program release notes and installation books. This book guides you in estimating overall install space for successful installation of PSSP.

To calculate how much disk space you need on the control workstation and the PSSP nodes for a successful install, sum the estimated sizes of all the licensed

programs you plan to install . As you decide which images to install, you might develop a list similar to the example in Table 44 on page 100 for your installation plan. Include the required images plus your choices of the optional PSSP components and other licensed program images to install. This example includes images of:

- The minimum AIX file sets (**spimg**)
- The required PSSP components (**ssp**)
- The optional PSSP components and graphical user interfaces, in this example the Resource Center (**ssp.resctr**), IBM Virtual Shared Disk, Hashed Shared Disk, and Recoverable Virtual Shared Disk (**vsd, ssp.vsdgui**)
- Each optional PSSP-related licensed program, in this example LoadLeveler (**LoadL**).

The example depicted in Table 44 on page 100 is used during this discussion to explain how to determine your total space requirements for the install images on each control workstation and boot-install server.

Table 45 on page 101 has the approximate sizes of the images and file sets shipped with the PSSP software.

Note: The RSCT file sets are delivered with AIX 5L 5.1. Be sure to install them because they are required by the PSSP software. RSCT filesets (rsct.*) require approximately 36 MB of total disk space in **/usr**. See the book *AIX 5L 5.1: Packaging Guide for LPP Installation*.

Estimating requirements for lppsource

The **lppsource** is a required resource for NIM, the network installation management facility used to install AIX on the nodes. The lppsource contains the AIX file sets. The amount of space this resource needs depends on how you use the resource. For instance, if you plan to use DCE authentication, the DCE install files must be added to the lppsource directory.

If you plan to install DCE file sets, they are required to also be in the lppsource directory on the control workstation and to be exported to all the nodes on which you plan to use DCE security services. Follow normal procedures for installing licensed programs. See the book *IBM Distributed Computing Environment for AIX: Quick Beginnings* for planning and installing DCE.

You can download all of the AIX file sets from the AIX installation media. Although this takes more space than the minimal required file sets, it saves time and effort if you intend to use **installp** for additional file sets that are not already installed on your nodes. IBM suggests this method because it makes it easier to perform additional **installp** installations.

Alternatively, you can download only the AIX file sets required by NIM to perform the mksysb installations on the nodes. The list of the minimal AIX file sets required appears in the book *PSSP: Installation and Migration Guide*, which tells you how to download the file sets.

Downloading all of the AIX file sets requires approximately 2GB of disk space. Downloading only the minimal AIX file sets requires approximately 1.5GB.

You also need to determine which lppsource levels you need. In general, you need one lppsource level for each AIX level that you intend to install. The AIX 5L 5.1 level uses approximately 1MB more than AIX 4.3.3.

Estimating the node installation image requirements

When installing the nodes, a **mksysb** image (spimg) is installed. The **mksysb** image is stored on the control workstation. A typical **mksysb** image is generally in a size range from about 91MB to about 1.5GB. If you intend to install one image on some nodes and another image on other nodes then you must also account for the extra space required by multiple images.

Other installp image requirements

If you want to install additional licensed programs that are not part of the AIX installation media and are not included in PSSP, then you should also include the space that they require in your calculations. Find their sizes in the respective publications. For example, LoadLeveler and IBM C and C++ compilers are additional licensed programs that require space.

Combining the space requirements

All these resources reside in a directory typically called **/spdata**. Use the following algorithm to estimate the amount of additional space you need for **/spdata** on each control workstation and boot-install server:

```
lppsource + mksysb_image + pssp_image + optional_images = total_space
```

For example, the space needed on the control workstation for **/spdata** to contain lppsource plus only the required images for PSSP 3.4 with AIX 4.3.3 is approximately:

```
  1500MB   +   91MB   +   168MB = 1759MB
lppsource + spimg mksysb + ssp = minimum space
```

Table 44 summarizes the amount of space required for the install images chosen for the example. Determine the size of your lppsource. Use the spimg value in the algorithm directly. Sum the rest of the images to get the `ssp` value. Then use the algorithm to estimate the total space required for **/spdata**.

Table 44. Example of listing installp images

Space required for storing installp images		
installp image	**Space required**	**Description**
spimg.510	442 MB AIX 5L bos.obj.ssp.510, bos.obj.ssp.510_64	This has the minimal AIX 5L 5.1 image.
ssp	150 MB	This has the base PSSP components. It must be on the control workstation.
ssp.resctr	4 MB	The Resource Center image is optional on the control workstation and nodes.
vsd and ssp.vsdgui	8 MB	The IBM Virtual Shared Disk, Hashed Shared Disk, and IBM Recoverable Virtual Shared Disk optional image must be on the control workstation and nodes that will have or use IBM Virtual Shared Disks. The ssp.vsdgui image is the IBM Virtual Shared Disk Perspective. It must be on the control workstation and is optional on nodes.

Table 45 shows the approximate install image and individual file set sizes. Decide which you plan to install on each control workstation and PSSP node and use them to help estimate your disk space and install image space requirements.

Table 45. PSSP 3.5 file set and install image sizes

Image or file set name	Image size	File set size
PSSP minimal AIX image spimg	442MB (AIX 5L)	
bos.obj.ssp.510		206MB
bos.obj.ssp.510_64		236MB
PSSP image ssp	150MB	
ssp.authent		660KB
ssp.basic		6.8MB
ssp.cediag		460KB
ssp.clients		12.7MB
ssp.css		90MB
ssp.docs		41MB
ssp.gui		20MB
ssp.ha_topsvcs.compat		1KB
ssp.perlpkg		1.6MB
ssp.pman		553KB
ssp.public		12MB
ssp.spmgr		151KB
ssp.st		1MB
ssp.sysctl		1.5MB
ssp.sysman		1.3MB
ssp.tecad		204KB
ssp.top		1.4MB
ssp.top.gui		1.4MB
ssp.ucode		1MB
PSSP image ssp.hacws	150KB	
ssp.hacws		150KB
PSSP image vsd	4.6MB	
vsd.cmi		191KB
vsd.hsd		213KB
vsd.rvsd.hc		1.6MB
vsd.rvsd.rvsdd		331KB
vsd.rvsd.scripts		326KB
vsd.sysctl		463KB
vsd.vsdd		2MB
PSSP image for graphical user interface		
ssp.vsdgui		2.6MB
PSSP image ssp.resctr for the Resource Center		

Table 45. PSSP 3.5 file set and install image sizes (continued)

Image or file set name	Image size	File set size
ssp.resctr.rte		4MB
PSSP images ssp.msg.en_US.* – US English ISO8859-1, ISO8859-15		
ssp.msg.en_US.authent		16KB
ssp.msg.en_US.basic		262KB
ssp.msg.en_US.cediag		33KB
ssp.msg.en_US.clients		160KB
ssp.msg.en_US.pman		24KB
ssp.msg.en_US.spmgr		7KB
ssp.msg.en_US.sysctl		34KB
ssp.msg.en_US.sysman		71KB
PSSP images ssp.msg.En_US.* – US English IBM-850		
ssp.msg.En_US.authent		16KB
ssp.msg.En_US.basic		262KB
ssp.msg.En_US.cediag		33KB
ssp.msg.En_US.clients		160KB
ssp.msg.En_US.pman		24KB
ssp.msg.En_US.spmgr		7KB
ssp.msg.En_US.sysctl		34KB
ssp.msg.En_US.sysman		71KB
Note: The total storage can cross multiple file systems.		

Planning your system network

This section contains some hints, tips and other information to help in tuning the SP system. These sections provide specific information on the SP and its subsystems. By no means is this section complete and comprehensive, but it addresses some SP-specific considerations. See the book *AIX 5L: Performance Management Guide* for additional AIX tuning information.

System topology considerations

When configuring larger systems, you need to consider several topics when setting up your network. These are the SP Ethernet, the outside network connections, the routers, the gateways, and the switch traffic.

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

If you connect the SP Ethernet to your external network, you must make sure that the user traffic does not overload the SP Ethernet network. If your outside network is a high speed network like FDDI or HIPPI, routing the traffic to the SP Ethernet can overload it. For gateways to FDDI and other high speed networks, you should

route traffic over the switch network. You should configure routers or gateways to distribute the network traffic so that one network or subnet is not a bottleneck. If the SP Ethernet is overloaded by user traffic, move the user traffic to another network.

If you expect a lot of traffic, then you should configure several gateways. You can monitor all the traffic on these networks using the standard network monitoring tools. For more information on these tools, refer to the *AIX 5L: Performance Management Guide* publication.

Boot-Install server requirements

When planning your SP Ethernet admin LAN topology, consider your network install server requirements. The network install process uses the SP Ethernet for transferring the install image from the install server to the SP nodes. Running lots of concurrent network installs can exceed the capacity of the SP Ethernet admin LAN. The following are suggested guidelines for designing the SP Ethernet topology for efficient network installs. Many of the configuration options will require additional network hardware beyond the minimal node and control workstation requirements. There are also network addressing and security issues to consider.

The p670 or p690 server gets network connected and does not require you to use en0. For all other nodes, you must use the en0 adapter to connect each node to the SP Ethernet admin LAN. The following requirements pertaining to the SP Ethernet admin LAN exist for all configurations:

- Each boot-install server's Ethernet adapter must be directly connected to each of the control workstations' Ethernet adapters.
- The Ethernet adapter must always be in the SP node's lowest hardware slot of all Ethernets. This does not pertain to the model p670 and p690.
- The NIM clients that are served by boot-install servers must be on the same subnet as the boot-install server's Ethernet adapter.
- NIM clients must have a route to the control workstation over the SP Ethernet.
- The control workstation must have a route to the NIM clients over the SP Ethernet.

Notes:

1. Certain security options have limitations with multiple boot-install servers. See "Limitations when using restricted root access" on page 149 before you decide.
2. Do not install cascading levels of boot/install servers. Every boot/install server node must have the control workstation as its boot/install server.

Single frame systems

For small systems, you can use the control workstation as the network install server. This means that the SP Ethernet admin LAN is a single network connecting all nodes to the control workstation. When installing the nodes, limit yourself to installing eight nodes at a time because that is the limit of acceptable throughput on the Ethernet. Figure 7 on page 104 shows an Ethernet topology for a single-frame system.

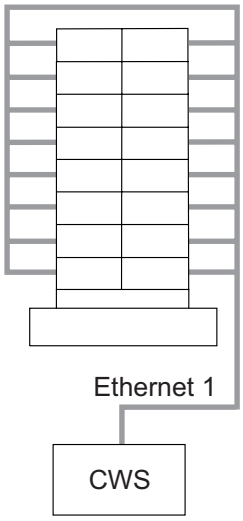


Figure 7. Ethernet topology with one adapter for a single-frame SP system

An alternate way to configure your system is to install a second Ethernet adapter in your control workstation, if you have an available I/O slot, and use two Ethernet segments to the SP nodes. Connect each network to half of the SP nodes. When network installing the frame, you can install all 16 nodes at the same time. Figure 8 shows this alternate Ethernet topology for a single-frame system.

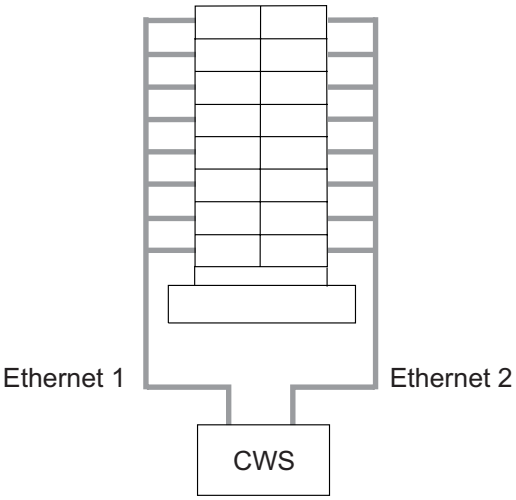


Figure 8. Ethernet topology with two adapters for single-frame SP system

Set up your SP Ethernet admin LAN routing so nodes on one Ethernet can communicate to nodes on the other network. Set up your network mask so that each SP Ethernet is its own subnet within a larger network address. Consult your local network administrator about getting and assigning network addresses and network masks.

Multiple frame systems

For multiple frame systems, you might want to spread the network traffic over multiple Ethernets, and keep the maximum number of simultaneous installs per network to eight. You can use the control workstation to network install specific SP nodes which will be the network install servers for the rest of nodes.

Following are three ways to accomplish this.

1. The first method uses a control workstation with one Ethernet adapter for each frame of the system, and one associated SP Ethernet per frame. So, if you have a system with four frames as in Figure 9, the control workstation must have enough I/O slots for four Ethernet adapters, and each adapter connects one of the four SP frame Ethernet segments to the control workstation. Using this method, you install the first eight nodes on a frame at a time, or up to 32 nodes if you use all four Ethernet segments simultaneously. Running two installs will install up to 64 nodes. Figure 9 shows an Ethernet topology for this multi-frame system.

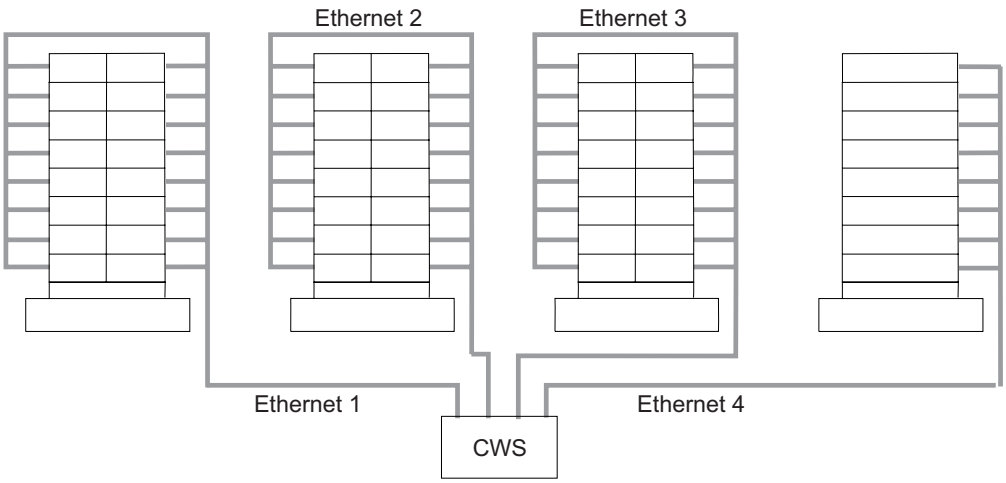


Figure 9. Method 1 Ethernet topology for multi-frame SP system

   Once again, set up your SP Ethernet routing so nodes on one Ethernet can communicate to nodes on another. Set up your network mask so that each SP Ethernet is its own subnet within a larger network address. Consult your local network administrator about getting and assigning network addresses and network masks.

   This method is applicable up to the number of slots your control workstation has available.

2. A second approach designates the first node in each frame as a network install server, and then the remaining nodes of that frame are set to be installed by that node. This means that, from the control workstation, you will have an SP Ethernet segment connected to one node on each frame. Then the network install node in each frame has a second Ethernet card installed which is connected to an Ethernet card in the rest of the nodes in the frame. Figure 10 on page 106 shows an Ethernet topology for this multi-frame system.

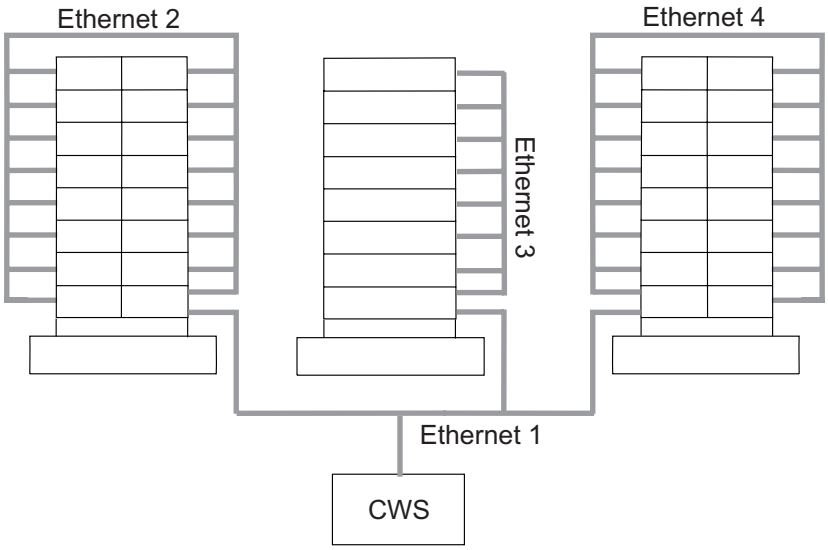


Figure 10. Method 2 Ethernet topology for multi-frame SP system

When using this method, installing the nodes requires that you first install the network install node in each frame. The second set of installs will install up to eight additional nodes on the frame. The last install, if needed, installs the rest of the nodes in each frame.

Be forewarned that this configuration usually brings performance problems due to two phenomena:

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

b. An application running on a node which produces a high volume of SP Ethernet traffic (for example, LoadLeveler) causes all subnet routing to go through the one control workstation Ethernet adapter. Moving the subject application to the control workstation can cut that traffic in half, but the control workstation must have the capacity to accommodate that application.

You can improve the performance here by adding an external router, similar to that described in method 3.

3. A third method adds an external router to the topology of the previous approach. This router is made part of each of the frame Ethernets, so that traffic to the outside need not go through the control workstation. You can do this only if the control workstation can also be attached externally, providing another route between nodes and the control workstation. Figure 11 on page 107 shows this Ethernet topology for such a multi-frame system.

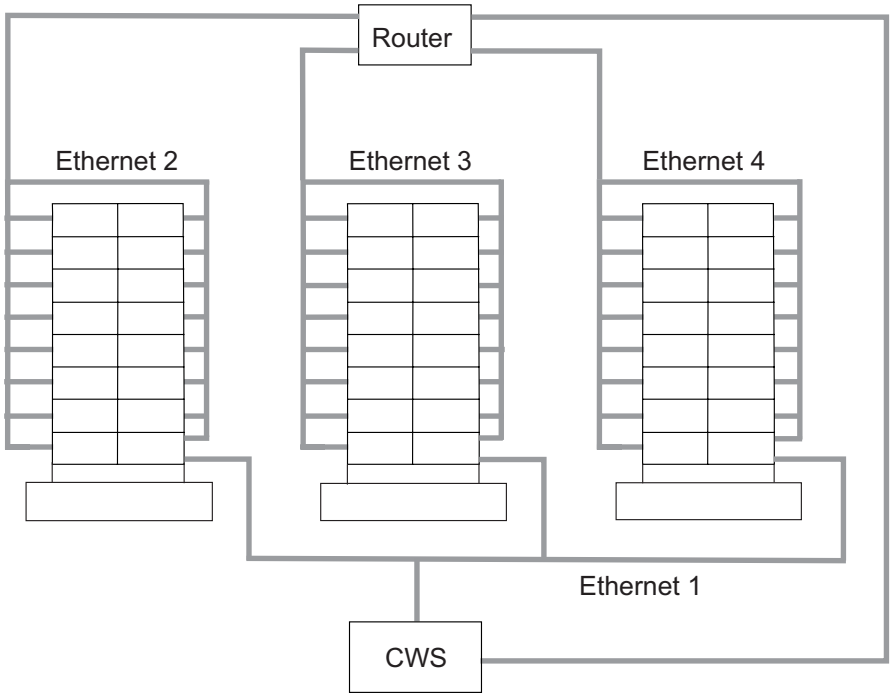


Figure 11. Method 3 Ethernet topology for multi-frame SP system

An alternative to the router in this configuration is an Ethernet switch, which could have a high-speed network connection to the control workstation.

Future expansion considerations and large scale configuration

If your configuration will grow over time to a large configuration, you might want to dedicate your network install nodes in a different manner.

For very large configurations you might want to dedicate a frame of nodes as designated network install nodes, as shown in Figure 12 on page 108. In this configuration, each SP Ethernet from the control workstation is connected to up to eight network install nodes in a frame. These network install nodes are in turn connected to additional frames.

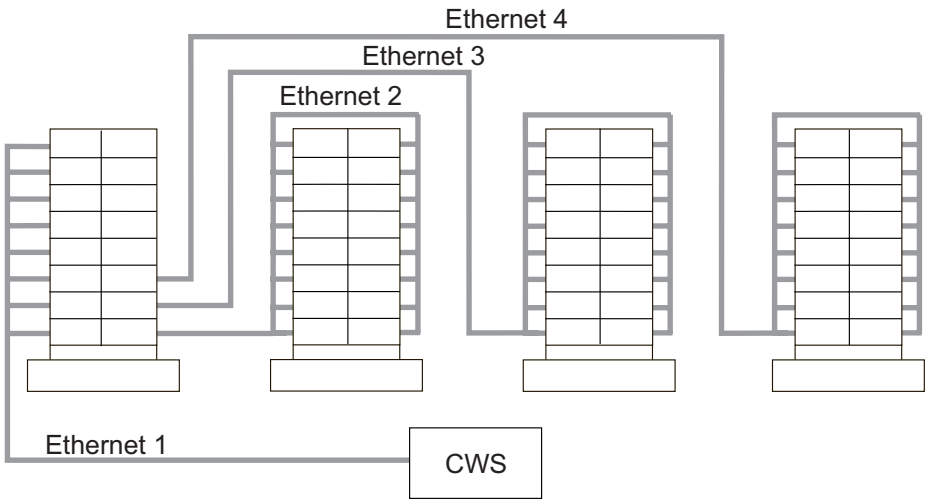


Figure 12. Boot-server frame approach

The advantage of this is that when you add an additional frame to your SP configuration, all you need to do is connect the new frame to one of the network install nodes, and reconfigure the system.

The network install procedure for this system is the same as for multiple frame systems. You first install the network install servers at a rate of eight per SP Ethernet segment. The network install servers then install eight other nodes until all nodes are installed.

The network address usually used for the SP Ethernet is a class C internet address. This address has a limit of 256 individual addresses before you need to add additional network addresses for the SP Ethernet. If your system is expected to grow beyond this number of nodes, you should plan with your Network Administrator additional network addresses for future SP Ethernet expansion. This will save you from having to re-assign the SP Ethernet addresses when you reach the address limit.

Location and reference rate of customer data

Customer application data can be delivered to applications running on the SP from file servers. These file servers can be either internal SP nodes or separate external systems. The location of the data, how often you refer to it, and whether it is accessed in read-only or both read and write modes affect the performance of applications using this data. Applications that have a high data reference rate, especially those that read and write data, benefit from having the data closely located to the node on which the application executes. The *co-location* of data and applications minimizes the amount of network processing required to move the data to and from its file server.

Home directory server planning

When planning for home directory servers, you must determine how much traffic will be generated by requests from the nodes to the server. Because some home directories are NFS, AFS, or DFS-mounted, you need to determine the amount of traffic in operations per second.

If the amount of traffic is greater than the capacity of a single network, you need to add additional networks and divide the number of nodes per network to the server.

If the amount of traffic is greater than the capacity on the server, you need to configure additional servers, each connected to all networks.

Authentication servers

When you install the Cluster 1600 system PSSP software, you can define one or more authentication servers. Authentication provides a more secure Cluster 1600 system managed by PSSP by verifying the identity of clients that access key systems management facilities.

You can install, configure, and use the DCE security services, the PSSP implementation of Kerberos V4 authentication, or you can integrate your Cluster 1600 system managed by PSSP into an existing authentication domain, such as a DCE or AFS cell. If you choose to use AFS authentication servers, note in particular the section on the assignment of TCP/IP port numbers in the **/etc/services** file.

Carefully plan the location of master and replica authentication servers. The following are some considerations:

- You might want to set up your servers on independent IBM @server pSeries or RS/6000 workstations that are isolated by physical location or have limited network access.
- DCE master and replica servers can be installed on the control workstation, on any PSSP node, or on an external network-attached IBM @server pSeries or RS/6000 node. It might not be good to place the DCE master server on a PSSP node but it could be good for a replica server.
- Kerberos V4 servers can be on the control workstation or on an external network-attached IBM @server pSeries or RS/6000 server.
- Making the control workstation the master DCE authentication server might not be good for performance. It is still good to have the Kerberos V4 master authentication server on the control workstation.
- Your DCE master authentication server must be installed and operating before you install and configure your control workstation, unless the control workstation is to be the master server.

You have more authentication options than before, but any node still running a pre-PSSP 3.2 release requires use of Kerberos V4. See Chapter 6, “Planning for security” on page 145 for more options and planning information.

If you need still more information, see the books *AIX Version 4: System Management Guide* and *PSSP: Administration Guide*.

Understanding node hard disk choices

The *pv_list* attribute in the SDR specifies on which disks to create the root volume group (**rootvg** by default) and to transfer the **mksysb** image during AIX network installation of a node. The default value of this attribute is **hdisk0** with one exception. For the POWER3 SMP high node, the default is the SCSI disk locations of the two internal disks delivered with the node. Also by default for POWER3 SMP high nodes, the two internal disks will be used for mirroring with two copies of the root volume group. If your POWER3 SMP high node is not configured with the two internal disks, you must override the default. Depending on your nodes and environment, you might have the installation include additional hard disks.

You can use more than one disk when the **mksysb** image is larger than the disk can hold or when you need the root volume group to span multiple disks. If you do not have either of these circumstances, you should not install on more than one

disk. If you do use more than one disk, keep in mind that the first disk in a node is not necessarily **hdisk0**. When you boot up a node, the first disk found is **hdisk0**. If you have a fast, wide external disk attached to a node, it can come up as **hdisk0**.

If you have another disk, you can define a different root volume group on that disk and import it. This lets you reinstall the node and import the volume group without having to back up and restore the data on the non-install disk.

Check your disks to ensure your install image is on internal disks. With bootable external disk support like SSA or Fibre Channel disks, you can have the install image on external disk as long as that image is not shared across nodes. However, IBM suggests you still keep the install image on internal disk for efficiency.

To see the value of the *pv_list* attribute, run the **splstdata** command with the **-v** option. To change the *pv_list* attribute, use the **spchvgobj** command with the **-h** option. See the book *PSSP: Command and Technical Reference* for more information on the **spchvgobj** command. See the book *PSSP: Administration Guide* for more information about managing root volume groups, including mirroring the root volume group and alternate root volume groups.

Planning your network configuration

This section discusses what you need to know to plan your network configuration. Instructions for completing the remaining system planning worksheets begin in Chapter 2, “Defining the system that fits your needs” on page 19. All the worksheets are summarized in Appendix C, “SP system planning worksheets” on page 281.

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. This repeats some information contained in “Completing the node layout worksheets” on page 56. This repetition is valuable because of the importance of this information.

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
- How you will establish these default and network routes (that is, using **routed** or **gated** daemons or using explicit route statements).

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/host** files or some other method).

Note

All names and addresses of all IP interfaces on your nodes must be resolvable on the control workstation and on independent workstations set up as authentication servers before you install and configure the SP system. The SP system uses only IPv4 addresses. Some PSSP components tolerate IPv6 aliases for IPv4 network addresses but not with DCE, HACMP, HACWS, or an SP switch. For information about the SP system tolerating IPv6 aliases for some IPv4 network addresses, see the appendix on the subject in the book *PSSP: Administration Guide*.

Once you have set the host names and IP addresses on the control workstation, you should not change them.

Some name resolution facilities let you map multiple IP interfaces to the same host name. For the SP system, IBM suggests that you assign unique host names to each IP interface on your nodes.

Understanding the SP networks

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

- "The SP Ethernet admin LAN"
- "Additional LANs" on page 112
- "Firewall LANs" on page 112

You also need to understand the considerations that are relevant to certain hardware or system features as they relate to planning the SP networks:

- "Considering IP over the switch" on page 113
- "Considering subnetting" on page 113
- "Considering an HMC trusted network" on page 113

The SP Ethernet admin LAN

The SP Ethernet is the administrative LAN that connects all nodes in one system running PSSP to the control workstation. It is used for PSSP installation and communication among the control workstation, boot-install servers, and other nodes in the network. For p690 and p670 servers, it can also be used to connect the HMC to the control workstation for hardware control and monitoring. See "Considering an HMC trusted network" on page 113.

For each node, ensure that the SDR reliable_hostname attribute is identical to the default host name returned by the host command for its SP Ethernet IP addresses. With the exception of the p690 nodes, the PSSP components expect that en0 is the connection from the node to the SP Ethernet admin LAN for installs and other PSSP functions. For example, if the en0 IP address of a node is 129.40.133.75, and 'host 129.40.133.75' shows the default host name is k65n11.ppd.pok.ibm.com, then it also should be the host name set in the reliable_hostname attribute in the SDR.

For a p690 node, you can use any Ethernet adapter that is supported for connecting to the SP Ethernet admin LAN and identify it by name or by physical location. For all other nodes, you must connect the SP Ethernet admin LAN to the Ethernet adapter in the node's lowest hardware slot of all the Ethernet adapters on

that node. When a node is network booted, it selects the lowest Ethernet adapter from which to perform the install. This Ethernet adapter must be on the same subnet of an Ethernet adapter on the node's boot-install server. In nodes that have an integrated Ethernet adapter, it is always the lowest Ethernet adapter. Be sure to maintain this relationship when adding Ethernet adapters to a node.

You can attach the SP Ethernet to other site networks and use it for other site-specific functions. You assign all addresses and names used for the SP Ethernet admin LAN.

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

- Single-subnet, single-stage SP Ethernet in which one interface on the control workstation connects to all SP nodes.
- Multi-subnet, single-stage SP Ethernet. There is more than one interface on the control workstation and each connects to a subset of the SP nodes.
- Multi-subnet, multi-stage SP Ethernet. A set of nodes, acting as routers to the remaining nodes on separate subnets, connects directly to the control workstation.

See "System topology considerations" on page 102 for sample configurations illustrating these connection methods.

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

Also, when booting from the network, nodes broadcast their host request over their SP Ethernet admin LAN interface. Therefore, that interface must be the Ethernet adapter on the node that is connected to the boot-install network. For all nodes except those on a p670 or p690, it must be en0.

Additional LANs

The SP Ethernet admin LAN can provide a means to connect all nodes and the control workstation to your site networks. However, it is likely that you will want to connect your SP nodes to site networks through other network interfaces. If the SP Ethernet admin LAN is used for other networking purposes, the amount of external traffic must be limited. If too much traffic is generated on the SP Ethernet admin LAN, the administration of the SP nodes might be severely impacted. For example, problems might occur with network installs, diagnostic function, and maintenance mode access. In an extreme case, if too much external traffic occurs, the nodes will hang when broadcasting for the network.

Additional Ethernet, Fiber Distributed Data Interface (FDDI), and token-ring networks can also be configured by the PSSP software. Other network adapters must be configured manually. These connections can 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.

Firewall LANs

You can plan to run PSSP on a system with a firewall. For instructions, see *Implementing a Firewalled RS/6000 SP System*.

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

If you are not enabling ARP on the switch, specify the switch network subnet mask and the IP address. See “Understanding placement and numbering” on page 118. Unlike all other network interfaces, which can have sets of nodes divided into several different subnets, the SP switch IP network must be one contiguous subnet that includes all of the nodes in the system. If you use the SP Switch2 and all nodes are running PSSP 3.4 or greater, you have optional connectivity so some nodes can be left off the switch.

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

Considering subnetting

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. See the relevant edition of the AIX book *System Management Guide: Communications and Networks* for information about Internet addresses and subnets.

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

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

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

Considering an HMC trusted network

Cluster 1600 managed by PSSP configurations that contain p690 or p670 servers require additional security planning and configuration to protect password data transferred from the control workstation to a p690 or p670 Hardware Management Console (HMC). Cluster 1600 configurations managed by PSSP that do not contain p690 or p670 servers do not require additional security planning and configuration.

PSSP Hardware Monitor control of p690 and p670 servers is established through an IP network connection between the control workstation and an HMC. This IP network connection is also used to transfer PSSP Kerberos V4 keytab files (password files), the PSSP secure file collections password file, and HMC user ID

and password login data. In order to protect password data transferred from the control workstation to an HMC, a trusted network must be established between the control workstation and HMC.

A trusted network is one where all hosts on the same network (LAN) are regarded as trusted, according to site security policies and procedures governing the hosts. Data on a trusted network can be seen by all trusted hosts and users on the trusted hosts, but the implied trust among and between the hosts assumes that the data will not be intercepted or modified. By way of implied mutual trust, traffic flowing across the trusted network is regarded as safe from unwanted or unintended interception or tampering. However, it does not imply that the data on the trusted network is itself private or encrypted.

You will need a trusted network that best suits your environment:

- ***Either the SP Ethernet admin LAN is a trusted network***

  If you already consider the SP Ethernet admin LAN a trusted network, no additional setup or configuration is needed to satisfy the security requirement. Follow instructions for p690 and p670 SP-attached server support in this book and the books *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* and *PSSP: Installation and Migration Guide*.

- ***Or you can establish another trusted network between the control workstation and the HMC***

  If you do not consider the SP Ethernet admin LAN a trusted network, you must establish a trusted network between the control workstation and the HMC in order to satisfy the security requirement. The rest of this discussion helps you plan to establish a trusted network between the control workstation and the HMC.

A trusted network between the control workstation and an HMC requires that both hosts are connected via a network other than the SP Ethernet admin LAN. This other network is referred to as the *HMC trusted network*. IBM strongly suggests that you connect only the SP control workstation and p690 or p670 HMC systems to the *HMC trusted network*.

The HMC trusted network requires a set of IP addresses all in the same subnet. IBM suggests that you reserve the subnet for use only by hosts that will be connected to the HMC trusted network. If you plan on connecting multiple control workstations and multiple HMC systems to the HMC trusted network, ensure the subnet can accommodate the total number of trusted hosts (actual or expected).

In order for a control workstation to be connected to both the SP Ethernet admin LAN and the HMC trusted network, each control workstation requires the installation and configuration of an additional network adapter to be configured for the HMC trusted network. See “Additional LANs” on page 112.

Each HMC must have a network adapter configured for and connected to the HMC trusted network. Connecting an HMC to the HMC trusted network will require the reconfiguration of an existing HMC network adapter. How the existing HMC network adapter is reconfigured depends on whether your p690 or p670 is a new SP-attached server install, or an existing SP-attached server:

- New SP-attached server install

  When your p690 or p670 is already installed and configured, but not yet connected to your SP system, it is a new SP-attached server install. Therefore, you must reconfigure the existing network adapter in the HMC for the HMC trusted network.

- Existing SP-attached server

  When your p690 or p670 is already an SP-attached server, the SP Ethernet admin LAN is used to connect the HMC to the control workstation. Specifically, the network adapter in the HMC is configured for the SP Ethernet admin LAN.

  In this case, you must unconfigure the network adapter in the HMC, disconnect it from the SP Ethernet admin LAN, connect it to the HMC trusted network, and then configure it for the HMC trusted network.

Considering network router nodes

If you plan to use an SP Switch Router and the SP Switch Router Adapter for routing purposes in your environment, the next few paragraphs on using standard nodes as a network router might not be applicable to your SP configuration. However, if you are not using the SP Switch Router, you might be interested in some considerations for using your nodes as network routers.

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

For nodes that use low bandwidth networks (such as, 10Mb Ethernet or token ring) as the routed network, a customer network running at full bandwidth results in a lightly loaded CPU on the router node. For nodes that use high bandwidth networks (such as, Gb Ethernet or FDDI) as the customer routed network, a customer network running at or near maximum bandwidth results in high CPU utilization on the router node. For this reason, you should not assign any additional role in the computing environment, such as a node in a parallel job, to a router using a high bandwidth network as the customer network. You also should not connect more than one high bandwidth network card to a router node.

Applications, such as POE, should run on nodes other than high bandwidth routers. However, low bandwidth gateways can run with these applications.

For systems that use low bandwidth routers, traffic can be routed through the SP Ethernet, but careful monitoring of the SP Ethernet will be needed to prevent traffic coming through the router from impacting other users of the SP Ethernet. For high bandwidth networks, traffic should be routed across the switch to the destination nodes. The amount of traffic coming in through the high bandwidth network can be up to 10 times the bandwidth the SP Ethernet can handle.

Information about configuring network adapters and the various network tunables on the nodes is in the book *PSSP: Administration Guide*.

Considering the SP Switch Router

The SP Switch Router is something you can use in a system with the SP Switch. It is not supported with the SP Switch2. It is by type an extension node, more specifically a dependent node. The SP Switch Router gives you high speed access to other systems. Without the SP Switch Router, you would need to dedicate a standard node to performing external network router functions. Also, because the SP Switch Router is external to the frame, it does not take up valuable processor space.

The SP Switch Router has two optional sizes. The smaller unit has four internal slots and the larger unit has sixteen. One slot must be occupied by an SP Switch

Router Adapter card which provides the SP connection. The other slots can be filled with any combination of network connection cards including the types:

- Ethernet
- FDDI
- ATM
- SONET
- HIPPI
- HSSI
- Additional SP Switch Router Adapters

Additional SP Switch Router Adapters are needed for communicating between system partitions and other SP systems. These cards provide switching rates of from 4 to 16 GB per second between the router and the external network.

To attach an extension node to an SP Switch, configuration information must be specified on the control workstation. Communication of switch configuration information between the control workstation and the SP Switch Router takes place over the SP system's administrative Ethernet and requires use of the UDP port number 162 on the control workstation. If this port is in use, a new communication port will have to be configured into both the control workstation and the SNMP agent supporting the extension node.

You can improve throughput of data coming into and going out of the SP system by using the SP Switch Router. The SP Switch Router can be connected with the SP Switch Router Adapter to an SP Switch, 8-port or 16-port. Each SP Switch Router Adapter in the SP Switch Router requires a valid unused switch port in the SP system. See "Choosing a valid port on the SP Switch" on page 117.

Considering a clustered server configuration

A clustered server is any of the servers discussed in "Question 8: Which and how many nodes do you need?" on page 43. It is not mounted in an SP frame and has no SP frame or node supervisor, though some do have comparable function enabling hardware control and monitoring. It is directly attached to the SP Ethernet admin LAN and to the control workstation. The means of connection differ with the server hardware.

In a clustered server system configuration you can assign frame numbers in any order. However, if you add SP frames with SP nodes or with SP switches, your system will then be subject to all the rules of an SP system ***and these clustered servers become SP-attached servers***. Remember that those terms reflect only the system configuration in which the servers participate.

If you might use the SP Switch2 or the SP Switch, you need to plan the respective switch network. If you might use the SP Switch, plan your system with suitable frame numbers and switch port numbers in advance so you can expand to an SP system without having to totally reconfigure existing servers. See "Switch port numbering for a switchless system" on page 127. Also read and understand the information regarding SP-attached servers in "Understanding placement and numbering" on page 118.

Considering an SP-attached server

An SP-attached server is also any of the servers discussed in "Question 8: Which and how many nodes do you need?" on page 43 but configured to be attached to the SP system.

If the SP system has the SP Switch, an SP-attached server requires an available node slot within an SP frame to reserve a valid unused switch port on the switch in the same SP frame. An SP-attached server is not supported with an SP Switch-8. You must connect the server to the SP Switch network with an adapter. That adapter connects to the valid unused switch port in the SP frame. For a p690 node, each LPAR attaches to the switch. See “Choosing a valid port on the SP Switch”.

In a switchless SP system where you will never use SP system partitioning, you can force the system to be non-partitionable and simply assign switch port numbers sequentially. See “Understanding partitionability choices” on page 97.

If the system has the SP Switch2, an SP-attached server and each p690 LPAR node can connect with an adapter to any available switch port in the SP frame. If all nodes are running PSSP 3.4 or greater, you can leave some nodes not connected to the switch.

You must assign a frame number to an SP-attached server. Be sure to read and understand the information regarding SP-attached servers in “Understanding placement and numbering” on page 118.

Choosing a valid port on the SP Switch

Each SP Switch Router Adapter in the SP Switch Router and each adapter for an SP-attached server requires a valid unused port of an SP Switch in the SP system. A valid unused switch port is a switch port that meets the rules for configuring frames and switches.

There are two basic sets of rules for choosing a valid switch port:

1. Rules for selecting a valid switch port associated with ***an empty node slot***.
2. Rules for selecting a valid switch port associated with ***an unused node slot*** created by a wide or high node position which is either the second half of a wide node or one of the last three positions of a high node.

These rules are discussed further in this chapter.

Examples of using an empty node slot position

One example of using an empty node slot position is a single frame system with fourteen thin nodes located in slots 1 through 14. This system has two unused node slots in position 15 and 16. These two empty node slots have corresponding switch ports which provide valid connections for an SP Switch Router Adapter or the adapter of an SP-attached server.

Another example is a logical pair, two frame system with one shared switch. The first frame is fully populated with eight wide nodes. The second frame has three wide nodes in slots 1, 3, and 5 (see later sections in this chapter for explanations of node numbering schemes). The only valid switch ports in this configuration would be those switch ports associated with node slots 7, 9,11, 13, and 15 in the second frame.

In a logical system with four frames holding fourteen high nodes sharing one switch, there will only be two empty node positions (see the Frames section of Chapter 1 for clarification). In this example, the first three frames are fully populated with four high nodes in each frame. The last frame has two high nodes and two empty high node slots. This means the system has two valid switch ports associated with node slot numbers 9 and 13.

Examples of using node slot positions covered by a wide or high node

The first example is a single frame fully populated with eight wide nodes. These wide nodes occupy the odd numbered node slots. Therefore, all of the even number slots are said to be unoccupied and would have valid switch ports associated with them. These ports can be used for an SP Switch Router Adapter or the adapter for an SP-attached server.

A second example is a single frame system with twelve thin nodes in slots 1 through 12 and a high node in slot 13. A high node occupies four slots but only uses one switch port. Therefore, the only valid switch ports in this configuration are created by the three unused node slots occupied by the high node. In other words, the switch ports are associated with node slots 14, 15, and 16.

Understanding placement and numbering

Thin, wide, and high nodes can coexist in the same frame and in the same SP system partition. Whether or not you use nodes with varied physical node sizes, you must very carefully consider the set of supported frame configurations. Extension nodes (like an SP Switch Router), SP-attached servers, and SP Expansion I/O Units must also be accommodated. Use the information in this section for:

- Deciding where to place your nodes, servers, and expansion I/O units
- Deciding how to sequence your frames
- Assigning IP addresses to the nodes and servers
- Assigning IP addresses to the interfaces of the nodes and the SP Switch or SP Switch2

Hardware planning is described in Volume 1.

This book covers switch planning only in the context of system configuration. For physical planning regarding switch wiring, cabling, and allowing for future hardware expansion, see the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*.

Slot numbers

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

The placement of a node in an SP system is sensed by the hardware. The node 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. (See thin nodes in Figure 13 on page 119.)
- A wide node occupies two slots and its slot number is the odd-numbered slot. (See the wide nodes in Figure 13 on page 119.)
- A high node occupies four consecutive slots in a frame. Its slot number is the first (lowest number) of these slots.
- An SP Expansion I/O Unit is not a node but a unit that expands the I/O capacity of the node to which it is connected. It can be connected to a POWER3 SMP

high node housed in the same frame or in a separate frame. In either case, it occupies a slot in a frame that a thin node might otherwise occupy and its slot number is the corresponding slot.

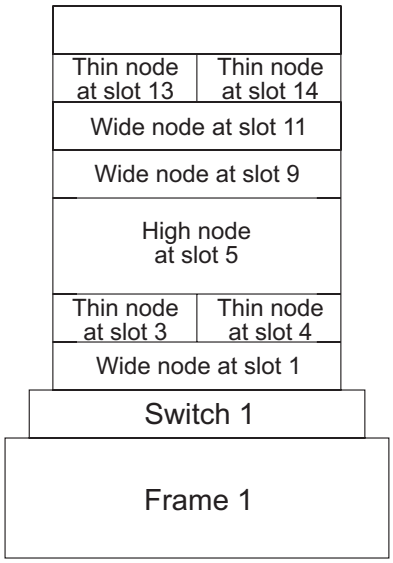


Figure 13. Node slot assignment

Frame numbers and switch numbers

The administrator establishes the frame numbers when the system is installed. To allow for growth you can skip frame numbers, but the highest number you can use for frames with nodes is 128. You can use higher frame numbers for additional frames that contain only switches or SP Expansion I/O Units. 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 by the PSSP software during the configuration process. This order is used to assign global identifiers to the switch ports, nodes, and SP Expansion I/O Units. This is also the order used to determine which frames share a switch.

For this discussion the types of frames are the following:

- A frame that has nodes and a switch is a *switched frame*.
- A frame with nodes and no switch is a *non-switched expansion frame*.
- A frame that has only switches for switch-to-switch connections is a *switch-only frame*. This is also called an intermediate switch board (ISB) frame.
- A frame that has only SP Expansion I/O Units, no nodes and no switch, is a *non-node frame*.
- A frame that has SP Switch2 multiple node switch boards (NSBs) and no nodes, is a *multiple NSB frame*.

If you have multiple NSB, switch-only, or non-node frames, configure them as the last frames in your SP system with numbers above 128. Assign high frame numbers to allow maximum configuration of nodes and for future expansion.

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

If you are planning a system of clustered servers, you can assign frame numbers in any order within the noted ranges. See “Considering a clustered server configuration” on page 116.

Node numbering

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 regardless of node or frame type. Replace *node number* with *expansion number* for calculating the global ID of an SP Expansion I/O Unit. Global IDs are automatically assigned using the following formula:

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

where *slot_number* is the lowest slot number occupied by the node or unit. Each type is considered to occupy one slot or a consecutive sequence of slots. For each type, there is an integer *n* such that a thin node, an LPAR, or an I/O expansion unit occupies slot *n*, a wide node occupies slots *n, n+1*, and a high node occupies *n, n+1, n+2, n+3*. An SP-attached server is considered to be one node in one frame, unless it is a p690 with multiple LPARs, in which case each LPAR is a node in that frame. For single thin nodes (not in a pair), wide nodes, and high nodes, *n* must be odd. For an SP-attached server, *n* is 1. Use *n* in place of *slot_number* in the formula. In LPAR-capable machines (for example, p670 or p690), the slot number is equivalent to the LPAR number assigned during LPAR configuration from the HMC.

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

Node numbers are assigned independent of whether the frame is fully populated. Figure 14 demonstrates node numbering. Frame 4 represents an SP-attached server in a position where it does not interrupt the switched frame and companion non-switched expansion frame, in other words the switch capsule 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 formula.

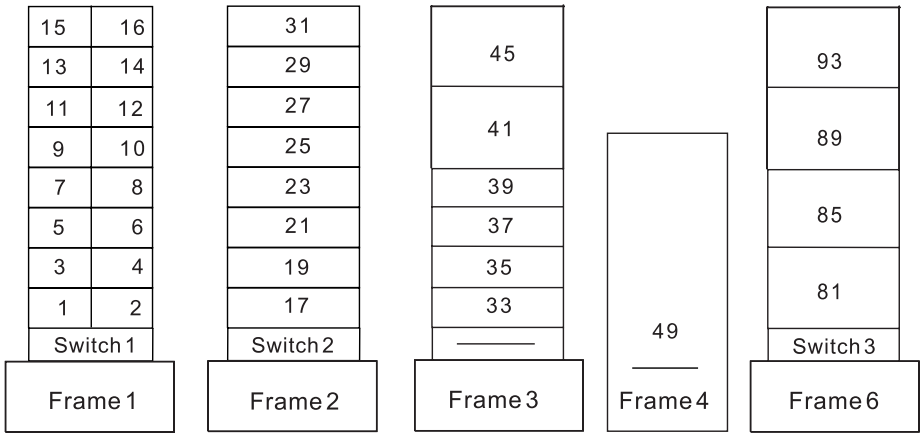


Figure 14. Node numbering

Node placement with the SP Switch

In order to properly plan an SP system with the SP Switch, you must understand the supported frame and switch configurations and the distribution of the switch port assignments in each of the supported configurations.

The PSSP system with the SP Switch supports the four possible frame and switch configurations shown in Figure 15 on page 122. Think of each configuration as a *switch capsule* which is comprised of a switched frame and its possible companion non-switched expansion frames. A non-switched expansion frame is a successor frame within the switch capsule that has SP nodes using switch ports of the switched frame.

Figure 15 on page 122 has four switch capsules, illustrating the four frame and switch configurations that are supported and the switch port number assignments in each. In the figure, no shading indicates a valid slot in which a node can be placed, the number in the slot represents that node's switch port assignment, and shading indicates that a node cannot be placed in that slot.

The four switch capsules can be repeated and mixed throughout your SP system. For example, consider an SP system with a switched frame followed by two non-switched expansion frames. They in turn might be followed by another switched frame and one more non-switched expansion frame. That SP system is therefore comprised of one switch capsule matching configuration 2 followed by another switch capsule matching configuration 1.

In every configuration that follows the first, the switch port numbers assigned increment to allow for the switches that reside in numerically lower frames. In the previous example, the frames in the first switch capsule (matching configuration 2) are assigned switch port numbers 0 through 15 and the frames in the second switch capsule (matching configuration 1) are assigned switch port numbers 16 through 31.

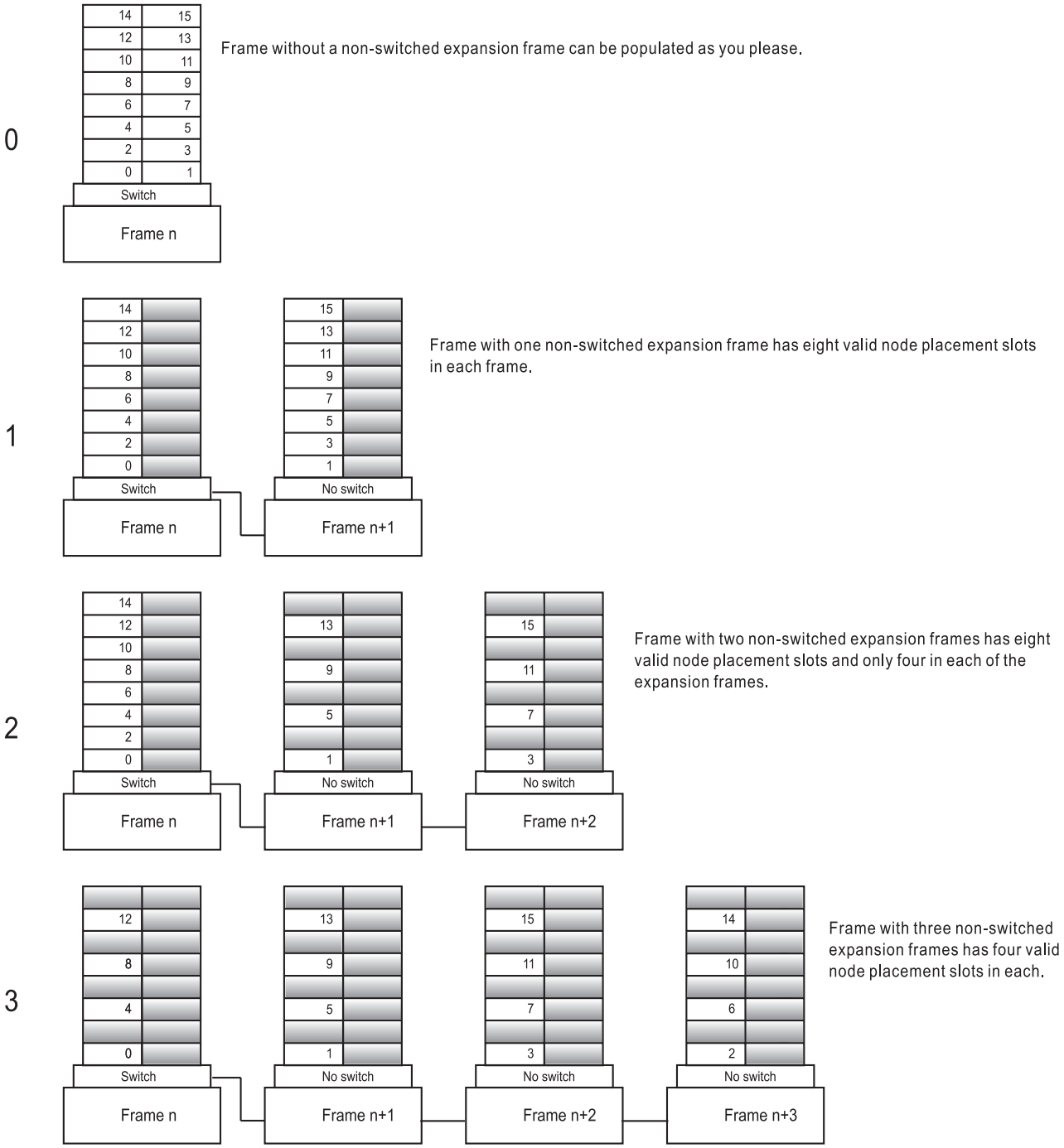


Figure 15. Supported SP Switch configurations showing switch port assignments

Keep in mind that any non-switched expansion frames must have frame numbers that immediately follow their associated switched frame without any gaps. For instance, if a system has a switched frame numbered 1, and two non-switched expansion frames attached to the switch on frame 1, the non-switched expansion frames must be numbered frame 2 and frame 3.

Frame numbers can be skipped between switched frames. It is a good idea to skip numbers to allow for future expansion. For example, consider a system that has a switched frame with four high nodes and another switched frame with 16 thin nodes. At this point the switched frame with four high nodes is a switch capsule that

is not fully populated. To accommodate future expansion, you would be wise to assign frame number 1 to the high node frame and number 5 to the thin node frame. That allows for the future addition of up to three non-switched expansion frames to the high node frame without disrupting the system. If the thin node frame had been numbered frame 2, the addition of a non-switched expansion frame would require you to reconfigure the thin node frame and all of its nodes.

An SP-attached server is managed by the PSSP components as though it is in a frame of its own. Servers with multiple LPARs are managed as a separate node for each LPAR within that frame. However, a server does not enter into the determination of the frame and switch capsule configuration of your SP system. It has the following additional characteristics:

- A p690 with multiple LPARs is one frame with one node for each LPAR within that frame. Any other type of server is one frame with one node. 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 derive the next node number in the next frame in sequence.
- If it uses the SP Switch2, it connects to any switch port in an SP switched frame with or without SP nodes or in a multiple NSB frame.
- If it uses the SP Switch, it connects to a switch port of an existing SP switched frame with or without SP nodes.
- It cannot be within a switch capsule in an SP Switch configuration – in other words, between a switched frame and any non-switched expansion frame using that switch. Give it a frame number that fits between switch capsules.
- It is not supported with the 8-port SP Switch.

SP Expansion I/O Units also do not enter into the determination of the frame and switch capsule configuration. They do not connect to a switch so the switch port corresponding to the slot occupied by an SP Expansion I/O Unit is available to be otherwise used, for instance by an SP-attached server or SP Switch Router. The nodes to which SP Expansion I/O Units are connected must honor the placement convention. These expansion units however can be placed in any unoccupied slot of a frame with nodes or in a non-node frame that is used specifically for them. In a system with the SP Switch, give a non-node frame a number that fits between switch capsules.

Node placement with the SP Switch2

An SP Switch2 system is more flexible than an SP Switch system in regard to node placement. Actually the SP Switch2 has no restrictions on node placement. A node can be placed anywhere it is allowed by the physical constraints. The only SP Switch2 rules have to do with the system configuration and with nodes connecting to the SP Switch2, which are the following:

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

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

4. You can override the default configuration. If you are not going to set up the system based on the default, you need to create an **/etc/plane.info** file to let the system know how the switches are to be configured. See the section on the **/etc/plane.info** file in the book *PSSP: Command and Technical Reference*.

Figure 16 shows an example of a system with a two-plane SP Switch2 configuration.

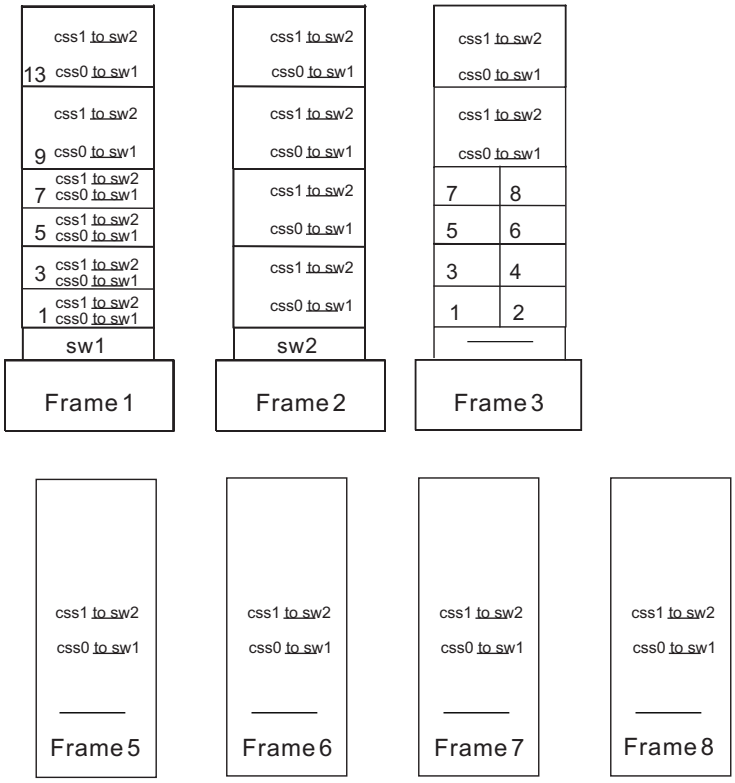


Figure 16. Nodes in a two-plane SP Switch2 configuration

Every node in the system has two switch adapters, **css0** connected to sw1 and **css1** connected to sw2. The **css0** network is **plane 0** and the **css1** network is **plane 1**. Additionally the system is comprised of the following:

- Frame 1 with SP Switch 2 sw1 contains four wide nodes in slots 1, 3, 5, and 7 and two high nodes in slots 9 and 13. These nodes collectively use six switch ports in sw1 and in sw2. Since slots 2, 4, 6, 8, 10, 11, 12, 14, 15, and 16 cannot be occupied by more nodes within the frame, ten switch ports remain available at this point. The high nodes in slots 9 and 13 connect to SP Expansion I/O Units in Frame 3, slots 1 and 2 respectively.
- Frame 2 with SP Switch sw2 has four high nodes at slots 1, 5, 9, and 13. Each of them connect to SP Expansion I/O Units in Frame 3, slots 3, 4, 5, and 6 respectively. Each node also connects to sw1 and sw2, so six switch ports now remain to be claimed on each switch.
- Frame 3 has no switch. It is a non-switched expansion frame. This frame has eight SP Expansion I/O Units, of which six are used by nodes in Frames 1 and 2. It also has two high nodes at slots 9 and 13 connected to the I/O units in slots 7 and 8 and two switch ports in sw1 and sw2, leaving four ports yet to be claimed on each switch.
- Frames 5 through 8 are SP-attached servers, each connected to one of the remaining four ports in each switch. The first of these has frame number 5.

There is still room for upgrades in the future. For instance, by replacing the wide nodes in Frame 1 with two high nodes, another SP expansion frame as Frame 4 can house two more high nodes and SP Expansion I/O Units. With the SP Switch2, node numbers and switch port numbers are automatically generated.

Switch port numbering

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) must be introduced to provide 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 Switch2

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

Switch port 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 evaluate 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. This is demonstrated in Figure 18 on page 128.

For additional explanation with switch port numbers, see Chapter 7, “Planning SP system partitions”, particularly “Example 3 – An SP with 3 frames, 2 SP Switches, and various node sizes” on page 184.

Switch port numbering for an SP Switch-8

Node numbers for short and tall frames are assigned using the same algorithm. See “Node numbering” on page 120.

Note: Extension nodes must be placed into a valid switch port location as verified in the SDR Syspar_map. See “Choosing a valid port on the SP Switch” on page 117.

However, for the SP Switch-8, a different algorithm is used for assigning nodes their switch port numbers. A system with this switch contains only switch port numbers 0 through 7.

The following algorithm is used to assign nodes their switch port numbers in systems with eight port switches:

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

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

Figure 17 and Table 46 contain sample switch port numbers for a system with a short frame and an eight port switch.

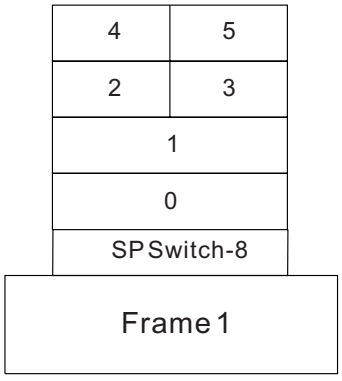


Figure 17. Switch port numbering for an SP Switch-8 in a short frame

Table 46. Sample switch port numbers for the SP Switch-8

Slot Number	Populated?	Node Number	Switch Port Number
1	Yes	1	0
2	No		
3	Yes	3	1

Table 46. Sample switch port numbers for the SP Switch-8 (continued)

Slot Number	Populated?	Node Number	Switch Port Number
4	No		
5	Yes	5	2
6	Yes	6	3
7	Yes	7	4
8	Yes	8	5
9 - 16*	No		

* Slot numbers 9-16 are used only for tall models.

Switch port numbering for a switchless system

You need to plan a switch network, even if you do not plan to use an SP Switch, whenever you plan to use any of the following:

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

unless you force the system to be non-partitionable. See "Understanding partitionability choices" on page 97.

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

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

where *frame_number* is the number of the frame in which the node is located and *slot_number* is the lowest numbered slot the node occupies from 1 to 16. For SP-attached servers, slot number is that of the associated slot in an SP frame where the switch port will not be otherwise used.

Even in a switchless system, you need to fill in the switch worksheet to set Switch Port Number when you plan to use SP-attached servers. You can do this for a clustered server as if it were an SP-attached server if you might add an SP Switch or expand to a standard SP system in the future. During the system installation and configuration process, you will be asked to enter the value. This is because PSSP cannot dynamically determine the value as it can for SP nodes. See "Completing the switch configuration worksheet" on page 63.

IP address assignment

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

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

Figure 18 on page 128 illustrates the switch port numbers for an SP system that uses the SP Switch. It also illustrates how the switch port numbers are set for a non-switched expansion frame and for an SP-attached server. In Figure 18 on page 128

page 128, Switch 2 connects to the nodes in Frame 3. Specifically, the nodes of Frame 3 use respective ports of Switch 2 which are not used by nodes in Frame 2. To determine the switch port number for nodes that are not in a switched frame, use the following formula:

switch_port_number = (*switch_number* - 1) X 16 + *port_number*
switch_port_number = (2 - 1) X 16 + 1
switch_port_number = 17

Based on the formula, the first slot of Frame 3 has switch port number **17**. The formula also results in switch port number 27 for the SP-attached server as frame 4 using switch port 11 in switch number 2.

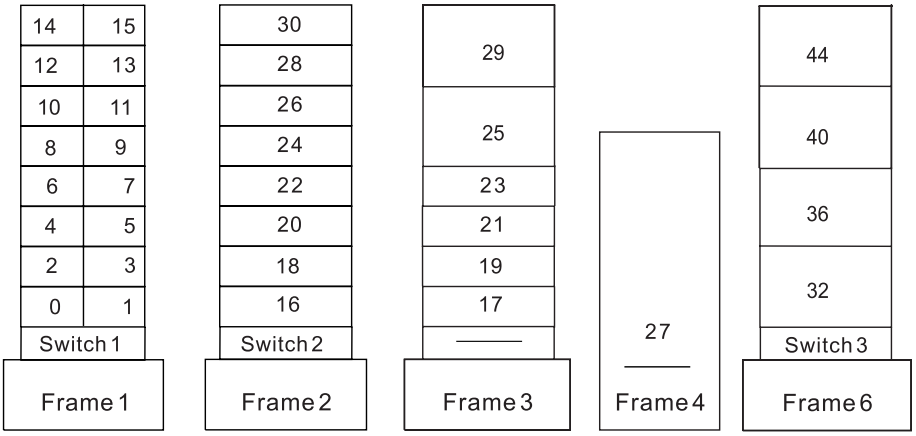


Figure 18. Switch port numbering sequence

Chapter 4. Planning for a high availability control workstation

Planning for a high availability control workstation requires planning for both hardware and software. For planning what to use as your control workstation, see “Question 10: What do you need for your control workstation?” on page 70. For hardware planning information see the book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*. That book describes the hardware components and cabling you need to install and run the High Availability Control Workstation (HACWS) software successfully. For the software requirements, see “Requirements for HACWS configurations” on page 136.

The design of the HACWS component is modeled on the HACMP licensed program. HACWS uses HACMP running on two control workstations in a two-node rotating configuration. HACWS uses an external DASD that is accessed non-concurrently between the two control workstations for storage of SP-related data. There is also a dual RS-232 frame supervisor card with a connection from each control workstation to each SP frame in your configuration. This HACWS configuration provides automated detection, notification, and recovery of control workstation failures.

Note: HACWS has limits and restrictions. Be sure to read “Limits and restrictions” on page 134.

Overall system view of an HACWS configuration

The system looks similar except that there are two control workstations connected to the SP Ethernet admin LAN and the TTY network. The frame supervisor TTY network is modified to add a standby link. The second control workstation is the backup. Figure 19 on page 130 shows a logical view of a high availability control workstation. The figure shows disk mirroring, an important part of high availability planning.

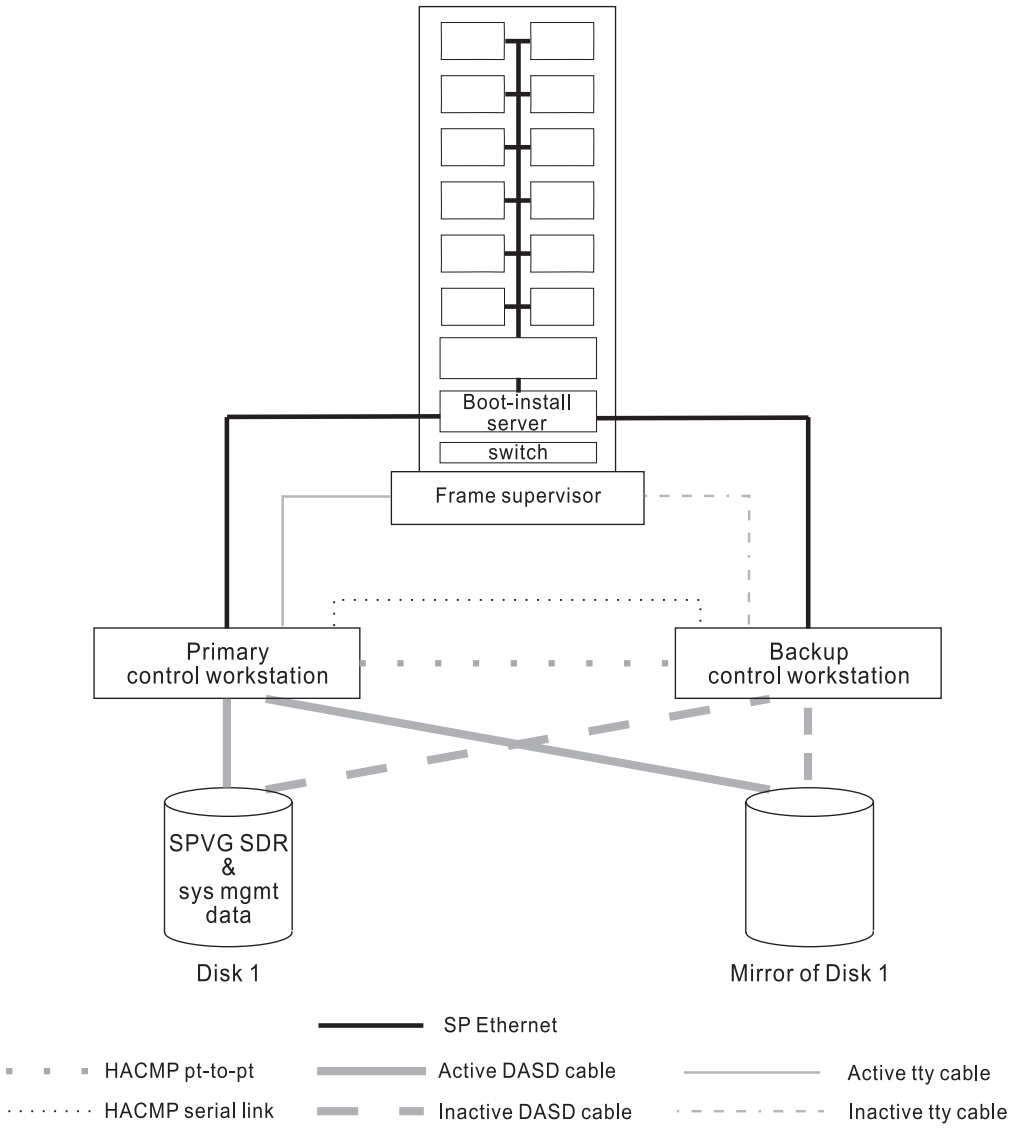


Figure 19. High Availability Control Workstation with disk mirroring

If the primary control workstation fails, there is a disruptive failover that switches the external disk storage, performs IP and hardware address takeover, restarts the control workstation applications, remounts file systems, resumes hardware monitoring, and lets clients reconnect to obtain services or to update control workstation data. This means that there is only one active control workstation at any time.

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 SP data is stored in a separate volume group on the external disk storage.

The backup control workstation can run other unrelated applications if desired. However, if the application on the backup control workstation takes significant resource, that application might have to be stopped during failover and reintegration periods.

Benefits of a high availability control workstation

HACWS is a major component of the effort to reduce the possibility of single point of failure opportunities in the SP system. There are already redundant power supplies and replaceable nodes. However, there are also many elements of hardware and software that could fail on a control workstation. With a high availability control workstation, your SP system will have the added security of a backup control workstation. Also, HACWS allows your control workstation to be powered down for maintenance or updating without affecting the entire SP system.

Difference between fault tolerance and high availability

Before planning whether to use a high availability control workstation, read the following section to understand the difference between high availability and fault tolerance.

Fault tolerance

The *fault tolerant* or *continuous availability* model relies on specialized hardware to detect a hardware fault and instantaneously switch to a redundant hardware component – whether the failed component is a processor, memory board, power supply, I/O subsystem, or storage subsystem.

Although this failover is apparently seamless and offers non-stop service, a high premium is paid in both hardware cost and performance because the redundant components do no processing.

More importantly, the fault tolerant model does not address software failures, by far the most common reason for down time.

High availability

The *high availability* or *fault resiliency* model views availability not as a series of replicated physical components, but rather as a set of system-wide, shared resources that cooperate to provide essential services.

High availability combines software with industry-standard hardware to minimize down time by quickly restoring services when a system, component, or application fails. While not instantaneous, restoring services is rapid, often less than a minute.

The distinguishing factor between fault tolerance and high availability is that a fault tolerant environment offers no service interruption, versus a minimal service interruption in a highly available environment. Many sites are willing to absorb a small amount of down time with high availability rather than pay the much higher cost of providing fault tolerance. Moreover, in most highly available configurations, the backup processors are available for use during normal operation.

IBM's approach to high availability for control workstations

For the reasons already mentioned, IBM has taken the high availability approach to control workstation support for the SP system. The control workstation is a suitable candidate for high availability because it can typically withstand a short interruption, but must be restored quickly. In the SP configuration, the control workstation has been a possible single point of failure.

Eliminating the control workstation as a single point of failure

A *single point of failure* exists when a critical function is provided by a single component. If that component fails, the system has no other way to provide that function and essential services become unavailable.

The key facet of a highly available system is its ability to detect and respond to changes that could impair essential services. The SP with the HACWS software lets a system continue to provide services critical to an installation even though a key system component — the control workstation — is no longer available. When the control workstation becomes unavailable, through either a planned or inadvertent event, the SP high availability component is able to detect the loss and shift that component's workload to a backup control workstation.

Refer to the following tables for some of the consequences of failure of a control workstation that has not been backed up.

Table 47. Effect of CWS failure on mandatory software in a single-CWS configuration

Major software component	**Effect on SP System**
Hardware Monitor	1. No control of SP hardware except for the on/off switch on a node, and the use of the service laptop connected to a frame supervisor cable. 2. Nodes cannot be hot-plugged in or out of the frames controlled by the failed control workstation.
SDR	1. No hardware or software configuration changes can occur. 2. No installations can be started. 3. A switch fault will not complete processing and the switch will remain in service mode if a fault occurs while the control workstation is unavailable. 4. No cluster shutdowns can occur. 5. A node can still be powered off and on manually, but this causes a switch fault.
Kerberos V4 Authentication Server (if no backup server exists)	1. Users cannot obtain new tickets via kinit. 2. Background processes using rcmdtgt to get ticket will fail. 3. Users cannot change passwords. 4. New users cannot be added to the authentication database.
Diagnostics	Diagnostics cannot be run on node boot disks.
File Collections Master	No new distributed file updates can occur.
Availability subsystems (hats, hags, haem)	These subsystems will not restart upon node reboot.

Table 48. Effect of CWS failure on user data on the CWS

Major software component	Effect on SP System
User Management	You cannot make changes to a user data base stored on the control workstation.
Hardware Logging Daemon	1. Hardware logging immediately stops. 2. Nodes cannot be hot plugged.
Error Logging Alerts	If sent by **mail** will be put in the node mail spool.
Accounting Master	1. No consolidated accounting records are kept during down time. 2. Records are consolidated after the control workstation comes up.
User File Server	1. Running jobs might fail. 2. Jobs might not be able to access needed data.

Consequences of a high availability control workstation failure

When a failure occurs in a high availability control workstation, the following steps take place automatically:

- The external disk storage is switched to the backup control workstation.
- The hardware and IP addresses are switched to the backup control workstation.
- The control workstation applications are restarted.
- The file systems are remounted.
- Hardware monitoring is resumed.
- Clients are allowed to reconnect to obtain data or to update control workstation data.

Note: See “Limits and restrictions” on page 134 for limitations with respect to RS/6000 and pSeries servers.

System stability with HACWS

When a control workstation fails, it causes significant loss of function in configuration, systems management, hardware monitoring, and the ability to handle a switch fault. The reliability of the whole system is compromised by the chance of a switch fault during a control workstation outage. Using the high availability control workstation increases the mean time before failure (MTBF) of the entire system.

The failover is disruptive. Applications at the control workstation that are interrupted will not resume automatically and must be restarted. The interruption is momentary. Applications within nodes, that require no communication with the control workstation might not notice the failover. Applications relying on data from the SDR will be momentarily interrupted. Having a backup control workstation available prevents this problem.

Occasionally, you might need to take a control workstation down to maintain the hardware or software or to repair or update a component of the system. Using high availability control workstation lets you schedule this upkeep without taking the entire system down. The serviceability of the SP is increased by the service time for the control workstation, which increases the mean time to repair (MTTR) of the system as a whole.

Related options and limitations for control workstations

Some configuration options that can make your control workstation more available are separate from the high availability control workstation program. They include disk mirroring, uninterruptible power supplies, and dual disk controllers both internal and external. You must also be aware of any frame supervisor changes, HACWS limits and restrictions, and complete the related HACMP planning worksheets.

Uninterruptable power supply

An uninterruptable power supply can supply electricity to a device to keep it running when main power is interrupted or is unreliable. Usually an uninterruptible power supply is not the sole source of power. Rather, it is typically used to smooth a fluctuating source or to provide enough power to enable a device to shut down gracefully. You can use an uninterruptible power supply in conjunction with all other means of assuring control workstation reliability. See the RS/6000 General Services Document *Site and Hardware Planning Information* for the power consumption requirements of your control workstation.

Power independence

Each control workstation should be attached to a different electrical power source or breaker panel if possible. They should at least be on separate circuits so that maintenance or failures in main power can affect only one control workstation.

Single control workstation with disk mirroring

The process of mirroring occurs when each block of data written to one disk is also written to another disk. You always have a copy of your data in case one disk or disk adapter fails. As a middle ground to availability you can decide to have a single control workstation and mirror the root volume group to provide better availability of the control workstation. This involves two or three times the number of disks needed for the root volume group, depending on whether you want one or two mirrors of the original. See the discussion on mirroring in the book *AIX System Management Guide: Operating System and Devices*. That book describes the AIX support for mirroring root volume groups. See the book *PSSP: Administration Guide* for additional SP-specific support of mirroring root volume groups.

Spare Ethernet adapters

You can cable spare SP Ethernet adapters into the existing Ethernet LAN segments for the SP and leave them in a defined but unavailable configuration state. When an Ethernet adapter fails, you can unconfigure the failing adapter and configure the spare Ethernet adapter for that LAN segment. You can use the spare adapter until the failed one is repaired or replaced. (Note that with older 10 Mb Ethernets, the spare Ethernet adapter still counts as one of the stations in the 30 total stations you can have on an Ethernet LAN segment.)

Frame supervisor changes

Check with your IBM representative for information about ordering the necessary hardware.

Limits and restrictions

The high availability control workstation support has the following limitations and restrictions:

- DCE restriction:

If you plan to have DCE authentication enabled, you cannot use HACWS. If you already use HACWS, do not enable DCE authentication.

- There are no problems in the authentication of both systems on the nodes with the restricted root access option enabled. However, you have to copy the authorization files, **/.rhosts** or **/.klogin**, to the backup control workstation. Also, it is important that you activate restricted root access from the active primary control workstation, since it is the Kerberos V4 Master.
- IPv6 restriction:

  HACWS does not tolerate IPv6 aliases for IPv4 addresses.
- You cannot split the load across a primary and backup control workstation. Only one or the other provides all the function at any given time.
- The primary and backup control workstations must each be a supported and independent control workstation. You cannot use a PSSP node 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 your primary control workstation is a PSSP Kerberos V4 authentication server, the backup control workstation must be a secondary authentication server.

HACWS is not supported in any configuration of clustered servers. HACWS is also not supported with SP-attached servers that use the CSP or HMC hardware protocol. Generally, it is not a good idea to run HACWS on systems with SP-attached servers.

However, if you already have HACWS running on your SP system, have reasoned that you cannot do without HACWS, accept all the limitations, have specialized experience, and can implement your own manual intervention procedures for fail over, you might be able to make HACWS continue to work in your SP system with SP-attached servers that use the SAMI hardware protocol. IBM @server pSeries 680 and RS/6000 Enterprise Servers S70, S7A, and S80 connect to the control workstation by two serial connections, making them SP-attached servers in an SP system. One connection is for hardware monitoring and control and the other is for serial terminal support. These servers use the SAMI hardware protocol for those connections. Only one control workstation at a time can be connected to each server, so there cannot be automatic physical failover done by the HACWS software. When the primary control workstation fails over to the backup control workstation, hardware control and monitoring support and serial terminal support are not available for these servers.

The following apply if you use high availability control workstation support with SAMI protocol RS/6000 or pSeries servers in an SP-attached or clustered servers configuration:

- Each server is directly attached to the control workstation. Depending on the type of server, you might only be able to connect to one control workstation at a time. Using HACWS might require manual intervention or might not be an option at all. Check the control workstation hardware specifications carefully.

- 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 the servers:
  - The 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.
  - You cannot use PSSP to make configuration changes related to the servers. For example, you cannot add new servers.
  - You cannot use PSSP to reboot a server.
  - You can use PSSP to shutdown or restart the SP system but it will not effect SP-attached servers.
  - You cannot use PSSP to shutdown or restart a system of clustered servers.

For a list of the monitoring and control functions that might be lost depending on your server hardware, see the discussion of SP-attached and clustered servers in the HACWS chapter of the book *PSSP: Administration Guide*.

Completing planning worksheets for HACWS

You will have to complete the following worksheets in the HACMP documentation:

- Shared Volume Group/File System Worksheet (Non-Concurrent)
- Defining Shared LVM Components for Non-Concurrent Access

As you complete the HACMP planning and installation steps, take the Non-Concurrent option whenever you are given the choice.

See the book *HACMP: Planning Guide*, for complete planning information.

Requirements for HACWS configurations

The software requirements for the high availability control workstation include:

- Two AIX server licenses, one for each control workstation.

  AIX 5L 5.1 is required for PSSP 3.5. See the *Read This First* document for the latest information on what modification levels of AIX are supported with PSSP 3.5.
- Two licenses for IBM C for AIX or the batch C and C++ compilers and runtime libraries of VisualAge C++ Professional 6.0 for AIX or later.

  If the compiler's license server is on the control workstation, the backup control workstation should also have a license server with at least one license. If there is no license server on the backup control workstation, an outage on the primary control workstation will not allow the SP system access to a compiler license.
- Two licenses and software sets for High Availability Cluster Multi-Processing for AIX (HACMP or HACMP/ES).

  Both the client and server option must be installed on both control workstations. See the appropriate HACMP documentation for the latest information on which levels of HACMP are supported with AIX 5L 5.1 and PSSP 3.5.
- PSSP optional component HACWS

  This is the customization software that is required for HACMP support of the control workstation. It comes with your order of PSSP 3.5 as an optionally installable component. You will need to install a copy on both control workstations.

Planning and using the backup control workstation will be simpler if you configure your backup control workstation identical to the primary control workstation. Be sure you have all the related cables, hardware, and software your SP system might need. Some components must be identical, others can be similar. For example, the TTY assignments on each must be identical and should be configured in the same slots on each. If you have the same number and type of disks on each, your planning and operation will be simpler. Otherwise you might have to plan recovery scripts that address HD0 on one control workstation and HD3 on the other.

Planning your HACWS network configuration

Planning your HACWS network configuration is a complex task which requires understanding the basic HACMP concepts. These concepts are explained in the HACMP publications. This section demonstrates how to plan your HACWS network configuration through a hypothetical situation. Additional specific HACWS network requirements are also described in this section.

Assume that your system has a single control workstation named *dutchess.xyz.com* and it will serve as the primary control workstation after you install HACWS. The workstation you add will become the backup control workstation. The name of the backup control workstation is *ulster.xyz.com*.

The SP nodes get control workstation services by accessing the network interface whose name matches the host name of the primary control workstation. In this example, the SP nodes get control workstation services by accessing *dutchess.xyz.com*. If the primary control workstation fails and the backup control workstation takes over, the backup control workstation assumes the network identity of *dutchess.xyz.com*.

The *dutchess.xyz.com* network interface gets configured on the control workstation currently providing the control workstation services. HACMP refers to *dutchess.xyz.com* as a **service address** (or service interface). The primary control workstation must use a different network address when it reboots in order to avoid a network address conflict between the two control workstations. HACMP refers to this alternate network address as a **boot address** (or boot interface). In this example, the boot address of the primary control workstation is *dutchess_bt.xyz.com*.

In addition, HACWS requires that the backup control workstation must always be reachable via a network interface whose name matches its host name. In this example, this name is *ulster.xyz.com*. This network interface does not get identified to HACMP. If you have no available adapter upon which to configure the *ulster.xyz.com* network interface, you can use an IP address alias.

Each control workstation in this example configuration contains one Ethernet adapter, connected to the SP Ethernet network. After the two control workstations are booted and before HACMP is started, their network configuration looks like the one illustrated in Figure 20 on page 138.

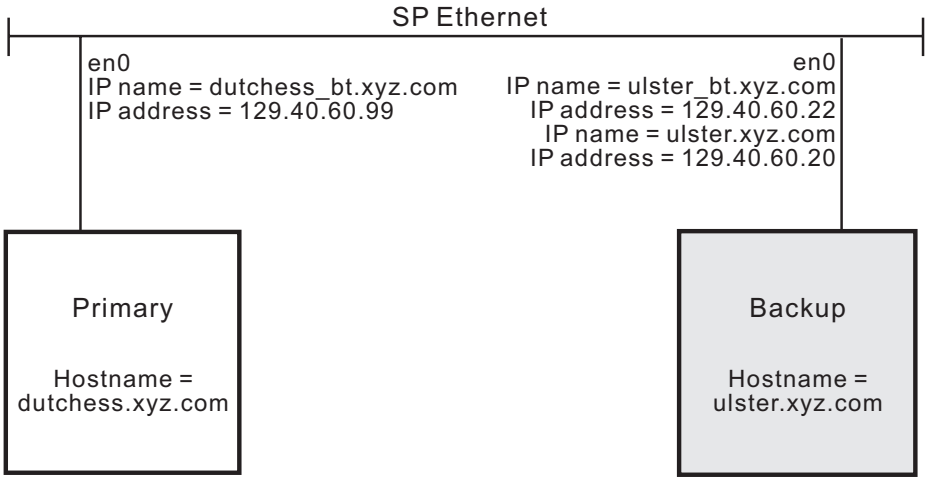


Figure 20. Initial control workstation network configuration

At this point, neither machine is providing control workstation services, so the *dutchess.xyz.com* network interface is not available. The Ethernet adapter on the primary is configured with its boot address *dutchess_bt.xyz.com* and the Ethernet adapter on the backup is configured with its boot address *ulster_bt.xyz.com*. Since there is only one network adapter, the network interface *ulster.xyz.com* must be configured as an IP address alias on the backup control workstation.

Note: Both IP addresses 129.40.60.22 and 129.40.60.20 are assigned to the adapter **en0** on the backup control workstation. If another network adapter is available, you do not have to use an IP address alias.

When the operator starts HACMP on both control workstations, the first control workstation to start HACMP becomes the active control workstation. (The operator selects the machine to become the active control workstation by starting HACMP on it first.) If HACMP is first started on the primary control workstation and then on the backup control workstation, the network configuration looks like the one illustrated in Figure 21.

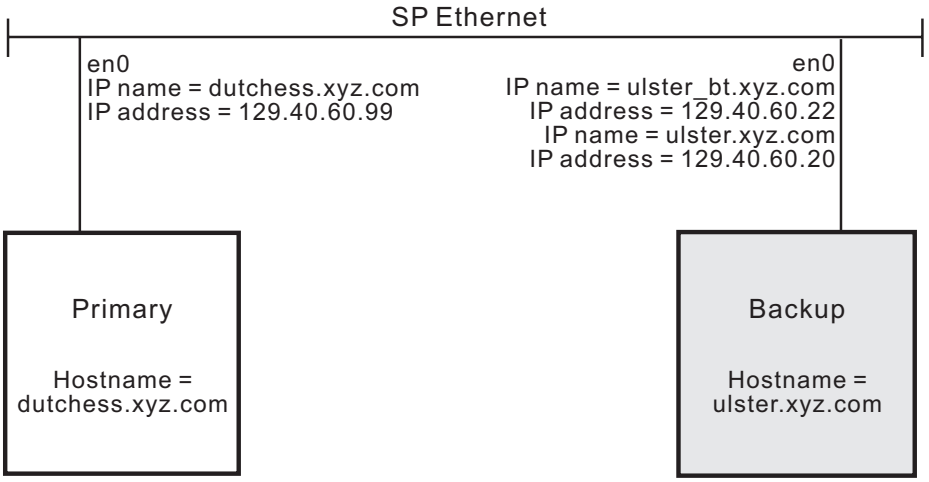


Figure 21. Starting HACMP

The only change to the network configuration is that the **boot address** *dutchess_bt.xyz.com* on the primary control workstation has been replaced by the **service address** *dutchess.xyz.com*.

If the primary control workstation should fail and the backup control workstation take over, the network interface looks like the one illustrated in Figure 22.

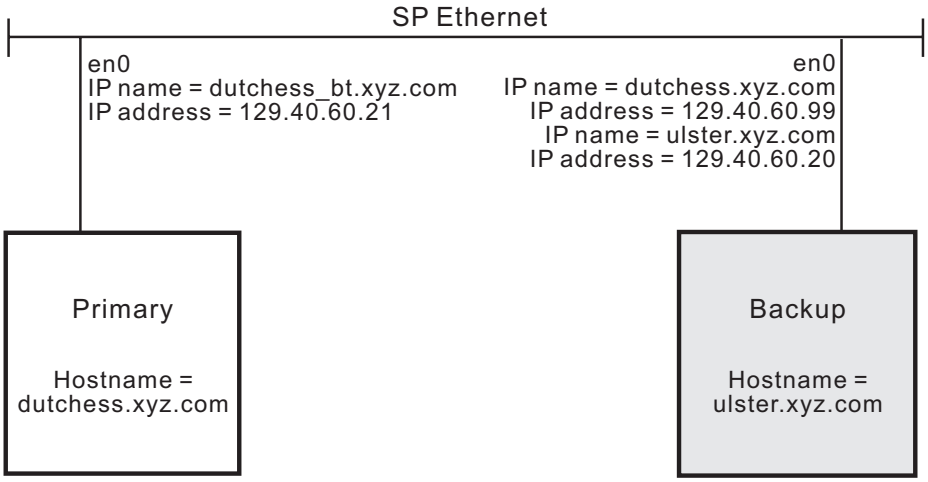


Figure 22. Control workstation failover

If the primary control workstation is still running, then its Ethernet adapter is back on its boot address *dutchess_bt.xyz.com*, and the boot address *ulster_bt.xyz.com* on the backup control workstation has been replaced by the service address *dutchess.xyz.com*. The SP nodes continue to get control workstation services by accessing *dutchess.xyz.com*.

Note: The network interface *ulster.xyz.com* remains configured on the backup control workstation.

You can identify multiple network interfaces to move back and forth between the two control workstations along with the control workstation services. Some possible reasons for doing this are:

- You have an SP system with a large number of nodes and multiple Ethernet adapters on the control workstation connected to the SP Ethernet network.
- You want the control workstation to provide a separate network interface for each SP system partition.
- You want a network interface on an external network to allow workstations outside of the SP system to transparently access the active control workstation.

Each of these network interfaces is effectively a service address. However, the number of service addresses identified to HACMP cannot exceed the number of network adapters. Use IP address aliases to make up the difference.

In this example, each control workstation has only one network adapter. Since *dutchess.xyz.com* is defined to HACMP as a service address, any additional "effective" service addresses must be configured using IP address aliases. If you added an SP system partition whose network interface name on the control workstation is *columbia.xyz.com* to this example configuration, it would look like Figure 23 on page 140 when the backup control workstation is active.

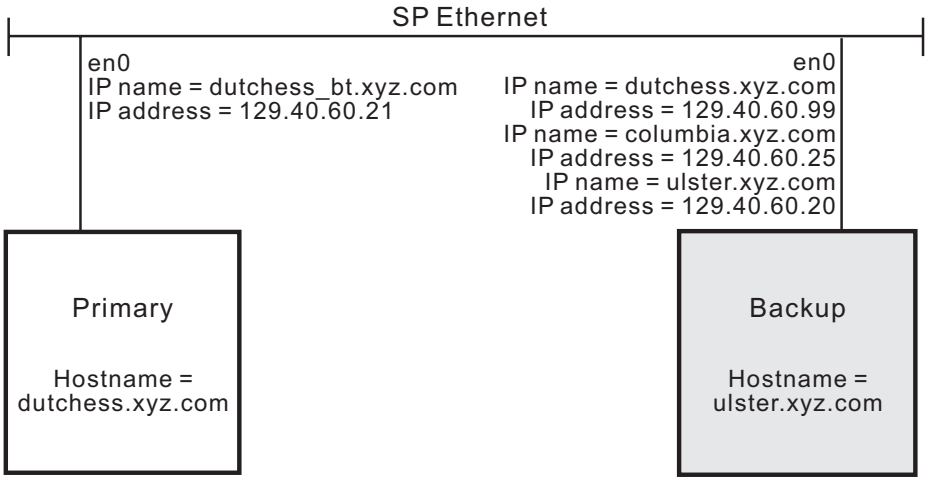


Figure 23. Adding an SP system partition

The HACMP service address *dutchess.xyz.com* is configured on adapter **en0** on the backup control workstation and the network interfaces *columbia.xyz.com* and *ulster.xyz.com* are configured on adapter **en0** as IP address aliases. The service address *dutchess.xyz.com* is identified to HACMP. **For each service address that is identified to HACMP, there must be boot addresses for both control workstations.** The boot address *dutchess_bt.xyz.com* is identified to HACMP for the primary control workstation, and the boot address *ulster_bt.xyz.com* is identified to HACMP for the backup control workstation

At this point, if you have not done so already, you need to do the following:

1. Determine the control workstation service addresses for your configuration.
2. Determine which service addresses should be identified to HACMP and which service addresses need to be configured using IP address aliases. **The host name of the primary control workstation** (*dutchess.xyz.com*) **must always be identified to HACMP as a service address.** Remember the number of service addresses identified to HACMP cannot exceed the number of network adapters.
3. Determine the boot addresses for your configuration. The number of boot addresses on each control workstation will match the number of service addresses defined to HACMP. For example, if you identify three service addresses to HACMP, then you need to identify six boot addresses — three boot addresses on each control workstation.
4. Make sure the host name of the backup control workstation (*ulster.xyz.com*) is always a valid network interface on the backup control workstation.
5. If your site uses a name server, make sure that all of these network interfaces have been added to your name server.

Chapter 5. Planning for IBM Virtual Shared Disks

Without special programming, a physical disk connected to a node can only be accessed by applications running on that node. The IBM Virtual Shared Disk component lets you define IBM Virtual Shared Disks. It provides the special programming that allows applications running on multiple nodes within the same switch partition to access the data on a raw logical volume as if it were local at each of the nodes. Actually, the logical volume is located at *one* of the nodes called a *server* node. The concurrent IBM Virtual Shared Disk support allows you to use the concurrent disk access environment supplied by AIX.

The IBM Recoverable Virtual Shared Disk subsystem is a required component for IBM Virtual Shared Disk support that enhances IBM Virtual Shared Disk availability by multi-host attaching the physical disk to another node which takes over the I/O service if the IBM Virtual Shared Disk node or communication adapter fails. You configure nodes as primary and secondary server nodes of IBM Virtual Shared Disks. It offers continuous access to data with transparent recovery in the event of the failure of a node, disk, disk adapter, disk cable, or communication adapter. Recovery from an SP switch adapter failure is the same as from a node failure – control of connected multi-host attached volumes is passed to the secondary server.

Note: IBM Virtual Shared Disks are supported only on system configurations with an SP switch or SP Switch2. In an SP Switch2 system where some nodes are not on the switch, IBM Virtual Shared Disks can work only with those nodes that are on the switch.

Enhanced security options:

You have the option of running PSSP with an enhanced level of security. The restricted root access option removes the dependency PSSP system management software otherwise has to internally issue **rsh** and **rcp** commands as a root user from a node. With restricted root access active, any such actions can only be run from the control workstation or from nodes configured to authorize them and PSSP does not automatically grant authorization for a root user to issue **rsh** and **rcp** commands from a node. If you enable this option some procedures might not work as documented. For example, to run HACMP an administrator must grant the authorizations for a root user to issue **rsh** and **rcp** commands that PSSP would otherwise grant automatically. See “Considering restricted root access” on page 146 for a description of this option.

You can use a secure remote command process to be run by the PSSP system management software in place of the **rsh** and **rcp** commands. See “Considering a secure remote command process” on page 151 for a description of this option.

IBM Virtual Shared Disks are not supported in this enhanced security environment. See “Limitations when using restricted root access” on page 149 and “Considering choosing none for AIX remote command authorization” on page 152 for a complete list of limitations.

This chapter is the first step in planning to use the optional components of PSSP that help you create and use IBM Virtual Shared Disks. After reading this chapter, if you plan to use IBM Virtual Shared Disks on a system you already have, see

Chapter 11, “Planning for migration” on page 213 for versions supported, coexistence, and migration information. Whether you plan to use them on an existing or a new system, see the book *PSSP: Managing Shared Disks* for additional planning information, to prepare for migrating to or installing PSSP 3.5, for creating and using IBM Virtual Shared Disks, and for more information about running under various PSSP security configurations.

Planning for the IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk optional components of PSSP

Plan how you are going to use the IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk software before you install the hardware. Consider the following:

- The IBM Virtual Shared Disk subsystem lets you assign volume groups located on any physical disks in any node within a partition. Volume groups assigned to be used by the IBM Virtual Shared Disk subsystem ought to be used only to define IBM Virtual Shared Disks.
- Each IBM Virtual Shared Disk cluster must be in the same system partition. You can have separate IBM Virtual Shared Disk clusters in separate system partitions, but they cannot communicate directly with each other. See Chapter 7, “Planning SP system partitions” on page 171 for system partition planning information.
- In order to provide recovery, the IBM Recoverable Virtual Shared Disk software requires multi-host attached disk storage, which must be installed before you define or use the IBM Virtual Shared Disks. It also requires a minimum of two SP nodes.

See the book *PSSP: Managing Shared Disks* for additional explanation of these and other dependencies and restrictions.

Planning for IBM Virtual Shared Disk communications

When you define IBM Virtual Shared Disks, you specify the SP Switch or other connection method. See the book *PSSP: Administration Guide* for detailed node and disk connection information. See the relevant edition of the AIX book *System Management Guide: Operating System and Devices* to read about the Logical Volume Manager component if you are not already familiar with it.

By design of the Logical Volume Manager, each logical partition maps to one physical partition and each physical partition maps to a number of disk sectors. The Logical Volume Manager limits the number of physical partitions that it can track per disk to 1016. In most cases, not all the 1016 tracking partitions are used by a disk. The default size of each physical partition during a **mkvg** command is 4 MB, which implies that individual disks up to 4 GB can be included into a volume group.

If a disk larger than 4 GB is added to a volume group (based on usage of the default 4 MB size for the physical partition), the disk addition fails. The warning message provided will be: `The physical partition size of`*number A* `requires the creation of` *number B* `partitions for hdisk`*X*.

The system limitation is 1016 physical partitions per disk. Specify a larger physical partition size in order to create a volume group on this disk. Note that the size of the partition determines the granularity by which logical volumes (and file systems) could be increased in size in a given volume group definition. Moreover, this setting

could not be overridden once a volume group is defined. If you intend to dedicate DASD for a database with large tables on external DASD, you should consider using a large partition size.

There are two instances where this limitation is enforced:

1. You try to use **mkvg** to create a volume group and the number of physical partitions on a disk in the volume group exceeds 1016. A work-around to this limitation is to select from the physical partition size ranges of: 1, 2, (4), 8, 16, 32, 64, 128, 256 Megabytes and use the **mkvg -s** option.
2. The disk that violates the 1016 limitation attempts to join an existing volume group with the **extendvg** command.

You can recreate the volume group with a larger partition size allowing the new disk to work or create a stand-alone volume group consisting of a larger physical size for the new disk. If the install code detects that the rootvg drive is larger than 4 GB, it will change the **mkvg -s** value until the entire disk capacity can be mapped to the available 1016 tracks. This install change also implies that all other disks added to rootvg, regardless of size, will also be defined at that physical partition size. For RAID systems, the /dev/hdiskX name used by the Logical Volume Manager might really consist of many non-4 GB disks. In that case, the 1016 requirement still exists. Logical Volume Manager is not aware of the size of the individual disks that really make up /dev/hdiskX. Logical Volume Manager bases the 1016 limitation on the AIX-recognized size of /dev/hdiskX, and not the real physical disks that make up /dev/hdiskX.

In some instances, you will experience a problem adding a new disk to an existing volume group or in creating a new volume group. The warning message provided by Logical Volume Manager will be: *Not enough descriptor area space left in this volume group.*

Either try adding a smaller PV or use another volume group. On every disk in a volume group, there exists an area called the volume group descriptor area (VGDA). This space allows you to take a volume group to another AIX system and importvg the volume group into the AIX system. The VGDA contains the names of disks that make up the volume group, their physical sizes, partition mapping, logical volumes that exist in the volume group, and other pertinent LVM management information. When you create a volume group, the mkvg command defaults to allowing the new volume group to have a maximum of 32 disks in a volume group. However, as bigger disks have become more prevalent, this 32 disk limit is usually not achieved because the space in the VGDA is used up faster, as it accounts for the capacity on the bigger disks. This maximum VGDA space, for 32 disks, is a fixed size which is part of the LVM design. Large disks require more management mapping space in the VGDA, causing the number and size of available disks to be added to the existing volume group to shrink. When a disk is added to a volume group, not only does the new disk get a copy of the updated VGDA, but all existing drives in the volume group must be able to accept the new, updated VGDA. The exception to this description of the maximum VGDA is rootvg. In order to provide users more free disk space, when rootvg is created, mkvg does not use the maximum limit of 32 disks that are allowed into a volume group. Instead, the number of disks picked in the AIX install menu is used as the reference number by mkvg -d during the creation of rootvg. This limit does not prohibit you from adding more disks to rootvg during post-install. The amount of free space left in a VGDA, and the number size of the disks added to a volume group, depends on the size and number of disks already defined for a volume group. If you require more VGDA

space in the rootvg, then use the **mksysb** and **migratepv** commands to reconstruct and reorganize your rootvg (the only way to change the -d limitation is recreation of a volume group).

Note: IBM suggests that you do not place user data onto rootvg disks. This separation provides an extra degree of system integrity.

The logical volume control block (LVCB) is the first 512 bytes of a logical volume. This area holds important information such as the creation date of the logical volume, information about mirrored copies, and possible mount points in the journaled file system (JFS). Certain Logical Volume Manager commands are required to update the LVCB, as part of the algorithms in Logical Volume Manager. The old LVCB is read and analyzed to see if it is a valid. If the information is valid LVCB information, the LVCB is updated. If the information is not valid, the LVCB update is not performed and the following warning message is issued: *Warning, cannot write lv control block data*

Most of the time, this is a result of database programs accessing raw logical volumes (and bypassing the JFS) as storage media. When this occurs, the information for the database is literally written over the LVCB. Although this might seem fatal, it is not the case. Once the LVCB is overwritten, you can still do the following:

- Expand a logical volume
- Create mirrored copies of the logical volume
- Remove the logical volume
- Create a journaled file system to mount the logical volume.

There are limitations to deleting LVCBs. The logical volumes with deleted LVCB's face possible, incomplete importation into other AIX systems. During an importvg, the Logical Volume Manager command scans the LVCB's of all defined logical volumes in a volume group for information concerning the logical volumes. If the LVCB is deleted, the imported volume group will still define the logical volume to the new AIX system, which, is accessing this volume group, and you can still access the raw logical volume. However, any journaled file system information is lost and the associated mount point will not be imported into the new AIX system. You must create new mount points and the availability of previous data stored in the file system is not assured. Also, during this import of logical volume with an erased LVCB, some non-JFS information concerning the logical volume, which is displayed by the **lslv** command, cannot be found. When this occurs, the system uses default logical volume information to populate the logical volume's ODM information. Therefore, some output from **lslv** will be inconsistent with the real logical volume. If any logical volume copies still exist on the original disks, the information will not be correctly reflected in the ODM database. Use **rmlvcopy** and **mklvcopy** commands to rebuild any logical volume copies and synchronize the ODM.

Chapter 6. Planning for security

Analyzing company resources and protecting these resources is usually part of an administrator's duties. Your resources pertaining to PSSP include the following:

- System and application data, user programs, and user data.
- Communication devices and communication access methods.
- Login permissions – specifically, who can log in, when, and how much resource each user can have.
- Authentication, passwords, and credentials.

This chapter describes the choices you can make and what to prepare when planning for security before installing and configuring, or migrating to and configuring PSSP 3.5. The following are prerequisites for making educated choices for security with PSSP:

1. You ought to have a clearly expressed and understood security policy for your organization.
2. You ought to already understand security of computer systems in general and be familiar particularly with the security services that you can use on the Cluster 1600 system managed by PSSP (SP system) to help enforce that security policy:
   - The Distributed Computing Environment (DCE) security services.
   - The PSSP implementation of Kerberos V4.
   - The standard AIX security service.
   - The IETF Secure Shell protocol.
3. You ought to understand the factors on which to base your choices – the degree and granularity of protection necessary across the entire Cluster 1600 system managed by PSSP and within each SP system partition – and which services offer the level of protection you require.

To become familiar with security terminology and concepts as applied on a system managed by PSSP, see the book *PSSP: Administration Guide*. The chapter called ″Security features of the PSSP software″ explains terminology and basic concepts of security services on the Cluster 1600 system managed by PSSP. It includes suggested reading for DCE and for a secure remote command environment.

Choosing authentication options

Networked applications have client and server parts executing on different hosts. An authentication service allows networked applications to determine their mutual identities for security. This authentication service is provided by one or more authentication servers (daemons) running on systems that are accessible from application client and server systems. An authentication server provides the credentials for the application client and server to perform the authentication task. In a given authentication implementation, when there is more than one authentication server, one is the primary server and all others are replica or secondary servers. You can choose which authentication services to use on your system from the set of those supported in PSSP 3.5.

To consider which authentication service is right for you, answer the following questions:

1. How restrictive does your system need to be?
2. Do you require a level of security that AIX alone does not offer?

3. Does the PSSP implementation of Kerberos V4 offer the level of security you need?
4. Do the DCE security services offer the level of security you need?
5. Do you want to restrict the PSSP installation, configuration, and system management software, hereafter called simply the PSSP system management software, from using the **rsh** and **rcp** commands as a root user from a node? See "Considering restricted root access".
6. Do you want the PSSP system management software to use a secure remote command process of your choice in place of **rsh** or **rcp** commands? See "Considering a secure remote command process" on page 151.

In considering how restrictive you need to be, think about how users connect to your system. Are they connected directly, through the Internet, another external network, a protected LAN? Even if you want all the data available to everyone, you probably still want to protect it from being inadvertently changed or destroyed. Only you can judge which security services offer enough control and protection for your needs. Of those supported on the Cluster 1600 system managed by PSSP, the DCE security services offer the greatest granularity of control. There are also other options that you can choose in addition to the security services.

DCE restriction:
If you plan to have DCE authentication enabled, you cannot use HACWS. If you already use HACWS, do not enable DCE authentication. Do not use IPv6 aliases if you plan to use DCE.

The AIX trusted computing base option:
The AIX mksysb image (SPIMG) that comes with PSSP does not have the AIX trusted computing base option enabled. The AIX trusted computing base option and PSSP security management do not conflict. If you want to enable the AIX trusted computing base option, you must create and use your own customized mksysb. Customization to enable that option can only be done during your installation phase prior to creating the mksysb. See the information about working with the AIX image in the book *PSSP: Installation and Migration Guide*.

Considering secure File Collections

The supman user ID is used by PSSP when distributing file collections. Not managing a supman AIX user ID password is a security risk. PSSP 3.5 provides commands to initially set and routinely change the supman password. See the section ″Establish password for secure file collections,″ in the ″Installing and Configuring a new RS/6000 SP system″ chapter of *PSSP: Installation and Migration Guide*.

Considering restricted root access

The PSSP and PSSP-related software supports using the AIX authenticated remote commands (**rsh, rcp, rlogin**), plus the **ftp** and **telnet** commands for system management. Those commands are available for general use and support multiple authentication methods. Several PSSP components and PSSP-related licensed programs rely on the ability to use the **rsh** command to issue remote commands as the *root* user from a node to be run on the control workstation or from the control

workstation to be run on nodes. Likewise, several components use the **rcp** command to copy files between the control workstation and nodes. In order to provide this capability, PSSP has effectively defined *one root user across the entire system*. Gaining access as root on any node implies that you can gain access as root on the control workstation and all other nodes in the system.

When, for instance, the Cluster 1600 system managed by PSSP is used as a server consolidation system, this single root user might not be desirable. The ***restricted root access*** option is available to limit the use of **rsh** and **rcp** within PSSP system management. It restricts the PSSP and PSSP-related software from automatically using **rsh** and **rcp** commands as root user from a node. When restricted root access is active, any such actions can only be run from the control workstation or from nodes explicitly configured to authorize them. It restricts root **rsh** and **rcp** authorizations from the nodes to the control workstation, but permits control workstation to node **rsh** and **rcp** access.

At the time restricted root access is activated, the PSSP and the root-owned remote command authorizations are reconfigured such that:

- When run as a root process on the control workstation, PSSP software can continue to issue **rsh** and **rcp** commands to any node. Authorization file entries are created by PSSP on the nodes for which a root process on the control workstation can obtain the necessary credentials to issue these commands.
- PSSP software run as a root process on a node no longer requires the capability to issue **rsh** and **rcp** commands to the control workstation or to any other PSSP node. PSSP no longer creates authorization file entries on the control workstation or nodes that grant remote command access to a root process on a node.

When PSSP system management on the nodes needs to copy or access files or run commands on other nodes or on the control workstation, sysctl is used to perform the task from the control workstation. The restrictions imposed by using restricted root access have implications for the functionality of some PSSP and PSSP-related components as well as system management.

Restricted root access can be activated under all the security configurations supported. However, restricted root access is most relevant when you use a secure remote command process and with PSSP security methods `dce`, `dce:compat`, and `compat`. Otherwise, restricted root access under the minimal security state `none/std` is of little value.

For the remainder of this section, all references to access control lists (ACLs) apply to root owned sysctl and AIX remote command ACLs because they apply to the different PSSP security configurations.

How does it work?

Restricted root access can be enabled by setting a site environment attribute in the SDR. Because the SDR in non-DCE environments can be changed by root automatically, the `SP_Restricted` class was created with the `restrict_root_rcmd` attribute. It can only be changed by user `root` on the control workstation. The default value of `restrict_root_rcmd` is `false`.

You can change the attribute by using the *Site environment* SMIT panel or by using the **spsitenv** command. When changed, the authorization files on the control workstation and the nodes are immediately updated. All of the PSSP commands that require **rsh** or **rcp** access check the attribute each time access is required. They automatically use the sysctl method when restricted root access is enabled.

AIX and PSSP remote commands: AIX remote command authorization files for root on the nodes and the control workstation are generated and updated by the **updauthfiles** script. This script is run at various times including when the restricted root access state is changed, when the security settings on the system are modified, and when a node is installed or booted.

When restricted root access is enabled, **updauthfiles** removes all PSSP-generated entries in the remote command authorization files and adds the following entries, which are dependent on the authentication method set for `auth_root_rcmd`:

- Standard AIX: **/.rhosts** is modified to contain only:
  - *cwsname*
  - *additional cws interface names*
- Kerberos V4: **/.klogin** is modified to contain only:
  - rcmd.*cwsname*@*realm* on both the control workstation and nodes.
  - root.admin@*realm* in the **/.klogin** file on the control workstation only.
- Kerberos V5: **/.k5login** is modified to contain only:
  - ssp/*cwsname*/spbgroot@*realm*
  - hosts/*cwsname*/self@*realm*

Any manual changes previously made to these files are not removed automatically. Before PSSP 3.2, the **/.klogin** entries generated by PSSP were not saved in a separate file. If you have migrated your system from an earlier release than PSSP 3.2, there might be PSSP-generated entries in your **/.klogin** file that are no longer recognized. It is a good idea to check the content of all root remote command authorization files after activating restricted root access.

The sysctl facility: The sysctl facility is an authenticated client-server system for running commands and TCL scripts remotely and in parallel. The sysctl server daemon sysctld processes all sysctl client requests for the node on which it runs. There is a sysctld daemon running on each PSSP node as well as on the control workstation.

By default, the **/etc/sysctl.conf** configuration file is read and interpreted each time the sysctl server is started or restarted. This file contains the definitions of the commands that can be executed by sysctl. Additional sysctl commands are supplied with PSSP to enable restricted root access. Since the contents of the **/etc/sysctl.conf** file on the control workstation are different from the one on the nodes, the control workstation version should not be included in any file distribution scheme.

Access to the new sysctl commands is provided through a set of sysctl access control list (ACL) files shipped with PSSP. For `compat` mode, these ACLs are installed in the **/etc** directory and contain a Kerberos V4 entry for the global SP principal `root.SPbgAdm`. For DCE mode, the sysctl daemon initializes new DCE ACL objects relating to the new sysctl commands with entries for the PSSP DCE principal, `ssp/.../spbgroot`, which allow the principal to access the new sysctl commands. (For minimal mode, none/std, with restricted root access enabled, sysctl ACL entries must be manually updated to allow the new sysctl commands to be executed. However, this is not a suggested security configuration under which to enable restricted root access.

For sysctl to function correctly with restricted root access enabled, it is important that the sysctl daemon use the same default TCP/IP port number on all nodes and on the control workstation.

Overall, sysctl becomes a critical part of the SP environment when operating with restricted root access.

Limitations when using restricted root access

Restricted root access is a feature that creates new functionality within PSSP and fundamentally changes authorization methods. Obviously, this has consequences for some PSSP components and some AIX software programs that are used in an SP environment.

Coexistence: You can enable restricted root access after PSSP 3.2 or later is on the control workstation and all the nodes have PSSP 3.2 or later. At activation time, the **spsitenv** command checks whether all nodes are at a correct level. If there are nodes at a level earlier than PSSP 3.2, restricted root access cannot be activated. That also affects the ability to activate a secure remote command process.

IBM Virtual Shared Disk and GPFS: During configuration and run time the IBM Virtual Shared Disk component of PSSP and the GPFS licensed program rely on existing combinations of **rsh** and **sysctl** commands and nested **sysctl** calls to access information on the control workstation as well as on other nodes (node-to-node). The existing architecture relies on the existence of a common PSSP root identity that can be authorized successfully under **rsh** and **sysctl** ACLs. When restricted root access is enabled, the common PSSP root access required by IBM Virtual Shared Disk and GPFS is disabled. When attempting to activate restricted root access, if IBM Virtual Shared Disk adapters are defined or if GPFS is installed on the control workstation, the activation will fail.

HACMP: The HACMP licensed program can use Kerberos V4 as an authentication and authorization method that requires additional principals and interfaces defined in the Kerberos V4 database. When restricted root access is not enabled, adding additional principals to the Kerberos V4 database can be done manually or by running the `cl_setup_kerberos` script. This script is run from an HACMP node and does the following:

- It reads the HACMP topology.
- It uses **rsh** and **rcp** to update the Kerberos database on the control workstation.
- It creates new client keys for the HACMP nodes.
- It copies the keys from the control workstation to the HACMP nodes.

When restricted root access is enabled, the actions performed by `cl_setup_kerberos` are no longer possible due to the HACMP **rsh** and **rcp** requirements from a node to the control workstation. Also, the HACMP nodes can no longer use **rsh** and **rcp** to synchronize the HACMP ODM, given that the remote command authorization files on the control workstation (and other nodes) will no longer have common root authorization from node-to-node or node-to-control workstation. These authorizations will have to be added manually to the appropriate root-level remote command authorization files.

The synchronization capability can be restored by editing the ***/*.klogin** files on the HACMP nodes, for a compat/k4 enabled security mode, explicitly allowing both systems to have root access to each other. The restricted root access function will not interfere with custom settings.

Updating the Kerberos V4 database using the cl_setup_kerberos script requires root access to the control workstation, which is exactly what restricted root access was designed to eliminate. Manual updating of the Kerberos V4 database is the most secure way to ensure that the database is updated, but this is a complex and

error-prone method. A work-around is to temporarily add the HACMP nodes to the /.klogin file on the control workstation, for a compat/k4 enabled security mode, run the cl_setup_kerberos script, and then remove the authorizations immediately afterwards.

See the HACMP documentation for details on running with HACMP under various PSSP security configurations.

HACWS: When using the HACWS component of PSSP, there are no problems in the authentication of both systems on the nodes. However, you have to copy the authorization files, **/.rhosts** or **/.klogin**, to the backup control workstation. Also, it is important that you activate restricted root access from the active primary control workstation, since it is the Kerberos V4 Master.

There are other limitations and restrictions for HACWS, for instance it cannot run with DCE security mode enabled. For the complete list see “Limits and restrictions” on page 134.

Boot-install servers: Using multiple boot-install servers is not suggested and is not automatically supported by PSSP with restricted root access, a secure remote command process, or with AIX authorization for remote commands set to `none`. Boot-install servers are NIM Masters and they need to have remote command access to the control workstation when changing PSSP nodes. Boot-install servers also need **rsh** capability on their client nodes during installation time. At runtime, remote command authorization on the client nodes is only needed when a software or patch update is required.

However, depending on the size of your system and network loads, it might not be possible to have a single boot-install server. Though PSSP does not automatically create the correct entries in the authorization files to allow the remote commands to function, you can create the authorizations.

To use multiple boot-install servers, you can manually establish the correct authorizations on your system:

1. On the control workstation, change the authorization files, depending on the setting of the auth_root_rcmd attribute:
   - If auth_root_rcmd = standard, add an entry for the boot-install server node hostname in the **/.rhosts** file.
   - If auth_root_rcmd = k4, add an entry for the boot-install server node remote command principal in the **/.klogin** file.
   - If auth_root_rcmd = DCE, add an entry for the self-host and the **spbgroot** principal for the boot-install server node.
2. On the boot-install server node, edit the **/etc/sysctl.conf** file and include these entries:

   ```
   /usr/lpp/ssp/sysctl/bin/install.cmds
   /usr/lpp/ssp/sysctl/bin/switch.cmds
   ```
3. If you are initiating a node install customization, add this entry:

   ```
   /usr/lpp/ssp/samples/sysctl/firstboot.cmds
   ```
4. For these changes to take effect, stop and restart sysctld on both the control workstation and all boot-install servers.

Ecommands: With restricted root access, the following switch management commands can only be run from the control workstation:

CSS_test

Eclock
Efence
Eprimary
Equiesce
Estart
Eunfence
Eunpartition
mult_senders_test
switch_stress
wrap_test

System management commands: With restricted root access, certain system management commands should only be run from the control workstation. If run from a node, authorization failures might prevent successful completion. The following commands should only be run from the control workstation:

dsh
lppdiff
pcat
pcp
pexec
pexescr
pfind
pkill
pls
pmv
ppred
pps
prm
ptest
setup_authent
spacctnd
spgetdesc
splstdata (-d, -h, -i options only)
spmkuser

Additional security implications: Using restricted root access will not, in itself, make your system more secure. The fact that PSSP does not automatically authorize root to use **rsh** and **rcp** commands from a node and gain default access on other PSSP nodes does not prevent root from logging in to another node using the **telnet** command.

Even if you disable the ability for root to remotely log in to nodes and the control workstation, root can still use the **su** command to switch to another user and then exploit that user's remote command or other AIX and PSSP remote command capability to access other nodes or the control workstation.

You will have to adapt your user management for root to fit into the new access policy.

Considering a secure remote command process

You can have the PSSP system management software use a ***secure remote command process*** in place of the AIX **rsh** and **rcp** commands when restricted root access is enabled. You can acquire, install, and configure any secure remote command software of your choice. With restricted root access and a secure remote command process enabled, the PSSP system management software has no

dependency to internally issue **rsh** and **rcp** commands as a root user from the control workstation to nodes, from nodes to nodes, nor from nodes to the control workstation.

You must enable the restricted root access option before or when you enable the secure remote command process. You can then proceed to the rest of the PSSP security configuration.

There are three environment variables to let you pick whether you want the PSSP system management software to use AIX **rsh** and **rcp** remote commands or a secure remote command process for parallel remote commands like dsh, pcp, and others. The following are the environment variables and how to use them:

RCMD_PGM
: Enable use of the executables named by the `DSH_REMOTE_CMD` and `REMOTE_COPY_CMD` environment variables. The default is `rsh`. Set to `secrshell` to enable a secure remote command process.

DSH_REMOTE_CMD
: Specify the path and name of the remote command executable. The default with `rsh` is `/bin/rsh`. The default with `secrshell` is `/bin/ssh`.

REMOTE_COPY_CMD
: Specify the path and name of the remote copy command executable. The default with `rsh` is `/bin/rcp`. The default with `secrshell` is `/bin/scp`.

Like the restricted root access option, a secure remote command process can also be enabled by setting site environment attributes in the SDR. It is extremely important to keep these environment variables consistent and set to the remote command process you want to use. See “Understanding remote command choices” on page 96.

Note: Keep in mind that this support of a secure remote command process is enabled ***only*** in the PSSP installation, configuration, and system management software. You still need to explicitly grant the authorization that PSSP would otherwise have granted automatically for other applications that require root ability to issue **rsh** and **rcp** commands, like LoadLeveler and Parallel Environment. Some PSSP components like IBM Virtual Shared Disks and problem management, and licensed programs like GPFS, do not support this environment. See “Limitations when using restricted root access” on page 149 and “Considering choosing none for AIX remote command authorization”.

Considering choosing none for AIX remote command authorization

After you install the secure remote command software of your choice on the control workstation and enable the restricted root access and secure remote command process options, when setting your security configuration within each SP system partition a choice of `none` is offered for AIX remote command authorization. If you choose `none`, the PSSP system management software does not generate any root entries in the **.klogin**, **.k5login**, or **.rhosts** files on the nodes in the partition. To be able to set AIX authorization for remote commands to `none` in any partition, PSSP 3.4 must be installed on all nodes in that partition.

A reason you might want to set `none` for AIX remote command authorization is to add a greater level of security such that even the root user is no longer automatically authorized to issue **rsh** or **rcp** commands from the control workstation to the nodes. Enabling the secure remote command process does not remove the

capability of root to issue **rsh** or **rcp** commands from the control workstation to the nodes. It only ensures that the PSSP installation, configuration, and system management scripts that run as root use the secure remote command process instead. However, the root user or other processes running as root can still run **rsh** or **rcp** commands from the control workstation to the nodes. If the AIX remote command authorization is `none`, some applications that still rely on using **rsh** or **rcp** commands can no longer run.

Certain PSSP components and licensed programs are affected when you choose `none` for AIX remote command authorization:

- LoadLeveler and Parallel Environment

  You will need to configure security and authentication options for these licensed programs.

  Root user will not be able to issue the `llctl -g start` command from the nodes or control workstation. You can use the **dsh** command to run the `llctl start` command in parallel to the appropriate nodes or you can create a new user id to perform LoadLeveler administrative functions with the correct authorization to run **rsh** commands.
- Problem Management

  Some subscriptions to the Problem Management component of PSSP run using PSSP-generated principals for the **rsh** and **rcp** commands. Those principals are removed when AIX remote command authorization is `none` and the subscriptions might fail on the nodes and the control workstation.

Recording your authentication options

Prepare your choices for the security configuration within each SP system partition. Use a separate copy of Table 74 on page 297 for each SP system partition to record your choices for the following:

1. **The authentication services that PSSP is to automatically install.**
   - The DCE security services
   - The PSSP implementation of Kerberos V4

   Consider the following:
   - If you plan to choose DCE for SP trusted services in step 3 on page 154, choose DCE in this step.
   - If you plan to choose Kerberos V4 or compatibility in step 3 on page 154, choose Kerberos V4 in this step.
2. **The authentication service for remote commands.**

   Choose which authentication services are to be used for the AIX authenticated remote commands. This involves the following:

   a. ***The authorization files for AIX remote command authorization for the root user***:
      - .k5login for Kerberos V5 authentication
      - .klogin for Kerberos V4 authentication
      - .rhosts for standard AIX authentication

      Note: You can operate with none of these files if you plan to enable the restricted root access option and use a secure remote command process. In that case you have the option to choose `none`. You will probably have a secure login instead of rlogin. The ftp and telnet commands do not require these files. See “Considering restricted root

access" on page 146, "Considering a secure remote command process" on page 151, and "Considering choosing none for AIX remote command authorization" on page 152 before you decide.

b. ***The authentication methods for AIX remote commands*** to be enabled in each SP system partition. Choose any combination of the following:
   - Kerberos V5
   - Kerberos V4
   - standard AIX

Consider the following:

- For at least one of the authentication methods in 2b that you choose, you must choose the corresponding authorization file in 2a on page 153.
- Be sure to choose at least one authentication method that can operate on all the PSSP nodes and external hosts in your installation on which you expect to run AIX remote commands from the control workstation.
- If any node in a system partition is running an earlier version than PSSP 3.2, you must use Kerberos V4 for remote commands.
- DCE security services has Kerberos V5 protocol compatibility and can be the Kerberos V5 authentication provider for AIX remote commands.
- When determining which authentication method to use for remote commands, AIX examines the order of precedence set by the AIX **chauthent** command. This order determines which authentication method is used when a remote command is issued on the workstation. If the first method fails authorization, the second method is tried, and so on. The order of precedence, from the highest level of security to the lowest, is Kerberos V5, Kerberos V4, then Standard AIX.

3. **The authentication methods for SP trusted services.**

   The SP System Monitor command-line interface, the SP Perspectives graphical user interfaces, the PSSP remote execution facilities **dsh** and **sysctl**, and other SP trusted services all use authentication services. Choose which authentication methods are to be enabled for mutual authentication by SP trusted services within each SP system partition:
   - DCE
   - Compatibility (compat)
   - DCE and compat
   - None

   The compatibility method lets the SP trusted services use whatever means of authentication they used before PSSP 3.2. Some used Kerberos V4 with access control lists, some had their own independent means, some simply required root access, and some did not require authentication.

   Consider the following:

   - If you plan to enable the DCE authentication method for use by SP trusted services, the authentication methods for use by AIX remote commands in step 2b that you plan to enable must include Kerberos V5.
   - If you plan to enable the compat authentication method for use by SP trusted services, the authentication methods for use by AIX remote commands in step 2b that you plan to enable must include Kerberos V4.

Understanding which security services software to acquire

Your choices determine which security services software you need to obtain in addition to the AIX and PSSP software, where it is to be installed and configured, and at what point in the process. The PSSP implementation of Kerberos V4 authentication comes with the PSSP software and the authenticated remote commands come with AIX. You have to plan to obtain, install, and configure any other security services software that you choose to use. Neither the secure remote command, the DCE, nor the Kerberos V5 software comes with AIX or PSSP. If you decide to use DCE, PSSP 3.5 requires DCE 3.2, which can run with PSSP 3.4 and PSSP 3.2 nodes as well.

You need to obtain, install, and configure any secure remote command software that you want to use. PSSP 3.5 was tested with OpenSSH using the default install configuration and with **StrictHostKeyChecking no**. The following must be true for your secure remote command process to work:

- It conforms to the IETF Secure Shell protocol.
- The secure remote commands can be run by the PSSP system management as root from the control workstation to the nodes without prompts for a password or passphrase. This normally means the following:
  - The root public key generated on the control workstation is installed on the control workstation and all the nodes. An example of this is given in **/usr/lpp/ssp/samples/script.cust** file.
  - The root public key generated on a boot-install server (BIS) node is installed on the control workstation, the BIS node, and on all the nodes it serves. An example of this is given in **/usr/lpp/ssp/samples/sysctl/firstboot.cmds** file.
  - A **known_host** file has been generated on the control workstation and the BIS nodes for all nodes or the **StrictHostKeyChecking** variable is disabled to ensure that the PSSP scripts and commands are not prompted for passwords or passphrases. Prompting to the scripts will cause a hang.
- The boot-install server node still has **rsh** and **rcp** authorization for root access to NIM. See “Boot-install servers” on page 150.

See the discussion about setting up your secure remote command environment in the book *PSSP: Installation and Migration Guide*. See also the **dsh** command in the book *PSSP: Command and Technical Reference*.

To use DCE or Kerberos V5, you will have to do the following *before* you begin to install and configure PSSP:

1. Obtain the server software and licenses.
2. Have AIX installed and operating on the control workstation.
3. Have the DCE core server software installed, configured, and operational.
4. Have the control workstation operational in the DCE cell as a client or a server.
5. Place the DCE client software in the lppsource file on the control workstation.

The system installation scripts support the automatic installation and configuration of DCE clients and of the PSSP implementation of Kerberos V4 servers and clients. You are fully responsible for planning, installing, configuring, and operating any other security services software that you might choose to use.

There are many ways to use DCE, both internal and external to PSSP. Consider the following:

- When the set of SP trusted services authentication methods include DCE, then DCE V3.1 for AIX is required for both servers and clients. The base DCE client services are required on the control workstation, if it is not a server, and on the nodes in partitions in which DCE is enabled as an authentication method.

  If you want to run PSSP DCE administration commands on a cell administrator workstation remote from the Cluster 1600 system managed by PSSP, you need to install the **ssp.clients** file set and its prerequisites on that cell administrator workstation.
- When the SP trusted services authentication methods do not include DCE (your choice is compat or none), then DCE is not used by SP trusted services. In that case:
  - If you use DCE only for authentication within the AIX remote commands, a minimum level of DCE 2.2 is required.
  - If you use DCE only for authentication within DFS or some other non-SP-specific services, including locally developed custom DCE applications, then you can use whatever version of DCE is appropriate.

Protecting your authentication database

A primary consideration when you install and configure your system is the security of the workstations on which the authentication servers are to run. Because the authentication server is the repository of the secret keys for all principals, you need to protect it. Consider taking the following steps:

- Locate authentication server systems in physically secure areas, with access limited to administrators authorized to perform tasks related to their maintenance.
- Do not give user IDs for these systems to users, including PSSP system administrators, who are not authorized as security administrators. Do not enable remote access to them using **telnet**, **rlogin**, or **ftp**.
- Enable AIX system auditing to record security events on these systems.
- Implement an appropriate password selection and aging policy for the master password, and change the password regularly.
- Compromise of the master password would expose all private keys stored in the database. Establish a recovery plan that includes changing all passwords, replacing all server key files, and destroying all outstanding tickets.

Planning for DCE authentication

This section covers what you might need to plan after you decide to use DCE security services. Basically, after deciding to use DCE, there are certain actions you need to complete on your own, and others that PSSP helps you with. You need to obtain the DCE products to be installed, install the master and replica servers, and establish connectivity to the control workstation and nodes. The PSSP installation services help you install and configure the DCE clients on the nodes. Develop your own procedure by addressing each task and considering how to apply it in your situation. For instance:

- You might be adding your Cluster 1600 system managed by PSSP to an existing distributed system network with DCE security services already operating. This means you already have DCE and the servers are external to your control workstation. You need to establish connectivity and configure your control workstation as a client in the existing DCE cell. Place the DCE client code in the lppsource file on the control workstation. You will do additional configuration of

DCE on the control workstation and the DCE client software will be installed and configured on the nodes as part of the PSSP installation and configuration process.

- You might be planning a new Cluster 1600 system managed by PSSP where DCE security services need to be established. This means you need to obtain the software, plan, install, and configure the master and replica servers and clients, and place the DCE client code in the lppsource file on the control workstation. You will do additional configuration of DCE on the control workstation and the DCE client software will be installed and configured on the nodes as part of the PSSP installation and configuration process.

The following tasks are involved in planning to use DCE authentication:

- Deciding in which cell to configure DCE authentication.
- Considering to exclude network interfaces from DCE remote procedure call binding.
- Planning location of DCE servers.
- Establishing authorization to install and configure DCE.
- Preparing to configure SP trusted services to use DCE.
- Deciding on granularity of access to the SDR.
- Deciding on granularity of access to SP System Monitor objects.
- Planning use of AIX remote commands.

Deciding in which cell to configure DCE authentication

The Cluster 1600 system managed by PSSP must be part of only one cell. For instance, you cannot have one partition in cell A and another in cell B. Install and configure the control workstation as a DCE client within that cell. You can also configure the control workstation as a DCE server but it might not be best for performance. Both the cell and the control workstation must be configured and running DCE before performing any PSSP installation and configuration tasks that set up DCE on the Cluster 1600 system managed by PSSP.

Considering to exclude network interfaces from DCE Remote Procedure Call binding

Each time a DCE Remote Procedure Call (RPC) server is started, the Cell Directory Service (CDS) bindings are newly created, binding the server to all the network interfaces in the machine. To avoid performance degradation, you might want to exclude some interfaces from DCE RPC binding if some clients do not have a route to those interfaces. DCE provides the following environment variables with which you can exclude specific interfaces from being used:

- RPC_UNSUPPORTED_NETIFS
- RPC_UNSUPPORTED_NETADDRS

You can use either one depending on whether you choose to identify them by interface name or by IP address. You can set these environment variables by using an export statement in a shell script before the servers are invoked from that same shell script. If the servers are to be started during the boot process, either of these environment variables must be in the **/etc/environment** file. Environment variables in that file are made available to all the processes after the next boot.

To exclude the interfaces with the names en0 and en1, for example, use the statement:

```
# export RPC_UNSUPPORTED_NETIFS=en0:en1
```

To exclude the interfaces with IP addresses 123.45.67.89 and 123.45.125.23, for example, use the statement:

```
# export RPC_UNSUPPORTED_NETADDRS=123.45.67.89:123.45.125.23
```

Planning location of DCE servers

The control workstation and PSSP nodes need connectivity to the servers. If you plan to configure the Cluster 1600 system managed by PSSP into an existing DCE cell, record the locations of the DCE master and replica servers. If the system will have many nodes used as part of parallel jobs with many tasks using DCE authentication, consider having DCE replica servers for better performance. For new DCE cells, plan and record where to install your DCE master and replica servers. Your choices are:

- On the control workstation.
- On PSSP nodes.
- On external workstations.

The DCE master servers for a given cell must exist either on the control workstation or on a system external to the Cluster 1600 system managed by PSSP. They cannot be on PSSP nodes if any SP system partitions are configured for DCE within that cell. PSSP nodes or the control workstation can be configured to run replica (secondary) DCE servers.

See “Authentication servers” on page 109 for related information.

Establishing authorization to install and configure DCE

Only the *root* user on the PSSP control workstation can select to install and configure DCE for an SP system partition. On the control workstation, certain configuration tasks require DCE cell administrator authority. This might not be the person with root authority on the PSSP control workstation. Plan activities to authorize the appropriate people and assign the installation and configuration tasks respectively.

Preparing to configure SP trusted services to use DCE

A DCE cell administrator must prepare for configuring the SP security services, which are in the **ssp.clients** file set, before the **config_spsec** command is run either from SMIT or from the command line during installation and configuration.

The service configuration file **/usr/lpp/ssp/config/spsec_defaults** contains the information that is used during configuration for automatic creation of DCE entities in the DCE registry and the DCE Cell Directory Service (CDS) database. The file has the following information:

- The default name of the DCE organization and group used to create DCE accounts for the service principals used by the SP trusted services.
- The default name of the DCE group with authority to administer the DCE ACLs of the trusted services. The DCE cell administrator is a member of this group and can add other members.
- The default DCE name and attributes for each SP trusted service. This name forms the constant part in deriving the full principal name of the service, its account name, its key file path name, and its CDS path name.

Before you install a Cluster 1600 system managed by PSSP into an existing DCE cell, check those names to prevent conflicts with names that already exist in your DCE database. You might also want to consider if you have different security

requirements in different SP system partitions. If you do, you might want to override some of the settings in the **spsec_defaults** file to set up access groups for partition-sensitive authentication.

Do not update the **spsec_defaults** file. If you want to replace a default name or set up partition-sensitive access groups, update the **/spdata/sys1/spec/spsec_overrides** file, which is also in the **ssp.clients** file set.

To apply any overrides, update the copy of the **spsec_overrides** file that is on the control workstation before the **config_spsec** command is run for the first time. That updated file is automatically copied to the nodes during the PSSP installation and configuration processing. For file content and override examples, see the book *PSSP: Command and Technical Reference*.

Note: If you override any default DCE entities, be sure to instruct users in your organization that the respective default names used in the PSSP publications have been replaced. Also, if you plan to install PSSP security services on any independent network-attached pSeries or RS/6000 workstations that are to function as PSSP clients, be sure to copy the **spsec_overrides** file to each of them before configuring security services on them. After the **config_spsec** command is run, do not modify the **spsec_overrides** file.

Deciding on granularity of access to the SDR

All users can read the SDR. Write and admin access can be restricted. If you use DCE-only authentication, administrators whose duties involve performing tasks that create or update SDR information, by using PSSP Perspectives, SMIT, or SDR commands, will have to use the **dce_login** command with a principal that has been added to the appropriate SDR user access groups. IBM suggests that write and admin SDR access be limited to administrators who need to run scripts to install or configure the Cluster 1600 system managed by PSSP. That includes those who manage IBM Virtual Shared Disks.

You can also grant SDR access on a partition basis by changing the **spsec_overrides** file to specify separate groups for different partitions.

Therefore, you need to consider and plan the granularity of protection you want to establish for giving principals write or admin access to the SDR. See the appendix about the SDR in the book *PSSP: Administration Guide* for more information.

Deciding on granularity of access to SP System Monitor objects

The SP System Monitor allows authorized users to control and monitor the status of the frames, nodes, and switches of the system. It is a client-server application. The server, the hardmon daemon, executes only on the PSSP control workstation and controls the SP supervisor subsystem through the device special files for the RS-232 connections to each frame supervisor. Authorized users can use the SP Hardware Perspective or commands to see status information or control the power and the logical key switch position for node processors.

The objects defined and controlled by the system monitor are the following:

1. a single system object
2. frame objects
3. slot objects
4. a hardmon object (the system monitor daemon)

The system object is the primary object and is the container for frame and slot objects. Monitor and control operations are performed on frame and slot objects. The hardmon object represents the administration function.

These objects are hierarchical: the system object contains frames and frame objects contain slot objects which represent node or switch processors. You can have potentially hundreds or thousands of objects and an access control list for each object. On the other hand, since the object structure is hierarchical you can have only the system object ACL and set principals and groups with authority to control the entire system. You can have system level, frame level, and slot level authorization. Authorization is checked from the lowest level to the highest level. If an ACL for the slot object does not exist, hardmon checks for the frame object. If an ACL for the frame object does not exist, hardmon uses the system object. The ACLs of a container object are copied to the contained object by default when the object is created.

Therefore, you need to consider and plan the granularity of object protection you want to establish for authorizing principals to monitor or control the system. The book *PSSP: Administration Guide* tells you about setting the authorizations.

Planning use of AIX remote commands

To use Kerberos V5 authentication within the AIX remote commands, DCE principals must be authorized to access a PSSP node with an AIX user id. This authorization is controlled through the use of a .k5login file in the AIX user's home directory. The .k5login file can be created and maintained by each user or by a system administrator.

Planning for Kerberos V4

Planning the installation and configuration of Kerberos V4 involves:

- Establishing Authorization to Install and Administer Kerberos V4
- Deciding on your authentication configuration
- Selecting authentication options to install
- Creating the authentication configuration files
- Deciding on authentication realms

Establishing authorization to install and administer Kerberos V4

When you set up your Kerberos V4 authentication realm and install and configure your Cluster 1600 system managed by PSSP, the roles of system installer, system administrator, and security administrator become intertwined. This is because some of the PSSP installation and customization scripts are designed to automatically recognize the authentication entities that have to be created for the remote commands and the SP trusted services. The **setup_server** script creates service principals whenever it discovers a new network interface defined on the control workstation for which it has no Kerberos V4 principal defined. At these times, the script must be running in an environment where the invoking user has authentication credentials as a database administrator. If the script needs credentials but finds that you are not authorized as an **admin** principal, it will fail. If this happens, obtain the appropriate credentials and issue the command or invoke it from SMIT again.

Other installation and customization tasks invoke the **setup_server** script internally. Any of these tasks which involve adding or renaming of adapters or host names

also require credentials as a database administrator. Examples of installation scripts used to perform these tasks are **spadaptrs**, **sphostnam**, **spbootins**, and **spethernt**.

Other administration tasks that simply use the remote command facilities of **dsh**, **rsh**, or **rcp** require normal user credentials for the principals, since they do not involve changes to the authentication database.

Plan activities to authorize the appropriate people and assign the installation and configuration tasks respectively. You must define at least one user principal authorized to perform installation tasks. A system administrator, logged on as **root**, must assume the identity of this principal. When you use authentication services provided by AFS or another Kerberos implementation, this principal should already exist in the authentication database. For SP authentication services and other Kerberos V4 implementations, the administrator's principal can have any name with an instance of **admin**. An AFS principal has administrative authority if it has an **admin** attribute in its definition.

Deciding on Kerberos V4 authentication configuration

This section describes the various ways you can configure a Cluster 1600 system managed by PSSP in an authentication realm. The following sections illustrate the possible authentication configurations used with the control workstation. The configurations also include other independent network-attached pSeries or RS/6000 workstations on which you install SP authentication services, and non-RS/6000 workstations, when the authentication servers are configured in each of the supported manners. The control workstation and the PSSP nodes are always in a single authentication realm, which can optionally include other workstations and even other Cluster 1600 systems managed by PSSP. The authentication servers can be on any workstation in the realm, but not on PSSP nodes. If you have AFS installed on your workstations, you can choose to use AFS servers for authentication, but are not required to do so. If you do not use AFS, you can use SP authentication servers or other Kerberos V4 servers. The PSSP nodes will have authentication services installed for all authentication configurations.

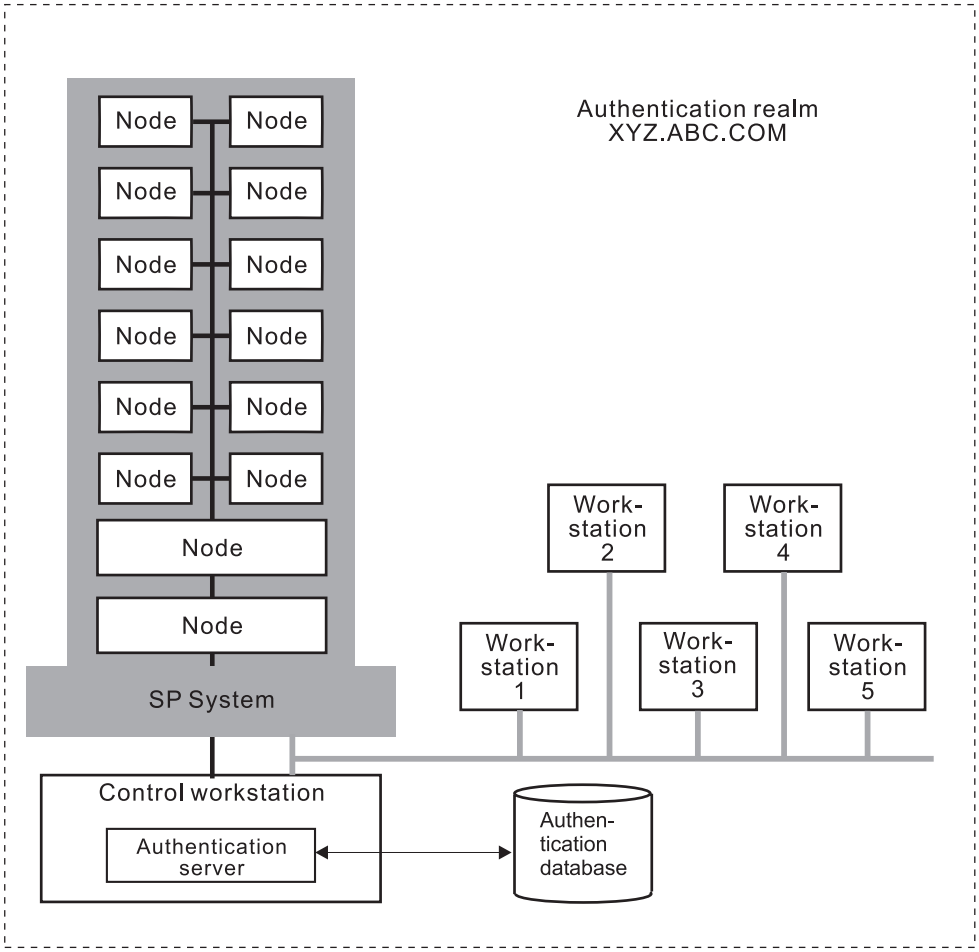


Figure 24. The control workstation as primary Kerberos V4 authentication server

CWS-Primary

Figure 24 illustrates this configuration as follows:

- The control workstation is the primary authentication server, with the SP authentication server (file set ssp.authent) and authenticated services (file set ssp.clients) installed.
- Other pSeries or RS/6000 workstations can be secondary authentication servers, with the SP authentication server installed.
- Other pSeries or RS/6000 workstations can have SP authenticated services installed.

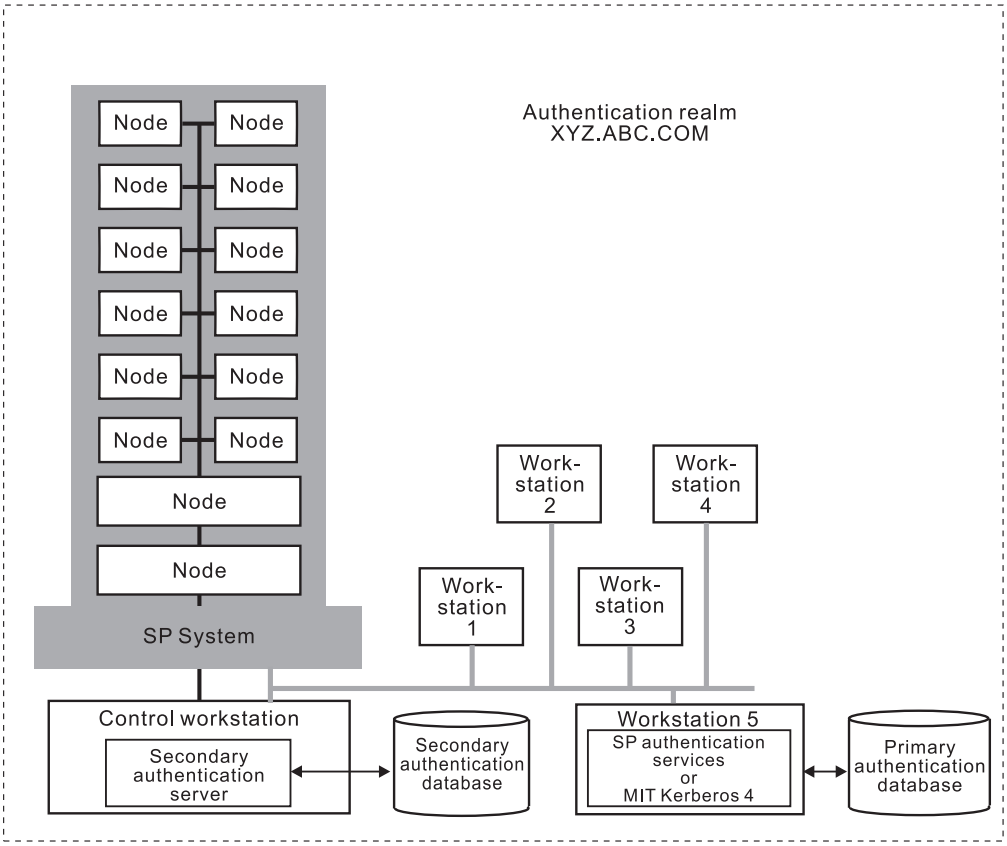


Figure 25. The control workstation as secondary Kerberos V4 authentication server

CWS-Secondary

Figure 25 illustrates this configuration as follows:

- The primary authentication server is one of the following:
  - A pSeries or RS/6000 workstation with the SP authentication server (file set ssp.authent) and authenticated services (file set ssp.clients) installed
  - A workstation with another Kerberos V4 implementation
- The control workstation is a secondary authentication server, with the SP authentication server and authenticated services installed.
- Other pSeries or RS/6000 workstations can be secondary authentication servers, with the SP authentication server installed.
- Other pSeries or RS/6000 workstations can have SP authenticated services installed.

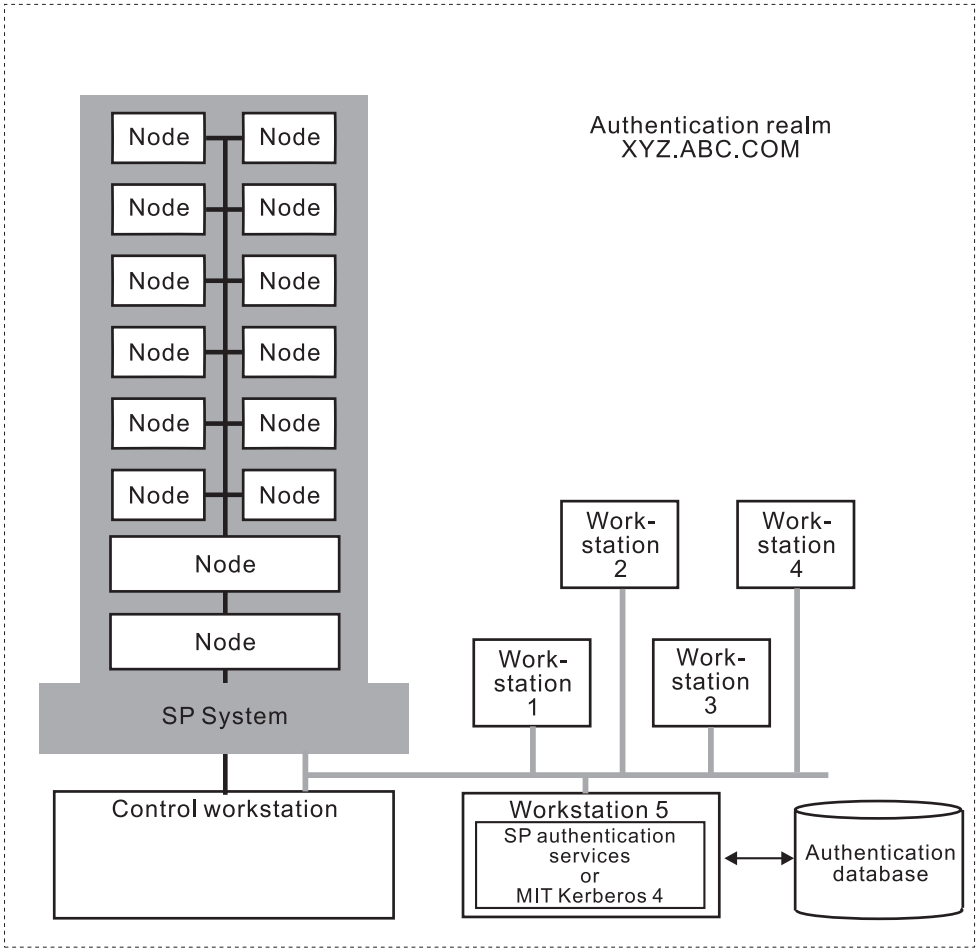


Figure 26. The control workstation as client of Kerberos V4 authentication server

CWS-Client

Figure 26 illustrates this configuration as follows:

- The primary authentication server is one of the following:
  - A pSeries or RS/6000 workstation with the SP authentication server (file set ssp.authent) and authentication services (file set ssp.clients) installed
  - A workstation with another Kerberos V4 implementation
- Other pSeries or RS/6000 workstations can be secondary authentication servers, with the SP authentication server installed.
- The control workstation has SP authenticated services installed.
- Other pSeries or RS/6000 workstations can have SP authenticated services installed.

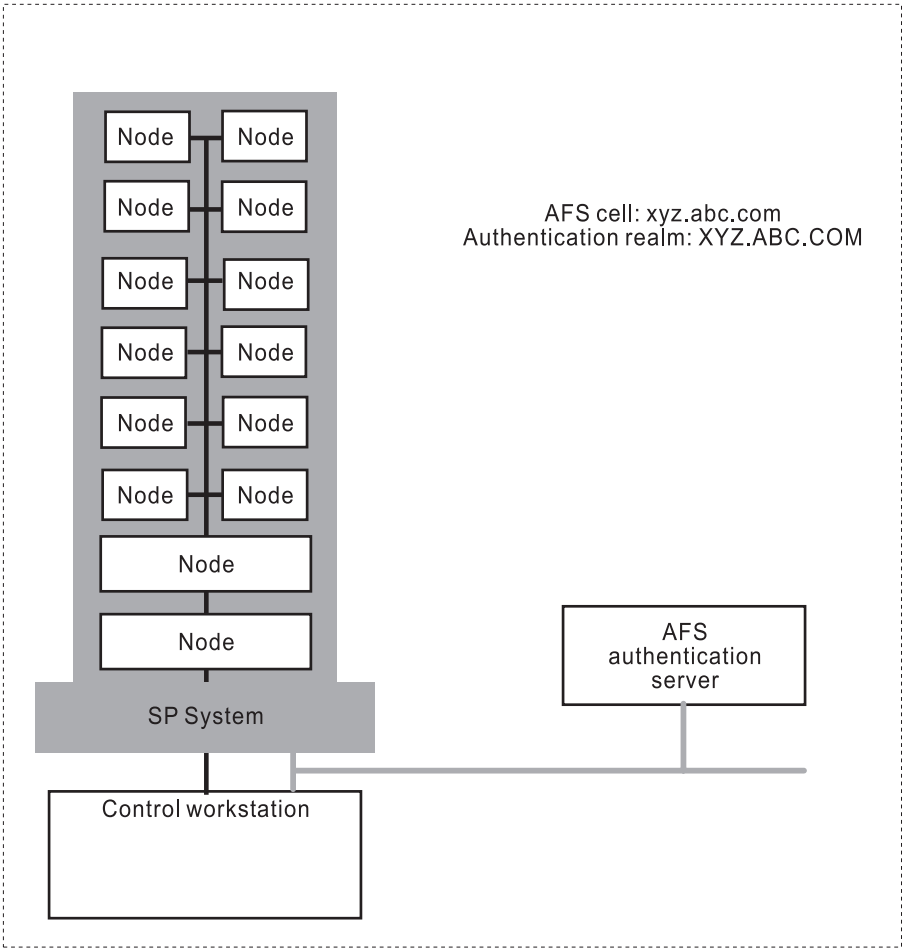


Figure 27. Using AFS authentication services

AFS Server

Figure 27 illustrates this configuration as follows:

- The authentication servers are AFS servers, on the control workstation or other workstations.
- All workstations, including the control workstation, have AFS client services installed.
- The control workstation has SP authenticated services (file set ssp.clients) installed. installed.
- Other pSeries or RS/6000 workstations can have SP authenticated services installed.

Selecting the Kerberos V4 authentication options to install

Selecting options for installation depends on where the authentication server is located. SP authentication code is distributed in two separately installable options. The **ssp.authent** file set contains only parts required on a system that is to be an authentication server. The remainder of Kerberos V4 authenticated services is distributed in the **ssp.clients** file set.

You must install **ssp.authent** on the control workstation, if it is to be a Kerberos V4 authentication server, either primary or secondary. You can also install **ssp.authent** on any other pSeries or RS/6000 workstation that you plan to be an authentication server. You will not be able to install it if the system already has a Kerberos V4

implementation installed. If you want to install the SP authentication facilities, you must first remove the other Kerberos implementation.

You must install **ssp.clients** on the control workstation, even in the case where you intend to use AFS or Kerberos V4 authentication. You need to also install it on any other pSeries or RS/6000 workstation that you plan to be an authentication server or from which you plan to perform system management tasks using System Monitor commands, AIX remote commands, or the **sysctl** remote command execution facility. Workstations using Kerberos V4 authentication do not require **ssp.clients** if they are not using PSSP system management tools. All PSSP nodes will have **ssp.clients** installed.

Creating the Kerberos V4 configuration files

For some of these configurations, you need to create a configuration file (**/etc/krb.conf**) that lists the local realm name and all server host names. The following list identifies the cases in which you must provide the **/etc/krb.conf** file, and shows simple examples:

CWS-Primary
: Optional - you must supply the file only if there will be one or more secondary servers on pSeries or RS/6000 workstations. If the **/etc/krb.conf** file is not supplied, **setup_authent** creates a file listing the local host as the primary server. For example:

```
XYZ.COM
XYZ.COM spcw.xyz.com admin server
XYZ.COM ksecondary.xyz.com
```

CWS-Secondary
: Required - control workstation is listed as a secondary server. This requires that the **krb.conf** file is first created on the primary authentication server and is copied to the control workstation. For example:

```
XYZ.COM
XYZ.COM kprimary.xyz.com  admin server
XYZ.COM spcw.xyz.com
```

CWS-Client
: Required - control workstation is not listed in configuration file. This requires that the **krb.conf** file is first created on the primary authentication server and is copied to the control workstation. For example:

```
XYZ.COM
XYZ.COM kprimary.xyz.com  admin server
XYZ.COM ksecondary.xyz.com
```

CWS-AFS
: None - file is derived automatically from AFS configuration files. AFS uses the configuration files **/usr/vice/etc/CellServDb** and **/usr/vice/etc/ThisCell**.

See the **krb.conf** file format man pages in *PSSP: Command and Technical Reference* for more information and examples.

Deciding on authentication realms

If you are using AFS authentication servers, your authentication realms are the same as your AFS cells. The name of the local realm for the Cluster 1600 system managed by PSSP will be automatically set to the name of the AFS cell of the control workstation, and is converted to upper case.

When you are not using AFS, the following considerations apply. A Cluster 1600 system managed by PSSP must be installed in a single Kerberos V4 authentication realm. This is the case if you are installing PSSP Kerberos V4 authentication on only the control workstation. The authentication realm could be an existing realm, consisting of systems using another Kerberos implementation, to which you add the Cluster 1600 system managed by PSSP. You can give the realm any name you like or default the authentication realm name to the domain part of the server's host name, converted to upper case.

Whenever you have additional Cluster 1600 systems managed by PSSP or other workstations using authenticated services, you must decide whether you want them all in the same realm, sharing a single set of principals in one master authentication database. Generally a single realm is easier to manage and easier for users who don't have to concern themselves with selecting the correct realm when identifying a principal.

If there is to be any use of authenticated services between two different authentication realms, each realm must have a unique name. If you choose to have multiple realms, and there are systems in both whose host names have the same domain part, you can configure only one using the default authentication realm name. If there is any chance that you would add additional authentication realms and want to use authenticated services between systems in them, it is best to create your own non-default and meaningful realm names when you plan your configuration.

See the section on working with authentication realms in *PSSP: Administration Guide* for more information.

Planning for standard AIX authentication

There are no special installation or configuration considerations to use Standard AIX authentication. It comes with your AIX 4.3.3 operating system. Standard AIX authentication is based on IP address or a user ID and password. Access is based on the contents of a file in the user's home directory. The *.rhosts* file contains the list of authorized source host names and user names. A system administrator who is logged in as root can perform all installation tasks.

Checklists for authentication planning

Use each checklist that applies to an authentication method that you plan to enable.

Using DCE security services

For DCE-specific tasks, see the DCE publications. The following checklist summarizes what you need to do from the PSSP point of view:

1. Obtain the product to install. See the book *IBM DCE Version 3.1 for AIX: Quick Beginnings* which describes DCE and explains how to plan for, install, and configure the product.
2. Plan the cell.
3. Plan the master and replica servers.
4. Ensure the PSSP control workstation and nodes have connectivity.
5. Plan which security services to install on the control workstation and nodes.
6. Plan if you need to exclude any SP network interfaces.
7. Plan which authentication methods to enable in each SP system partition for root user execution of remote commands.

8. Plan which authentication methods to enable in each SP system partition for the SP trusted services.
9. Considering the granularity of access control you require and the existing names in your DCE database, plan any partition-specific or other DCE group and principal names for which you might want to override the PSSP default names.
10. Ensure authorizations are established in the DCE database in order to be able to install PSSP with DCE security services.
11. Ensure the DCE servers are completely installed and functional before installing PSSP.

Using Kerberos V4 authentication servers

If you plan to use Kerberos V4 authentication, you need to decide what authentication realms your network will have.

Complete this checklist for each realm:

1. Decide on the name of the realm.
2. Determine the administrative principal you will use for installing the SP authentication on the control workstation and other pSeries or RS/6000 workstations. Either this administrative user or another that you define later must be assigned UID 0 in order to perform SP installation tasks that require both root privileges and Kerberos administrative authority.
3. Decide which system is the primary server.

   If it will be an SP authentication server:

   - Make sure no other Kerberos system is installed.

   Otherwise, it must be an existing (primary) Kerberos server.

   - Make sure the authentication server is installed and running.
   - Make sure the kshell service (rsh/rcp daemon) is available.
   - Make sure that network interfaces and name resolution are set up to allow it to access the primary server.
4. Decide which systems will be secondary servers.
5. Make sure that network interfaces and name resolution are set up to allow it to access the primary server and the SP system.

   If any

   - Decide how you will order the entries in the **/etc/krb.conf** configuration file.
   - Decide how often you want to automatically propagate the authentication database from the primary server to the secondaries.
   - For each secondary server
     - Make sure no other Kerberos system is installed.
     - Make sure that network interfaces and name resolution are set up to allow it to access the primary server.
6. Identify any other pSeries or RS/6000 systems that will be clients.

   If any other pSeries or RS/6000 systems will be clients:

   - Decide how you will order the entries in the **/etc/krb.conf** configuration file.
   - Make sure that network interfaces and name resolution are set up to allow it to access the primary server and the SP system.

Using AFS authentication

If you plan to use AFS authentication servers with your SP system, take into account the unique considerations in the following checklist:

1. Any pSeries or RS/6000 workstation on which you are installing the SP authentication support, including the control workstation, must have already been set up as either an AFS client system or as an AFS server.
2. If the AFS configuration files, **CellServDB** and **ThisCell**, are installed in a directory other than **/usr/vice/etc**, or if the **kas** program is not installed in **/usr/afsws/etc** or **/usr/afs/etc**, you must create symbolic links at the directory level so the SP **setup_authent** program can find these files.
3. You must have a user defined with the AFS **admin** attribute that can be used during SP authentication setup and installation. This user will also be the default user defined with administrative authority in the System Monitor's access control list file. You can add other administrators later.
4. In order for users to use the authentication service on the SP nodes, you must also install AFS client services on those systems. See the instructions for AFS client customization of the SP nodes in the sample file **afsclient.cust** in the *PSSP: Administration Guide*
5. The authentication server (**kaserver**) in AFS 3.4 for AIX 4.1 accepts Kerberos V4 protocol requests using the well-defined **udp** port assigned to the **kerberos** service. AIX 4.1 assigns the Kerberos V5 port number 88 to work with DCE. PSSP authentication services based on Kerberos V4, uses a default port number of 750. The PSSP commands use the service name **kerberos4** to avoid this conflict with the Kerberos V5 service name. For PSSP authentication commands to communicate with an AFS 3.4 **kaserver** on AIX 4.1, you must do one of the following steps:
   - Stop the **kaserver**, redefine the **udp** port number for the **kerberos** service to 750 on the AFS Authentication server system, then restart the **kaserver**.
   - Add a statement to **/etc/services** that defines the **udp** port for the **kerberos4** service as 88 on the SP control workstation and on any other independent workstation that will be a client system for PSSP authenticated services.

Authentication worksheets

For planning your authentication, make one copy for each SP system partition of Table 74 on page 297 and complete it. For each SP system partition, do the following:

- If you use DCE authentication, copy and complete Table 75 on page 297.
- If you use PSSP Kerberos V4 or other Kerberos authentication servers, copy and complete Table 76 on page 298 and Table 77 on page 299.
- If you use an AFS authentication server, copy and complete Table 78 on page 299.

Note: Be aware that you are recording passwords in some of these worksheets. Remember to classify and secure the completed worksheets.

Chapter 7. Planning SP system partitions

This chapter describes planning considerations for system partitioning. It describes the predefined system partitioning layouts shipped with the SP system software and introduces you to the System Partitioning Aid which allows you to create new system partitioning layouts that better suit your needs.

Note:

- SP system partitioning is supported by default on SP systems without a switch and on Cluster 1600 systems managed by PSSP with the SP Switch.
- SP system partitioning on a switchless SP system can be disabled.
- SP system partitioning is not supported on clustered server systems without a switch or on Cluster 1600 systems managed by PSSP with the SP Switch2.

For more specific information on how to partition your system, see the book *PSSP: Administration Guide*.

What is system partitioning?

System partitioning is the process of dividing your system into non-overlapping sets of nodes in order to make your system more efficient and more tailored to your needs. System partitions are usually relatively static and long-lived entities.

A system partition is, at the most elementary level, a group of nodes (not including the control workstation). In essence, a system partition is a subset of an SP system which consists of sufficient pieces (nodes, control workstation, data, commands, and so on) to form a logical SP subsystem.

With system partitions, you can ensure that applications using the SP Switch running on one group of nodes are not inadvertently affected by activity on other nodes in the system.

Dependent nodes and SP-attached servers should be considered the same as standard nodes when planning a system partition.

System partitioning affects communication which occurs over the SP Switch only; other communication paths are not affected. System partitioning also provides environmental controls that allow the system administrator to control and monitor only the current system partition.

Why would you partition the system?

Your SP system has a particular configuration defined by its frames and nodes. The SP comes with a single default system partition and a set of predefined system partition layouts for each standard configuration. These layouts have been selected in a way which meets minimal throughput capabilities. In addition, you to construct your own layouts.

Carefully consider:

1. “The default system partition” on page 172
2. “Benefits of multiple system partitions” on page 172

If you are planning a switchless SP system and decide you do not need multiple system partitions and will never want to use the SP Switch or your need for more SP-attached servers is greater, see “Understanding partitionability choices” on page 97. If you do want multiple system partitions and none of the predefined layouts meets your system partitioning needs, you can define your own using the System Partitioning Aid or you can submit a Request for Price Quote (RPQ) to IBM to request additional layouts. See your IBM representative for more information on the RPQ process.

The default system partition

Taking advantage of multiple system partitions is something you do by choice. However, the partitioning atmosphere is always present to some extent. In the beginning, when you have installed the PSSP software, but before you explicitly partition your system, there is one system partition that contains all of the nodes and its name is the same as the name of the control workstation. This is the *default* or *persistent* system partition. It always exists. When you choose a different partition layout, one of the resulting partitions is this default system partition. A new system partition is formed by taking nodes from existing system partitions and collecting them as a new group.

Benefits of multiple system partitions

You gain several benefits from using system partitions. You can:

- Run switch-based applications on a set of nodes without interfering with switch work on another set, regardless of application or node failures. In particular, you can isolate SP Switch traffic, preventing it from affecting switch traffic in another system partition.
- Separate a test area for application development from your production area.
- Install and test new releases and migrate applications without affecting current work.
- Have one operator manage, at a system level, more than one logical system from a single control workstation.
- Separate system administration for each partition.

Change management and non-disruptive migration

You can test new levels of AIX, PSSP, other IBM licensed programs, local application programs, or other software on a system currently running a production workload without disrupting that workload. Such a system partitioning solution assumes that there are spare nodes available to set aside in a test system partition. This solution lets you run migration scenarios on the test partition nodes without interfering with day-to-day operations on the rest of the system. You can form and manage system partitions and then customize the partitions with software.

Multiple production environments

You might also need to create multiple production environments with the same non-interfering characteristics as in “Change management and non-disruptive migration”. With system partitions these environments are sufficiently isolated so that the workload in one environment is not adversely affected by the workload in another environment. This is especially true for services whose usage is not monitored and for which there is no charge, but which have critical impact on performance of jobs, such as the SP Switch. System partitions let you isolate SP Switch traffic in one system partition from the SP Switch traffic in other system partitions.

Security across system partitions

There are rules that govern the setting of security attributes within a partition. These rules apply to the default single system partition as well as to each additional partition that you might create. “Choosing authentication options” on page 145 explains those rules. The setting for AIX remote command authentication on the control workstation is the union of the related settings in all partitions. The setting of authentication method for SP trusted services on the control workstation is the union of the related settings in all partitions.

If you are planning to change your partition configuration, consider the following:

- Adding Kerberos V4 authentication to a node requires the node be *customized* which is a procedure described in the book *PSSP: Installation and Migration Guide*. A node can have Kerberos V4 authentication added in one of the following ways:
  - Adding Kerberos V4 authentication to the security configuration of the system partition in which the node resides.
  - Repartitioning such that the node moves to a partition having Kerberos V4 authentication already set in the security configuration.
- Changing security authentication settings for a partition requires all the affected nodes to be rebooted.
- There are partitioning commands and SMIT panels available for partitioning your SP system. Although the commands provide interfaces to allow setting the security attributes in the SDR, do not use them to change the settings for existing partitions. Use the security commands and SMIT panels listed in the book *PSSP: Installation and Migration Guide* to change these settings.

When an SP system is to be partitioned, you might have different security requirements in different partitions. If you use DCE security services, partition names can be appended to the group names in the **spsec_overrides** file to define different groups in different partitions. To accomplish this, use the **:p** option in the **spsec_overrides** file before configuring SP security services. See Chapter 6, “Planning for security” on page 145, particularly “Preparing to configure SP trusted services to use DCE” on page 158.

Example 1 – The basic 16-node system

Figure 28 on page 174 shows a simple 16-node system that contains one frame, one switch board, and 16 thin nodes installed. In this example, the nodes are named Node01, Node02, and so on up through Node16. You can name your nodes any way you want, but the nodes are also known by *node numbers*, and the node numbers are assigned in the same manner as they are named in this example: from bottom left to top right.

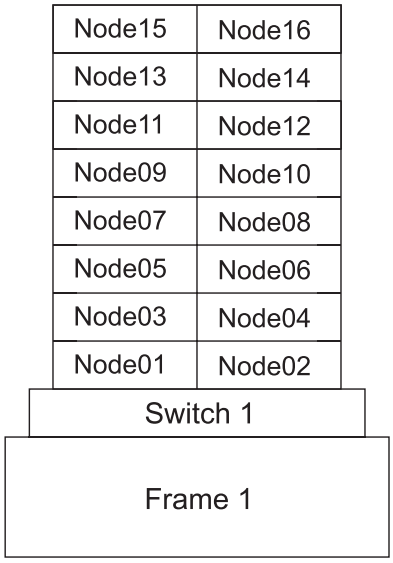


Figure 28. A simple 1-frame SP system

Assume that you own this system, and that your day-to-day operations revolve around software called Application A, Version 1. Also, assume that you are interested in upgrading to Version 2 of Application A, and want to try out the new version while still relying on Version 1.

After evaluating your current workload, you determine that any 12 nodes are sufficient to perform your normal activity and, therefore, you decide to set 4 nodes aside to try out Version 2. This means you want to partition your 16-node system into 2 subsystems: a 12-node system partition and a 4-node system partition.

When you consult the predefined layouts shipped with your system, you find that several 4_12-layouts are provided for your 16-node system, and you decide to go with the following (listing node numbers rather than node names):

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

You adopt this configuration using a simple SMIT panel, and begin running your production load on Partition 1. Your choice is pictured in Figure 29 on page 175.

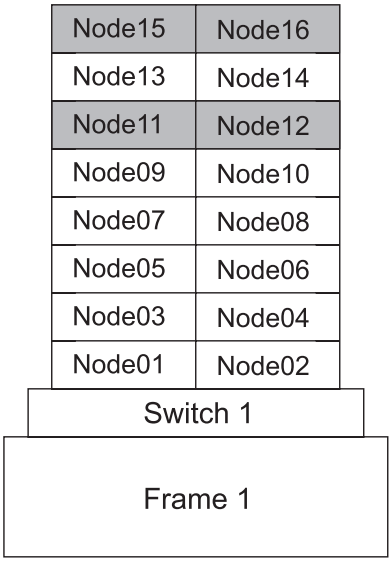


Figure 29. A partitioned 1-frame SP system

Next you install Version 2 of Application A (together with any prerequisite software and hardware) on the nodes of Partition 2, provide Partition 2 with suitable test data, and begin executing trial runs of Version 2 on Partition 2.

Again, the switch-intensive portions of the applications of interest (Application A, Version 1 and Application A, Version 2) will run independently in their respective partitions. That is, your daily production runs and the Version 2 trial runs will not affect each other — in regard to switch performance. This is because the 4_12-layouts provided were constructed with that goal.

Using an SP switch in a partition

The SP system partitioning can be used with the 16-port or 8-port SP Switch. It is not supported with the SP Switch2.

The physical makeup of a switch board

Actually, your choice in Example 1 was not necessarily as simple as suggested. A full *switch board* consists of 8 *switch chips* as shown in Figure 30 on page 176. Each chip has 8 *ports* to which nodes and other switch chips can connect.

Precisely 4 of the switch chips can have nodes connected to them, as on the left side of the board in Figure 30 on page 176. These chips are called *node switch chips*. Due to physical choices made in the SP frame, the nodes are connected as shown in the figure. Notice the following:

1. Nodes 1, 2, 5 and 6 are attached to switch chip 5.

   Note: Nodes connected to the same chip can communicate with each other via that chip.
2. Nodes 3, 4, 7 and 8 are attached to switch chip 6.
3. Nodes 9, 10, 13 and 14 are attached to switch chip 4.
4. Nodes 11, 12, 15 and 16 are attached to switch chip 7.
5. There are no direct links among chips 4-7, nor among chips 0-3.

6. Each of chips 4-7 is directly connected to all of chips 0-3. Therefore, for example, the nodes on switch chip 4 can communicate with the nodes on switch chip 7 via any of chips 0-3.

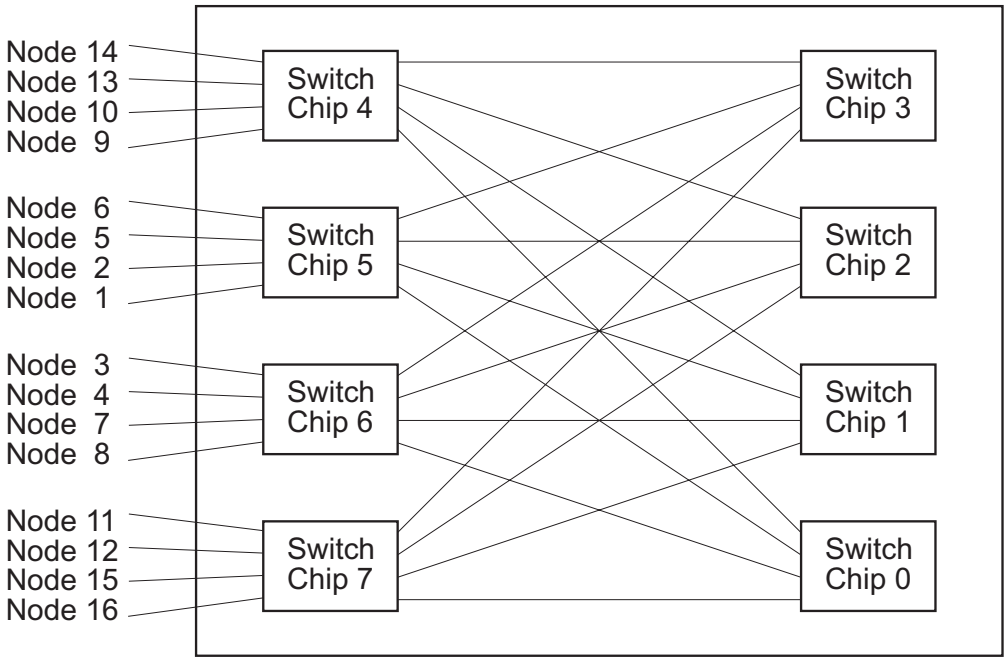


Figure 30. Full switch board

Chips 0-3 are called *link switch chips*, and are also used in multi-frame systems to connect the various switch boards to each other using ports not shown in the figure.

Systems with switches are assumed to be used in performance-critical parallel computing. One major objective in partitioning a system with a switch is to keep the switch communication traffic in one switch partition from interfering with that of another. In order to ensure this, each switch chip is placed completely in one system partition.

Any link which joins switch chips in different partitions is disabled, so traffic of one partition cannot enter the physical bounds of another partition. The result of the partitioning choice you made in Example 1 is shown in Figure 31 on page 177. Notice that the links from Chip 7 are missing in the diagram, indicating they have been logically removed from the active configuration, or disabled.

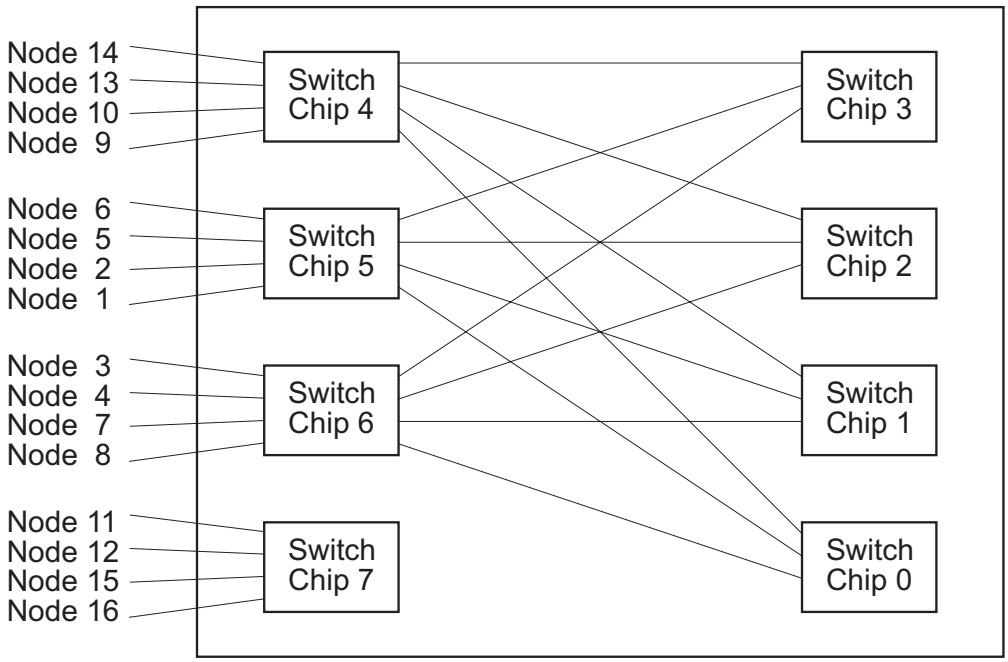


Figure 31. Nodes 11, 12, 15, and 16 partitioned off

Systems with a low cost switch

If your system contains the low cost SP Switch-8, your system partitioning capabilities are restricted. The SP Switch-8, has only 2 chips with nodes attached. So, if you have the maximum 8 nodes attached to the switch, you have 2 possible configurations: a single-partition 8-node system, or 2 system partitions of 4 nodes each.

Switchless systems

One main consideration when planning for system partitions is the use of a switch. Partitioning, however, is also applicable to switchless systems. If you have a switchless system, and later want to add a switch, you might have to rethink your system partition choice. In fact you might have to reinstall the ssp.top file set so that any special switchless configurations you have constructed are removed from the system.

If you choose one of the supplied layouts, your partitioning choice is `switch smart`: your layout will still be usable when the switch arrives. This is because the predefined layouts are constrained to be usable in a system with an SP Switch. Such a layout might be unsatisfactory, however, for your switchless environment, in which case you can use the System Partitioning Aid to build your own layout.

On the other hand, if you are certain that you will never want to use the SP Switch or SP system partitioning, or your need for more SP-attached servers is greater, see “Understanding partitionability choices” on page 97.

Example 2 – A switchless system

Figure 32 shows a switchless system having one frame and only 7 nodes. Partitioning this system might be helpful for migration testing similar to that discussed in Example 1. In this case, since there is no switch, we are not bound by switch chip-related rules. We can assign nodes to partitions in any way we want.

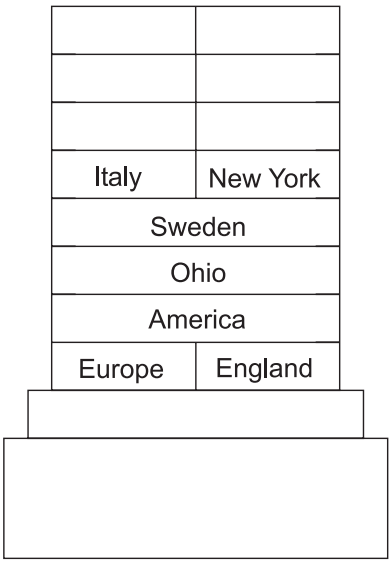


Figure 32. One sparse frame with no switch

For example, suppose you wanted to divide the system into 2 pieces as follows:

1. In Partition 1, group the Europe node and its affiliates, which are Italy, Sweden, and England.
2. In Partition 2, group the America node and its affiliates, which are New York and Ohio.

Using node numbers, you have:

```
              Partition 1                Partition 2
              -----------                -----------
        nodes  1,2,7,9               nodes  3,5,10
```

This configuration does not match any of the predefined layouts. Therefore, you would use the System Partitioning Aid to construct it.

The System Partitioning Aid

The System Partitioning Aid allows you to create a new system partition layout. In other words, if none of the layouts shipped with the SP meets your needs, you can use the System Partitioning Aid to generate one that does; and you can save this new layout for future reference.

The System Partitioning Aid provides both a Graphical User Interface (GUI) and command line interface. If you are experienced with partitioning or you have a simple system environment, the command line interface might serve your needs. While learning, or for more complex situations, you might find the GUI interface more beneficial.

The System Partitioning Aid supports the partitioning of systems with up to 128 nodes, whether switchless or switched (contains one or more switches). However, any SP system must be switch-wise homogeneous: system partitioning does not support the joining of switched and switchless systems.

Details on the System Partitioning Aid appear in the books *PSSP: Administration Guide* and *PSSP: Command and Technical Reference*. The system partitioning examples in this chapter should help you understand the value of the System Partitioning Aid.

Accessing data across system partitions

In addition to the restrictions on switch traffic, as illustrated in Example 1, data cannot generally be shared across system partitions. Therefore:

- Access to IBM Virtual Shared Disks and pseudo-tape devices across system partitions is not supported.
- multi-host attached disks cannot span system partitions.
- A physical file system, that is, the logical volumes containing the files, cannot span system partitions.
- You can use a distributed file system to mount file systems across partition boundaries, just as you would use a distributed file system from one SP to another. Keep in mind that doing this might effect nodes in both partitions in terms of both compute and network utilization.

The relationship of SP resources to system partitions

The SP can have a variety of both hardware and software resources associated with it. This section discusses how these resources interact with each other with regard to system partitions.

Single point of control with system partitions

You manage a partitioned SP system from a single point of control using the control workstation. There is one common administrative domain from which you can restrict interaction to one system partition.

The common administrative domain

From an administrative point of view, each partition is a logical SP system within one common administrative domain. This means that:

- Only one control workstation is needed. (If using a high availability control workstation, two workstations are available, but only one is used as the control workstation at any point in time.)
- The hardware monitor allows an administrator to control and monitor the entire system or a system partition. The administrator can issue commands that affect one, several, or all system partitions.
- There is one Kerberos V4 realm for the entire system.
- There is one DCE cell (Kerberos V5 realm) for the entire system.
- There is one user name space for the entire system.
- There is one accounting master for the entire system.
- The boot/install functions of a server node ignore system partition boundaries. However, a boot/install server must be at the same or later AIX and PSSP level as the nodes it is serving.

The SP_NAME environment variable

The entire SP is one administrative domain for the system administrator, who manages the system partitions as logical SP systems. An administrator restricts interaction to a specific system partition by setting the SP_NAME environment variable to the name or IP address of that system partition.

On the control workstation, the administrator is in an environment for one system partition at a time, as defined by the SP_NAME environment variable. Any task performed at the control workstation that requires information from the SDR gets the information for the current system partition. The operator must either set the SP_NAME environment variable or issue a command that sets it. If SP_NAME is not set, the environment is the default (or persistent) system partition.

The SDR in a partitioned system

The SDR contains data about the entire SP system. Generally, this data is separated into *system* (global) and *partitioned* classes. Requests made to the SDR, whether in software or manually, require an appropriate name or IP address for the system partition. If no such identifier is specified, the value of SP_NAME is used.

On the control workstation, the administrator is in an environment for one system partition at a time as identified by the SP_NAME environment variable. Any task performed at the control workstation that gets information from the SDR gets the information for the current system partition. Also, all global data (data affecting **all** system partitions) is accessible from any system partition.

Networking considerations

System partitioning does not require physical changes to the networking configurations of a system. You should consider certain effects that might warrant a physical change.

Ethernet interference, causing slower performance, might occur between nodes on the same physical Ethernet subnetwork. If these nodes are in different system partitions, an action such as booting all the nodes in one system partition might adversely affect the other system partition. You should consider creating system partitions aligned on the physical Ethernet subnetwork boundaries. This is fairly straightforward for system partitioning where the partitioning is on frame boundaries.

There is no connectivity over the switch between system partitions. This means that a gateway node with routing set up to the switch network might require routing changes if the gateway is to remain a gateway for more than one system partition. You can do this using explicit host routes on the gateway node, or by enabling ARP on all system partitions and redefining the IP addresses within a system partition as a different subnetwork.

Running multiple levels of software within a partition

Remember that a system partition is an SP system — essentially a smaller SP carved out of the whole one. You cannot expect the smaller SP to do what the larger cannot. However, more flexibility was introduced with *coexistence* to support migration and make it easier to upgrade production applications.

With coexistence, nodes can still be divided into partitions. However, coexistence lets each node within that partition run its own individual version of PSSP. Within that node, any software that operates under that node's version of PSSP will still function. Even though the nodes are running a variety of PSSP levels, the SP

system still functions normally. Therefore, depending on migration and coexistence limitations of the software you have or plan to install, you might be able to migrate your SP system one node at a time.

For additional information on this support, including supported levels and limitations, see Chapter 11, “Planning for migration” on page 213.

Overview of rules affecting resources and system partitions

The SP resources must conform to certain rules if they are to be a part of a system partition. The following list provides an overview of these rules:

- An unpartitioned SP is treated as a single system partition.
- Each node within a system partition must use the same set of authentication methods enabled in that partition. Each system partition can use a different set of authentication methods. The set of authentication methods enabled on the control workstation must be the union of all the authentication methods enabled in each system partition plus any other method currently set.
- The number of system partitions you can define depends upon the size of your SP and on the way that nodes are connected. In order to achieve isolation between system partitions, the nodes connected to the same switch chip belong to the same partition.
- Each system partition in a system with an SP Switch has a primary node, for switch initialization, and a backup primary node.
- Each system partition has an associated topology file which defines the portion of the switch network that it owns. Switch initialization occurs within a system partition for that portion of the switch fabric that is defined by the corresponding topology file.
- Switch operations and message traffic are managed within a system partition.
- You can have multiple system partitions in a LoadLeveler cluster and user space jobs can run in any suitably configured system partition, but any one user space job must run within a system partition: it cannot span system partitions.
- The IBM Virtual Shared Disk support and the pseudo-tape device driver cannot cross system partition boundaries. The IBM Recoverable Virtual Shared Disks and multi-host attached disks must be connected to nodes within the same system partition.

  A physical file system, that is, the logical volumes containing the files, cannot span system partitions.
- Each system partition has subsystems that are system partition-sensitive because they operate within a partition rather than throughout the entire system. These subsystems (such as hats and hags) are managed by the Syspar Controller which operates through the **syspar_ctrl** command. This command provides a single interface to control system partition-sensitive subsystem scripts. For more information see the book *PSSP: Administration Guide*.
- HACMP clusters do not span system partition boundaries.

System partitioning for systems with multiple node types

The physical types supported in the SP system for running PSSP are: thin, wide, and high nodes in SP frames, and SP-attached servers. The physical type affects the membership possibilities in a system partition. To understand how you can run multiple node types within a system partition, you need to understand the concepts of *node slots* and *switch chips*. A ***node slot*** is the space that one thin node can occupy in an SP frame. Every node or SP-attached server gets connected to one port on the ***switch chip***.

Thin-node frames

There are 16 node slots in an SP frame. Figure 33 shows how the slots are numbered in a frame. In Example 1, we considered a 1-frame system of 16 thin nodes. In that case, there is one node per slot, and the number of a node is precisely the number of the slot it occupies.

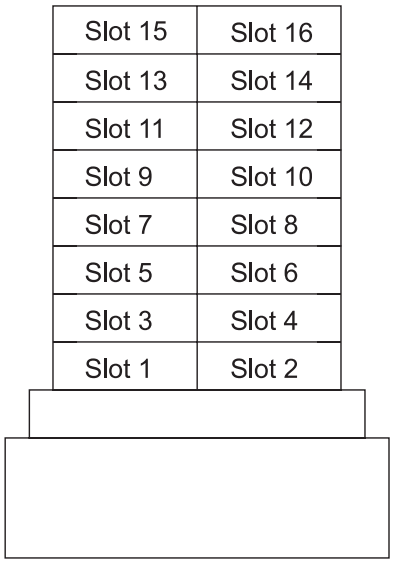


Figure 33. One SP frame with slots numbered

Partitioning with wide and high nodes

A wide node occupies two adjacent slots (a drawer) and a high node occupies four adjacent slots (2 adjacent drawers). The correspondence between node numbers and slot numbers is a topic of Example 3. For now, remember a node's node number is the lowest numbered slot that it occupies. As you plan your system partitions, think in terms of slots. Then you can decide what combination of thin, wide, and high nodes you want to occupy those slots.

Figure 34 on page 183 shows a frame populated with 3 wide, 1 high, and 6 thin nodes. The nodes in that figure have been given simple names using their node number. Note that nodes 2, 8, 10, 11, 12, and 14 do not exist. The preceding discussion expands to the following complete summary for the slots of the frame in Figure 33:

- Slots 1 and 2 contain wide node 1
- Slot 3 contains thin node 3
- Slot 4 contains thin node 4
- Slot 5 contains thin node 5
- Slot 6 contains thin node 6
- Slots 7 and 8 contain wide node 7
- Slots 9, 10, 11, and 12 contain high node 9
- Slots 13 and 14 contain wide node 13
- Slot 15 contains thin node 15
- Slot 16 contains thin node 16

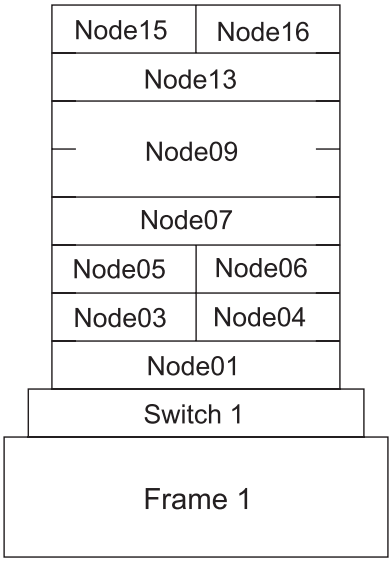


Figure 34. Varied nodes, 1-frame SP system

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

A system partition can be no smaller than a switch chip and the nodes attached to it which occupy some number of slots in the SP system. Following are examples of possible scenarios of nodes attached to a single switch chip:

- Four thin nodes attached (4 slots)
- Three thin nodes attached (3 slots) and one unused node switch port
- Two wide nodes attached (4 slots) and 2 unused node switch ports
- One wide node and two thin nodes attached (4 slots) and 1 unused node switch port
- One wide node and one thin node attached (2 slots) and 2 unused node switch ports
- One high node and two thin nodes attached (6 slots) and 1 unused node switch port

  Note: A high node occupies 4 adjacent slots. The high node is attached to one switch chip at one port.
- One high node and one wide node attached (6 slots) and 2 unused switch ports

In practice, every slot is assigned to some chip, via fictitious nodes if necessary, so that if that slot is later filled with a node, it is not a major reconfiguration event.

SP-attached server

An SP-attached server is not in an SP frame, but it is managed by the PSSP components as though it is in a frame. The SP-attached server is always viewed as occupying slot one in its frame. Its frame is considered to have 16 node slots, just like SP frames. However, because it must attach to an existing SP frame, it must occupy a port in the switch chip whether or not it also connects to an SP Switch.

Example 3 – An SP with 3 frames, 2 SP Switches, and various node sizes

Figure 35 shows a 3-frame system containing wide nodes, thin nodes and high nodes. The nodes have been named in accordance with their frame and slot location.

Note: Keep in mind that you probably cannot order the system discussed in this example. This system has nodes located in legitimate locations, but the models available to order from IBM might not include this configuration. Over time, however, you might add, delete, and move nodes of your system such that you might arrive at a similar system and be faced with similar considerations.

There is an SP Switch in each of frames one and three. Frame two is sharing the first frame's SP Switch, which is possible because the configuration of frames one and two is an example of configuration number 1 in Figure 15 on page 122. You can connect a maximum of 16 nodes to a switch board. Since Frames 1 and 2 have only 11 nodes total, there is room for future expansion.

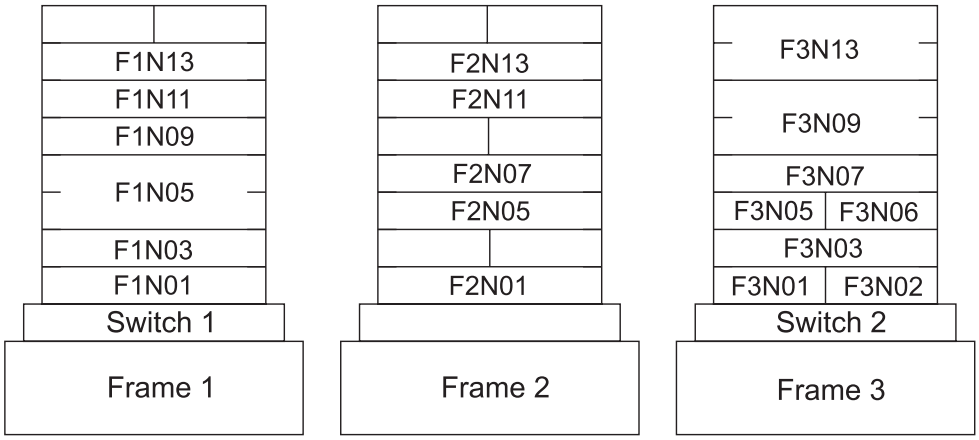


Figure 35. Three SP frames with 2 SP Switches

The nodes in a system are assigned node numbers sequentially across the frames, bottom left to top right, except that node numbers are skipped to accommodate later expansion and node shifting. Put another way, the first 16 node numbers are assigned to the 16 slots of the first frame, the next 16 node numbers to the 16 slots of the second frame, and so on. The following are cases where node numbers are skipped:

1. A wide node takes up two slots

   For example:
   - there is no F1N14, or node number 14, because wide node F1N13 occupies both slots 13 and 14 of Frame 1
   - there is no F2N08, or node number 24, because wide node F2N07 occupies both of slots 7 and 8 of Frame 2
2. A high node takes up four slots

   For example:
   - F1N06, F1N07 and F1N08 (node numbers 6-8) cannot exist because high node F1N05 takes up all of slots 5-8 in Frame 1
3. A slot is left empty

For example:

- there is no F1N15, or node number 15, because slot 15 in Frame 1 is unoccupied
- there is no F2N03, or node number 19, because slot 3 of Frame 2 is unoccupied

So, how do the nodes in this system attach to the SP Switches? Each switch can have 16 nodes attached. Therefore, the system has 32 *node switch ports*. The system needs to know to which of these ports each node is connected. These node switch ports are numbered 0 through 31. The system understands the *switch port number* of a node to be the number of the node switch port to which the node is connected. The switch port number of a node is sometimes called its *switch node number*.

In the 16-node system of Example 1, a node's switch port number is one less than its node number, because switch port numbers start at zero. Therefore, node number 1 has switch port 0, and so on through node number 16 which has switch port number 15.

Note: Although this discussion might sound complicated, it really isn't. Just keep in mind that a node generally sits in the midst of a large system and, at any point in time, you might care about any one of the following:

- Where does the node sit in its frame? (slot number)
- What is the node's position relative to all the rest of the nodes in the system? (node number)
- Where does the node connect to the switch fabric? (switch port number, or switch node number)

You can ascertain the current node number mapping for a system under operation by issuing the command **sysparaid -i**.

When possible, switch connections are made as illustrated in Example 1. Therefore, in Frame 1 of Figure 35 on page 184, F1N03 sits in slot 3, is node number 3 and has switch port number 2. The wide node F1N01 is node number 1 and uses switch port number 0.

There is no node number 2, so switch port number 1 is not used by Frame 1. However, F2N01 in Frame 2 needs a switch port, and is a likely candidate to take the place of the missing node number 2 on Switch 1. So, F2N02 occupies slot 1 of Frame 2, is node number 17 in the system, and uses switch port number 1.

Continuing along this track, F1N05 uses switch port number 4 and F2N05 uses switch port 5. Switch port 6 is unused since there is no F1N07, but switch port 7 is used by F2N07.

Switch port numbers continue with the next switch of the system. So, F3N01 uses switch port 16, F3N02 uses switch port 17, and so on. However, the F3N01 is node number 33 and F3N02 is node number 34.

Now, assume you want to partition this system as follows:

```
Partition 1 - F1N01, F2N01, F1N05, F2N05,
              F1N03, F2N07
Partition 2 - F1N09, F1N13, F2N13
              F1N11, F2N11
Partition 3 - F3N01, F3N02, F3N05, F3N06,
              F3N03, F3N07,
              F3N09, F3N13
```

The nodes are listed in this order on purpose – by switch chip. This layout is not among the predefined ones shipped with the SP. You can use the System Partitioning Aid to help specify this layout. First, recognize that for this system to ever have been operational, the system was installed and its specific makeup (existing frames, existing switches, node names, node types, node numbers, switch port numbers), was stored in the SDR. The System Partitioning Aid has that data to build upon. To specify the system partitioning layout you want, do one of the following:

- Invoke the System Partitioning Aid from the command line, specifying the partitions via node lists in an input file. For more information see the book *PSSP: Command and Technical Reference*.
- Bring up the graphical user interface of the System Partitioning Aid, by using the **spsyspar** command, and select the nodes for each partition using a pointer device.

  This interface is also available under the *SP Perspectives* graphical user interface.

You can plan a system partitioning layout before it is realized. This topic is discussed in Appendix A, “The System Partitioning Aid - A brief tutorial” on page 239.

The System Partitioning Aid will not allow you to do something inappropriate like split a switch chip among partitions; nor define a partition having extremely poor bandwidth or reliability over the switch. (See the PSSP: Administration Guide for additional information on such restrictions.) When you are satisfied, the System Partitioning Aid will save your layout information in an appropriate directory. Note that layouts are classified based on chip assignments and the maximum number of nodes which can be attached to those chips. Therefore, this layout would be saved as an 8_8_16-layout. (8+8+16 = 32 is the maximum number of nodes you can attach to 2 switches.)

System partitioning configuration directory structure

System partitioning is supported by the SP software in the **ssp.top** install file set. You can choose to install this support when you install PSSP on the control workstation. This provides the system with a directory of predefined system partitioning layouts, as well as the System Partitioning Aid, a tool for building additional layouts. The directory is represented in Figure 36 on page 187. An introduction to the System Partitioning Aid is provided in Appendix A, “The System Partitioning Aid - A brief tutorial” on page 239.

For system partitioning purposes, a system is cataloged by its switch configuration. The number of used node slots in the system and the type of nodes the system contains play a role in how you want to partition your system. *However, the quantity and kinds of switches determine your options.*

A switch board to which nodes are connected is called a *Node Switch Board* (NSB). In larger systems, it becomes impossible to adequately connect all pairs of NSBs to each other. Additional switch boards are inserted to provide additional connectivity. These *extra* switch boards have no nodes attached, just other switch boards. Such a switch board is called an *Intermediate Switch Board* (ISB).

For example, the 1-frame system considered in Example 1 is classified as a 1nsb0isb system. It has 1 NSB and 0 ISBs. The system of Example 3 had 3 frames, but only 2 switch boards. It is a 2nsb0isb system.

The **syspar_configs** directory within the **spdata** file system contains all system partition configuration information. Figure 36 shows this directory structure. In this figure, subdirectory 2nsb0isb is expanded to illustrate the predefined layouts available for such systems:

1. The system has a maximum of 32 nodes, 2 switches with up to 16 nodes each.
2. Using the predefined layouts shipped with the SP, the system can be configured as (partitioned into) 4_28, 8_24, or 16_16 subsystems; or it can be used as an undivided 32-node system.

   "Example 3 – An SP with 3 frames, 2 SP Switches, and various node sizes" on page 184 illustrates how to construct a new 8_8_16 layout. You could use the System Partitioning Aid to save this layout, in which case, the System Partitioning Aid would introduce a corresponding new config.8_8_16 directory in the 2nsb0isb subtree. Within that new config-level directory, a layout subdirectory would be introduced named **layout.***name_desired* where *name_desired* is a name you specified to the System Partitioning Aid.

   "Example 2 – A switchless system" on page 178 illustrates (implicitly) how to construct a new, switchless 4_12 layout. Although only 7 nodes were available, you had a full frame for which the maximum size system is 16; categorization is based on the maximum number of nodes and any unlisted nodes go in the last partition. If you use the System Partitioning Aid to save this layout, the System Partitioning Aid would save it in the **1nsb0isb** subtree as **1nsb0isb/config.4_12/layout/layout.***name_desired* where *name_desired* is a name you specify to the System Partitioning Aid.
3. For the 4_28 case, there are 8 different layouts available, one for each of the 8 node switch chips in the 2 switches.
4. For each available layout, the corresponding subdirectory contains a description file (**layout.desc**) and the specifics of the individual system partitions; the partition's **nodelist** file and topology file.

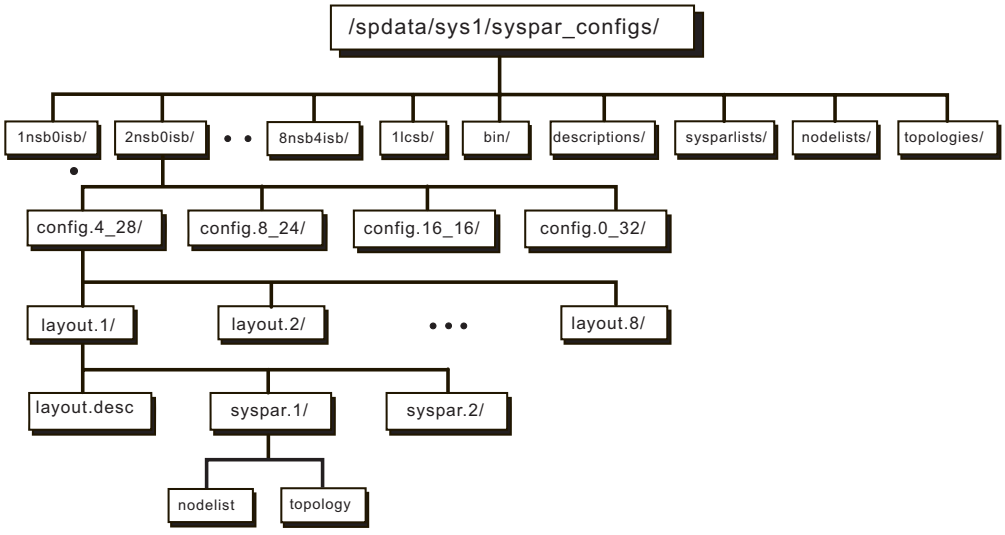


Figure 36. The directory structure of system partition information

The higher-level directories **descriptions**, **sysparlists**, **nodelists**, and **topologies** contain the files common to various configuration layouts. For predefined layouts, the low-level files **layout.desc**, **nodes.syspar**, **nodelist**, and **topology** are actually links into these higher-level directories. For layouts constructed via the System Partitioning Aid, no links are used; the actual files are stored at these lower levels.

The specifics of each of the predefined configurations are in Appendix B, “System Partitioning” on page 259. Consult that information as you complete the worksheets in Appendix C, “SP system planning worksheets” on page 281.

Chapter 8. Planning to record and diagnose system problems

This chapter discusses the log files you can keep, how to get help, and the diagnosis tools that are available.

Configuring the AIX error log

The AIX Error Log facility is configured by default to be 1 MB in size. When the log fills up, it wraps around and overwrites existing entries. The PSSP software utilizes the AIX Error Log frequently. Therefore, it is good to set the size to at least 4 MB after all nodes are installed. You can do this once for all nodes with the command:

```
dsh -a /usr/lib/errdemon -s 4096000
```

Configuring the BSD syslog

The PSSP software uses the Berkeley Software Distribution (BSD) syslog subsystem.

The control workstation

The PSSP install process configures the BSD syslog subsystem on the control workstation only to write all syslog messages for its daemons to the **daemon.notice** facility. All kernel error messages are logged via the AIX Error Log Facility. If the control workstation already has the **daemon.notice** facility configured, it does not change that configuration.

PSSP nodes

The BSD syslog facility is not configured on the PSSP nodes when it is installed. By default, AIX does not configure the BSD syslog. The configuration file for BSD syslog is **/etc/syslog.conf**. Configure the syslog on all nodes where you want entries made. Note also that any PSSP error logs are also written to the AIX Error Log and usually contain more information about probable cause and possible recovery or diagnostic actions. In AIX, the predominant error logging facility is the AIX error log and only application software that was ported from 'other sources' contains the calls to syslog and logger. The AIX kernel does not use syslog for error logging.

The PSSP File Collections facilities can be used to manage the **/etc/syslog.conf** file if all the nodes have the same configuration file. Be aware that with the amount of information syslog collects, logs might consume the network resources on the system if they are forwarded to a single node. Additionally, the system multiplexes the **/dev/console** tty cables onto a single cable per frame. If **/dev/console** is used for syslog messages performance problems might occur. If you want syslog messages, IBM suggests that they be logged on a per-node basis, and that you use tools like **dsh** and **sysctl** to view and manage them.

See the *PSSP: Diagnosis Guide* for more information about error logging.

SP system logs

The various PSSP and PSSP-related components create logs during normal operations in the following directories:

/var/adm/SPlogs/*****
/var/adm/SPlogs/auth_install/*****

/var/adm/SPlogs/auto/*
/var/adm/SPlogs/cs/*
/var/adm/SPlogs/csd/*
/var/adm/SPlogs/css/*
/var/adm/SPlogs/css0/*
/var/adm/SPlogs/css1/*
/var/adm/SPlogs/filec/*
/var/adm/SPlogs/filec/logs/*
/var/adm/SPlogs/get_keyfiles/*
/var/adm/SPlogs/kerberos/*
/var/adm/SPlogs/kfserver/*
/var/adm/SPlogs/pman/*
/var/adm/SPlogs/sdr/*
/var/adm/SPlogs/SPconfig/*
/var/adm/SPlogs/spacs/*
/var/adm/SPlogs/spmgr/*
/var/adm/SPlogs/spmon/*
/var/adm/SPlogs/spmon/nc/*
/var/adm/SPlogs/st/*
/var/adm/SPlogs/sysctl/*
/var/adm/SPlogs/sysman/*
/var/ha/log/hags/*
/var/ha/log/em/*
/var/ha/log/hats/*
/var/ha/log/nim.*
/var/ha/run/*

See the *PSSP: Diagnosis Guide* book for details on log information.

Finding and using error messages

Most error messages generated by the SP are listed and explained in the *PSSP: Messages Guide*. The book lists the messages in numerical order. Each message should have a part labeled “User Response” that describes the actions, if any, that you should take when you encounter the message. If the information in a message does not help resolve the problem, you should have users follow a predefined path for resolving the problem.

Getting help from IBM

Before you call for help, check if all the latest service has been applied to your system. Then see the relevant discussion in the book *PSSP: Diagnosis Guide* to help you diagnose problems before placing a call. If you still need help from IBM in resolving a Cluster 1600 system managed by PSSP problem, you can call IBM. You might be asked to send relevant data and to open a problem management record (PMR) for tracking purposes.

Finding service information

A Web site contains all the service bulletins and flashes as well as PTF and APAR reports for all current releases of PSSP, LoadLeveler, and Parallel Environment. The Web address is:

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

Calling IBM for help

You can get assistance by calling IBM Support. Before you call, be sure you have the following information:

1. Your access code (customer number). This number was entered on Worksheet 4, “Major system hardware components ” in Table 63 on page 284.
2. The IBM product number, for example:
   - For a problem with PSSP, use product number: 5765-D51
   - For a problem with LoadLeveler, use product number: 5765-E69

   Similarly, each product has its own number that will speed the correct routing of your call. See Table 50 on page 215.
3. The name and version of the operating system you are using.
4. Any relevant machine type and serial numbers.
5. A telephone number where you can be reached.

The person with whom you speak will ask for the above information and give you a time period during which an IBM representative will call you back.

In the United States:
The number for IBM software support is **1-800-237-5511**.
The number for IBM AIX support is **1-800-CALL-AIX**.
The number for IBM hardware support is **1-800-IBM-SERV**.
The number for the IBM PC Help Center is **1-800-772-2227**

Outside the United States, contact your local IBM Service Center.

Sending problem data to IBM

You might be asked to produce a system dump and send it to the IBM support office. See the book *PSSP: Administration Guide* for instructions on how to produce this information.

Customers within the United States

To send the data to IBM, label the tape or diskette with the problem number and mail it to:

IBM AIX Customer Service and Support
Dept. 39KA, Mail Station P961, Bldg. 415
2455 South Road
Poughkeepsie, N.Y. 12601-5400

ATTN: APAR Processing

Customers outside the United States

Your local IBM Service Center can provide you with the address to use.

Opening a Problem Management Record (PMR)

A PMR is an online software record used to keep track of software problems reported by customers.

Follow your local support or service procedures for opening a PMR.

Note: To aid in quick problem determination and resolution, it will be very useful to have the SDR data specific to the problem included in the PMR. You can

obtain the SDR data using the **splstdata** command. Use the appropriate command flag to view data relevant to the problem. For example:

splstdata -e
Lists environment choices

splstdata -n
Lists node information

splstdata -s
Lists switch information

For more information on **splstdata**, refer to *PSSP: Command and Technical Reference*.

IBM tools for problem resolution

IBM offers several tools to help you with efficient problem resolution. Service Director for RS/6000 is standard with the SP while others are separate software packages.

Inventory Scout

Inventory Scout is an AIX tool that surveys servers and workstations for hardware and software information. Inventory Scout is used by the following Web services:

- Microcode Discovery Service, which provides a customized report indicating if installed microcode is at the latest level.
- VPD Capture Service, which can collect your server's vital product data and transmit this information to IBM for matching with a Miscellaneous Equipment Specification (MES) upgrade.

Inventory Scout runs on AIX POWER Versions 4.1.5 or later. The tool can be invoked by Java applets or run from the command line.

Service Director for RS/6000

Service Director is a set of IBM software applications that monitor the health of your SP system. Service Director analyzes AIX error logs and runs diagnostics against those error logs. You can define which systems have the Service Director clients and servers. You can also define the level of error log forwarding or network access.

During error conditions, Service Director analyzes the severity of the fault and determines whether or not to capture fault information. Depending on how you configure Service Director, the IBM support center and the responsible system administrator at your location receive E-mail containing the fault information. If a Service Request Number is created, a record of that is created, the product automatically sends a message (call home) to IBM and a PMR is opened. Upon receiving the fault notification, IBM will automatically dispatch a service engineer with the parts needed to correct the problem, if such an action is needed.

Planning the Service Director's physical environment

Service Director requires a local server. The local server must have:

- An available S1 serial port.
- 5 MB of free disk space.

Typically, the local server is the control workstation. However, if the control workstation does not have an available serial port, any other workstation on the LAN can act as the local host.

The local host uses the required serial port for a modem interface. The modem is then used to transmit fault messages to IBM and your system administrator over local phone lines. All new RS/6000 SP systems include a modem package as part of the ship group. This package includes:

- An IBM compatible modem (minimum 9600 bps baud rate).
- A 9 pin to 25 pin serial cable.
- A 25 pin extension cable fifteen meters long.

You **must** supply the following:

- An external, analog phone line.
- A telephone extension cable capable of reaching the modem from the phone jack.

In addition to the local host's physical requirements, **all** nodes in your SP system **must** have the client version of Service Director installed. This **requires** 1.5 MB of free disk space on each node.

Planning Service Director's Software Environment

SP systems require Service Director 2.1 (or later). The disks and documentation you will need to install Service Director are included with the modem equipment in the ship group package sent with all new SP systems. In addition to the disk space requirements listed above, Service Director 2.1 has the following prerequisites:

- AIX 4.1 or later.
- IBM Diagnostics must be active on all workstations and nodes.
- Error logging must be active on all workstations and nodes.

Service Director can be installed concurrently. During installation, you will be presented with several customization options for system analysis scheduling and error notification. Once installed, Service Director runs dynamically under AIX and is capable of using the local server to display a structured view of problem management information. This information includes:

- Recent hardware events.
- A history of past hardware events.
- Statistical analysis of problems logged into Service Director.
- Client node status may be viewed remotely from the local server.

Service Director error logs are maintained in each node and are not consolidated in the local server.

Note: Operating system upgrades can introduce new error logs. Therefore, Service Director software upgrades might be needed when you upgrade PSSP and AIX.

Considering security

Service Director accesses only system error logs, system diagnostic data, and vital product data (VPD) files. All information that is transmitted to IBM is also routed by E-mail to the system administrator that you assign. Service Director disables the login capability on the assigned serial port and the modem configuration will not permit auto-answer. **No customer-unique data is ever accessed**.

NetView for AIX

NetView for AIX manages multi-vendor networks by polling the base AIX SNMP daemon agents to gather information for display and action by network control desk. It performs the following functions:

- Automatic discovery of the network (creating and maintaining topological network maps)
- Performance management for monitoring network status, displaying critical network resource status and statistical summaries for analysis and corrective actions
- Fault management for verifying the integrity of the network, utilizing threshold and filtering algorithms for easier alert notification, and defining and implementing corrective actions to SNMP traps

Note that NetView for AIX is not supported on the control workstation.

EMEA Service Planning applications

The EMEA Service Planning offering, available directly from EMEA, runs a set of application programs managed by **cron** and the AIX Error Notification Facility to collect data from the ErrorLog, Syslog, **/var/adm**, and **/tmp** from individual nodes. The data is stored at the control workstation. The application, if required by events in the logs, calls the support center and opens a PMR.

Chapter 9. Planning for PSSP-related licensed programs

This chapter briefly discusses planning information for PSSP-related licensed programs that are offered as Cluster 1600 software building blocks. You might want to consider them when planning your Cluster 1600 system managed by PSSP. See Chapter 11, “Planning for migration” on page 213 for versions supported, coexistence, or migration information. For complete detailed information on the individual licensed programs, see their books which are listed in “Bibliography” on page 313.

Enhanced security options:
You have the option of running your system with an enhanced level of security. The restricted root access option removes the dependency PSSP system management software otherwise has to internally issue **rsh** and **rcp** commands as a root user from a node. With restricted root access active, any such actions can only be run from the control workstation or from nodes configured to authorize them and PSSP does not automatically grant authorization for a root user to issue **rsh** and **rcp** commands from a node. If you enable this option some procedures might not work as documented. For example, to run HACMP an administrator must grant the authorizations for a root user to issue **rsh** and **rcp** commands that PSSP would otherwise grant automatically.

You can use a secure remote command process to be run by the PSSP system management software in place of the **rsh** and **rcp** commands.

Some of the licensed programs discussed in this chapter might be affected in some way with these enhanced security options enabled. See “Considering restricted root access” on page 146, “Considering a secure remote command process” on page 151, and “Considering choosing none for AIX remote command authorization” on page 152 for descriptions of these options. See the respective licensed program publications for information about running them under various PSSP security configurations.

Planning for Parallel Environment

The IBM Parallel Environment for AIX program product is designed to help you develop parallel programs and execute them on the Cluster 1600 system managed by PSSP (SP system) or a networked cluster of RS/6000 processors. See “Parallel Environment” on page 28 for a functional description.

Parallel Environment supports the Message Passing Library (MPL) subroutines, the Message Passing Interface (MPI) standard, and the Communications Low-Level Applications Programming Interface (LAPI). If you plan to develop applications using the LAPI component of PSSP 3.5, you must use Parallel Environment 3.2. See Chapter 11, “Planning for migration” on page 213 for more release level compatibility information.

Be aware that parts of the Parallel Environment installation steps might interact with or be affected by PSSP component installations, particularly the switch services component of PSSP (**ssp.css**) and the SP security services (**ssp.clients**). See the book *PE: Installation Guide* for details on planning for and installing Parallel Environment, particularly if you are interested in any of the following:

- Installing Parallel Environment on a control workstation.
- Installing Parallel Environment to run off the rack, with the **ssp.clients** file set.
- Installing the switch services component of PSSP (**ssp.css**) after Parallel Environment has been installed.

Be sure to address the following considerations:

1. Ticket and credentials lifetimes.
2. What ticket and credential lifetimes need to be set with respect to PE and LoadLeveler. The lifetime needs to be long enough to allow for the expected longest running job to complete.
3. DCE server replication when large numbers of tasks are to be used.
4. The DCE user id is used if it is different from the AIX user id. IBM suggests you use DCE integrated login.

Planning for Parallel ESSL

Parallel ESSL is a scalable mathematical subroutine library that supports parallel processing applications on the Cluster 1600 system managed by PSSP (SP system). See “Parallel Engineering and Scientific Subroutine Library” on page 29 for a functional description.

The subroutines run under the AIX operating system and can be called from application programs written in Fortran, C, C++, and High Performance Fortran (HPF). On the SP system, PSSP is required.

The design of Parallel ESSL centers on exploiting operational characteristics and the parallel processing architecture of the Cluster 1600 system managed by PSSP (SP system). The latest release, Parallel ESSL 2.3, is designed to run on the SP Switch2 or the SP Switch. To take advantage of this increased performance, programs that already use the routines need only be relinked, not recompiled.

For communication, Parallel ESSL includes the Basic Linear Algebra Communications Subprograms (BLACS), which use the Parallel Environment (PE) Message Passing Interface (MPI). Communications using the User Space (US) require use of an SP switch. Communications using the Internet Protocol (IP) can use Ethernet, Token Ring, FDDI, SP Switch2, SP Switch, or SP Switch-8.

If you want to use this licensed program with PSSP 3.5, you also need IBM ESSL and Parallel Environment. See “Parallel ESSL” on page 234 for software level dependencies, coexistence, or migration information.

Planning for High Availability Cluster Multi-Processing (HACMP)

IBM’s tool for building UNIX-based mission-critical computing platforms is the High Availability Cluster Multi-Processing (HACMP) for AIX software package. HACMP ensures that critical resources are available for processing. See “High Availability Cluster Multi-Processing” on page 27 for a functional description.

Typically, HACMP or HACMP/ES is run on the control workstation only if HACWS is being used. HACMP/ES is run on the nodes.

Note: HACMP does not tolerate IPv6 aliases for IPv4 addresses.

For complete planning information and for running with HACMP under various PSSP security configurations, see the book *HACMP: Planning Guide*.

Planning for LoadLeveler

LoadLeveler is an IBM software product that provides workload management of both interactive and batch processing on a Cluster 1600 system managed by PSSP (SP system) or pSeries and RS/6000 workstation. The LoadLeveler software lets you build, submit, and process both serial and parallel jobs. See “LoadLeveler” on page 27 for a brief functional description.

Compatibility

The latest release is LoadLeveler 3.1. For compatibility information see “LoadLeveler” on page 231 in Chapter 11, “Planning for migration”.

For the most up-to-date migration instructions, see the README file distributed with LoadLeveler 3.1.

Planning for a highly available LoadLeveler cluster

LoadLeveler provides features within the product for automatic recovery in the event of failure of the central manager in the batch configuration and of the domain name server running Interactive Network Dispatcher in the interactive configuration. Additionally, the availability of individual compute nodes and file systems in the LoadLeveler cluster can be enhanced by using the High Availability Cluster Multi-Processing (HACMP) product as well as the High Availability Control Workstation (HACWS) optional feature of PSSP. For details on how to configure LoadLeveler for high availability, see the ITSO Redbook *Implementing High Availability on the RS/6000 SP*.

Planning your LoadLeveler configuration

In general, planning the LoadLeveler installation for workload management requires making the following configuration decisions. You must decide what is suitable to your environment. Be sure to address the following:

- Consider ticket and credentials lifetimes.
- Decide what ticket and credential lifetimes need to be set with respect to PE and LoadLeveler. The lifetime needs to be long enough to allow for the expected longest running job to complete.
- Consider using DCE server replication when large numbers of tasks are to be used in parallel jobs.
- The DCE user id is used if it is different from the AIX user id. IBM suggests you use DCE integrated login.
- Select a node to serve as central manager and one or more alternate central managers. The central manager can be any node in the cluster. In selecting one, consider the current workload and network access. Note that no new work can be performed while the central manager is down, and no queries can be made about any of the running jobs without the central manager.
- Determine which nodes will be scheduling nodes, execution nodes, submit-only nodes, and public submit nodes.
- Determine where to locate home and local directories. For maximum performance, keep the log, spool, and execute directories in a local file system.
- Determine if LoadLeveler daemons should communicate over the switch. It may not be desirable in your environment to have the daemons communicate over the switch. You need to evaluate the network traffic in your system to determine if LoadLeveler IP communications over the switch is desirable.

- Determine if HACMP is necessary to provide failover capability of individual compute nodes or the switch. If using LoadLeveler in conjunction with HACMP, decide which nodes will be grouped together for backup purposes. (HACMP can only provide capability for up to eight nodes.) Each backup node needs to know which set of seven nodes it will back up. This relationship is defined in the form of HACMP resource groups.
- Determine if your SP workload includes parallel jobs and if they will involve the SP Switch. If so, you will need to perform additional configuration activities. See the LoadLeveler publication for details.

Other planning considerations:

1. LoadLeveler requires a **common name space** for the entire LoadLeveler cluster. To run jobs on any machine in the LoadLeveler cluster, you must have the same uid (system ID number for a user) and gid (system ID number for a group) on every machine in the cluster. If you do not have a user ID on one machine, your jobs will not run on that machine.
2. LoadLeveler works in conjunction with the NFS or AFS file systems. Allowing users to share file systems to obtain a single, network-wide image, is one way to make managing LoadLeveler easier.
3. Some nodes in the LoadLeveler cluster might have special software installed that you might need to run your jobs successfully. You should configure LoadLeveler to distinguish those nodes from other nodes using, for example, job classes.

Planning for General Parallel File System (GPFS)

General Parallel File System for AIX provides concurrent shared access to files spanning multiple disk drives located on multiple nodes. This licensed program provides file system service to parallel and serial applications in a number of environments, including SP or HACMP/ES clusters:

- An SP environment is any Cluster 1600 system managed by PSSP with an SP switch network and PSSP running as the primary system management software with the IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk components.
- An HACMP/ES cluster environment is any Cluster 1600 system managed by PSSP or other system running the HACMP/ES licensed program.

This discussion is limited to using GPFS in an SP environment.

In the SP environment, the boundaries of the GPFS cluster depend on the switch type being used. With the SP Switch2, the GPFS cluster is equal to all of the nodes in the system. With the SP Switch, the GPFS cluster is equal to the SP system partition in which it is configured. Within a GPFS cluster, the nodes are divided into one or more GPFS nodesets. A *nodeset* is a group of nodes that all run the same level of GPFS and operate on the same file system. The nodes in each nodeset share a set of file systems that are not accessible by the nodes in any other nodeset.

You can modify your GPFS configuration after it has been set, but your reward for a little consideration before installing it is a more efficient file system.

Hardware and operating environment considerations:

- GPFS requires the IBM Recoverable Virtual Shared Disk component of PSSP. If IBM Recoverable Virtual Shared Disk is on multiple nodes within a system partition, each node must have the same level of PSSP.

- If you are using twin-tailed disks, you must select an alternate node as a backup IBM Virtual Shared Disk server.
- Do you have sufficient disks and adapters to provide the needed storage capacity and required I/O time?
- GPFS is supported only in systems with the SP Switch or the SP Switch2.

File size considerations:

- How much data will be stored and how large will your files become?
- How often will the files be accessed?
- Do your applications handle large amounts of data in single read/write operations or is the opposite true?
- How many files do you anticipate handling in the future?

Data recovery considerations:

- Node Failure: GPFS automatically reconfigures itself to continue operations without the failing node.
- IBM Virtual Shared Disk server and disk failure: Your recovery strategy depends on your answer to the question, 'Is your primary concern loss of data, loss of access to data, or do you need protection from both server and disk failure?'
  1. If data loss is your concern, a RAID device might be the best solution.
  2. If data access is your concern, twin-tailed disks could be your solution.
  3. If both data loss and access are potential problems, first consider mirroring at the logical volume manager for data recovery. If mirroring does not fit your system needs, another option is *replication*, which automatically creates and maintains copies of all file information.
- Connectivity failure: an adapter failure is treated as a node failure on SP Switch and one switch plane SP Switch2 systems. On SP Switch2 systems with two switch planes, a condition with two adapter failures on the same node is treated as a node failure.

Details on planning and implementing these strategies and other methods can be found in the book *IBM General Parallel File System for AIX: Concepts, Planning, and Installation Guide*.

Part 2. Customizing your system

Chapter 10. Planning for expanding or modifying your system

As your organization's processing needs and resources change, you might find that your current system setup no longer meets your needs. You might want to add, remove, or upgrade nodes, frames, or switches. Your changing needs might require you to perform other hardware or software modifications to your system. Planning ahead when you first configure your system can make future changes easier.

This chapter discusses the most common topics to consider prior to expanding or modifying your system. In addition, several sample scenarios illustrate the most common ways of expanding your system.

The book *PSSP: Installation and Migration Guide* discusses how to add, delete, or replace hardware in your system. Prior to expanding or modifying your system in any way, you should read this chapter to understand how to plan for the change. Careful planning will help ensure your system is back up and running as soon as possible.

The book *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* discusses site planning considerations such as planning for additional floor space or power concerns. Be sure to consult that book prior to expanding or modifying your system.

Note: There are many different ways that you can configure your system and each configuration requires you to plan for system setup. IBM tests and supports the most common configurations. Keep in mind that the more complex your specific configuration, the chances are less that IBM has tested that configuration. If you decide to expand or modify your configuration in a manner that is not addressed in this chapter or book, you should consult with your IBM representative prior to modifying your setup.

Questions to answer before expanding/modifying/ordering your system

This section poses some of the most common questions to consider prior to ordering or changing your system. These topics are illustrated in the scenarios presented later in this chapter.

To use an example, consider the expansion of the existing 3-frame system pictured in Figure 37 on page 204. This system has frames numbered 1, 2, and 4. Each frame has several unused node slots. Frames 1 and 4 have a switch, but Frame 2 does not. Frame 2 is a *non-switched expansion frame* whose nodes use the switch in Frame 1.

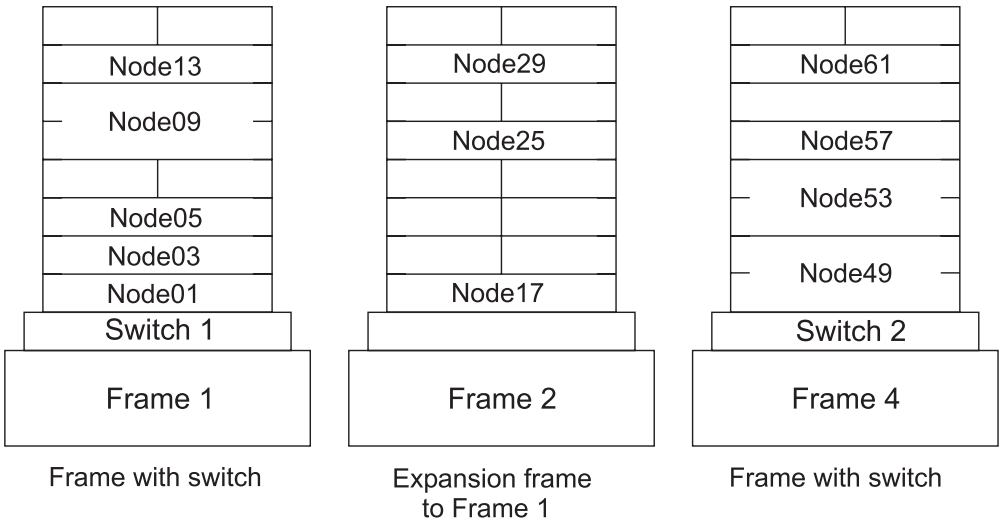


Figure 37. Sample SP Switch system: 3-frames, 1-switch

How large do I want my system to grow?

Before you expand your system, plan ahead for how large you want your system to eventually grow. Planning will encourage you to leave unused frame numbers for future expansion, and will help you avoid having to move nodes between frames. You can make the Sample SP Switch System grow by doing any of the following:

- Add nodes or SP Expansion I/O Units in empty slots
- Add a frame (perhaps Frame 3, 5, or 6) which might have more SP nodes. It might be an SP-attached server, or a non-node frame with only SP Expansion I/O Units. You might be expanding to beyond 5 switch frames so you need a switch-only frame of intermediate switch board (ISB) switches for more switch-to-switch connections.
- Install an SP Switch in Frame 2.
- Add an SP Switch Router.
- Drop the SP Switch and convert to the SP Switch2.

If you are planning a system of clustered enterprise servers, it can become a larger scale SP system. You can add an SP Switch2 or SP Switch in a clustered enterprise server configuration. Your system will then be subject to all the rules for that switch and these servers will become SP-attached servers. Consider the following:

- If you use the SP Switch2 you can add a multiple NSB frame to contain up to 8 node switch boards, each of which can accommodate up to 16 nodes. You need not be concerned with switch capsule or frame sequencing rules. You can number your frames in any order.
- If you want to add the SP Switch you have to add an SP frame to contain the switch which can accommodate up to 16 nodes. In the case of the SP Switch, plan your system honoring the switch capsule rules, with appropriate frame numbers and switch port numbers, so you can expand your system without having to totally reconfigure existing servers.

How do I reduce system down time?

Expanding a system can require that the system be shut down for an extended period of time. When adding a frame or switch to the system, there is often a great deal of cable wiring required. If you know that you want your system to grow in the future by adding nodes, frames, or switches, you might want to consider purchasing some of the hardware in advance. By purchasing in advance, you can set up the hardware and cables with the future in mind to avoid probable cable rewiring, node movement, and reconfiguration complexities at a later date. This can substantially reduce the amount of time your system will be down during future expansion activity.

Notice in the Sample SP Switch System that all 8 of the nodes in Frames 1 and 2 could reside in a single frame, but then many expansion choices would require adding a frame, moving nodes, cabling frames to each other, and so on. Such modifications cannot be done without considerable down time. However, the chosen configuration allows for some expansion without any major difficulties.

What must I understand before adding switches?

If you are considering adding a switch to a system that does not have any, keep in mind that the SP Switch2 does not support system partitioning though it does have advantages. See “Choosing a switch” on page 21 for more considerations.

If you are thinking about increasing the number of switches in your system, at least one of the following is pertinent when expanding from:

- 1 switch to 2 switches

  Cables need to be added, but the first switch could continue running until the new switch is ready for installation tests.
- 2 switches to 3 switches, or 3 switches to 4 switches

  Cables must be rerouted. This scenario can cause a significant amount of system down time. For highest availability, consider installing more frames initially, with empty slots for future node additions.
- 4 switches to 5 or more switches

  Addition of a switch-only frame requires recabling, typically taking several days to accomplish. This is a complex scenario that requires detailed planning.
- 6 to 8 switches

  Once configured with a switch-only frame, additional frames can be added without cabling changes to other frames. With careful planning, system outages can be reduced, although installation tests do require checking the entire switch network.

What network topology topics do I need to consider?

Whenever you modify your system by adding additional hardware, your network topology is affected. This section discusses networking topics you should consider prior to adding any hardware to your system:

- Name server

  Every node in your system has a name assigned to it which is resolved by the name server you are using. The name server translates the symbolic name assigned to a node into an Internet address. As you contemplate adding nodes to your system, plan ahead for the names you will assign to the nodes and how your name server will resolve them.
- Available addresses

Nodes have Internet addresses assigned to them, as well as names. While you are planning to add nodes to your system, you need to also plan for the additional Internet addresses that will be assigned to these nodes. In addition, while planning the addresses for these network interfaces, you might reserve additional addresses for expansion the next time your system grows.

If you are using a netmask that limits the number of addresses you can have, you can change your netmask to free up addresses, or you can elect to use a different subnet.

What control workstation topics do I need to consider?

You need to be certain that the control workstation is capable of supporting whatever you plan to add. It must have sufficient processor speed, DASD, and other hardware, such as serial ports and Ethernet adapters. You might need to replace the control workstation. See Chapter 2, “Question 10: What do you need for your control workstation?” on page 70.

What system partitioning topics should I consider?

You can consider system partitioning in an SP or clustered server system that uses the SP Switch, or in a switchless SP system. Migration install enhancements do not require the system to be partitioned, but there are many situations when partitions might be advantageous, including the following:

- Testing new levels of software or equipment in isolation.
- Grouping common resources together for critical production workloads. Isolation might be necessary, all or part of the time, for; security, separation of workloads, reduced performance interference between workloads, and to allow for more orderly migration.
- Handling changes in total system workloads, particularly when large parallel jobs are being run.
- Introducing major new applications.

The simplest planning guideline with regard to partitioning is to group nodes together in a common frame if they are to belong to the same partition. Even with the *System Partitioning Aid* (see Chapter 7, “Planning SP system partitions” on page 171) there are some restrictions on subdividing switches. Using frames for system partition boundaries makes adding expansion frames easier, and keeps the system more available.

What expansion frame topics should I consider?

In some configurations, a frame can exist that contains nodes and a switch, but the nodes do not use all of the node switch ports. For example, a frame filled with eight wide nodes only uses eight node switch ports, leaving eight ports free. You can add one or more non-switched expansion frames immediately after such a frame to allow the nodes in the non-switched expansion frames to take advantage of these unused switch ports. In the Sample SP Switch System, Frame 2 is a non-switched expansion frame and frame number 3 has been reserved for the addition of a second non-switched expansion frame to share Frame 1’s switch.

Similarly, if a frame having a switch is filled with four high nodes, only 4 node switch ports are occupied, leaving 12 unused. Up to 3 non-switched expansion frames can be inserted to make use of these 12 ports. For example, a single frame might be inserted containing any of 4 wide nodes, 4 high nodes, or some mixture of wide and high node types.

Note that the non-switched expansion frame's number is dependent upon the frame to which it is attached. If the frame containing a switch is number 1, the first associated non-switched expansion frame must be numbered 2, the second 3, and the third 4. Therefore, if you foresee adding non-switched expansion frames to your system in the future, number your frames to allow for the insertion of non-switched expansion frames. Otherwise, the frames which immediately follow must be completely reconfigured.

If your system is organized for partitioning, you might want to leave unused slots for additional nodes, adding an extra frame if necessary; or by leaving gaps in the frame numbers to allow specific frame additions. This is particularly useful if the partition needs a mix of thin, wide, and high nodes.

Again, plan ahead for growth when you assign network addresses. This is easier to manage if you have reserved space for growth in your frame and partition layout.

What boot-install server topics should I consider?

Generally, you might want to have one boot-install server for every sixteen nodes.

Scenario 1: Expanding the sample SP Switch system by adding a node

Note in the Sample SP Switch System that slots 5, 6, 7 and 8 of Frame 2 are empty. You can install a wide or high Node 21 at slot 5 of Frame 2. Further, if proper cabling has been used, only Nodes 1, 17 and 5 are connected to the switch chip to which Node 21 would normally connect; that is, Node 21's normal switch port on Switch 1 is unused. (See Chapter 7, "Planning SP system partitions" on page 171 for more information on node switch port assignment.) So, you can indeed physically install Node 21 in this system as if it were there originally.

To install this new node in the system, and start it running on the switch:

1. Physically install the new node, including cabling to the switch.
2. Enter the new node's network data into the SDR.
3. Install the software on the node using a mksysb image.
4. Perform post-install customization.
   - Add required PTFs
   - Adjust file systems
   - Configure applications
   - Perform installation tests.
5. Bring the new node up on the switch by using the **Eunfence** command.
6. Perform switch installation test.

Scenario 2: Expanding the sample SP Switch system by adding a frame

Before addressing specific examples for the Sample SP Switch System, review the possibilities for frame expansion, and general concerns.

Frame expansion possibilities

When you add a frame to your system, you can add the frame at the end of your system, between two existing frames, or even at the beginning. A special case of the first two possibilities is a non-switched expansion frame.

Non-switched expansion frames

In some configurations, a frame might exist that contains nodes and a switch but the nodes do not use up all the switch ports. One or more non-switched expansion frames can be added, immediately following this frame, whose nodes will share the preceding frame's switch.

Adding a frame at the end of the system

If you have not planned ahead for other expansion, IBM suggests that you add frames only to the end of your system. Otherwise you will have to reconfigure the SDR, and perhaps have to move nodes to accommodate your needs.

Adding a frame in between two existing frames

This is fairly straight forward if the frame number was reserved. This is true whether the new frame is a switched or a non-switched expansion frame. However, if the frame number was not reserved, there can be much work to do. The new frame splits the old system into 2 pieces, and the second piece (the higher numbered frames) must be redefined to the system. Further, for a switched system, some amount of recabling will be necessary, prior to the cabling of the new frame to the existing system.

Adding a frame to the beginning of a system

If your system has a switch, the first frame in the system must have a switch. Therefore, if you plan on inserting the additional frame in the first position in your system, that frame must contain a switch.

If your system does not have a switch, you can insert the additional frame in the first position without any such restriction.

Beyond this item, this case has some of the same overhead as the previous case: the entire old system is the ″second piece″.

General concerns for adding a frame

The following are topics you need to consider when adding a new frame to your system.

1. Control workstation

   When adding a frame to a system, you need to ensure that the control workstation has enough spare serial ports to support the additional frames. One serial port is required for each additional frame, two for an SP-attached server. If you do not have enough ports, you need to upgrade the control workstation.

   If you use HACWS, there are two control workstations to consider.

2. Frames in the existing configuration

   You need to consider the existing configuration and plan your new configuration by following the rules explained in “Understanding placement and numbering” on page 118. Be careful to not interrupt a switch capsule.

3. Types of nodes in the existing configuration

   You need to consider what types of nodes you already have and what types you will be adding in the additional frame. For example, consider how thin, wide, and high nodes work together.

4. Switch

   You need to consider the implications involved if your system has a switch.

   - If you currently have one switch and are adding a second switch, you need to add the cables for the second switch. During this time, the first switch might be able to run until the new frame is ready for installation tests.

- If you currently have two switches and add a third switch or you have three switches and add a fourth switch, you need to add and, perhaps, reroute cables. This scenario can cause a significant amount of system down time. For highest availability, consider installing more frames with empty slots for future node additions.

5. SP Ethernet Network

   You need to consider the Ethernet network being used. Ask yourself whether you want to separate the Ethernet into multiple subnets. For example, do you want to have one network per frame with one boot-install server per frame or do you want to boot all of the frames from the control workstation?

   Also, consider the bandwidth of the default thin wire Ethernet. This Ethernet can load approximately 8 nodes at a time. With larger systems, there are higher technology Ethernets available that can allow you to load software at a faster rate than with the thin wire Ethernet.

6. IP Addresses

   Your decision for the previous concern will play a role in planning for IP addresses. You need to ensure that the nodes that will occupy the additional frame will have IP addresses. If you are using a netmask that limits the number of addresses you can have, you can either modify your netmask to free up addresses or you can use a different subnet.

7. System Partitioning

   If you have a partitioned switched system, and the new frame is a non-switched expansion frame, you might not need to re-partition, because partitioning for a switched system assumes the maximum number of nodes are present; so the nodes in the non-switched expansion frame are already accommodated. However, at this point you might decide you do not like where the new nodes have implicitly resided, in which case you must re-partition.

   If the new frame has its own switch, then you are increasing the number of switches in the system. If your system is partitioned, in this case you will need to re-partition the system because partitioning had not previously accounted for these new nodes.

   If you have a partitioned switchless system, you must re-partition, because partitioning in this case is based on the number of nodes actually installed.

Scenario 2-A: Adding a non-switched expansion frame to the sample SP Switch system

Note: See “Node placement with the SP Switch” on page 120, particularly Figure 15 on page 122, for the specifics on valid node placement, and Chapter 7, “Planning SP system partitions” on page 171 for more information on assignment of nodes to switch ports.

Consider Frame 1 of the Sample SP Switch System. It has only 5 nodes so 11 node switch ports are available for other nodes to use. Given Frame 1’s configuration, 8 ports are actually set aside for Frame 1 so 8 ports are available for expansion frames. Frame 2 uses only 3, but reserves at least 4. Specifically, Frame 2’s nodes are located such that a second expansion frame of 4 nodes is valid. You can insert a Frame 3 with up to 4 nodes and cable all these nodes to Frame 1’s switch.

Therefore, you need to accomplish the following:

1. Install the new hardware, and attach the new frame to the control workstation via a 232 port.
2. Cable the new nodes to the Frame 1 switch.

3. Run the **spframe** command to establish the SDR entries for the new nodes.
4. Enter the new nodes' network data into the SDR.
5. Install the software on the new nodes using a mksysb image.
6. Perform post-install customization.
   - Add required PTFs
   - Adjust file systems
   - Configure applications
7. Bring the new nodes up on the switch by using the **Eunfence** command.
8. Perform installation tests.

Scenario 2-B: Adding a frame at the end of the sample SP Switch system

For the Sample SP Switch System, you could add a Frame 5 which is a non-switched expansion frame for Frame 4. Frame 4's switch has several unused ports and Frame 4 has only 4 nodes located such that Frame 4 can be expanded by as many as 3 frames. So, Frame 5 would be the first of these expansion frames. This expansion would be done like that in Scenario 2-A.

Alternatively, you might want to add a frame after Frame 4 which has its own switch. Given the preceding discussion, you might want to designate the new switched expansion frame as Frame 8 (or 6 or 7). You should do this to reserve space for non-switched expansion frames to come later. This case is more complicated, because you are adding a new switch, thereby changing an important part of the system. The following modifications must be made to the 2-A list:

- In addition to cabling the new nodes to the new switch, the new switch must be cabled to the existing switches.
- After adjusting the file systems in post-install customization, you must select a new switch configuration to indicate the new switch structure.
- Before bringing up the switch, use the **Eclock** command to get the system switches synchronized.
- Bringing the new nodes up on the switch requires use of the **Estart** command, at least on any partition containing new nodes.

Scenario 2-C: Adding a frame in between two existing frames

Suppose you wanted to insert a frame between Frames 1 and 2, where this new frame will also be a non-switched expansion frame to Frame 1. To accomplish this expansion, first delete Frame 2 from the system, then add Frame 2 (the new frame) and Frame 3 (the previous Frame 2) to the system. Note that the old Frame 2 nodes will be rebuilt as Frame 3 nodes. You must:

1. Save mksysb images of the original Frame 2 nodes; one image per unique node.
2. Use the **spdelfram** command to remove Frame 2 configuration data from the SDR.
3. Add the new Frame 2 as in Scenario 2-A above.
4. Add the new Frame 3 as in Scenario 2-A, using the newly saved mksysb images as appropriate.

Scenario 3: Expanding the sample SP Switch system by adding a switch

Before going through the scenario, review the list of topics to consider when planning to add a switch. See "The physical makeup of a switch board" on page 175 to understand how a switch works.

1. Switch type

What type of switch will you be adding? The table below describes the types available. You cannot add an SP Switch to expand an SP system that already has an SP Switch-8. You must convert the SP Switch-8 to an SP Switch.

If you are adding any of the switches in the table, an IBM Customer Engineer installs the switch hardware on your system.

2. Frame support

   Prior to adding the switch, you need to consider which frames the switch will support and record your information on the Switch Configuration Worksheet.

The switch scenario

The Sample SP Switch System has 3 frames, but only 2 switches. Frame 2 has no switch since it is a non-switched expansion frame using Frame 1's switch. Suppose you choose to give Frame 2 its own switch – apparently a preliminary step to further changes. So, Frame 2 will no longer be a non-switched expansion frame. To synchronize the switches, do the following:

1. Quiesce switch traffic.
2. Install the new switch in Frame 2.
3. Re-cable the nodes of Frame 2 to the new switch.
4. Cable the new switch (now Switch 2) in Frame 2 to the switches in Frames 1 and 4, and re-cable the switch in Frame 4 (now Switch 3) to the switch in Frame 1.
5. Choose a new switch configuration which matches the expanded system.
6. Use **Eclock** to synchronize the switches.
7. Set the nodes of Frame 2 to the ″customize″ boot status. Then reboot the Frame 2 nodes, or run **pssp_script**, to get the these nodes recustomized for their new switch.
8. Use the **Estart** command, once for each system partition, to bring up the new switch fabric.
9. Perform install tests to assure the new hardware and connections perform correctly.

Chapter 11. Planning for migration

This chapter includes factors to consider when planning to migrate the software on an existing SP system. Migration addresses upgrading the software from supported levels of PSSP and AIX to PSSP 3.5 and AIX 5L 5.1. See other chapters in this book for information pertaining to reconfiguring or expanding an existing SP system or planning for new system installations.

Migrating an SP to newer software levels is a relatively complex task, but the complexities and risks can be minimized by thoroughly planning each migration phase before beginning the migration.

The book *PSSP: Installation and Migration Guide* describes the specific steps to be completed in implementing a software migration. Other books that might be helpful during your planning phase include:

- Other PSSP books (such as *PSSP: Administration Guide*, *PSSP: Managing Shared Disks*).
- The *AIX: Installation Guide* for each version of AIX you plan to run. For PSSP 3.5, you must install AIX 5L 5.1 on the control workstation. You can also load an AIX 4.3.3 install image on the control workstation in order to use it as a boot-install server to install AIX 4.3.3 on some nodes.
- The ITSO Redbooks:
  - *AIX 5L Differences Guide Version 5.1*
  - *AIX Differences Guide Version 4.3*
  - *AIX 5L Porting Guide*
  - *AIX 4.3 Migration Guide*
- Books for the IBM licensed programs and other products you might be using.

The underlying migration support provided in PSSP has not changed. The base support for the mechanics of performing a migration also have not changed. However, there might be some new considerations for you that arise from varying levels of PSSP, AIX, PSSP-related licensed programs, and hardware that you might already have or that you plan to install on your system. For instance, a pSeries 690 or p670 node requires PSSP 3.4 or later with AIX 5L 5.1. There are more dependencies so be sure to carefully consider all the software requirements relative to your hardware and to read information throughout this chapter pertaining to the PSSP components and PSSP-related licensed programs that you use, particularly in “Migration and coexistence limitations” on page 220.

This chapter discusses the principle migration planning phases:

- “Developing your migration goals” on page 214:

  Briefly discusses considerations such as what software is supported at various migration checkpoints and how to plan your SP system configuration in preparation for migration.
- “Developing your migration strategy” on page 218:

  Briefly discusses system requirements and migration options you need to consider while planning your migration goals and the steps you need to complete to achieve those goals. Understanding coexistence and the advantages and disadvantages of system partitioning will help you refine your migration strategy.
- “Reviewing your migration steps” on page 237:

Summarizes the migration steps, providing a transition to the detailed migration instructions in the book *PSSP: Installation and Migration Guide*.

Developing your migration goals

Before you begin planning the actual system migration steps, you must understand your current system configuration and the system requirements that led you to that configuration. Also, before planning begins, review earlier system plans for goals that have not yet been met. Assessing the priority of the goals or why they were not met can influence how you will conduct the current system migration.

Similarly, while the configuration worksheets in this book are generally not required for performing a software migration, there can be merit in reviewing your previous set, and possibly reviewing or completing the current worksheets. For example, this might be appropriate when evaluating the use of system partitions or coexistence in your current systems or as part of your planned migration strategy, or in determining any changes to your boot-install server configuration.

The underlying task in planning your migration is to determine where you want to be and what staging will allow you to ultimately reach that goal. There are general factors that drive the requirement for migrating to new software levels, including both advantages (such as, new function, performance) and possible impacts or disadvantages (such as, production down time, stability). The fact that you are planning a migration implies that these factors have already been considered.

Another factor that will influence your migration plans involves the dependencies and limitations that exist between applications. For example, if you plan to run the General Parallel File System licensed program, you must also run either PSSP with the IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk optional components installed or you must run with the HACMP/ES licensed program. Besides co-requisite software limitations, other limitations might involve operating systems, system software, and applications which might operate in your current system environment but not in the migrated environment.

Understanding coexistence support and possibly having SP system partitions can help you improve efficiency on your SP system. However, you must fully assess your system so that you will have all of the information that you need to plan the steps of your migration.

A full migration plan involves breaking your migration tasks down into distinct, verifiable, and recoverable steps, and planning the requirements for each migration step. A well-planned migration has the added benefit of minimizing system downtime.

The software requirements, weighed against your SP system workload, generally drive three key components of your migration goals:

1. “Planning base software requirements”
2. “Planning how many nodes to migrate” on page 216
3. “Planning migration stages” on page 217

Planning base software requirements

The topics addressed in this section are:

- “Supported migration paths” on page 215
- “Supported software levels” on page 215

Supported migration paths

A direct migration path for the control workstation is supported from PSSP 3.4 with AIX 5L 5.1 to PSSP 3.5 with AIX 5L 5.1. A control workstation with earlier releases than PSSP 3.2 must first be migrated to PSSP 3.4 and AIX 4.3.3, then AIX 4.3.3 to AIX 5L 5.1, then PSSP 3.4 to PSSP 3.5. PSSP 3.2 can be migrated directly to PSSP 3.5 if it is done in one service window by following the standard migration steps:

- Quiescing the system
- Migrating AIX from 4.3.3 to 5L 5.1
- Migrating to PSSP

Table 49 lists the migration paths that are available for nodes to PSSP 3.5.

Table 49. Migration paths for nodes

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

If your system contains a control workstation or node that is currently at a PSSP and AIX level not listed in the **From** column, you need to migrate it to one of the listed combinations before you can use one of the listed migration paths. Afterward nodes can migrate to PSSP 3.5 and AIX 5L 5.1 in one step. How to actually migrate is documented in the book *PSSP: Installation and Migration Guide*. You need to be prepared before you get started.

Some optional components of PSSP and PSSP-related licensed programs have dependencies on certain levels of other components or programs. Be sure to read “Migration and coexistence limitations” on page 220 in this chapter.

Supported software levels

PSSP 3.5 is supported on AIX 5L 5.1 with a 64-bit kernel and a 32-bit kernel. AIX 5L 5.1 enables you to install both kernels and switch from operating with one to operating with the other. PSSP 3.5 supports that ability. Evaluate your installation’s current operational requirements to understand your software and hardware requirements before you migrate to PSSP 3.5.

In addition to the operational requirements placed on your system software, some IBM licensed programs also have PSSP release level dependencies. The following table summarizes those dependencies.

Table 50. Supported IBM licensed programs per supported PSSP and AIX release

PSSP and AIX	IBM licensed programs
PSSP 3.5 (5765-D51) and AIX 5L 5.1 (5765-E61)	• General Parallel File System 2.1 (5765-B95) • HACMP 4.5 (5765-E54) • HAGEO/GeoRM 2.3 (5765-E82) • LoadLeveler 3.1 (5765-E69) • Parallel Environment 3.2 (5765-543) • Engineering and Scientific Subroutine Library (ESSL) 3.3 (5765-C42) • Parallel ESSL 2.3 (5765-C41)

Table 50. Supported IBM licensed programs per supported PSSP and AIX release (continued)

PSSP and AIX	IBM licensed programs
PSSP 3.4 (5765-D51) and AIX 5L 5.1 (5765-E61)	• General Parallel File System 1.5 (5765-B95) • HACMP 4.4.1 (5765-E54) • HAGEO/GeoRM 2.3 (5765-E82) • LoadLeveler 3.1 (5765-E69) • Parallel Environment 3.2 (5765-543) • ESSL 3.3 (5765-C42) • Parallel ESSL 2.3 (5765-C41)
PSSP 3.4 (5765-D51) and AIX 4.3.3 (5765-C34)	• General Parallel File System 1.5 (5765-B95) • HACMP 4.4.1 (5765-E54) • HAGEO/GeoRM 2.3 (5765-E82) • LoadLeveler 2.2 (5765-E69) • Parallel Environment 3.1 (5765-543) • ESSL 3.2 (5765-C42) • Parallel ESSL 2.2 (5765-C41)
PSSP 3.2 (5765-D51) and AIX 4.3.3 (5765-C34)	• General Parallel File System 1.4 (5765-B95) • HACMP 4.4 (5765-E54) • HACMP with HAGEO/GeoRM 2.2 (5765-E64) • LoadLeveler 2.2 (5765-D61) • Parallel Environment 3.1 (5765-543) • ESSL 3.2 (5765-C42) • Parallel ESSL 2.2 (5765-C41)

See the “Bibliography” on page 313 for other IBM documentation with information on AIX requirements for other licensed programs in the IBM RS/6000 software catalog.

Planning how many nodes to migrate

Subject to your requirements, you might migrate your entire SP system or just part of it. Migration addresses upgrading the software on an existing SP system to AIX 5L 5.1 and PSSP 3.5 from earlier supported levels of AIX, PSSP, and PSSP-related licensed programs. IBM offers flexibility with features that help when migrating your system – coexistence and system partitioning:

- Coexistence refers to support within each licensed program that allows for mixed levels of PSSP and AIX in the same SP system partition. Coexistence is independent of system partitioning. Coexistence is an important factor in the ability to migrate one node at a time and, as such, is a key feature of migration.
- System partitioning is a mechanism for dividing an SP system that uses the SP Switch into logical systems. The definition of these logical systems is a function of the switch chip which results in the system partitions being isolated across the switch.

Consider coexistence and system partitioning while evaluating your system requirements. Think about what applications you need to run and what levels of PSSP and AIX are needed to support those applications. Then, factoring in your current SP configuration, determine how many nodes you will need to run each type of workload. Important considerations and other relevant information on these two features is provided in “Developing your migration strategy” on page 218.

Note: Before migrating any nodes, the control workstation must be migrated to the highest PSSP and AIX levels you plan to run on any one of the nodes.

Planning migration stages

Some migrations have service prerequisites of program temporary fixes (PTFs) that need to be applied to your system. See the *Read This First* document for specific information. These services can be applied well in advance and they must be done before migrating to PSSP 3.5.

When possible, plan your migration in multiple stages, breaking them down into distinct steps that can be easily defined, executed, and verified. Plan a reasonable amount of time to complete each step, define validation steps and periods, and be prepared for recovery or to back out should a step not go as planned. Proper migration staging can better ensure an effective and successful migration, while minimizing system down time. You can also distribute system down time over a longer period by migrating a few nodes at a time, subject to your needs.

There are three main high-level suggestions for doing this:

1. Migrate the control workstation then validate the system.
2. Migrate a subset of the nodes then validate the system.
3. Migrate and validate the remainder of your system according to your plan.

AIX 5L 5.1 does not preserve binary compatibility for applications because it introduced a new 64-bit ABI. Because of this binary incompatibility, if you want to migrate to AIX 5L 5.1, you must migrate the control workstation first to PSSP 3.4 and AIX 4.3.3 if it does not already have that level. Then migrate to AIX 5L 5.1 and PSSP 3.5. Alternately, you can migrate your control workstation from PSSP 3.2 and AIX 4.3.3 directly to PSSP 3.5 and AIX 5L 5.1 if both AIX and PSSP migrations are done in a single service window.

The nodes can subsequently be migrated from any supported base release to PSSP 3.5 on AIX 5L 5.1 with a single **nodecond** operation. To migrate nodes with earlier versions like PSSP 2.4 on AIX 4.2.1, first migrate to PSSP 3.4 on AIX 4.3.3. For example, to minimize the amount of change to your control workstation at one time and also minimize your service window, you might want to do the following:

1. Upgrade earlier levels to PSSP 3.4 and AIX 4.3.3 (node by node is possible), leaving other PSSP-related licensed programs at the current level.
2. Upgrade HACMP and GPFS to the newest level. These can be done independently.
3. Upgrade AIX 4.3.3 to AIX 5L 5.1 (node by node is possible) and upgrade LoadLeveler and Parallel Environment to the new level. Both upgrades are necessary if you want the new support for parallel jobs.
4. Upgrade other licensed programs.

Table 51 on page 218 might be helpful. The programs shown in bold letters are migrated within the migration stage or have already been migrated in an earlier stage. Those not bold remain at the current level within the stage. After stage 3 is complete, all the programs are at the latest release levels supported with PSSP 3.5.

Table 51. Suggested migration stages

From level	Migration stage 1	Migration stage 2	Migration stage 3	Migration stage 4
AIX 4.3.3	AIX 4.3.3	AIX 4.3.3	***AIX 5.1***	***AIX 5.1***
PSSP 3.2	***PSSP 3.4***	***PSSP 3.4***	***PSSP 3.4***	***PSSP 3.5***
GPFS 1.3 or 1.4	GPFS 1.3 or 1.4	***GPFS 1.5***	***GPFS 1.5***	***GPFS 1.6***
HACMP 4.4	HACMP 4.4	***HACMP 4.4.1***	***HACMP 4.4.1***	***HACMP 4.5***
LL 2.2	LL 2.2	LL 2.2	***LL 3.1***	***LL 3.1***
PE 3.1	PE 3.1	PE 3.1	***PE 3.2***	***PE 3.2***
ESSL 3.2	ESSL 3.2	ESSL 3.2	***ESSL 3.3***	***ESSL 3.3***
PESSL 2.2	PESSL 2.2	PESSL 2.2	***PESSL 2.3***	***PESSL 2.3***

Developing your migration strategy

The intent of this stage of your migration planning activity is to focus primarily on the scope of your migration in terms of deciding on which and how many nodes to migrate which licensed programs and to which levels. You should be entering this planning stage with a clear understanding of your migration goals.

If you are migrating an entire SP system or existing system partition, you might be able to skip the remainder of this section. If, on the other hand, you are interested in migrating a subset of your system, resulting in a system running mixed levels of PSSP, the information discussed further in this section might be crucial.

SP system partitioning and coexistence are capabilities that offer flexibility in the number of nodes that you need to migrate at any one time. Your operational and performance goals might necessitate the use of multiple system partitions, coexistence, a combination of the two, or neither. Understanding the advantages and disadvantages of system partitioning and coexistence will help you assess their suitability for your needs.

Other factors that will influence your migration strategy include:

- “Boot-install servers and other resources” on page 219
- “Root volume group mirroring” on page 220
- “Migration and coexistence limitations” on page 220
- “IP performance tuning” on page 234
- “Changes in recent levels of PSSP” on page 235
- “AIX and PSSP migration options” on page 236

Each of those factors is discussed later in this section.

Using system partitions for migration

The SP system supports multiple system partitions, except in SP Switch2 system configurations. A Cluster 1600 system managed by PSSP with no SP node frame or SP Switch does not support SP system partitions. SP system partitions effectively subdivide an SP system into logical systems. These logical systems have two primary features:

1. SP Switch traffic in a system partition is isolated to nodes within that system partition.
2. Multiple system partitions can run different levels of AIX and SP software.

These features facilitate migrating nodes in relative isolation from the rest of the system. Using these features, you can define a system test partition for newly migrated nodes. After the migration is complete and you have validated system performance, the nodes can be returned to production.

You have the ability to use system partitioning for migration due to the fact that the SP Switch traffic in a system partition is isolated to nodes within that system partition. This factor stems from the SP Switch architecture, in which the switch chip connects nodes in a specific sequence. The switch chip therefore becomes the basic building block for a system partition and establishes a minimum partition size that depends on the partition's node types. It is this partition size that sets the granularity with which an SP system can be upgraded to new software levels. Coexistence, described in the next section, can provide even finer granularity within a system partition.

See Chapter 7, "Planning SP system partitions" on page 171 for additional information on the use of SP system partitions.

Using coexistence for migration

In traditional SP system partitions, all nodes within a single system partition generally run the same levels of operating system and system support software. However, different partitions can run different levels of operating system and system support software. Therefore multiple release levels of licensed programs like Parallel Environment can run on an SP without restriction as long as each different release level is within a separate SP system partition.

For many installations with the need to migrate only a few nodes, for switchless clustered servers systems, or for SP Switch2 system configurations, the system partition approach is not viable. This is true in a small system (in terms of number of nodes), or a system with a migration requirement that includes migrating less nodes than can be represented by a system partition, possibly only one node, such as for LAN consolidation. It might be that you do not want the switch isolation function. Coexistence is aimed specifically at providing additional flexibility for migration scenarios where system partitioning is not available or not wanted.

Coexistence support is provided for multiple levels of PSSP and coordinating levels of AIX in the same system partition. However, there are requirements and certain limitations that must be understood and adhered to in considering the use of coexistence. Some of the PSSP-related licensed programs are not supported or are restricted in a mixed system partition. For example, parallel processing licensed programs like Parallel Environment are generally not supported in mixed system partitions. Inter-node communication over the SP Switch using TCP/IP is supported, but user space communication is not available in a coexistence configuration. The supported coexistence configurations and the limitations that apply to these coexistence configurations are described in the remainder of this section.

Boot-install servers and other resources

Your boot-install server must be at the highest level of AIX and PSSP that it is to serve.

Your migration plan must also consider additional resources; for example, additional DASD to support multiple levels of software, particularly if you plan to use coexistence. In that case, plan on allocating 2 GB of disk for each level of AIX and PSSP being served by your control workstation or boot-install server. The space is used for additional directories, specifically:

- multiple AIX mksysb subdirectories under: **/spdata/sys1/install/images/...**
- multiple AIX subdirectories under: **/spdata/sys1/install/default/...** or **/spdata/sys1/install/***customized_name***/...**

  which include: **lppsource/** and **spot/**
- multiple PSSP subdirectories under: **/spdata/sys1/install/pssplpp/...**

Root volume group mirroring

If you already have root volume group mirroring on SP nodes or on servers that you now want to attach to the SP system, enter the mirroring information into the SDR ***before*** either migrating a node to PSSP 3.5 or attaching a server. Failure to enter existing root volume group mirroring information will result in the root volume group being ***unmirrored*** during migration to PSSP 3.5 or attaching a server to a PSSP 3.5 system.

Switching between volume groups on a node requires that you run the **syspar_ctrl -r** command in the system partition where the node is located.

There are SMIT panels and commands for mirroring, unmirroring, and modifying the Volume_group objects. You can use the **spchvgobj** command to change attributes like install_image, lppsource_name, code_version, install_disk, and boot_server. You can use the **spbootins** command to set the bootp_response attribute and the current volume group of a node.

Migration and coexistence limitations

PSSP 3.5 is supported on AIX 5L 5.1. The PSSP software has some 64-bit support features. See “Support for applications with 64-bit addressing” on page 235.

PSSP 3.5 supports multiple levels of AIX and PSSP in the same system partition. Keep in mind that an unpartitioned system is actually a single default system partition. However, only certain combinations of PSSP and AIX are supported to coexist in a system partition. Some licensed programs state that multiple levels can coexist but not interoperate. When coexistence does not include interoperability, it is explicitly stated where applicable in the subsections that follow.

Coexistence is supported in the same system partition or a single default system partition (the entire Cluster 1600 system managed by PSSP). A mixed system partition consists of the control workstation running PSSP 3.5 and AIX 5L 5.1 and the nodes running any combination of:

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

Table 52 lists the levels of PSSP and corresponding levels of AIX that are supported in a mixed system partition.

Table 52. Levels of PSSP and AIX supported in a mixed system partition

	AIX 4.3.3	AIX 5L 5.1
PSSP 3.5		S
PSSP 3.4	S	S
PSSP 3.2	S	

Generally, any combination of the PSSP levels listed in Table 52 on page 220 can coexist in a system partition and you can migrate to a new level of PSSP or AIX one node at a time. However, some PSSP components, related licensed programs, and hardware impose limitations. Also, some software program features have PSSP and AIX level dependencies – you must ensure that the proper release levels of these licensed programs are used on nodes running the coordinating supported PSSP and AIX levels.

For example:

- To have a system with a two-plane SP Switch2 configuration, all nodes connected to the switch must run PSSP 3.4 or later.
- A coexistence combination that is ***not supported*** is GPFS 2.1 and HACMP 4.4.1.

The PSSP components and related licensed programs discussed in this section have notable exceptions that might restrict your ability to migrate one node at a time or might limit your coexistence options.

Note: Before migrating any node, the control workstation must be migrated to the highest level of PSSP and AIX that you intend to have on your system. Program Temporary Fixes (PTFs) might be necessary. See the PSSP 3.5 *Read This First* document for the latest requirements.

Special considerations apply to the following:

Communications Low-Level Application Programming Interface

The Communications Low-Level Application Programming Interface (LAPI) provides a more primitive interface to the SP Switch and SP Switch2 than either MPI or IP. Parallel Environment is required for using LAPI. The LAPI 64-bit support is available only on AIX 5L 5.1.

The LAPI libraries are also shipped with PE 3.2 for use with standalone RS/6000 or pSeries servers without an SP or PSSP installed. When both PSSP and POE are installed, the PSSP LAPI libraries take precedence. This adds support for LAPI shared memory, with ssp.css not needed.

Coexistence: LAPI cannot run in a mixed environment of PSSP levels.

Migration: User space LAPI will not interfere with node by node migration with the currently supported levels of PSSP. However, applications that use pre-PSSP 3.2

LAPI cannot migrate one node at a time. LAPI users need to zero any unused fields in the `lapi_info_t` structure and recompile the programs to use LAPI in PSSP 3.2 or later. Migrate such applications on all nodes in the same service window.

Communications Kernel Low-Level Application Programming Interface

Communications Kernel Low-Level Application Programming Interface (KLAPI) provides LAPI plus a copy avoidance communication interface to other kernel subsystems like the IBM Virtual Shared Disk component of PSSP.

Coexistence: KLAPI cannot run in a mixed environment of PSSP levels.

Migration: See "IBM Virtual Shared Disk" on page 226.

Switch management and TCP/IP over the switch

The switch support component of PSSP (CSS), provides for switch management and TCP/IP over the switch between nodes in a mixed partition. Certain hardware, new nodes and adapters, might require a specific level of PSSP and related licensed programs. Be very careful to coordinate both your current and new hardware requirements with the software requirements. With PSSP 3.5 or 3.4, it is possible to configure an SP Switch2 system with a mix of nodes on the switch, and some nodes not on the switch.

Coexistence statement: For IP communication over the switch:

- Nodes attached to an SP switch can coexist in a mixed system partition running any supported combination of PSSP.
- Different switch types, like the SP Switch2 and SP Switch, **cannot** coexist in the same SP or clustered servers system.
- In an SP system with an SP Switch2, all nodes exist in the same, singular default system partition.
- To have a system with a two-plane SP Switch2 configuration, all nodes connected to the switch must run PSSP 3.4 or later. Nodes that do not support two planes can be in the system, but not connected to the switch.
- In an SP system without an SP switch, not a clustered servers configuration, the system can be partitioned.
- PSSP 3.5 CSS supports 64-bit applications communicating over the switch. An application on a node running a 64-bit kernel could communicate with the application on a node running a 32-bit kernel, depending on the application. CSS supports the attachment to the switch of nodes running both 32-bit and 64-bit kernels. It is up to the applications running on the nodes to decide how they support a mix of 32-bit and 64-bit nodes.

Note: If you are using Parallel Environment, see "Parallel Environment" on page 232 for more limitations that might apply.

The SP Switch Router is supported in SP systems with the SP Switch as an extension node, not in switchless or SP Switch2 systems. Extension node support will function in a mixed system partition of nodes running any combination of the supported PSSP levels.

Migration statement: Switch support software does not interfere with migrating one node at a time. Converting from one type of switch to another is a hardware configuration change, not a migration change, but consider the sequence of activities:

- Convert a High Performance switch to the SP Switch2 or SP Switch before migrating PSSP.
- Converting from one type of switch hardware to another is not in the scope of this discussion. It is considered a configuration change not a migration. However, be aware at this point that if you plan to convert to the SP Switch2, the system must first be configured as a single SP system partition before the switch is installed and the PSSP software is migrated. The SP Switch2 does not support SP system partitioning.
- To have a two-plane SP Switch2 configuration, migrate all nodes connected to the switch to PSSP 3.4 or later. Nodes that do not support two planes can be in the system, but not connected to the switch.

More of switch management has been automated. If you have switch commands in local scripts and procedures, consider removing them and rely on the automation available in PSSP. On the other hand, if you prefer, you can turn off automatic switch management. You will have to turn it off any time you boot the control workstation.

Security support

You have the option of running PSSP with an enhanced level of security. See “Considering restricted root access” on page 146, “Considering a secure remote command process” on page 151, and “Considering choosing none for AIX remote command authorization” on page 152 for option descriptions and a complete list of limitations.

The AIX remote command suite supports DCE security services. The suite includes **rsh**, **rcp**, **rlogin**, **telnet**, and **ftp**. PSSP uses these enhanced AIX commands. For PSSP migration purposes, the AIX remote commands, **rsh** and **rcp**, were enhanced to call a PSSP-supplied Kerberos V4 set of **rsh** and **rcp** subroutines. The AIX **/usr/bin/rsh** and the **/usr/bin/rcp** with PSSP supports multiple authentication methods, including Kerberos V5, Kerberos V4, and standard AIX authentication.

With PSSP, the **/usr/lpp/ssp/rcmd/bin/rsh** and **/usr/lpp/ssp/rcmd/bin/rcp** commands are symbolic links to the AIX **/usr/bin/rsh** and **/usr/bin/rcp** commands respectively. Be aware of the following with respect to the **rsh** and **rcp** commands:

- The Kerberos V4 authentication method must be configured on the control workstation and in any node running earlier than PSSP 3.2 for the SP system management commands that use the **rsh** and **rcp** commands to function properly. After the control workstation and all the nodes have AIX 4.3.3 and PSSP 3.2 or later, you can choose whether to use DCE (Kerberos V5), Kerberos V4, and standard AIX authentication methods for remote command authentication.
- The authentication methods for SP trusted services can be set for both DCE and compatibility with previous releases of PSSP. If that is configured, the SP trusted services provide both DCE authentication and any service-specific authentication as was supported on pre-PSSP 3.2 nodes.
- Before enabling DCE authentication for SP trusted services within an SP system partition, migrate all the nodes within the partition to PSSP 3.2 or later. If you do not, there will be failures within the SP trusted services.
- If you install DCE and choose it as an authentication method, any applications which do not support DCE will experience authentication error messages from DCE before the system proceeds to try using the next authentication method configured. You can change those applications to support DCE, handle the messages, or take a risk and use the environment variable K5MUTE which suppresses ***all*** messages.

If you decide to use DCE, PSSP 3.5 requires DCE 3.2, which can run with PSSP 3.4 and PSSP 3.2 nodes as well.

DCE and HACWS restriction:
If you plan to have DCE authentication enabled, you cannot use HACWS. If you already use HACWS, do not enable DCE authentication.

Coexistence statement: For coexistence with AIX 5L 5.1 and PSSP 3.5 on the control workstation, the following considerations apply:

- The PSSP 3.5 security services do not affect the ability of the system to interoperate within an existing AFS cell and to use the AFS servers as a Kerberos V4 authentication server.
- A PSSP 3.5 mixed system partition can coexist and interoperate. Nodes can be configured with any of the supported security services, which include DCE, Kerberos V4, and standard AIX authorization. Restricted root access and secure remote commands may be enabled or disabled.
- Before or when you enable a secure remote command process for PSSP 3.5, your secure remote command software must be installed and running on the control workstation, you must have restricted root access enabled, and all the nodes must have PSSP 3.2 or later.
- Before or when you set the authorization methods for AIX remote commands to `none`, all the nodes must have PSSP 3.4 or later.
- The authentication methods used by SP trusted services can be set to indicate that both DCE and compatibility with previous levels of PSSP is required. If this is configured, then all SP trusted services will provide support for the following:
  - DCE authentication
  - any service-specific authentication as was supported on pre-PSSP 3.2 nodes

Migration statement: PSSP 3.5 security does not interfere with a node by node migration of PSSP or AIX levels. The following criteria apply:

- When migrating pre-PSSP 3.2 nodes, continue to configure and use Kerberos V4 until all the nodes have PSSP 3.2 or later. The impacts of the DCE security exploitation project on system migration from releases earlier than PSSP 3.2 are minimal:
  - The authentication methods used by SP trusted services will default to `compat` so that all system services will continue to operate in the same manner as in a pre-PSSP 3.2 system.
  - Although the installation and use of Kerberos V4 is optional in PSSP 3.2 or later, systems migrating from PSSP 3.1 or earlier already have Kerberos V4 installed and enabled for use with the AIX remote commands.
- Before or when you enable a secure remote command process for PSSP 3.5, your secure remote command software must be installed and running on the control workstation, you must have restricted root access enabled, and all of the nodes must have PSSP 3.2 or later. An attempt to enable restricted root access or a secure remote command process is denied if all of the nodes are not at PSSP 3.2 or later.
- Before or when you set the authorization methods for AIX remote commands to `none`, all the nodes must have PSSP 3.4 or later. If not, an attempt to set authorization methods for AIX remote commands to `none` is denied.

- Before disabling `compat` for SP trusted services within an SP system partition, first migrate all SP nodes within that partition to PSSP 3.2 or later. Failure to do that will result in failures within the SP trusted services.
- Kerberos V5 can be configured on the SP even though the SP does not support Kerberos V5. You can have Kerberos V5 installed and configured on an SP system before a migration.

The RS/6000 Cluster Technology components

RS/6000 Cluster Technology (RSCT) is the package of high availability support programs that began as components of PSSP:

- Event Management
- Group Services
- Topology Services

They have been separated from PSSP as a package to be made available with AIX for broader use:

- For AIX 5L 5.1, RSCT and Resource Monitoring and Control (RMC) is part of the package on the base AIX CDs. When you install AIX 5L for PSSP 3.5, be sure to include the file sets:
  rsct.basic.rte
  rsct.basic.sp
  rsct.compat.basic
  rsct.compat.clients
  rsct.msg.*locale*
- For AIX 4.3.3, RSCT is a separate licensed program shipped with the PSSP 3.4 package. That is, you need to include RSCT file sets when you install or migrate to PSSP 3.4.

Coexistence Statement: The RSCT subsystem can coexist in a mixed system partition with any combination of supported levels of PSSP. Both PSSP and HACMP require the RSCT components.

Migration Statement: RSCT does not interfere with the PSSP being able to migrate one node at a time. RSCT also supports node by node migration from the supported levels of PSSP to PSSP 3.5. However, you need to be aware of several considerations.

Event Management: When you monitor hardware or IBM Virtual Shared Disks using the problem management or SP Perspectives component, they are using the Event Management services. PSSP 3.5 and 3.4 support some hardware that was not supported in earlier releases. In order to monitor such hardware after you migrate from an earlier release than PSSP 3.4 to PSSP 3.5 on the control workstation, you need to perform the procedure *Activating the Configuration Data in the SDR*. That procedure is documented in the Event Management subsystem chapter of the book *PSSP: Administration Guide*. The new function is not enabled until all nodes within the SP system partition are running PSSP 3.5 or PSSP 3.4.

Performance Toolbox, Agent Component (perfagent): This function needed by the Event Management component of RSCT comes with AIX 5L 5.1 or AIX 4.3.3 in file set **perfagent.tools**, which you must be sure to install. See “PAIDE (perfagent)”.

PAIDE (perfagent)

PSSP requires the **perfagent.tools** fileset that comes with AIX. The correct level of perfagent needs to be installed on the control workstation and copied to the

lppsource directory so it can be installed on the nodes. Perfagent must be installed to obtain events from the AIX resource monitor. The level of perfagent required is dependent upon the level of AIX.

To verify whether the correct level of perfagent is installed, issue the command:

```
lslpp -l perfagent.tools
```

Table 53 shows the supported levels.

Table 53. Supported PAIDE levels

	AIX 4.3.3	AIX 5L 5.1	PSSP 3.2	PSSP 3.4	PSSP 3.5
perfagent.tools 5.1.x		S		S	S
perfagent.tools 2.2.33.x	S		S	S	

IBM Virtual Shared Disk

IBM Virtual Shared Disk is an optional component of PSSP. If you have IBM Virtual Shared Disks and you choose to migrate to the PSSP 3.5 IBM Virtual Shared Disk component, the IBM Recoverable Virtual Shared Disk component is required as well.

IBM Virtual Shared Disks are supported only on a system with an SP Switch or SP Switch2. In an SP Switch2 system where some nodes are not on the switch, IBM Virtual Shared Disks can work only with those nodes that are on the switch.

Enhanced security options restrictions:
The IBM Virtual Shared Disk component is not supported with restricted root access enabled. See “Considering restricted root access” on page 146, “Considering a secure remote command process” on page 151, and “Considering choosing none for AIX remote command authorization” on page 152 for option descriptions and a complete list of limitations.

Coexistence statement: The IBM Virtual Shared Disk component of PSSP 3.5 can coexist and interoperate in a mixed system partition with any combination of the supported levels of PSSP, but the level of IBM Virtual Shared Disk function available is that of the earliest version of PSSP in the system partition. For example, if you have a mixed system partition with nodes running PSSP 3.5 and PSSP 3.4, the level of IBM Virtual Shared Disk function available is that of PSSP 3.4.

The IBM Virtual Shared Disk component of PSSP only supports the IP protocol when communicating with a PSSP 3.2 or 3.4 node. For example, if a node running the PSSP 3.4 IBM Virtual Shared Disk component is configured to use KLAPI, it transparently uses IP to communicate with a node that has the PSSP 3.5 IBM Virtual Shared Disk component. After the PSSP 3.4 IBM Virtual Shared Disk node is migrated to PSSP 3.5 IBM Virtual Shared Disk, communication over KLAPI resumes.

Coexistence of nodes with the ability to run 32-bit or 64-bit mode is supported only between nodes running PSSP 3.5. Coexistence of PSSP 3.5 IBM Virtual Shared Disk with nodes running earlier versions of PSSP IBM Virtual Shared Disk requires that all nodes run in 32-bit mode.

Migration statement: The IBM Virtual Shared Disk component does not interfere with migrating one node at a time.

Before migrating to PSSP 3.5, install the additional fileset **bos.clvm.enh** when you migrate to AIX 5L 5.1. It is required by the IBM Virtual Shared Disk component of PSSP 3.5, but is not part of the default AIX installation.

IBM Recoverable Virtual Shared Disk

IBM Recoverable Virtual Shared Disk is an optional component of PSSP, but if you choose to install the IBM Virtual Shared Disk component, the IBM Recoverable Virtual Shared Disk component is required. It also has dependencies on the RSCT components of AIX. See "The RS/6000 Cluster Technology components" on page 225.

Enhanced security options restrictions:
IBM Recoverable Virtual Shared Disk is not supported with restricted root access enabled. See "Considering restricted root access" on page 146, "Considering a secure remote command process" on page 151, and "Considering choosing none for AIX remote command authorization" on page 152 for option descriptions and a complete list of limitations.

IBM Recoverable Virtual Shared Disk coexistence: The IBM Recoverable Virtual Shared Disk component of PSSP 3.5 can coexist and interoperate with any combination of the supported levels of PSSP. However, IBM Recoverable Virtual Shared Disk functions at the earliest installed level of PSSP in the system partition.

The **rvsdrestrict** command does not dynamically change IBM Recoverable Virtual Shared Disk run levels across the system. An instance of IBM Recoverable Virtual Shared Disk only reacts to this information after being restarted. If your cluster runs at a given level, and you want to override that level, stop IBM Recoverable Virtual Shared Disk on all nodes, run the **rvsdrestrict** command to change the level, then restart.

Coexistence of nodes with the ability to run 32-bit and 64-bit mode is supported only among nodes running PSSP 3.5. Coexistence of PSSP 3.5 IBM Recoverable Virtual Shared Disk with nodes running earlier versions requires that all those nodes run in 32-bit mode.

IBM Recoverable Virtual Shared Disk migration: The following migration criteria apply:

- In order to exploit the new functions available in the IBM Recoverable Virtual Shared Disk component of PSSP 3.5, you need to migrate all nodes in a system partition using IBM Recoverable Virtual Shared Disk to PSSP 3.5. IBM Recoverable Virtual Shared Disk requires the IBM Virtual Shared Disk optional component of PSSP 3.5 and the RSCT components of AIX 5L 5.1 be running also.
- After the last node is migrated to PSSP 3.5 with the IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk components, you need to reset the IBM Recoverable Virtual Shared Disk subsystem on all nodes in the system partition. That requires stopping and starting applications like Oracle. You can do that in a service window of less than four hours, usually approximately ten minutes. You must use the **rvsdrestrict** command to choose the specific level that IBM Recoverable Virtual Shared Disk is to run in a mixed system partition and there is no need to reinstall that level. If a node has a lower level of the IBM

Recoverable Virtual Shared Disk software installed than what is set with this command, the IBM Recoverable Virtual Shared Disk subsystem will not start on that node.

General Parallel File System (GPFS)

GPFS does not conform to PSSP migration and coexistence practices. Refer to *IBM General Parallel File System for AIX: Concepts, Planning, and Installation Guide* for complete details on GPFS migration.

GPFS is supported in systems with the SP Switch or the SP Switch2. See "Bibliography" on page 313 for the GPFS publications to reference about which adapters are supported for connecting nodes to the SP Switch2 and for all other GPFS-related information.

Enhanced security options restrictions:

GPFS is not supported with restricted root access enabled. See "Considering restricted root access" on page 146, "Considering a secure remote command process" on page 151, and "Considering choosing none for AIX remote command authorization" on page 152 for option descriptions and a complete list of limitations.

GPFS coexistence in an SP environment: GPFS will coexist in a mixed SP system partition under certain circumstances. A file system managed by the GPFS licensed program can only be accessed from within the GPFS nodeset to which it belongs. In order to run multiple levels of GPFS within an SP system partition, you must create multiple GPFS nodesets as described in the GPFS documentation. All nodes within a nodeset must run the same level of GPFS. Nodes do not share access to file systems across nodesets.

GPFS 2.1 will not coexist or interoperate with earlier releases of GPFS in the same nodeset. All applications that run on earlier releases of GPFS can run on GPFS 2.1. All file systems created with earlier releases of GPFS can be used with GPFS 2.1 and can be upgraded to GPFS 2.1 file systems.

Migration statement: Nodes cannot be migrated one node at a time to GPFS 2.1. The following considerations apply:

- All nodes within a GPFS nodeset that use a given file system must have the same level of GPFS and must be migrated to the same new level at the same time. However, the migration can be completed within a 4 hour time frame. All nodes must be rebooted to pick up kernel extensions.
- If any node will have GPFS 2.1, the control workstation must have GPFS 2.1 installed.
- New file system functions existing in GPFS 2.1 are not usable until you explicitly authorize these changes by issuing the **mmchfs -V** command.
- GPFS 1.5 is dependent on the IBM Recoverable Virtual Shared Disk component of PSSP 3.4. Because GPFS 1.4 is only supported on AIX 4.3.3, a migration to GPFS 1.5 must happen before a migration to AIX 5L 5.1.
- In order to use the 64-bit versions of the GPFS programming interfaces, you must recompile your code using the appropriate 64-bit options for your compiler.
- The GPFS kernel extensions are now shipped in both 32-bit and 64-bit formats. GPFS 2.1 supports a mix of nodes running in either 32-bit or 64-bit kernel mode.
- GPFS use of DCE for PSSP security is only supported with the 32-bit AIX kernel and with GPFS running in 32-bit mode.

GPFS levels supported: Table 54 shows the supported levels. See Table 51 on page 218 for possible migration stages.

Table 54. Supported GPFS levels

GPFS	AIX 4.3.3	AIX 5L 5.1	PSSP 3.2	PSSP 3.4	PSSP 3.5
2.1		S			S
1.5	S	S		S	S
1.4	S		S	S	

HACWS

The High Availability Control Workstation (HACWS) optional component of PSSP only runs on the SP control workstations. An HACWS configuration at the PSSP 3.5 level requires the following software on both control workstations:

- PSSP 3.5 including the **ssp.hacws** file set.
- AIX 5L 5.1.

  See the PSSP 3.5 *Read This First* document for the latest information on the required modification levels of AIX.
- Any level of HACMP supported with AIX 5L 5.1 and PSSP 3.5.

  HACMP 4.5 is supported with AIX 5L 5.1 and PSSP 3.5, but AIX and HACMP are often on a different release schedule than PSSP. Over time there might be newer supported releases. See the appropriate HACMP documentation for the latest information on which levels of HACMP are supported with AIX 5L 5.1 and PSSP 3.5.

For more information see the book *PSSP: Installation and Migration Guide*.

DCE and HACWS restriction:

If you plan to have DCE authentication enabled, you cannot use HACWS. If you already use HACWS, do not enable DCE authentication.

Enhanced security options restrictions:

Consider the following:

- HACWS has no problems in the authentication of both control workstations on the nodes. However, you have to copy the authorization files, ***/*.rhosts** or ***/*.klogin**, to the backup control workstation.
- If you plan to enable restricted root access, it is important to do that from the active primary control workstation, since it is the Kerberos V4 Master.
- If you use HACWS you might also use HACMP which has more restrictions.

See “Considering restricted root access” on page 146, “Considering a secure remote command process” on page 151, and “Considering choosing none for AIX remote command authorization” on page 152 for option descriptions and a complete list of limitations.

High Availability Cluster Multi-Processing

IBM’s tool for building UNIX-based mission-critical computing platforms is the High Availability Cluster Multi-Processing (HACMP) licensed program. The HACMP program ensures that critical resources are available for processing. Currently

HACMP is one licensed program with several features, one of which is the Enhanced Scalability feature called HACMP/ES.

Note: Except in statements that are explicitly about HACMP and HACMP/ES separately, all statements about HACMP apply to HACMP/ES as well.

While PSSP has no direct requirement for HACMP, if you already use or are planning to use HACMP and PSSP, there are some cross dependencies. For example, if you have hardware like the p690 or p670 that requires PSSP 3.4 or later, you must also have the correct level of HACMP. See the appropriate HACMP documentation for the latest information on which levels of HACMP you need for the hardware on your system.

If you have existing HACMP clusters, you can migrate to the HACMP/ES feature and re-use all of your existing configuration definitions and customized scripts.

HACMP can be run on all PSSP nodes in any Cluster 1600 system managed by PSSP (SP system) configuration. Do not run HACMP and HACMP/ES on the same node. Typically, HACMP or HACMP/ES is run on the control workstation only if HACWS is being used. Like PSSP, HACMP has a dependency on the RSCT software. See “The RS/6000 Cluster Technology components” on page 225.

Enhanced security options restrictions:
The HACMP licensed program is not automatically supported with restricted root access enabled. See “Considering restricted root access” on page 146, “Considering a secure remote command process” on page 151, and “Considering choosing none for AIX remote command authorization” on page 152 for option descriptions and a complete list of limitations.

Coexistence statement: HACMP 4.5.0 is not compatible with any of the earlier versions. While there is a version compatibility function to allow HACMP 4.5 to temporarily coexist in a ***cluster*** with mixed releases, it is intended as a migration aid only. With PSSP, an HACMP ***cluster*** equals an SP system partition or an entire unpartitioned Cluster 1600 system managed by PSSP (SP system). Once the migration is completed, each node within a cluster must be at the same AIX and HACMP release levels, including all PTFs.

HACMP and HACMP/ES can coexist in a mixed system partition containing nodes running the supported combinations of PSSP with the following conditions:

- The functions provided by HACMP 4.5 are not available until all nodes in the cluster are running HACMP 4.5. They are not available in a mixed version cluster. The commands that create sites will ensure the entire cluster is at the HACMP 4.5 level.
- *HACMP and HACMP/ES clusters do not interoperate*. HACMP nodes must be in a separate cluster from HACMP/ES nodes.

Migration statement: HACMP 4.5 will not interfere with node-to-node migration from prior releases of the HACMP program. The functions provided in this release will not be available until all nodes in the cluster have HACMP 4.5. Table 55 on page 231 lists the levels of PSSP and corresponding levels of AIX in which HACMP levels can coexist during migration only.

Table 55. HACMP Levels supported during migration only

HACMP	AIX 4.3.3	AIX 5L 5.1	PSSP 3.2	PSSP 3.4	PSSP 3.5
4.5.0		S		S	S
4.4.1	S	S	S	S	
4.4.0	S		S		
Note: HACMP 4.5.0 is not compatible with previous releases. The HACMP version compatibility function exists only to ease migration, not to provide long-term compatibility between versions of the program.					

LoadLeveler

The LoadLeveler (LL) licensed program supports scheduling and load balancing of parallel jobs on the SP system. LoadLeveler 3.1 supports coexistence and node by node migration with some restrictions.

Coexistence Statement: AIX 5L 5.1 is a prerequisite to installing LoadLeveler 3.1. LoadLeveler 3.1 will not coexist and interoperate with AIX 4.3.3 on a node. But to support migration, LoadLeveler 3.1 can coexist with LoadLeveler 2.2 and AIX 4.3.3 in an SP system partition with the following restrictions:

- The checkpoint-restart function for parallel jobs cannot be used until all nodes within the SP system partition or cluster are running LoadLeveler 3.1. Checkpoint-restart does not run on nodes running the 64-bit kernel.
- New functions provided in LoadLeveler 3.1 cannot be used until all nodes are running LoadLeveler 3.1.
- All nodes running LoadLeveler 2.2 must have the minimum PTF level described in the README file.

LoadLeveler and Parallel Environment on one node can coexist in these combinations:

- LoadLeveler 3.1 with Parallel Environment 3.2, AIX 5L 5.1, and PSSP 3.5 or PSSP 3.4
- LoadLeveler 2.2 with Parallel Environment 3.1, AIX 4.3.3, and PSSP 3.2 or PSSP 3.4

Migration Statement: LoadLeveler 3.1 supports node by node migration from LoadLeveler 2.2:

- To migrate from LoadLeveler 2.2 to 3.1, first migrate the Central Manager node. After the Central Manager is running 3.1, all other nodes can be migrated node by node. Do not use the **llacctmrg** command while there are mixed levels of LoadLeveler on the system.
- LoadLeveler 3.1 and Parallel Environment 3.2 have PTFs to support 64-bit kernels in PSSP 3.5 so they can run a combination of nodes with PSSP 3.4 and 32-bit kernel, or with PSSP 3.5 and 32 or 64-bit kernel. After a node is upgraded, there are 32-bit and 64-bit versions of the LoadLeveler kernel extension. At run time, depending on whether a 32-bit or a 64-bit kernel is active, the correct version of LoadLeveler is loaded with no need for configuration changes or operator intervention. The first node selected for the 64-bit LoadLeveler upgrade does not have to be the Central Manager node.

Table 56 shows the supported levels.

Table 56. Supported LoadLeveler levels

LL	AIX 4.3.3	AIX 5L 5.1	PSSP 3.2	PSSP 3.4	PSSP 3.5
3.1		S		S	S
2.2	S		S	S	
Note: Running LL 2.2 on AIX 4.3.3 requires that LL be at the PTF level that is noted in the PSSP 3.5 *Read This First* document. For the latest LoadLeveler migration instructions, see the README file distributed with LoadLeveler 3.1.					

Parallel Environment

Parallel processing support applications, like IBM Parallel Environment (PE), are not supported in a mixed system partition. This applies to their use of either IP or user space communication. All the nodes involved in a parallel job must be running the same level of Parallel Environment.

Parallel Environment is comprised of:

- Parallel Operating Environment (POE)
- Message Passing Libraries (MPI and MPL)
- Parallel Utilities which facilitate file manipulation (MPI sample programs)

Coexistence: A single job uses a homogeneous environment. All of the nodes involved in a single parallel job must be running the same level of PE, PSSP, and AIX. PE 3.2 supports PSSP 3.5 and PSSP 3.4 with AIX 5L 5.1.

The PE MPI libraries used to run a job on a node must be compatible with the PSSP MPCI libraries on that node. The following combinations are compatible:

- PE(MPI) 3.2 and PSSP(MPCI) 3.5 or 3.4 and AIX 5L 5.1
- PE(MPI) 3.1 and PSSP(MPCI) 3.4 or 3.2 and AIX 4.3.3

See “LoadLeveler” on page 231 for associations between Parallel Environment and LoadLeveler. Other considerations apply:

- 64-bit support

  All nodes in a parallel job must have the same level of PE and PSSP – there is no mix and match. Applications that use LAPI must be at the same level of PSSP throughout. The 64-bit and 32-bit libraries coexist. Tasks of 32-bit applications can run concurrently with tasks of 64-bit applications on the same node. Tasks running 64-bit and 32-bit can coexist on the same node. Any one MPI job must consist of all 64-bit or all 32-bit tasks. LAPI jobs that do not use MPI can consist of a mixture of 64-bit and 32-bit tasks. All applications wanting PSSP services that use the 64-bit kernel must run with all nodes running AIX 5L 5.1 and PSSP 3.5. 64-bit applications can run with AIX 5L 5.1 and PSSP 3.4, but PSSP 3.4 services do not use the 64-bit kernel.
- Checkpoint/Restart

  Different versions of the checkpoint/restart function cannot coexist on the same system. Applications built to use the new kernel level checkpointing must only run on nodes running PSSP 3.5 or PSSP 3.4 with the correlating versions of LoadLeveler and PE.

  On a system with mixed levels of LoadLeveler, PE, and PSSP, the user can direct a job to run on the appropriate set of nodes by using the ***requirements*** statement in a job command file or by using a host list. LoadLeveler and POE cannot automatically detect the level of checkpointing that is necessary to run the application.

Migration: PE 3.2 supports PSSP 3.5 and PSSP 3.4 with AIX 5L 5.1. To migrate to PE 3.2, the following considerations apply:

- Parallel Environment does not support node by node migration. All the nodes in an SP system partition must be migrated to a new level of PE within the same service window.
- Applications using **threads** have binary compatibility without recompiling.
- You can upgrade to PSSP 3.4 before migrating to PE 3.2. A migration to PSSP 3.5 requires a migration to PE 3.2 in the same service window, if PE was not migrated previously. PSSP contains MPCI. The PE package contains POE, PMD, and MPI. The LAPI libraries are also included in the PE 3.2 ppe.poe, when ssp.css is not installed, for support of LAPI shared memory on standalone RS/6000 or pSeries systems without an SP or PSSP installed.
- 64-bit support

  User applications might need to be recompiled with the **-q64** flag, but that is application dependent. LoadLeveler 3.1 and Parallel Environment 3.2 have PTFs to support 64-bit kernels in PSSP 3.5 so they can run a combination of nodes with PSSP 3.4 and 32-bit kernel, or with PSSP 3.5 and 32 or 64-bit kernel.
- Checkpoint-restart

  The checkpoint-restart function supported with AIX 5L 5.1 runs only on a system running PSSP 3.5 or PSSP 3.4 and the correlating level of LoadLeveler and Parallel Environment. The earlier checkpoint function will not work on an upgraded node. Checkpoint-restart supports 64-bit applications, but it is not supported on nodes running the 64-bit kernel.
- MPCI

  The MPCI library is shipped with the CSS component of PSSP. There are modifications in the MPCI packet header format. No mixed levels of nodes/releases is allowed for running a MPI job.
- SP switch

  The checkpoint-restart function is supported with the SP Switch or the SP Switch2, but parallel user space jobs checkpointed on a system with one type of switch cannot be restarted on another system with a different type of switch.
- See “LoadLeveler” on page 231 for associations between Parallel Environment and LoadLeveler.

Table 57 shows the supported levels. See Table 51 on page 218 for possible migration stages.

Table 57. Supported Parallel Environment levels

PE	AIX 4.3.3	AIX 5L 5.1	PSSP 3.2	PSSP 3.4	PSSP 3.5
3.2		S		S	S
3.1	S		S	S	
Note: PE has some Program Temporary Fix (PTF) dependencies. See the PSSP 3.5 *Read This First* and the PE README documents for the latest information on which PTF levels are necessary.					

Parallel tools

Parallel tools include:

- PE Benchmarker
  - Performance Collection Tool (ppe.perf)
  - Profile Visualization Tool (ppe.pvt)
- Parallel debugger (PDBX) (Dependent on POE)

- Application performance analysis tool (Xprofiler)
- Dynamic Probe Class Library (DPCL) (Supported in open source)

These tools are shipped with Parallel Environment which requires that all the nodes involved in a parallel job be running the same level of Parallel Environment. These have the same coexistence limitations as stated for “Parallel Environment” on page 232 with the exception of PE Benchmarker and Xprofiler.

PE Benchmarker does not prevent node-by-node migration from supported PSSP levels. PE Benchmarker is only supported in AIX 5L 5.1. It does not depend on PSSP, but it will use the switch clock API in PSSP if it is present.

Xprofiler has no dependency on PSSP. It does not interoperate with other instances of Xprofiler, but it does not interfere with PSSP coexistence or migration. Table 58 lists the supported levels.

Table 58. Supported Xprofiler levels

Xprofiler	AIX 4.3.3	AIX 5L 5.1	PSSP 3.2	PSSP 3.4	PSSP 3.5
1.2		S		S	S
1.1	S		S	S	
1.0	S		S	S	

Parallel ESSL

Parallel ESSL is not supported in a mixed system partition. PESSL can only run in a system partition with all nodes at the same AIX and PSSP level. Which level of Parallel ESSL runs on a particular level of PSSP and AIX is based on which level of PE runs on a particular level of PSSP and AIX:

- Parallel ESSL 2.3 requires ESSL 3.3, PE 3.2 on AIX 5L 5.1 and PSSP 3.5 or PSSP 3.4
- Parallel ESSL 2.2 requires ESSL 3.2, PE 3.1 on AIX 4.3.3 and PSSP 3.4 or PSSP 3.2
- Parallel ESSL is not directly dependent on a level of PSSP or AIX.

Parallel ESSL coexistence and migration is the same as for Parallel Environment because it is dependent on it. It also requires the ESSL licensed program. See “Parallel Environment” on page 232.

IP performance tuning

This section presents some high-level considerations related to performance of TCP/IP over the switch in a coexistence environment. Note that these are simply important factors to be considered in approaching tuning, and that the SP organization has not conducted significant performance evaluation studies in this area.

In general, with all else being equal, the goal for performance achieved between nodes running different levels of PSSP should be the performance delivered by the earlier level of PSSP (each release of PSSP has included performance improvements). Traditional tuning considerations, such as those derived from the performance characteristics of different SP node types and installation/application communication patterns will still apply. For example, the switch throughput is limited to the speed of the slowest node in an IP connection. With coexistence, tuning activities might also need to reflect the levels of PSSP on the particular nodes running (communicating) in a mixed system partition.

There are two main areas where this might come into play:

1. Tuning for AIX - tuning methodologies typically employed for different releases.
2. Tuning for the switch - appropriate settings for the adapter device driver buffer pools.

In tuning for the switch, the values used for the switch adapter or device driver IP buffer pools are the primary considerations. The rpoolsize and spoolsize parameters are changed using the **chgcss** command. The aggregate pool size is a function of the size of kernel memory.

In summary, the suggested approach for factoring coexistence into your overall SP tuning strategy is to begin with the above general approach to tuning for mixed levels of AIX and PSSP. Consider the other characteristics that influence performance for your specific configuration, making trade-offs if necessary. Then, as with any performance tuning strategy, make refinements based on your results or as your SP migration strategy progresses.

Note: See performance tuning information on the Web at:

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

Changes in recent levels of PSSP

Here are some recent changes in PSSP, PSSP-related licensed programs, or AIX support that might affect your migration plans.

Support for applications with 64-bit addressing

PSSP 3.5 supports applications that use 64-bit addressing. User applications might require recompile with the **-q64** flag, but that is application dependent.

All nodes in a job must run the same level of PE and PSSP, there is no mix and match. Applications that use LAPI must be at the same level of PSSP throughout. The 64-bit and 32-bit libraries coexist. Tasks running 32-bit and 64-bit can coexist on the same node. An MPI job must consist of all 32-bit or all 64-bit tasks. LAPI jobs can consist of a mixture of 32-bit and 64-bit tasks. Applications that use 64-bit addressing must use AIX 5L 5.1 and PSSP 3.5.

Refer to the following Web site to view the various AIX processors that support the 64-bit kernel:

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

then select ″AIX 5L release notes″ and select the HTML option for ″AIX 5L for POWER Version 5.1 Release Notes.″ Using the ″Contents,″ search for the 64-bit Kernel entry under the ″Base Operating System (BOS)″ heading.

Support for checkpoint-restart

The latest checkpoint-restart function can run only on a system running the latest versions of LoadLeveler, Parallel Environment, and PSSP 3.5 or PSSP 3.4 with AIX 5L 5.1. The older checkpoint function will not work on any upgraded node. Checkpoint-restart supports 64-bit applications, but it is not supported on nodes running the 64-bit kernel.

RSCT packaging

RS/6000 Cluster Technology (RSCT) is the package of high availability support programs that began as components of PSSP:

- Event Management
- Group Services
- Topology Services

They have been separated from PSSP as a package to be made available with AIX for broader use. See “The RS/6000 Cluster Technology components” on page 225.

Performance Toolbox Parallel Extensions

The Performance Toolbox Parallel Extensions for AIX (PTPE) software, an optional component of PSSP 3.2, has been withdrawn from service and is not in PSSP 3.5 or PSSP 3.4. This does not prevent the PTPE software from running on earlier PSSP nodes in mixed SP configurations.

SP TaskGuides

The SP TaskGuides component available in PSSP 3.2 has been withdrawn from service and is not in PSSP 3.5 or PSSP 3.4.

SP Job Manager Package

The SP Job Manager Package available in PSSP 3.4 has been withdrawn from service and is not in PSSP 3.5.

AIX support

AIX 5L 5.1 does not preserve binary compatibility for 64-bit applications because it introduces a new 64-bit ABI. Therefore PSSP 3.2 and earlier releases are not supported on AIX 5L 5.1.

Information about AIX can be found in the relevant edition of the book *Differences Guide*. See the “Bibliography” on page 313for how to find AIX information on the Web.

TCP/IP Internet Protocol Version 6 (IPv6) extends the maximum number of IP addresses from 32-bit addressing to 128-bit addressing. IPv6 is compatible with the current base of IPv4 host and routers. IPv6 and IPv4 hosts and routers can *tunnel* IPv6 datagrams over regions of IPv4 routing topology by encapsulating them within IPv4 packets. IPv6 is an evolutionary change from IPv4 and allows a mixture of the new and the old to coexist on the same network.

Restriction: IPv6 is not supported for use by the PSSP components. It cannot be used with SP adapters and is incompatible with the RSCT components. If you are using an SP-attached or clustered server, be sure that the server does not use IPv6. Some PSSP components tolerate IPv6 aliases for IPv4 network addresses but not with DCE, HACMP, HACWS, or an SP switch. For more information about the SP system tolerating IPv6 aliases for IPv4 network addresses, see the appendix on the subject in the book *PSSP: Administration Guide*.

AIX and PSSP migration options

There are three main ways to migrate your system each with their own advantages:

1. Migration install - preserves base configuration
2. Overwrite install - provides a clean start
3. Migration then re-install - migrate one node then use this image to re-install remaining nodes

After performing any needed system preparation steps, the next step in migrating your SP system is to migrate the control workstation to the appropriate level of AIX

and PSSP. That is, the control workstation must be migrated to AIX 5L 5.1 and PSSP 3.5 before migrating any of the nodes. The control workstation must be at the highest release levels to be used on any of the nodes.

Migrating from PSSP 3.2 and AIX 4.3.3 to PSSP 3.5 and AIX 5L 5.1 must be done in one service window. Afterward, the nodes can be migrated from any supported base release to PSSP 3.5 on AIX 5L 5.1 with a single **nodecond** operation.

For systems running unsupported levels of PSSP or for an overwrite install, the control workstation migration must include both AIX and PSSP upgrades in the same service window before the SP system can be returned to production. To migrate from unsupported levels like PSSP 2.4 or PSSP 3.1, first migrate the control workstation and nodes to PSSP 3.4 on AIX 4.3.3. Then, migrate the control workstation to AIX 5L 5.1 and migrate the nodes to AIX 5L 5.1 with a second **nodecond** operation.

After the control workstation has been migrated to AIX 5L 5.1 and PSSP 3.5, and the system has been validated, the nodes can be migrated. Migrate nodes beginning with any boot-install servers you might have. The basic migration options for migrating the nodes are:

- AIX Migration
- PSSP Migration
- AIX and PSSP Migration (done in the same service window)

Also, you can optionally migrate one node, then use the mksysb from that node to install the remaining nodes to be migrated.

Reviewing your migration steps

This section summarizes the key components of a migration. Review and assess these components, consider them from a sizing and impact point of view, and qualify them with respect to your overall migration goals and strategy. Additional details on these steps can be found in the *PSSP: Installation and Migration Guide*.

1. Determine your migration goals (which nodes, how many nodes)
2. Determine your migration strategy. Be certain that you understand coexistence limitations.
3. Plan your migration windows
4. Plan your recovery procedures
5. Gather necessary materials:
   - new release levels of AIX and PSSP
   - documentation - for AIX, PSSP, other licensed programs, required AIX and PSSP service (for older levels)
   - new release levels of other licensed programs used
   - any additional DASD required, resources (like tape) for backups
6. Create system backups - control workstation, nodes to be migrated
7. Conduct the migration, in stages as applicable:
   a. Quiesce the system, stopping all production.
   b. If necessary, to migrate the control workstation first to PSSP 3.4 and AIX 4.3.3, do the following:
      1) Apply required service to nodes (required PTFs are necessary for coexistence to work).

2) Prepare the control workstation (DASD, PTF service, SDRScan, archive the SDR).
3) Migrate the control workstation to AIX 4.3.3 and PSSP 3.4.
4) If migrating from pre-PSSP 3.2, you must merge any customization data in the **/etc/sysctl*.acl** file with the new **/etc/sysctl*.acl** files delivered with PSSP 3.4. You can find the saved **/etc/sysctl*.acl** files in the **/lpp/save.config/etc** directory after the install.
5) If migrating from PSSP 3.2 or later, you can simply copy the **/lpp/save.config/etc/sysctl*.acl** files to the **/etc** directory after the install. No new entries were put in these files for PSSP 3.4.
6) Validate.
7) If any nodes are at PSSP 3.1.1 or earlier, migrate them to PSSP 3.2 or later.

c. To migrate the control workstation to PSSP 3.5 and AIX 5L 5.1, do the following:
1) Prepare the control workstation (DASD, PTF service, archive the SDR).
2) Apply required service to nodes (required PTFs are necessary for coexistence to work).
3) Migrate the control workstation to AIX 5L 5.1 and PSSP 3.5.
4) Validate.

d. Partition the system if necessary due to coexistence limitations or switch configuration.
e. Migrate a test node to PSSP 3.5 and AIX 5L 5.1 (optional but strongly suggested) and validate.
f. Migrate any boot-install servers to the latest level of PSSP and AIX that is to be used on any nodes they serve.
g. Migrate any other nodes to PSSP 3.5 (and AIX 5L 5.1 if necessary).
h. Migrate GPFS and HACMP if they are being used.
i. Migrate LoadLeveler and Parallel Environment if they are being used and not at the correct levels.

Note: Complete and verify each step before going on to the next step.

8. Perform post-migration activities.

Appendix A. The System Partitioning Aid - A brief tutorial

PSSP includes a tool to facilitate system partitioning activity. You can partition any SP system that does not use the SP Switch2 and that is not a system of clustered enterprise servers. The objectives of this application are to enhance understanding of system partitioning, and to allow you to create system partitioning configurations beyond those provided with PSSP. This application, called the *System Partitioning Aid*, is provided in two forms:

sysparaid a command line interface (CLI) which is text-file based;

spsyspar a graphical user interface (GUI) which provides capability to view graphical representations of system partitioning layout alternatives, and to dynamically create new alternatives.

The GUI makes use of the command line interface, and uses the *SP Perspectives* code for graphics support. Both interfaces allow you to verify candidate layouts, and allow you to save a new, valid layout to disk. The new layout is then available to be made the active configuration at a later date. This allows you to plan ahead for configuration changes.

This appendix describes the GUI, then the CLI version of this application. The GUI is the interface suggested for a person not experienced with partitioning. In addition, this appendix presents a partitioning exercise which addresses the example in Chapter 5 of this document.

The GUI - spsyspar

The GUI version of the System Partitioning Aid provides a dynamic view of the system partitioning layout, allowing you to modify the layout interactively.

The command **spsyspar** brings up the window shown in Figure 38 on page 240. This window consists of five screen areas:

Pull Down Menu Bar Menus provide pull down access to actions.

Tool Bar Icons provide immediate execution of certain actions.

Nodes Pane Graphic representation of targeted system partitioning layout.

System partitions Pane Iconic representation of system partitions in the current layout.

Information Area Displays information about the object or screen area at the current cursor location. (Resides at very bottom of window.)

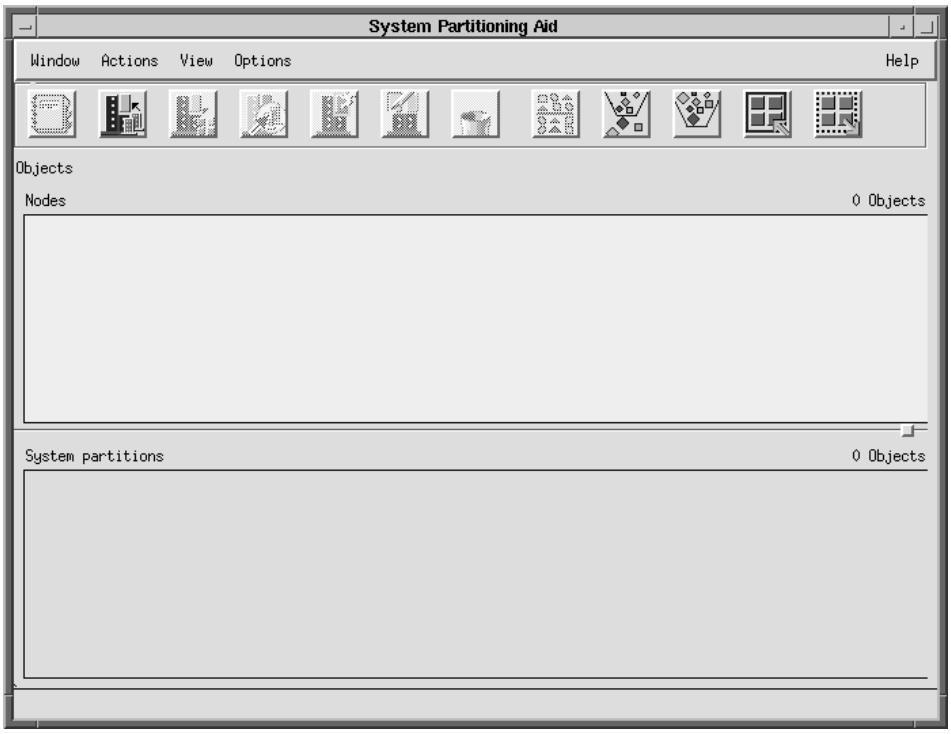


Figure 38. System Partitioning Aid main window

In Figure 38, the Nodes and System Partitions panes are empty. If an SDR exists, **spsyspar** treats the active system partitioning layout as the current target, and pictures it in the object panes. So, on an active system, **spsyspar** does not come up with empty panes. The Nodes pane contains the frames and nodes of the system, and the System partitions pane contains system partition icons.

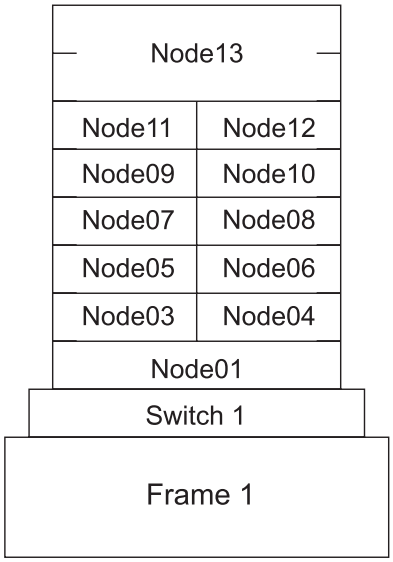


Figure 39. Sample 1-frame system (1 wide, 10 thin, and 1 high nodes)

For example, assume you invoked **spsyspar** on the control workstation for the 1-frame system pictured in Figure 39, where there is 1 wide node, 10 thin nodes and 1 high node. If the active system partitioning layout has the bottom half of the

frame in system partition Alpha and the top half in system partition Beta, then **spsyspar** presents the window shown in Figure 40. A single frame is presented in the Nodes pane, with the nodes pictured as defined (thin, wide, or high) in the SDR. Icons for partitions Alpha and Beta are shown in the System partitions pane.

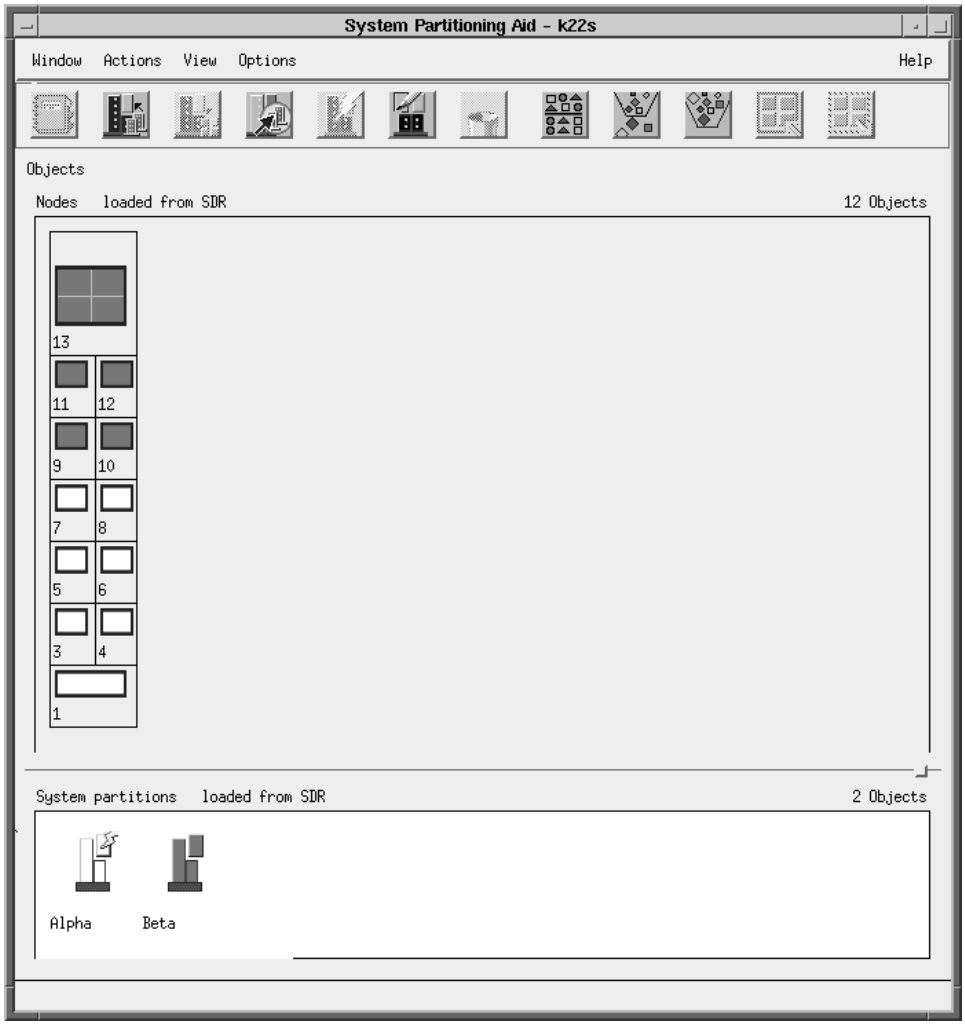


Figure 40. Main window for sample system

The **spsyspar** window is a standard window which you can move and size like any other window. The Nodes and System partitions panes become scrollable when appropriate. Also, the division of real estate between these two panes is controlled via the small box located between them and at the right side of the window; that box is called a *sash*.

In Figure 40, notice that the title of the window contains ″ - k22s″. ″k22s″ is the name of the control workstation of the target system. Also, if you look closely at the ″System partitions″ pane of Figure 40, you will see that the Alpha partition is marked with a ″lightening bolt″. This signifies that Alpha is the *active partition*. Any partition-specific activity, such as assignment of nodes, would be done for objects in partition Alpha. In addition, the brighter colored ″System partitions″ pane is the pane of *focus*. This determines the choices available from the Tool Bar and the Pull Down Menu – items not applicable to the current focus are grayed out and not accessible.

Tool bar actions

The Tool Bar consists of several icons which allow you to execute important actions. These actions are also available through the Pull Down Menu Bar.

View and modify information about selected objects (Notebook)

The availability of the icons of the Tool Bar is generally affected by the nodes and system partitions previously selected. Actions which are not available appear grayed out. For example, if you click on node 8 in the Nodes pane, then select the first Tool Bar icon, which pictures a notebook, a new window comes up named ″View Node 8″ containing data relevant to node 8. This window appears in Figure 41.

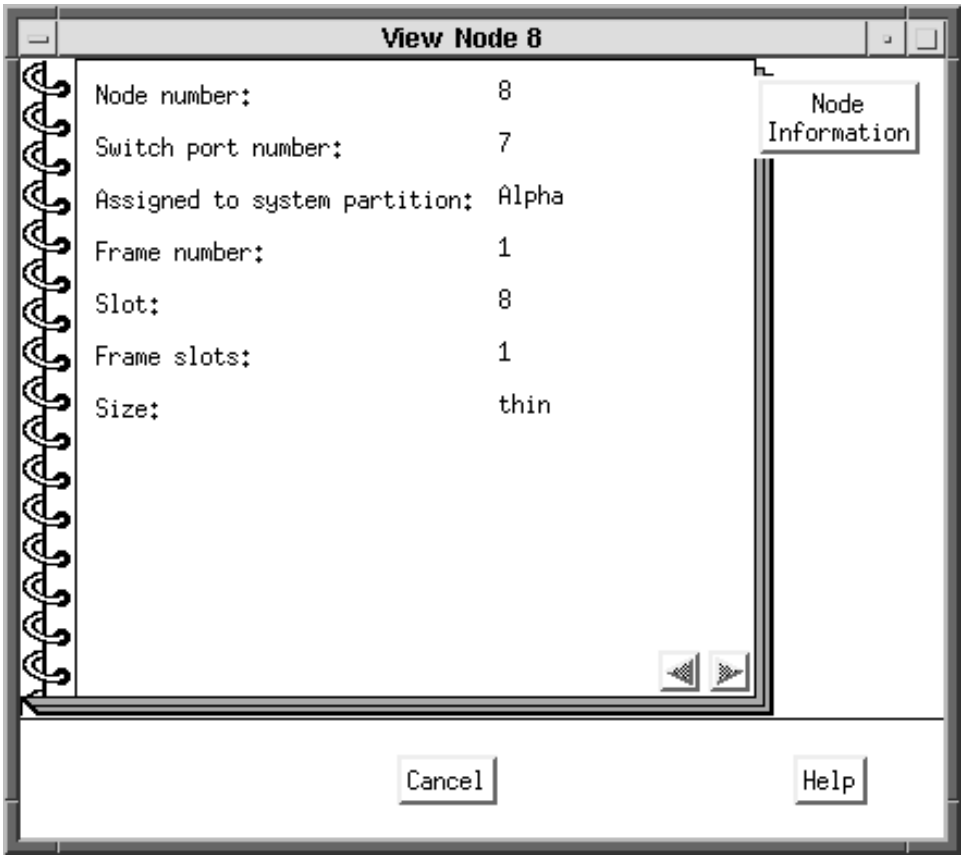


Figure 41. Notebook for node 8 of sample system

If you instead click on partition `Alpha` in the system partitions pane, then select the notebook icon, you get a window named ″View/Modify System Partition Alpha″, which contains data for system partition `Alpha`. This system partition notebook is more complicated than a Node notebook, and contains each of the following pages, which are shown in Figure 42 on page 243 for this example:

Definition
: partition name and description, together with current **spsyspar** session parameters

Nodes
: a list of information for the nodes in this partition

Topology File
: view of the topology file specifying this partition

Chip Allocation
switch chips allocated to this partition, if the configuration was not shipped by IBM

Performance
performance numbers for this partition, if the configuration was not shipped by IBM

Note: A configuration is either one of those shipped by IBM with PSSP in the directory **/spdata/sys1/syspar_configs**, or it was added later by a user of the System Partitioning Aid. The configurations shipped by IBM satisfy certain minimal bandwidth criteria, but partitions created using the System Partitioning Aid might not satisfy that criteria. Configurations created by using the System Partitioning Aid are evaluated for correctness and performance. The ″Chip Allocation″ and ″Performance″ pages of the system partition notebook record such data for a user-created layout.

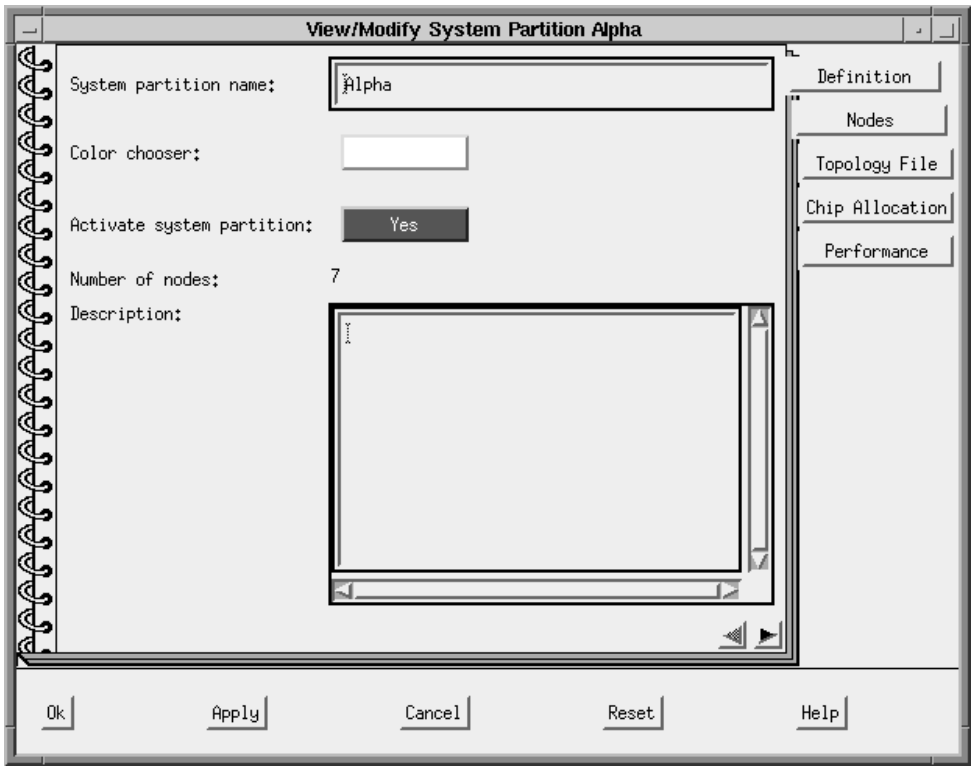


Figure 42. Notebook for partition Alpha of sample system

You can modify each attribute on the ″Definition″ page of the partition notebook, except the number of nodes. The other pages of the notebook are read-only.

Display previously defined and user generated system configurations

Select the second icon on the Tool Bar to display available system partitioning configurations. The resulting dialog box appears in Figure 43 on page 244 and displays the configurations that you can select. Clicking on one of these configurations expands that configuration to show the corresponding layouts available - both those shipped by IBM and the ones created by users. In Figure 43 on page 244, configuration 8_8 has been expanded showing there are three layouts available under this configuration.

If you click on a layout, then the Open button, **spsyspar** now treats that layout as the target system. This makes **spsyspar** useful in planning for future expansion. If the layout is for a configuration that matches the real system, you have a choice of seeing nodes pictured as defined in the SDR. The default, the only possibility if the SDR is not available, is to show only thin nodes with all slots populated. Since **spsyspar** cannot know the correct node types to show, it depicts all nodes as thin.

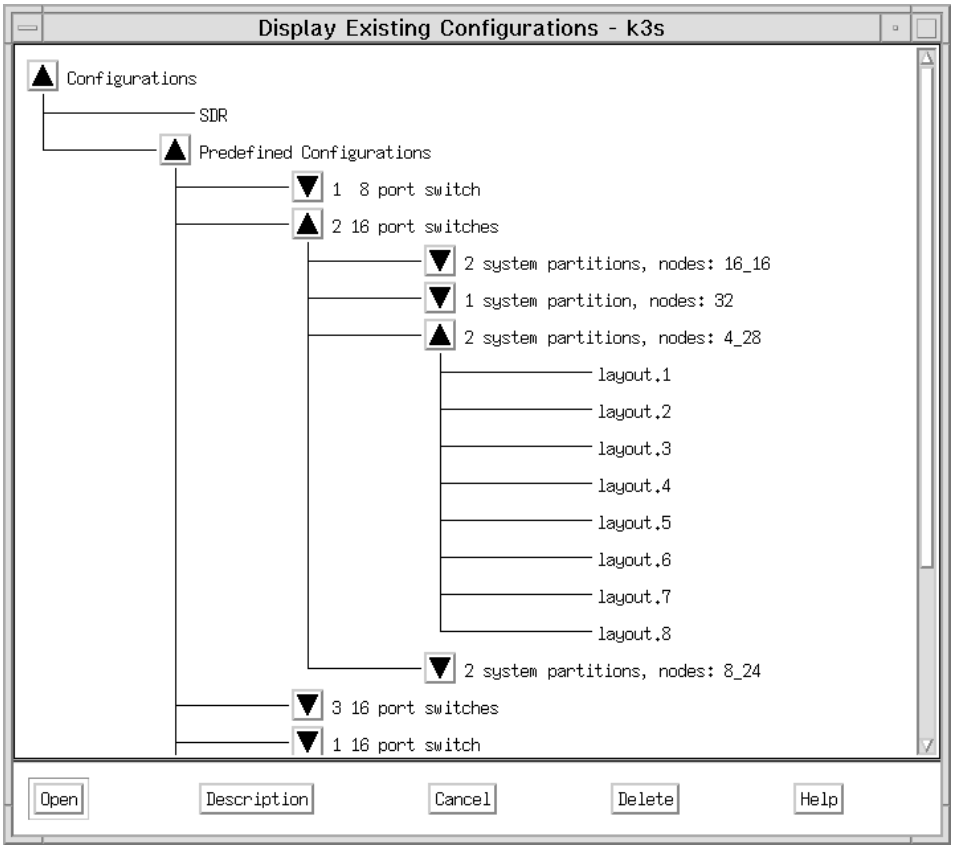


Figure 43. Alpha Notebook for sample system

You also have the opportunity to read the description of a layout or delete a layout created by a user. By looking at the description for layout.3 under configuration '2 system partitions, nodes: 4_28', you would see it is equivalent to the layout depicted in Figure 40 on page 241.

Place selected nodes into an active partition

You can set the active partition by selecting a partition in the System Partitions pane and then choosing ″Select Active″. Then, under the ″Actions″ pull down, select ″System Partitions″. (See also the description for the fifth icon below.) Once an active partition is set, you may select nodes in the nodes pane and then select the third icon. This moves any selected nodes into the active partition. In addition, any nodes attached to the same switch chip(s) as the node(s) selected are also placed in the active partition. A message appears informing the user that this has happened.

In our example, if `Beta` is the active partition, and node 1 is selected, then clicking on the third icon moves nodes 1, 5, and 6 from partition `Alpha` to partition `Beta`.

Generate files used to define system configuration

The fourth icon checks whether the current system partition layout is equivalent to one which already exists, and if not, builds the corresponding layout in the appropriate location on disk. Then this new layout may be chosen as the active configuration at a later time.

Activate a system partition for node assignment

The fifth icon provides an alternate way of setting the active partition. This is equivalent to choosing ″Select Active″ under the ″Actions″ pull down. The current active partition is marked with a lightening bolt.

Define a new system partition

The sixth icon brings up a ″Define System Partition″ dialog box which is actually the ″Definition″ page in a new system partition's notebook. You can specify the name, description, and color of the new partition. Of course, this new partition has no nodes yet, because you must first perform a ″Place selected nodes...″ for this new partition. The new partition is also set as the active one to prepare for specifying member nodes.

Remove selected system partition

The seventh icon deletes the selected system partition from the current layout. If the selected partition has nodes assigned and is currently the active partition, you cannot delete the partition until all nodes of the partition have been reassigned to another partition(s). If the selected partition has no nodes assigned and is currently the active partition, it cannot be deleted until a different partition becomes the active partition.

Sort the objects in the current pane

The eighth icon sorts the node or system partition objects in the respective pane, depending on which pane is currently active. For the Nodes pane, this makes sense and is only available for use if the icon view of the nodes has been set via the ″View″ Pull Down Menu item. The icon view dispenses with frames and simply represents all the nodes as independent entities. The icon view of the Nodes pane has been selected in Figure 44 on page 246, and the nodes have sorted in descending order.

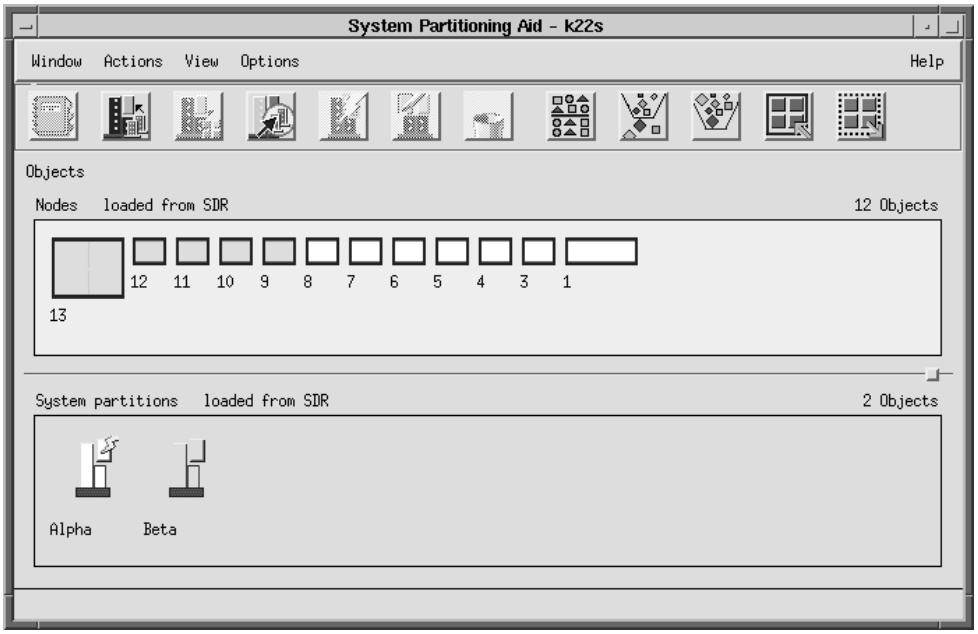


Figure 44. Descending sort in Nodes pane (icon view)

Filter the objects in the current pane

The ninth icon allows you to define a filter, and uses that filter to control which objects in the active pane are seen. In our example, if the node pane is selected, specifying the filter ″1*″ for inclusion as shown in Figure 45 causes the frame to be redrawn with only nodes 1, 10, 11, 12, and 13 shown. Alternatively, you may select those nodes in the Nodes pane, and choose the ″Filter by what is selected″ option on the ″Filter Nodes″ dialog window.

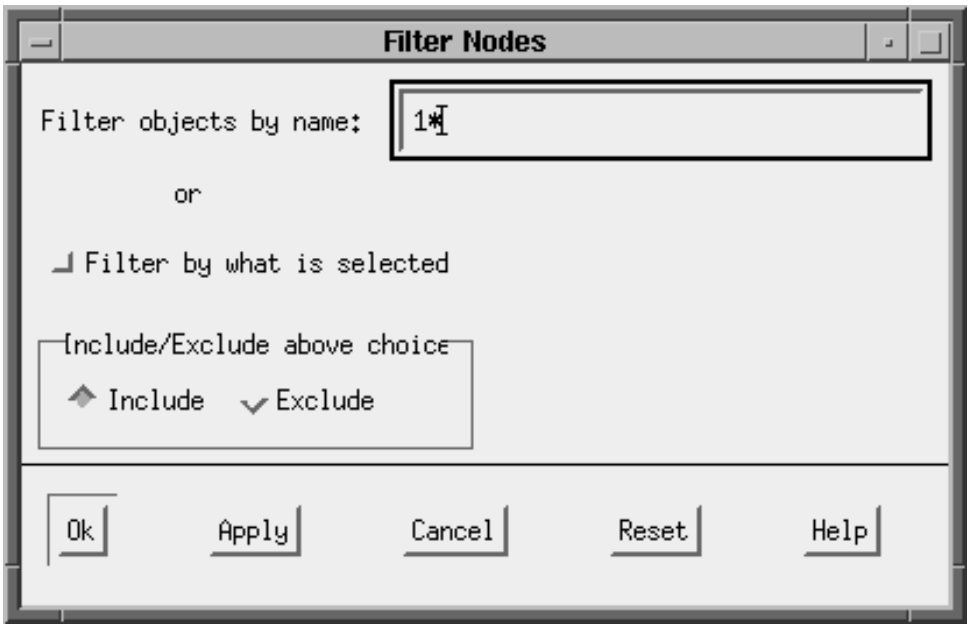


Figure 45. Filter menu with ″1″ filter specified for Nodes pane*

If you select the System partitions pane, specifying the filter B* for inclusion results in only the `Beta` system partition being shown: both in the Nodes pane and the partitions pane. A filter may be imposed on each pane.

Remove any filter being applied to the objects in the current pane

The tenth icon undoes any filtering for the currently active pane.

Select all objects in the current pane

The eleventh icon applies only to the Nodes pane. It marks all the nodes as if they had been sequentially selected. Then, you may deselect nodes one at a time to achieve the desired combination.

Deselect all objects in the current pane

The twelfth icon also applies only to the Nodes pane. It clears all selections from the pane so you can start from the beginning again.

The CLI - sysparaid

Use the command **sysparaid** to verify the validity of a system partitioning configuration without invoking the GUI. Optionally, you may request the corresponding layout files be constructed and saved for activation later.

The CLI **sysparaid** is invoked by the GUI **spsyspar** to handle a graphically specified layout. In that case the **spsyspar** code constructs the necessary input data and option specifications for the user.

When working with **sysparaid** directly, you must provide these inputs and options. The syntax for the CLI is shown below. For complete syntax, see the book *PSSP: Command and Technical Reference*.

```
sysparaid [-s layout_name | a_fully_qualified_path]
 input_file [topology_file]
```

where:

- *input_file*

  is the input file specifying the system partitions.

- *topology_file*

  is an optional topology file to be used in evaluating the candidate system partitioning layout. This file is the master topology file for the target system, and is necessary when this file is not present in the **/spdata/sys1/syspar_configs/topologies** directory.

- **-s**

  Specifies that the configuration layout data is to be saved for later use. If *layout_name* is specified as a simple string, the results are stored at the appropriate location in the system partition directory tree, under the directory named layout.*layout_name*. If *a_fully_qualified_path* is specified, the results are stored at that location only.

The input file must specify the size of the system, number of partitions to be used, which nodes are in which partition and so on. The format of the input file is shown in Figure 46 on page 248 and the file shown is shipped with PSSP in the**ssp.top** file set as the **inpfile.template** file in the directory **/spdata/sys1/syspar_configs/bin**.

Recall the Sample System of Figure 39 on page 240, and the `Alpha` and `Beta` partitions of Figure 40 on page 241. An input file for **sysparaid** which specifies that layout is the **my_part_in** file presented in Figure 47.

```
This file is a template for the input file to the System Partitioning Aid.
Copy this into a new file,  fill all fields as described.
Frame Type of 16 slot frames is tall and that of 8 slot frames is short.
Select one of the four keywords provided for Switch type.
Nodes may be identified using either node numbers or switch port
numbers.
Select one of the two options provided for Node Numbering Scheme.
System Partition Name, Number of Nodes in the System Partition
and list of nodes in the System Partition must be provided for
all system partitions. The node list can be provided in one of
the following formats:
      - A list with one entry on each line
      - A range of the form X - Y
      - A combination of the above options
      - For the last partition the keyword remaining_nodes may be
        used provided all nodes or switch ports not in the last
        system partition have been specified in other system
        partitions.
Comment lines enclosed between /* and */ may be deleted.
New comments may be added provided they follow the comment
convention.
********************************************************************
   Number of Nodes in System:
   Number of Frames in System:
   Frame Type: tall   short
   Switch Type:  HiPS   SP   LC8   SP8   NA
   Number of Switches in Node Frames:
   Number of Switches in Switch Only Frames:
   Number of System Partitions:
   Node Numbering Scheme:  node_number    switch_port_number
   System Partition Name:
   Number of Nodes in System Partition:
   List of nodes in system partition
```

Figure 46. File inpfile.template provided with PSSP

```
   Number of Nodes in System:  12
   Number of Frames in System:  1
   Frame Type:  tall
    Switch Type:  SP
       Number of Switches in Node Frames:  1
       Number of Switches in Switch Only Frames:  0
       Number of System Partitions:  2
       Node Numbering Scheme:  node_number
       System Partition Name:  Alpha
       Number of Nodes in System Partition:  7
       List of nodes in system partition
       1
       3 - 8
       System Partition Name:  Beta
       Number of Nodes in System Partition:  5
       List of nodes in system partition
       9 - 13
```

Figure 47. File my_part_in

You could execute **sysparaid** as follows to check for validity:

```
sysparaid  my_part_in
```

(If the global system topology file is not present in the **/spdata/sys1/syspar_configs/topologies** directory, you must provide that topology file.) **sysparaid** examines the inputs and recognizes that this layout is equivalent to the layout shipped by IBM as:

```
/spdata/sys1/syspar_configs/1nsb0isb/config.8_8/layout.3
```

If **sysparaid** did not find an existing equivalent layout, it would report that the layout is valid, and you could rerun **sysparaid** specifying the **-s** (save) option with a directory in which to place the results. The results would consist of

layout.desc
file describing this system partitioning layout;

nodes.syspar
file with shorthand listing of partition contents;

spa.snapshot
file listing ownership of switch chips by partition;

syspar.1.Alpha
directory for `Alpha` - node list, topology, snapshot, metric files;

syspar.2.Beta
directory for `Beta` - node list, topology, snapshot, metric files.

Example 3 of Chapter 5

The picture of the 3-frame system discussed in Chapter 5 is reproduced in Figure 48 below. Suppose you plan to have this system at some point in the future, and wish to partition it in the manner described in “Example 3 – An SP with 3 frames, 2 SP Switches, and various node sizes” on page 184:

```
Partition 1 - F1N01, F2N01, F1N05, F2N05,
              F1N03, F2N07
Partition 2 - F1N09, F1N13, F2N13
              F1N11, F2N11
Partition 3 - F3N01, F3N02, F3N05, F3N06,
              F3N03, F3N07,
              F3N09, F3N13
```

This layout is not one of those shipped by IBM, You would create it using the System Partitioning Aid. Further, if this system is not ″in hand″, then **spsyspar** cannot picture the system correctly, and shows only thin nodes.

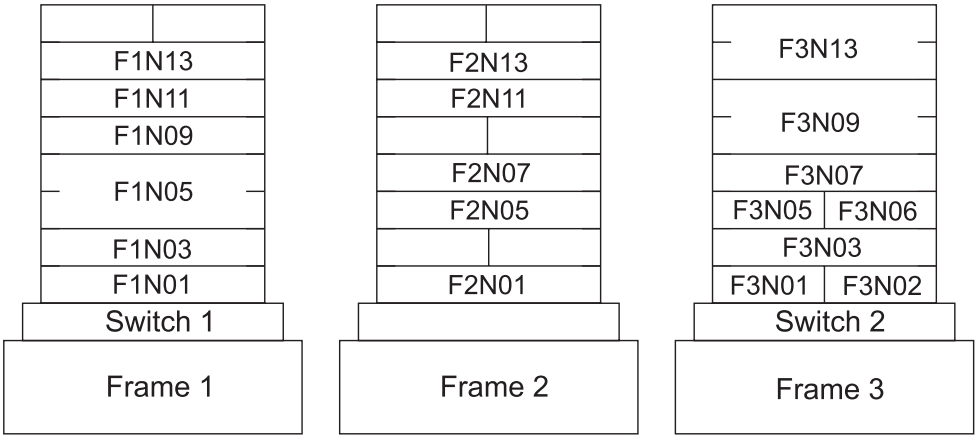


Figure 48. Three frames with 2 switches

1. Start by bringing up **spsyspar**.
2. Click on the ″Display previously defined ...″ icon. (The second Tool Bar icon.)

3. Select the ″2 16 port switches″ and select the ″32″ configuration. You find there is only one such layout. Select this layout and open it. You now have a 2-frame, 32-node system as shown in Figure 49. The system partition name `alice blue` is a default choice, which matches the default color chosen by the tool.

Understand that the first frame in the figure really represents both of Frames 1 and 2: Frame 2 is an expansion frame for Frame 1 since it shares Frame 1's switch. Also, the nodes in the second frame pictured would be in Frame 3 of the real system, and would be numbered starting at 33, rather than 16. Once you complete this exercise, you will save a layout which you can use correctly once the real system is available. Partitioning is based on switch chips, not on node numbers.

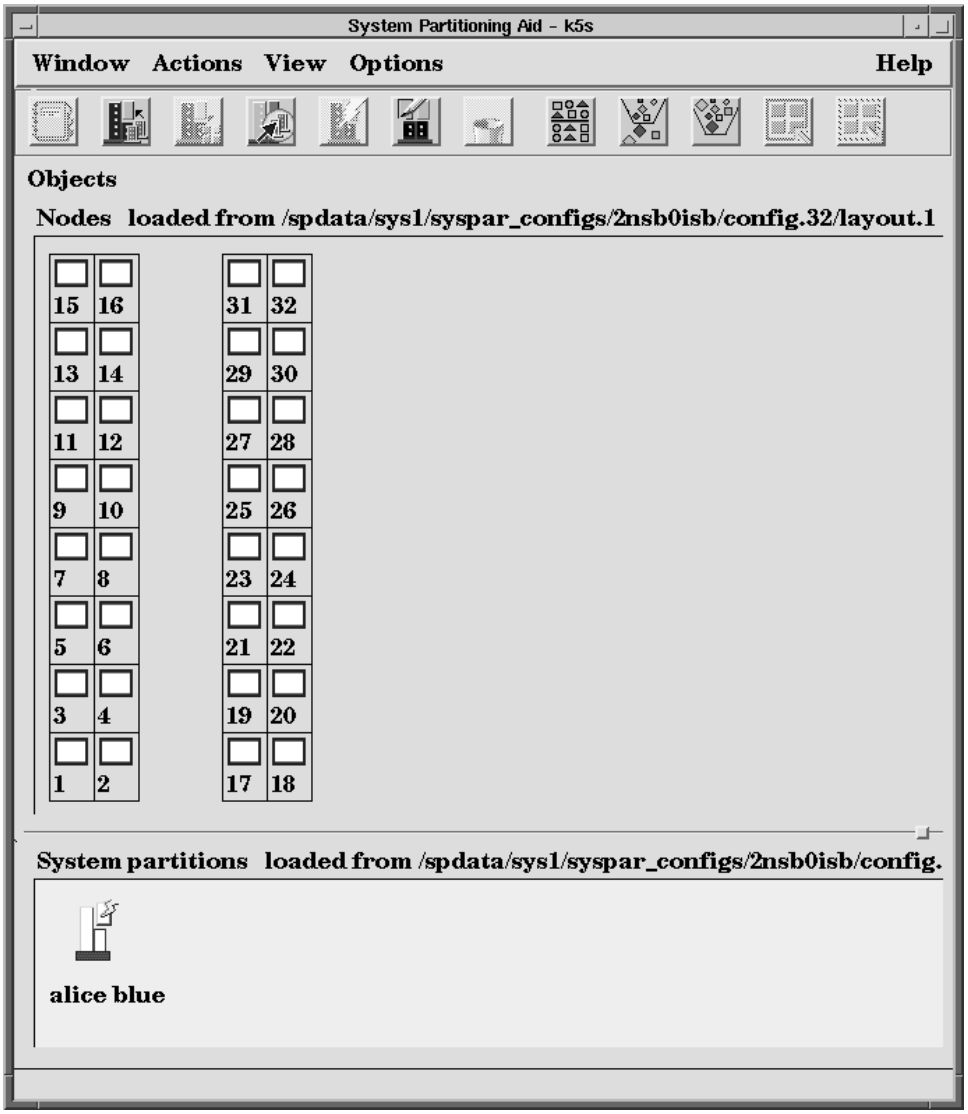


Figure 49. Main window for Example 3 of Chapter 5

Your objective for system partitions is to divide the system pictured in Figure 49 into 3 pieces: the lower half of Frame 1, the upper half of Frame 1, and Frame 2. You may perform the following tasks to accomplish this, and arrive at Figure 50 on page 252:

1. In the notebook for the existing partition (the default partition) change the partition name to `Par1`.
2. Select the ″Define a new system partition″ icon (the one with the pencil) and define a new partition with name `Par2`.
3. Repeat the previous step for `Par3`.
4. Make `Par2` active. (Use the lightning bolt icon)
5. Select node 9. Then assign it to `Par2`. (Third icon.) Note that nodes 9, 10, 13 and 14 move to `Par2` because they all connect to the same switch chip.
6. Select node 12. Then assign it to `Par2`. (Third icon.) Nodes 11, 15 and 16 also join `Par2`.
7. Make `Par3` active. (Use the lightning bolt icon)
8. Select nodes 21, 23, 25, and 27. Then assign these nodes to `Par3` by clicking on the third icon. Notice that all the Frame 3 nodes are placed in `Par3` due to the sharing of switch chips.

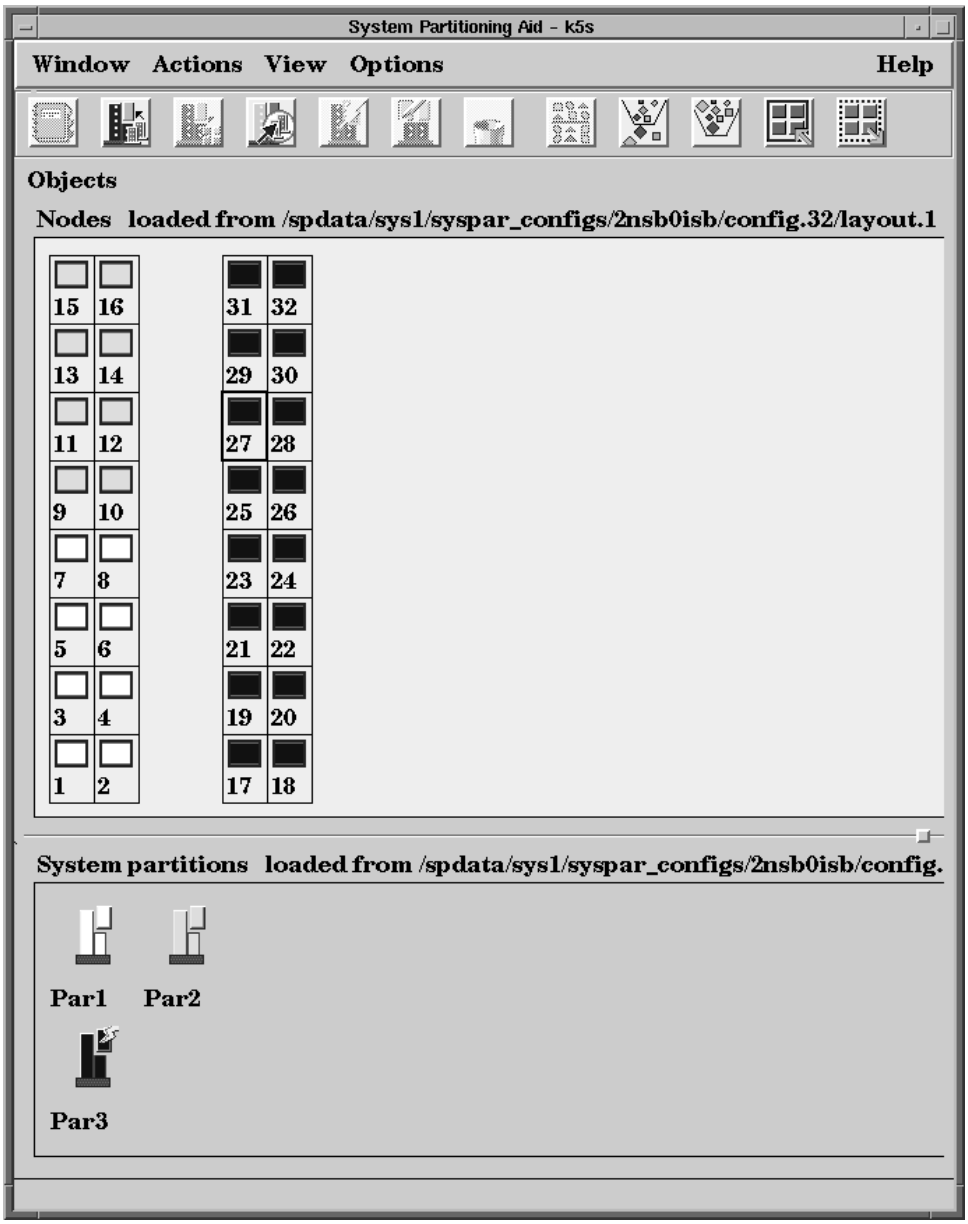


Figure 50. System partitioning for Example 3 of Chapter 5

To make the system represented look more like our system, you can use filtering on the Nodes pane. To do this, follow these steps:

1. Select all the nodes which should be in the system.
2. Select the filtering icon, and choose ″Filter by what is selected.″

The result is depicted in Figure 51 on page 253. For the real system, Nodes 5, 25 and 29 will be high. Figure 51 on page 253 looks good in this respect. However, Nodes 6 and 8 distort our perception of Node 5.

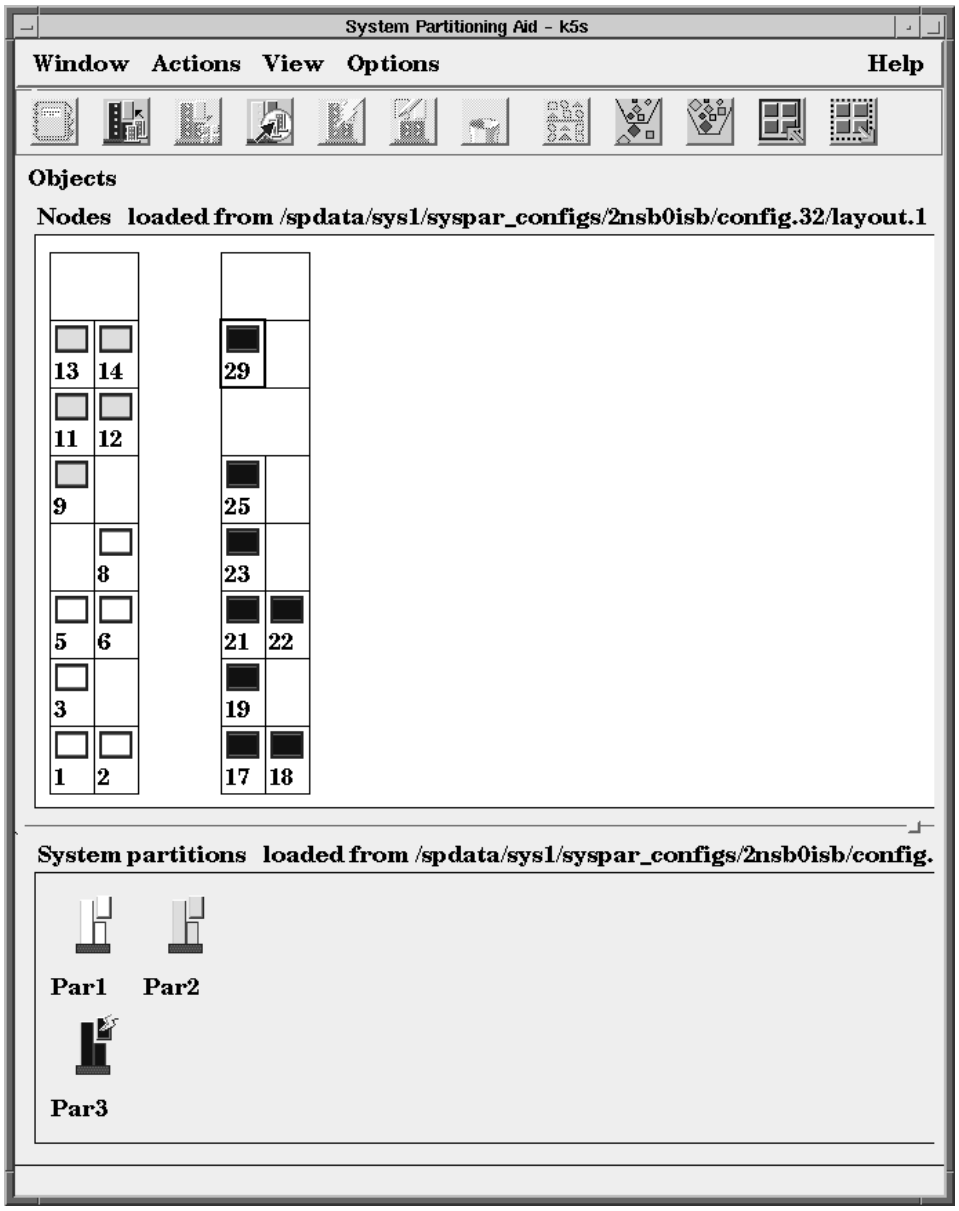


Figure 51. System partitioning for Example 3 of Chapter 5

Validate and save the new layout by clicking on the fourth Tool Bar icon, ″Generate files used to define system configuration.″ The resulting window appears in Figure 51. The code wants to store this new layout as an 8_8_16 configuration of a `2nsb0isb` system, which is correct. (If you remove the filter you applied earlier, you indeed see partitions of 8, 8 and 16 nodes.) You can choose the directory extension, (the example uses directory extension ″mine_1″). Therefore, the name of the directory containing the new layout is ″layout.mine_1″.

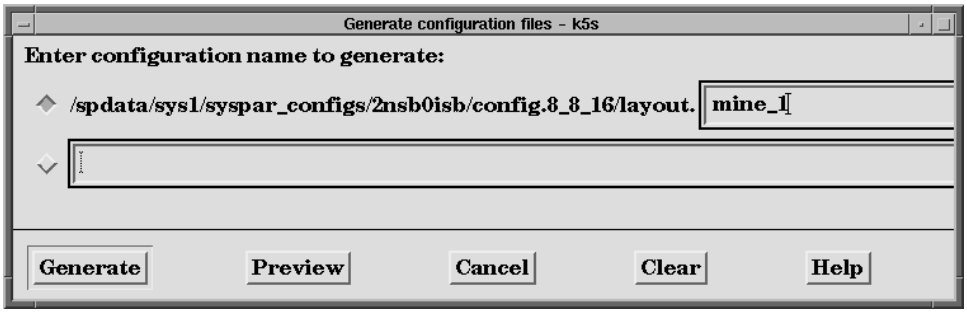


Figure 52. Dialog box for specifying name of new layout

Click on ″Generate″ and receive the message in the following figure. Note the warning about losing the configuration. You should backup the layouts you create before reinstalling PSSP or **ssp.top**.

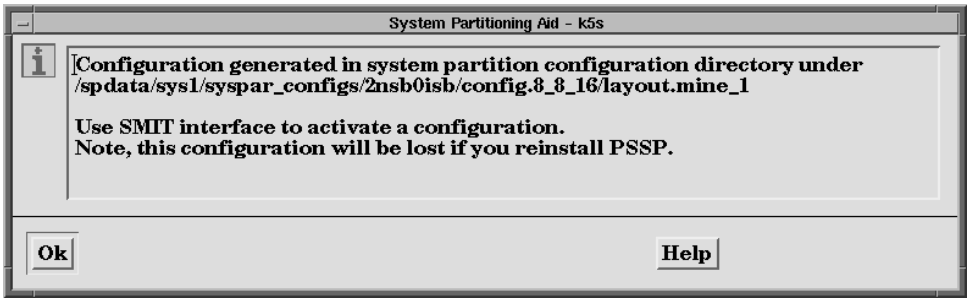


Figure 53. Message issued when new layout is saved

The CLI

Recall that the GUI (**spsyspar**) invokes the CLI (**sysparaid**) to validate and save a new layout. The previous GUI activity finished the job by issuing the command:

```
spsyspar  -s mine_1  inputfile
```

where `inputfile` is as shown in Figure 54 on page 255. (**spsyspar** chooses the correct global topology file based on the ″Number of Switches ...″ entries in this input file.)

```
Number of Nodes in System:  32
Number of Frames in System:  2
Frame Type:  tall
Switch Type:  SP
Number of Switches in Node Frames:  2
Number of Switches in Switch Only Frames:  0
Number of System Partitions:  3
Node Numbering Scheme:  switch_port_number
System Partition Name:  Par1
Number of Nodes in System Partition:  8
0-7
System Partition Name:  Par2
Number of Nodes in System Partition:  8
8 - 15
System Partition Name:  Par3
Number of Nodes in System Partition:  16
16 - 31
```

Figure 54. CLI input file from spsyspar

If you use the CLI directly, you can use an input file similar to that in Figure 54, but representing the facts more precisely:

- The system has 3 frames and 2 switches.
- Existing nodes in the bottom half of Frames 1 and 2 are in `Par1`.
- Existing nodes in the top of Frames 1 and 2 are in `Par2`.
- Existing nodes in Frame 3 are in `Par3`.

Figure 55 is the appropriate input file.

```
Number of Nodes in System:  19
Number of Frames in System:  3
Frame Type:  tall
Switch Type:  SP
Number of Switches in Node Frames:  2
Number of Switches in Switch Only Frames:  0
Number of System Partitions:  3
Node Numbering Scheme:  switch_port_number
System Partition Name:  Par1
Number of Nodes in System Partition:  6
0-2
4-6
System Partition Name:  Par2
Number of Nodes in System Partition:  5
8
10-13
System Partition Name:  Par3
Number of Nodes in System Partition:  8
16 - 18
20 - 22
24
28
```

Figure 55. Alternate CLI input file

Other files and data

When you save a new layout, supplemental files are saved in the respective directory. These include chip allocation files and performance files. For example, if you look at the **layout.mine_1** directory saved earlier, the **syspar.2.Par1** subdirectory contains the files **spa.snapshot** and **spa.metrics**.

The **spa.snapshot** file data is available for viewing in the GUI as the ″Chip Allocation″ page of the notebook for `Par1`. (First icon.) This GUI presentation is

produced in Figure 56. Par1 is completely contained in Frames 1 and 2 and so only uses Switch 1, denoted NSB 1 (Node Switch Board 1) in the **spa.snapshot** file. The 2 chips on the left are the node-attached chips, and the 2 chips on the right provide connectivity between those chips. A rule which **sysparaid** adheres to is any 2 node chips in a partition must have 2 link switch chips through which to communicate. This guarantees minimal, acceptable bandwidth and reliability characteristics.

A summary of the chip assignments for all partitions is stored in an **spa.snapshot** file in the **layout.mine_1** directory level.

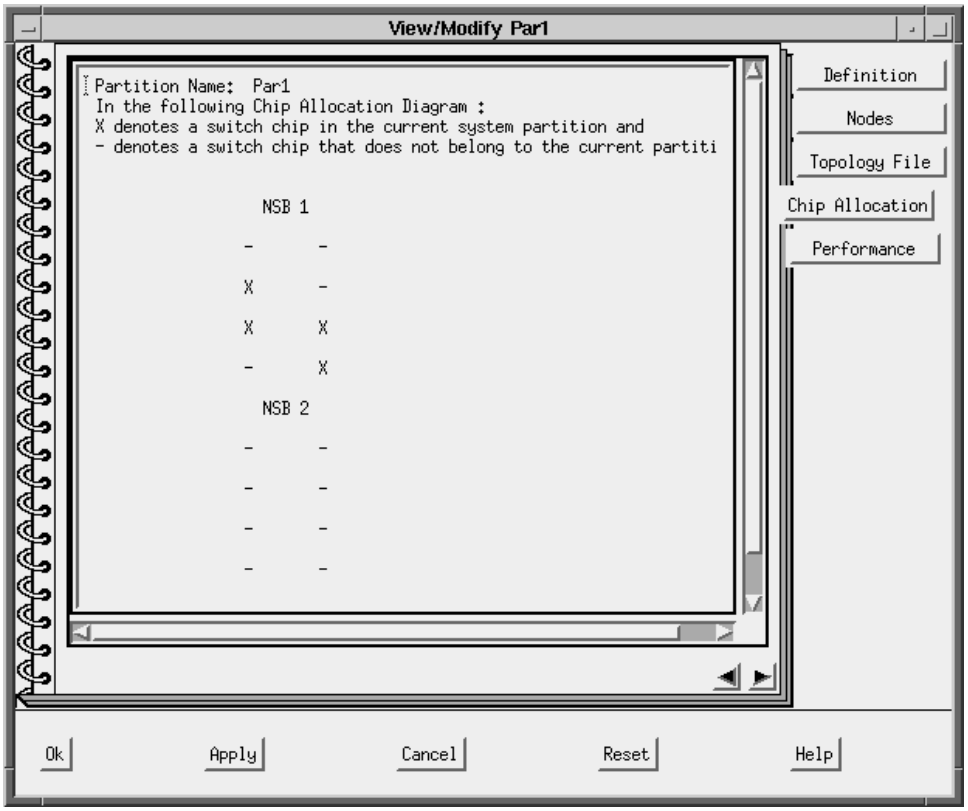


Figure 56. Switch chips allocated to system partition Par1

The **spa.metrics** data is available in the GUI on the ″Performance″ page of the notebook for Par1. This GUI presentation is given in Figure 57 on page 257. Chips 5 and 6 are the node chips of Figure 56. The bandwidth numbers for Par1 are less than 100%. This measure is a comparison to the unpartitioned case where all 4 link switch chips would be available for the nodes on chips 5 and 6 to communicate through. So, in some cases, total traffic throughput between nodes of Par1 is cut by as much as half from the unpartitioned case. On average, that communication is only cut to 87.5%, since some of the nodes are on the same chip.

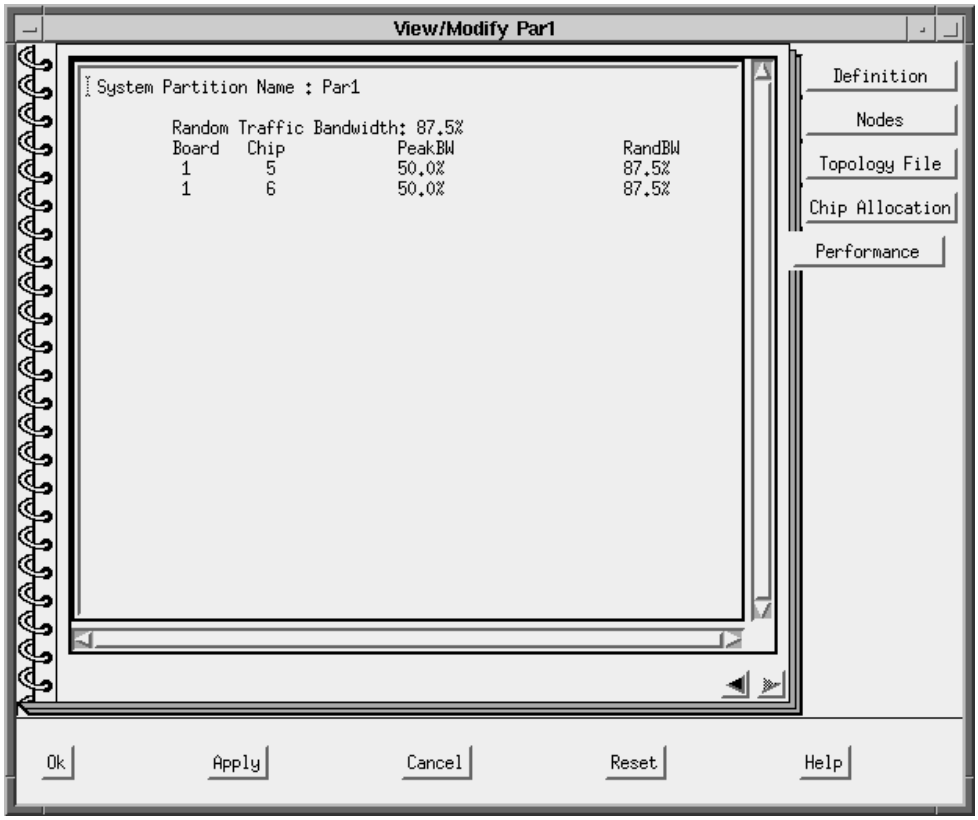


Figure 57. Performance numbers for system partition Par1

Appendix B. System Partitioning

This appendix contains a description for each of the system partitioning layouts, ordered by system size, that IBM provides. There are none for an SP system that uses the SP Switch2 or a system comprised of clustered enterprise servers since it supports only a single system partition configuration.

8 Switch Port System

Layout for 4_4 Partition of 8 Switch Port System with an SP Switch-8

This layout is the only layout choice for a 4_4 system partition configuration of an 8 switch port system with no intermediate switch boards.

Layout 1

This is the description of the only layout choice for a 4_4 system partition configuration of an 8 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5

Partition 2 contains switch_port_numbers: 2, 3, 6, 7

Layout for 8 Partition of 8 Switch Port System with an SP Switch-8

This layout is the only layout choice for an 8 system partition configuration of an 8 switch port system with no intermediate switch boards.

Layout 1

This is the description of the only layout choice for an 8 system partition configuration of an 8 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 7

16 Switch Port System

Layouts for 8_8 Partition of 16 Switch Port System

The following are the layout choices for an 8_8 system partition of a 16 switch port system with no intermediate switch boards:

Layout 1

This is the description of one of the layout choices for an 8_8 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5, 8, 9, 12, 13

Partition 2 contains switch_port_numbers: 2, 3, 6, 7, 10, 11, 14, 15

Layout 2

This is the description of the layout choices for an 8_8 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5, 10, 11, 14, 15

Partition 2 contains switch_port_numbers: 2, 3, 6 - 9, 12, 13

Layout 3

Partition 1 contains switch_port_numbers: 0 - 7

Partition 2 contains switch_port_numbers: 8 - 15

Layouts for 4_4_8 Partition of 16 Switch Port System

The following are the layout choices for a 4_4_8 system partition of a 16 switch port system with no intermediate switch boards.

Layout 1

This is the description of the layout choices for a 4_4_8 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5

Partition 2 contains switch_port_numbers: 8, 9, 12, 13

Partition 3 contains switch_port_numbers: 2, 3, 6, 7, 10, 11, 14, 15

Layout 2

This is the description of the layout choices for a 4_4_8 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5

Partition 2 contains switch_port_numbers: 2, 3, 6, 7

Partition 3 contains switch_port_numbers: 8 - 15

Layout 3

This is the description of the layout choices for a 4_4_8 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5

Partition 2 contains switch_port_numbers: 10, 11, 14, 15

Partition 3 contains switch_port_numbers: 2, 3, 6 - 9, 12, 13

Layout 4

This is the description of the layout choices for a 4_4_8 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 2, 3, 6, 7

Partition 2 contains switch_port_numbers: 8, 9, 12, 13

Partition 3 contains switch_port_numbers: 0, 1, 4, 5, 10, 11, 14, 15

Layout 5

Partition 1 contains switch_port_numbers: 2, 3, 6, 7

Partition 2 contains switch_port_numbers: 10, 11, 14, 15

Partition 3 contains switch_port_numbers: 0, 1, 4, 5, 8, 9, 12, 13

Layout 6

This is the description of the layout choices for a 4_4_8 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 8, 9, 12, 13

Partition 2 contains switch_port_numbers: 10, 11, 14, 15

Partition 3 contains switch_port_numbers: 0 - 7

Layouts for 4_12 Partition of 16 Switch Port System

The following are the layout choices for a 4_12 system partition of a 16 switch port system.

Layout 1

This is the description of the layout choices for a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5

Partition 2 contains switch_port_numbers: 2, 3, 6 - 15

Layout 2

This is the description of the layout choices for a 4_12 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 8, 9, 12, 13

Partition 2 contains switch_port_numbers: 0 - 7, 10, 11, 14, 15

Layout 3

This is the description of the layout choices for a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 2, 3, 6, 7

Partition 2 contains switch_port_numbers: 0, 1, 4, 5, 8 - 15

Layout 4

This is the description of the layout choices for a 4_12 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 10, 11, 14, 15

Partition 2 contains switch_port_numbers: 0 - 9, 12, 13

Layouts for 4_4_4_4 Partition of 16 Switch Port System

This layout is the only layout choice for a 4_4_4_4 system partition of a 16 switch port system.

Layout 1

This is the description of the layout choices for a 4_4_4_4 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5

Partition 2 contains switch_port_numbers: 8, 9, 12, 13

Partition 3 contains switch_port_numbers: 2, 3, 6, 7

Partition 4 contains switch_port_numbers: 10, 11, 14, 15

Layouts for 16 Partition of 16 Switch Port System

This layout is the only layout choice for a 16 system partition of an 16 switch port system.

Layout 1

This is the description of the layout choices for a 16 system partition configuration of a 16 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15

---

32 Switch Port System

Layouts for 8_24 Partition of 32 Switch Port System

The following are the layout choices for an 8_24 system partition of a 32 switch port system.

Layout 1

This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5, 8, 9, 12, 13

Partition 2 contains switch_port_numbers: 2, 3, 6, 7, 10, 11, 14 - 31

Layout 2

This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 2, 3, 6, 7, 10, 11, 14, 15

Partition 2 contains switch_port_numbers: 0, 1, 4, 5, 8, 9, 12, 13, 16 - 31

Layout 3

This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5, 10, 11, 14, 15

Partition 2 contains switch_port_numbers: 2, 3, 6 - 9, 12, 13, 16 - 31

Layout 4

This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 2, 3, 6 - 9, 12, 13

Partition 2 contains switch_port_numbers: 0, 1, 4, 5, 10, 11, 14 - 31

Layout 5

This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 8 - 15

Partition 2 contains switch_port_numbers: 0 - 7, 16 - 31

Layout 6
This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 7

Partition 2 contains switch_port_numbers: 8 - 31

Layout 7
This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16, 17, 20, 21, 24, 25, 28, 29

Partition 2 contains switch_port_numbers: 0 - 15, 18, 19, 22, 23, 26, 27, 30, 31

Layout 8
This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 18, 19, 22, 23, 26, 27, 30, 31

Partition 2 contains switch_port_numbers: 0 - 17, 20, 21, 24, 25, 28, 29

Layout 9
This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16, 17, 20, 21, 26, 27, 30, 31

Partition 2 contains switch_port_numbers: 0 - 15, 18, 19, 22 - 25, 28, 29

Layout 10
This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 18, 19, 22 - 25, 28, 29

Partition 2 contains switch_port_numbers: 0 - 17, 20, 21, 26, 27, 30, 31

Layout 11
This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 24 - 31

Partition 2 contains switch_port_numbers: 0 - 23

Layout 12
This is the description of the layout choices for an 8_24 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 23

Partition 2 contains switch_port_numbers: 0 - 15, 24 - 31

Layouts for 4_28 Partition of 32 Switch Port System

The following are the layout choices for a 4_28 system partition of a 32 switch port system.

Layout 1

This is the description of the layout choices for a 4_28 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0, 1, 4, 5

Partition 2 contains switch_port_numbers: 2, 3, 6 - 31

Layout 2

This is the description of the layout choices for a 4_28 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 8, 9, 12, 13

Partition 2 contains switch_port_numbers: 0 - 7, 10, 11, 14 - 31

Layout 3

This is the description of the layout choices for a 4_28 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 2, 3, 6, 7

Partition 2 contains switch_port_numbers: 0, 1, 4, 5, 8 - 31

Layout 4

This is the description of the layout choices for a 4_28 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 10, 11, 14, 15

Partition 2 contains switch_port_numbers: 0 - 9, 12, 13, 16 - 31

Layout 5

This is the description of the layout choices for a 4_28 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16, 17, 20, 21

Partition 2 contains switch_port_numbers: 0 - 15, 18, 19, 22 - 31

Layout 6

This is the description of the layout choices for a 4_28 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 24, 25, 28, 29

Partition 2 contains switch_port_numbers: 0 - 23, 26, 27, 30, 31

Layout 7

This is the description of the layout choices for a 4_28 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 18, 19, 22, 23

Partition 2 contains switch_port_numbers: 0 - 17, 20, 21, 24 - 31

Layout 8

This is the description of the layout choices for a 4_28 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 26, 27, 30, 31

Partition 2 contains switch_port_numbers: 0 - 25, 28, 29

Layouts for 16_16 Partition of 32 Switch Port System

This layout is the only layout choice for a 16_16 system partition of a 32 switch port system.

Layout 1

This is the description of the layout choices for a 16_16 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15

Partition 2 contains switch_port_numbers: 16 - 31

Layouts for 32 Partition of 32 Switch Port System

This layout is the only layout choice for a 32 system partition of a 32 switch port system.

Layout 1

This is the description of the layout choices for a 32 system partition configuration of a 32 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 31

---

48 Switch Port System

Layouts for 16_32 Partition of 48 Switch Port System

The following are the layout choices for a 16_32 system partition of a 48 switch port system.

Layout 1

This is the description of the layout choices for a 16_32 system partition configuration of a 48 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 31

Partition 2 contains switch_port_numbers: 32 - 47

Layout 2

This is the description of the layout choices for a 16_32 system partition configuration of a 48 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 32 - 47

Partition 2 contains switch_port_numbers: 16 - 31

Layout 3

This is the description of the layout choices for a 16_32 system partition configuration of a 48 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 47

Partition 2 contains switch_port_numbers: 0 - 15

Layouts for 48 Partition of 48 Switch Port System

This layout is the only layout choice for a 48 system partition of a 48 switch port system.

Layout 1

This is the description of the layout choices for a 48 system partition configuration of a 48 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 47

---

64 Switch Port System

Layouts for 16_48 Partition of 64 Switch Port System

The following are the layout choices for a 16_48 system partition of a 64 switch port system.

Layout 1

This is the description of the layout choices for a 16_48 system partition configuration of a 64 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 47

Partition 2 contains switch_port_numbers: 48 - 63

Layout 2

This is the description of the layout choices for a 16_48 system partition configuration of a 64 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 31, 48 - 63

Partition 2 contains switch_port_numbers: 32 - 47

Layout 3

This is the description of the layout choices for a 16_48 system partition configuration of a 64 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 32 - 63

Partition 2 contains switch_port_numbers: 16 - 31

Layout 4

This is the description of the layout choices for a 16_48 system partition configuration of a 64 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 63

Partition 2 contains switch_port_numbers: 0 - 15

Layouts for 32_32 Partition of 64 Switch Port System

The following are the layout choices for a 32_32 system partition of a 64 switch port system.

Layout 1

This is the description of the layout choices for a 32_32 system partition configuration of a 64 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 31

Partition 2 contains switch_port_numbers: 32 - 63

Layout 2

This is the description of the layout choices for a 32_32 system partition configuration of a 64 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 32 - 47

Partition 2 contains switch_port_numbers: 16 - 31, 48 - 63

Layout 3

This is the description of the layout choices for a 32_32 system partition configuration of a 64 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 48 - 63

Partition 2 contains switch_port_numbers: 16 - 47

Layouts for 64 Partition of 64 Switch Port System

This layout is the only layout choice for a 64 system partition of a 64 switch port system.

Layout 1

This is the description of the layout choices for a 64 system partition configuration of a 64 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63

---

80 Switch Port System With 0 Intermediate Switch Boards

Layouts for 16_64 Partition

The following are the layout choices for a 16_64 system partition of a 80 switch port system.

Layout 1

This is the description of the layout choices for a 16_64 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63

Partition 2 contains switch_port_numbers: 64 - 79

Layout 2

This is the description of the layout choices for a 16_64 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 47, 64 - 79

Partition 2 contains switch_port_numbers: 48 - 63

Layout 3
This is the description of the layout choices for a 16_64 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 31, 48 - 79

Partition 2 contains switch_port_numbers: 32 - 47

Layout 4
This is the description of the layout choices for a 16_64 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 32 - 79

Partition 2 contains switch_port_numbers: 16 - 31

Layout 5
This is the description of the layout choices for a 16_64 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 79

Partition 2 contains switch_port_numbers: 0 - 15

Layouts for 32_48 Partition
The following are the layout choices for a 32_48 system partition of an 80 switch port system.

Layout 1
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 31

Partition 2 contains switch_port_numbers: 32 - 79

Layout 2
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 32 - 47

Partition 2 contains switch_port_numbers: 16 - 31, 48 - 79

Layout 3
This is the description of the layout choices for a 32_48 system partition configuration of a 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 48 - 63

Partition 2 contains switch_port_numbers: 16 - 47, 64 - 79

Layout 4
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 64 - 79

Partition 2 contains switch_port_numbers: 16 - 63

Layout 5
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 47

Partition 2 contains switch_port_numbers: 0 - 15, 48 - 79

Layout 6
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 31, 48 - 63

Partition 2 contains switch_port_numbers: 0 - 15, 32 - 47, 64 - 79

Layout 7
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 31, 64 - 79

Partition 2 contains switch_port_numbers: 0 - 15, 32 - 63

Layout 8
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 32 - 63

Partition 2 contains switch_port_numbers: 0 - 31, 64 - 79

Layout 9
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 32 - 47

Partition 2 contains switch_port_numbers: 64 - 79

Layout 10
This is the description of the layout choices for a 32_48 system partition configuration of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 48 - 79

Partition 2 contains switch_port_numbers: 0 - 47

Layouts for 80 Partition

This layout is the only layout choice for an 80 system partition of an 80 switch port system.

Layout 1

This is the description of the layout choices for an 80 partition of an 80 switch port system with no intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 79

---

80 Switch Port System With Intermediate Switch Boards

Layouts for 16_16_48 Partition

The following are the layout choices for a 16_16_48 system partition of an 80 switch port system.

Layout 1

This is the description of the layout choices for a 16_16_48 partition of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15

Partition 2 contains switch_port_numbers: 16 - 63

Partition 3 contains switch_port_numbers: 64 - 79

Layout 2

This is the description of the layout choices for a 16_16_48 system partition configuration of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 31

Partition 2 contains switch_port_numbers: 0 - 15, 32 - 63

Partition 3 contains switch_port_numbers: 64 - 79

Layout 3

This is the description of the layout choices for a 16_16_48 partition of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 32 - 47

Partition 2 contains switch_port_numbers: 0 - 31, 48 - 63

Partition 3 contains switch_port_numbers: 64 - 79

Layout 4

This is the description of the layout choices for a 16_16_48 partition of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 48 - 63

Partition 2 contains switch_port_numbers: 0 - 47

Partition 3 contains switch_port_numbers: 64 - 79

Layouts for 16_64 Partition

The following are the layout choices for a 16_64 system partition of an 80 switch port system.

Layout 1

This is the description of the layout choices for a 16_64 partition of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 64 - 79

Partition 2 contains switch_port_numbers: 0 - 63

Layout 2

This is the description of the layout choices for a 16_64 partition of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 48 - 63

Partition 2 contains switch_port_numbers: 0 - 47, 64 - 79

Layout 3

This is the description of the layout choices for a 16_64 partition of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 32 - 47

Partition 2 contains switch_port_numbers: 0 - 31, 48 - 79

Layout 4

This is the description of the layout choices for a 16_64 partition of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 31

Partition 2 contains switch_port_numbers: 0 - 15, 32 - 79

Layout 5

This is the description of the layout choices for a 16_64 partition of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15

Partition 2 contains switch_port_numbers: 16 - 79

Layouts for 80 Partition

This layout is the only layout choice for an 80 system partition of an 80 switch port system.

Layout 1

This is the description of the layout choices for an 80 system partition configuration of an 80 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 79

96 Switch Port System

Layouts for 32_64 Partition

This layout is the only layout choice for a 32_64 system partition of a 96 switch port system.

Layout 1

This is the description of the layout choices for a 32_64 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 64 - 95

Partition 2 contains switch_port_numbers: 0 - 63

Layouts for 16_32_48 Partition

The following are the layout choices for a 16_32_48 system partition of a 96 switch port system.

Layout 1

This is the description of the layout choices for a 16_32_48 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15

Partition 2 contains switch_port_numbers: 16 - 63

Partition 3 contains switch_port_numbers: 64 - 95

Layout 2

This is the description of the layout choices for a 16_32_48 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 31

Partition 2 contains switch_port_numbers: 0 - 15, 32 - 63

Partition 3 contains switch_port_numbers: 64 - 95

Layout 3

This is the description of the layout choices for a 16_32_48 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 32 -47

Partition 2 contains switch_port_numbers: 0 - 31, 48 - 63

Partition 3 contains switch_port_numbers: 64 - 95

Layout 4

This is the description of the layout choices for a 16_32_48 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 48 - 63

Partition 2 contains switch_port_numbers: 0 - 47

Partition 3 contains switch_port_numbers: 64 - 95

Layouts for 16_80 Partition

The following are the layout choices for a 16_80 system partition of a 96 switch port system.

Layout 1

This is the description of the layout choices for a 16_80 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15

Partition 2 contains switch_port_numbers: 16 - 95

Layout 2

This is the description of the layout choices for a 16_80 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 31

Partition 2 contains switch_port_numbers: 0 - 15, 32 - 95

Layout 3

This is the description of the layout choices for a 16_80 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 32 - 47

Partition 2 contains switch_port_numbers: 0 - 31, 48 - 95

Layout 4

This is the description of the layout choices for a 16_80 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 48 - 63

Partition 2 contains switch_port_numbers: 0 - 47, 64 - 95

Layout 5

This is the description of the layout choices for a 16_80 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 64 - 79

Partition 2 contains switch_port_numbers: 0 - 63, 80 - 95

Layout 6

This is the description of the layout choices for a 16_80 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 80 - 95

Partition 2 contains switch_port_numbers: 0 - 79

Layouts for 96 Partition

This layout is the only layout choice for a 96 system partition of a 96 switch port system.

Layout 1

This is the description of the layout choices for a 96 system partition configuration of a 96 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 95

---

112 Switch Port System

Layouts for 48_64 Partition

This layout is the only layout choice for a 48_64 system partition of a 112 switch port system.

Layout 1

This is the description of the layout choices for breakup is one of the layout choices for a 48_64 partition with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63

Partition 2 contains switch_port_numbers: 64 - 111

Layouts for 16_48_48 Partition

The following are the layout choices for a 16_48_48 system partition of a 112 switch port system.

Layout 1

This is the description of the layout choices for a 16_48_48 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 63

Partition 2 contains switch_port_numbers: 64 - 111

Partition 3 contains switch_port_numbers: 0 - 15

Layout 2

This is the description of the layout choices for a 16_48_48 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains node slots: 0 - 15, 32 - 63

Partition 2 contains switch_port_numbers: 64 - 111

Partition 3 contains switch_port_numbers: 16 - 31

Layout 3

This is the description of the layout choices for a 16_48_48 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 31, 48 - 63

Partition 2 contains switch_port_numbers: 64 - 111

Partition 3 contains switch_port_numbers: 32 - 47

Layout 4

This is the description of the layout choices for a 16_48_48 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 47

Partition 2 contains switch_port_numbers: 64 - 111

Partition 3 contains switch_port_numbers: 48 - 63

Layouts for 16_96 Partition

The following are the layout choices for a 16_96 system partition of a 112 switch port system.

Layout 1

This is the description of the layout choices for a 16_96 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15

Partition 2 contains switch_port_numbers: 16 - 111

Layout 2

This is the description of the layout choices for a 16_96 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 31

Partition 2 contains switch_port_numbers: 0 - 15, 32 - 111

Layout 3

This is the description of the layout choices for a 16_96 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 32 - 47

Partition 2 contains switch_port_numbers: 0 - 31, 48 - 111

Layout 4

This is the description of the layout choices for a 16_96 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 48 - 63

Partition 2 contains switch_port_numbers: 0 - 47, 64 - 111

Layout 5

This is the description of the layout choices for a 16_96 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 64 - 79

Partition 2 contains switch_port_numbers: 0 - 63, 80 - 111

Layout 6

This is the description of the layout choices for a 16_96 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 80 - 95

Partition 2 contains switch_port_numbers: 0 - 79, 96 - 111

Layout 7
This is the description of the layout choices for a 16_96 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 96 - 111

Partition 2 contains switch_port_numbers: 0 - 95

Layouts for 112 Partition
This layout is the only layout choice for a 112 system partition of a 112 switch port system.

Layout 1
This is the description of the layout choices for a 112 system partition configuration of a 112 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 111

---

128 Switch Port System

Layouts for 16_48_64 Partition
The following are the layout choices for a 16_48_64 system partition of a 128 switch port system.

Layout 1
This is the description of the layout choices for a 16_48_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 64 - 127

Partition 2 contains switch_port_numbers: 16 - 63

Partition 3 contains switch_port_numbers: 0 - 15

Layout 2
This is the description of the layout choices for a 16_48_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 64 - 127

Partition 2 contains switch_port_numbers: 0 - 15, 32 - 63

Partition 3 contains switch_port_numbers: 16 - 31

Layout 3
This is the description of the layout choices for a 16_48_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 64 - 127

Partition 2 contains switch_port_numbers: 0 - 31, 48 - 63

Partition 3 contains switch_port_numbers: 32 - 47

Layout 4
This is the description of the layout choices for a 16_48_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 64 - 127

Partition 2 contains switch_port_numbers: 0 - 47

Partition 3 contains switch_port_numbers: 48 - 63

Layout 5

This is the description of the layout choices for a 16_48_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63

Partition 2 contains switch_port_numbers: 80 - 127

Partition 3 contains switch_port_numbers: 64 - 79

Layout 6

This is the description of the layout choices for a 16_48_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63

Partition 2 contains switch_port_numbers: 64 - 79, 96 - 127

Partition 3 contains switch_port_numbers: 80 - 95

Layout 7

This is the description of the layout choices for a 16_48_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63

Partition 2 contains switch_port_numbers: 64 - 95, 112 - 127

Partition 3 contains switch_port_numbers: 96 - 111

Layout 8

This is the description of the layout choices for a 16_48_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63

Partition 2 contains switch_port_numbers: 64 - 111

Partition 3 contains switch_port_numbers: 112 - 127

Layouts for 16_112 Partition

The following are the layout choices for a 16_112 system partition of a 128 switch port system.

Layout 1

This is the description of the layout choices for a 16_112 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 16 - 127

Partition 2 contains switch_port_numbers: 0 - 15

Layout 2

This is the description of the layout choices for a 16_112 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 15, 32 - 127

Partition 2 contains switch_port_numbers: 16 - 31

Layout 3

This is the description of the layout choices for a 16_112 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 31, 48 - 127

Partition 2 contains switch_port_numbers: 32 - 47

Layout 4

This is the description of the layout choices for a 16_112 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 47, 64 - 127

Partition 2 contains switch_port_numbers: 48 - 63

Layout 5

This is the description of the layout choices for a 16_112 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63, 80 - 127

Partition 2 contains switch_port_numbers: 64 - 79

Layout 6

This is the description of the layout choices for a 16_112 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 79, 96 - 127

Partition 2 contains switch_port_numbers: 80 - 95

Layout 7

This is the description of the layout choices for a 16_112 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 95, 112 - 127

Partition 2 contains switch_port_numbers: 96 - 111

Layout 8

This is the description of the layout choices for a 16_112 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 111

Partition 2 contains switch_port_numbers: 112 - 127

Layouts for 64_64 Partition

This layout is the only layout choice for a 64_64 system partition of an 128 switch port system.

Layout 1

This is the description of the layout choices for a 64_64 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 63

Partition 2 contains switch_port_numbers: 64 - 127

Layouts for 128 Partition

This layout is the only layout choice for a 128 system partition of a 128 switch port system.

Layout 1

This is the description of the layout choices for a 128 system partition configuration of a 128 switch port system with 4 intermediate switch boards.

Partition 1 contains switch_port_numbers: 0 - 127

Appendix C. SP system planning worksheets

This chapter contains the following SP system planning worksheets:

Table 59. List of SP planning worksheets

Make copies of these worksheets as required. Instructions for using the worksheets are contained in Chapter 2, “Defining the system that fits your needs” on page 19, in Chapter 3, “Defining the configuration that fits your needs” on page 89, and in Chapter 6, “Planning for security” on page 145.

Table 60. Preliminary list of applications

Preliminary list of applications - Worksheet 1		
Application	**Parallel**	**Need switch?**
Enter the name of each application you plan to install. Use ***y*** if the application is to use parallel processing or a switch, ***n*** where it will not, ***?*** if you do not yet know.		

Table 61. IBM licensed programs to order

IBM licensed programs to order – Worksheet 2			
Order		**Program number**	**Level for AIX 5L 5.1**
	IBM VisualAge C++ Professional (batch C and C++)	41L3180	5.0.2 or later
	IBM DCE	5765-D17	
	IBM DCE Base Services (6693, 41L2819) and Servers (6688, 41L2813)	5801-AAR	
	IBM Parallel System Support Programs (PSSP)	5765-D51	3.5
	IBM Parallel Environment	5765-D93	3.2
	IBM Parallel Engineering and Scientific Subroutine Library (Parallel ESSL)	5765-C41	2.3
	IBM ESSL	5765-C42	3.3
	IBM High Availability Cluster Multi-Processing (HACMP features HAS, CRM, ES, ESCRM)	5765-E54	4.5.0
	IBM HACMP features HAGEO or GeoRM	5765-E82	2.3
	IBM LoadLeveler	5765-E69	3.1
		5765-D61	
	IBM General Parallel File System	5765-B95	1.5

Table 62. External disk storage

External disk storage – Worksheet 3			
Disk subsystem	**Adapters (# - type)**	**Number of disks**	**Disk size**

Table 63. Major system hardware components

Major system hardware components – Worksheet 4			
Company name:			
Customer number:			**Date:**
Customer contact:			**Phone:**
IBM contact:			**Phone:**
Complete the following by entering quantities to order...			
Frames	**Nodes**	**Attached Servers**	
550 (tall):	375/450 MHz Thin:	p680:	
1550 (tall):	375/450 MHz Wide:	p660 6H1:	
500 (short):	375 MHz High:	p660 6M1:	
1500 (short):		p660 6H0:	
		p690 (number LPARs:)	
		p670 (number LPARs:)	
Switch subsystem components			
SP Switch2:	SP Switch2 Adapter:	SP Switch2 PCI-X Attachment Adapter:	
	SP Switch2 MX2 Adapter:	SP Switch2 PCI Attachment Adapter:	
SP Switch 16-port:	SP Switch MX Adapter:	SP Switch MX2 Adapter:	
SP Switch 8-port:	SP Switch Adapter:	RS/6000 SP System Attachment Adapter:	
SP Switch Router:	SP Switch Router Adapter:		
External storage units:	*Type*	*Quantity*	
Network media cards:	*Type*	*Quantity*	
Fill in after you place your order...			
Cluster model number:		Purchase order number:	
Cluster serial number:		SP model number:	
Control workstation:		SP serial number:	
Peripherals:			

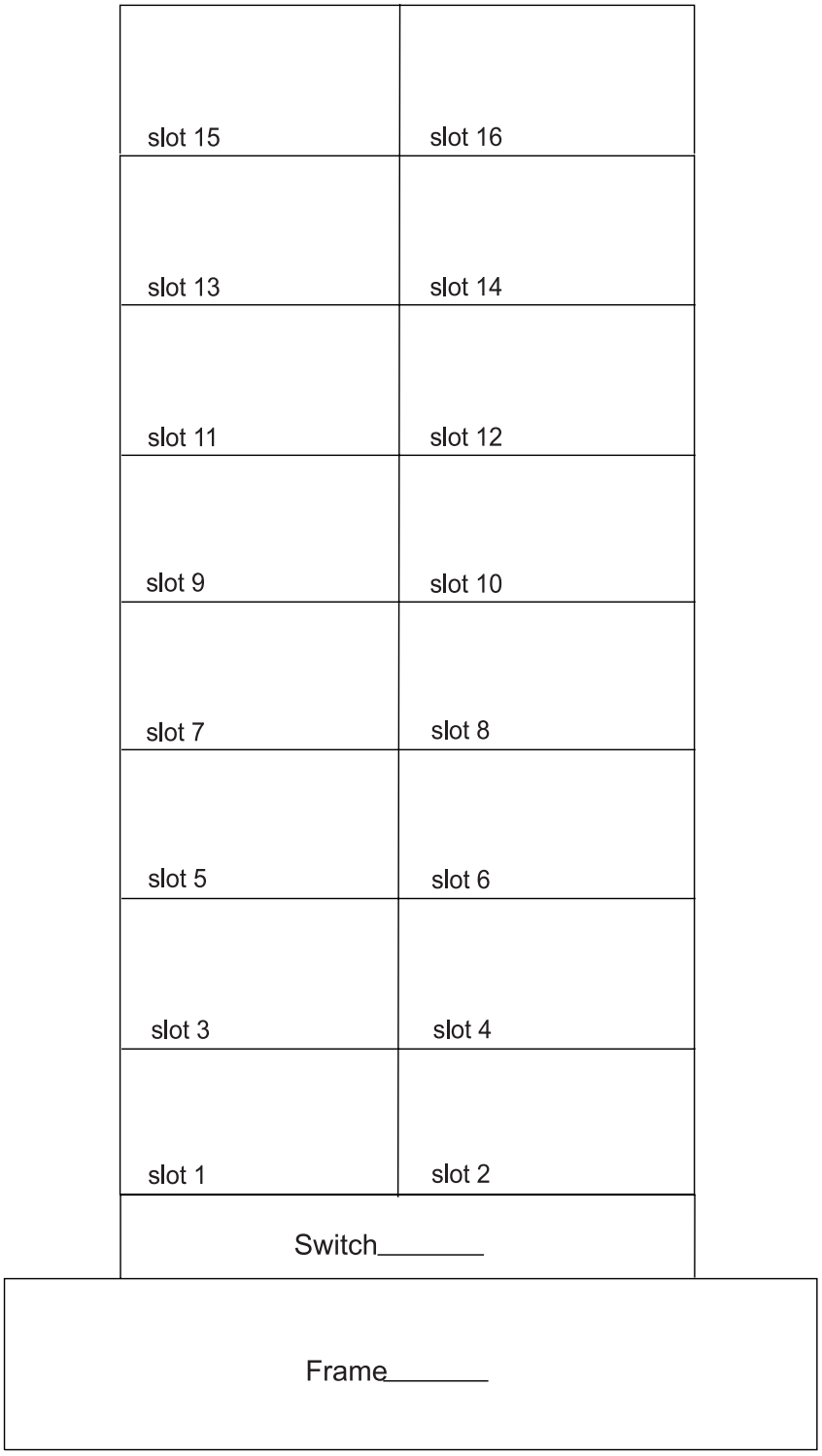


Figure 58. Node layout – Worksheet 5

Table 64. Hardware configuration by frame

Hardware configuration by frame – Worksheet 7					
Frame number:		**Hardware protocol:**		**Switch number:**	
p690/p670 name:			**HMC hardware monitor user id:**		
HMC trusted network adapter names or IP addresses:					
Slot or **LPAR** or **frame**	**Node** or **Expansion number**	**Node type** or **Associated frame/slot**	**Number processors, memory**	**Internal disk**	**Additional adapters**
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
Note: For p690 or p670 server name, use the name you assigned to the server with the HMC user interface.					

Table 65. PSSP admin LAN

PSSP admin LAN configuration – Worksheet 8				
Company name:			**Date:**	
Frame number:			**p690/p670 server name:**	
Associated node slot or **frame**	**Admin LAN netmask:** *(must be en0 unless p690/p670)*			**Default route**
	Hostname	**Adapter name** or **physical location**	**IP Address**	
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				

Notes:

1. AIX is case sensitive. If name-to-address resolution is provided by DNS, NIS or some other means, the names in the SDR must match exactly. Otherwise, use lower case for the host name and addresses.
2. Wide nodes occupy two slots and use the *odd-numbered* slot.
3. High nodes occupy four slots (2 drawers) and use the lowest odd-numbered slot.
4. Use adapter name or physical location for p690 and p670 nodes only. All other nodes must use en0.
5. For p690 or p670 server name, use the name you assigned to the server with the HMC user interface.

Table 66. Additional adapters node network configuration

Additional adapters node network configuration – Worksheet 9				
Company name:			**Date:**	
Frame number:			**p690/p670 server name:**	
Token ring speed:				
Associated node slot or **frame**	**Additional adapters netmask:**			**Default route**
	Adapter name or **physical location**	**Hostname**	**IP address**	
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				

Notes:

1. AIX is case sensitive. If name-to-address resolution is provided by DNS, NIS or some other means, the names in the SDR must match exactly. Otherwise, use lower case for the host name and addresses.
2. Wide nodes occupy two frame slots and use the *odd-numbered* slot.
3. High nodes occupy four frame slots (2 drawers) and use the lowest odd-numbered slot.
4. Use adapter physical location for p690 and p670 nodes only. All other nodes must use the adapter name.
5. For p690 or p670 server name, use the name you assigned to the server with the HMC user interface.

Table 67. Switch configuration worksheet

Switch configuration – Worksheet 10							
Frame number:	**Switch number:**	**css0 netmask:**		**css1 netmask:**		**ml0 netmask:**	
Slot number	**Switch port number**	**css0 hostname**	**css0 IP address**	**css1 hostname**	**css1 IP address**	**ml0 hostname**	**ml0 IP address**
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
Note: Switch port number is necessary only with the SP Switch. Use of **css1** and **ml0** are options with the SP Switch2 only.							

Table 68. Specifying the system images (SPIMG)

Specifying system images – Worksheet 11	
System image name	
AIX level	
Partition number	
Install on node numbers	
Specify *internal* disks where you want to install rootvg	
Check here if you want only the SPIMG minimal image of AIX ____	
IBM licensed programs	
	AIX
	PSSP
Additional AIX software	
Other applications	

Table 69. File set list for PSSP 3.5

PSSP 3.5 file sets – Worksheet 12		
System image name `spimg1`		
	File set	**Description**
	Image of AIX: spimg	
	bos.obj.ssp.510	File with mksysb image of minimal AIX 5L 5.1 system with 32-bit kernel
	bos.obj.ssp.510_64	File with mksysb image of minimal AIX 5L 5.1 system with 64-bit kernel
	PSSP image: ssp Base components of PSSP	
	ssp.authent	SP Kerberos V4 Server
	ssp.basic	SP System Support Package
	ssp.cediag	SP CE Diagnostics
	ssp.clients	SP Client Programs
	ssp.css	SP Communication Subsystem Package
	ssp.docs	SP man pages, PDF files, and HTML files
	ssp.gui	SP System Monitor Graphical User Interface
	ssp.ha_topsvcs.compat	Compatibility for ssp.ha and ssp.topsvcs clients
	ssp.perlpkg	SP PERL distribution package
	ssp.pman	SP Problem Management
	ssp.public	Public Code compressed tarfiles
	ssp.spmgr	SP Extension Node SNMP Manager
	ssp.st	Switch Table API package
	ssp.sysctl	SP sysctl package
	ssp.sysman	Optional System Management programs
	ssp.tecad	SP HA TEC Event Adapter package
	ssp.top	SP Communication Subsystem Topology package
	ssp.top.gui	SP System Partitioning Aid
	ssp.ucode	SP Supervisor microcode package
	PSSP image: ssp.hacws Optional component of PSSP	
	ssp.hacws	SP High Availability Control Workstation
	PSSP image: vsd Components for managing IBM Virtual Shared Disks	
	vsd.cmi	IBM Virtual Shared Disk Centralized Management Interface (SMIT)
	vsd.hsd	Hashed Shared Disk data striping device driver
	vsd.rvsd.hc	IBM Recoverable Virtual Shared Disk Connection Manager
	vsd.rvsd.rvsdd	IBM Recoverable Virtual Shared Disk daemon
	vsd.rvsd.scripts	IBM Recoverable Virtual Shared Disk recovery scripts
	vsd.sysctl	IBM Virtual Shared Disk sysctl commands
	vsd.vsdd	IBM Virtual Shared Disk device driver
	PSSP image: ssp.vsdgui IBM Virtual Shared Disk Perspectives GUI	
	PSSP image: ssp.resctr Resource Center with links to online publications and other information.	
	ssp.resctr.rte	Cluster Resource Center
	PSSP image: ssp.en_US.* US English ISO8859-1, ISO8859-15	
	ssp.msg.en_US.authent	SP Authentication Server Messages

Table 69. File set list for PSSP 3.5 (continued)

PSSP 3.5 file sets – Worksheet 12		
System image name `spimg1`		
	File set	**Description**
	ssp.msg.en_US.basic	SP System Support Messages
	ssp.msg.en_US.cediag	SP CE Diagnostic Messages
	ssp.msg.en_US.clients	SP Authenticated Client Messages
	ssp.msg.en_US.pman	SP Problem Management Messages
	ssp.msg.en_US.spmgr	SP Extension Node Manager Messages
	ssp.msg.en_US.sysctl	SP Package Messages
	ssp.msg.en_US.sysman	Optional System Management Messages
	PSSP image: ssp.En_US.* US English IBM-850	
	ssp.msg.En_US.authent	SP Authentication Server Messages
	ssp.msg.En_US.basic	SP System Support Messages
	ssp.msg.En_US.cediag	SP CE Diagnostic Messages
	ssp.msg.En_US.clients	SP Authenticated Client Messages
	ssp.msg.En_US.pman	SP Problem Management Messages
	ssp.msg.En_US.spmgr	SP Extension Node Manager Messages
	ssp.msg.En_US.sysctl	SP Package Messages
	ssp.msg.En_US.sysman	Optional System Management Messages
Note: You can choose to install complete images or only selected file sets. Keep in mind that some optional components require others. See the respective planning and migration sections in this book for dependencies. For information on which PSSP file sets to install on the control workstation and which to install on a node, see the books *PSSP: Installation and Migration Guide* and *PSSP: Managing Shared Disks*.		

Table 70. Control workstation worksheet

Control workstation - Worksheet 13	
Control workstation image:	
Control Workstation Name	
Model	
Install rootvg on disk	
Disk Space	
Memory Size	
Hardware options and adapters:	
Type	***Quantity***
ATM	
Ethernet	
FDDI	
Token ring (speed 16Mbps)	
Multiport Serial Adapters	
8 mm tape drive	
CD-ROM	
IBM licensed programs:	
Other applications:	

Table 71. Time zones

Select a time zone - Worksheet 14				
Select	**Time zone**	**Select**	**Time zone**	**Description**
	(CUT0)		(CUT0GDT)	Coordinated Universal Time (CUT)
	(GMT0)		(GMT0BST)	United Kingdom (CUT)
	(AZOREST1)		(AZOREST1AZOREDT)	Azores; Cape Verde (CUT -1)
	(FALKST2)		(FALKST2FALKDT)	Falkland Islands (CUT -2)
	(GRNLNDST3)		(GRNLNDST3GRNLNDDT)	Greenland; East Brazil (CUT -3)
	(AST4)		(AST4ADT)	Central Brazil (CUT -4)
	(EST5)		(EST5EDT)	Eastern U.S.; Colombia (CUT -5)
	(CST6)		(CST6CDT)	Central U.S.; Honduras (CUT -6)
	(MST7)		(MST7MDT)	Mountain U.S. (CUT -7)
	(PST8)		(PST8PDT)	Pacific U.S.; Yukon (CUT -8)
	(AST9)		(AST9ADT)	Alaska (CUT -9)
	(HST10)		(HST10HDT)	Hawaii; Aleutian (CUT-10)
	(BST11)		(BST11BDT)	Bering Straits (CUT-11)
	(NZST-12)		(NZST-12NZDT)	New Zealand (CUT+12)
	(MET-11M)		(MET-11METDT)	Solomon Islands (CUT+11)
	(EET-10E)		(EET-10EETDT)	Eastern Australia (CUT+10)
	(JST-9)		(JST-9JDT)	Japan (CUT +9)
	(KORST-9)		(KORST-9KORDT)	Korea (CUT +9)
	(WAUST-8)		(WAUST-8WAUDT)	Western Australia (CUT +8)
	(TAIST-8)		(TAIST-8TAIDT)	Taiwan (CUT +8)
	(THAIST-7)		(THAIST-7THAIDT)	Thailand (CUT +7)
	(TASHST-6)		(TASHST-6TASHDT)	Tashkent; Central Asia (CUT +6)
	(PAKST-5)		(PAKST-5PAKDT)	Pakistan (CUT +5)
	(WST-4)		(WST-4WDT)	Gorki; Central Asia; Oman (CUT +4)
	(MEST-3)		(MEST-3MEDT)	Turkey (CUT +3)
	(SAUST-3)		(SAUST-3SAUDT)	Saudi Arabia (CUT +3)
	(WET-2)		(WET-2WET)	Finland (CUT +2)
	(USAST-2)		(USAST-2USADT)	South Africa (CUT +2)
	(NFT-1)		(NFT-1DFT)	Norway; France (CUT +1)

Table 72. Control workstation connections worksheet

Control workstation connections - Worksheet 15							
Company name:				**Date:**			
System name:				**Control workstation name:**			
Frame hardware control connections				**Control workstation network connections**			
Frame number	**Serial port for RS233 control line**	**tty device**		**Adapter**	**Hostname**	**IP address**	**Netmask**

Note:

Column 2 applies to nodes with CSP or SAMI hardware protocol. Record one serial port for nodes with CSP and two for nodes with SAMI.

Servers with the HMC protocol require a network connection only. Columns 2 and 3 do not apply.

Table 73. Site environment worksheet

Site environment - Worksheet 16			
Company name:		**Date:**	
System name:		**Control workstation name:**	
SMIT dialog field name (1)	**Site attribute (2)**	**Default value**	**Your choice**
Default Network Install Image	install_image	bos.obj.ssp.510	
Remove Install Image After Installs	remove_image	false	
NTP Installation	ntp_config	consensus	
NTP Server Hostname	ntp_server		
NTP Version	ntp_version	3	
Automounter Configuration	amd_config	true	
User Administration Interface	usermgmt_config	true	
Password File Server Hostname	passwd_file_loc	CWS hostname	
Password File	passwd_file	/etc/passwd	
Home Directory Server Hostname	homedir_server	CWS hostname	
Home Directory Path	homedir_path	/home/*cws*	
File Collection Management	filecoll_config	true	
File Collection daemon uid	supman_uid	102	
File Collection port	supfilesrv_port	8431	
SP Accounting Enabled	spacct_enable	false	
SP Accounting Active Node Threshold	spacct_node	80	
SP Exclusive Use Accounting Enabled	spacct_exclusive_enable	false	
Accounting Master	acct_master	0	
CWS LPP source name	cw_lppsource_name	default (3)	
SP administrative locale	admin_locale	AIX locale of CWS	
SDR may contain ASCII data only	SDR_ASCII_only	true	
Root remote command access restricted	restrict_root_rcmd	false	
Remote command method	rcmd_pgm	rsh	
Remote command executable	dsh_remote_cmd	`/bin/rsh` with `rsh`. `/bin/ssh` with `secrshell`.	
Remote copy executable	remote_copy_cmd	`/bin/rcp`with `rsh`. `/bin/scp` with `secrshell`.	
SP Model Number	SP_type_model		
SP Serial Number	SP_serial_number		
Force non-partitionable	force_non_partitionable	false	
Cluster machine type and model	Cluster_mtm		
Cluster machine serial number	Cluster_ms		

Notes:
1. This is the name that appears on the SMIT dialog.
2. This is the attribute name to use on the **spsitenv** command.
3. Make sure the AIX level of the LPP source (indicated by this value) matches the AIX level installed on the control workstation.

Table 74. Authentication planning worksheet

Authentication planning – Worksheet 17			
Syspar:	**DCE or Kerberos V5**	**Compatibility or Kerberos V4**	**Standard AIX**
1. Security services to install			N/A
2a. AIX remote command authorization methods for the root user			
2b. Authentication methods for AIX remote commands			
3. Authentication methods for SP trusted services			N/A
Note: The rows are ordered to correspond with the related discussion in “Choosing authentication options” on page 145.			

Table 75. DCE authentication

DCE authentication – Worksheet 18	
DCE cell name	
Cell admin user name	
Cell admin password	
Master security server hostname	
Replica security server hostnames	
Master CDS server hostname	
Replica CDS server hostname	
Names of network interfaces to exclude	
Lan Profile id	
Note: After you have written the password on the worksheet, be sure to keep the worksheet in a secure environment.	

Table 76. PSSP Kerberos V4 or other Kerberos authentication servers

Kerberos authentication servers – Worksheet 19			
	Hostname (long)	**Default realm**	**Control workstation**
Primary server			
Secondary servers			
Client systems			

Note:

hostname
Fully qualified hostname. For example, `kgn.east.abc.com`

default realm
Domain portion of hostname in upper case. For example, `EAST.ABC.COM`

control workstation?

y This workstation is the SP control workstation for the system being installed.

n This workstation is *not* the SP control workstation for the system being installed.

Any of the secondary servers or client systems could be control workstations for other SP systems, but enter `y` only for this system's control workstation.

Table 77. PSSP Kerberos V4 or other Kerberos local realm information

Kerberos realm information – Worksheet 20		
	Name	**Password**
Local realm		(Master)
Administrative principal	.admin	
other principals		

Note:

local realm
If you leave the name blank, the local realm is the default realm you entered for the primary server. The password in this row is the **master** password of the primary authentication server using SP authentication.

administrative principal
The name you will use as the primary administrator of the authentication database. This includes the .admin suffix.

password
After you have written these passwords on the worksheet, be sure to keep the worksheet in a secure environment.

Table 78. AFS authentication server

AFS authentication server – Worksheet 21	
Administrative principal	
Password	
Directory containing CellServDB, ThisCell files	
Directory containing kas command	

Note: After you have written the password on the worksheet, be sure to keep the worksheet in a secure environment.

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

IBM World Trade Asia Corporation
Licensing
2-31 Roppongi 3-chome, Minato-ku
Tokyo 106, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

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

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

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

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
Intellectual Property Law
2455 South Road, P386
Poughkeepsie, New York 12601
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

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

COPYRIGHT LICENSE:

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

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

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

AFS
AIX
AIX 5L
AIXwindows
DB2
DFS
Enterprise Storage Server
ESCON
@server
e (logo) Server
IBM
IBM (logo)

IBMLink
LoadLeveler
MVS
NetView
POWERparallel
pSeries
Redbooks
RS/6000
Service Director
SP
System/390
TURBOWAYS
VisualAge

Intel is a trademark of Intel Corporation in the United States, other countries, or both.

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

Lotus Notes is a registered trademarks of Lotus Development Corporation.

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

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

Publicly available software

PSSP includes software that is publicly available:

expect
Programmed dialogue with interactive programs

Perl Practical Extraction and Report Language

SUP Software Update Protocol

Tcl Tool Command Language

TclX Tool Command Language Extended

Tk Tcl-based Tool Kit for X-windows

This book discusses the use of these products only as they apply specifically to the RS/6000 SP system. The distribution for these products includes the source code and associated documentation. **/usr/lpp/ssp/public** contains the compressed **tar** files of the publicly available software. (IBM has made minor modifications to the versions of Tcl and Tk used in the SP system to improve their security characteristics. Therefore, the IBM-supplied versions do not match exactly the versions you may build from the compressed **tar** files.) All copyright notices in the documentation must be respected. You can find version and distribution information for each of these products that are part of your selected install options in the **/usr/lpp/ssp/README/ssp.public.README** file.

Glossary of Terms and Abbreviations

A

ACL. Access Control List. A list that defines who has permission to access certain services; that is, for whom a server may perform certain tasks. This is usually a list of principals with the type of access assigned to each.

adapter. An adapter is a mechanism for attaching parts. For example, an adapter could be a part that electrically or physically connects a device to a computer or to another device. In the SP system, network connectivity is supplied by various adapters, some optional, that can provide connection to I/O devices, networks of workstations, and mainframe networks. Ethernet, FDDI, token-ring, HiPPI, SCSI, FCS, and ATM are examples of adapters that can be used as part of an SP system.

address. A character or group of characters that identifies a register, a device, a particular part of storage, or some other data source or destination.

AFS. A distributed file system that provides authentication services as part of its file system creation.

AIX. Abbreviation for Advanced Interactive Executive, IBM's licensed version of the UNIX operating system. AIX is particularly suited to support technical computing applications, including high function graphics and floating point computations.

API. Application Programming Interface. A set of programming functions and routines that provide access between the Application layer of the OSI seven-layer model and applications that want to use the network. It is a software interface.

application. The use to which a data processing system is put; for example, a payroll application, an airline reservation application.

application data. The data that is produced using an application program.

ARP. Address Resolution Protocol.

ATM. Asynchronous Transfer Mode. (See *TURBOWAYS 100 ATM Adapter*.)

authentication. The process of validating the identity of either a user of a service or the service itself. The process of a principal proving the authenticity of its identity.

authorization. The process of obtaining permission to access resources or perform tasks. In SP security services, authorization is based on the principal identifier. The granting of access rights to a principal.

authorization file. A type of ACL (access control list) used by the IBM AIX remote commands and the IBM PSSP Sysctl and Hardmon components.

B

batch processing. (1) The processing of data or the accomplishment of jobs accumulated in advance in such a manner that each accumulation thus formed is processed or accomplished in the same run. (2) The processing of data accumulating over a period of time. (3) Loosely, the execution of computer programs serially. (4) Computer programs executed in the background.

BOS. The AIX Base Operating System.

C

call home function. The ability of a system to call the IBM support center and open a PMR to have a repair scheduled.

CDE. Common Desktop Environment. A graphical user interface for UNIX.

charge feature. An optional feature for either software or hardware for which there is a charge.

CLI. Command Line Interface.

client. (1) A function that requests services from a server and makes them available to the user. (2) A term used in an environment to identify a machine that uses the resources of the network.

CMI. Centralized Management Interface provides a series of SMIT menus and dialogues used for defining and querying the SP system configuration.

Concurrent Virtual Shared Disk. A virtual shared disk that can be concurrently accessed by more than one server.

connectionless. A communication process that takes place without first establishing a connection.

connectionless network. A network in which the sending logical node must have the address of the receiving logical node before information interchange can begin. The packet is routed through nodes in the network based on the destination address in the packet. The originating source does not receive an acknowledgment that the packet was received at the destination.

control workstation. A single point of control allowing the administrator or operator to monitor and manage the SP system using the IBM AIX Parallel System Support Programs.

credentials. A protocol message, or part thereof, containing a ticket and an authenticator supplied by a client and used by a server to verify the client's identity.

css. Communication subsystem.

D

daemon. A process, not associated with a particular user, that performs system-wide functions such as administration and control of networks, execution of time-dependent activities, line printer spooling and so forth.

DASD. Direct Access Storage Device. Storage for input/output data.

DCE. Distributed Computing Environment.

DFS. distributed file system. A subset of the IBM Distributed Computing Environment.

DNS. Domain Name Service. A hierarchical name service which maps high level machine names to IP addresses.

E

Error Notification Object. An object in the SDR that is matched with an error log entry. When an error log entry occurs that matches the Notification Object, a user-specified action is taken.

ESCON. Enterprise Systems Connection. The ESCON channel connection allows the RS/6000 to communicate directly with a host System/390; the host operating system views the system unit as a control unit.

Ethernet. (1) Ethernet is the standard hardware for TCP/IP local area networks in the UNIX marketplace. It is a 10-megabit per second baseband type LAN that allows multiple stations to access the transmission medium at will without prior coordination, avoids contention by using carrier sense and deference, and resolves contention by collision detection (CSMA/CD). (2) A passive coaxial cable whose interconnections contain devices or components, or both, that are all active. It uses CSMA/CD technology to provide a best-effort delivery system.

Ethernet network. A baseband LAN with a bus topology in which messages are broadcast on a coaxial cabling using the carrier sense multiple access/collision detection (CSMA/CD) transmission method.

event. In Event Management, the notification that an expression evaluated to true. This evaluation occurs each time an instance of a resource variable is observed.

expect. Programmed dialogue with interactive programs.

expression. In Event Management, the relational expression between a resource variable and other elements (such as constants or the previous value of an instance of the variable) that, when true, generates an event. An example of an expression is `X < 10` where X represents the resource variable `IBM.PSSP.aixos.PagSp.%totalfree` (the percentage of total free paging space). When the expression is true, that is, when the total free paging space is observed to be less than 10%, the Event Management subsystem generates an event to notify the appropriate application.

F

failover. Also called fallover, the sequence of events when a primary or server machine fails and a secondary or backup machine assumes the primary workload. This is a disruptive failure with a short recovery time.

fall back. Also called fallback, the sequence of events when a primary or server machine takes back control of its workload from a secondary or backup machine.

FDDI. Fiber Distributed Data Interface.

FFDC. First Failure Data Capture.

Fiber Distributed Data Interface (FDDI). An American National Standards Institute (ANSI) standard for 100-megabit-per-second LAN using optical fiber cables. An FDDI local area network (LAN) can be up to 100 km (62 miles) and can include up to 500 system units. There can be up to 2 km (1.24 miles) between system units and concentrators.

file. A set of related records treated as a unit, for example, in stock control, a file could consist of a set of invoices.

file name. A CMS file identifier in the form of 'filename filetype filemode' (like: TEXT DATA A).

file server. A centrally located computer that acts as a storehouse of data and applications for numerous users of a local area network.

File Transfer Protocol (FTP). The Internet protocol (and program) used to transfer files between hosts. It is an application layer protocol in TCP/IP that uses TELNET and TCP protocols to transfer bulk-data files between machines or hosts.

firewall node. A node that physically and logically separates both sides of a firewalled SP system. The firewall node is connected to the Trusted nodes through

one Ethernet network (en0, the original SP Ethernet LAN). It is connected to the Untrusted nodes through a second Ethernet network (en1 for the second SP Ethernet LAN). The firewall node, by way of the firewall application's services and rules, permits or blocks (grants or denies) the flow of communication between the Trusted and Untrusted sides. This includes traffic to and from the firewall node, and to and from the control workstation. The firewall node is considered both an Untrusted and Trusted node.

firewall rule. A condition that permits or blocks a communication flow between two points, based on a combination of values, such as the source host (hostname and IP address), destination host (or target host, hostname and IP address), protocol type/service type, and port number (for example, sysctl 6680/tcp).

firewall service definitions. See firewall services.

firewall services. Services needed for PSSP to run and manage an SP system through a firewall. The service descriptions are modeled after the file format definition for the **/etc/services** file.

firewalled SP system. An RS/6000 SP system that has been modified to consist of a firewall node, Untrusted nodes, and Trusted nodes, in accordance with the instructions provided in: *Implementing a Firewalled RS/6000 SP System*

First Failure Data Capture (FFDC). A set of utilities used for recording persistent records of failures and significant software incidents. It provides a means of associating failures to one another, thus allowing software to link effects of a failure to their causes and thereby facilitating discovery of the root cause of a failure.

foreign host. Any host on the network other than the local host.

FTP. File transfer protocol.

G

gateway. An intelligent electronic device interconnecting dissimilar networks and providing protocol conversion for network compatibility. A gateway provides transparent access to dissimilar networks for nodes on either network. It operates at the session presentation and application layers.

H

HACMP. High Availability Cluster Multi-Processing for AIX.

HACWS. High Availability Control Workstation function, based on HACMP, provides for a backup control workstation for the SP system.

Hardware Management Console (HMC). The *IBM Hardware Management Console for pSeries* is an installation and service support processor that runs only the HMC software. For an IBM @server pSeries 690 or 670 server to run the PSSP software, an HMC is required with a trusted network connection to the PSSP control workstation. A trusted network can be the SP Ethernet admin LAN or a specifically configured HMC trusted network comprised only of HMC systems and control workstations. The HMC provides the following functions for the p690 or p670 server:

- Creating and maintaining a multiple partition environment
- Detecting, reporting, and storing changes in hardware conditions
- Acting as a focal point for service representatives to determine an appropriate service strategy

Hashed Shared Disk (HSD). The data striping device for the IBM Virtual Shared Disk. The device driver lets application programs stripe data across physical disks in multiple IBM Virtual Shared Disks, thus reducing I/O bottlenecks.

help key. In the SP graphical interface, the key that gives you access to the SP graphical interface help facility.

High Availability Cluster Multi-Processing. An IBM facility to cluster nodes or components to provide high availability by eliminating single points of failure.

HiPPI. High Performance Parallel Interface. RS/6000 units can attach to a HiPPI network as defined by the ANSI specifications. The HiPPI channel supports burst rates of 100 Mbps over dual simplex cables; connections can be up to 25 km in length as defined by the standard and can be extended using third-party HiPPI switches and fiber optic extenders.

home directory. The directory associated with an individual user.

host. A computer connected to a network, and providing an access method to that network. A host provides end-user services.

HMC. Hardware Management Console.

HMC Trusted Network. A trusted network is one where all hosts on the same network (LAN) are regarded as trusted, according to site security policies and procedures governing the hosts. Data on a trusted network can be seen by all trusted hosts and their users, but the implied trust among and between the hosts assumes that the data will not be intercepted or modified. Therefore, by way of implied mutual trust, traffic flowing across the trusted network is regarded as safe from unwanted or unintended interception or tampering. However, it does not imply that the data on the trusted network is itself private or encrypted.

Two types of trusted networks are available in Cluster 1600 configurations with p690 or p670 servers:

1. The SP Ethernet admin LAN is a trusted network.

   Since the SP Ethernet admin LAN is considered a trusted network, an HMC connected to it becomes part of the trusted network.

2. A trusted network established between the control workstation and the HMC.

   This requires that both hosts are connected via a network other than the SP Ethernet admin LAN. This ″other″ network is called the HMC trusted network.

I

instance vector. Obsolete term for resource identifier.

Intermediate Switch Board. Switches mounted in the switch expansion frame.

Internet. A specific inter-network consisting of large national backbone networks such as APARANET, MILNET, and NSFnet, and a myriad of regional and campus networks all over the world. The network uses the TCP/IP protocol suite.

Internet Protocol (IP). (1) A protocol that routes data through a network or interconnected networks. IP acts as an interface between the higher logical layers and the physical network. This protocol, however, does not provide error recovery, flow control, or guarantee the reliability of the physical network. IP is a connectionless protocol. (2) A protocol used to route data from its source to it destination in an Internet environment.

IP address. A 32-bit address assigned to devices or hosts in an IP internet that maps to a physical address. The IP address is composed of a network and host portion.

ISB. Intermediate Switch Board.

K

Kerberos. A service for authenticating users in a network environment.

kernel. The core portion of the UNIX operating system which controls the resources of the CPU and allocates them to the users. The kernel is memory-resident, is said to run in “kernel mode” and is protected from user tampering by the hardware.

Kernel Low-Level Application Programming Interface (KLAPI). KLAPI provides transport service for communication using the SP Switch.

L

LAN. (1) Acronym for Local Area Network, a data network located on the user’s premises in which serial transmission is used for direct data communication among data stations. (2) Physical network technology that transfers data a high speed over short distances. (3) A network in which a set of devices is connected to another for communication and that can be connected to a larger network.

LAPI. Low-level Communication API.

local host. The computer to which a user’s terminal is directly connected.

log database. A persistent storage location for the logged information.

log event. The recording of an event.

log event type. A particular kind of log event that has a hierarchy associated with it.

logging. The writing of information to persistent storage for subsequent analysis by humans or programs.

Low-level Communication API (LAPI). A low-level (low overhead) message passing protocol that uses a one-sided active message style interface to transfer messages between processes. LAPI is an IBM proprietary interface designed to exploit the SP switch adapters.

M

mask. To use a pattern of characters to control retention or elimination of portions of another pattern of characters.

menu. A display of a list of available functions for selection by the user.

Message Passing Interface (MPI). An industry standard message passing protocol that typically uses a two-sided send-receive model to transfer messages between processes.

Motif. The graphical user interface for OSF, incorporating the X Window System. Also called OSF/Motif.

MPI. Message Passing Interface.

MTBF. Mean time between failure. This is a measure of reliability.

MTTR. Mean time to repair. This is a measure of serviceability.

N

naive application. An application with no knowledge of a server that fails over to another server. Client to server retry methods are used to reconnect.

network. An interconnected group of nodes, lines, and terminals. A network provides the ability to transmit data to and receive data from other systems and users.

Network Interface Module (NIM). A process used by the Topology Services daemon to monitor each network interface.

NFS. Network File System. NFS allows different systems (UNIX or non-UNIX), different architectures, or vendors connected to the same network, to access remote files in a LAN environment as though they were local files.

NIM. (1) Network Installation Management is provided with AIX to install AIX on the nodes. (2) Network Interface Module is a process used by the Topology Services daemon to monitor each network interface.

NIM client. An AIX system installed and managed by a NIM master. NIM supports three types of clients:
- Standalone
- Diskless
- Dataless

NIM master. An AIX system that can install one or more NIM clients. An AIX system must be defined as a NIM master before defining any NIM clients on that system. A NIM master managers the configuration database containing the information for the NIM clients.

NIM object. A representation of information about the NIM environment. NIM stores this information as objects in the NIM database. The types of objects are:
- Network
- Machine
- Resource

NIS. Network Information System.

node. In a network, the point where one or more functional units interconnect transmission lines. A computer location defined in a network. The SP system can house several different types of nodes for both serial and parallel processing. These node types can include thin nodes, wide nodes, 604 high nodes, as well as other types of nodes both internal and external to the SP frame.

Node Switch Board. Switches mounted on frames that contain nodes.

NSB. Node Switch Board.

NTP. Network Time Protocol.

O

ODM. Object Data Manager. In AIX, a hierarchical object-oriented database for configuration data.

P

parallel environment. An execution and development environment for parallel processing programs where message passing services are commonly used.

Parallel Environment. An IBM licensed program that provides an execution and development environment for parallel C, C++, and FORTRAN programs. It includes tools for debugging, profiling, and tuning parallel programs.

parallel processing. A multiprocessor architecture which allows processes to be allocated to tightly coupled multiple processors in a cooperative processing environment, allowing concurrent execution of tasks.

parameter. (1) A variable that is given a constant value for a specified application and that may denote the application. (2) An item in a menu for which the operator specifies a value or for which the system provides a value when the menu is interpreted. (3) A name in a procedure that is used to refer to an argument that is passed to the procedure. (4) A particular piece of information that a system or application program needs to process a request.

partition. See system partition.

Perl. Practical Extraction and Report Language.

perspective. The primary window for each SP Perspectives application, so called because it provides a unique view of an SP system.

pipe. A UNIX utility allowing the output of one command to be the input of another. Represented by the | symbol. It is also referred to as filtering output.

PMR. Problem Management Report.

POE. Formerly Parallel Operating Environment, now Parallel Environment for AIX.

port. (1) An end point for communication between devices, generally referring to physical connection. (2) A 16-bit number identifying a particular TCP or UDP resource within a given TCP/IP node.

predicate. Obsolete term for expression.

Primary node or machine. (1) A device that runs a workload and has a standby device ready to assume the primary workload if that primary node fails or is taken out of service. (2) A node on the switch that initializes, provides diagnosis and recovery services, and performs other operations to the switch network. (3)

In IBM Virtual Shared Disk function, when physical disks are connected to two nodes (twin-tailed), one node is designated as the primary node for each disk and the other is designated the secondary, or backup, node. The primary node is the server node for IBM Virtual Shared Disks defined on the physical disks under normal conditions. The secondary node can become the server node for the disks if the primary node is unavailable (off-line or down).

Problem Management Report. The number in the IBM support mechanism that represents a service incident with a customer.

process. (1) A unique, finite course of events defined by its purpose or by its effect, achieved under defined conditions. (2) Any operation or combination of operations on data. (3) A function being performed or waiting to be performed. (4) A program in operation. For example, a daemon is a system process that is always running on the system.

protocol. A set of semantic and syntactic rules that defines the behavior of functional units in achieving communication.

R

RAID. Redundant array of independent disks.

rearm expression. In Event Management, an expression used to generate an event that alternates with an original event expression in the following way: the event expression is used until it is true, then the rearm expression is used until it is true, then the event expression is used, and so on. The rearm expression is commonly the inverse of the event expression (for example, a resource variable is on or off). It can also be used with the event expression to define an upper and lower boundary for a condition of interest.

rearm predicate. Obsolete term for rearm expression.

remote host. *See foreign host.*

resource. In Event Management, an entity in the system that provides a set of services. Examples of resources include hardware entities such as processors, disk drives, memory, and adapters, and software entities such as database applications, processes, and file systems. Each resource in the system has one or more attributes that define the state of the resource.

resource identifier. In Event Management, a set of elements, where each element is a name/value pair of the form `name=value`, whose values uniquely identify the copy of the resource (and by extension, the copy of the resource variable) in the system.

resource monitor. A program that supplies information about resources in the system. It can be a command, a daemon, or part of an application or subsystem that manages any type of system resource.

resource variable. In Event Management, the representation of an attribute of a resource. An example of a resource variable is `IBM.AIX.PagSp.%totalfree`, which represents the percentage of total free paging space. `IBM.AIX.PagSp` specifies the resource name and `%totalfree` specifies the resource attribute.

Restricted Root Access (RRA). Restricted root access (RRA) limits the uses of the **rsh** and **rcp** commands within PSSP software. When RRA is enabled, it restricts root **rsh** and **rcp** authorizations from the nodes to the control workstation, and from one node to another. However, control workstation to node **rsh** and **rcp** access is still permitted.

RISC. Reduced Instruction Set Computing (RISC), the technology for today's high performance personal computers and workstations, was invented in 1975. Uses a small simplified set of frequently used instructions for rapid execution.

rlogin (remote LOGIN). A service offered by Berkeley UNIX systems that allows authorized users of one machine to connect to other UNIX systems across a network and interact as if their terminals were connected directly. The rlogin software passes information about the user's environment (for example, terminal type) to the remote machine.

RPC. Acronym for Remote Procedure Call, a facility that a client uses to have a server execute a procedure call. This facility is composed of a library of procedures plus an XDR.

RRA. Restricted Root Access.

RSH. A variant of RLOGIN command that invokes a command interpreter on a remote UNIX machine and passes the command line arguments to the command interpreter, skipping the LOGIN step completely. See also *rlogin*.

S

SCSI. Small Computer System Interface.

Secondary node. In IBM Virtual Shared Disk function, when physical disks are connected to two nodes (twin-tailed), one node is designated as the primary node for each disk and the other is designated as the secondary, or backup, node. The secondary node acts as the server node for the IBM Virtual Shared disks defined on the physical disks if the primary node is unavailable (off-line or down).

Secure File Collections. Implementation of the PSSP File Collections feature with an option that limits access to the supman userid password, to the system administrator.

server. (1) A function that provides services for users. A machine may run client and server processes at the same time. (2) A machine that provides resources to the network. It provides a network service, such as disk storage and file transfer, or a program that uses such a service. (3) A device, program, or code module on a network dedicated to providing a specific service to a network. (4) On a LAN, a data station that provides facilities to other data stations. Examples are file server, print server, and mail server.

shell. The shell is the primary user interface for the UNIX operating system. It serves as command language interpreter, programming language, and allows foreground and background processing. There are three different implementations of the shell concept: Bourne, C and Korn.

Small Computer System Interface (SCSI). An input and output bus that provides a standard interface for the attachment of various direct access storage devices (DASD) and tape drives to the RS/6000.

Small Computer Systems Interface Adapter (SCSI Adapter). An adapter that supports the attachment of various direct-access storage devices (DASD) and tape drives to the RS/6000.

SMIT. The System Management Interface Toolkit is a set of menu driven utilities for AIX that provides functions such as transaction login, shell script creation, automatic updates of object database, and so forth.

SNMP. Simple Network Management Protocol. (1) An IP network management protocol that is used to monitor attached networks and routers. (2) A TCP/IP-based protocol for exchanging network management information and outlining the structure for communications among network devices.

socket. (1) An abstraction used by Berkeley UNIX that allows an application to access TCP/IP protocol functions. (2) An IP address and port number pairing. (3) In TCP/IP, the Internet address of the host computer on which the application runs, and the port number it uses. A TCP/IP application is identified by its socket.

standby node or machine. A device that waits for a failure of a primary node in order to assume the identity of the primary node. The standby machine then runs the primary's workload until the primary is back in service.

subnet. Shortened form of subnetwork.

subnet mask. A bit template that identifies to the TCP/IP protocol code the bits of the host address that are to be used for routing for specific subnetworks.

subnetwork. Any group of nodes that have a set of common characteristics, such as the same network ID.

subsystem. A software component that is not usually associated with a user command. It is usually a daemon process. A subsystem will perform work or provide services on behalf of a user request or operating system request.

SUP. Software Update Protocol.

switch capsule. A group of SP frames consisting of a switched frame and its companion non-switched frames.

Sysctl. Secure System Command Execution Tool. An authenticated client/server system for running commands remotely and in parallel.

syslog. A BSD logging system used to collect and manage other subsystem's logging data.

System Administrator. The user who is responsible for setting up, modifying, and maintaining the SP system.

system partition. A group of nonoverlapping nodes on a switch chip boundary that act as a logical SP system.

T

tar. Tape ARchive, is a standard UNIX data archive utility for storing data on tape media.

Tcl. Tool Command Language.

TclX. Tool Command Language Extended.

TCP. Acronym for Transmission Control Protocol, a stream communication protocol that includes error recovery and flow control.

TCP/IP. Acronym for Transmission Control Protocol/Internet Protocol, a suite of protocols designed to allow communication between networks regardless of the technologies implemented in each network. TCP provides a reliable host-to-host protocol between hosts in packet-switched communications networks and in interconnected systems of such networks. It assumes that the underlying protocol is the Internet Protocol.

Telnet. Terminal Emulation Protocol, a TCP/IP application protocol that allows interactive access to foreign hosts.

ticket. An encrypted protocol message used to securely pass the identity of a user from a client to a server.

Tk. Tcl-based Tool Kit for X Windows.

TMPCP. Tape Management Program Control Point.

token-ring. (1) Network technology that controls media access by passing a token (special packet or frame) between media-attached machines. (2) A network with a ring topology that passes tokens from one attaching device (node) to another. (3) The IBM Token-Ring LAN connection allows the RS/6000 system unit to

participate in a LAN adhering to the IEEE 802.5 Token-Passing Ring standard or the ECMA standard 89 for Token-Ring, baseband LANs.

transaction. An exchange between the user and the system. Each activity the system performs for the user is considered a transaction.

transceiver (transmitter-receiver). A physical device that connects a host interface to a local area network, such as Ethernet. Ethernet transceivers contain electronics that apply signals to the cable and sense collisions.

transfer. To send data from one place and to receive the data at another place. Synonymous with move.

transmission. The sending of data from one place for reception elsewhere.

trusted network. See HMC trusted network.

Trusted node.

A node located on the Trusted side of a firewalled SP system. See the definition for Trusted side.

Trusted side.

The Trusted side of a firewalled SP system covers those SP hosts that sit after the firewall node. This includes Trusted nodes and the control workstation. The Trusted side may also be considered the private side.

TURBOWAYS 100 ATM Adapter. An IBM high-performance, high-function intelligent adapter that provides dedicated 100 Mbps ATM (asynchronous transfer mode) connection for high-performance servers and workstations.

U

UDP. User Datagram Protocol.

UNIX operating system. An operating system developed by Bell Laboratories that features multiprogramming in a multiuser environment. The UNIX operating system was originally developed for use on minicomputers, but has been adapted for mainframes and microcomputers. **Note:** The AIX operating system is IBM's implementation of the UNIX operating system.

Untrusted node.

A node located on the Untrusted side of a firewalled SP system. See the definition for Untrusted side.

Untrusted side.

The Untrusted side of a firewalled SP system consists of Untrusted nodes. These nodes sit before the firewall node. The Untrusted side may also be considered the public side, depending on your view of how a firewalled SP system is used.

user. Anyone who requires the services of a computing system.

User Datagram Protocol (UDP). (1) In TCP/IP, a packet-level protocol built directly on the Internet Protocol layer. UDP is used for application-to-application programs between TCP/IP host systems. (2) A transport protocol in the Internet suite of protocols that provides unreliable, connectionless datagram service. (3) The Internet Protocol that enables an application programmer on one machine or process to send a datagram to an application program on another machine or process.

user ID. A nonnegative integer, contained in an object of type *uid_t*, that is used to uniquely identify a system user.

V

Virtual Shared Disk, IBM. The function that allows application programs executing at different nodes of a system partition to access a raw logical volume as if it were local at each of the nodes. In actuality, the logical volume is local at only one of the nodes (the server node).

W

workstation. (1) A configuration of input/output equipment at which an operator works. (2) A terminal or microcomputer, usually one that is connected to a mainframe or to a network, at which a user can perform applications.

X

X Window System. A graphical user interface product.

Bibliography

This bibliography helps you find product documentation related to the RS/6000 SP hardware and software products.

You can find most of the IBM product information for RS/6000 SP products on the World Wide Web. Formats for both viewing and downloading are available.

PSSP documentation is shipped with the PSSP product in a variety of formats and can be installed on your system. The man pages for public code that PSSP includes are also available online.

Finally, this bibliography contains a list of non-IBM publications that discuss parallel computing and other topics related to the RS/6000 SP.

Information formats

Documentation supporting RS/6000 SP software licensed programs is no longer available from IBM in hardcopy format. However, you can view, search, and print documentation in the following ways:

- On the World Wide Web
- Online from the product media or the Cluster Resource Center

Finding documentation on the World Wide Web

Most of the RS/6000 SP hardware and software books are available from the IBM Web site at:

| http://www.ibm.com/servers/eserver/pseries/library

You can view a book or download a Portable Document Format (PDF) version of it. At the time this manual was published, the Web address of the ″RS/6000 SP hardware and software documentation″ page was:

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

However, the structure of the RS/6000 Web site can change over time.

You can also use the IBM Publications Center, which offers customized search functions, to help you find the publications you need. At the time this manual was published, the Web address of the IBM Publications Center was:

| http://www.ibm.com/shop/publications/order

Accessing PSSP documentation online

On the same medium as the PSSP product code, IBM ships PSSP man pages, HTML files, and PDF files. In order to use these publications, you must first install the **ssp.docs** file set.

To view the PSSP HTML publications, you need access to an HTML document browser such as Netscape. The HTML files and an index that links to them are installed in the **/usr/lpp/ssp/html** directory. Once installed, you can also view the HTML files from the Cluster Resource Center.

If you have installed the Cluster Resource Center on your SP system, you can access it by entering the **/usr/lpp/ssp/bin/resource_center** command. If you have the Cluster Resource Center on CD-ROM, see the **readme.txt** file for information about how to run it.

To view the PSSP PDF publications, you need access to the Adobe Acrobat Reader. The Acrobat Reader is shipped with the AIX Bonus Pack and is also freely available for downloading from the Adobe Web site at:

http://www.adobe.com

To successfully print a large PDF file (approximately 300 or more pages) from the Adobe Acrobat reader, you may need to select the “Download Fonts Once” button on the Print window.

Manual pages for public code

The following manual pages for public code are available in this product:

SUP /usr/lpp/ssp/man/man1/sup.1

Perl (Version 4.036)

/usr/lpp/ssp/perl/man/perl.man

/usr/lpp/ssp/perl/man/h2ph.man

/usr/lpp/ssp/perl/man/s2p.man

/usr/lpp/ssp/perl/man/a2p.man

Manual pages and other documentation for **Tcl**, **TclX**, **Tk**, and **expect** can be found in the compressed **tar** files located in the **/usr/lpp/ssp/public** directory.

System planning publications

This section lists the IBM product documentation for planning for the IBM RS/6000 SP hardware and software and for an IBM @server Cluster 1600.

IBM @server *Cluster 1600:*

- *Planning, Installation, and Service*, GA22-7863

IBM RS/6000 SP:

- *Planning, Volume 1, Hardware and Physical Environment*, GA22-7280
- *Planning, Volume 2, Control Workstation and Software Environment*, GA22-7281

RS/6000 SP hardware publications

This section lists the IBM product documentation for the IBM RS/6000 SP hardware.

IBM RS/6000 SP:

- *Planning, Volume 1, Hardware and Physical Environment*, GA22-7280
- *Planning, Volume 2, Control Workstation and Software Environment*, GA22-7281
- *Installation and Relocation*, GA22-7441
- *System Service Guide*, GA22-7442
- *SP Switch Service Guide*, GA22-7443

- *SP Switch2 Service Guide*, GA22-7444
- *Uniprocessor Node Service Guide*, GA22-7445
- *604 and 604e SMP High Node Service Guide*, GA22-7446
- *SMP Thin and Wide Node Service Guide*, GA22-7447
- *POWER3 SMP High Node Service Guide*, GA22-7448
- *Safety Information*, GA22-7467

RS/6000 SP Switch Router publications

The RS/6000 SP Switch Router is based on the Ascend GRF switched IP router product from Lucent Technologies. You can order the SP Switch Router as the IBM 9077.

The following publications are shipped with the SP Switch Router. You can also order these publications from IBM using the order numbers shown.

- *Ascend GRF GateD Manual*, GA22-7327
- *Ascend GRF 400/1600 Getting Started*, GA22-7368
- *Ascend GRF Configuration and Management*, GA22-7366
- *Ascend GRF Reference Guide*, GA22-7367
- *SP Switch Router Adapter Guide*, GA22-7310

Related hardware publications

For publications on the latest IBM @server pSeries and RS/6000 hardware products, see the Web site:

http://www.ibm.com/servers/eserver/pseries/library/hardware_docs/

That site includes links to the following:

- General service documentation
- Guides by system (pSeries and RS/6000)
- Installable options
- IBM Hardware Management Console for pSeries guides

RS/6000 SP software publications

This section lists the IBM product documentation for software products related to the IBM RS/6000 SP system. These products include:

- IBM Parallel System Support Programs for AIX (PSSP)
- IBM LoadLeveler for AIX 5L (LoadLeveler)
- IBM Parallel Environment for AIX (Parallel Environment)
- IBM General Parallel File System for AIX (GPFS)
- IBM Engineering and Scientific Subroutine Library (ESSL) for AIX
- IBM Parallel ESSL for AIX
- IBM High Availability Cluster Multi-Processing for AIX (HACMP)
- IBM High Availability Geographic Clustering for AIX (HAGEO)
- IBM Geographic Remote Mirroring for AIX (GeoRM)

PSSP Publications

IBM RS/6000 SP:

- *Planning, Volume 2, Control Workstation and Software Environment*, GA22-7281

PSSP:

- *Installation and Migration Guide*, GA22-7347
- *Administration Guide*, SA22-7348
- *Managing Shared Disks*, SA22-7349
- *Diagnosis Guide*, GA22-7350
- *Command and Technical Reference*, SA22-7351
- *Messages Reference*, GA22-7352
- *Implementing a Firewalled RS/6000 SP System*, GA22-7874

RS/6000 Cluster Technology (RSCT):

- *Event Management Programming Guide and Reference*, SA22-7354
- *Group Services Programming Guide and Reference*, SA22-7355
- *First Failure Data Capture Programming Guide and Reference*, SA22-7454

LoadLeveler Publications

LoadLeveler:

- *Using and Administering*, SA22-7881
- *Diagnosis and Messages Guide*, GA22-7882
- *Installation Memo*, GI11-2819

GPFS Publications

GPFS for AIX 5L:

- *AIX Clusters Concepts, Planning, and Installation*, GA22-7895
- *AIX Clusters Problem Determination Guide*, GA22-7897
- *AIX Clusters Administration and Programming Reference*, SA22-7896
- *AIX Clusters Data Management API Guide*, GA22-7898
- *PSSP Clusters Concepts, Planning, and Installation*, GA22-7899
- *PSSP Clusters Problem Determination Guide*, GA22-7901
- *PSSP Clusters Administration and Programming Reference*, SA22-7900
- *PSSP Clusters Data Management API Guide*, GA22-7902

Parallel Environment Publications

Parallel Environment:

- *Installation Guide*, GA22-7418
- *Messages*, GA22-7419
- *MPI Programming Guide*, SA22-7422
- *MPI Subroutine Reference*, SA22-7423
- *Hitchhiker's Guide*, SA22-7424
- *Operation and Use, Volume 1*, SA22-7425
- *Operation and Use, Volume 2*, SA22-7426

Parallel ESSL and ESSL Publications

- *ESSL Products: General Information*, GC23-0529
- *Parallel ESSL: Guide and Reference*, SA22-7273
- *ESSL: Guide and Reference*, SA22-7272

HACMP Publications

HACMP:
- *Concepts and Facilities*, SC23-4276
- *Planning Guide*, SC23-4277
- *Installation Guide*, SC23-4278
- *Administration Guide*, SC23-4279
- *Troubleshooting Guide*, SC23-4280
- *Programming Locking Applications*, SC23-4281
- *Programming Client Applications*, SC23-4282
- *Master Index and Glossary*, SC23-4285
- *Enhanced Scalability Installation and Administration Guide*, SC23-4306

HAGEO Publications

HAGEO:
- *Concepts and Facilities*, SC23-1922
- *Planning and Administration Guide*, SC23-1886

GeoRM Publications

GeoRM:
- *Concepts and Facilities*, SC23-4307
- *Planning and Administration Guide*, SC23-4308

AIX publications

You can find links to the latest AIX publications on the Web site:

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

The IBM Reliable Scalable Cluster Technology (RSCT) software is packaged with AIX 5L 5.1. You can find links to the RSCT publications on the Web site:

http://www.ibm.com/servers/eserver/clusters/library

The publications are listed here for your convenience:

IBM Reliable Scalable Cluster Technology for AIX 5L:
- *RSCT Guide and Reference*, SA22-7889
- *Technical Reference*, SA22-7890
- *Messages*, GA22-7891

DCE publications

The DCE library consists of the following books:

- *IBM DCE for AIX: Administration Commands Reference*
- *IBM DCE for AIX: Administration Guide—Introduction*
- *IBM DCE for AIX: Administration Guide—Core Components*
- *IBM DCE for AIX: DFS Administration Guide and Reference*
- *IBM DCE for AIX: Application Development Guide—Introduction and Style Guide*
- *IBM DCE for AIX: Application Development Guide—Core Components*
- *IBM DCE for AIX: Application Development Guide—Directory Services*
- *IBM DCE for AIX: Application Development Reference*
- *IBM DCE for AIX: Problem Determination Guide*
- *IBM DCE for AIX: Release Notes*

You can view the relevant version of a DCE book or download a Portable Document Format (PDF) file of it from the IBM DCE Web site at:

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

Redbooks

IBM's International Technical Support Organization (ITSO) has published a number of redbooks related to the RS/6000 SP. For a current list, see the ITSO Web site at:

http://www.ibm.com/redbooks

Non-IBM publications

Here are some non-IBM publications that you might find helpful.

- Almasi, G., Gottlieb, A., *Highly Parallel Computing*, Benjamin-Cummings Publishing Company, Inc., 1989.
- Foster, I., *Designing and Building Parallel Programs*, Addison-Wesley, 1995.
- Gropp, W., Lusk, E., Skjellum, A., *Using MPI*, The MIT Press, 1994.
- Message Passing Interface Forum, *MPI: A Message-Passing Interface Standard, Version 1.1*, University of Tennessee, Knoxville, Tennessee, June 6, 1995.
- Message Passing Interface Forum, *MPI-2: Extensions to the Message-Passing Interface, Version 2.0*, University of Tennessee, Knoxville, Tennessee, July 18, 1997.
- Ousterhout, John K., *Tcl and the Tk Toolkit*, Addison-Wesley, Reading, MA, 1994, ISBN 0-201-63337-X.
- Pfister, Gregory, F., *In Search of Clusters*, Prentice Hall, 1998.
- Barrett, D., Silverman, R., *SSH The Secure Shell The Definitive Guide*, O'Reilly, 2001.

Index

Numerics

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

W

Readers' comments – We'd like to hear from you

RS/6000 SP
Planning Volume 2, Control Workstation and Software Environment

Publication No. GA22-7281-07

Overall, how satisfied are you with the information in this book?

	Very Satisfied	Satisfied	Neutral	Dissatisfied	Very Dissatisfied
Overall satisfaction	☐	☐	☐	☐	☐

How satisfied are you that the information in this book is:

	Very Satisfied	Satisfied	Neutral	Dissatisfied	Very Dissatisfied
Accurate	☐	☐	☐	☐	☐
Complete	☐	☐	☐	☐	☐
Easy to find	☐	☐	☐	☐	☐
Easy to understand	☐	☐	☐	☐	☐
Well organized	☐	☐	☐	☐	☐
Applicable to your tasks	☐	☐	☐	☐	☐

Please tell us how we can improve this book:

Thank you for your responses. May we contact you? ☐ Yes ☐ No

When you send comments to IBM, you grant IBM a nonexclusive right to use or distribute your comments in any way it believes appropriate without incurring any obligation to you.

Name _____ Address _____

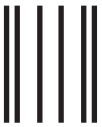
Company or Organization _____

Phone No. _____

Readers' Comments — We'd Like to Hear from You
GA22-7281-07

Cut or Fold
Along Line

Fold and Tape **Please do not staple** Fold and Tape



NO POSTAGE
NECESSARY
IF MAILED IN THE
UNITED STATES

BUSINESS REPLY MAIL

FIRST-CLASS MAIL PERMIT NO. 40 ARMONK, NEW YORK

POSTAGE WILL BE PAID BY ADDRESSEE

IBM Corporation
Department 55JA, Mail Station P384
2455 South Road
Poughkeepsie NY 12601-5400

Fold and Tape **Please do not staple** Fold and Tape

Cut or Fold
Along Line

Program Number: 5765-D51

GA22-7281-07