IBM RS/6000 SP

# Planning Volume 2, Control Workstation and Software Environment

GA22-7281-02

IBM RS/6000 SP

# Planning Volume 2, Control Workstation and Software Environment

GA22-7281-02

**Note!**

Before using this information and the product it supports, be sure to read the general information under “Notices” on page xi

| **Third Edition (February 1998)**

| This is a major revision of GA22–7281–01.

| This edition applies to Version 2 Release 4 of the IBM Parallel System Support Programs for AIX (PSSP) Licensed Program
| (5765–529), which runs on the IBM RS/6000 SP, and to all subsequent releases and modifications until otherwise indicated in new
| editions. Significant changes or additions to the text and illustrations are indicated by a vertical line (|) to the left of the change.

Order publications through your IBM representative or the IBM branch office serving your locality. Publications are not stocked at the address below.

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 the following address:

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

FAX (United States and Canada): 1+914+432-9405
FAX (Other Countries):
Your international Access Code +1+914+432+9405

IBMLink (United States customers only): KGNVMC(MHVRCFS)
IBM Mail Exchange: USIB6TC9 at IBMMAIL
Internet e-mail: mhvrcfs@vnet.ibm.com
World Wide Web: http://www.rs6000.ibm.com

If you would like 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.

© **Copyright International Business Machines Corporation 1997, 1998. All rights reserved.**
Note to U.S. Government Users — Documentation related to restricted rights — Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corp.

# Contents


**Notices** . . . . . xi
Trademarks . . . . . xi
Publicly Available Software . . . . . xii

**About This Book** . . . . . xiii
Who Should Use This Book . . . . . xiv
How This Book is Organized . . . . . xiv
Typographic Conventions . . . . . xv
Software Level Notation . . . . . xv
Accessing Online Information . . . . . xv
Obtaining Documentation . . . . . xvi
Related Publications . . . . . xvi
RS/6000 SP Publications . . . . . xvi
Parallel System Support Programs for AIX Publications . . . . . xvi
IBM Recoverable Virtual Shared Disk Publication . . . . . xvii
General Parallel File System for AIX Publication . . . . . xvii
LoadLeveler Publications . . . . . xvii
Performance Toolbox (PTX) Publication . . . . . xvii
Performance Monitor Publication . . . . . xvii
Parallel Environment for AIX Publications . . . . . xvii
Client Input Output/Sockets (CLIO/S) Publications . . . . . xvii
Parallel I/O File System Publications . . . . . xvii
Network Tape Access and Control System for AIX (NetTAPE) Publications xviii
Other IBM Publications . . . . . xviii
International Technical Support Organization Publications (Red Books) . . xviii
AIX and RS/6000 Publications . . . . . xviii
Service . . . . . xix
Network Queueing System/MVS (NQS/MVS) . . . . . xx
Network Connectivity . . . . . xx
RS/6000 SP Switch Router . . . . . xx
Adapters . . . . . xx
Non-IBM Publications . . . . . xxi
Parallel Computing . . . . . xxi
Tcl . . . . . xxi
Manual Pages for Public Code . . . . . xxi

**Part 1. Planning Your System** . . . . . 1

**Chapter 1. Introduction to System Planning** . . . . . 3
Planning Services . . . . . 3
Hardware Overview . . . . . 4
Software Overview . . . . . 6
What's New in AIX and PSSP? . . . . . 7
SP Planning Issues . . . . . 9
Using SP Books for Planning . . . . . 10

**Chapter 2. Defining the System that Fits Your Needs** . . . . . 11
Question 1: Why Do You Need an SP? . . . . . 11
Question 2: Do You Want a Preloaded SP or the Default Version? . . . . . 14


© Copyright IBM Corp. 1997, 1998

Question 3: What Related IBM Program Products Do You Need? . . . . . 18
Question 4: What Levels of AIX Do You Need? . . . . . 22
Question 5: What Type of Network Connectivity Do You Need? . . . . . 23
Question 6: What are Your Disk Storage Requirements? . . . . . 25
Question 7: What are Your Reliability and Availability Requirements? . . . . . 29
Question 8: How Many Nodes Do You Need? . . . . . 30
Question 9: Defining Your System Images . . . . . 43
Question 10: What Do You Need for Your Control Workstation? . . . . . 46

**Chapter 3. Defining the Configuration that Fits Your Needs** . . . . . 55
The Impact of Software Planning on Site Planning . . . . . 55
Planning Your Site Environment . . . . . 55
Planning Your System Network . . . . . 64
Determining Space Requirements . . . . . 71
Planning Your Network Configuration . . . . . 74
Understanding Node Numbering and Switch Node Numbering . . . . . 79

**Chapter 4. Planning for a High Availability Control Workstation** . . . . . 87
Overall System View of a High Availability Control Workstation . . . . . 87
Benefits of a High Availability Control Workstation . . . . . 89
Difference Between Fault Tolerance and High Availability . . . . . 89
IBM's Approach to High Availability for Control Workstations . . . . . 90
Related Reliability Options for Control Workstations . . . . . 92
Software Requirements for HACWS Control Workstation Configurations . . . . 94
Planning Your High Availability Control Workstation Network Configuration . . 95

**Chapter 5. Planning SP System Partitions** . . . . . 99
What is System Partitioning . . . . . 99
How Do You Partition the System? . . . . . 99
Example 1 -The basic 16-node system . . . . . 101
Using a Switch in a Partition . . . . . 102
Example 2 - A Switchless System . . . . . 105
The System Partitioning Aid . . . . . 105
Accessing Data Across System Partitions . . . . . 106
The Relationship of SP Resources to System Partitions . . . . . 106
Example 3 - An SP with 3 frames, 2 switches, and various node sizes . . . . 111
System Partitioning Configuration Directory Structure . . . . . 114

**Chapter 6. Planning for Security** . . . . . 117
Authentication in the SP System . . . . . 117
Authentication Worksheets . . . . . 127

**Chapter 7. Planning to Record and Diagnose System Problems** . . . . . 129
Configuring the AIX Error Log . . . . . 129
Configuring the BSD Syslog . . . . . 129
System Error Logs . . . . . 130
Finding and Using Error Messages . . . . . 131
Getting Help from IBM . . . . . 131
IBM Tools for Problem Resolution . . . . . 132

**Chapter 8. Planning for PSSP-Related LPPs** . . . . . 135
Planning for IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk . . . . . 135
Planning for Virtual Shared Disk Communications . . . . . 135

Planning for Parallel ESSL . . . 138
Planning for Parallel Environment . . . 138
Planning for Performance Toolbox Parallel Extension (PTPE) . . . 139
Planning for High Availability Cluster Multi-Processing (HACMP) . . . 140
Planning for LoadLeveler . . . 140
Planning for Parallel I/O File System . . . 142
Planning for NetTAPE . . . 142
Planning for IBM Client Input Output/Sockets (CLIO/S) . . . 143
Planning for General Parallel File System (GPFS) . . . 143

**Part 2. Customizing Your System** . . . 145

**Chapter 9. Planning for Expanding or Modifying Your System** . . . 147
Questions to Answer Before Expanding/Modifying/Ordering Your System . . . 147
Scenario 1: Expanding the Sample System by Adding a Node . . . 151
Scenario 2: Expanding the Sample System by Adding a Frame . . . 151
Scenario 3: Expanding the Sample System by Adding a Switch . . . 155

**Chapter 10. Planning for Migration** . . . 157
Developing Your Migration Goals . . . 158
Developing Your Migration Strategy . . . 162
Reviewing Your Migration Steps . . . 174

**Part 3. Appendixes** . . . 177

**Appendix A. The System Partitioning Aid - A Brief Tutorial** . . . 179
The GUI - "spsyspar" . . . 179
The CLI - "sysparaid" . . . 188
Example 3 of Chapter 5 . . . 191

**Appendix B. System Partitioning** . . . 203
8 Switch Port System . . . 203
16 Switch Port System . . . 203
32 Switch Port System . . . 206
48 Switch Port System . . . 210
64 Switch Port System . . . 211
80 Switch Port System With 0 Intermediate Switch Boards . . . 212
80 Switch Port System With Intermediate Switch Boards . . . 215
96 Switch Port System . . . 217
112 Switch Port System . . . 219
128 Switch Port System . . . 222

**Appendix C. SP System Planning Worksheets** . . . 227

**Glossary** . . . 245
Terms and Abbreviations . . . 245

**Index** . . . 253

# Figures


1. Typical SP Uses (not to scale) . . . 12
2. Node Layout Example For The ABC Corporation . . . 35
3. The ABC Corporation Node Layout Example With Communications Information . . . 36
4. Ethernet topology for a single-frame SP . . . 65
5. Ethernet topology for a single-frame SP . . . 65
6. Method 1 Ethernet topology for a multiple-frame SP . . . 66
7. Method 2 Ethernet topology for a multiple-frame SP . . . 67
8. Method 3 Ethernet topology for a multiple-frame SP . . . 68
9. Boot Server Frame Approach . . . 69
10. Node-slot assignment . . . 79
11. Supported frame configurations showing switch port assignments . . . 81
12. Node numbering and switch to switch frame connections when thin nodes are in use . . . 82
13. Node Numbering for an SP System . . . 83
14. Switch Port Numbering for an SP Switch-8 and Short Frame . . . 85
15. Switch Node Numbering Sequence . . . 86
16. High Availability Control Workstation With Disk Mirroring . . . 88
17. Initial control workstation Network Configuration . . . 96
18. Starting HACMP . . . 97
19. control workstation Failover . . . 98
20. Adding a System Partition . . . 98
21. Simple 1-frame system . . . 101
22. Partitioned 1-frame system . . . 102
23. Full switch board . . . 103
24. Nodes 11,12,15,16 partitioned off . . . 104
25. One sparse frame with no switch . . . 105
26. One frame with slots numbered . . . 109
27. Varied Node, 1-frame System . . . 110
28. Three frames with 2 switches . . . 111
29. The Directory Structure Of System Partition Information . . . 115
30. The control workstation as primary authentication server . . . 120
31. The control workstation as secondary authentication server . . . 121
32. The control workstation as client of authentication server . . . 122
33. Using AFS authentication services on the SP system . . . 123
34. Sample System: 3-frame, 1-switch . . . 148
35. spsyspar's Main Window . . . 180
36. Sample 1-Frame System (1 wide, 10 thin, and 1 high nodes) . . . 181
37. Main Window for Sample System . . . 182
38. Notebook for Node 8 of Sample System . . . 183
39. Notebook for Partition Alpha of Sample System . . . 184
40. Alpha Notebook for Sample System . . . 185
41. Descending sort in Nodes pane (Icon View) . . . 187
42. Filter menu with "1*" filter specified for Nodes pane . . . 188
43. inpfile.template provided with PSSP . . . 190
44. my_part_in file. . . . 190
45. Three frames with 2 switches . . . 192
46. Main Window for Example 3 of Chapter 5 . . . 193
47. System Partitioning for Example 3 of Chapter 5 . . . 195
48. System Partitioning for Example 3 of Chapter 5 . . . 196


© Copyright IBM Corp. 1997, 1998

49. Dialog box for specifying name of new layout . . . . . 197
50. Message issued when new layout has been saved . . . . . 197
51. CLI input file from "spsyspar" . . . . . 198
52. Alternate CLI input file . . . . . 199
53. Switch chips allocated to system partition Par1 . . . . . 200
54. Performance numbers for system partition Par1 . . . . . 201
55. SP Node Layout Worksheet For One Frame. . . . . 230
56. Extra SP Node Layout Worksheet For One Frame . . . . . 230
57. Node Layout Worksheet For Two Frames . . . . . 231

# Tables


1. Preliminary List of Applications For ABC Corporation . . . 18
2. IBM Program Products Ordered by ABC Corporation . . . 22
3. Direct Migration Paths . . . 23
4. Operating System Level Selected By ABC Corporation . . . 23
5. ABC Corporations's External Disk Storage Needs . . . 28
6. Requirements for the High Availability Control Workstation . . . 29
7. Function Checklist . . . 30
8. The Basic SP Frame and Switch Topology . . . 32
9. SP Node Summary . . . 32
10. Overall System Information . . . 33
11. ABC Corporations's Choices For Hardware Configuration By Node . . . 37
12. ABC Corporation's SP Ethernet . . . 38
13. ABC Corporation's Additional Adapters . . . 39
14. ABC Corporation's Choices For The Switch Configuration Worksheet . . . 40
15. PCI Adapters Supported Only on 332MHz Nodes . . . 41
16. MCA Adapters Supported on all but 332MHz Nodes . . . 41
17. ABC Corporation's Specifying the System Images . . . 44
18. File Set List for PSSP 2.4 . . . 45
19. SP Control Workstation . . . 52
20. SP Control Workstation Network . . . 53
21. Network Install Image Choices . . . 56
22. Time Service Choices . . . 58
23. User Directory Mounting Choices - System automounter support . . . 59
24. Print Management Choices . . . 60
25. User Account Management Choices . . . 61
26. System File Management Choices . . . 62
27. Accounting Choices . . . 63
28. LPP Source Directory Choices . . . 63
29. Space Required by installp Images . . . 72
30. Space Used by Individual File Sets . . . 73
31. Sample Switch Port Numbers for the SP Switch–8 . . . 85
32. Effect of Failure of Non-High Availability Control Workstation on Mandatory Software . . . 90
33. Effect of control workstation Failure on User Data on the Control Workstation . . . 91
34. Direct Migration Paths Supported . . . 159
35. Crossover Release Migration Paths Supported . . . 159
36. Levels of PSSP and AIX Supported in a Mixed System Partition . . . 164
37. Supported HACMP LPPs . . . 166
38. Supported Recoverable Virtual Shared Disk LPPs . . . 168
39. Supported Parallel Environment LPP Levels . . . 170
40. List of SP Planning Worksheets . . . 227
41. Preliminary List of Applications . . . 228
42. IBM Program Products to Order . . . 228
43. External Disk Storage Needs . . . 229
44. Overall System Information . . . 229
45. Hardware Configuration By Node . . . 231
46. SP Ethernet Node Network Configuration . . . 232
47. SP Additional Adapters Node Network Configuration . . . 233
48. Switch Configuration Worksheet . . . 234


© Copyright IBM Corp. 1997, 1998

49. PCI Adapters Supported on 332 MHz Nodes . . . . . 234
50. MCA Adapters Supported on all but 332 MHz Nodes . . . . . 235
51. Specifying the System Images (SPIMG) . . . . . 237
52. File Set List for PSSP 2.4 . . . . . 238
53. Control Workstation Image Worksheet . . . . . 239
54. Time Zones . . . . . 240
55. Control Workstation Network Worksheet . . . . . 241
56. Site Environment Worksheet . . . . . 242
57. PSSP Or Other Kerberos Authentication Servers . . . . . 243
58. Local Realm Information, PSSP Authentication Server . . . . . 244
59. AFS Authentication Servers . . . . . 244

# Notices

References in this publication to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program, or service. Evaluation and verification of operation in conjunction with other products, except those expressly designated by IBM, are the user's responsibility.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
500 Columbus Avenue
Thornwood, NY 10594
USA

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independent created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
Mail Station P300
522 South Road
Poughkeepsie, NY 12601-5400
USA
Attention: Information Request

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

## Trademarks

The following terms are trademarks of the IBM Corporation in the United States or other countries or both:

ADSTAR
AIX
APL2
C Set ++
DATABASE 2
DB2
ES/9000
ESCON
HACMP/6000
IBM
LoadLeveler

© Copyright IBM Corp. 1997, 1998

MVS
MVS/ESA
Magstar
Micro Channel
NQS/MVS
NetView
PROFS
POWERparallel
PowerPC
RS/6000
RS/6000 Scalable POWERparallel Systems
SP
Scalable POWERparallel Systems
Service Director
System/370
System/390
TURBOWAYS

AFS and DFS are trademarks of Transarc Corporation.

Approach, Domino, and Lotus Notes are trademarks of Lotus Development Corporation.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

Other company, product, and service names, which may be denoted by a double asterisk (**), may be trademarks or service marks of others.

## Publicly Available Software

This product includes software that is publicly available:

| | |
|---|---|
| **expect** | Programmed dialogue with interactive programs |
| **Kerberos** | Provides authentication of the execution of remote commands |
| **NTP** | Network Time Protocol |
| **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 SP system. The distribution for these products includes the source code and associated documentation. (Kerberos does not ship source code.) **/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.

# About This Book

This book helps you plan an IBM RS/6000 Scalable POWERparallel (SP) system installation that meets your programming requirements. 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. *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* provides information to help you understand hardware requirements, plan your physical environment, and prepare your site for hardware installation.

Other books that help you administer and use the SP system include:

- The *IBM Parallel System Support Programs for AIX: Administration Guide* provides information to help you understand and perform RS/6000 SP system administration. It includes concepts and instructions pertaining to:
  - Understanding IBM RS/6000 SP Systems
  - Performing routine system administration tasks
  - Making the system secure
  - Managing sets of nodes
  - Monitoring the system
  - Providing for system availability and problem management
  - Understanding and using the Communications Low-level Application Programming Interface (LAPI)
- The *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* helps you understand and perform SP installation, configuration, maintenance, and migration. It includes concepts and instructions pertaining to:
  - Installing and configuring the IBM Parallel System Support Programs for AIX (PSSP)
  - Installing on an existing RS/6000 system
  - Reconfiguring the system
  - Performing system maintenance
  - Migrating the system
- The *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* provides information to help you diagnose SP system problems and describes error messages, their meaning, and appropriate actions to take to recover from problems.
- The *IBM Parallel System Support Programs for AIX: Command and Technical Reference* provides detailed syntax and parameter information for all commands you can use to install, customize, and maintain the SP system.

This book applies to PSSP Version 2 Release 4. To find out what version of PSSP is running on your system, enter the following:

```
SDRGetObjects SP code_version
```

In response, the system displays something similar to:

```
code_version
PSSP-2.3
```

© Copyright IBM Corp. 1997, 1998

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 what version of PSSP is running on the nodes on your system, enter the following from your control workstation:

```
splst_versions -G -t
```

In response, the system displays something similar to:

```
1 PSSP-2.3
3 PSSP-2.3
7 PSSP-2.2
8 PSSP-2.2
```

**Save Your Old Manuals**

If you are running mixed levels of PSSP, be sure to maintain and refer to the appropriate documentation for whatever versions of PSSP you are running.

## Who Should Use This Book

This book is intended for those responsible for planning the network, control and system installation of an IBM RS/6000 SP 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.

## How This Book is Organized

This book is organized into several parts:

- Part 1, “Planning Your System” on page 1

  This section discusses a variety of topics including defining your system and configuration, planning for control workstations and system partitions, as well as security and problem diagnosis.

- Part 2, “Customizing Your System” on page 145

  This section discusses the planning considerations necessary when modifying your system by expanding it or migrating it to a different level of AIX and PSSP.

- Appendixes

  The first appendix contains a tutorial for using the System Partitioning Aid. The second appendix is a description of the system partitioning layouts. The third appendix contains system planning worksheets that you should copy, complete according to directions throughout the book, and save for future reference.

The back of the book includes a glossary and an index.

Vertical bars (|) to the left of the text in this book indicate changes or additions.

## 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.<br>• **Bold** words also indicate the first use of a term included in the glossary. |
| *Italic* | • *Italic* words or characters represent variable values that you must supply.<br>• *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.<br>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>**. |

## Software Level Notation

The meaning of notation when a software level is given in this book is demonstrated by the following examples:

**AIX 4.3** Any level of AIX version 4 release 3 and any modification level

**AIX 4.2.1** Only AIX version 4 release 2 modification level 1

**AIX 4.2.1 (or later)** AIX version 4 release 2 modification 1 or later modification levels

**PSSP 2.4 or later** PSSP version 2 release 4 with any modification level or later version and modification levels.

## Accessing Online Information

In order to use the PSSP man pages or access the PSSP online (HTML) publications, the **ssp.docs** file set must first be installed. To view the PSSP online publications, you also need access to an HTML document browser such as Netscape Navigator. An index to the HTML files that are provided with the **ssp.docs** file set is installed in the **/usr/lpp/ssp/html** directory.

## Obtaining Documentation

You can view this book or download a PostScript version of it from the IBM RS/6000 web site at **http://www.rs6000.ibm.com**. At the time this manual was published, the full path was **http://www.rs6000.ibm.com/resource/aix_resource/sp_books**. However, the structure of the RS/6000 web site can change over time.

---

## Related Publications

Here are some related publications.

## RS/6000 SP Publications

- *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*, GA22–7280
- *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22–7281
- *IBM RS/6000 SP: Maintenance Information, Volume 1, Installation and CE Operations*, GC23-3903
- *IBM RS/6000 SP: Maintenance Information, Volume 2, Maintenance Analysis Procedures and Parts Catalog*, GC23-3904

## Parallel System Support Programs for AIX Publications

- *IBM Parallel System Support Programs for AIX: Administration Guide*, GC23-3897
- *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*, GC23-3898
- *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide*, GC23-3899
- *IBM Parallel System Support Programs for AIX: Command and Technical Reference*, GC23-3900
- *IBM Parallel System Support Programs for AIX: Event Management Programming Guide and Reference*, SC23-3996
- *IBM Parallel System Support Programs for AIX: Group Services Programming Guide and Reference*, SC28-1675
- *IBM Parallel System Support Programs for AIX: Licensed Program Specification*, GC23-3901

As an alternative to ordering the individual books, you can use SBOF-8587 to order the above list from the SP software library. The following book includes the optional PSSP component Virtual Shared Disk and must be ordered separately.

- *IBM Parallel System Support Programs for AIX: Managing Shared Disks*, SA22–7279

## IBM Recoverable Virtual Shared Disk Publication

- *IBM Parallel System Support Programs for AIX: Managing Shared Disks*, SA22–7279

## General Parallel File System for AIX Publication

- *IBM General Parallel File System for AIX: Installation and Administration Guide*, SA22–7278

## LoadLeveler Publications

- *Using and Administering LoadLeveler*, SC23-3989
- *IBM LoadLeveler: Licensed Program Specification*, GH23-0040

## Performance Toolbox (PTX) Publication

- *IBM Performance Toolbox 1.2 and 2 for AIX: Guide and Reference*, SC23-2625

## Performance Monitor Publication

- *IBM Performance Toolbox Parallel Extensions for AIX: Guide and Reference*, SC23-3997

## Parallel Environment for AIX Publications

- *IBM Parallel Environment for AIX: Hitchhiker's Guide*, GC23-3895
- *IBM Parallel Environment for AIX: Operation and Use, Volume 1*, SC28-1979
- *IBM Parallel Environment for AIX: Operation and Use, Volume 2*, SC28-1980
- *IBM Parallel Environment for AIX: Installation Guide*, GC28-1981
- *IBM Parallel Environment for AIX: Messages*, GC28-1982
- *IBM Parallel Environment for AIX: MPL Programming and Subroutine Reference*, GC23-3893
- *IBM Parallel Environment for AIX: MPI Programming and Subroutine Reference*, GC23-3894
- *IBM Parallel Environment for AIX: Licensed Program Specification*, GC23-3896

As an alternative to ordering the individual books, you can use SBOF-8588 to order the entire Parallel Environment for AIX library.

## Client Input Output/Sockets (CLIO/S) Publications

- *IBM CLIO/S General Information*, GC23-3879
- *IBM CLIO/S Licensed Program Specification*, GC23-3789
- *IBM CLIO/S User's Guide and Reference*, GC28-1676

## Parallel I/O File System Publications

- *IBM AIX Parallel I/O File System: Fact Sheet*, G325-0649
- *IBM AIX Parallel I/O File System: Installation, Administration, and Use*, SH34-6065
- *IBM AIX Parallel I/O File System: Licensed Program Specification*, GH34-6066

## Network Tape Access and Control System for AIX (NetTAPE) Publications

- *IBM NetTAPE General Information*, GC23-3990
- *IBM NetTAPE User's Guide and Reference*, available from your IBM representative

---

## Other IBM Publications

Here are some other IBM publications that you may find helpful.

## International Technical Support Organization Publications (Red Books)

- *IBM International Technical Support Centers Implementing High Availability on RS/6000 SP*, SG24-4742
- *IBM International Technical Support Centers RS/6000 SP High Availability Infrastructure*, SG24-4838
- *IBM International Technical Support Centers RS/6000 SP PSSP 2.2 Technical Presentation*, SG24-4868
- *IBM International Technical Support Centers RS/6000 SP PSSP 2.3 Technical Presentation*, SG24-2080

## AIX and RS/6000 Publications

- *IBM AIX Version 4 Getting Started*, SC23-2527
- *IBM AIX Version 4.3 Quick Beginnings*, SC23-4114
- *IBM AIX General Concepts and Procedures for RS/6000*, GC23-2202
- *IBM AIX Version 4 Files Reference*, SC23-2512
- *IBM AIX Version 4.3 Files Reference*, SC23-4168
- *IBM AIX Version 4 System Management Guide: Communications and Networks*, SC23-2526
- *IBM AIX Version 4.3 System Management Guide: Communications and Networks*, SC23-4127
- *IBM AIX Version 4.1 Installation Guide*, SC23-2550
- *IBM AIX Version 4.2 Installation Guide*, SC23-1924
- *IBM AIX Version 4.3 Installation Guide*, SC23-4112
- *IBM AIX Version 4 Commands Reference*, SBOF-1851 (all volumes)
- *IBM AIX Version 4.3 Commands Reference*, SBOF-1877 (all volumes)
- *IBM AIX Versions 3.2 and 4 Performance Tuning Guide*, SC23-2365
- *IBM AIX Version 4.3 Messages Guide and Reference*, SC23-4129
- *IBM AIX Version 4 Messages Guide and Reference*, SC23-2641
- *IBM AIX Version 4.1 Network Installation Management Guide and Reference*, SC23-2627
- *IBM AIX Version 4.2 Network Installation Management Guide and Reference*, SC23-1926

- *IBM AIX Version 4.3 Network Installation Management Guide and Reference*, SC23-4113
- *IBM AIX Version 4 System Management Guide: Operating System and Devices*, SC23-2525
- *IBM AIX Version 4.3 System Management Guide: Operating System and Devices*, SC23-4126
- *IBM AIX Version 4 General Programming Concepts: Writing and Debugging Programs*, SC23-2533
- *IBM AIX Version 4.3 General Programming Concepts: Writing and Debugging Programs*, SC23-4128
- *IBM AIX Version 4 Communications Programming Concepts*, SC23-2610
- *IBM AIX Version 4.3 Communications Programming Concepts*, SC23-4124
- *Diskless Workstation Management Guide*, SC23-2433
- *C++ for AIX/6000: Language Reference*, SC09-1606
- *C++ for AIX/6000: Standard Class Library Reference*, SC09-1604
- *C++ for AIX/6000: User's Guide*, SC09-1605

## Service

- *IBM AIX Version 3.2 Problem Solving Guide and Reference*, SC23-2204
- *IBM AIX Version 4 Problem Solving Guide and Reference*, SC23-2606
- *IBM AIX Version 4.3 Problem Solving Guide and Reference*, SC23-4123
- *Information Manual for Diagnostics Micro Channel Bus*, SA23-2765
- *7012 300 Series Operator Guide*, SA23-2623
- *7012 300 Series Installation and Service Guide*, SA23-2624
- *7013 500 Series Operator Guide*, SA38-0530
- *7013 500 Series Installation and Service Guide*, SA38-0531
- *Electrical Safety for IBM Customer Engineers*, S229-8124
- *Service Information Binder*, SX33-6060
- *7015 Model R30 CPU Enclosure Operator Guide*, SA23-2742
- *7015 Model R30 CPU Enclosure Installation and Service Guide*, SA23-2743
- *Supplemental Information for 7012 G Series Models, 7013 J Series Models, 7015 Models R30 and R50* (No order number — this book comes as part of the ship group with the SP hardware)
- *7025 F30 Series User's Guide*, SA38-0504
- *7025 F30 Series Service Guide*, SA38-0505
- *7025 F40 Series User's Guide*, SA38-0514
- *7025 F40 Series Service Guide*, SA38-0515
- *7025 F50 Series User's Guide*, SA38-0540
- *7025 F50 Series Service Guide*, SA38-0541
- *7026 CPU Drawer User's Guide*, SA38-0534

- *7026 CPU Drawer Installation and Service Guide*, SA38-0535
- *7043 43P Series Models 140 and 240 User's Guide*, SA38-0511
- *7043 43P Series Models 140 and 240 Service Guide*, SA38-0512

## Network Queueing System/MVS (NQS/MVS)

- *IBM NQS/MVS Installing and Administering*, SC23-0232
- *IBM NQS/MVS Client User's Guide*, SC23-0231
- *IBM NQS/MVS General Information*, GC23-3684
- *IBM NQS/MVS Licensed Program Specifications*, GC23-0233

## Network Connectivity

- *IBM LAN Cabling System Planning and Installation Guide*, GA27-3361
- *IBM Cabling System Optical Fiber Planning and Installation Guide*, GA27-3943
- *IBM 8250/8260/8285 Planning and Site Preparation Guide*, GA33-0285
- *IBM 6611 Network Processor: Introduction and Planning Guide*, GK2T-0334

## RS/6000 SP Switch Router

Order according to RS/6000 SP Switch Router model:

- *Getting Started*
- *Configuration Guide*
- *Reference Guide*

You can order the RS/6000 SP Switch Router as the IBM 9077.

## Adapters

- *SP Switch Router Adapter Guide*, GA22-7310
- *FDDI Introduction and Planning Guide*, GA27-3892
- *FDDI User's Guide and Programming Reference*, SC28-2823
- *Planning for Fiber Optic Channel Links*, GA23-0367
- *IBM Token-Ring Network Introduction and Planning Guide*, GA27-3677
- *RS/6000 Token Ring Adapter Card*, G511-1681
- *HiPPI User's Guide and Programmer's Reference*, SA23-0369 and SA23-2488
- *AIX Parallel and ESCON Channel Tape Attachment/6000 Installation and User's Guide*, GA32-0311
- *9334 SCSI Expansion Units Operator Guide*, GA33-3232
- *9334 SCSI Model 010 and 011 Expansion Unit: Installation and Service Guide*, SY33-0165
- *9334 Model 500 and 501 SCSI Expansion Unit: Installation and Service Guide*, SY33-0167
- *SCSI-2 Fast/Wide Adapter*, SC23-2646
- *IBM SCSI-2 Fast/Wide Adapter/A Technical Reference*, S83G-7545

- *Turboways 100 User's Guide ATM*, GA27-4057
- *9333 Model 010 and 011 High-Performance Disk-Drive Subsystem Operator Guide*, GA33-3208
- *9333 Model 010 and 011 High-Performance Disk-Drive Subsystem Installation and Service Guide*, SY33-0161
- *9333 Model 010 and 011 High-Performance Disk-Drive Subsystem Hardware Technical Information*, SA33-3209
- *9333 Model 500 and 501 High-Performance Disk-Drive Subsystem Operator Guide*, GA33-3234
- *9333 Model 500 and 501 High-Performance Disk-Drive Subsystem Installation and Service Guide*, SY33-0168
- *9333 Model 500 and 501 High-Performance Disk-Drive Subsystem Hardware Technical Information*, SA33-3235
- *IBM SCSI Tape Drive, Medium Changer, and Library Device Drivers Installation and User's Guide*, GC35-0154

## Non-IBM Publications

Here are some non-IBM publications that you may find helpful.

## Parallel Computing

- Almasi, G., Gottlieb, A., *Highly Parallel Computing*, Benjamin-Cummings Publishing Company, Inc., 1989.
- 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.
- Foster, I., *Designing and Building Parallel Programs*, Adison Wesley, 1995.
- Pfister, Gregory, F., *In Search of Clusters*, Prentice Hall, 1995.

## Tcl

- Ousterhout, John K., *Tcl and the Tk Toolkit*, Addison-Wesley, Reading, MA, 1994, ISBN 0-201-63337-X.

## 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 |
| **NTP** | /usr/lpp/ssp/man/man8/xntpd.8 |
| | /usr/lpp/ssp/man/man8/xntpdc.8 |
| **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

**Perl (Version 5.003)** Man pages are in the /usr/lpp/ssp/perl5/man/man1 directory

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.

# Part 1. Planning Your System

© Copyright IBM Corp. 1997, 1998

# Chapter 1. Introduction to System Planning

IBM RS/6000 SP Systems are not just hardware and software. SP systems are also a continually changing set of human requirements. In order to get the highest level of performance out of your SP system, you need to plan for all of the internal and external activities. SP system planning produces the solid foundation you will need for managing your RS/6000 SP 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 partitions
- Migration and coexistence on existing systems
- Security and authentication
- Defining user accounts
- Backup procedures

This chapter gets your project team started on these planning tasks. As an added benefit, it familiarizes them with the SP System and how it can be integrated into your operational environment.

If you have an SP system and want to move to a different level of AIX and PSSP software, you might also need to plan for migration; possibly using coexistence or system partitioning.

If you need to, you can contract with IBM to plan and install your SP system. Contact your IBM representative if you want help with these tasks.

**Save Your Old Manuals**

If you are running mixed levels of PSSP in a system partition, be sure to maintain and refer to the appropriate documentation for whatever versions of PSSP you are running.

## Planning Services

This optional IBM service offering provides a specialist on site to assist you with planning your implementation. Activities included as part of this offering include:

- Planning for integrating 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.

© Copyright IBM Corp. 1997, 1998

IBM representatives and IBM Business Partners can also obtain information on selecting the right type of node for a customer's system by referring to the document *Node Selection for the IBM SP System - Factors to Consider*.

Author: Clive Harris
EMEA AIX Consultant
EMEA High-End Centre of Competence
IBM Basington, UK
Telephone (44–01256) 343184
PROFS: NHBVM1(HARRIC2)
email: clive_harris@uk.ibm.com

## Hardware Overview

The basic hardware components of an SP system are:

- Processor nodes
- Frames
- Optional switch
- A control workstation
- Network connectivity adapters

These components connect to your existing computer network through a Local Area Network (LAN), making the SP system accessible from any network-attached workstation.

**Hardware is described in Volume 1.**

This is merely an overview. Each type of hardware has its set of requirements. Be sure to see *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for physical specifications, requirements, and valid configurations.

## Processor Nodes

SP nodes are available in three types: thin nodes, wide nodes, and high 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 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. Systems that can have from 128 to 512 nodes are available by special bid.

SP processor nodes are uniprocessor or symmetric multiprocessor (SMP) systems available with varying levels of function, capacity and performance. Each processor node includes memory, direct access storage devices (DASD), and a method for Ethernet connection. The type of node and optional equipment it contains can lead to other requirements.

You should 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 data base operations. However, no rigid rule governs the logical configuration of a node. You can configure a physical node type for the logical functions that best serve your computing requirements.

A single tall frame can contain from one to 16 processor nodes, depending on the node type. The maximum number of high nodes that are supported on a 128–node system is 64.

## Extension Nodes

Extension nodes are non-standard nodes that extend the SP system's capabilities but cannot be used in all of the same ways as standard SP 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 standard nodes use is implemented on the SP Switch.

A physical dependent node, such as an SP Switch Router, can support multiple dependent node adapters. If a dependent node like an SP Switch Router contains more than one dependent node adapter, it can route data between SP system partitions. Data transmission is accomplished by linking the dependent node adapters in the SP Switch Router with valid switch ports on an SP Switch. If these SP Switches are located in different SP system partitions, data can be routed at high speed between the system partitions.

Also, 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.

Extension Nodes **require** a control workstation (CWS), primary node, and a primary backup node operating PSSP 2.3 or later.

## Frames

SP processor nodes are mounted in either a tall or short frame. The frame spaces that nodes fit into 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 wide node occupies one drawer (two slots) and a high node occupies two drawers (four slots). An internal power supply is included with each frame. Frames get equipped with the optional processor nodes and switches that you order.

## Switches

Switches are used to connect nodes, providing the message passing network through which SP processor nodes communicate with a minimum of four disjoint paths between any pair of nodes. The switches currently used are known as the SP Switch and the High Performance Switch.

### SP Switch

The SP Switch provides low latency, high-bandwidth communication between nodes. It consists of a switch assembly and the internal cables to support connection to sixteen processor nodes in a system (one switch per frame). The SP Switch offers the following improvements over the High Performance Switch series:

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

The SP Switch Adapter connects each SP node to the SP Switch subsystem. The SP Switch Router can be used to dramatically speed up TCP/IP, file transfers, remote procedure calls, and relational database functions.

**Note:** An SP Switch Router Adapter is needed to connect the SP Switch Router to the SP Switch.

### High Performance Switch

High Performance Switches predate and are not compatible with SP switches. You cannot mix High Performance switches with SP switches in an SP system, not even in separate system partitions. To take advantage of the SP Switch's performance, you must upgrade all your switches to SP Switches.

High Performance Switches are being phased out and are available for MES upgrade orders only. They cannot be ordered for new SP system builds.

## Control Workstation

The SP system requires a customer-supplied IBM RS/6000 workstation with a color monitor. The control workstation serves as a central point-of-control for managing and maintaining the SP processor nodes. The control workstation connects to each frame via an RS-232 line to provide hardware control functions. A system administrator can log in to the control workstation from any other workstation on the network to perform system management, monitoring, and control tasks.

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

## Network Connectivity Adapters

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.

---

# Software Overview

The SP system software infrastructure includes:

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

## IBM AIX

AIX provides operating system functions such as the AIXwindows user interface, extended real-time support, network installation management, advanced file system support, physical disk space management, and a platform for application development and execution. AIX provides UNIX functionality and conformance to industry standards for open systems.

## IBM Parallel System Support Programs for AIX (PSSP)

PSSP provides the functions required to manage an IBM RS/6000 SP System as a full-function parallel system. PSSP provides a single point of control for administrative tasks and helps increase productivity by letting administrators view, monitor, and control system operation. The following are also available as priced features of PSSP:

- Performance Toolbox Parallel Extension for AIX (PTPE)
- High Availability Control Workstation Connectivity (HACWS)

## Extension Node Support

Extension nodes are non-standard nodes that you can add to your system for higher-bandwidth communications. To assist you in attaching an SP Switch Router with one or more SP Switch Router Adapters to the SP Switch, configuration is supported via SMIT panels and SNMP, and operation is supported via standard SP Switch commands.

---

## What's New in AIX and PSSP?

New RS/6000 SP systems come with AIX 4.2.1 (or later) or AIX 4.3, and PSSP 2.4 installation tapes. The following sections briefly describe the new and changed functions in AIX 4.2, AIX 4.3, and PSSP 2.4. You should plan to order the system level that supports the functions you need.

## What's New in AIX 4.2?

AIX 4.2 adds new features that increase your system's usefulness. For complete information, see your IBM representative. The following is a partial summary:

- Support for newly announced and most current RS/6000 systems and adapters.
- Support for the IBM Network Station.
- Support for file sizes up to 64GB, allowing users greater freedom in application development and in the creation and use of larger data sets.
- Device support up to 1 terabyte, increasing attachment options and system capabilities.
- Graphical Workspace Manager iconifies active applications on a Common Desktop Environment to increase user productivity through ease-of-use features.
- Auto-loading X-Windows simplify installation.
- Year 2000 compliant
- Network Installation Manager guides users through software installation and distribution on SP nodes and external workstations.

- Support for definable multiple boot images and root volume groups as a fall back mechanism in case of problems.
- Includes Adobe Acrobat and Netscape commercial products giving users easy access to online documentation.
- A 60–day evaluation copy of ADSM is added.
- The IBM Internet Connections Secure Server 4.2 for AIX is added.

## What's New in AIX 4.3?

AIX 4.3 is the next evolutionary step in scalability, compatibility, connectivity, interoperability, and usability. For complete information, see your IBM representative. The following is a partial summary:

- A new computing environment supporting 64–bit exploitation.
- 32–bit or 64–bit application coexistence and concurrent execution for those who plan to implement 64–bit technology in the future.
- A high level of binary compatibility that helps protect your investment in existing RS/6000 computing environments. The following programs compiled on AIX 4.1 or 4.2 are among those supported on AIX 4.3 by exploiting AIX's binary compatibility:
  - High Availability Cluster Multi-Processing for AIX, Enhanced Scalability (HACMP/ES)
  - LoadLeveler
  - Network Tape Access and Control System for AIX (NetTAPE)
  - NetTAPE Tape Library Connection for AIX (NetTAPE TLC)
  - Parallel Environment for AIX
  - Parallel Engineering and Scientific Subroutine Library (ESSL) for AIX
  - Parallel System Support Programs for AIX
  - Recoverable Virtual Shared Disk
- An Internet- and intranet-ready operating environment.
- Online HTML-based AIX publications.

## What's New in PSSP 2.4?

Version 2 Release 4 of PSSP provides enhanced quality and support for the enablement of new SP nodes, AIX 4.2.1 (or later), and AIX 4.3. Functional enhancements in this release include:

- New Hardware Support:
  - 332 MHz SMP Wide and Thin nodes

    These are new 2– or 4–way PCI-based SMPs, utilizing PowerPC technology. These nodes provide outstanding commercial price/performance as a follow-on to the micro channel-based (MCA) SMP nodes of RS/6000 SP systems that are currently available. They are well suited to address:
    - mission-critical online transaction processing (OLTP)

- collaborative computing comprised of Lotus Notes, Domino Server, Internet, intranet, and groupware solutions
- mixed integer/floating point computing such as engineering analysis and petroleum research
- high I/O bandwidth requirements
- data mining and data warehouse applications which benefit from the integer computing performance of the new nodes

– SP Switch MX Adapter

Provides switch communication with high SP Switch bandwidth and low protocol latency.

– Frames have been simplified to two models, short and tall.

– New PCI adapters are introduced to support the new PCI-based nodes.

See *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for complete hardware descriptions.

- Migration and coexistence

  Support is provided for upgrading to PSSP 2.4 running on AIX 4.2.1 (or later) or on AIX 4.3.

  Support is provided for the coexistence of earlier releases of PSSP in a system partition with PSSP 2.4.

- Improved switch diagnostics

  Includes diagnostic documentation, messages, and more detailed entries in the AIX error log and switch log for better problem isolation.

While not part of PSSP, the following are related product enhancements:

- Recoverable Virtual Shared Disk

  Communication adapter recovery has been added, allowing SP Switch adapter or Ethernet adapter failure to be treated the same as node failure; that is, control of connected twin-tailed volumes is passed to the backup server.

- Service Director for RS/6000

  This set of IBM software applications that monitor the health of your SP system is now a standard deliverable with SP systems. It analyzes AIX error logs, runs diagnostics, and notifies you and IBM if certain problems occur.

## SP Planning Issues

If you are new to SP systems, you should read all of this book. The planning steps you take depend on where you are now and what system you want to end up with. Chapter 2, “Defining the System that Fits Your Needs” on page 11 helps you define an SP that meets your needs. The following list contains some of the major issues you need to consider when setting up your SP system.

- The type of computing your SP system will perform.
- Future expansion (scaling) plans for your system.
- The number of nodes you will need.
- The type of nodes you will need.

- The type of RS/6000 you can use for a control workstation.
- High Availability (system backup) requirements for data and hardware.
- Migration for system upgrades.
- The ability to use partitioning and coexistence as migration tools.
- Coexistence requirements and limitations.
- Possible operational benefits from using the Parallel Environment.
- The amount and type of data storage.
- External network connections for your SP system.

Remember, as your planning begins to shape your SP system, you will need to work closely with your hardware planners. Each software decision you make will create a corresponding requirement in hardware such as:

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

## Using SP Books for Planning

To help you plan your SP system, there are two volumes in the library. Work closely with your hardware planner and these two volumes because choices of hardware can lead to requirements on software, and choices of software can lead to requirements on hardware.

Planning books are available in the *IBM RS/6000 SP Planning* library:

- Use *Volume 1, Hardware and Physical Environment* to make sure you have the correct physical configuration and environment for your SP system.
- Use this book, *Volume 2, Control Workstation and Software Environment*, to plan your control workstation and software environment. Make your decisions about what components to install, which nodes to use for what purposes, and how to plan system upgrades using migration, coexistence, and system partitioning.

To help you administer and use your SP system, the folllowing are available in the *IBM Parallel System Support Programs for AIX* library:

- Use the *Administration Guide* book for the day-to-day running and management of your system.
- Use the *Installation and Migration Guide* book for new installations and system updates.
- Use the *Command and Technical Reference* book as a command and technical reference.
- Use the *Diagnosis and Messages Guide* book to help diagnose problems and to get more information about messages.

# Chapter 2. Defining the System that Fits Your Needs

This chapter helps you define a new RS/6000 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 prompted 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 SP to a network has important benefits because networking information is critical to the success of the RS/6000 SP installation. Network planning might seem at times to be complex but it is a 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 SP you'll make many decisions. The remainder of this chapter poses several questions for you to answer. Review these questions and become familiar with the types of information you will need to gather throughout the planning process:

This chapter contains sample worksheets for a hypothetical corporation called the `ABC Corporation`. Review these sample worksheets to see the decisions that the `ABC Corporation` made. Whenever decisions made by the `ABC Corporation` are shown, they will be in `constant width` format to help distinguish them from the regular text.

Your decisions will most likely be different from the `ABC Corporation`'s. This is natural since every company is different and the decisions you make should meet your corporation's needs.

As you go through these questions, fill in the worksheets in Appendix C, “SP System Planning Worksheets” on page 227 with the information about your system. Make several copies of all the worksheets first. You can change your mind as you go through the worksheets. You'll need these worksheets later when you want to add to your system.

## Question 1: Why Do You Need an SP?

Why do you need an SP? For LAN consolidation? For data mining? For engineering or scientific computing? See Figure 1 on page 12 for some typical SP uses.

© Copyright IBM Corp. 1997, 1998

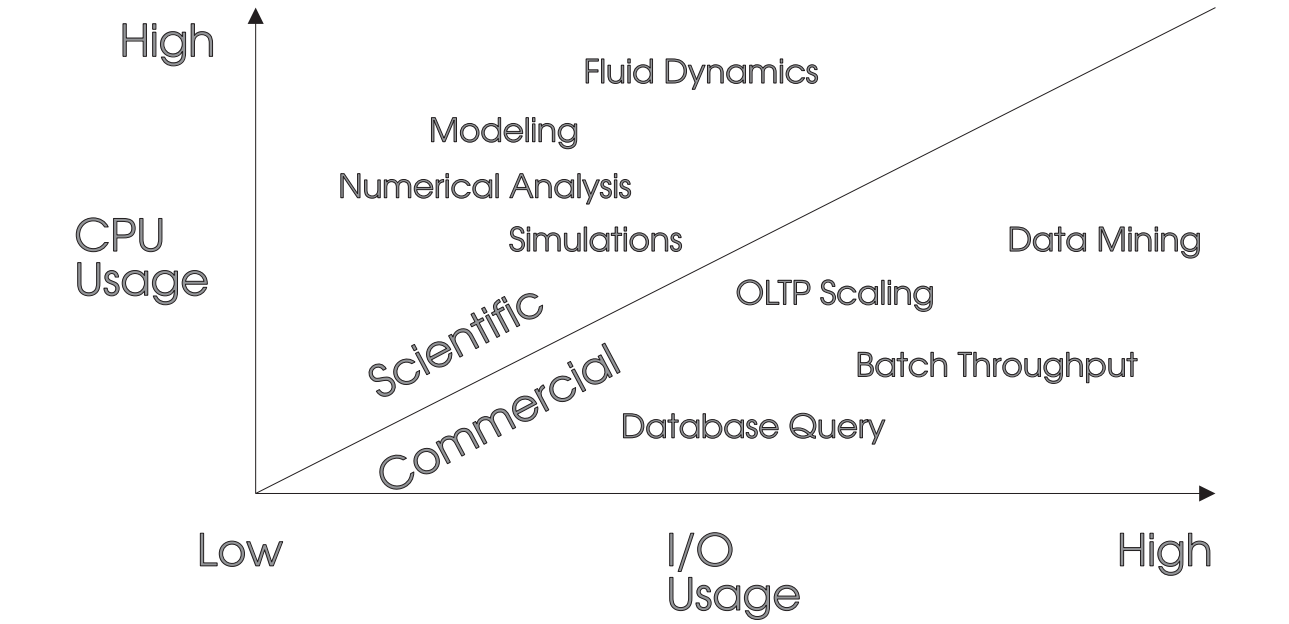



*Figure 1. Typical SP Uses (not to scale)*

What applications do you want to run on this system? Each SP node is a powerful uniprocessor or SMP running the AIX operating system and some special SP support programs. Thousands of IBM RS/6000 applications can run unchanged on the RS/6000 SP with a single point of control for system management.

## 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 nodes of the RS/6000 SP system. In this case, various parallel tools such as compilers, libraries and debuggers are required.

## Choosing a Switch

Now that you have considered the types of applications you will run on the system, you are ready to decide whether you can benefit from an SP Switch.

| Switch Feature | Description |
|---|---|
| SP Switch | This switch (feature code 4011) offers 32 connections, 16 internal and 16 external. It connects all the processor nodes, providing enhanced scalable high-performance communication between processor nodes for parallel job execution. |
| SP Switch-8 | This switch (feature code 4008) offers 8 internal connections to provide enhanced functions for small systems (up to 8 total nodes). It does not support scaling to larger systems. |

### SP Switch

The optional SP Switch provides low latency, high-bandwidth communication between nodes, supplying a minimum of four paths between any pair of nodes. Switches dramatically speed up TCP/IP, file transfers, remote procedure calls, and relational database functions. It consists of a switch assembly and the internal cables to support connection to 16 processor nodes in a system (one switch per frame). The SP Switch Adapter connects each SP node to the SP Switch subsystem. SP Switch capabilities include:

- Interframe connectivity and communication
- Scalability up to 128 node connections, including intermediate switch frames
- Constant bandwidth and latency between node pairs
- Support for Internet Protocol (IP) communication between nodes
- IP Address Resolution Protocol (ARP) support
- Support for dedicated user space access (via message passing libraries) and multiuser environments (via IP)
- Error detection and retry
- Fault isolation
- Concurrent maintenance for nodes

### High Performance Switch

High Performance Switch technology predates the SP Switch series. High Performance Switch networks (Feature 4007/10 and 4018) are available for MES upgrade orders only. They cannot be ordered for new SP system builds. If you plan on adding any of the new 332 MHz SMP wide and thin nodes supported by PSSP 2.4, you must upgrade all your switches to the SP Switch.

### Switch Incompatibility

Although all currently supported levels of PSSP support both the SP Switch and the High Performance Switch, the two switch networks are not compatible physically and cannot be mixed within an SP system, not even in a partitioned SP system.

## Question 2: Do You Want a Preloaded SP or the Default Version?

You have the option of ordering your SP 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 several manufacturing-based services available from which you can choose when you decide to order an SP:

1. Standard Default Tape Service

   By default, every SP system is delivered with the most current software level, including Program Temporary Fix (PTF), of AIX and PSSP. You will receive a MKSYSB backup tape for the control workstation and nodes. You also receive a node image tape for backup purposes. The control workstation and node tapes are delivered with your SP system. The tape for the control workstation contains installed versions of AIX and PSSP and a LoadLeveler install image.

   The version of AIX and PSSP installed is determined by one of the following feature codes:

   - FC #9434: AIX 4.3.0 and PSSP 2.4
   - FC #9424: AIX 4.2.1 and PSSP 2.4
   - FC #9422: AIX 4.2.1 and PSSP 2.2
   - FC #9410: AIX 4.1.5 and PSSP 2.2

   **Note:** If you do not specify a feature code, IBM will provide AIX 4.2.1 and PSSP 2.4.

   Choosing this option means you will do the loading and customization work including network configuration and the installation of any additional Licensed Program Products (LPPs). You will have to do a complete installation of **all** MKSYSB images from tape for **both** the nodes **and** the control workstation.

2. Customized Preload Service Feature Code 1250

   When you order this service, IBM installs AIX and PSSP components, as well as other IBM LPPs that you specify, onto the SP nodes and onto the control workstation image. This service also performs network customization on the SP prior to its arrival at your site. You can obtain a list of supported LPPs from the the internet. To access the web site enter the following URL:

   `http://www.rs6000.ibm.com/software/Apps/LPPmap.html`

   Customized preload is free of charge provided you have an installation service contract**. To get this service, you must also order the following:**

   - The installation service contract offered by the AIX Family of Services, Availability Center, or a certified IBM business partner.
   - Feature code 1250 on the SP (no charge).

   IBM will deliver customized MKSYSB backup tapes that you can use to install the control workstation and restore each of the nodes.

   When you order this feature code, the AIX Family of Services representative will work with you to provide the necessary customization information to Manufacturing. To ensure complete customization, you must complete and return online forms at least ten business days prior to the scheduled ship date. These are the forms that are sent to you by the SP2INST Service Machine after you place your order for an SP system. If returning the online forms at least ten business days prior to the scheduled ship date is not possible, IBM

might be able to perform network customization including the standard AIX default image, at least five business days prior to the scheduled ship date from manufacturing. If no information is received within five business days of the ship date, the order will be altered to remove feature code 1250 and provide the standard default tape service instead.

3. Customized Preload Service Feature Code 1251

   IBM offers feature code 1251 (additional charge) which preloads the RS/6000 SP Mission Critical POWERsolution for Oracle and BEA Tuxedo onto a customized SP system (one which has first been preloaded using Feature Code 1250). The solution combines the IBM RS/6000 SP server, the Oracle Parallel Server database, and the BEA Tuxedo transaction processing monitor. Additional optional software can also be preloaded with Feature Code 1251, either IBM ADSTAR Distributed Storage Manager (ADSM) or Spectra Logic Alexandria backup and restore software.

   This preload service **must** be part of the service contract. Ordering this feature assumes that you have obtained the appropriate licensing either from IBM, IBM Business Partners, or directly from the respective software vendors for your SP install. Feature Code 1250 is a **prerequisite** to this feature.

   After IBM receives your customization information, IBM manufacturing will customize network information (including hostnames and IP addresses), install customer selected AIX and PSSP components and selected LPPs on the control workstation image and on the SP nodes.

4. Advanced SP Customization Services (based on cost-recovery charges)

   Additional customization services are available both in-house and on-site according to your requirements. The Customized Solutions Group has a team comprised of skilled resources who are able to deliver the following services:

   - Pre-install Planning which includes a review of the control workstation, frame, network, and node configurations, the control workstation support matrix, and planning for SP frame upgrades.
   - Design reviews focused on how to implement application suites on specific SP hardware configuration.
   - 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 Lotus Notes implementation with HACMP.
   - DB/2 Parallel Edition implementation.

## Steps to Receive Customized Preload Service

1. Order the installation services contract by contacting the AIX Support Family Project Office at 1-800-CALLAIX (1-800-225-5249) or a certified IBM business partner.
2. Select feature code 1250 from the configurator.
3. Optionally, select feature code 1251 from the configurator.

4. After placing the order for an SP (machine type 9076) on the configurator, having selected feature codes 1250 and 1251, IBM sends a note and two required files to the person who configured the order. That person is requested to pass these on to the AIX Family of Services representative assigned to the order. The two files that are sent are:

   - *orderno* WKBKSOFT
   - *orderno* WKBKNET

   where *orderno* corresponds to the six digit manufacturing order number.

   If these files were not received, generically named versions are available to your IBM representative using a request command during a VM user session (see "Contacting the Customized Solution Organization" on page 17).

   After receiving the files, you should rename them so that the file name corresponds to the six digit manufacturing order number as follows:

   ```
   RENAME SP2INST WKBKSOFT A orderno WKBKSOFT A
   RENAME SP2INST WKBKNET A orderno WKBKNET A
   ```

5. The account team and installation service representative will work with you to fill in the information required in the two files. These files are online forms requesting information similar to the worksheets contained in this book. You can prepare to complete the online forms by first following the guidance in this book and completing the worksheets.
6. Edit these files and fill in the information that is required. Ensure the files have a file name that corresponds to the six digit purchase order number for the order.
7. Return the two files to SP2INST at KGNVMC or SP2CUST@VNET.IBM.COM at least ten business days prior to the scheduled ship date from manufacturing.

If the two files are not returned at least ten business days prior to the scheduled ship date from manufacturing, IBM might be able to perform network customization including the standard AIX default image, at least five business days prior to the scheduled ship date from manufacturing. However, in some cases, it might not be possible to perform any of the customization requested. In that case, the order will be altered to remove feature code 1250 and IBM manufacturing will provide the standard default tape service instead. The IBM installation service representative can perform any further customization that is required at the customer site.

## What You Get as a Package

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

- If you get the Standard Default Tape Service, the SP nodes are not preloaded. You receive tapes that contain the PSSP and AIX software with instructions for you to load the workstation and the nodes. Use the MKSYSB backup tape for the control workstation and nodes. You also receive a node image tape for backup purposes.
- If you get any variety of customized preload, the SP nodes are preloaded with the software and networking parameters per your order. The control workstation is not preloaded. You receive a customized MKSYSB control workstation tape

and must follow the instructions to load the workstation. You also receive a node image tape for backup purposes.

To further modify or customize your nodes, see *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* and *IBM Parallel System Support Programs for AIX: Administration Guide*.

## Contacting the Customized Solution Organization

If you have questions about these preload services, please contact the Poughkeepsie Customized Solution Organization. You can call, send a note, or request an instructional package. 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 VM userid JUSTASK at PKEDVM9
  - have your IBM representative request instruction files by typing on the command line of a VM session:

    ```
    REQUEST SP2INST FROM SP2INST AT KGNVMC
    ```
- from EMEA, do any of the following:
  - call 33–4–6734–4679
  - have your IBM representative send a note to VM userid 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 started considering the software, you should begin a list of the applications you want on your SP system. If you already know that you want to use the preload service, request the required forms. Whether or not you use the preload service forms, it helps to use the worksheets in this book to develop your list as you progress in the planning process.

List on Worksheet 1, “Preliminary List of Applications” in Table 41 on page 228, the applications you are considering. Check **parallel** √ and **Need Switch** √ if you know you need these functions. Put “?” 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, `Corporation ABC` looked at their application requirements and filled in Table 1 on page 18.

| *Table 1. Preliminary List of Applications For ABC Corporation* | | |
|---|---|---|
| **SP Preliminary List of Applications - Worksheet 1** | | |
| **Application** | **Parallel** | **Need Switch** |
| | | |
| DB2 Parallel Edition | √ | √ |
| AIX Performance Toolkit | | |
| Customer Written Application | √ | √ |
| | | |
| | | |
| | | |
| | | |
| | | |

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

## Question 3: What Related IBM Program Products Do You Need?

There are two sets of IBM Program Products that you must decide on. One set includes programs that are part of your SP environment and the other includes programs that are for the AIX operating system and IBM C or C Set ++ compilers that run on each node.

IBM C for AIX, V3 (5765-423) or IBM C Set ++ for AIX, V3 (5765–421) is a prerequisite of PSSP. It is necessary for service of PSSP. Without the compiler's preprocessor, dump diagnosis tools such as **crash**, will not function fully. You need at least a one-user license, but if you intend to do C development work, you will have to decide how many users you want to support at a given time.

Parallel System Support Programs (PSSP) for AIX is comprised of software components that are an integral part of your RS/6000 SP. It is the software that makes a collection of RS/6000 nodes into an SP. PSSP helps a system administrator manage the RS/6000 SP. It provides a single point of control for administrative tasks and helps increase productivity by letting administrators view, monitor, and control system operation.

There are many more LPPs available to run on your SP that can add to your enterprise's productivity. You can see a list on the Internet at URL

`http://www.rs6000.ibm.com/software/Apps/LPPmap.html`

However, the SP software suite has a number of optional LPPs that are closely related to PSSP. Each of those products is briefly described here. If you think one of the following products might provide a service you want on your SP, see Chapter 8, “Planning for PSSP-Related LPPs” on page 135 for planning information. If you have more questions, ask your IBM representative. At the end of this section, you will find a worksheet where you can list the program products you want.

## IBM LoadLeveler

(Program Number 5765-145.) LoadLeveler is an IBM software product that lets you build, submit, and process both serial and parallel jobs on your RS/6000 SP system or RS/6000 workstations. LoadLeveler is recommended, but not required, for batch processing on the SP system, and is included with the SP.

## IBM Client Input Output/Sockets (CLIO/S)

(Program Number 5648-129.) Client Input Output/Sockets provides high-speed transparent data transfer and tape access between MVS/ESA systems and AIX systems or between AIX systems. It provides a set of user commands and application programming interfaces that run on either MVS or AIX.

## Network Tape Products

IBM provides two complementary network tape products for AIX:

- **IBM Network Tape Access and Control System for AIX (NetTAPE)**(program number 5765-637)
- **IBM NetTAPE Tape Library Connection (NetTAPE TLC)** (program number 5765-643)

NetTAPE, along with NetTAPE TLC, improves and simplifies tape operations management and tape device access in RS/6000 SP systems. These products are intended for:

- Customers who want transparent tape access across their remote AIX systems, especially accounts with 3494/3495 Tape Library Data Server
- Customers who need to transfer data to and from tape servers and tape clients

Flexibility benefits include:

- Access to remote tape devices on workstations
- Centralized tape device operation
- Support of multiple tape formats, allowing mainframe-created tapes to be processed on workstations and vice versa
- Application programming interfaces (APIs) available for C and FORTRAN
- Communication with MVS tape systems when CLIO/S is installed

## IBM Parallel Environment for AIX

(Program Number 5765-543.) The IBM Parallel Environment for AIX program product provides support for parallel application development and execution for the RS/6000 SP or on a single RS/6000 processor or a TCP/IP-networked cluster of IBM RS/6000 processors. The Parallel Environment product contains tools to support the development and analysis of parallel applications written in FORTRAN, C, or C++, and also provides a user-friendly runtime environment for their execution. Parallel Environment Ver. 2.3 has support for the Message Passing Library (MPL) subroutines, the Message Passing Interface (MPI) standard, and the Low Level Applications Programming Interface (LAPI).

High nodes operating at PSSP 2.3 or later and AIX 4.2.1 or later, support the Parallel Environment.

## Parallel Libraries

Multiple parallel libraries are available to 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 the SP. Two IBM libraries are available for use on the RS/6000 SP: Parallel Optimization Subroutine Library (OSLp) and Parallel Engineering and Scientific Subroutine Library (Parallel ESSL).

### IBM Parallel Optimization Subroutine Library

(Program Number 5765-392.) The Parallel Optimization Library (OSLp) is a collection of high-performance subroutines for use by mathematical application programs that solve optimization problems. Parallel OSL helps you find the optimal solutions to several types of problems using linear programming, mixed-integer programming, and quadratic programming. OSLp subroutines can be called from FORTRAN, PL/1, APL2, and C programs at different levels.

### IBM Parallel Engineering and Scientific Subroutine Library (Parallel ESSL)

(Program Number 5765-C41.) Parallel ESSL accelerates applications by substituting comparable math subroutines and in-line code with high performance, highly-tuned subroutines. Both new and current numerically intensive applications can call Parallel ESSL subroutines. The design of Parallel ESSL centers on exploiting SP operational characteristics and the architecture of the SP.

## IBM Parallel I/O File System for AIX (PIOFS)

(Program Number 5765-297.) Parallel I/O File System is a high performance file system for the SP. PIOFS exploits the architecture of the SP on AIX 4.1 and AIX 4.2 but is not supported on AIX 4.3. It scales in file input/output performance just as the RS/6000 SP scales in computing performance by striping files across multiple server nodes.

## IBM High Availability Cluster Multi-Processing

(Program Number 5765-A86 ) 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. Currently there are two variations of the product, HACMP and HACMP/ES. HACMP/ES is newer and has enhanced scalability capability in addition to the function provided by HACMP.

Typically, HACMP is run only on the control workstation if HACWS is being used. HACMP can also be run on the SP nodes. HACMP/ES does not run on the CWS, it only runs on the SP nodes.

## IBM Recoverable Virtual Shared Disk

(Program Number 5765-646 ) The IBM Recoverable Virtual Shared Disk program product lets you configure nodes as primary and secondary Virtual Shared Disk server nodes. (The Virtual Shared Disk support is a base component of PSSP.) Recoverable Virtual Shared Disk provides transparent failover to a secondary server node if the primary server node for a set of virtual shared disks fails. Communication adapter recovery is also provided, allowing SP Switch adapter or

Ethernet adapter failure to be treated the same as node failure; that is, control of connected twin-tailed volumes is passed to the backup server.

## IBM General Parallel File System for AIX

(Program Number 5765-B95 ) General Parallel File System (GPFS) for AIX provides concurrent shared access to files spanning multiple disk drives located on multiple nodes. This LPP provides file system service to parallel and serial applications on the SP. Your SP system must be utilizing IBM Recoverable Virtual Shared Disk with the Virtual Shared Disk option of PSSP installed.

Using GPFS, your SP system performance improves by:

- Allowing multiple processes simultaneous access to the same file through standard AIX file calls.
- Increasing aggregate bandwidth of file system and balancing disk loading.
- Allowing concurrent read and write actions, important in parallel processing.
- Guaranteeing data consistency by a sophisticated token management system.
- Simplifying administration through simple, multiple node file system commands that function across the entire SP system.

## IBM Performance Toolbox Parallel Extensions for AIX

IBM Performance Toolbox Parallel Extensions for AIX (PTPE) collects and provides performance data for SP hardware and software. It is a priced feature of PSSP designed as an enhancement to Performance Toolbox for AIX, the preferred performance monitor for AIX systems. PTPE gives Performance Toolbox access to SP hardware and software statistics, and makes Performance Toolbox easier to use when monitoring a large systems. All this is done while retaining the same familiar interfaces to the performance data that you have come to expect from Performance Toolbox for AIX.

## Selecting IBM Program Products

Our sample customer, `Corporation ABC`, selected the products checked off in Table 2 on page 22. You can check off the products that you want in Worksheet 2, "IBM Program Products To Order" in Table 42 on page 228.

*Table 2. IBM Program Products Ordered by ABC Corporation*

| IBM Program Products - Worksheet 2 | | | |
|---|---|---|---|
| **Order** | **Program Product** | **Program Number** | **Level** |
| | IBM C for AIX | 5765-423 | 3.1 |
| √ | IBM C Set ++ for AIX | 5765-421 | 3.1 |
| √ | IBM Parallel System Support Programs for AIX (PSSP) | 5765-529 | 2.4 |
| | IBM LoadLeveler | 5765-145 | 1.3 |
| √ | IBM Client Input Output/Sockets (CLIO/S) | 5648-129 | 2.2 |
| √ | IBM Parallel Environment for AIX | 5765-543 | 2.3 |
| | | 5765-543 | 2.2 |
| | IBM Parallel Optimization Subroutine Library | 5765-392 | 1.3 |
| | IBM Parallel Engineering and Scientific Subroutine Library (Parallel ESSL) for AIX | 5765-C41 | 2.1 |
| | | 5765-B50, 5765-422 | 1.2 |
| | | 5765-422 | 1.1 |
| | IBM Parallel I/O File System for AIX | 5765-297 | 1.2 |
| | IBM High Availability Cluster Multi-Processing / Enhanced Scalability for AIX | 5765-A86 | 4.2 |
| | IBM High Availability Cluster Multi-Processing for AIX | | |
| √ | IBM Recoverable Virtual Shared Disk for AIX | 5765-646 | 2.1 |
| | | 5765-444 | 1.2 |
| | IBM Network Tape Access and Control System (NetTAPE) for AIX | 5765-637 | 1.2 |
| | IBM NetTAPE Tape Library Connection for AIX | 5765-643 | 1.2 |
| | IBM General Parallel File System for AIX | 5765-B95 | 1.1 |
| | IBM Interactive Network Dispatcher for AIX | 5765-B67 | 1.2 |

## Question 4: What Levels of AIX Do You Need?

New RS/6000 SPs come with AIX 4.2.1 (or later) or AIX 4.3 and PSSP 2.4 installation tapes. You might need AIX 4.1 to support specific hardware.

## Considering New Features Released

You should plan to order the system level which supports the functions you need. PSSP 2.4 is supported on both AIX 4.2.1 (or later) and on AIX 4.3. PSSP 2.4 is compiled on AIX 4.2.1 and is supported by binary compatibility on AIX 4.3. PSSP does not yet take advantage of the new 64–bit environment of AIX 4.3 but you might want it so you can exploit it. Before deciding on software levels, consider the new features available in the new releases. See “What's New in AIX and PSSP?” on page 7 for partial summaries of new features.

## Considering Migration and Coexistence

Support is provided for migrating to PSSP 2.4 running on AIX 4.2.1 (or later) or on AIX 4.3. PSSP 2.4 is compiled on AIX 4.2.1 and is supported by binary compatibility on AIX 4.3. Migration and coexistence support is provided.

Table 3 lists the direct migration paths to PSSP 2.4 that are available.

*Table 3. Direct Migration Paths*

| From | To |
|---|---|
| PSSP 2.1 and AIX 4.1.5 | PSSP 2.4 and AIX 4.2.1, 4.3 |
| PSSP 2.2 and AIX 4.1.5, 4.2.1 | PSSP 2.4 and AIX 4.2.1, 4.3 |
| PSSP 2.3 and AIX 4.2.1 | PSSP 2.4 and AIX 4.2.1, 4.3 |
| PSSP 2.3 and AIX 4.3 | PSSP 2.4 and AIX 4.3 |

Coexistence is supported in the same system partition or a single default system partition (the entire SP system) for nodes running any combination of:

- PSSP 2.1 and AIX 4.1.5
- PSSP 2.2 and AIX 4.1.5 or 4.2.1
- PSSP 2.3 and AIX 4.2.1 or 4.3
- PSSP 2.4 and AIX 4.2.1 or 4.3

For information on planning for migration and coexistence, see Chapter 10, “Planning for Migration” on page 157.

## Recording Your Decision for Question 4

Now you should know which level of AIX you need. If you need more than one level, you might require system partitions which are discussed in more detail in Chapter 5, “Planning SP System Partitions” on page 99.

ABC's worksheet appears in Table 4.

*Table 4. Operating System Level Selected By ABC Corporation*

| Check | AIX | PSSP |
|---|---|---|
| | AIX 4.3.0 | PSSP 2.4 |
| √ | AIX 4.2.1 | PSSP 2.4 |
| | AIX 4.2.1 | PSSP 2.2 |
| | AIX 4.1.5 | PSSP 2.2 |

## 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 *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*).

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 this network to the network of the SP and its control workstation?
- What type of physical ethernet LAN do you require? The default is BNC thin but thick is available.
- Do you have a TCP/IP address range? What is the address range and how is it subnetted? What subnet masks are employed? Will the address range be sufficient to cover the addresses of the SP and RS/6000 control workstations? Remember here to think about the future—if you are starting with a 10-node system but plan to grow to 100 nodes in the future—be sure to define an address range that is big 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 RS/6000 workstations and the SP into the network

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

- The RS/6000 SP with a switch has a minimum of two networks, the SP Ethernet network that connects each node to the control workstation and the switch network. The two networks must each be assigned unique TCP/IP network addresses. If you have a switch, you must also plan for the switch network.
- In addition to the SP Ethernet network and the switch network, often additional communications adapters such as ATM or FDDI adapters are installed in the SP nodes. If this is the case, then separate TCP/IP network addresses need to be assigned. Have these networks been considered?
- What domain name will be assigned to the SP?
- What IP networks, addresses, subnet masks and default gateways will be assigned to the SP networks?
- Will machines be configured as primary and secondary name servers?
- What MVS considerations are there? If you currently have an MVS system, and if you plan to move large amounts of data (many gigabytes) between the MVS system and the SP, you might need CLIO/S. CLIO/S provides high-speed, low-overhead transfers over fast channel-to-channel connections. Planning for CLIO/S requires participation by both MVS and SP system planners. For more information on CLIO, refer to “Planning for IBM Client Input Output/Sockets (CLIO/S)” on page 143.

You might not have laid out your system requirements quite far enough yet to be able to answer all these questions fully. 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 the layout plan is complete. Also, read “Planning Your Network Configuration” on page 74 which will help you understand the SP network capabilities.

The worksheets for this question are with the worksheet for Question 8.

## Network Connections Using an Extension Node

To provide higher bandwidth communications, an SP Switch Router can be connected to the SP Switch via the SP Switch Router Adapter. The SP Switch Router Adapter provides a high performance, 100 MB, full duplex interface between the SP Switch and the SP Switch Router.

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:

- Ethernet
- 10/100BaseT
- FDDI
- ATM OC3c
- SONET OC3c
- ATM OC12c
- 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 55. 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, you should be aware of the following considerations that relate to internal and external disk storage.

Study these considerations and record your answers. Later you'll fill them in on the worksheet, “Hardware Configuration By Node” in Table 45 on page 231.

## Disk Space for Users' Home Directories

You need to decide whether you will serve your users' 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 products consumes disk storage on each node. Use the tables in “Determining Space Requirements” on page 71 to calculate the disk storage needed for AIX and PSSP and the related products.

Think about the program products and applications you plan to install. How much space do you need for /**usr**, /, and other file systems? For the /**spdata** file system, note that you will need extra space on the control workstation and boot/install servers if more than one level of AIX or PSSP is maintained on the system.

Will your applications be installed in rootvg with the base AIX programs or will they be installed elsewhere? Decisions like these might help you decide whether to add additional internal 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 twin-tailed disks? What is the data protection strategy? Do you require disk mirroring or RAID 1 or RAID 5?

For each database you need to consider:

- How much temporary space is needed?
- How much space is needed for logging?
- How much space is needed for the database definition and data dictionary?
- How much space is required for rollback?

If you plan to use twin-tailed 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 thin nodes have fewer adapter slots.

## Disk Requirements for the Virtual Shared Disk Component of PSSP

IBM Virtual Shared Disk is a subsystem that lets you assign volume groups located on any physical disks in any node within a partition. Once the volume group has been assigned to a Virtual Shared Disk, application programs running in any node within that partition can access the Virtual Shared Disk as if it were a disk located in the node running the application.

If an application exploits the use of a Virtual Shared Disk, you should **not** place anything on that Virtual Shared Disk except volume groups. Other file systems located on Virtual Shared Disk portions of a physical disk will cause problems. Similarly, you should not create AIX Journaled File Systems (JFSs) on volume groups that contain Virtual Shared Disks.

An optional LPP, the IBM Recoverable Virtual Shared Disk, enhances Virtual Shared Disk function by recovering information that would otherwise be lost if a

Virtual Shared Disk node or communication adapter were to fail. However, Recoverable Virtual Shared Disk will only support recovering information from Virtual Shared Disks. It will not recover data from non-Virtual Shared Disk portions of the physical disk. HACMP/ES is another mechanism which you can use to failover file systems.

All Virtual Shared Disks should be defined on external disk storage drives. Data residing on an internal disk that is not twin-tailed to another disk will be lost if the node containing that internal disk fails. That data loss will occur whether or not the Recoverable Virtual Shared Disk is in use.

## File System Requirements

You should plan ahead of time for expected growth of all your file systems. Also, you should monitor your file system growth periodically and readjust your plans when necessary.

## 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 64 for recommended boot/install configurations for various system sizes.

## Multiple Boot Requirements

Definable multiple boot images and 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 two disks, each holding a complete copy of the operating system. If you want alternate boot system images, make sure you plan for enough disk space.

## External Disk Storage

If external disk storage is part of your system solution, you need to decide which type of the external disks offered for the SP best satisfies your needs.

One of the major decisions here is whether you want RAID (Redundant Arrays of Independent Disks) technology or not. RAID subsystems offer improved reliability over JBOD (Just a Bunch of Disks) subsystems. However, JBOD subsystems provide low cost disk storage media. Refer to the following table for more information on disk storage choices.

| Disk Storage | Description. |
|---|---|
| 7134 | The 7134 subsystem, designed for a mix of I/O and throughput, provides low cost disk storage media and is employed where performance is not the critical characteristic of the parallel environment. The 7134 with up to 16 disks can support from 4.5GB to 72GB of data. |
| 7133 | If you require better performance, the 7133 Serial Storage Architecture (SSA) Disk might be the subsystem of choice for you. It provides more megabytes per second throughput at a higher cost than the 7134, but does not provide for high reliability. This is a good storage subsystem for transaction processing with small, random reads and writes, and for scientific environments. The 7133 can support from 4.5GB to 72GB of data per each SSA adapter. You can also connect up to 16 SSA adapters to a ring of disk drives. |
| 7137 | The 7137 subsystem supports both RAID 0 and RAID 5 modes. It can hold from 4 to 33 gigabytes of data (29GB maximum in RAID 5 mode). The 7137 is the low end model of RAID support. If performance is not critical but reliability and low cost are important, this is a good choice. |
| 7135 | If performance and high reliability are key requirements, the 7135 is the subsystem of choice. It supports RAID modes R0, R1, R3, and R5. It has a 20MB/second SCSI-DE interface.<br>**R0** 4.5 to 135GB of data<br>**R1** 4.5 to 67.5GB of data<br>**R3, R5** 4.5 to 108GB of data<br>The 7135's high availability features include:<br>• No single point of failure<br>• Redundant power supplies and fans<br>• A dynamically switched “spare” controller<br>• Hot-pluggable disk drives and controllers |

## External Disk Storage Worksheet

The following table shows how the ABC Corporation specified their external disk storage needs. Record your external disk storage needs on Worksheet 3, Table 43 on page 229. You'll fold that information into Worksheet 4, the “SP Planning Worksheet” in Table 44 on page 229.

*Table 5. ABC Corporations's External Disk Storage Needs*

| Worksheet 3 | | |
|---|---|---|
| **Quantity** | **External Disk Subsystem** | **Disk Space (MB)** |
| | 7133 (4.5 to 72GB) | |
| | 7134 (4.5 to 72GB) | |
| 1 | 7135 (4.5 to 135GB) | 90 MB |
| | 7137 (4 to 33GB) | |
| **Note:** Complete the quantity and disk space columns for the external disk subsystems you require. | | |

## 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 their production system up and running. Two of the functions in PSSP that assist in reliability and availability are the High Availability Control Workstation and system partitions.

SP Switch systems provide enhanced availability.

## High Availability Control Workstation

One function providing enhanced reliability is the High Availability Control Workstation (HACWS) which basically provides a second control workstation that, in effect, 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, the SP high availability component detects the loss and shifts that component's workload 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 6.

*Table 6. Requirements for the High Availability Control Workstation*

| | |
|---|---|
| Software | An additional AIX license |
| | 1 HACWS software feature 3936 |
| | 2 licenses for HACMP; 5050 High Availability feature (1) |
| Hardware | A second control workstation |
| | HACWS Connectivity Feature |
| (1) You do not need feature 5051. | |

Planning and using the HACWS will be simpler if you configure your backup control workstation identical to the primary control workstation. Some components must be identical, others can be similar. 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, refer to Chapter 4, “Planning for a High Availability Control Workstation” on page 87.

## System Partitions

Partitioning your system can also aid in system availability. This support lets you logically divide the SP into non-overlapping groups of nodes called system partitions. You can then use a system partition to test new levels of AIX, PSSP, LPPs, application programs, or other software on a system that is currently running a production workload without disrupting that workload. The partitioning solution assumes that there are nodes available for the test system partition. A minimum system partition must consist of at least two drawers (or four slots).

An alternate method uses coexistence support (provided as of PSSP 2.2) which allows you to migrate one node of your system at a time. With coexistence, you are permitted to have multiple levels of PSSP operating within a single system partition.

Another 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 these 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.

Initially, the system is a single partition. The number of system partitions you can define is dependent upon the size of your SP. See Chapter 5, “Planning SP System Partitions” on page 99 for complete information about system partitions. If you decide you want system partitions, study that chapter in more detail before completing your system plan. For now, you need to decide only if it is something you want or need and how many system partitions you think you'll need (you can come back and modify these answers if you learn new information that affects your answer).

*Table 7. Function Checklist*

| Check | Function |
|---|---|
| | Do you want the redundancy of a High Availability Control Workstation? |
| | How many system partitions do you want? |
| √ | • Will you run PSSP 2.4? |
| | • Will you run PSSP 2.3? |
| √ | • Will you run PSSP 2.2? |
| | • Will you run PSSP 2.1? |
| **Note:** Review coexistence limitations in Chapter 10, “Planning for Migration” on page 157 to help you decide if you must partition your system. | |

## Question 8: 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 the SP is “scalable” which means that you can add more nodes later. Your answers to the prior questions should have helped you determine the type of work for which you will be using the SP. For example, if you previously determined that you want to divide your system into partitions, this can affect the number of nodes you require. Since the SP is scalable, you can select fewer nodes now and add more later or select more now and scale down later.

Some helpful hardware reference information is included here to help you select nodes. For more hardware information see *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*.

Along with deciding how many nodes you want, you must also decide what physical types of nodes you need. There are three physical types of nodes; thin, wide, and high.

- Thin Node

  Thin nodes are designed for users who require the highest number of processors per frame at the higher levels of computational performance. There

are three types of thin nodes; 120 MHz and 160 MHz uniprocessor, and 332 MHz SMP nodes.

The 120 MHz nodes have 64 kilobytes of data cache and a 64 bit memory bus.

The 160 MHz nodes have 128KB data cache and the bandwidth has been doubled with the addition of a second path for memory and the integer and floating point units (128 bit memory bus). The 160 MHz node performs at increased levels for most commercial applications and has better floating point performance due to the larger bus bandwidths and cache. Both 120 MHz and 160 MHz node types support four micro channel adapter (MCA) slots and up to two SCSI disks packaged internally.

The 332 MHz SMP Thin node is an IBM RS/6000 PowerPC containing the following components:

- A node supervisor with two or four 604+LX symmetric multiprocessors.
- A mezzanine (MX) system bus with one slot for an SP Switch MX Adapter.
- A PCI I/O bus with two 32–bit PCI slots.

**Note:** Adding a 332 MHz SMP thin node to an existing frame requires a power upgrade.

- Wide Node

  Wide nodes greatly expand the I/O and network server functions of the SP. Wide nodes occupy two slots in a frame and have more MCA or PCI slots to allow greater attachment options. (Deciding between them is basically the same as choosing between a desktop or a deskside model in the RS/6000 line.)

  The wide nodes with MCA have 256 kilobytes of data cache, eight MCA slots and up to four SCSI disks packaged internally.

  The 332 MHz SMP wide node has the same components as the 332 MHz SMP thin node but with ten PCI slots; three 64–bit slots and seven 32–bit slots. Each assembly has its own power supply and cooling fans.

  **Note:** Adding a 332 MHz SMP wide node to an existing frame requires a power upgrade.

- High Node

  The high node is an SMP system that can have 2, 4, 6, or 8 POWERPC 604e processors running at 200 MHz. The node includes 2 micro channel buses for I/O attachment including attachment to the SP Switch using the SP Switch adapters. The high node also includes a node supervisor through which hardware control and node conditioning are provided.

  The high node occupies two full drawers of a frame. This means that a maximum of four high nodes can fit in a tall frame and only two in a short frame. The high node is supported in both frames with or without the switch. Power and Power2 nodes can exist in the same frame and in the same partition as the high nodes. However, the different physical sizes results in changes to the set of configurations which are supported.

- Extension Node

  Configuring a system using extension nodes requires special planning regarding standard nodes.

  For all three standard node types, if you have a fully populated switch and you are ordering an SP Switch Router, you will have to reduce your node count by

one. This action is needed to open up a switch port for the SP Switch Router Adapter.

If you have decided to include extension nodes in your system, you must ensure that your system provides the switch ports needed for each logical node. For instance, each SP Switch Router Adapter in an SP Switch Router must be connected to a valid switch port on the SP switch. In other words, each dependent node logically occupies a frame slot and physically occupies the corresponding switch port. A standard node must not be assigned to the same slot, although it can overlap the slot. Read from “Planning Considerations for the SP Switch Router” on page 77 in Chapter 3, Defining the Configuration that Fits Your Needs for a discussion of valid extension node slots.

There are two frame models for the RS/6000 SP system which you can populate with optional nodes and switches to create the system configuration of your choice. Your layout can range from a single-frame entry system to a highly-parallel, large-scalable system. The frame models are listed in Table 8.

*Table 8. The Basic SP Frame and Switch Topology*

| Frame Model | Description |
|---|---|
| 500 | Short frame, power supply, additional equipment:<br>• up to eight nodes, type is optional, one drawer required<br>• SP Switch or SP Switch-8 optional |
| 550 | Tall frame, power supply, additional equipment:<br>• up to sixteen nodes, type is optional, one drawer required<br>• SP Switch or SP Switch-8 optional<br>• scalable up to 128 nodes with SP Switch |

Table 9 summarizes nodes that are currently orderable. See *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for more node information.

*Table 9. SP Node Summary*

| Speed in MHz | Node Type (# of nodes / # of drawers) | Minimum-Maximum Memory Per Node | Minimum-Maximum Disk Space Per Node |
|---|---|---|---|
| 120 | Thin (2 / 1) | 64 MB to 1 GB | 2 GB to 18 GB |
| 160 | Thin (2 / 1) | 64 MB to 1GB | 4.5 GB to 18.2 GB |
| 135 | Wide (1 / 1) | 64 MB to 2 GB | 2 GB to 36 GB |
| 200 | High (1 / 2) | 256 MB to 4 GB | 4.5 GB to 18 GB |
| 332 | Thin (2 / 1) | 256 MB to 3 GB | 4.5 GB to 18.2 GB |
| 332 | Wide (1 / 1) | 256 MB to 3 GB | 4.5 GB to 36.4 GB |

## System-Wide Worksheets

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 so already, make copies of the worksheets in Appendix C, "SP System Planning Worksheets" on page 227. The worksheets in this chapter have been filled out for a hypothetical customer, the `ABC Corporation`. `ABC`'s system-wide selections are in the "SP Planning Worksheet" in Table 10.

**Note:** Be sure to check *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for requirements of other hardware based on your choice of SP model. For instance, you cannot have a combination of High Performance Switch and SP Switch adapters.

*Table 10. Overall System Information*

| **SP Planning - Worksheet 4** | | |
|---|---|---|
| **Customer Name**_ABC Corporation___ | | **Date** _August 20, 1998____ |
| **Customer Number** _999999___ | | |
| **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** | **Nodes** |
| 500 (short): | 120 MHz Thin: | 200 MHz High: |
| 550 (tall): 1 | 160 MHz Thin: | 135 MHz Wide: |
| | 332 MHz Thin: 4 | 332 MHz Wide: 2 |
| **SP Switches** | | |
| 8-port: | 16-port: 1 | |
| SP Switch Router: 1 | SP Switch Router Adapter: 1 | |
| **External Storage Units:** | *Type* | *Quantity* |
| | 7135 | 2 |
| | | |
| **Network Media Cards:** | *Type* | *Quantity* |
| | | |
| | | |
| *Fill in the remainder of this chart after you place your order.* | | |
| | *System Number* | *Purchase Order Number* |
| RS/6000 SP: | | |
| Control Workstation: | | |
| Peripherals: | | |
| | | |
| | | |

Fill in Worksheet 4, "SP Planning" in Table 44 on page 229 with the heading information, the SP model, the number of frames and switches, and the number of each node type you need. If you selected an external disk in "Question 6: What are

Your Disk Storage Requirements?" on page 25, copy the information from that table to Worksheet 4 in Table 44 on page 229. Once you place your order you can fill in the order numbers for handy reference.

## Completing the SP Node Layout Worksheets

To complete the Node Layout Worksheets, first draw a diagram of your SP system. Then add network information to that diagram. After that, write your network information into the worksheets.

`ABC Corporation` drew the network shown in Figure 2 on page 35 and Figure 3 on page 36. You'll fill in as many copies of Worksheets 5 and 6, Figure 55 on page 230 and Figure 57 on page 231, as you need.

Complete the SP Node Layout Worksheets as follows:

1. For each frame, fill in the frame number and the switch number on the line marked **Frame Number** or **Switch Number** at the bottom of the diagram.
2. Indicate whether each node is a wide, thin, or high node using a unique identifier for each. For example, you could represent wide nodes with a *w*, thin nodes with a *t*, and high nodes with an *h*. Slot numbers have been indicated on each frame diagram. Wide nodes occupy two slots and use the odd-numbered slot number. Cross out the even slot numbers in all wide nodes. High nodes occupy four slots. Figure 2 on page 35 shows a single frame with numbered slots (those parentheses are switch port numbers) for the `ABC Corporation`.

   **Note:** If you are attaching extension nodes, create an indicator for each type. Using these indicators, mark the node slots that each extension node will logically occupy.

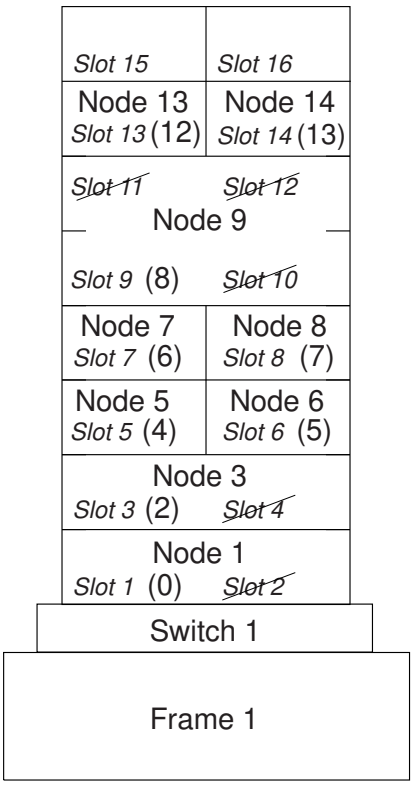



*Figure 2. Node Layout Example For The ABC Corporation*

3. Refer to "Understanding Node Numbering and Switch Node Numbering" on page 79 to learn more about node and switch numbering.

   At this point, your layout should look something like Figure 3 on page 36.

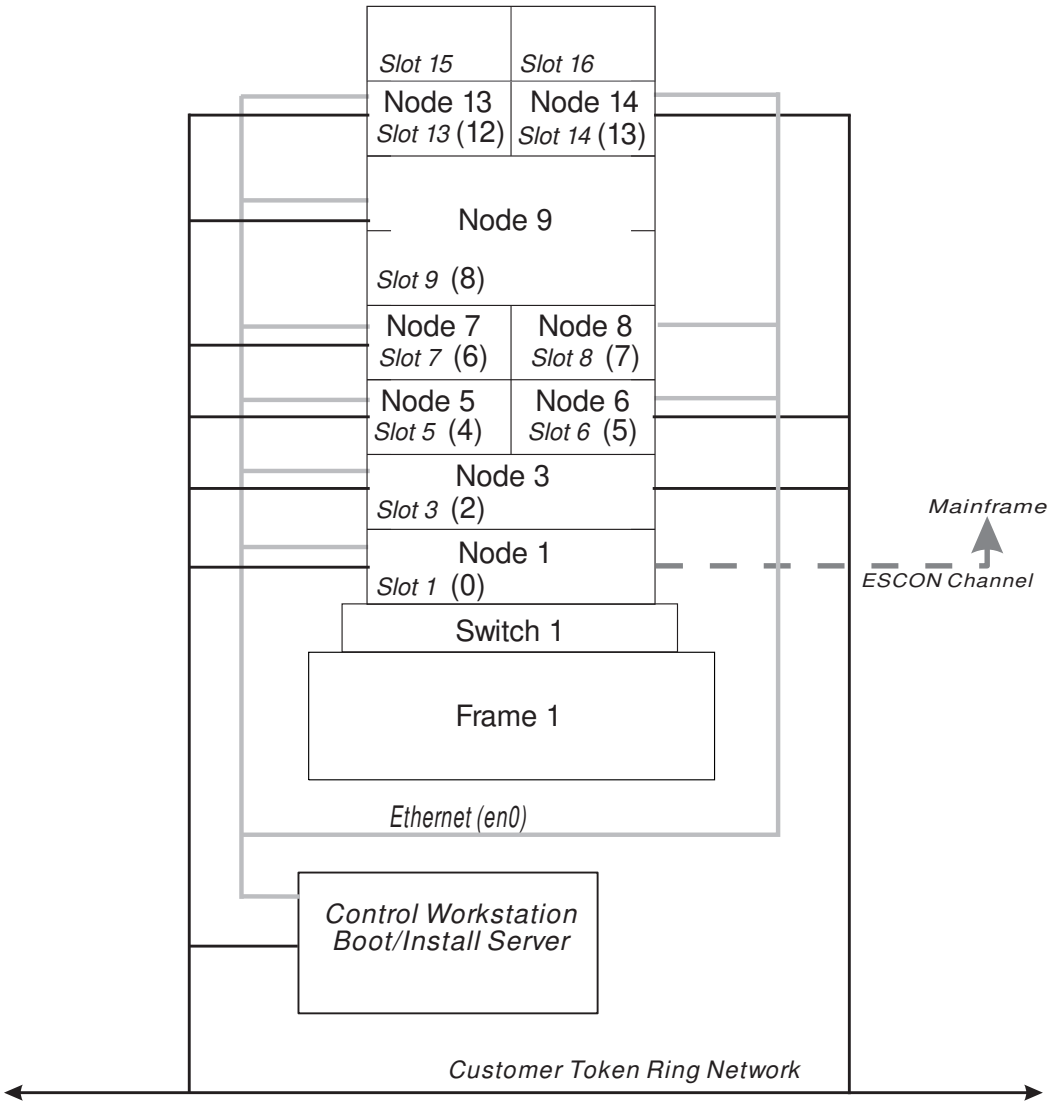



*Figure 3. The ABC Corporation Node Layout Example With Communications Information*

4. Sketch your SP Ethernet connections to each node and to the Control Workstation. Indicate specific adapter connections (for example, en0 and en1 connections). Refer to "System Topology Considerations" on page 64 for Ethernet tuning considerations.
5. Sketch your additional network connections.
6. Sketch connections to any routers, gateways, or networks.
7. Indicate network addresses, netmasks and hostnames for each subnet and node address on each node interface.

## Processor Memory

At the same time you decide what types of nodes and how many you want, you also need to decide how much processor memory each node will have and how much internal disk storage. Each of these values will affect the performance of your system, so choose carefully. `ABC Corporation` made the choices in Table 11 on page 37.

Once you decide on this information, fill in Worksheet 7, “Hardware Configuration by Node” in Table 45 on page 231. You might need multiple copies of this worksheet depending on the number of nodes you plan to install.

*Table 11. ABC Corporations's Choices For Hardware Configuration By Node*

| **SP Hardware Configuration by Node - Worksheet 7** | | | | | | |
|---|---|---|---|---|---|---|
| **Frame Number** __1___ | | | **Switch Number** __1___ | | | |
| **Slot Number** | **Node Number** | **Node Type** | **Processor Memory** | **Internal Disk** | **L2 Cache** | **Adapters** |
| Slot 1 | 1 | wide | 256MB | 18GB | | token ring |
| Slot 2 | -- | | | | | |
| Slot 3 | 3 | wide | 256MB | 18GB | | token ring |
| Slot 4 | -- | | | | | |
| Slot 5 | 5 | thin | 256MB | 9GB | 1MB | token ring |
| Slot 6 | 6 | thin | 256MB | 9GB | 1MB | token ring |
| Slot 7 | 7 | thin | 256MB | 9GB | 1MB | token ring |
| Slot 8 | 8 | thin | 256MB | 9GB | 1MB | |
| Slot 9 | 9 | high node | 2GB | 6GB | | token ring<br>FDDI |
| Slot 10 | — | | | | | |
| Slot 11 | — | | | | | |
| Slot 12 | — | | | | | |
| Slot 13 | 13 | thin | 256MB | 9GB | 1MB | token ring |
| Slot 14 | 14 | thin | 256MB | 9GB | 1MB | token ring |
| Slot 15 | — | | | | | |
| Slot 16 | — | | | | | |

# Networking Information

Each adapter in each node, workstation, and router has an IP address. Each of these addresses should have a separate name associated with it.

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, then they need to plan for the addition of all these names to their servers before the system arrives.

## Host Names

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

While completing these tables, keep in mind that the host name in the table is referring to the name given to that adapter. You need to select which of these

adapter host names should be the one given to the processor. An application might require that the processor host name be the name associated with the adapter over which its traffic will flow. Put a check (√) in the column of the adapter that will be the host name.

ABC Corporation completed the "Node Network Configuration Worksheets" starting with Table 12. Review your network topology and fill in worksheets starting with Table 46 on page 232. Be sure to make extra copies before you complete them. You need a copy for each frame you configured. If you have additional network adapters planned for some or all of your nodes, you need to plan their network information also.

*Table 12. ABC Corporation's SP Ethernet*

**SP Ethernet Node Network Configuration - Worksheet 8A**

**Company**
**Name**__ABC Corporation____
**Date**__January 25, 1998
**Frame Number**_1__
**Token Ring Speed**__16

| Slot | **SP Ethernet** *(en0 adapters)* **Netmask**__255.255.255.192__ | | Default Route |
|---|---|---|---|
| | **Hostname (note 1)** | **IP Address** | |
| 1 | spnode01 | 129.40.60.1 | 129.40.60.125 |
| 2 | -- | | |
| 3 | spnode03 | 129.40.60.3 | 129.40.60.125 |
| 4 | -- | | |
| 5 | spnode05 | 129.40.60.5 | 129.40.60.125 |
| 6 | spnode06 | 129.40.60.6 | 129.40.60.125 |
| 7 | spnode07 | 129.40.60.7 | 129.40.60.125 |
| 8 | spnode08 | 129.40.60.8 | 129.40.60.125 |
| 9 | spnode09 | 129.40.60.9 | 129.40.60.125 |
| 10 | -- | | |
| 11 | -- | | |
| 12 | -- | | |
| 13 | spnode13 | 129.40.60.13 | 129.40.60.125 |
| 14 | spnode14 | 129.40.60.14 | 129.40.60.125 |
| 15 | -- | | |
| 16 | -- | | |

**Notes:**

1. AIX is case sensitive. Use lower case for the hostname and addresses.
2. Wide nodes occupy two frame slots and use the *odd-numbered* slot number.
3. High nodes occupy four frame slots (2 drawers) and use the lowest odd-numbered slot number.

*Table 13. ABC Corporation's Additional Adapters*

**SP Additional Adapters Node Network Configuration - Worksheet 8B**

**Company**
**Name**__ABC Corporation____
**Date**__January 25, 1998
**Frame Number**_1__
**Token Ring Speed**__16

| Slot | Additional Adapters Netmask__255.255.255.192__ | | | Default Route |
|---|---|---|---|---|
| | Adapter Name | Hostname (note 1) | 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. Use lower case for the hostname and addresses.
2. Wide nodes occupy two frame slots and use the *odd-numbered* slot number.
3. High nodes occupy four frame slots (2 drawers) and use the lowest odd-numbered slot number.

## Switch Worksheet

If you are planning to install an SP with a switch, remember that the advantage of the switch is that it, too, has its own subnet. You need to plan for this network.

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 node number.

Make copies of Worksheet 9, “Switch Configuration ” in Table 48 on page 234 before you start. The hypothetical ABC Corporation filled out the following chart.

If you do not plan to have a switch, you can skip this worksheet.

*Table 14. ABC Corporation's Choices For The Switch Configuration Worksheet*

| **Switch Configuration - Worksheet 9** | | | |
|---|---|---|---|
| **Frame Number** 1 **Switch Number** 1 **Netmask** 255.255.255.192 | | | |
| Slot Number | Switch Node Number | Hostname | IP Address |
| Slot 1 | 0 | spsw01 | 129.40.62.1 |
| Slot 2 | -- | | |
| Slot 3 | 2 | spsw03 | 129.40.62.3 |
| Slot 4 | -- | | |
| Slot 5 | 4 | spsw05 | 129.40.62.5 |
| Slot 6 | 5 | spsw06 | 129.40.62.6 |
| Slot 7 | 6 | spsw07 | 129.40.62.7 |
| Slot 8 | 7 | spsw08 | 129.40.62.8 |
| Slot 9 | 8 | spsw08 | 129.40.62.9 |
| Slot 10 | -- | | |
| Slot 11 | -- | | |
| Slot 12 | -- | | |
| Slot 13 | 12 | spsw13 | 129.40.62.13 |
| Slot 14 | 13 | spsw14 | 129.40.62.14 |
| Slot 15 | -- | | |
| Slot 16 | -- | | |
| **Note:** Refer to Table 12 on page 38 to see how this table would be completed. | | | |

## Specifying Adapters

If you want to order other adapters for your nodes when you place your SP order, you can use Worksheet 10A or Worksheet 10B, depending on whether the adapter is for a 332 MHz node. The choices selected by ABC Corporation are noted on the following chart. Copy Table 49 on page 234 or Table 50 on page 235 and record your choices.

*Table 15. PCI Adapters Supported Only on 332MHz Nodes*

| PCI Adapters Supported on 332 MHz Nodes - Worksheet 10A | | | | | | |
|---|---|---|---|---|---|---|
| √ | **PCI Adapter Name** | **Feature Code** | **Maximum Per Wide Node** | **Maximum Per Thin Node** | **PCI Slots Required** | **AIX 4.2.1/4.3** |
| | FDDI SK-NET LP SAS | 2741 | 4 | 2 | 1 | yes |
| | FDDI SK-NET LP AS | 2742 | 4 | 2 | 1 | yes |
| | FDDI SK-NET UP SAS | 2743 | 4 | 2 | 1 | yes |
| | WAN RS232 128-port | 2944 | 7 | 2 | 1 | yes |
| √ | Ethernet 10/100 MB | 2968 | 4 | 2 | 1 | yes |
| | Token Ring Auto Lanstream | 2979 | 8 | 2 | 1 | yes |
| | Ethernet 10 MB BNC | 2985 | 8 | 2 | 1 | yes |
| | Ethernet 10 MB AUI | 2987 | 8 | 2 | 1 | yes |
| | ATM 155 MMF | 2988 | 4 | 2 | 1 | yes |
| √ | SCSI-2 Single-Ended | 6208 | 9 | 2 | 1 | yes |
| | SCSI-2 Differential | 6209 | 9 | 2 | 1 | yes |
| | SSA RAID EL | 6215 | 6 | 2 | 1 | yes |
| | SSA Fast-Write Cache Option | 6222 | Mounts on F/C 6215 | | 0 | yes |

*Table 16 (Page 1 of 2). MCA Adapters Supported on all but 332MHz Nodes*

| MCA Adapters Supported on all but 332 MHz Nodes - Worksheet 10B | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| √ | **Adapter** | **Feature code** | **Quantity per wide node** [1] | **Quantity per thin node** [2] | **Quantity per high node** [3] | **Number of slots Required** | **AIX 4.1** | **AIX 4.2** |
| | Internal Ethernet | Standard | N/A | 1 | N/A | 0 | yes | yes |
| | FCS Dwtr | 1902 [7/8/11] | 0 - 2 | 0 -1 | N/A | 0 | yes | yes |
| | FCS 1GB | 1904 [8/11] | N/A | N/A | N/A | 1 | 4.1.4 | no |
| | NetW TA 256 | 2402 | 0 - 7 | 0 - 4 | 0 - 4 | 1 | no | yes |
| | NetW TA 2048 | 2403 | 0 - 7 | 0 - 4 | 0 - 4 | 1 | no | yes |
| | SCSI-2 Ext I/O | 2410 | 0 - 7 | 0 - 4 | N/A | 1 | 4.1.4 | yes |
| | SCSI Turbo | 2412 | 0 - 7 | 0 - 4 | 0 -14 | 1 | 4.1.3 | yes |
| | SCSI F/W DIF | 2415 | 0 - 7 | 0 - 4 | 1 -14 | 1 | 4.1.1 | yes |
| | SCSI F/W DIF | 2416 | 0 - 7 | 0 - 4 | 0 - 14 | 1 | 4.1.1 | yes |
| | SCSI EXT I/O | 2420 | 0 - 7 | 0 - 4 | N/A | 1 | 4.1.4 | yes |
| | 4 port Multi Comm | 2700 | 0 - 7 | 0 - 3 | 0 - 8 | 1 | 4.1.1 | yes |
| | FDDI D/R | 2723 [4] | 0 - 3 | 0 - 2 | 0 - 8 | 1 | 4.1.1 | yes |
| | FDDI S/R | 2724 | 0 - 6 | 0 - 2 | 0 - 8 | 1 | 4.1.1 | yes |
| | HIPPI[5/6] | 2735 | 0 - 1 | N/A | 0 - 2[6] | 5[5] | 4.1.4 | yes |
| | ESCON Chan Em. | 2754 | 0 - 2 | 0 - 1 | 0 - 4 | 2 | 4.1.4 | yes |
| | BMCA | 2755 | 0 - 2 | 0 - 2 | 0 - 2 | 1 | 4.1.4 | yes |

*Table 16 (Page 2 of 2). MCA Adapters Supported on all but 332MHz Nodes*

**MCA Adapters Supported on all but 332 MHz Nodes - Worksheet 10B**

| √ | Adapter | Feature code | Quantity per wide node [1] | Quantity per thin node [2] | Quantity per high node [3] | Number of slots Required | AIX 4.1 | AIX 4.2 |
|---|---|---|---|---|---|---|---|---|
| | ESCON CNTRL | 2756 | 0 - 2 | 0 - 1 | 0 - 4 | 2 | 4.1.4 | yes |
| | 8-port async 232 | 2930[9] | 0 - 7 | 0 - 4 | 0 - 14 | 1 | 4.1.1 | yes |
| | 8-port async 422 | 2940[9] | 0 - 7 | 0 - 4 | 0 - 14 | 1 | 4.1.1 | yes |
| | X.25 inter co-p | 2960 | 0 - 7 | 0 - 4 | 0 - 8 | 1 | 4.1.3 | yes |
| | Token Ring | 2970 | 0 - 7 | 0 - 4 | 0 - 12 | 1 | 4.1.1 | yes |
| | Token Ring | 2972 | 0 - 7 | 0 - 3 | 0 - 12 | 1 | 4.1.1 | yes |
| | Ethernet | 2980 | | 0 - 3 | 1 - 8 | 1 | 4.1.1 | yes |
| | ATM 100 | 2984 | 0 - 2 | 0 - 2 | 0 - 2 | 1 | 4.1.4 | yes |
| \| | ATM 155 | 2989 [8] | 0 - 4 | 0 - 2 | 0 - 4 | 1 | 4.1.4 | yes |
| | Ether TP | 2992 | 0 - 7[13] | 0 - 3 | N/A | 1 | 4.1.4 | yes |
| | Ether BNC | 2993 | 0 - 7[13] | 0 - 3 | N/A | 1 | 4.1.4 | yes |
| | Ether BNC | 2994 | x | x | — | 1 | 4.1 | — |
| | SPSw (TB3) | 4020 | 0 - 1 | 0 - 1 | 0 - 1 | 1 | 4.1.1 | yes |
| | Enet 10baseT | 4224 | 0 -8 | 0 - 4 | 0 - 15 | 0 | 4.1.1 | yes |
| | HI-P Subsys | 6212 | 0 - 4 | 0 - 2 | 0 - 8[10] | 1 | 4.1.1 | yes |
| | SSA | 6214 | 0 - 4 | 0 - 2 | 0 - 8[10] | 1 | 4.1.4 | yes |
| | SSA | 6216 [8] | 0 - 4 | 0 - 2 | 0 - 8[10] | 1 | 4.1.4 | yes |
| | SSA 4RD | 6217 | 0 - 4 | 0 - 2 | 0 - 8 | 1 | 4.1.5 | yes |
| | SSA RAID EL | 6219 | 0 - 4 | 0 - 2 | 0 - 8 | 1 | 4.1.5 | yes |
| | Fast Write Cache Option | 6222 | Mounts on FC 6219 | | | | | |
| | Digital Truck | 6305 | 0 - 6 | 0 - 3 | 0 - 2 | 1 | 4.1.1 | yes |
| | Prtmstr 1MB | 7006[12] | 0 - 7 | 0 - 4 | 0 - 8 | 1 | 4.1.1 | yes |
| | 128 prt Cntrl | 8128 | 0 - 7 | 0 - 7 | 0 - 7 | 1 | 4.1.1 | yes |

**Note:**

[1]: There are a total of 7 MCA slots available per wide node.
[2]: There are a total of 4 MCS slots available per thin node.
[3]:There are a total of 16 MCA slots per high node.
[4]: FDDI D/R adapters (F/C 2723) have a mandatory prerequisite of FDDI S/R adapters (F/C 2724)
[5]: The HIPPI feature uses 3 physical MCA slots and requires a total of 5 MCA slots to satisfy power and thermal requirements.
[6]: HIPPI cannot be populated across the 2 micro channel bus on high nodes.
[7]: FCS Daughter card F/C 1902 does not require a micro channel slot.
[8]: These adapters are not supported on any 62MHz node.
[9]: This adapter has a co-requisite of 2995 feature cable.
[10]:The SSA Adapters in a high node are limited to a total count of 8 in any combination
[11]:1902, 1904, 1906 FCS adapters are not supported in the P2SC nodes which are the 135MHz wide nodes (F/C 2007) , and the 120MHz thin nodes (F/C 2008), nor are they supported on the SMP high nodes (F/C 2006 ).
[12]:7006 portmaster card requires the selection of 7042, 7044, 7046, or 7048.
[13]:The maximum of 2992 and 2993 in any combination is 8.

**Note:** If you order the SP Switch, you must order F/C 4020 for every node.

## Question 9: Defining Your System Images

After determining the quantity and the type of nodes you need, you now decide what system image you want installed on which nodes. The system image is the collection of SP 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 biggest implication here is that if all the node images are the same, the installation and backup/restore functions are much easier. 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 are quicker, they require separate upgrades and system backups.

If you decided to have system partitions, you need to decide how many partitions you want and what nodes go with what partition. To fully understand partitioning, read Chapter 5, “Planning SP System Partitions” on page 99 before you make any decisions about system partitions.

## 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 as many copies of the worksheets as you require; fill out one for each image you plan.

IBM provides one or more minimal system images (SPIMG) with PSSP . 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 "*PSSP Memo to Users*" will give you the latest information on the minimal images 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 recommends that the SPIMG be placed on an internal drive.

To specify system images, the `ABC Corporation` filled out Worksheet 11, Table 17 on page 44. To specify PSSP components, they filled out Worksheet 12, Table 18 on page 45.

*Table 17. ABC Corporation's Specifying the System Images*

| **Specifying System Images – Worksheet 11** | |
|---|---|
| **System Image Name** | SPIMG1 |
| **AIX Level** | 4.2.1 |
| **Partition Number** | 1 |
| **Install on Node Numbers** | 1, 3, 5, 6, 7, 8, 9, 13, 14 |
| **Specify** *internal* **disks where you wish to install rootvg** disk 1 | |
| **Check here if you want only the SPIMG minimal image** ____ | |
| **IBM Licensed Products** | |
| | PSSP |
| | GPFS |
| | C Set ++ |
| | LoadLeveler |
| | CLIO/S |
| | Parallel Environment |
| | Virtual Shared Disk (option of PSSP) |
| | Recoverable Virtual Shared Disk |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| **Additional AIX Software** | |
| | bosext2.ate.obj- async terminal emulator |
| | bosext2.dosutil.obj - DOS utilities |
| | |
| **Other Applications** | |
| | NFS |
| | |
| | |
| | |
| | |
| | |
| | |
| | |

*Table 18. File Set List for PSSP 2.4*

| **PSSP 2.4 File Sets — Worksheet 12** | | |
|---|---|---|
| **System Image Name** __spimg1_____ | | |
| √ | **File Set** | **Description** |
| √ | ssp.st | Application programming interface for loading, unloading, and querying the job switch resource table. |
| √ | ssp.ha | Availability subsystems which include heartbeat, group services, and event management. |
| √ | ssp.perlpkg | Perl4 and Perl5 |
| √ | ssp.pman | Problem management |
| √ | ssp.clients | SP Authenticated Client Commands |
| √ | ssp.basic | SP System Support Package |
| √ | ssp.css | SP Communication Subsystem Package (only if switch installed) |
| √ | ssp.sysman | Optional System Management Programs |
| √ | ssp.sysctl | SP Sysctl Package |
| √ | ssp.authent | SP Authentication Server |
| √ | ssp.public | Public Code Compressed Tarfiles |
| √ | ssp.docs | PostScript, man pages, and HTML files for PSSP documentation |
| √ | ssp.gui | SP System Monitor Graphical User Interface and SP Perspectives |
| √ | ssp.jm | SP Resource Manager Package |
| √ | ssp.top | SP System Partition Support |
| √ | ssp.csd.vsd | PSSP Virtual Shared Disk Package |
| √ | ssp.csd.cmi | SP Centralized Management Interface Package for PSSP Virtual Shared Disk |
| √ | ssp.csd.hsd | PSSP Virtual Shared Disk Data Striping package |
| √ | ssp.csd.gui | Perspectives for PSSP Virtual Shared Disk |
| √ | ssp.csd.sysctl | PSSP Virtual Shared Disk Usability Improvement |
| √ | ssp.hacws | High Availability Control Workstation Support |
| √ | ssp.ptpegui | PTPE graphical user interface |
| √ | ssp.topsvcs | Topology services |
| √ | spimg | Contains a single file with the mksysb image of a minimal AIX 4.2 system. |
| √ | ssp.top.gui | Graphical user interface for System Partitioning Aid |
| √ | ssp.spmgr | Extension Node SNMP Manager Support |
| √ | ssp.ucode | SP supervisor microcode package |
| **Note:** For information on whether these file sets are installed on the control workstation and the node, see chapter 2 of *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*. | | |

## Question 10: What Do You Need for Your Control Workstation?

When planning your control workstation you can view it as a server to the SP system applications. The subsystems running on the control workstation are the SP server applications for the SP nodes. The nodes are clients of the control workstation server applications. The control workstation server applications provide configuration data, security, hardware monitoring, diagnostics, a single point of control service, and, optionally, job scheduling data and a time source.

As in all servers the reliability of the servers will affect the availability of the clients. In this case the availability of the SP system as a whole is affected. See "Eliminating the Control Workstation as a Single Point of Failure" on page 90 for more details and what happens when a single control workstation configuration has a control workstation failure. When configuring your control workstation, availability of the resources should be a key consideration.

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, refer to Chapter 4, "Planning for a High Availability Control Workstation" on page 87.

## Software and Hardware Requirements for Control Workstations

### Required Software

- 5765-C34 AIX 4.2.1 (or later) or 4.3 server
- 5765–654 IBM Performance Toolbox for AIX, Agent Component 2.2.

  Only the Agent Component (perfagent) is required. This product provides the capability to monitor your SP system's performance, collects and displays statistical data for SP hardware and software, and simplifies run-time performance monitoring of a large number of nodes.

**Note:** Be cautious when you order perfagent. It has several feature codes. You must choose the one relative to the level of AIX that you are running.

- 5765-529 Parallel System Support Programs for AIX 2.4
- 5765–423 IBM C for AIX, V3 or 5765-421 C Set ++ for AIX, V3.

  One license is required for the SP. Concurrent licensing is recommended so the one license can float across the SP nodes and the control workstation. It is needed for **crash** to work effectively and to obtain IBM software support for the SP system. You can order the license as part of the SP system. It is not specifically required on the control workstation if a license server for AIX for C or C++ exists some place in the network and the SP is included in the license server's cell.

## Optional Software

The control workstation and its software are *not* 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.

The HACWS software option, a priced feature of PSSP, must be purchased in addition to the SP hardware, and is supplied on separate install media. For software requirements for HACWS, refer to “Software Requirements for HACWS Control Workstation Configurations” on page 94.

Service Director for RS/6000 is provided with the SP but is optional to be run on the control workstation. 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 and the level of error log forwarding or network access. See “Service Director for RS/6000” on page 132 for requirements and planning.

## Supported Control Workstations

The SP system requires a IBM RS/6000 workstation as a point-of-control for managing, monitoring, and maintaining the SP frames and individual processor nodes. See Chapter 4, “Planning for a High Availability Control Workstation” on page 87 for more planning information about the control workstation. The control workstation you supply connects to each frame through an RS-232 cable and the SP Ethernet.

The following RS/6000s are supported as **MCA** control workstations:

- RS/6000 7013 Model 5XX
- RS/6000 7012 Model 3XX
- RS/6000 7012 Model GXX
- RS/6000 7030 Model 3XX

  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.
- RS/6000 7015 Model 9XX, and RXX in either the 7015-99X or 7015-R00 rack.

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.

- RS/6000 7013 Model JXX

  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.

- RS/6000 7011 Model 25X (not recommended, total system performance is slow)
- RS/6000 7009 Model C10 and C20 (not recommended beyond one frame due to slow tty response)

The following RS/6000s are supported as **PCI** control workstations:

- RS/6000 7025–F30 PSSP 2.2 and later

  On new systems, either the 8–port (#2493) or 128–port (#2944) PCI bus asynchronous adapter should be used for frame controller connections.

  Use of the support processor option (#1001) is highly recommended. If this option is used, 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.

- RS/6000 7024–E20 PSSP 2.2 and later
- RS/6000 7024–E30 PSSP 2.2 and later
- RS/6000 7025–F40 and 7025–F50

  The native RS232 ports on the system planar can not be used as tty ports for the hardware controller interface. The 8-port asynchronous adapter EIA-232/ RS-422, PCI bus (#2943) or the 128–port Asynchronous Controller (#2944) are the only RS232 adapters that are supported. These adapters require AIX 4.2.1 or AIX 4.3 on the control workstation.

  Use of the support processor option (#1001) is highly recommended.

- RS/6000 7026–H10

  The native RS232 ports on the system planar can not be used as tty ports for the hardware controller interface. The 8-port Asynchronous Adapter EIA-232/ RS-422, PCI bus (#2943) or the 128–port Asynchronous Controller (#2944) are the only RS232 adapters that are supported. These adapters require AIX 4.2.1 or AIX 4.3 on the control workstation.

- RS/6000 7043–140 and 7043–240

  The native RS232 ports on the system planar can not be used as tty ports for the hardware controller interface. The 8-port asynchronous adapter EIA-232/ RS-422, PCI bus (#2943) or the 128–port Asynchronous Controller (#2944) are the only RS232 adapters that are supported. These adapters require AIX 4.2.1 or AIX 4.3 on the control workstation.

  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 includes expansion frames.

### Supported Control Workstations in Limited Environments

The following RS/6000 models are supported in limited environments:

- RS/6000 7006-41T/W is a supported control workstation if the High Availability Control Workstation (HACWS) feature is never going to be configured.
- RS/6000 7006-42T/W is a supported control workstation if the High Availability Control Workstation (HACWS) feature is never going to be configured.

### Unsupported Control Workstations

The following RS/6000 models are **not supported** as control workstations:

- RS/6000 7007 Model N40
- RS/6000 7008 Model XXX
- RS/6000 7011 Model 22X
- RS/6000 7011 Model M2X
- RS/6000 7011 Model 23X
- RS/6000 7020 Model 40P
- RS/6000 7247-82X
- RS/6000 7248-1XX Model 43P
- RS/6000 7249-851
- RS/6000 7586-43P

**Please Note:**

Any RS/6000 not listed as either supported, supported in limited environments, or as unsupported, will not function as a 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 recommended.
- Four GB of disk storage. If the SP is going to use an HACWS configuration, you can configure 2GB of disk storage in the rootvg volume group and 2GB 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 2GB 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 LPP growth. Keep in mind that if you have nodes at different levels of PSSP or AIX , each node requires its own LPP source which 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 1GB for each AIX release and modification level for lppsource files. Additional disk space should be added for mksysb images for the nodes.

If you plan on using rootvg mirroring, then double the number of physical disks you estimated so far.

- Physically installed with the RS-232 cable to within 12 meters of each SP frame.
- Equipped with the following I/O devices and adapters:
  - Three and a half inch diskette drive
  - Four or eight millimeter (or equivalent) tape drive
  - 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. A backup control workstation does not satisfy this requirement. This additional connection is used to access the control workstation from the network when the SP nodes are down.
  - SP Ethernet adapters for connection to the SP Ethernet

    The number of ethernet adapters required depends completely on the ethernet topology you use on your SP. 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 MB ethernet adapters. A network hub or switch is required.
  - A SCSI CD-ROM device

## HACWS Minimum Requirements

The following requirements are in addition to the previous requirements:

- 2 Supported RS/6000 workstations

  Each of these RS/6000s must have the same set of I/O required for control workstations as listed above. 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.
- External disk storage that is supported by HACMP and the control workstation being used:
  - 2 external disk controllers and mirrored disks are strongly recommended 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.

- The HACWS connectivity feature #1245 on each SP frame
- An additional RS232 connection for HACMP communication is needed if target mode SCSI is not being used for the HACMP communication.

## Hardware Controller Interface Planning

Each frame of the SP must be attached to the control workstation by an RS-232 line connected to a serial port on the workstation.

**Note:** IBM highly recommends that you use asynchronous adapter cards instead of the native RS232 ports on the system units. Several RS/6000 systems do not support the use of the native serial ports for the frame controller connections. See *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment* for specific hardware.

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

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

Complete one set of worksheets for each control workstation you will configure. The `ABC Corporation` completed Worksheet 13 in Table 19 on page 52, and Worksheet 15, in Table 20 on page 53.

You should complete Worksheet 13, "SP Control Workstation" in Table 53 on page 239, Worksheet 14, "Select a Time Zone" in Table 54 on page 240, and Worksheet 15, “SP Control Workstation Network” in Table 55 on page 241.

*Table 19. SP Control Workstation*

| **SP Control Workstation - Worksheet 13** | | |
|---|---|---|
| **SP Control Workstation Image:** | | |
| | Control Workstation Name | cws01 |
| | Model | 7025- F30 |
| | Install rootvg on disk | 1 |
| | Disk Space | 4GB |
| | Memory Size | 128MB |
| **Hardware Options and Adapters:** | | |
| | *Type* | *Quantity* |
| | ATM | |
| | Ethernet | 1 |
| | FDDI | |
| | Token Ring; Speed 16MB | 1 |
| | Multiport Serial Adapters | |
| | 8 mm tape drive | 1 |
| | CD-ROM | 1 |
| **IBM Licensed Products:** | | |
| | AIX | |
| | PSSP | |
| | C Set ++ for AIX | |
| | LoadLeveler | |
| | perfagent | |
| | | |
| | | |
| | | |
| | | |
| **Other Applications:** | | |
| | NFS | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Table 20. SP Control Workstation Network*

**SP Control Workstation Network - Worksheet 15**

**Name**__ABC Corporation_____
**Date**__January 25, 1998
**System Name**__spsystem1__
**Control Workstation Name**__cws01__

| Frame Hardware Control Connections (RS-232) | | | | Control Workstation Network Connections (note 2) | | | |
|---|---|---|---|---|---|---|---|
| **Frame Number** | **Serial Port for RS-233 Control Line (note 1)** | **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 |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

**Notes:**

1. Use the SMIT **tty** menu or the **mkdev** command to configure these ports.
2. Use the SMIT **mkinet** menu or the **mkdev** command to configure your SP Ethernet connection.

# Chapter 3. Defining the Configuration that Fits Your Needs

This chapter provides the information you need to plan to configure your system before installing it. There is an SP Site Environment Planning Worksheet in Appendix C, “SP System Planning Worksheets” on page 227 to use with this chapter. Once completed, you'll use this worksheet to:

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

Make copies of Worksheet 16, **SP Site Environment** (page 242) before you begin.

## The Impact of Software Planning on Site Planning

Planning and configuring your SP system software has an impact on the SP site plan you select. The following sections discuss some of the system planning decisions you need to make, and their impact on performance and site planning.

## Planning Your Site Environment

You plan your site environment by entering site configuration information on the control workstation through SMIT panels or by using the **spsitenv** command. SMIT is the System Management Interface Tool, supplied as part of the PSSP software.

The installation and configuration scripts read the configuration information data and customize the SP configuration according to your choices. The entries you put on the worksheet are the entries you'll make on the SMIT panels. Site environment data includes:

- The name of the default network install image
- Your method of time service, the name of your time servers, and the version of NTP in use
- Whether you want to have the SP services configure and manage the Automounter
- User Admin information and how you want to use RS/6000 SP User Management
- Whether you want RS/6000 SP File Collection Management and where it will run
- Whether you want to use RS/6000 SP Accounting
- 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.)

You can easily change the choices discussed in the following sections any time after the installation. If you are unsure about any of these options, you can safely select the defaults, then change your selections later.

© Copyright IBM Corp. 1997, 1998

## 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 **SP Site Environment Worksheet** on page 242. A brief description of the function of each area along with a discussion of the alternatives should give you enough information to fill out the worksheet. Detailed information about these and other system administration issues is in the section on managing the SP system in *IBM Parallel System Support Programs for AIX: Administration Guide*.

Remember, the defaults are designed to provide a workable SP system. You can change them later, if 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 SP node when the install image field is not set. The default is **bos.obj.ssp.421**, shipped with the SP System.

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.

You can reclaim this disk space by setting *remove_image* 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 21. Network Install Image Choices*

| | **Worksheet Entries To Be Filled In** |
|---|---|
| **To do this....** | *remove_image* |
| Remove the network install image after all installs have completed | **true** |
| Do not remove the network 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)

By default the SP system uses Network Time Protocol (NTP) to synchronize the time-of-day clocks on the control workstation and SP nodes. There are several ways in which you might currently be synchronizing the time-of-day in your existing computing environment:

- You might already be using NTP, either locally or through the Internet

- You might be using some other time service software
- You might not have an established method for synchronizing the system clocks on the computing systems throughout your environment.

Kerberos ticket expiration depends on proper time synchronization, so the SP system provides several options for time keeping:

- If you have an established NTP time server, you can use it to synchronize and manage time on the SP system.
- You can choose an NTP time server from the Internet.
- You can run NTP locally on the SP system to generate a consensus time.
- You can choose not to use NTP at all, relying on another method at your site.

> **Note**
>
> The SP machines do not have system batteries. If you choose not to use NTP, you must have another way to manage clock synchronization.

- You cannot choose the control workstation or backup control workstation to be the time master.

See *IBM Parallel System Support Programs for AIX: Administration Guide* for managing NTP.

## High Availability Control Workstation Considerations

If you install a High Availability 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 install a High Availability Control Workstation and use the Internet configuration, both control workstations get time from the Internet. 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 the SP System). If your installation is using an earlier version of NTP, change this value. The other two attributes are described in Table 22 on page 58.

*Table 22. 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.
- * Refer to **README.public** in **/usr/lpp/ssp/public** for information on Internet time servers.

# Understanding User Directory Mounting Choices—Automount Daemon

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.

## Automount Considerations

As of PSSP 2.3, the automounter daemon known as Amd has been replaced by an AIX resident automounter daemon called Automount. Both the AIX automounter and Amd provide the same basic functions, but Amd offers more customization options than the AIX automounter. Therefore, if you have a complex Amd configuration, you might not find equivalent functions using the AIX automounter.

Implementation of the AIX resident automounter requires PSSP 2.3 (or later) on the control workstations. Booting the control workstation creates all automounter directories, map files, and logs needed by the system. Booting the CWS also converts any existing user directory Amd map files into AIX resident automounter map files. If you have modified the user directory map files prior to upgrading your system, these conversion utilities might fail. All other map files will need to be converted manually by the customer.

Booting the SP nodes invokes a similar process creating node directories and logs. Map files are downloaded from the CWS 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 Work Station.

Automount uses NFS (Network File Systems) to mount or AIX to link directories. Nodes running PSSP 2.3 (or later) operate Automount by default. Pre-PSSP 2.3 nodes still run Amd and Amd will still be included with pre-PSSP 2.3 packages. However, Amd is no longer supported by IBM for PSSP 2.3 and above. Therefore, it is up to your System Administrator to supply and maintain this software if you wish to run Amd on nodes operating at PSSP 2.3 (or later). As an alternative to Automount and Amd, you can also 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 *IBM Parallel System Support Programs for AIX: Administration Guide*.

### Site Environment Worksheet Entries

Only one attribute applies to the Automount option.

*Table 23. User Directory Mounting Choices - System automounter support*

| | **Worksheet Entries To Be Filled In** |
|---|---|
| **To do this....** | *amd_config* |
| Use **AIX Automounter** supplied with SP Parallel System Support Programs | **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 Print Management Choices

The SP Print Management System has been removed as of PSSP 2.3. That is, the SP Print Management System cannot be configured on nodes running PSSP 2.3 (or later). IBM recommends the use of Printing Systems Manager (PSM) for AIX as a more general solution to managing printing on the SP system.

However, if you are running earlier versions of PSSP on some of your nodes, the SP Print Management System is still supported on those nodes. Because of that, SP systems with pre-PSSP 2.3 nodes will have Print Management configured on the control workstation (even if the control workstation is at PSSP 2.3 or later) for coexistence support.

If you are running mixed levels of PSSP in your system, be sure to maintain and refer to the appropriate documentation for whatever versions of PSSP you are running.

On nodes running PSSP 2.2 or earlier, the SP Print Management System programs bypass the standard AIX print command subsystem and route print output to one or

more print servers. When spooling printer output, Print Management operates in either **open** or **secure** mode.

Open mode requires all users to have **rsh** privileges to the print hosts. Secure mode denies users **rsh** privileges; however, it requires the use of a special user account, with a default userid of **prtid**, to transfer the jobs to the print host. In secure mode, all print jobs are *owned* by this special user account. Your third option is to not use the SP print subsystem and use some other means of handling print output from your SP system. Standard AIX print queue support may be adequate for small systems.

### Site Environment Worksheet Entries

You have three options for handling print output.

*Table 24. Print Management Choices*

| | **Worksheet Entries To Be Filled In** | |
|---|---|---|
| **To do this....** | *print_config* | *print_id* |
| Use the SP print subsystem in open mode | **open** | |
| Use the SP print subsystem in secure mode | **secure** | **prtid** (default) (userid of your print host) |
| Do not use the SP print subsystem | **false** (default) | |

**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

User account management for the SP system 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 file collection technology. If you are using Network Information Service (NIS), these commands will utilize NIS. A set of customer commands is provided to interface to this function.

These options are offered to help you manage user accounts. These 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 hostname 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 hostname 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 *IBM Parallel System Support Programs for AIX: Administration Guide*.

### Site Environment Worksheet Entries

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

*Table 25. 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 the SP user account management software | **false** | | | | |
| Use the SP user account management software | **true**(default) | password server hostname (ctl wkstn - default) | name of the password file<br>**(/etc/passwd)**<br>(default) | hostname of the home directory server<br>*(ctl wkstn)*<br>(default) | **/home/**<name of your home directory server> |

**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 SP file collection technology 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 *IBM Parallel System Support Programs for AIX: 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) userid for the file collection daemon along with a unique (unused) port through which to communicate.

*Table 26. System File Management Choices*

| | **Worksheet Entries To Be Filled In** | | |
|---|---|---|---|
| **To do this....** | *filecoll_config* | *supman_uid* | *supfilesrv_port* |
| Do not use the SP file collection technology | **false** | | |
| Use the SP file collection technology | **true** (default) | unique user ID (default **102**, username **supman**) | unique port number (default **8431**) |

## 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 may be enabled or disabled for accounting.

### Site Environment Worksheet Entries

The following attributes apply to SP accounting, 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 for jobs having 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 27. Accounting Choices*

| | **Worksheet Entries To Be Filled In .....** | | | |
|---|---|---|---|---|
| **To do this...** | *spacct_enable* | *spacct_actnode_thresh* | *spacct_exclusive_enable* | *acct_master* |
| Do not use the SP accounting | **false** (default) | | | |
| Use the SP 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 on this utility and how to set up an accounting system, see the chapter on accounting in *IBM Parallel System Support Programs for AIX: Administration Guide* and in *AIX Version 4 System Management Guide*.

## Understanding LPP Source Directory Name Choices

The *cw_lppsource_name* attribute lets you specify the name of the directory to which the AIX file sets (the lpp source) will be copied.

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 lpp source name for the nodes (such as aix421 if that is what you choose to call the subdirectory with the AIX 4.2.1 lpp source). Keep in mind that the setup_server program looks for this name later in the installation process. By default, it is set to the string "default", so that if you use that as your subdirectory name, you do not have to change the value of *cw_lppsource_name*. If you do not provide a name, setup_server assumes the value is **default**.

See the chapter on preparing the control workstation in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

### Site Environment Worksheet Entries

Only one attribute applies to the LPP source directory name option.

*Table 28. LPP Source Directory Choices*

| | **Worksheet Entries To Be Filled In** |
|---|---|
| **To do this....** | *cw_lppsource_name* |
| Use **aix421** to uniquely identify the new lpp source subdirectory | **aix421** |
| Use **default** as the default lpp source subdirectory | |

**Note:**

- Change default attribute values to suit your environment.
- Blank entries imply that you make no substitutions for these values.

## 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 *AIX Version 4 Performance Tuning 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 CWS to each of the nodes in the SP that are to be operated and managed by that CWS 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 Version 4 Performance Tuning Guide* publication.

## Boot/Install Server Requirements

When planning your SP Ethernet topology you should 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. The following are recommended 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 issues to consider.

The following requirements exist for all configurations:

- Each boot/install server's en0 ethernet adapter must be directly connected to each of the CWS's ethernet adapters.
- The NIM clients that are served by boot/install servers, must be on the same subnet as the boot/install server's ethernet adapter.
- NIM clients must have a route to the CWS over the SP Ethernet.

## Single Frame Systems

For small systems, you can use the control workstation as the network install server. This means that the SP Ethernet is a single network connecting all nodes to the control workstation. When installing the nodes, you should limit yourself to installing 8 nodes at a time because this is the limit of acceptable throughput on the Ethernet. Figure 4 shows an Ethernet topology for a single-frame system.

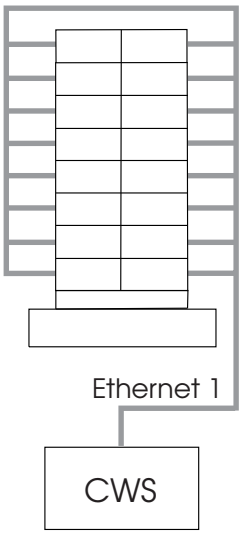



*Figure 4. Ethernet topology for a single-frame SP*

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. Each network should be connected to half of the SP nodes. When network installing the frame, you can install all 16 nodes at the same time. Figure 5 shows this alternate Ethernet topology for a single-frame system.

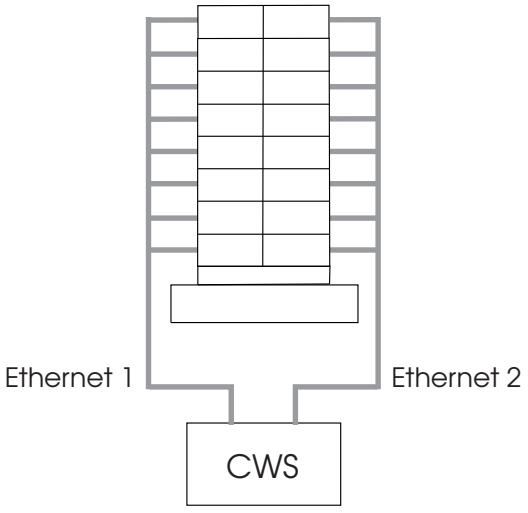



*Figure 5. Ethernet topology for a single-frame SP*

You have to set up your SP Ethernet routing so nodes on one Ethernet can communicate to nodes on the other network. You also need to 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 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 6, 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 6 shows an Ethernet topology for this multiple-frame system.

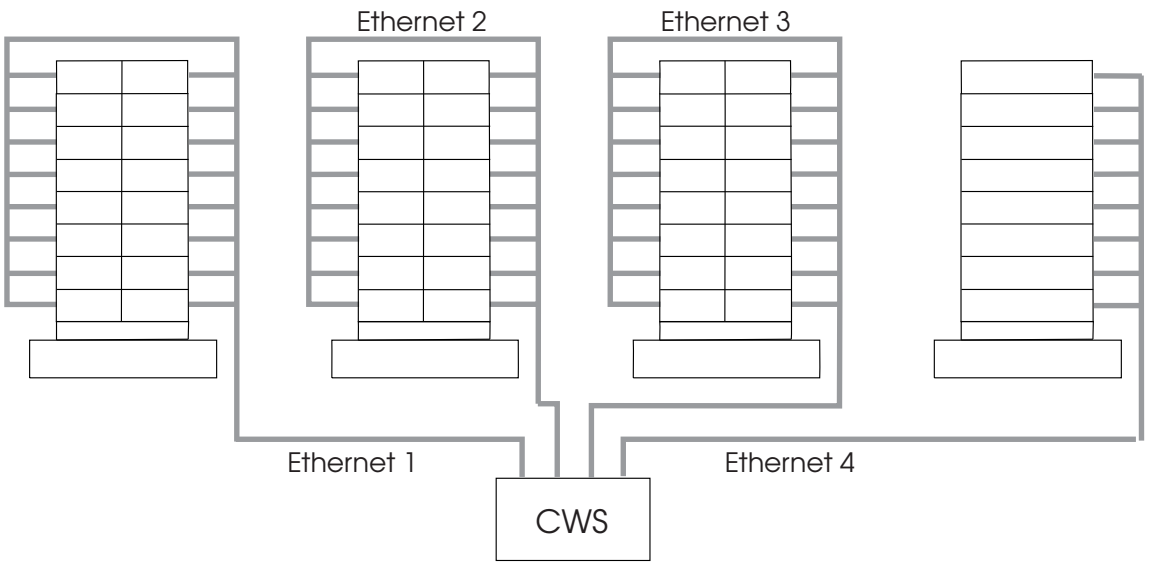



*Figure 6. Method 1 Ethernet topology for a multiple-frame SP*

   Once again, you will have to set up your SP Ethernet routing so nodes on one Ethernet can communicate to nodes on another. You also need to 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 7 on page 67 shows an Ethernet topology for this multiple-frame system.

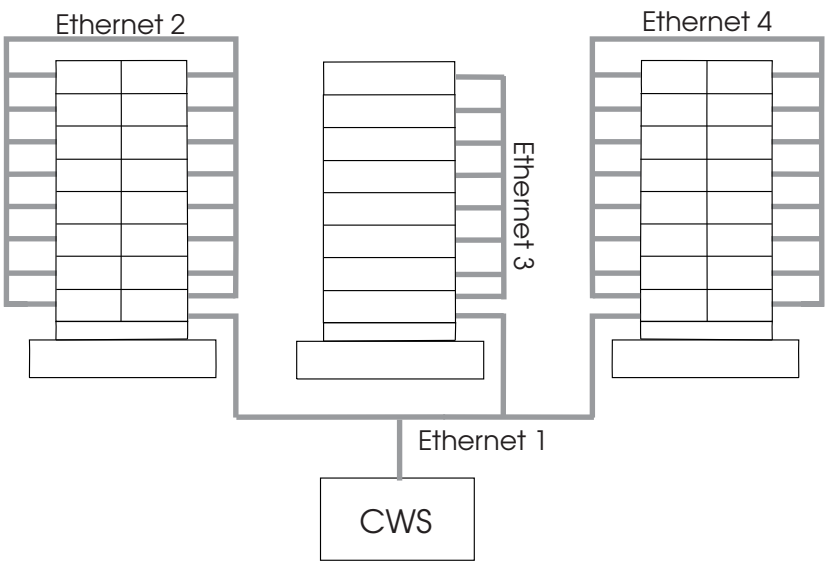



*Figure 7. Method 2 Ethernet topology for a multiple-frame SP*

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 traffic (installs, SDR activity, POE, etc.) 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 may cut that traffic in half, but the control workstation must be large enough 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. If the control workstation may also be attached externally, providing another route between nodes and the control workstation. Figure 8 on page 68 shows this Ethernet topology for such a multiple-frame system.

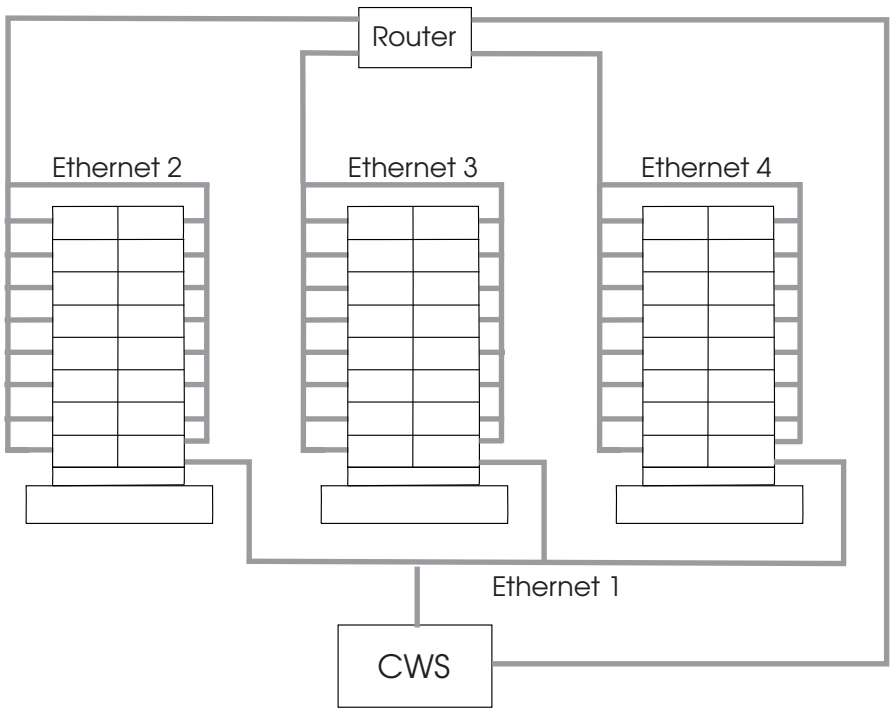



*Figure 8. Method 3 Ethernet topology for a multiple-frame SP*

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 9 on page 69. 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.

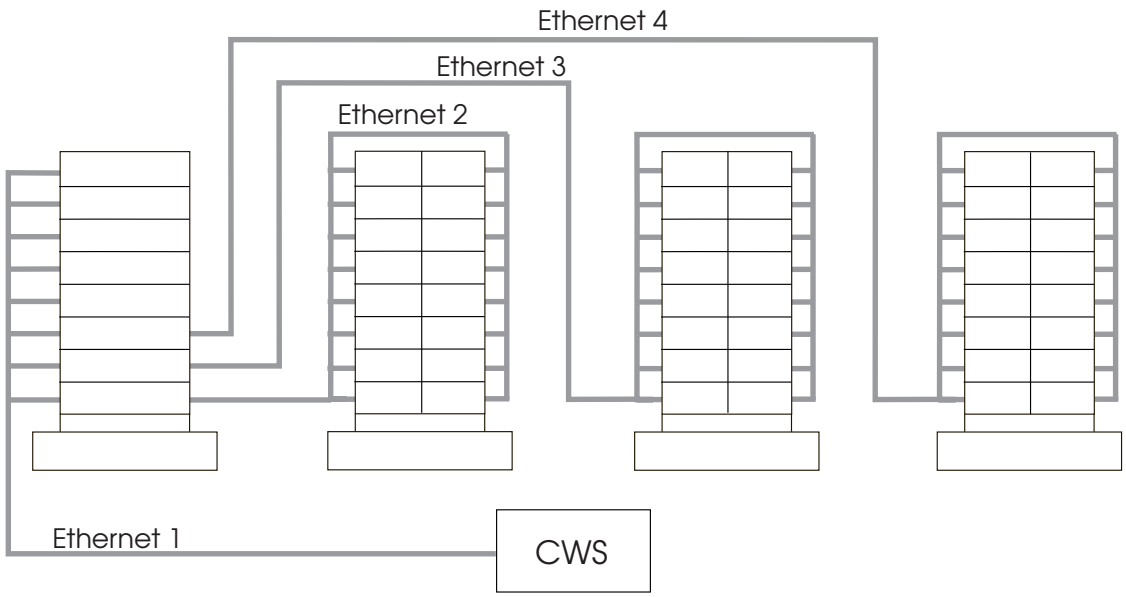



*Figure 9. Boot Server Frame Approach*

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. In a system that consists of just Ethernet and the High Performance Switch, you cannot optimize so that you achieve peak performance on both networks. You have to decide which network you want to be affected.

## 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 SP system, you must define one or more authentication servers. Authentication provides a more secure SP system by verifying the identify of clients that access key systems management facilities. Your SP system's control workstation can be an authentication server, as can other independent workstations. The SP nodes should not be used as secondary servers. At least one secondary server is recommended for improved availability and possibly improved performance. You can install and configure PSSP authentication servers or integrate your SP system into an existing authentication domain, such as an AFS cell. If you choose to use AFS (Version 3.4 for AIX) authentication servers, note in particular the section on the assignment of TCP/IP port numbers in the **/etc/services** file.

You should carefully consider whether the SP control workstation will be an authentication server. You may want to set up your servers on independent RS/6000 workstations that are isolated by physical location or have limited network access. Your primary authentication server must be installed and operating before you install and configure your control workstation, unless it will be the primary server.

See *AIX Version 4 System Management Guide* and *IBM Parallel System Support Programs for AIX: Administration Guide* for more information.

## Understanding Node Hard Disk Choices

The *install_disk* attribute determines which disks are used to create the root volume group (**rootvg**) and to transfer the **mksysb** image during AIX network installation of a node. The default value of this attribute is **hdisk0**. Depending on your environment, you may have the installation include another hard disk.

If you plan to use the Alternate Boot System Image ability, the choice of two levels of software to boot per node, you must use only one disk in the *install_disk* attribute.

You may use more than one disk when the **mksysb** image is larger than the disk or when you need the root volume group to span multiple disks. 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 may come up as **hdisk0**.

Check your disks to ensure your install image is on internal disks.

If you do not have either of these requirements, you should not install on more than one disk. If you have another disk, you can define a different 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.

To change the *install_disk* attribute, use the **spbootins** command with the **-h** option. See *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on the **spbootins** command.

## Determining Space Requirements

Your SP install package includes standard installation images plus optional images that you order. You must sum the estimated sizes of all the products you plan to run. The example installation in Table 29 on page 72 includes:

- An image comprised of the required PSSP components (**ssp**)
- An image comprised of the AIX file sets (**spimg**)
- An image of PSSP optional Virtual Shared Disk components (**ssp.csd**)
- An image for each optional PSSP-related LPP, in this case HACWS (**ssp.hacws**), and PTPE (**ssp.ptpegui**).

Use this as an example to learn how to determine your space requirements.

## 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.

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, this might save time and effort if you intend to use **installp** for additional file sets that are not already installed on your nodes. IBM recommends 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 IBM Parallel System Support Programs for AIX: Installation and Migration Guide, Chapter 2, which defines how to download the file sets.

Downloading all of the AIX file sets requires approximately 1.5GB of disk space. Downloading only the minimal AIX file sets requires approximately 500MB.

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 (AIX 4.2.1 or 4.3.0). Each level uses approximately the same amount.

## 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. **mksysb** images can vary in size from about 65MB to upwards of 700MB. 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 LPPs 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. For example, HACWS, PTPE, and C Set ++ are all additional LPP's 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 space you need for /spdata:

```
lppsource + mksysb_images + pssp_image + optional_images = total_additional_space
```

For example, the minimum space needed on the control workstation for /spdata to contain lppsource plus only the required images is:

```
500MB + 65MB + 124MB = 689MB
```

lppsource + spimg mksysb + ssp = minimum space

Table 29 summarizes the amount of space some images require. See Table 30 on page 73 for the amount of space used by individual file sets.

*Table 29. Space Required by installp Images*

| Space Required for Storing installp Images | | |
|---|---|---|
| **installp Image** | **Space Required** | **Description** |
| ssp | 124 MB | This has the required PSSP components. It must be on the control workstation. |
| spimg | 75 MB (AIX 4.3.0)<br>65 MB (AIX 4.2.1) | This is the AIX image and must be on the control workstation. |
| ssp.csd | 4.3 MB | This is the PSSP image for the optional Virtual Shared Disk component. If used, it must be on the control workstation and can also be on other nodes. |
| ssp.hacws | 167 KB | This is an optional image which must be on the control workstation if used. |
| ssp.ptpegui | 2.4 MB | This is an optional image which must be on the control workstation if used. |

*Table 30. Space Used by Individual File Sets*

| Space Used by Individual File Sets | |
|---|---|
| **Space Used by PSSP image: ssp** | |
| **File Set** | **Total Storage** |
| ssp.clients | 13.5 MB |
| ssp.basic | 5.2 MB |
| ssp.sysman | 1.1 MB |
| ssp.css | 10.8 MB |
| ssp.jm | 1.1 MB |
| ssp.public | 14 MB |
| ssp.gui | 26 MB |
| ssp.docs | 64 MB |
| ssp.sysctl | 812 KB |
| ssp.authent | 602 KB |
| ssp.top | 1.4 MB |
| ssp.top.gui | 2.6 MB |
| ssp.st | 983 KB |
| ssp.ha | 17 MB |
| ssp.perlpkg | 8.3 MB |
| ssp.pman | 721 KB |
| ssp.spmgr | 606 KB |
| ssp.topsvcs | 2.3 MB |
| ssp.ucode | 620 KB |
| **Space Used by PSSP image: spimg** | |
| **File Set** | **Total Storage** |
| spimg.430 | 75 MB |
| spimg.421 | 65 MB |
| **Space Used by PSSP image: ssp.csd** | |
| **File Set** | **Total Storage** |
| ssp.csd.vsd | 450 KB |
| ssp.csd.cmi | 115 KB |
| ssp.csd.hsd | 156 KB |
| ssp.csd.sysctl | 370 KB |
| ssp.csd.gui | 3.5 MB |
| **Space Used by PSSP image: ssp.hacws** | |
| **File Set** | **Total Storage** |
| ssp.hacws | 167 KB |
| **Space Used by PSSP image: ssp.ptpegui** | |
| **File Set** | **Total Storage** |
| ssp.ptpegui | 2.4 MB |
| **Note:** The total storage can cross multiple file systems. | |

# 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 11 and are summarized in Appendix C, "SP System Planning Worksheets" on page 227.

# 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 information contained in "Completing the SP Node Layout Worksheets" on page 34. This repetition is necessary because of the information's importance.

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.

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 hostname. For the SP, IBM recommends that you assign unique hostnames to each IP interface on your nodes.

# Understanding the SP Networks

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

## SP Ethernet

SP Ethernet is the name of the LAN that connects all SP nodes to the control workstation. The Parallel System Support Programs (PSSP) use this connection for network installs and other SP functions.

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

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 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.
- Multiple-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.
- Multiple-subnet, multiple-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 64 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 multiple-stage, multiple-subnet SP Ethernet, 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 en0 interface. Therefore, en0 of the node must be the Ethernet that is connected to the boot/install network.

## Additional LANs

The SP Ethernet can provide a means to connect all nodes and the control workstation to your site networks. However, it is likely that you will want to connect your SP nodes to site networks through other network interfaces. If the SP Ethernet is used for other networking purposes, the amount of external traffic must be limited. If too much traffic is generated on the SP Ethernet, the administration of the SP nodes may be severely impacted. For example, problems may 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.

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

## IP over the Switch

If your SP has a switch and you want to use IP for communications over the switch, each node needs to have an IP address and name assigned for its switch interface, the **css0** adapter. The **css0** adapter applies to both the High Performance Switch and the SP Switch. If hosts outside the SP switch network need to communicate over the switch using IP with nodes in the SP, 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 starting node's IP address. After the first address is selected, subsequent node addresses are based on the switch port number. See “Understanding Node Numbering and Switch Node Numbering” on page 79. Unlike all other network

interfaces, which can have sets of nodes divided into several different subnets, the switch IP network must be one contiguous subnet which includes all the nodes in the system.

If you want to assign your switch IP addresses as you do your other adapters, you must enable ARP for the **css0** adapter. If you enable ARP for the **css0** adapter, you may 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.

### Subnetting Considerations

All but the simplest SP system configurations will likely include several subnets. Thoughtful use of netmasks in planning your networks can economize on the use of network addresses. Refer to *AIX Version 4 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).

## Planning Considerations for Network Router Nodes

If you are ordering 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 may not be applicable to your SP configuration. However, if you are not ordering the SP Switch Router, then this section describes 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 may vary depending on your network type. (In some environments, router nodes may also be called gateway nodes.)

For nodes that use 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 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 FDDI as the customer network. You also should not connect more than one FDDI to a router node.

Applications, such as POE and the Resource Manager, should run on nodes other than FDDI routers. However, Ethernet and token ring gateways can run with these applications.

For systems that use Ethernet or token-ring routers, traffic can be routed through the SP Ethernet 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 FDDI networks, traffic should be routed across the switch to the destination nodes. The amount of traffic coming in through the FDDI network can be up to 10 times the bandwidth the SP Ethernet can handle.

Information about configuring network adapters, and tuning the various network tunables on the nodes is in *IBM Parallel System Support Programs for AIX: Administration Guide*.

## Planning Considerations for the SP Switch Router

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

- HIPPI
- FDDI
- HSSI
- 10 or 100 BaseT
- ATM
- Ethernet
- 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 four to sixteen gigabits 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.

The SP Switch Router requires PSSP 2.3 or later on both the primary node and the primary backup node. Using the SP Switch Router Adapter, the SP Switch Router can be connected to an SP Switch (8–port or 16–port).

**Note:** The SP Switch Router cannot be connected to the High Performance series of switches.

## SP Switch Router Network Connections

The SP Switch Router Adapter in the SP Switch Router can be attached to an SP switch to improve throughput of data coming into and going out of the RS/6000 SP system. Each SP Switch Router Adapter in the SP Switch Router requires a valid unused switch port 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.

***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 the SP Switch Router Adapter.

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 within a wide node 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.

**Note:** In this example, attaching an SP Switch Router Adapter to an even number node slot will prevent you from using the frame as a non-switched expansion frame. Any future system expansion in this instance will require an SP switch.

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.

**Note:** The frame in this example is a mixed frame. Therefore the SP Switch Router Adapter can be attached to any valid switch port without penalty to

future expansion since a mixed frame cannot be used as a non-switched expansion frame.

These rules are discussed in more detail later in this chapter.

## Understanding Node Numbering and Switch Node Numbering

Use the information in this section for assigning IP addresses to the nodes and the node SP switch interface.

### 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 full-sized SP frame from the front, the 16 slots are numbered sequentially from bottom left to top right.

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

- A thin node occupies a single slot in a drawer and its slot number is the corresponding slot. (See thin nodes in Figure 10.)
- A wide node occupies two slots and its slot number is the odd-numbered slot. (See the wide nodes in Figure 10.)
- A high node occupies four consecutive slots in a frame. Its slot number is the first (lowest number) of these slots.

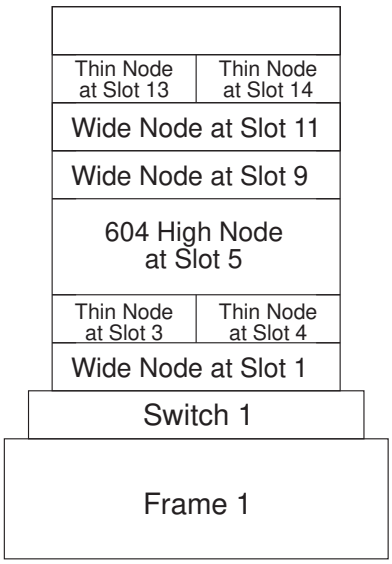



*Figure 10. Node-slot assignment*

## Node Placement

A frame full of thin nodes is called a *thin-node frame*. Similarly, you may have a *wide-node frame* or a *high-node frame*. In addition, a frame that contains nodes of different types is called a *mixed-node frame*.

In what follows, keep in mind that, so long as it physically fits, a wide node may always be placed where two thin nodes might be, and a high node may always be placed where two wide nodes, or four thin nodes, might be.

Currently, the SP supports four frame configurations. Each configuration corresponds to a frame and its possible companion *non-switched expansion frames*. A non-switched expansion frame is a successor frame which shares the subject frame's switch. The capability for sharing a switch is affected by the number of nodes in the frame having the switch. There are restrictions on what constitutes a "supported" frame configuration. Figure 11 on page 81 illustrates the supported configurations for switch ports.

In configuration 0, the model frame has a switch which uses all 16 of its switch ports. Since all switch ports are used, the frame supports zero non-switched expansion frames. Similarly, if a frame does not have a dependable pattern of unused slots, then it cannot share its switch and it also has zero non-switched expansion frames. For example, any frame containing a thin node "threatens" to eventually use all of its node slots. Therefore, no switch ports are available and the frame is not a candidate for expansion. For related reasons, thin nodes are not allowed in a non-switched expansion frame.

An example of a dependable pattern of unused switch ports is where the model frame has only wide nodes. It could use at most eight switch ports, and so has eight to share with other frames. Those other frames are allowed to be configured as in configuration one or two in Figure 11 on page 81. Configuration one has a single non-switched expansion frame using wide nodes while configuration two has two non-switched expansion frames.

In configuration 2, only high nodes are used in the two non-switched expansion frames. However, wide nodes can be substituted for high nodes **if and only if** the wide node is placed in the same address point that the high node would go into. In other words, only four wide nodes could be placed in either non-switched expansion frame in configuration two.

Another example of a dependable pattern of unused slots is where the switched frame is a *high-node frame*: every node is a high node, and they occupy switch ports 0, 4, 8, or 12 as shown in configuration 3 of Figure 11 on page 81. Such a frame uses only four switch ports, and, therefore, its switch can support 12 additional nodes. A restriction here is that each non-switched expansion frame is also considered a *high-node frame*. Therefore, there can be up to three non-switched expansion frames. Each of these non-switched expansion frames can house a maximum of four wide or high nodes. Once again, if wide nodes are used, they **must** be placed in the high node address points.

Mixed-node frames containing thin nodes **cannot** share a switch. Each frame must have its own switch for frame to frame data transfers. A typical system configuration for this instance is shown in Figure 12 on page 82.

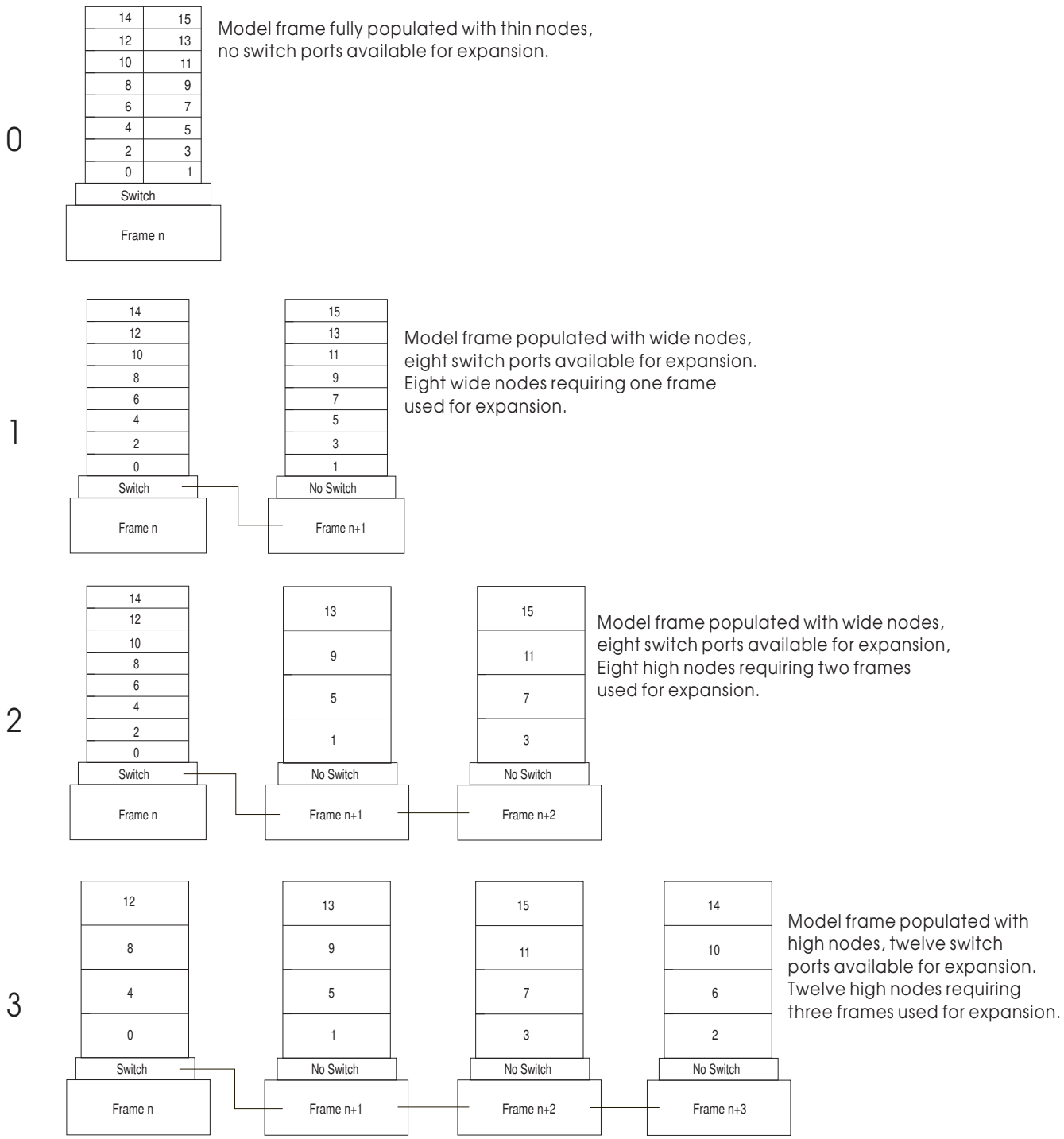
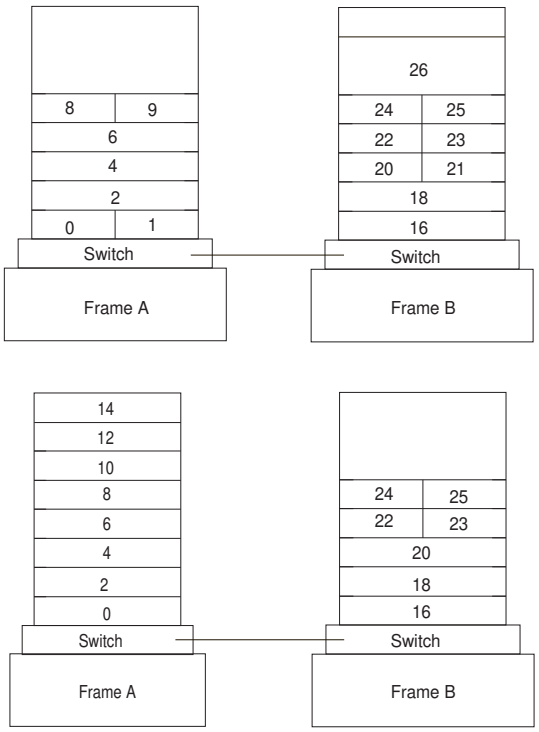



*Figure 11. Supported frame configurations showing switch port assignments*

## Frame Numbers and Switch Numbers

The administrator establishes the frame numbers when the system is installed. Each frame is referenced by the tty port to which the frame supervisor is attached and is assigned a numeric identifier. The order in which the frames are numbered defines the sequence in which frames are examined during the configuration process. This order is used to assign global identifiers to the switch ports and nodes. This is also the order used to determine which frames share a switch.




Node numbering and switch to switch frame connections when thin nodes are in use.

*Figure 12. Node numbering and switch to switch frame connections when thin nodes are in use*

For this discussion:

- A frame that has nodes and a switch is a switched frame.
- A frame with nodes and no switch is a non-switched frame.
- A frame that has only switches is a switch-only frame.

You can skip numbers between switched frames to allow for growth by the addition of non-switched expansion frames at a later time. Nodes in non-switched expansion frames will be attached to the switch in the first preceding switched frame. A maximum of three non-switched expansion frames are supported per switch.

Non-switched expansion frames must be numbered consecutively from the base frame number that contains a switch. For example, if switched frame *n* has three non-switched expansion frames, they must be numbered *n+1*, *n+2*, and *n+3*. If the frame numbers of existing frames had to be changed in order to add a non-switched expansion frame, then part or all of the system would have to be reconfigured.

If you use a switch-only frame, it must be the last frame in the SP system. If you use multiple switch-only frames, they must be numbered contiguously and must be the last frame numbers in the SP. Assign a high frame number to the switch-only frame to allow for the future addition of frames with frame numbers lower than that of the switch-only frame.

To number the switches, the configuration software automatically checks each frame from first to last for the presence of a switch and numbers them beginning

with 1, progressing consecutively to the last switch on the last frame. (See Figure 13 on page 83.) Within switch-only frames, each switch is numbered consecutively, starting with number 1001 at the bottom of the first switch-only frame and progressing to the top of the last switch-only frame.

## Node Numbering for Tall Frames With a Switch

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 by the following formula:

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

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

Node numbers are assigned independent of whether the frame is fully populated. (See Figure 13.)

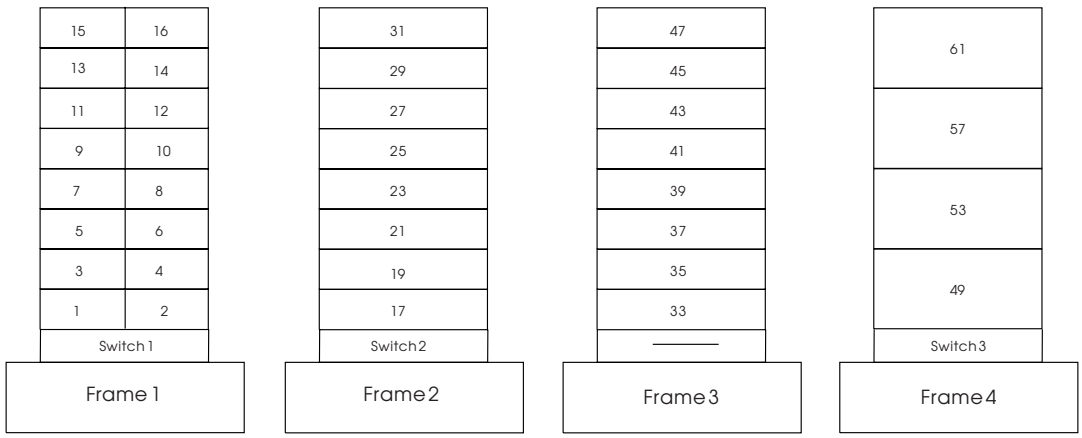



*Figure 13. Node Numbering for an SP System*

## 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 a full size switch, and 8 ports for each of the SP Switch-8 and LC-8 switches.) 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.

These node *switch ports* are numbered sequentially across the switch boards, starting with 0 on board 1. If a node is connected to switch port number *p* in this sequence, then the node *uses switch port number p*, or the node *is switch node number p*.

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

For the 16 port SP and High Performance switches, you can evaluate the following formulas to determine the switch port number to which a node is attached:

| If the node is: | Formula. |
|---|---|
| Connected to a switch within its frame | switch_port_number = (switch_number - 1) X 16 + (slot_number - 1) |
| Connected to a switch outside of its frame (Expansion Frame) | switch_port_number = (switch_number - 1) X 16 + port_number |

Here, *switch_number* is the number of the switch board to which the node is connected and *port_number* is the port position on the switch board to which the node is connected.

For more information on switch port numbers, see Chapter 5, particularly Example 3.

## SP Switch-8 and High Performance Switch LC-8, Port Numbering

The SP Switch-8 is compatible with high nodes, however, the High Performance Switch LC-8 **is not compatible** with high nodes.

Node numbers for short frames are assigned by the same algorithm used to assign node numbers in the tall frames. This formula is:

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

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

**Note:** Extension nodes must be placed into a valid switch port location as verified in the SDR Syspar_map.

For the SP Switch-8, and the High Performance Switch LC-8 a different algorithm is used for assigning nodes their switch port numbers. A system with this kind of switch contains only switch port numbers 0 through 7.

The following algorithm is used to assign nodes their switch port numbers for 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 14 on page 85 and Table 31 on page 85 contain sample switch port numbers for a system with a short frame and an eight port switch.

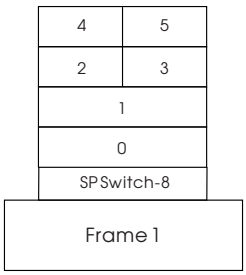



*Figure 14. Switch Port Numbering for an SP Switch-8 and Short Frame*

*Table 31. 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 |
| 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.

## IP 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, 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.

Figure 15 on page 86 illustrates the switch port numbers for an SP system. It also illustrates how the switch port numbers are set for a non-switched expansion frame. In Figure 15 on page 86, Switch 2 connects to the nodes in Frame 3. Specifically, the nodes of Frame 3 use the respective ports of Switch 2 not used by Frame 2. Based on the formula for a frame that does not have a switch, Frame 3's first slot has a switch port number of **17**:

*switch_port_number* = (*switch_number* - 1) X 16 + *port_number*
*switch_port_number* = ( 2 - 1) X 16 + 1
*switch_port_number* = 17

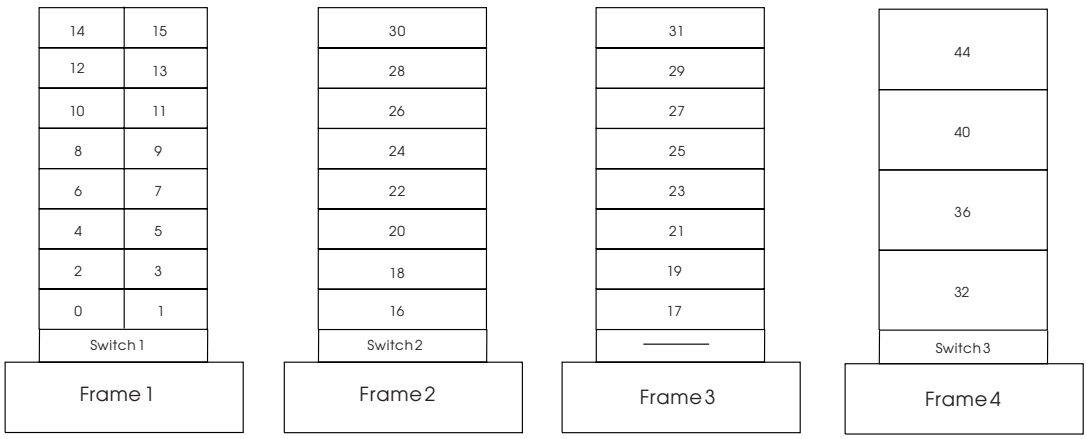



*Figure 15. Switch Node Numbering Sequence*

If ARP is enabled for the **css0** adapter, then the IP addresses may be assigned like any other adapter. That is, they may be assigned beginning and ending at any node, and they do not have to be contiguous addresses for all the **css0** adapters in the system.

# Chapter 4. Planning for a High Availability Control Workstation

**Note:** For specific information on planning your control workstation, refer to "Question 10: What Do You Need for Your Control Workstation?" on page 46.

Planning for a High Availability Control Workstation requires planning for both hardware and software. For hardware planning information read *IBM RS/6000 SP: Planning Volume 1*. That book describes the hardware components and cabling you need to install the High Availability Control Workstation successfully. For information on software requirements, refer to "Software Requirements for HACWS Control Workstation Configurations" on page 94.

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

## Overall System View of a High Availability Control Workstation

The SP system looks similar except that there are two control workstations connected to the SP Ethernet and TTY network. The frame supervisor TTY network is modified to add a standby link. The second control workstation is the backup. Figure 16 on page 88 shows a logical view of a High Availability Control Workstation. The figure shows disk mirroring, an important part of high availability planning.

© Copyright IBM Corp. 1997, 1998

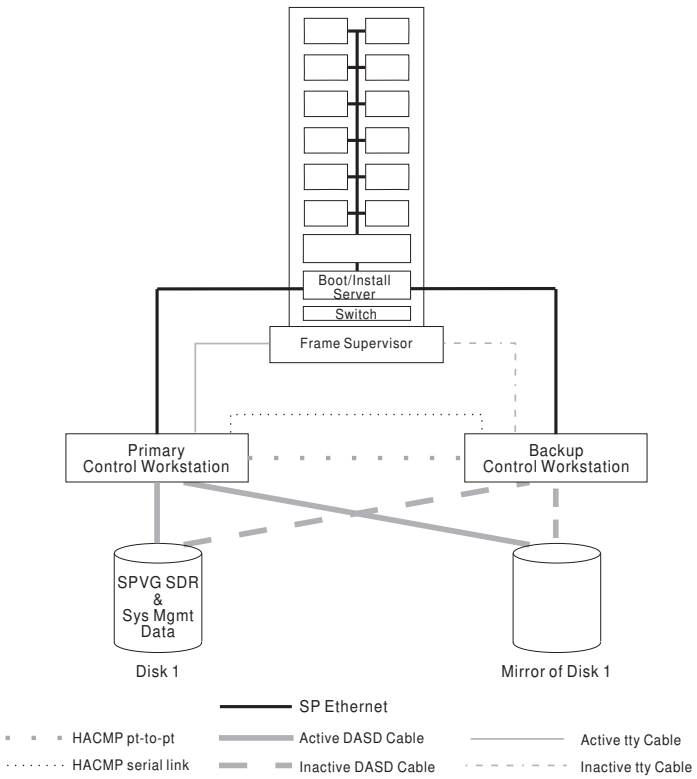



*Figure 16. High Availability Control Workstation With Disk Mirroring*

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 may have to be stopped during failover and reintegration periods.

## Benefits of a High Availability Control Workstation

High Availability Control Workstation is a major component of the effort to reduce the possibility of single point of failure opportunities in the SP. 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, High Availability Control Workstation 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 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 cutover 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 software with High Availability Control Workstation 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 32 (Page 1 of 2). Effect of Failure of Non-High Availability Control Workstation on Mandatory Software*

| **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.<br>2. Nodes cannot be hot-plugged in or out of the frames controlled by the failed control workstation. |
| SDR | 1. Current running jobs continue to completion.<br>2. No new parallel jobs can start.<br>3. The Resource Manager daemons die because they cannot make contact with the SDR.<br>4. Serial jobs can continue to be started.<br>5. No hardware or software configuration changes can occur.<br>6. No installations can be started.<br>7. 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.<br>8. No cluster shutdowns can occur.<br>9. A node can still be powered off and on manually, but this causes a switch fault. |
| Kerberos Authentication Server (if no backup server exists) | 1. Users cannot obtain new tickets via kinit.<br>2. Background processes using rcmdtgt to get ticket will fail.<br>3. Users cannot change passwords.<br>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. |

| *Table 32 (Page 2 of 2). Effect of Failure of Non-High Availability Control Workstation on Mandatory Software* | |
|---|---|
| **Major Software Component** | **Effect on SP System** |
| Availability subsystems (hats, hags, haem) | These subsystems will not restart upon node reboot. |

| *Table 33. Effect of control workstation Failure on User Data on the Control Workstation* | |
|---|---|
| **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.<br>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.<br>2. Records are consolidated after the control workstation comes up. |
| User File Server | 1. Running jobs may fail.<br>2. Jobs may not be able to access needed data. |
| Parallel Environment home node | Running parallel jobs will fail immediately. |

## 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.

## System Stability With High Availability Control Workstation

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 may not notice the failover. Applications relying on data from the SDR will be momentarily interrupted. In an environment where the resource manager allocates resources, an outage of hardware or software on the control workstation causes the entire system to stop running after a few minutes. The resource manager and its backup fail when they lose connection to the SDR. New parallel jobs cannot be started and existing parallel jobs cannot be controlled; they must either run to completion or be manually killed. Having a backup control workstation available prevents this problem.

Occasionally, you may 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 Reliability Options for Control Workstations

Some configuration options that can make your control workstation more available are not part of the High Availability Control Workstation product. They include disk mirroring, uninterruptible power supplies, and dual disk controllers both internal and external.

## Uninterruptable Power Supply (UPS)

A UPS can supply electricity to a device to keep it running when main power is interrupted or is unreliable. Usually a UPS 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 a UPS in conjunction with all other means of assuring control workstation reliability. See *IBM RS/6000 System Overview and Planning* 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 requires twice the number of disks in the root volume group. See “Mirroring rootvg for Maximum Operating System Availability” in *AIX System Management Guide: Operating System and Devices*. That book describes how to create and manage mirrored rootvg 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 the spare Ethernet adapter still counts as one of the stations in the 30 total stations you may have on an Ethernet LAN segment.

## Completing Planning Worksheets for High Availability Control Workstation

You'll need to complete the following worksheets in the High Availability Cluster Multi-Processing for RS/6000 documentation:

- Shared Volume Group/File System Worksheet (Non-Concurrent)
- Defining Shared LVM Components for Non-Concurrent Access

As you complete the High Availability Cluster Multi-Processing for RS/6000 planning and installation steps, take the Non-Concurrent option whenever you are given the choice.

See *IBM High Availability Cluster Multi-Processing for AIX: Planning Guide*, SC23–1939–02 for complete planning information and references to other HACMP publications.

## Limits and Restrictions

The High Availability Control Workstation has the following limitations and restrictions:

- You cannot split the load across a primary and backup control workstation. Either the primary or the backup provides all the function at one time.
- The primary and backup control workstations must each be a RS/6000. You cannot use a node at your SP as a backup control workstation.
- The backup control workstation cannot be used as the control workstation for another SP system.
- The backup control workstation cannot be a shared backup of two primary control workstations.

  There is a one-to-one relationship of primary to backup control workstations; a single primary and backup control workstation combination can be used to control only one SP system.
- The Resource Manager is *not* restarted when control is transferred to a backup control workstation. The operator must decide whether to kill the currently running parallel jobs (suspend will not work) and restart the Resource Manager, or to wait until all the running jobs complete and then restart the Resource Manager. New parallel jobs cannot start until the resource manager is restarted.
- If your primary control workstation is a PSSP authentication server, the backup control workstation must be a secondary authentication server.

## Frame Supervisor Changes

Check with your IBM representative for information about ordering the necessary hardware.

## Software Requirements for HACWS Control Workstation Configurations

The software requirements for the control workstation include:

- Two licenses for AIX server
- Two licenses for C or C Set ++ for AIX.

  If the C for AIX 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 C for AIX license.
- Two licenses and software sets for 5765–A86 High Availability Cluster Multi-Processing (HACMP).

  This is the high availability feature of HACMP (F/C 3081). Both the client and server option of feature #3081 must be installed on both control workstations. You must purchase two licenses.
- 5765-529 PSSP 2.4 Priced feature #3936 HACWS

  This is the customization software that is required for HACMP support of the control workstation. This is one license per SP system. A copy is installed on both control workstations.

## Required High Availability Control Workstation Components

Once you decide that the High Availability Control Workstation is right for your installation, you must order the following components:

- An AIX server license for each control workstation.

  AIX 4.2.1 (or later) or AIX 4.3 is required. Also, the AIX version you are using must be supported on PSSP 2.4. Refer to the "Memo to Users" for the latest information on what levels of AIX are supported with PSSP 2.4.
- The High Availability Control Workstation feature with its cables, hardware, and software. If you are installing a new SP, the High Availability Control Workstation software may be on the installation media, but you require the license to use it.
- Two licenses for HACMP/6000; 5050 High Availability feature. (You do not need the 5051 feature.)

  HACMP 4.2 or greater is required. Also, the HACMP version you are using must be supported on the level of AIX that you are using. Refer to the appropriate HACMP documentation to determine what levels of HACMP are supported with the level of AIX that you are using or considering.

Planning and using the backup control workstation 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. 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 High Availability Control Workstation 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 hostname 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 hostname. 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 17 on page 96.

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.

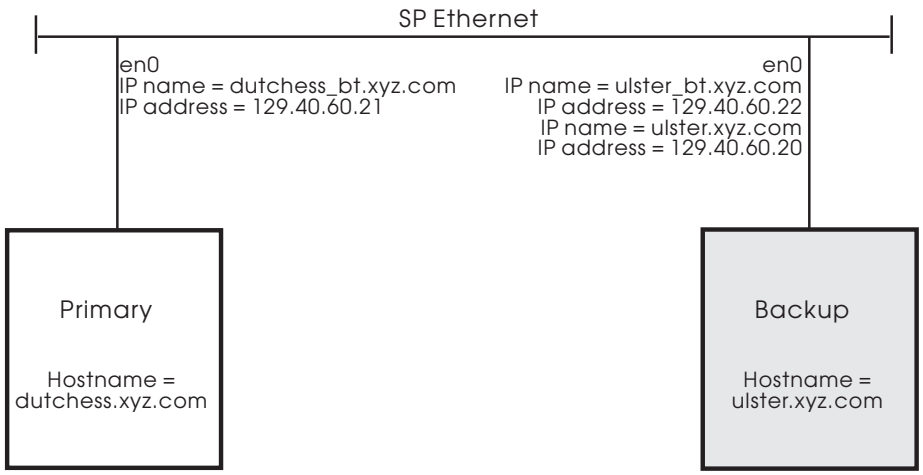



*Figure 17. Initial control workstation Network Configuration*

**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 18 on page 97.

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 19 on page 98.

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.

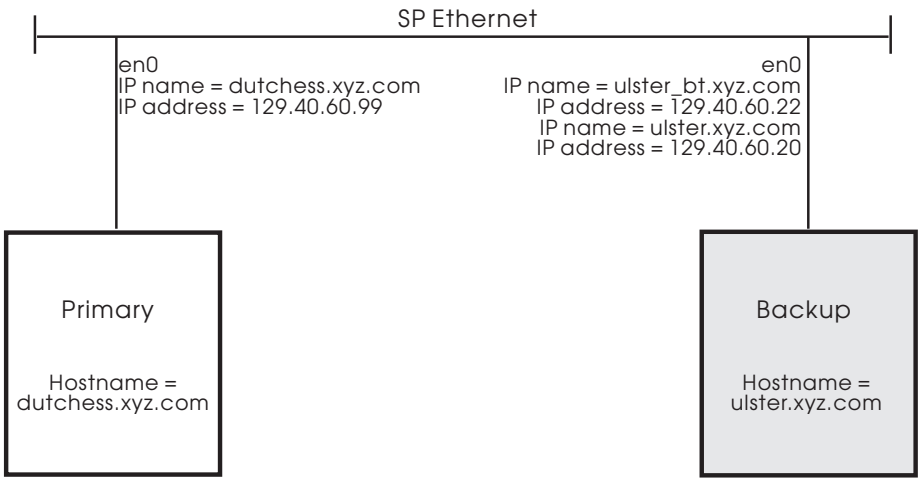



*Figure 18. Starting HACMP*

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 aliasing 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 20 on page 98 when the backup control workstation is active.

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 hostname 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 hostname of the backup control workstation (*ulster.xyz.com*) is always a valid network interface on the backup control workstation.

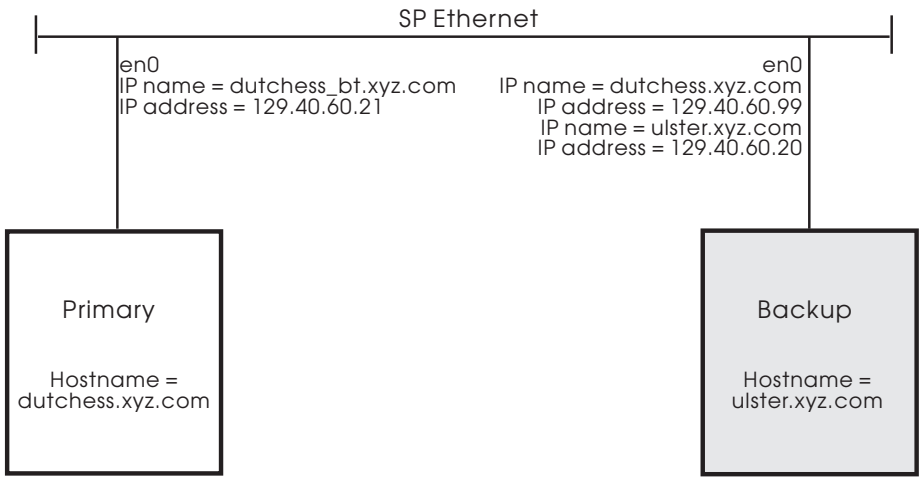



*Figure 19. control workstation Failover*

5. If your site uses a name server, make sure that all of these network interfaces have been added to your name server.

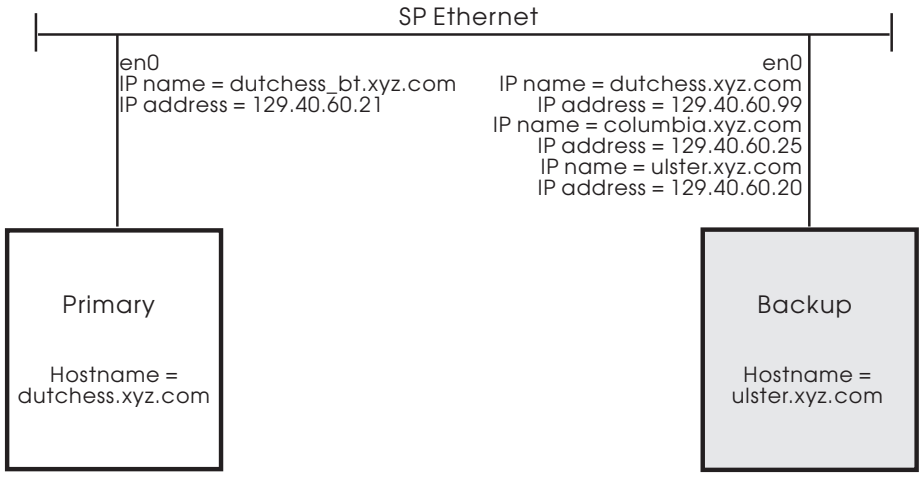



*Figure 20. Adding a System Partition*

# Chapter 5. Planning SP System Partitions

This chapter describes how to plan 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 which better suit your needs. System partitioning can apply to any system, whether it contains a switch or not, except for the HiPS-LC8 switch.

For more specific information on how to partition your system, see *IBM Parallel System Support Programs for AIX: 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.

A system partition is, at the most elementary level, a group of nodes (not including the control workstation). In essence, a system partition is an SP Subsystem which consists of sufficient pieces (nodes, control workstation, data, commands, and so on) of the SP system to form a logical SP.

With system partitions, you may ensure that switch applications running on one group of nodes are not inadvertently affected by activity on other nodes in the system.

Dependent nodes should be considered the same as standard nodes when planning a system partition.

System partitioning affects communication which occurs over the switch only; other communication paths are unaffected. System partitioning also provides environmental controls that allow the system administrator to control and monitor only the current system partition.

## How Do You Partition the System?

Your SP System has a particular configuration defined by its frames and nodes. The SP comes with 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, the SP comes with the System Partitioning Aid software that allows you to construct your own layouts. If none of the predefined layouts meets your system partitioning needs, you may define your own using the System Partitioning Aid or you may submit a Request for Price Quote (RPQ) to IBM to request additional layouts. See your IBM representative for more information on the RPQ process.

© Copyright IBM Corp. 1997, 1998

## Default System Partition

Taking advantage of system partitioning 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 intentionally 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 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 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, LPPs, 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. These types of partitions are usually relatively static and long-lived entities.

### 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 switch. System partitions let you isolate switch traffic in one system partition from the switch traffic in other system partitions.

## Example 1 -The basic 16-node system

Figure 21 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 may 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.

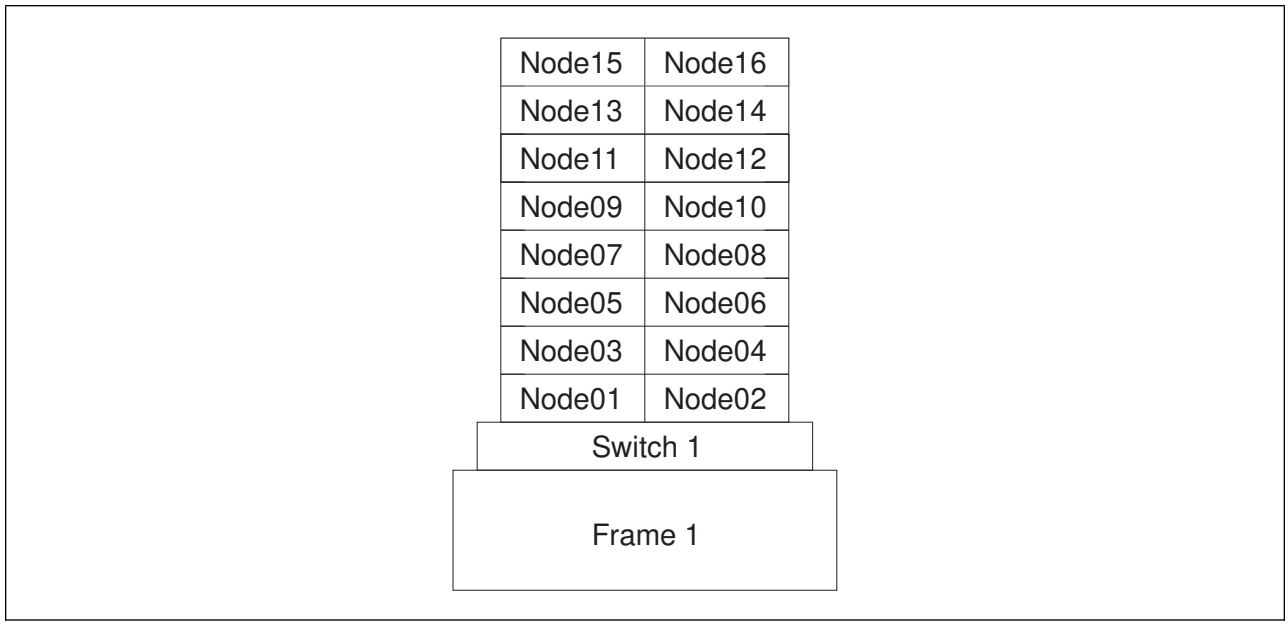



*Figure 21. Simple 1-frame system*

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 wish 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 would like to set 4 nodes aside to try out Version 2. This means you wish 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 22 on page 102.

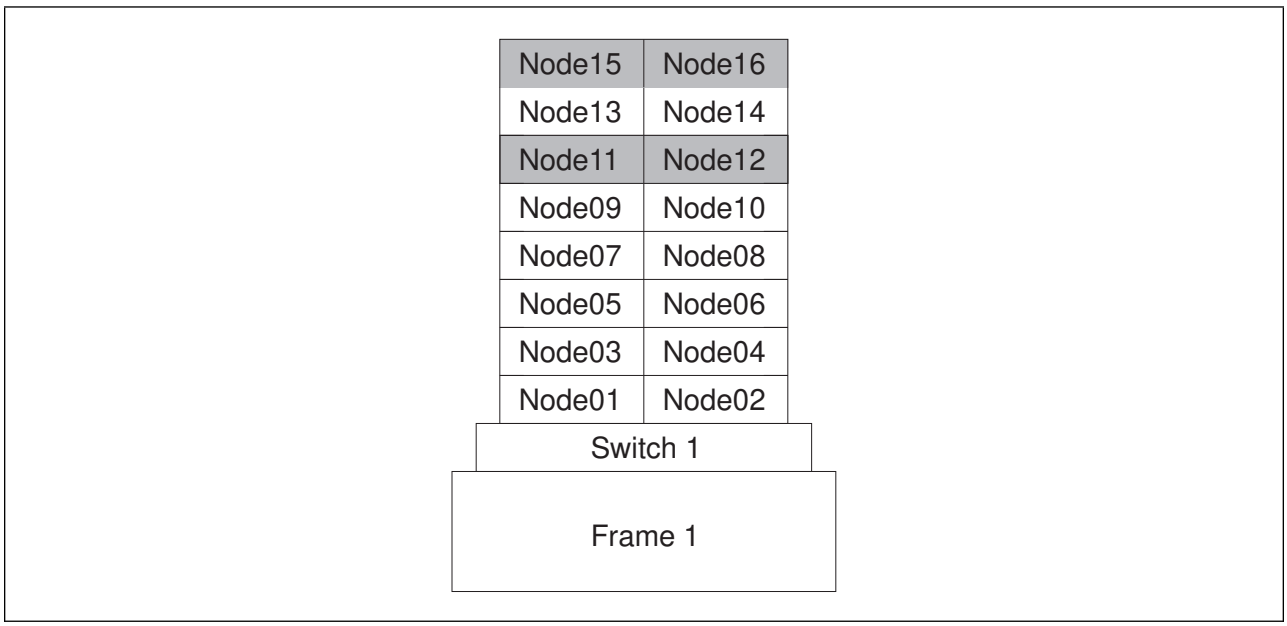



*Figure 22. Partitioned 1-frame system*

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 a Switch in a Partition

The SP supports the following switches:

- Scalable POWERparallel Switch (SP Switch)
- SP Switch-8
- High Performance Switch
- High Performance Switch-LC8 (cannot be partitioned)

## The Physical Makeup of a Switch Board

Actually, your choice in Example 1 was not necessarily as simple as suggested. A full *switch board* (whether the High Performance Switch or SP Switch) consists of 8 *switch chips* as shown in Figure 23 on page 103. Each chip has 8 *ports* to which nodes and other switch chips may connect.

Precisely 4 of the switch chips may have nodes connected to them, as on the left side of the board in Figure 23 on page 103. 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 may 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 may communicate with the nodes on switch chip 7 via any of chips 0-3.

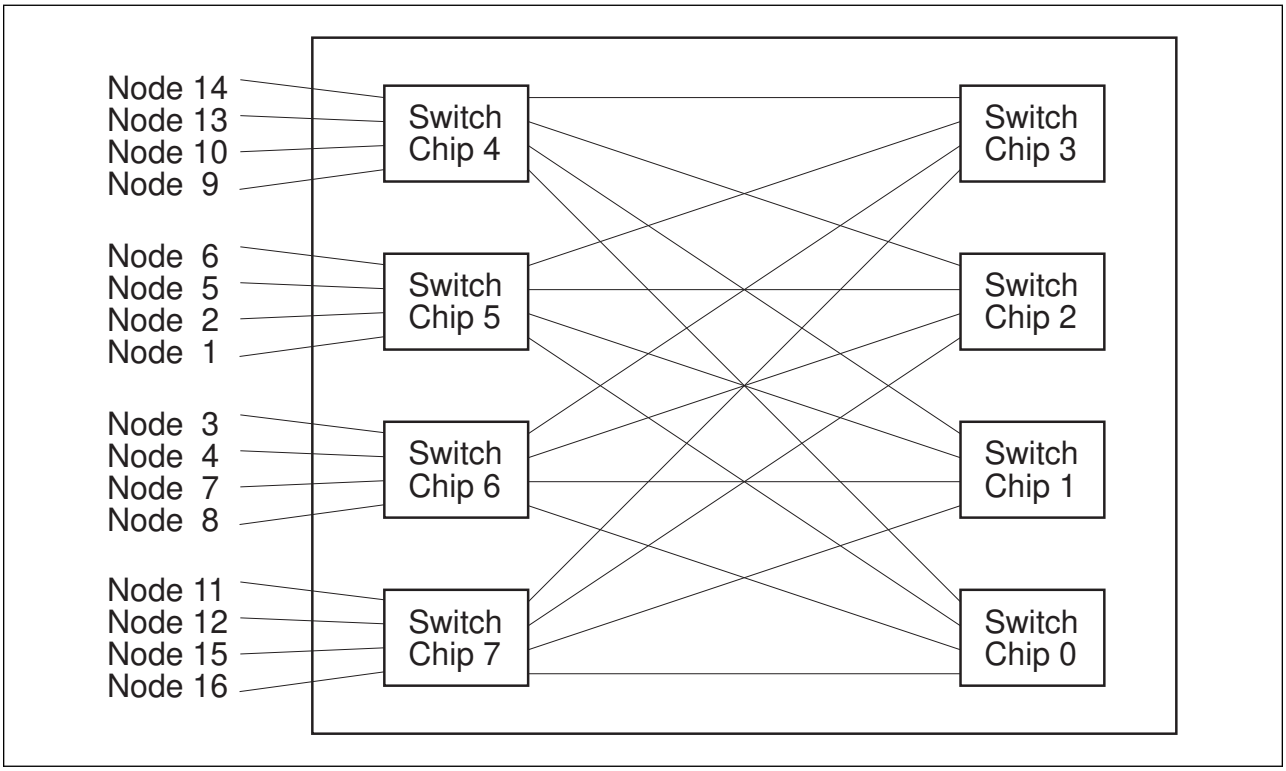



*Figure 23. Full switch board*

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 24 on page 104. Notice that the links from Chip 7 are missing in the diagram, indicating they have been logically removed from the active configuration, or disabled.

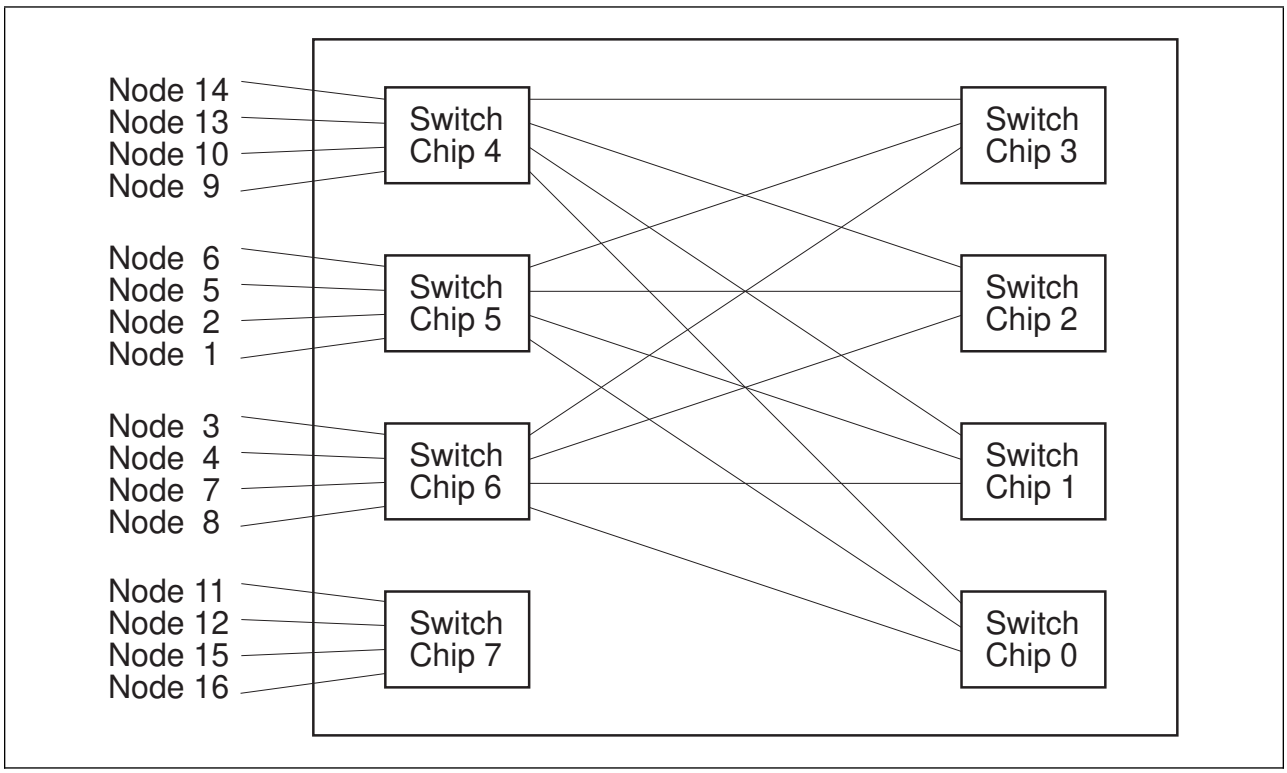



*Figure 24. Nodes 11,12,15,16 partitioned off*

## Systems With a Low Cost Switch

The SP Switch-8 is the only *low cost* switch available for new systems however the High Performance LC-8 switch is available for existing systems. If your system contains the 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.

If your system has a High Performance Switch-8, all of your nodes are connected to a single switch chip, and therefore, only a single partition is possible.

## 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 add a switch, you may have to rethink your system partition choice. In fact you may want to reinstall ssp.top 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 a switch.

Such a layout may be unsatisfactory, however, for your switchless environment, in which case you may use the System Partitioning Aid to build your own layout.

## Example 2 - A Switchless System

Figure 25 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.

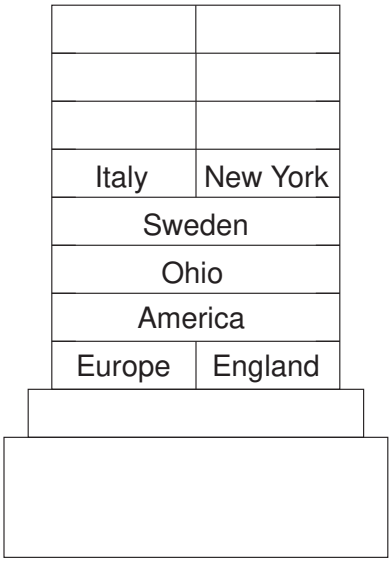



*Figure 25. One sparse frame with no switch*

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 may 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 an "experienced partitioner" 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 allow a High Performance Switch system to be joined with an SP Switch system to form a single system; nor does it support the joining of switched and switchless systems.

Details on the System Partitioning Aid appear in *IBM Parallel System Support Programs for AIX: Administration Guide*and in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*. This chapter provides examples to help you understand its value.

## 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 PSSP Virtual Shared Disks and pseudo-tape devices across system partitions is not supported.
- Twin-tailed 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 filesystems across partition boundaries, just as you would use a distributed file system from one SP to another. Keep in mind that doing this may affect nodes in both partitions in terms of both CPU 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. 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 server 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 may warrant a physical change.

Ethernet interference, causing slower performance, may 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 may 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 may 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 subsystem — essentially a smaller SP carved out of the real one. You cannot expect the smaller SP to do what the larger cannot. However, beginning with PSSP 2.2, more flexibility was introduced to support migration and make it easier to upgrade production applications. The added flexibility comes from coexistence support which has been added.

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 10, “Planning for Migration” on page 157.

## 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 un-partitioned SP is treated as a single system partition.
- The number of system partitions you can define is 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 having a switch has a primary node for switch initialization. If the switch is an SP Switch, then each system partition also has 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 defined by the corresponding topology file.
- Switch operations and message traffic are managed within a system partition.
- Jobs controlled by the Resource Manager are contained within a system partition and each system has both a primary and backup Resource Manager. The Resource Manager pools cannot cross system partition boundaries. Only one system partition can be part of any one LoadLeveler cluster when the Resource Manager is used by LoadLeveler. Also, a LoadLeveler class must be within a single system partition when the Resource Manager is used by LoadLeveler.
- The PSSP Virtual Shared Disk support and the pseudo-tape device driver cannot cross system partition boundaries. The IBM Recoverable Virtual Shared Disks and twin-tailed 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, hb, 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 *Administration Guide*.
- HACMP clusters do not span system partition boundaries.

## System Partitioning for Systems with Multiple Node Types

There are three physical node images supported in the SP: thin, wide, and high. These physical images affect the membership possibilities of system partitions. To understand how you can run multiple node types within a partition, you need to understand node slots. A *node slot* is the space that one thin node can occupy.

### Thin Node Frames

There are 16 node slots in an SP frame. Figure 26 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.

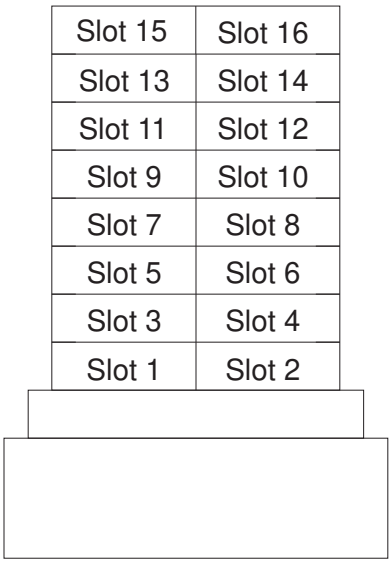



*Figure 26. One frame with slots numbered*

### 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 27 on page 110 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 26:

- 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

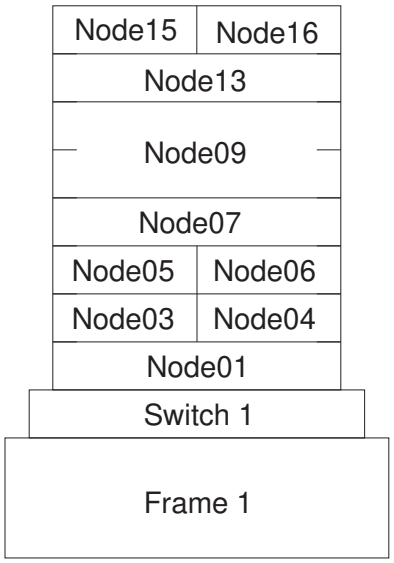



*Figure 27. Varied Node, 1-frame System*

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 comprised physically 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.

## Example 3 - An SP with 3 frames, 2 switches, and various node sizes

**Note:**

*Please keep in mind that you might not be able to order the system discussed in this example. This system has nodes located in legitimate locations. However, 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 do arrive at a similar system.*

Figure 28 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.

There is a switch in each of the first and third frames. The second frame is sharing the first frame's switch, which is possible because the configuration matches number 1 in Figure 11 on page 81. 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.

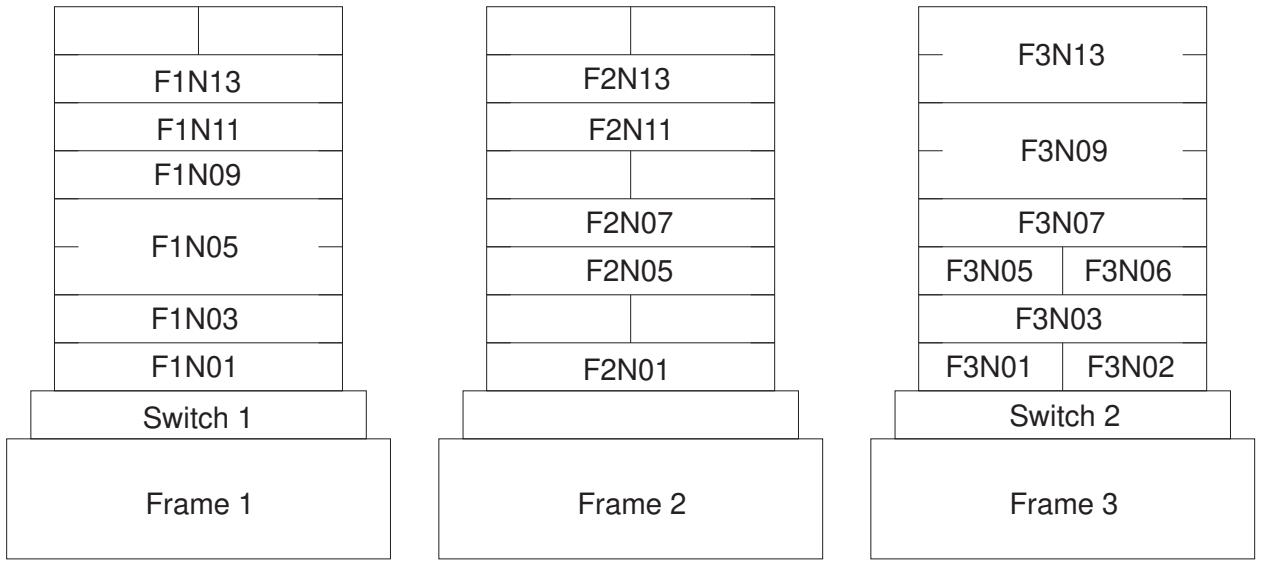



*Figure 28. Three frames with 2 switches*

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 a second slot

   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 three additional 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 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 28 on page 111, 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 unused 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 wished 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: *IBM Parallel System Support Programs for AIX: Command and Technical Reference*, GC23-3900
- Bring up the GUI version of the System Partitioning Aid, and select the nodes for each partition using a pointer device.

  **Note:** The GUI version of the System Partitioning Aid is available only under the *SP Perspectives* as of PSSP 2.2. Also, one can actually plan for such a system before it is realized. This topic is discussed in Appendix A, “The System Partitioning Aid - A Brief Tutorial” on page 179.

In either case, 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 IBM Parallel System Support Programs for AIX: 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's ssp.top option. 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 29 on page 115. An introduction to the System Partitioning Aid is provided in Appendix A, “The System Partitioning Aid - A Brief Tutorial” on page 179.

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 wish 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* or *NSB*. In larger systems, it becomes impossible to adequately connect all pairs of NSB switch boards 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* or *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 29 on page 115 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 switches, and various node sizes” on page 111 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 we specified to the System Partitioning Aid.

   “Example 2 - A Switchless System” on page 105 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.

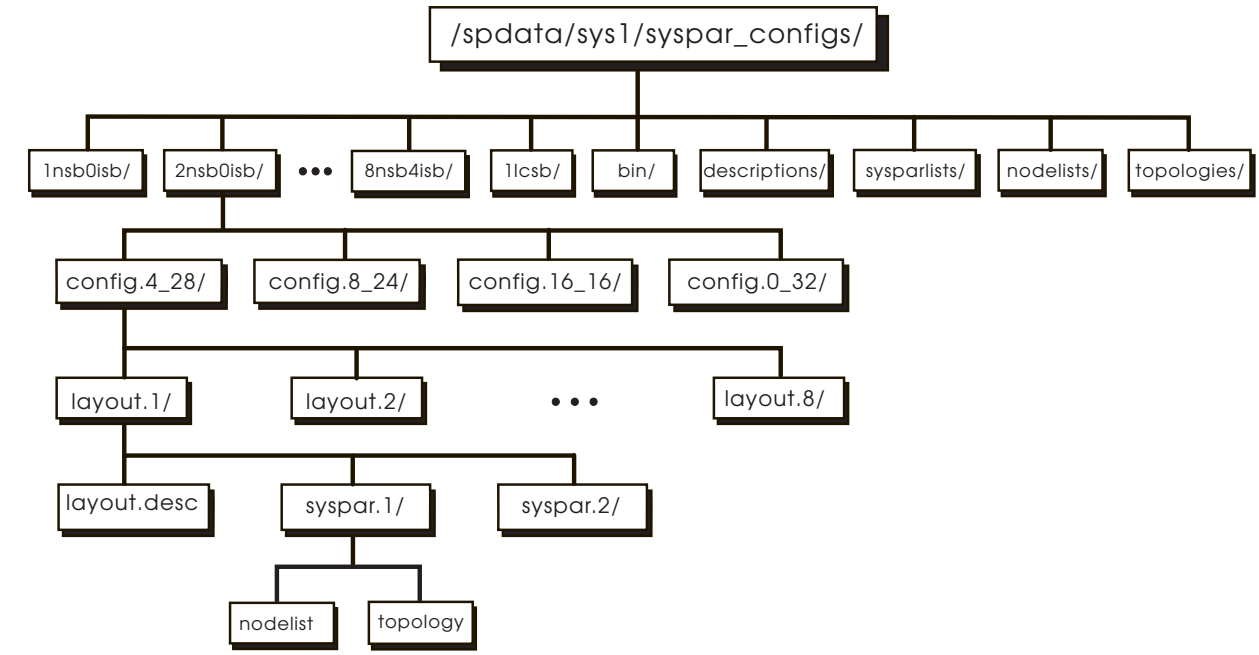



*Figure 29. The Directory Structure Of System Partition Information*

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 203. Consult that information as you complete the worksheets in Appendix C, “SP System Planning Worksheets” on page 227.

# Chapter 6. Planning for Security

Analyzing company resources and protecting these resources is part of the administrator's duties. Your resources with regard to PSSP include the following:

- System and application data, user programs, and user data
- Communications devices and communication access methods
- Login permissions. Specifically, who can log in, when, and how much resource each user can have.
- Authentication, passwords, and Kerberos tickets.

This chapter describes how authentication services work in the SP so that you can plan your authentication use.

## Authentication in the SP System

If you are unfamiliar with Kerberos terminology, you might want to review the definitions of terms used to describe the security facilities in the *Administration Guide*.

The SP authentication services are based on Version 4 of MIT's Kerberos security facility. It is a *de facto* standard authentication protocol that has several implementations, including use by AFS. Authentication allows networked applications (applications that have client and server parts executing on different hosts) to securely determine their mutual identities through a trusted third party. This includes the Kerberos authentication and key distribution center (KDC). This service is provided by one or more Authentication servers (daemons) running on systems that are accessible from application client server systems, providing credentials that they use to perform the authentication task.

When there is more than one authentication server system, one is the primary server and all others are secondary servers. The primary server maintains the master copy of the authentication database. Secondary servers operate using copies of the master database, which are updated periodically at the discretion of the system administrator. The authentication database contains an encrypted key (password) for each client and server entity, known in Kerberos terminology as a *principal*.

## Authentication Principals

The identities that the authentication service verifies are registered in the authentication database. There are basically two kinds of principals: users, who assume the role of clients when they invoke authenticated services; and the services that these users invoke. SP authentication principals are defined across a collection of systems known collectively as the local authentication realm. Each realm has a primary authentication server, any number of secondary servers, and client systems (systems that have PSSP authenticated services installed but are not authentication servers).

Each principal's full identifier is in the form: *name.instance@realm*

where:

© Copyright IBM Corp. 1997, 1998

*name*
is a user or service name

*instance*
is a qualifier used to identify multiple principals with the same name. For users, this allows the assignment of different roles and privileges to each instance. For services, instance is used to distinguish between identical application servers running on multiple hosts. For PSSP authenticated services, it is derived from the hostname used to access a particular instance of the server daemon.

*realm*
is the name of the realm in which the principal is defined. The entire SP system, including its control workstation is always in a single realm.

Ordinarily, a user principal is known by a simple name that is the same as the AIX login name. It is not required to be the same, however, because principal names are managed separately and logging into the SP system as an AIX user is independent of the procedure for establishing one's identity as an authentication principal. User principals are defined by authentication administrators; the first is defined when the primary authentication server is configured. Administrators of the PSSP authentication database have principal identities of <name>.admin. AFS authentication administration does not use instances in its principal identifiers.

Examples of user principals on an SP system are:

```
frank
root.admin
operator
melissa.admin
kelly.sysprog
```

## SP Service Principals

Because authenticated application services are provided by separate servers (daemons) on individual workstations or nodes, each instance of a daemon must be distinguished from others like it running on different hosts. Client programs identify the target of a request by a hostname that represents a particular network interface on the server host. The hostname is supplied explicitly by the client interface, such as on the rsh command; or it is provided implicitly as with the SP_NAME environment variable used by the System Monitor interface. The service principal associated with a particular target hostname is identified in the form: *service.host*

where:

*service*
is the name of the service

*host*
is the short form of the fully-resolved hostname, converted to lower-case. For example, if a target node has three network interfaces:

1. The internal ethernet with hostname SP12, which resolves to Node12
2. A token ring with hostname tr12.abc.com
3. An SP Switch with hostname sw12, which resolves to Switch12.abc.com

PSSP authenticated services will define three service principals for this node named rcmd.node12, rcmd.tr12, and rcmd.switch12, that it will use to perform authentication on client requests arriving through the respective interfaces.

The SP system includes versions of the **/usr/lpp/ssp/rcmd/bin/rsh** and **/usr/lpp/ssp/rcmd/bin/rcp** commands that use SP authentication services. These commands are used by SP installation and configuration scripts and are available for general use by system administrators. The SP System Monitor command-line and graphical interfaces, and the remote execution facilities of **dsh** and **sysctl** also use authentication services. This suite of administration tools uses two service principals: the System Monitor applications use the **hardmon** principal, and the other facilities use the **rcmd** principal. Instances of **hardmon** are defined on the SP control workstation and on other RS/6000 workstations on which the SP authentication services are initialized. Instances of the **rcmd** service principal are defined on the same systems as the **hardmon** principal, plus each SP node.

## Administrative Principals

Initializing authentication services requires defining at least one user principal who is 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 MIT Kerberos Version 4 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 Your Authentication Configuration

This section describes the various ways you can configure an SP system in an authentication realm. The following sections illustrate the possible authentication configurations used with the control workstation. The configurations also include other 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 SP nodes are always in a single authentication realm, which may optionally include other workstations and even other SP systems. The authentication servers can be on any workstation in the realm, but not on SP 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 MIT Kerberos Version 4 servers. The SP nodes will have authentication services installed for all authentication configurations.

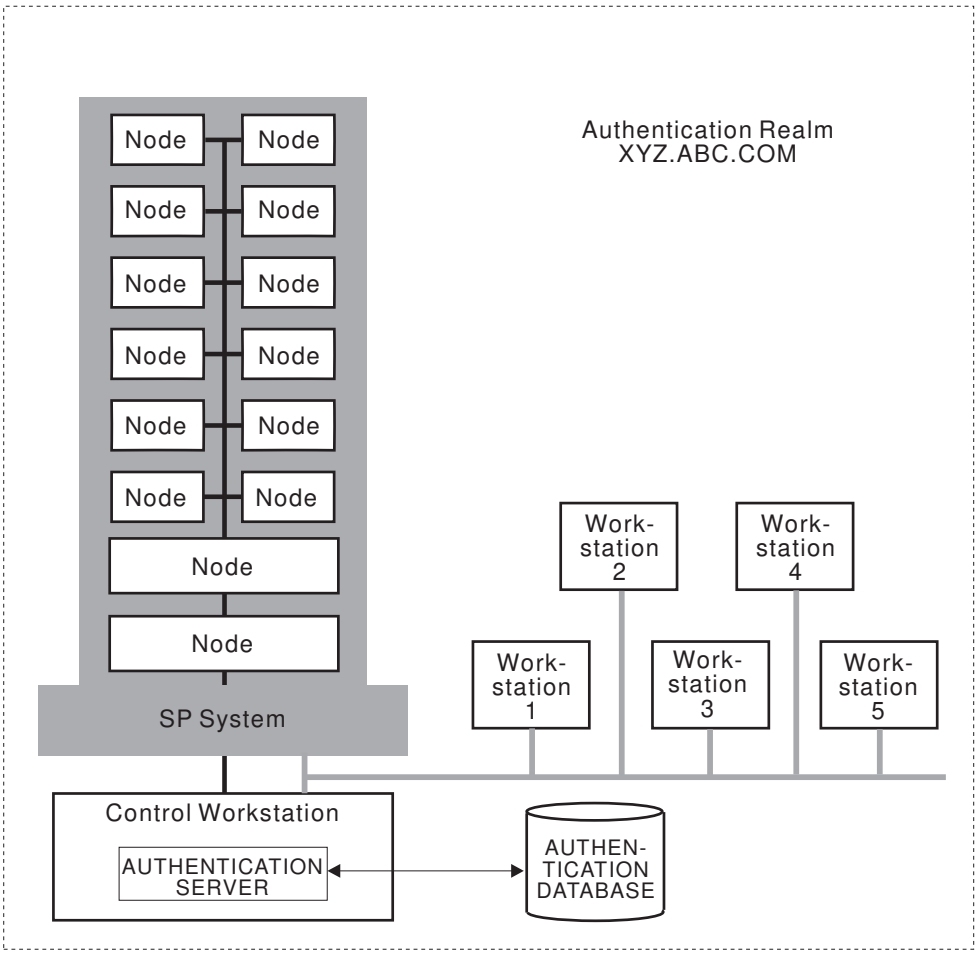



*Figure 30. The control workstation as primary authentication server*

**CWS-Primary**

Figure 30 illustrates this configuration as follows:

- The control workstation is the primary authentication server, with the SP authentication server (fileset ssp.authent) and authenticated services (fileset ssp.client) installed.
- Other RS/6000 workstations can be secondary authentication servers, with the SP authentication server installed.
- Other RS/6000 workstations can have SP authenticated services installed.

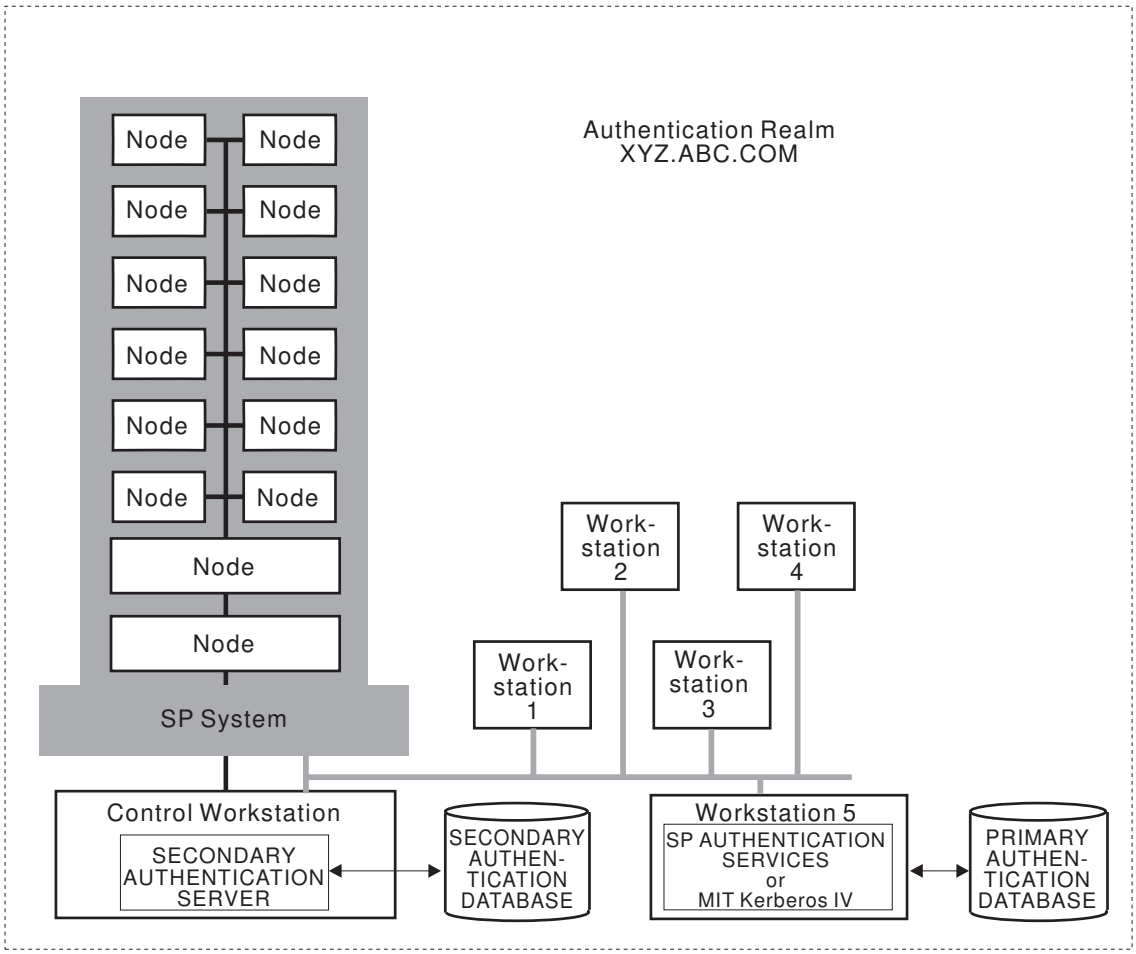



*Figure 31. The control workstation as secondary authentication server*

**CWS-Secondary**

Figure 31 illustrates this configuration as follows:

- The primary authentication server is one of the following:
  - An RS/6000 workstation with the SP authentication server (fileset ssp.authent) and authenticated services (fileset ssp.clients) installed
  - A workstation with another MIT Kerberos Version 4 implementation
- The control workstation is a secondary authentication server, with the SP authentication server and authenticated services installed.
- Other RS/6000 workstations can be secondary authentication servers, with the SP authentication server installed.
- Other RS/6000 workstations can have SP authenticated services installed.

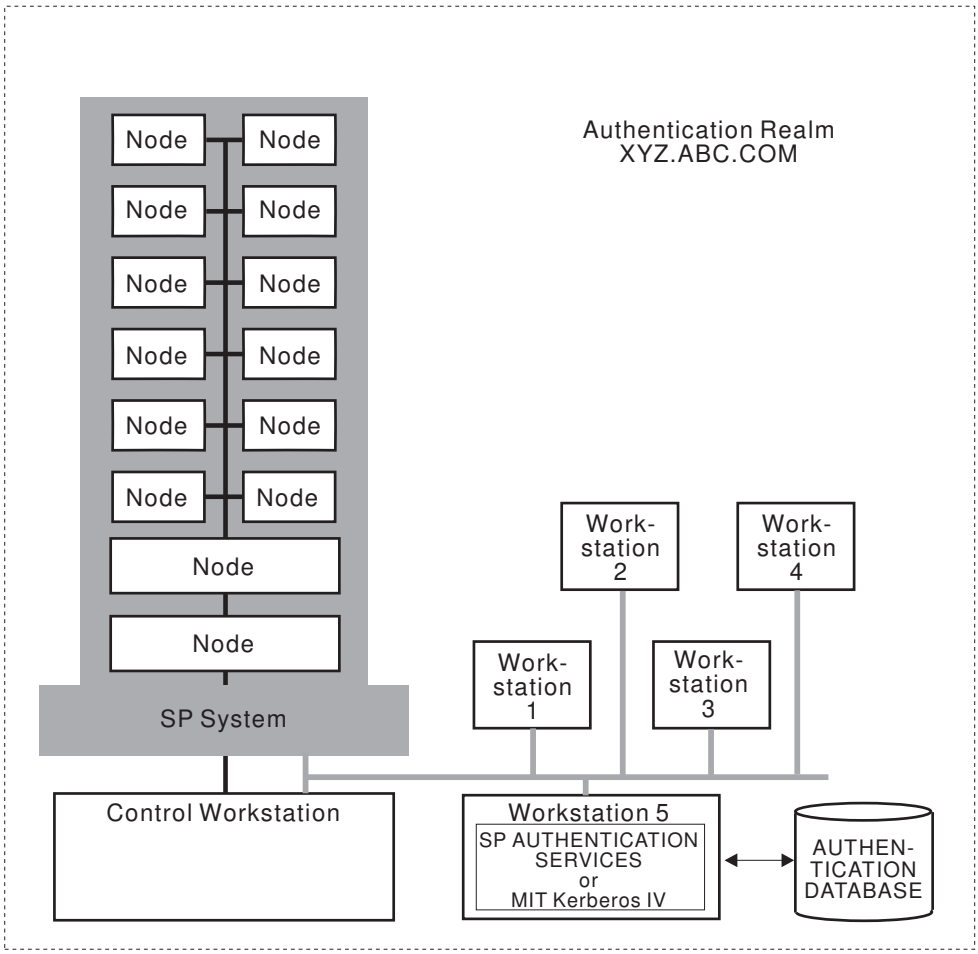



*Figure 32. The control workstation as client of authentication server*

**CWS-Client**

Figure 32 illustrates this configuration as follows:

- The primary authentication server is one of the following:
  - An RS/6000 workstation with the SP authentication server (fileset ssp.authent) and authentication services (fileset ssp.clients) installed
  - A workstation with another MIT Kerberos Version 4 implementation
- Other RS/6000 workstations can be secondary authentication servers, with the SP authentication server installed.
- The control workstation has SP authenticated services installed.
- Other RS/6000 workstations can have SP authenticated services installed.

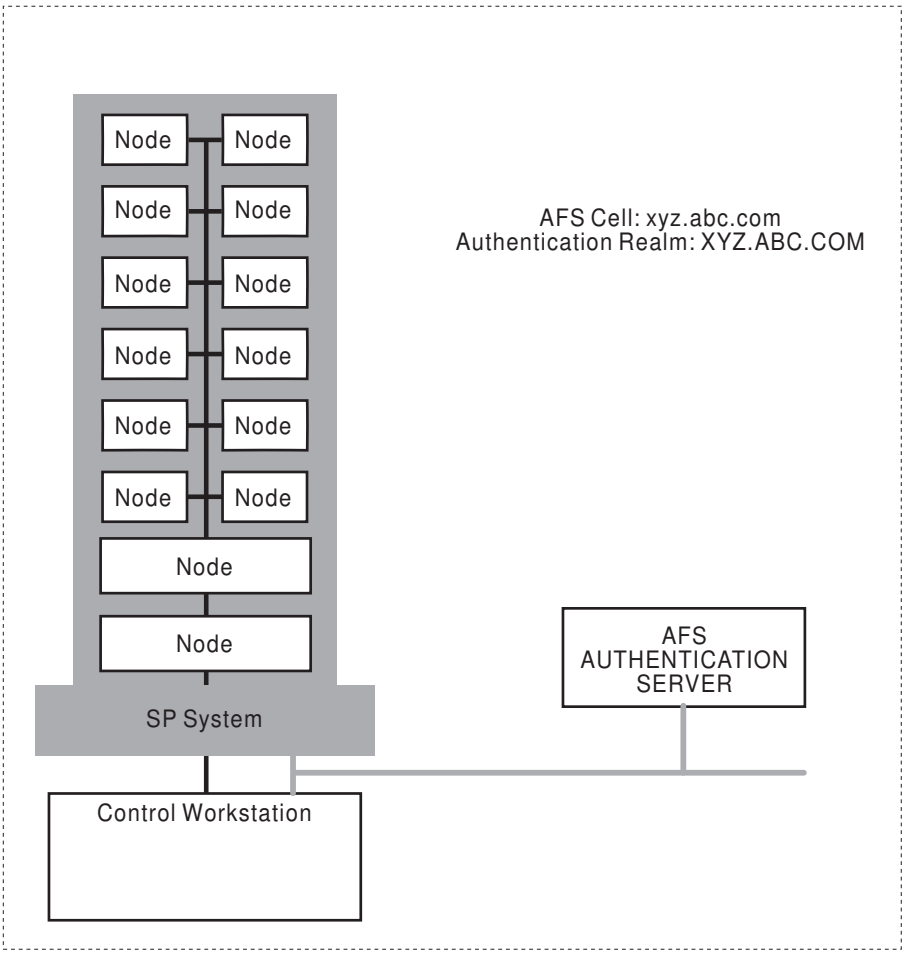



*Figure 33. Using AFS authentication services on the SP system*

**AFS Server**

Figure 33 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 (fileset ssp.clients) installed. installed.
- Other RS/6000 workstations can have SP authenticated services installed.

## 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 SP system 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 SP system must be installed in a single authentication realm. This is the case if you are installing SP 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 SP system. You can give the realm any

name you like or default the authentication realm name to the domain part of the server's hostname, converted to upper case.

Whenever you have additional SP systems 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 hostnames 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 *IBM Parallel System Support Programs for AIX: Administration Guide* for more information.

## Creating the Authentication 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 hostnames. 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 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**.

Refer to the file format man pages **krb.conf** in *Commands and Technical Reference* for more information and examples.

## Selecting the Authentication Options to Install

Selecting SP options for installation depends on where the authentication server is located. SP authentication code is distributed in two separately installable options. **ssp.authent** contains only parts required on a system that is to be a authentication server. The remainder of Kerberos and authenticated services is distributed in **ssp.clients**.

You must install **ssp.clients** on the SP control workstation, even in the case where you intend to use AFS or MIT Kerberos Version 4 authentication. You should also install it on any other 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 MIT Kerberos Version 4 authentication do not require **ssp.clients** if they are not using SP system management tools. All SP nodes will have **ssp.clients** installed.

You must install **ssp.authent** on the control workstation, if it is to be an authentication server, either primary or secondary. You can also install **ssp.authent** on any other RS/6000 workstation that you plan to be an authentication server. You will not be able to install it if the system already has an MIT Kerberos Version 4 implementation installed. If you want to install the SP authentication facilities, you must first remove the other Kerberos implementation.

## Checklists for Authentication Planning

### Using SP or MIT Kerberos 4 Authentication Servers

Decide what authentication realms your network will have.

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 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 RS/6000 systems that will be clients.

   If any other 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 Servers

If you choose to use AFS authentication services with your SP system, take into account the following unique considerations:

1. Any 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 *IBM Parallel System Support Programs for AIX: Administration Guide*
5. The authentication server (**kaserver**) in AFS Version 3.4 for AIX 4.1 accepts Kerberos Version 4 protocol requests using the well-defined **udp** port assigned to the **kerberos** service. AIX 4.1 assigns the Kerberos Version 5 port number 88 to work with DCE. PSSP authentication services based on MIT Kerberos Version 4, uses a default port number of 750. The PSSP commands use the service name **kerberos4** to avoid this conflict with the Kerberos 5 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

Copy and complete Worksheet 17 Table 57 on page 243 with your authentication information. If you use PSSP authentication servers, fill out Table 58 on page 244. If you use an AFS authentication server, fill out Table 59 on page 244.

Be aware that these worksheets ask you for passwords. Keep these worksheets, when filled out, in a secure location.

# Chapter 7. Planning to Record and Diagnose System Problems

## 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 SP software is utilizing the AIX Error Log frequently. Therefore, you should set the size to at least 4MB. You can do this once for all nodes with the **dsh** command after the nodes are installed.

```
dsh -a /usr/lib/errdemon -s 4096000
```

## Configuring the BSD Syslog

### Control Workstation

The SP installation configures the Berkeley Software Distribution (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 the previous configuration.

### SP Nodes

The BSD syslog facility is not configured on the SP 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 if you want entries made there. Note also that any SP 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 code that was ported from 'other sources' contains the calls to syslog and logger. The AIX kernel does not use syslog for error logging.

The File Collections facilities can be used to manage the /**etc**/**syslog.conf** file if all the nodes have the same configuration file. Administrators should be aware that the amount of information that syslog collects may consume network resources on an SP system if they are forwarded to a single node. Additionally, the SP 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 recommends that they be logged on a per-node basis, and that you use tools such as **dsh** and **sysctl** to view and manage them.

See the *Diagnosis and Messages Guide* for more information about error logging.

© Copyright IBM Corp. 1997, 1998

## System Error Logs

The following error logs are valid for PSSP 2.4.

## SP Specific Logs

SP creates the following logs during normal operations.

**/var/adm/SPlogs/cs/cstart**.*timestamp*.**pid**
**/var/adm/SPlogs/cs/cshut**.*timestamp*.**pid**
**/var/adm/SPlogs/sysman/***node***console.log**
**/var/adm/SPlogs/sysman/***node***config.log**
**/var/adm/SPlogs/sysman/***node***firstboot.log**
**/var/adm/SPlogs/spacs/spacs.log**
**/var/adm/SPlogs/csd/vsd.log**
**/var/adm/SPlogs/css/rc.switch.log**
**/var/adm/SPlogs/css/daemon.stderr**
**/var/adm/SPlogs/css/daemon.stdout**
**/var/adm/SPlogs/css/fs_daemon_print.file**
**/var/adm/SPlogs/css/worm.trace**
**/var/adm/SPlogs/css/flt**
**/var/adm/SPlogs/css/router.log**
**/var/adm/SPlogs/css/toplogy.data**
**/var/adm/SPlogs/css/out.top**
**/var/adm/SPlogs/css/dtbx.trace**
**/var/adm/SPlogs/spmon/nc/nc.***frameid.slotid*
**/var/adm/SPlogs/spmon/hmlogfile.***nnn*
**/var/adm/SPlogs/spmon/hm_frame_packet_dump**
**/var/adm/SPlogs/spmon/hm_response_dump**
**/var/adm/SPlogs/spmon/hm_sldata_dump**
**/var/adm/SPlogs/spmon/splogd.state_changes.***timestamp*
**/var/adm/SPlogs/SPdaemon.log**
**/var/adm/SPlogs/sdr/sdrlog.***nnn*
**/var/adm/SPlogs/sdr/SDR_config.log**
**/var/adm/SPlogs/sysctl/sysctld.log**
**/var/adm/SPlogs/kerberos/admin_server.syslog**
**/var/adm/SPlogs/kerberos/kerberos.log**
**/var/adm/SPlogs/kerberos/kerberos.slave_log**
**/var/adm/SPlogs/kerberos/kpropd.log**
**/var/adm/SPlogs/pman/pmand.log**
**/var/adm/SPlogs/pman/pmand**.*partition name*.**log**
**/var/adm/SPlogs/pman/pmand.log**
**/var/adm/SPlogs/pman/pmand**.*partition name*.**log**
**/var/adm/SPlogs/auto/auto.log**
**/var/adm/SPlogs/filec/logs/***
**/var/adm/SPlogs/spmgr/***
**/var/tmp/sp.configd.log**
**/var/tmp/vsd.debuglog**
**/var/ha/log/hags***
**/var/ha/log/em***
**/var/ha/log/hats***
**/tmp/jmd_out**
**/tmp/jmd_err**

## Finding and Using Error Messages

Most error messages generated by the SP are listed and explained in the *Diagnosis and Messages Guide*. The second section of that 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 pre-defined path for resolving the problem.

## Getting Help from IBM

If you require help from IBM in resolving an SP system problem, you can get assistance by calling IBM Support at the following numbers. Before you call, be sure you have the following information:

1. Your access code (customer number). This number was entered on Worksheet 4, "SP Planning" in Table 44 on page 229.
2. The SP product number, for example:
   - For a problem with PSSP 2.4, use product number: 5765–529
   - For a problem with LoadLeveler 1.3, use product number: 5765–145

   Similarly, each product has it own order number that will speed the correct routing of your call.
3. The name and version of the operating system you are using.
4. A telephone number where you can be reached.

The person with whom you speak will ask for the above information and then give you a time period during which an IBM SP 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**.

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. Refer to *IBM Parallel System Support Programs for AIX: 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 RS/6000 Scalable POWERparallel Systems
Dept. 39KA, M/S P961, Bldg. 415
522 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/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 *IBM Parallel System Support Programs for AIX: 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.

## 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 Problem Management Report (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:

- A Hayes 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.

**Note:** Specific Feature Codes for Service Director hardware items are detailed in *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*.

## 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 8. Planning for PSSP-Related LPPs

This chapter briefly discusses planning information for PSSP-related Licensed Program Products (LPPs) with regard to PSSP. See Chapter 10, “Planning for Migration” on page 157 for versions supported, coexistence or migration information . For complete detailed information on the individual LPPs, see their books which are listed in “Related Publications” on page xvi.

## Planning for IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk

Virtual Shared Disk is an optional component of PSSP that lets application programs executing on different nodes of a cluster access 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 Virtual Shared Disk component of PSSP can:

- Support fencing
- Determine the size of an underlying Logical Volume on a server and provide that information to applications

Recoverable Virtual Shared Disk is an LPP that works in cooperation with the Virtual Shared Disk component of PSSP. It offers continuous access to data with transparent recovery in the event of the failure of an SP node, disk, disk adapter, disk cable or communication adapter. Recoverable Virtual Shared Disk also provides fencing support and recovery by volume group, which allows proactive recovery of applications using Virtual Shared Disk.

You should plan how you are going to use Virtual Shared Disk and Recoverable Virtual Shared Disk before you install the hardware. Each Virtual Shared Disk cluster must be in the same system partition. You can have separate Virtual Shared Disks in separate system partitions, but they cannot communicate directly with each other. See Chapter 5, “Planning SP System Partitions” on page 99 for system partition planning information.

IBM Recoverable Virtual Shared Disk requires twin-tailed disk storage which must be installed before you define or use the Recoverable Virtual Shared Disks. It also requires a minimum of two SP nodes.

## Planning for Virtual Shared Disk Communications

When you define Virtual Shared Disks, you specify the SP switch or other connection method. See the *IBM Parallel System Support Programs for AIX: Administration Guide* for detailed Virtual Shared Disk planning information.

In the design of Logical Volume Manager, each logical partition maps to one physical partition and each physical partition maps to a number of disk sectors. The design of Logical Volume Manager limits the number of physical partitions that Logical Volume Manager 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.

© Copyright IBM Corp. 1997, 1998

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 hdiskX.*.

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. The workaround 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 a pre-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 Logical Volume Manager in AIX might really consist of many non-4 GB disks. In this case, the 1016 requirement still exists. Logical Volume Manager is unaware 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 AIX 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 in AIX 3.2, the number of disks picked in the install menu of AIX is used as the reference number by mkvg -d during the creation of rootvg. For AIX 4.1, this -d number is 7 for one disk and one more for each additional disk picked. For example, if two disks are picked, the number is 8 and if three disks are picked, the number is 9, and so on. 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 recommends 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 filesystem (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 filesystem 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 filesystem 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.

## Planning for Parallel ESSL

Parallel ESSL is a scalable mathematical subroutine library that supports parallel processing applications on IBM RS/6000 SP and on clusters of IBM RS/6000 workstations. Parallel ESSL supports the Single Program Multiple Data programming model and provides subroutines in six major areas of mathematical computations. It is tuned for optimal performance on the SP consisting of POWER2 nodes with the High Performance Switch, High Performance Switch-LC8, SP Switch, or SP Switch-8.

Parallel ESSL provides subroutines in the following computational areas:

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

The subroutines run under the AIX operating system and can be called from application programs written in Fortran, C, C Set ++, and High Performance Fortran (HPF). On the SP, Parallel System Support Programs (PSSP) is also required.

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 either the High Performance Switch, High Performance Switch-LC8, SP Switch, or SP Switch-8. Communications using the Internet Protocol (IP) may use Ethernet, Token Ring, FDDI, High Performance Switch, High Performance Switch-LC8, SP Switch, or SP Switch-8.

To order IBM Parallel ESSL for AIX, specify program number 5765-C41.

## 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 IBM RS/6000 SP System or a networked cluster of RS/6000 processors. The main Parallel Environment components are:

- **The Communications Low-Level Application Programming Interface (LAPI)**, which is designed to provide optimal communication performance on the SP Switch.
- **Message Passing and Collective Communications Application Programming Interface (API) Subroutine Library**, which helps application developers parallelize their code.
- **Parallel Operating Environment**, which provides the ability to create and execute parallel application programs.
- **Parallel Debuggers**, which assist in debugging parallel applications.

- **Visualization Tool**, a trace generation and display system to visualize performance characteristics of your program and system.

You should 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 authenticated services (**ssp.clients**). See *IBM Parallel Environment for AIX: Installation Guide* for details on planning 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** fileset.
- Installing the switch services component of PSSP (**ssp.css**) after Parallel Operating Environment has been installed.

## Planning for Performance Toolbox Parallel Extension (PTPE)

Performance Toolbox Parallel Extension (PTPE) for AIX is a performance monitor for the SP system. PTPE is a priced feature of PSSP. When installed on your SP, it allows easy access to performance information about both SP hardware and software (LPPs). This information is available as both run-time (current) and archived (historical) data that you can analyze, manipulate, print, and import to a database, should you so desire.

PTPE requires Performance Toolbox for AIX (PTX), program number 5765–654. PTPE builds on the capabilities of PTX, adding monitoring functions specific to the SP system. You can use PTPE to examine the current performance state of any node in your SP system. You decide what performance information to display, and view or print it from any node in the system.

PTPE collects and archives performance statistics for each SP node. It calculates averages for common performance information for all SP nodes. All PTPE data is available for display to help you evaluate performance of the SP at both the node and system level.

The translation table consumes 13,200,012 bytes (roughly 12.6 MB). As a result, the filesystem containing the **/var/adm/ptpe** directory needs to have at least 13 MB available space when PTPE is installed and started for the first time. If this space is not available, PTPE will not start. Once PTPE has successfully started, the complete table space is reserved in the directory even if 50,000 statistics aren't available (so PTPE can assimilate new statistics if they become available at a later time).

Administrators should set up a logical volume, containing at least 4 LP's (16 MB) on each node where PTPE is to run. Create a new filesystem for the logical volume, and mount the filesystem over the **/var/adm/ptpe** directory. This will ensure that PTPE has enough DASD to start.

Administrators also need to become part of the **perfmon** user group, and thoughtfully lay out the monitoring hierarchy.

For information on PTPE, refer to *IBM Performance Toolbox Parallel Extensions for AIX Guide and Reference* SC23-3997.

## 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. Currently there are two variations of the product, HACMP and HACMP/ES. HACMP/ES is newer and has enhanced scalability capability in addition to the function provided by HACMP.

Typically, HACMP is run only on the control workstation if HACWS is being used. HACMP can also be run on the SP nodes. HACMP/ES does not run on the CWS, it only runs on the SP nodes.

HACMP/ES has a dependency on PSSP's Group Services which is only in PSSP 2.4 or later.

See *IBM High Availability Cluster Multi-Processing for AIX: Planning Guide*, SC23–1939–02 for planning information and references to other HACMP publications.

## Planning for LoadLeveler

LoadLeveler is an IBM software product that lets you build, submit, and process both serial and parallel jobs on your RS/6000 workstation, Silicon Graphics systems, Sun SPARCstation systems, and Hewlett-Packard systems. LoadLeveler is recommended, but not required, for batch processing on the SP system, and is included with the SP.

LoadLeveler provides workload management capabilities on the RS/6000 SP and other heterogeneous workstations running UNIX operating systems. (RS/6000, Sun, HP, and SGI workstations) It allows you to view individual nodes as a single computational resource.

LoadLeveler is an integral piece of the total System Management solution on the RS/6000 SP. LoadLeveler can take advantage of features provided in PSSP, such as event management and performance monitoring. LoadLeveler will 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).

## Compatibility

The current release of LoadLeveler for AIX is 1.3. It is available for AIX 4.1 and later. LoadLeveler 1.3 for AIX allows OEM platforms to participate in the LoadLeveler 1.3 cluster as submit only nodes at this time.

LoadLeveler 1.2.x and LoadLeveler 1.3 are not compatible. There have been changes to the protocol used between daemons and changes in the format of the job_queue. You must upgrade all LoadLeveler nodes to LoadLeveler 1.3 or maintain two separate LoadLeveler clusters, one for LoadLeveler 1.3 and one for LoadLeveler 1.2.x.

In order to migrate from LoadLeveler 1.2.x to LoadLeveler 1.3, you must drain the startd and schedd daemons in the cluster, shutdown LoadLeveler, install

LoadLeveler 1.3, convert the job_queue file, and then restart LoadLeveler. For detailed instructions, refer the README file distributed with LoadLeveler 1.3.

## 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 nameserver running Interactive Network Dispatcher in the interactive configuration. Additionally, the availability of individual compute nodes and filesystems 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, refer to the ITSO Redbook *Implementing High Availability on the RS/6000 SP*, SG24-4742-00.

## Performance Monitoring of LoadLeveler

LoadLeveler can utilize the Performance Toolbox Parallel Extensions (PTPE) optional feature of PSSP 2.2 (or later) to collect performance information on scheduling and executing nodes.

In general, planning the LoadLeveler installation for batch processing requires making the following configuration decisions. You must decide what is suitable to your environment.

- 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 the SP workload includes parallel batch jobs. If you want to use the High Performance switch for parallel jobs in user mode, or if you want the Resource Manager to coordinate machine usage between interactive and batch jobs in the SP system you must configure the Resource Manager.

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 filesystems. Allowing users to share filesystems 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 users might need to run their jobs successfully. You should configure LoadLeveler to distinguish those nodes from other nodes using, for example, job classes.

## Planning for Parallel I/O File System

The Parallel I/O File System (PIOFS) for AIX is designed for serial or parallel applications that require large temporary files and high I/O bandwidth. PIOFS provides a global namespace and lets you create files as large as 128 Terabytes that span multiple server nodes.

## Planning for NetTAPE

IBM provides two network tape products for AIX that are supported across a network of RS/6000 SP or RS/6000 Family Systems workstations, or both:

- IBM Network Tape Access and Control System for AIX (NetTAPE)
- IBM NetTAPE Tape Library Connection (NetTAPE TLC)

NetTAPE improves and simplifies tape operations management and tape device access in networks, providing a single-system image to users of data stored on tape. NetTAPE supports multiple standard record formats such as Variable, Variable Block, Variable Block Spanned, and Fixed Block. There are APIs for customization as well as ADSM enhancements for remote tape support and external library support.

You must install the NetTAPE product on each node where NetTAPE will be used. The size of the installp image is about 38 MB. NetTAPE requires the following prerequisites:

- AIX Version 4.1, AIX Version 4.2, AIX 4.3, or subsequent compatible releases
- One or more of the following hardware:
  - RS/6000 SP
  - RS/6000 Family Systems
  - Any TCP/IP hardware
  - IBM tape devices
  - Ampex 310 Tape Device

NetTAPE TLC builds on NetTAPE, adding support for robotic tape library devices. These devices include the IBM 3494 and IBM 3495 Tape Library Data Servers, IBM Magstar MP 3575 Tape Library Dataserver, and many other libraries and autochanger devices supported by ADSM device drivers.

The NetTAPE TLC product only needs to be installed on the node where the library server is running. The size of the installp image is about 6 MB. NetTAPE TLC requires the following prerequisites:

- NetTAPE
- One or more of the following hardware:

- IBM Tape Libraries
- StorageTek Tape Libraries
- Other SCSI-attached libraries or autochangers

## Planning for IBM Client Input Output/Sockets (CLIO/S)

Client Input Output/Sockets provides high-speed transparent data transfer and tape access between MVS/ESA systems and AIX systems or between AIX systems. It provides a set of user commands and application programming interfaces that run on either MVS or AIX. CLIO/S is compatible with and complementary to NetTAPE when comprehensive tape access across both AIX and MVS is required.

If you currently have an MVS system, and if you plan to move large amounts of data (many gigabytes) between the MVS system and the SP, you might need CLIO/S. CLIO/S provides high-speed, low-overhead transfers over fast channel-to-channel connections. These channel to channel connections **require** the IBM ESCON Channel Adapter or the IBM Block Multiplexer Channel Adapter cards (and associated microcode) in the SP.

Planning for CLIO/S requires participation by both MVS and SP system planners. The main issues to consider in the planning stage include:

- Look at how frequently your data base is either loaded, backed up, or restored.
- Look at the current size and projected future size of your data base.
- CLIO/S data transfer is accomplished by moving MVS data files directly into AIX, bypassing the TCP/IP stacks in the MVS system.

  Some workloads might be distributed across several nodes. Doing so requires individual channel adapter cards for each node connected directly to the MVS system.

  Other nodes can be connected indirectly to MVS. This is done by routing node to node connections through the SP switch. In this case, the SP node that is directly attached, receives data from the MVS system. The data is then routed indirectly from MVS to other SP nodes via the SP switch.

  Some systems might require intermediate data storage between the MVS and AIX systems while other systems will allow direct data transfer.

## 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 LPP provides file system service to parallel and serial applications on the SP.

You can modify your GPFS configuration after it has been set, but a little consideration before installation rewards you with a more efficient file system.

Hardware and Operating Environment Considerations:

- GPFS requires Recoverable Virtual Shared Disk 2.1. Therefore Virtual Shared Disk, an optional component of PSSP, is also required. If Recoverable Virtual Shared Disk is on multiple nodes, they all must have 2.1.
- If you are using twin-tailed disks, you must select an alternate node as a backup Virtual Shared Disk server.

- Do you have sufficient disks and adapters to provide the needed storage capacity and required I/O time?

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:
  1. You must enable the High Availability Services option (mandatory for GPFS).
  2. GPFS automatically reconfigures itself to continue operations without the failing node.
- 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 data access, 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: Adapter failure is treated as a node failure.

Details on implementing these strategies and other methods can be found in the IBM publication *IBM General Parallel File System for AIX: Installation and Administration Guide* .

# Part 2. Customizing Your System

© Copyright IBM Corp. 1997, 1998

# Chapter 9. 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.

Chapter 4 of *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*, "Reconfiguring your System", 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.

*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 34 on page 148. 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.

© Copyright IBM Corp. 1997, 1998

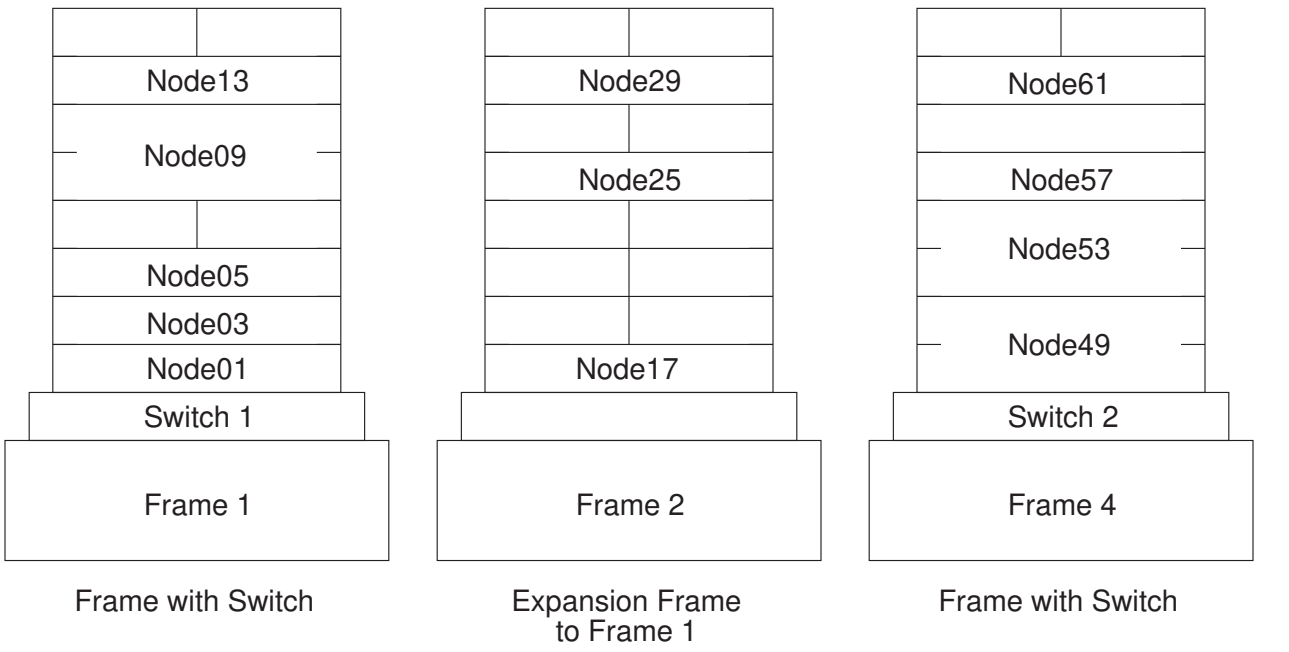



*Figure 34. Sample System: 3-frame, 1-switch*

## How large do I want my system to grow?

Prior to expanding your system, you should 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. The Sample System might grow in any of the following ways:

- Nodes - insert nodes in empty slots;
- Frame - add any of Frames 3, 5, 6, ...
- Switch - install a switch in Frame 2
- switched IP router

## 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 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 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 re-cabling, 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:

- Nameserver

  Every node in your system has a name assigned to it which is resolved by the nameserver you are using. The nameserver translates the symbolic name assigned to a node into its internet address. As you contemplate adding nodes to your system, plan ahead for the names you will assign to the nodes and how your nameserver 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 make sure that the control workstation you currently have is capable of supporting the larger system. It must have sufficient processor speed, DASD, and other hardware - serial ports, Ethernet adapters, and so on. See Chapter 2, “Question 10: What Do You Need for Your Control Workstation?” on page 46.

## What system partitioning topics should I consider?

Migration install enhancements do not require the system to be partitioned, but there are many situations when partitions may be advantageous, including:

- 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(s) if they are to belong to the same partition. Even with the new *System Partitioning Aid* (see Chapter 5, “Planning SP System Partitions” on page 99) there are some restrictions on subdividing switches. Bounding system partitions along frame boundaries also 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 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 may 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?

- You should (generally) add a boot/install server for every 16 nodes being added.
- You must create a primary (and an optional backup) boot/install server for nodes running PSSP 1.2.

## Scenario 1: Expanding the Sample System by Adding a Node

Note in the Sample System that slots 5, 6, 7 and 8 of Frame 2 are empty. You may 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 5, “Planning SP System Partitions” on page 99 for more information on node switch port assignment.) So, you may 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; how depends on your switch type:
   - High Performance Switch -- Perform an **Estart**.
   - SP Switch -- Use the **Eunfence** command.
6. Perform switch installation test.

## Scenario 2: Expanding the Sample System by Adding a Frame

Before addressing specific examples for the Sample 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 may exist that contains nodes and a switch but the nodes do not completely use up the switch ports. One or more non-switched expansion frames may 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 recommends that you add frames only to the end of your system. Otherwise you will have to reconfigure the System Data Repository (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(s)

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. If you do not have enough ports, you will need to upgrade the control workstation.

   If you use HACWS, there are 2 control workstations to consider here.

2. 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.

3. Switch

   You need to consider the implications involved if your system has a switch, either a High Performance Switch or an SP switch.

   - Switch incompatibility.

     High Performance Switches are not compatible with SP switches. You cannot mix High Performance switches with SP switches in an SP system, not even in separate system partitions. To take advantage of the SP switch's performance, you must upgrade all switches to SP switches. High Performance Switches are being phased out and are not available for new systems. They are available only for MES upgrade orders. Under those

circumstances, they are still supported by all currently supported levels of PSSP.

- 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.

4. 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.

5. 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 may need to use a different subnet.

6. System Partitioning

   If you have a partitioned switched system, and the new frame is an expansion frame, you may not need to re-partition, because partitioning for a switched system assumes the maximum number of nodes are present; so the expansion frame nodes are already handled. However, at this point you may 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 System

**Note:** See “Node Placement” on page 80, particularly Figure 11 on page 81, for the specifics on valid node placement, and Chapter 5, “Planning SP System Partitions” on page 99 for more information on assignment of nodes to switch ports.

Consider Frame 1 of the Sample 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 CWS 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; how depends on your switch type:
   - High Performance Switch -- Use the **Estart** command.
   - SP Switch -- Use the **Eunfence** command.
8. Perform installation tests.

## Scenario 2-B: Adding a Frame at the End of the Sample System

For the Sample 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 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 102 to understand how a switch works.

1. Switch incompatibility

   You cannot mix High Performance switches with SP switches in an SP system, not even in separate system partitions. To take advantage of the SP switch's performance, you must upgrade all switches to SP switches.

   High Performance Switch technology predates the SP Switch series. High Performance Switch networks (Feature 4007/10 and 4018) are available for MES upgrade orders only. They cannot be ordered for new SP system builds. Under those circumstances, they are still supported by all currently supported levels of PSSP.

2. Switch type

   What type of switch will you be adding? The table below describes the types available.

   If you are adding any of the switches in the table, an IBM Customer Engineer installs the switch hardware on your system.

3. 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.

| Switch Type | Description | Feature Code |
|---|---|---|
| Scalable POWERparallel Switch (SP Switch) | This switch connects all the processor nodes, providing enhanced scalable high-performance communication between processor nodes for parallel job execution. | 4011 |
| Scalable POWERparallel Switch-8 (SP Switch-8) | This switch provides enhanced high-performance switch function for small systems (up to 8 total nodes). | 4008 |
| High Performance Switch and High Performance LC-8 Switch | Available for existing systems only. | 4010 and 4007(LC-8) |

## The Switch Scenario

The Sample 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. You must:

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 10. Planning for Migration

This chapter includes factors to consider when planning to migrate an existing IBM RS/6000 SP system. Migration addresses upgrading the software from supported levels of PSSP and AIX to PSSP 2.4 and either AIX 4.2.1 (or later), or AIX 4.3. Refer to other chapters in this book for information pertaining to reconfiguring or expanding an existing SP system or to new SP system installations.

Migrating an IBM RS/6000 SP to newer software levels is a relatively complex task, but these complexities (and risks) can be minimized by thoroughly planning each migration phase before beginning the migration.

The principle migration planning phases are:

- Developing your migration goals:

  Briefly discusses considerations such as what software is supported at various migration endpoints and how to plan your SP system configuration in preparation for migration.

- Developing your migration strategy:

  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. Note that understanding coexistence and the advantages and disadvantages of system partitioning will help you refine your migration strategy.

- Reviewing your migration steps:

  Briefly summarizes the high level migration steps and provides a transition to the detailed migration information which is provided in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

The PSSP Installation and Migration Guide describes the specific steps to be completed in implementing a software migration. Other books that might be beneficial for the planning phase include:

- Other PSSP books (such as Administration Guide, Managing Shared Disks).
- The AIX Installation Guide.
- The ITSO Redbook "A Holistic Approach to AIX 4.1 Migration, Planning Guide."
- The ITSO Redbook "AIX Ver. 4.2 Differences Guide."
- Books for IBM LPPs and other products you might be using.

Note that the underlying migration support provided by PSSP 2.4 is basically the same since PSSP 2.2. There are some new considerations that arise from the varying release levels of PSSP, AIX, and related LPPs that either might already exist or that you plan to install on your SP, but the base support for the mechanics of performing a migration are not changed.

© Copyright IBM Corp. 1997, 1998

## 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, you should review prior system plans for unmet goals. 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 (e.g., new function, performance) and possible impacts or disadvantages (e.g., 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 General Parallel File System, you must run your system with Recoverable Virtual Shared Disk. 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.

These software requirements, matched against your IBM RS/6000 SP's workload, generally drive three key components of your migration goals:

1. Planning your base software requirements.
2. Planning how many nodes you will migrate.
3. Planning your migration in verifiable stages.

Understanding coexistence support and possibly doing some system partitioning can help you fully develop your SP systems's efficiency. 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.

# Planning Base Software Requirements

## Supported Migration Paths

A direct migration path to PSSP 2.4 is provided on AIX 4.2.1 (or later) and on AIX 4.3 from each of the PSSP and AIX levels shown in Table 34.

*Table 34. Direct Migration Paths Supported*

| From | To |
|---|---|
| PSSP 2.1 and AIX 4.1.5 | PSSP 2.4 and AIX 4.2.1 or AIX 4.3 |
| PSSP 2.2 and AIX 4.1.5 or AIX 4.2.1 | PSSP 2.4 and AIX 4.2.1 or AIX 4.3 |
| PSSP 2.3 and AIX 4.2.1 | PSSP 2.4 and AIX 4.2.1 or AIX 4.3 |
| PSSP 2.3 and AIX 4.3 | PSSP 2.4 and AIX 4.3 |

PSSP 2.4 supports four different combination crossover releases. A crossover release allows you to just migrate PSSP without changing AIX levels, or just migrate AIX without changing PSSP levels. A crossover release helps to minimize the amount of change caused to an SP CWS or node during migration. The crossover combinations supported are shown in Table 35.

*Table 35. Crossover Release Migration Paths Supported*

| From | To |
|---|---|
| PSSP 2.2 and AIX 4.2.1 | PSSP 2.4 and AIX 4.2.1 |
| PSSP 2.3 and AIX 4.2.1 | PSSP 2.4 and AIX 4.2.1 |
| PSSP 2.3 and AIX 4.3 | PSSP 2.4 and AIX 4.3 |
| PSSP 2.4 and AIX 4.2.1 | PSSP 2.4 and AIX 4.3 |

This means that if your SP contains a CWS or node that is currently at a PSSP or AIX level not listed in the **From** column of Table 34 or Table 35, you must migrate to one of the listed combinations before you can migrate to PSSP 2.4.

## Supported Software Levels

PSSP 2.4 is supported on both AIX 4.2.1 (or later) and on AIX 4.3. PSSP 2.4 is compiled on AIX 4.2.1 and is supported by binary compatibility on AIX 4.3. Your installation's current operational requirements should give you a good understanding of the software requirements that will exist in your IBM RS/6000 SP after it has been migrated to PSSP 2.4.

In addition to the operational requirements placed on your system software, IBM RS/6000 software products also have PSSP release level dependencies. The following table summarizes those dependencies.

| PSSP/AIX Level: | Supported IBM LPP Release: |
|---|---|
| PSSP 2.4 (5765-529) and AIX 4.2.1 (or later) or AIX 4.3 (5765–655 or 5765–C34) | • LoadLeveler 1.3 (5765-145)<br>• Parallel Environment 2.3 (5765-543)<br>• Parallel ESSL 2.1 (5765-C41)<br>• General Parallel File System 1.1 (5765-B95)<br>• Recoverable Virtual Shared Disk 2.1 (5765-646)<br>• PIOFS 1.2 (5765-297)<br>• Performance Toolbox Parallel Extensions (priced feature of PSSP)<br>• CLIO/S 2.2<br>• Network Tape Access and Control System 1.2 (5765-637)<br>• NetTAPE Tape Library Connection 1.2 (5765-643)<br>• HACMP/ES and HACMP 4.2 (5765–A86)<br>• HACWS (priced feature of PSSP, also requires HACMP) |
| PSSP 2.3 (no longer orderable) and AIX 4.2.1 (or later) or AIX 4.3 (5765–655 or 5765–C34) | • LoadLeveler 1.3 (5765-145)<br>• Parallel Environment 2.3 (5765-543)<br>• Parallel ESSL 2.1 (5765-C41)<br>• Recoverable Virtual Shared Disk 2.1 (5765-646)<br>• PIOFS 1.2 (5765-297)<br>• Performance Toolbox Parallel Extensions (priced feature of PSSP)<br>• CLIO/S 2.2<br>• Network Tape Access and Control System 1.2 (5765-637)<br>• NetTAPE Tape Library Connection 1.2 (5765-643)<br>• HACMP/ES and HACMP 4.2 (5765–A86)<br>• HACWS (priced feature of PSSP, also requires HACMP) |
| PSSP 2.2 (5765-529) and AIX 4.2.1 or AIX 4.2.0 (5765-655 or 5765-C34) | • LoadLeveler 1.3<br>• Parallel Environment 2.2<br>• PVMe 2.2<br>• Parallel ESSL 1.2 (5765-422)<br>• PIOFS 1.2<br>• Performance Toolbox Parallel Extensions (priced feature of PSSP)<br>• Recoverable Virtual Shared Disk 1.2 (5765-444)<br>• CLIO/S 2.2<br>• NetTAPE 1.1.1<br>• HACMP 4.2 |
| PSSP 2.2 (5765-529) and AIX 4.1.5 or AIX 4.1.4 (5765-393 or 5765-C34) | • LoadLeveler 1.2.1 and 1.3<br>• Parallel Environment 2.2<br>• PVMe 2.2<br>• Parallel ESSL 1.2<br>• PIOFS 1.2<br>• Performance Toolbox Parallel Extensions (priced feature of PSSP)<br>• Recoverable Virtual Shared Disk 1.2<br>• CLIO/S 2.2<br>• NetTAPE 1.1.1<br>• HACWS (priced feature of PSSP)<br>• HACMP 4.2 |

Refer to other IBM documentation for information on AIX requirements for other LPPs in the IBM RS/6000 software catalog.

## Planning How Many Nodes to Migrate

Subject to your requirements, you might migrate your entire IBM RS/6000 SP or just part of it. IBM has provided some features to help provide flexibility when migrating your system, two of which are coexistence and system partitioning.

1. Coexistence:

   Coexistence refers to support within each product that allows for mixed levels of PSSP and AIX in the same IBM RS/6000 SP system or 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 component of migration.

2. System partitioning:

   System partitioning is a mechanism for dividing an IBM RS/6000 SP 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 IBM RS/6000 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 162.

**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 that need to be applied to your system, refer to the Memo to Users for specific information. These services can be done well in advance but must be done before migrating to PSSP 2.4.

When possible, you should define your migration with multiple stages, breaking them down into distinct steps that can be easily planned and verified. You should plan a reasonable amount of time to complete each step, define validation steps and periods, and be prepared for recovery or 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. Note that you can also distribute system down time over a longer period by migrating a few nodes at a time, subject to your requirements.

There are three main high-level recommendations for doing this:

1. Migrate the control workstation then validate the IBM RS/6000 SP system.
2. Migrate a subset of the nodes then validate the IBM RS/6000 SP system.
3. Migrate and validate the remainder of your SP system according to plan.

For example, you want to minimize the amount of change to your CWS at one time and also minimize your service window. To migrate to PSSP 2.4 from PSSP 2.2 and AIX 4.1.5, first migrate only AIX. Then validate your SP system (still at the PSSP 2.2 level). Then upgrade PSSP after you are satisfied and scheduling

permits. This involves upgrading AIX 4.1.5 to AIX 4.2.1, the crossover release from which you can then migrate PSSP 2.2 to PSSP 2.4.

## 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 the number of nodes, and the methodology to be employed in doing this. You should be entering this planning stage with a basic definition of what you want to migrate (e.g., how many and which nodes), and possibly with some thoughts on how you'd like to go about this.

If you are migrating an entire IBM RS/6000 SP system or an existing system partition, the next two sections on coexistence and system partitions might be unnecessary. If, on the other hand, you are interested in migrating a subset of your IBM RS/6000 system and you are not familiar with the options available to you, the information in those sections on using multiple system partitions and coexistence for migration might be beneficial.

Coexistence and system partitioning are options that can provide flexibility in the number of nodes that you need to migrate at any one time. Your migration goals might suggest the use of multiple system partitions, coexistence, a combination of the two, or neither. Understanding the advantages and disadvantages of coexistence and system partitioning will help you assess their suitability for your needs.

Other factors that will influence your migration strategy include:

- Coexistence limitations on PSSP and the IBM LPPs that will run at each level.
- Setting up boot/install servers.
- Functional changes in PSSP 2.4.
- Migration approach options.

Each of these factors is discussed later.

## Using System Partitions for Migration

The IBM RS/6000 SP supports multiple system partitions, which effectively subdivides an IBM RS/6000 SP into logical systems. These logical systems have two primary features:

1. 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 IBM RS/6000 SP software.

These features facilitate migrating RS/6000 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 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 5, “Planning SP System Partitions” on page 99 for additional information on the use of system partitions.

## Using Coexistence for Migration

In traditional IBM RS/6000 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 LPPs like Parallel Environment or Recoverable Virtual Shared Disk can run on an IBM RS/6000 SP without restriction within the separate system partitions.

For many installations with the desire to migrate a small number of nodes, the system partition approach is not viable. This would apply to 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 also be that the switch isolation function is not desired. Coexistence is aimed specifically at providing additional flexibility for such migration scenarios.

With PSSP 2.4, coexistence support is provided for multiple levels of PSSP, with 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 IBM RS/6000 SP related LPPs are not supported or are restricted in a mixed system partition. For example, the IBM RS/6000 SP parallel processing products (such as Parallel Environment) are generally not supported in mixed system partitions. Inter-node communication over the 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.

## Migration and Coexistence Limitations

PSSP 2.4 is supported on both AIX 4.2.1 (or later) and on AIX 4.3. More specifically, it is compiled on AIX 4.2.1 and supported by binary compatibility on AIX 4.3. This means PSSP does not yet exploit the 64–bit features of AIX 4.3 but it does not prevent you from exploiting AIX 4.3 with your other software.

PSSP 2.4 supports multiple levels of AIX and PSSP in the same system partition (remember that an unpartitioned system is actually one default system partition). However, only certain combinations of PSSP and AIX are now supported to coexist in a system partition.

Coexistence is supported in the same system partition or a single default system partition (the entire SP system) for nodes running any combination of:

- PSSP 2.4 and AIX 4.2.1 (or later) or AIX 4.3
- PSSP 2.3 and AIX 4.2.1 (or later) or AIX 4.3

- PSSP 2.2 and AIX 4.1.5 or AIX 4.2.1
- PSSP 2.1 and AIX 4.1.5

Table 36 lists the levels of PSSP and corresponding levels of AIX that are supported in a mixed system partition.

*Table 36. Levels of PSSP and AIX Supported in a Mixed System Partition*

| Product | AIX 4.1.5 | AIX 4.2.1 | AIX 4.3 |
|---|---|---|---|
| PSSP 2.1 | S | | |
| PSSP 2.2 | S | B | |
| PSSP 2.3 | | S | B |
| PSSP 2.4 | | S | B |

where: S = Supported, B = Supported by Binary Compatibility

In general, any combination of the PSSP levels listed in Table 36 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 and related LPPs do still have some limitations. Also, many software products have PSSP and AIX dependencies — you must ensure that the proper release levels of these products are used on nodes running the coordinating supported PSSP and AIX levels. The following products have notable exceptions that might restrict your ability to migrate one node at a time or might limit your coexistence options.

## Switch Management and TCP/IP Over the Switch

The PSSP switch support (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. For instance, the 332 MHz SMP Wide and Thin Nodes and the SP Switch MX Adapter require PSSP 2.4 or later.

***Coexistence Statement:*** For IP communication over the switch:

- Nodes attached to an SP Switch can coexist in a mixed system partition running any combination of the supported levels of PSSP. If the SP Switch is attached to a dependent node, the primary node and backup primary node must be running PSSP 2.3 (or later). If a node with an earlier level of PSSP is allowed to become the primary node for the SP Switch, then the dependent node connection to the SP Switch will be disabled automatically. If the SP Switch is not attached to a dependent node, the primary node and backup primary node can be running PSSP 2.1 (or later).
- Nodes attached to a High Performance Switch can coexist in a mixed system partition running any combination of the supported levels of PSSP. The primary node can be PSSP 2.1 (or later).
- An SP Switch and a High Performance Switch **cannot** coexist in the same RS/6000 SP System.

**Note:** If you are using Parallel Environment, refer to the PE coexistence statement for more specific limitations that might apply.

***Migration Statement:*** Switch support does not interfere with migrating one node at a time. Converting from High Performance Switch to an SP Switch is a configuration change, not a migration change.

## Extension Node Support

Extension Node support in PSSP 2.4 will function in a mixed system partition of nodes running any combination of the supported PSSP levels however, the control workstation, the primary node and the primary backup node **must** be running PSSP 2.3 or later. If a node with an earlier level of PSSP installed (PSSP 2.1 or PSSP 2.2) is allowed to become the primary node for the SP Switch, the dependent node connection to the SP Switch will be disabled automatically. In the event of a failure it is the administrator's responsibility to override the newly assigned primary or primary backup (to ensure it is a node running PSSP 2.3 or later).

## The Group Services API

Programmers writing to the Group Services API and also Systems Administrators with systems using the Group Services API need to be aware that all nodes must be at PSSP 2.3 or later in order to utilize the Group Services API functions that were introduced in PSSP 2.3 . The Group Services API functions in PSSP 2.3 or later can interoperate with those in PSSP 2.2, but the level of function available in such a configuration is the PSSP 2.2 level. To exploit the newer Group Services API functions, all nodes in a system partition must be running PSSP 2.3 or later.

The Group Services API function is not available in PSSP 2.1.

## High Availability Cluster Multi-Processing

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. Currently there are two variations of the product, HACMP and HACMP/ES. HACMP/ES has enhanced scalability capability.

Typically, HACMP is run only on the control workstation if HACWS is being used. HACMP can also be run on the SP nodes. HACMP/ES does not run on the CWS, it only runs on the SP nodes.

HACMP/ES has a dependency on PSSP Group Services which implies also a dependency on PSSP 2.3 or later.

***Coexistence Statement:*** HACMP and HACMP/ES can coexist in a mixed system partition containing nodes running the supported combinations of PSSP with the following conditions:

- HACMP nodes and HACMP/ES nodes **cannot** coexist in the same cluster.
- HACMP nodes and HACMP/ES nodes can coexist in the same mixed system partition provided that the nodes running HACMP are in an HACMP cluster and those running HACMP/ES are in an HACMP/ES cluster.
- HACMP and HACMP/ES clusters **do not** interoperate.

***Migration Statement:*** Table 37 on page 166 lists the levels of PSSP and corresponding levels of AIX in which HACMP LPPs can be run.

| Table 37. Supported HACMP LPPs | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Product** | **AIX 4.1.5** | **AIX 4.2.1** | **AIX 4.3** | **PSSP 2.1** | **PSSP 2.2** | **PSSP 2.3** | **PSSP 2.4** |
| HACMP/ES 4.2.2 | | S | B | | | S | S |
| HACMP 4.2.2 | S | S | B | | | S | S |
| HACMP 4.2.1 | | S | B | | S | S | S |
| where: S = Supported, B = Supported by Binary Compatibility | | | | | | | |

- Migrating from HACMP to HACMP/ES requires the HACMP cluster to be brought down, all nodes are then migrated to HACMP/ES, and then the cluster is restarted.
- Migrating from HACMP 4.2.1 to 4.2.2 does not require bringing down the HACMP cluster.
- Due to HACMP dependencies on different levels of AIX, migrating one node at a time might require you to upgrade AIX, PSSP, and HACMP on the node all during the same service window.

## IBM Virtual Shared Disk

Virtual Shared Disk, an optional component of PSSP, can coexist and interoperate in a mixed system partition when all of the nodes are running any combination of PSSP 2.2, 2.3, and 2.4 but, the level of function available in the configuration is that associated with the oldest version of PSSP in the system partition. For example, if you have a mixed system partition with nodes running PSSP 2.4 and PSSP 2.2, the level of Virtual Shared Disk function available is that of PSSP 2.2. Virtual Shared Disk can coexist in a system partition running any of the supported combinations but is subject to the following limitations:

- Attempts to configure virtual shared disks on PSSP 2.2, 2.3, or 2.4 nodes for virtual shared disks served by nodes with PSSP 2.2 or later will succeed and requests to these virtual shared disks will be served.
- Attempts to configure virtual shared disks on PSSP 2.2, 2.3, or 2.4 nodes for virtual shared disks served by nodes with PSSP 2.1 will not succeed.
- Attempts to configure virtual shared disks on PSSP 2.1 nodes for virtual shared disks served by nodes with PSSP 2.2 or later will succeed, but requests to these virtual shared disks will not be served and will eventually time out.

For example, if you have nodes in the same system partition as follows:

1. PSSP 2.4, serving virtual shared disks a and b
2. PSSP 2.3, serving virtual shared disks c and d
3. PSSP 2.2, serving virtual shared disks e and f
4. PSSP 2.1, serving virtual shared disks g and h

If you were to configure the virtual shared disks on node 1, running *cfgvsd — a* on node 1, it would be able to configure virtual shared disks a, b, c, d, e, and f. It would not even see virtual shared disks g or h because they are served by a node with PSSP 2.1.

If you were to configure the virtual shared disks on node 4, running *cfgvsd — a* on node 4, it would be able to configure virtual shared disks a, b, c, d, e, f, g, and h but requests to virtual shared disks a, b, c, d, e, and f will not be served and will eventually time out. Only requests to the virtual shared disks g and h will be served.

Virtual Shared Disk does not interfere with migrating one node at a time from PSSP 2.2 or 2.3 to PSSP 2.4 but being able to migrate one node at a time from 2.1 to 2.4 is subject to the coexistence limitations already stated.

## IBM Recoverable Virtual Shared Disk

Recoverable Virtual Shared Disk calls for special consideration in several areas.

***Recoverable Virtual Shared Disk Coexistence:*** Recoverable Virtual Shared Disk 2.1 can coexist and interoperate with Recoverable Virtual Shared Disk 1.2 in a mixed system partition with any combination of the following groupings:

1. Recoverable Virtual Shared Disk 2.1 on either PSSP 2.3 or 2.4 nodes
2. Recoverable Virtual Shared Disk 1.2 on PSSP 2.2 nodes
3. Recoverable Virtual Shared Disk 1.2 and HACMP on PSSP 2.3 or 2.4 nodes

However, if the configuration includes any node running Recoverable Virtual Shared Disk 1.2 or any of the nodes are installed with PSSP 2.2, then the level of Recoverable Virtual Shared Disk function available is that of Recoverable Virtual Shared Disk 1.2.

Recoverable Virtual Shared Disk includes the following quorum rules and restrictions in a coexistence environment:

- Recoverable Virtual Shared Disk 2.1 and 1.2 will treat nodes running Recoverable Virtual Shared Disk 1.1 as down.
- Recoverable Virtual Shared Disk 1.1 will treat nodes running 2.1 or 1.2 as down.
- Quorum is evaluated accordingly:
  - Recoverable Virtual Shared Disk 2.1 and 1.2

    (*number of PSSP 2.4, 2.3 or 2.2 virtual shared disk nodes* +*cws*) / 2 + 1

    **Note:** quorum can be overridden by the administrator in Recoverable Virtual Shared Disk 2.1 or 1.2.
  - Recoverable Virtual Shared Disk 1.1

    (*number of all nodes* + *cws*) / 2 + 1

    **Note:** Upgrading more than half of the nodes to PSSP 2.2 (Recoverable Virtual Shared Disk 1.2) or later will cause the virtual shared disk group running on earlier releases to become inactive.

The following is an example of an 8–node single system partition SP, where Recoverable Virtual Shared Disk 2.1 will coexist and interoperate successfully at the Recoverable Virtual Shared Disk 1.2 level. It is operating at the 1.2 level because there are nodes in the system partition that are installed with PSSP 2.2 (for the example Recoverable Virtual Shared Disk is abbreviated rvsd and Virtual Shared Disk is abbreviated vsd):

1. PSSP 2.4 with rvsd 2.1
2. PSSP 2.4 with rvsd 2.1
3. PSSP 2.4, with no rvsd or vsd installed
4. PSSP 2.3 with rvsd 2.1
5. PSSP 2.3 with rvsd 2.1
6. PSSP 2.3, with no rvsd or vsd installed
7. PSSP 2.2, with no rvsd or vsd installed
8. PSSP 2.2, with no rvsd or vsd installed

***PSSP Migration***

- When migrating from Recoverable Virtual Shared Disk 1.1 and PSSP 2.1, you need to be aware of the coexistence and interoperability limitations in Recoverable Virtual Shared Disk and how quorum is determined.
- When migrating from Recoverable Virtual Shared Disk 1.2 and PSSP 2.2 to Recoverable Virtual Shared Disk 2.1 and PSSP 2.3 or 2.4, you can migrate one node at a time but you must migrate both Recoverable Virtual Shared Disk and PSSP at the same time.
- When migrating from Recoverable Virtual Shared Disk 2.1 and PSSP 2.3 to PSSP 2.4, you can migrate one node at a time just the PSSP level.

***Recoverable Virtual Shared Disk Migration***

- In order to exploit the new functions available in Recoverable Virtual Shared Disk 2.1, all nodes in a system partition must be running Recoverable Virtual Shared Disk 2.1 and PSSP 2.3 or 2.4. Enhancements first made available in Virtual Shared Disk in PSSP 2.3 are used by Recoverable Virtual Shared Disk 2.1.
- When the last Recoverable Virtual Shared Disk 1.2 node is migrated to Recoverable Virtual Shared Disk 2.1 and the last PSSP 2.2 or earlier node is migrated to PSSP 2.3 or 2.4, all nodes in the system partition must have Recoverable Virtual Shared Disk reset. This requires stopping and starting applications, but can occur in a service window of approximately ten minutes. This enables the nodes in that system partition to determine that they should be using the Recoverable Virtual Shared Disk 2.1 level of function.

Table 38 shows the supported levels.

*Table 38. Supported Recoverable Virtual Shared Disk LPPs*

| Recoverable Virtual Shared Disk | AIX 4.1.5 | AIX 4.2.1 | AIX 4.3 | PSSP 2.1 | PSSP 2.2 | PSSP 2.3 | PSSP 2.4 |
|---|---|---|---|---|---|---|---|
| 1.1 | S | | | S | | | |
| 1.2 | S | B | | | S | | |
| 2.1 | | S | B | | | S | S |
| where: S = Supported, B = Supported by Binary Compatibility | | | | | | | |

## General Parallel File System for AIX (GPFS)

GPFS 1.1 is supported on PSSP 2.4 with AIX 4.2.1 (or later) or AIX 4.3. GPFS 1.1 requires Recoverable Virtual Shared Disk 2.1. To get Recoverable Virtual Shared Disk 2.1 level of function, all nodes in the system partition that are running Recoverable Virtual Shared Disk must be running the 2.1 level and PSSP 2.3 or 2.4. If any node is running PSSP 2.2, even if Recoverable Virtual Shared Disk is not installed on that node, Recoverable Virtual Shared Disk will run at the 1.2 level and GPFS will not work in that configuration.

The following is an example of a 4–node single system partition SP where GPFS 1.1 and Recoverable Virtual Shared Disk 2.1 will coexist and operate successfully:

1. PSSP 2.4 with Recoverable Virtual Shared Disk 2.1 and GPFS 1.1
2. PSSP 2.4 with Recoverable Virtual Shared Disk 2.1 and GPFS 1.1
3. PSSP 2.3, with Recoverable Virtual Shared Disk 2.1
4. PSSP 2.3, with no Recoverable Virtual Shared Disk, Virtual Shared Disk, or GPFS installed

The following is an example of a 4–node single system partition SP where GPFS 1.1 does **not** work properly because Recoverable Virtual Shared Disk is running at the 1.2 level due to a node running PSSP 2.2:

1. PSSP 2.4 with Recoverable Virtual Shared Disk 2.1 and GPFS 1.1
2. PSSP 2.4 with Recoverable Virtual Shared Disk 2.1 and GPFS 1.1
3. PSSP 2.3, with Recoverable Virtual Shared Disk 2.1
4. PSSP 2.2, with no Recoverable Virtual Shared Disk, Virtual Shared Disk, or GPFS installed

## Parallel Environment Products

Parallel applications like IBM Parallel Environment for AIX, or Parallel ESSL for AIX, are not supported in a mixed partition. Parallel Environment (PE) includes POE, MPI, MPL, and Parallel Utilities. They require that all the nodes involved in a parallel job be running the same level of PE. This applies to their use for either IP or user space communication. Restrictions are as follows:

- Parallel Environment 2.3 can run in a mixed system partition with PSSP 2.3 and PSSP 2.4 nodes
- Parallel Environment 2.2 must have all nodes in the system partition at PSSP 2.2
- Parallel Environment 2.1 must have all nodes in the system partition at PSSP 2.1

Parallel Environment does not support node by node migration. Limitations are as follows:

- When migrating from Parallel Environment 2.1 and PSSP 2.1 to Parallel Environment 2.3 and PSSP 2.3 or 2.4, all nodes in the system partition need to be migrated within the same service window.
- When migrating from Parallel Environment 2.2 and PSSP 2.2 to Parallel Environment 2.3 and PSSP 2.3 or 2.4, all nodes in the system partition need to be migrated within the same service window.
- Since Parallel Environment 2.3 can run on both PSSP 2.3 and 2.4, a node by node migration from 2.3 to 2.4 can be performed.

*Table 39. Supported Parallel Environment LPP Levels*

| Parallel Environment | AIX 4.1.5 | AIX 4.2.1 | AIX 4.3 | PSSP 2.1 | PSSP 2.2 | PSSP 2.3 | PSSP 2.4 |
|---|---|---|---|---|---|---|---|
| 2.1 | S | | | S | | | |
| 2.2 | S | S | | | S | | |
| 2.3 | | S | B | | | S | S |

where: S = Supported, B = Supported by Binary Compatibility

## Parallel Tools

Parallel Tools include PDBX, PEDB, VT, and Xprofiler. PE requires that all the nodes involved in a parallel job be running the same level of PE. These have the same coexistence limitations as stated for “Parallel Environment Products” on page 169.

## Parallel ESSL for AIX

Parallel ESSL is not supported in a mixed system partition except for system partitions with PSSP 2.3 and PSSP 2.4.This applies to their use for either IP or user space communication. Which level of Parallel ESSL runs on a particular level of PSSP and AIX is based on which level of PE runs on that particular level of PSSP and AIX, as follows:

- Parallel ESSL 1.2.1 requires PE 2.2 or PE 2.3
- Parallel ESSL 2.1 requires PE 2.3

Parallel ESSL is not directly dependent on any level of PSSP or AIX.

Parallel ESSL migration is also dependent on the PE migration.

## LoadLeveler

LoadLeveler 1.3 can coexist in a mixed system partition of nodes running an combination of the supported PSSP levels as long as all nodes are running LoadLeveler 1.3. LoadLeveler is not compatible with earlier levels of LoadLeveler.

LoadLeveler does not support node by node migration. LoadLeveler provides other mechanisms for migration, including the use of separate LoadLeveler clusters.

## CLIO/S 2.2, PIOFS 1.2 and NetTAPE

These three products have the following support:

- Do not require a specific level of PSSP
- Do work in a mixed system partition of any combination of the supported levels of PSSP
- NetTAPE 1.1 and 1.2 can coexist in the same SP but in separate system partitions

### PAIDE (perfagent)

The Performance Toolbox for AIX, Agent Component (PAIDE) is required for PSSP 2.2 or later. The correct level of PAIDE (perfagent) needs to be installed on the control workstation and copied to the **lppsource** directory so it can be installed on the nodes.

To check whether the correct level of perfagent is installed, issue the command:

```
lslpp –l perfagent.server
```

**Note:** Be cautious when you order perfagent. It has several feature codes. You must choose the one relative to the level of AIX that you are running. See "Memo to Users" for latest updates.

## 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 IBM RS/6000 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 IBM RS/6000 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/device driver IP buffer pools are the primary considerations. The rpoolsize and spoolsize parameters available in PSSP 2.2, PSSP 2.3 (PTF needed, see "Memo to Users" for latest updates), and PSSP 2.4 are changed using the chgcss command. The aggregate pool size is a function of the size of kernel memory.

In summary, the recommended approach for factoring coexistence into your overall IBM RS/6000 SP tuning strategy is to start with the above general approach to tuning for mixed levels of AIX/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 IBM RS/6000 SP migration strategy progresses.

**Note:** See performance tuning information on the web at **http://www.rs6000.ibm.com/support/sp**

## 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.

One other area of migration planning is that of additional resources. For example, this would include your evaluation of the need for additional DASD to support multiple levels of software, particularly if you plan on using coexistence. For this, you should plan on having 2 GB of disk allocated for each level of AIX/PSSP being served by your control workstation or boot/install server. This is typically used for additional directories under the modified directory structure introduced in PSSP 2.2, 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/

## Changes in Recent Levels of PSSP

Here are some changes in PSSP support that could affect your migration plans.

### PVMe

PVMe is no longer supported. MPI is the strategic direction IBM has recognized for message passing codes. IBM does not support it, but if your application cannot be migrated to use the MPI protocol then you might want to look at the Oak Ridge version of PVM.

### Automounter

As of PSSP 2.3 , the Amd automount daemon, which is freely available under license, is replaced by the AIX automount daemon, which is available as part of NFS in the Network Support Facilities of AIX Base Operating System (BOS) Runtime. Amd uses map files to define the automounter control. These map files are not compatible with the AIX automounter and must be converted.

If your current installation has the Amd configuration turned on and is using the SP User management Services (SP site environment variables *amd_config* and *usermgmt_config* are both **true**), the SP maintains a user home directory map file for the/**u** file system. If you have not modified the/**etc/amd/amd-maps/amd.u map** file, the PSSP System Management Software will automatically convert this map file for you when migrating to PSSP 2.3.

If you have modified the **amd.u** Amd map file, added your own map files for additional automounter support, or in any other way customized your Amd installation, you will need to consider the impact of automounter conversion in planning your migration to PSSP 2.3. You will need to manually convert your Amd map files to AIX Automount map files. Please refer to the following AIX publications for information on the AIX automounter and map file format:

**AIX command reference:** *automount* command

**System Management Guide: Communications and Networks:** Mounting an NFS File System using the automount daemon

**System Management Guide: Communications and Networks:** *NIS Automounter*

If you find it impossible to convert your current installation to use the AIX automount daemon, you can provide your own automounter support through a set of user customization scripts. See "Managing the Automounter" chapter of the SP Administration Guide for more details.

### Print Management

As of PSSP 2.3, the SP Print Management Subsystem is no longer supported . IBM recommends the use of Printing Systems Manager (PSM) for AIX as a more general solution for managing printing on the IBM RS/6000 SP. Note that the SP Print Management Subsystem is still supported on nodes of an IBM RS/6000 SP that are running earlier levels of PSSP, even if the IBM RS/6000 SP system has been partially migrated to PSSP 2.3.

## HACWS Migration Strategy

High Availability Control Work Station (HACWS) only runs on the RS/6000 SP Control Workstations. A HACWS configuration at the PSSP 2.4 level requires the following software on both control workstations:

- PSSP 2.4 (including the ssp.hacws 2.4.0.0 file set).
- AIX 4.2.1 (or later) or AIX 4.3 and PSSP 2.4. Refer to the "Memo to Users" for the latest information on what levels of AIX are supported with PSSP 2.4.
- Any level of HACMP 4.2 or later that is supported with the level of AIX that you are using. Refer to the appropriate HACMP documentation to determine what levels of HACMP are supported with the level of AIX that you are using or considering.

Whether or not you need to upgrade all three of these at the same time depends on your software levels before migration. You can choose to upgrade your HACWS configuration a little at a time, stopping along the way to run your system long enough to become confident that it is stable before proceeding to the next phase.

For more information see *IBM Parallel System Support Programs for AIX: Installation and Migration 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 IBM RS/6000 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 4.2.1 (or later) or AIX 4.3 and PSSP 2.4 before migrating any of the nodes.

For example, to migrate to PSSP 2.4 from PSSP 2.2 and AIX 4.1.5, first migrate only AIX. Then validate your SP system (still at the PSSP 2.2 level). Then upgrade PSSP after you are satisfied and scheduling permits. This involves upgrading AIX

4.1.5 to AIX 4.2.1, the crossover release from which you can then migrate PSSP 2.2 to PSSP 2.4.

For systems running PSSP 1.2 or PSSP 2.1, or if an overwrite install is desired, the control workstation migration must include both AIX and PSSP upgrades before the SP system can be returned to production.

**Note:** Both upgrades must be done in the same service window.

After the control workstation has been migrated to AIX 4.2.1 (or later) or AIX 4.3 and PSSP 2.4, and the system has been validated, the nodes can be migrated. Start the node migration with boot/install servers if applicable. The same basic migration options exist for migrating the nodes, such as:

- 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. These components should be reviewed and assessed, considered from a sizing and impact point of view, and qualified with respect to your overall migration goals and strategy. Additional details on these steps can be found in the *IBM Parallel System Support Programs for AIX: 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 - AIX, PSSP, LPP and other products required AIX and PSSP service (for older levels)
   - new release levels of other products used
   - any additional DASD required, resources (e.g., tape) for backups
6. Create system backups - CWS, nodes to be migrated
7. Conduct the migration, in stages as applicable:
   a. Apply required service to nodes prior to migration (required PTFs are necessary for coexistence to work)
   b. Prepare the control workstation (such as DASD, PTF service, archive the SDR)
   c. Migrate the control workstation to the latest level of PSSP (and AIX if necessary), validate

d. Partition the system (might be necessary due to coexistence limitations)

e. Migrate a test node to the latest level of PSSP (optional but highly recommended)

f. Migrate boot/install servers to the latest level of PSSP (and AIX if necessary)

g. Migrate any remaining nodes to the latest level of PSSP (and AIX if necessary)

**Note:** Each step should be completed and verified before going on to the next step.

8. Perform post-migration activities

# Part 3. Appendixes

© Copyright IBM Corp. 1997, 1998

# Appendix A. The System Partitioning Aid - A Brief Tutorial

PSSP includes a tool to facilitate system partitioning activity. 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, and then the CLI version of this application. This is the order of exposure recommended for the inexperienced partitioner. 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 35 on page 180. 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.)

© Copyright IBM Corp. 1997, 1998

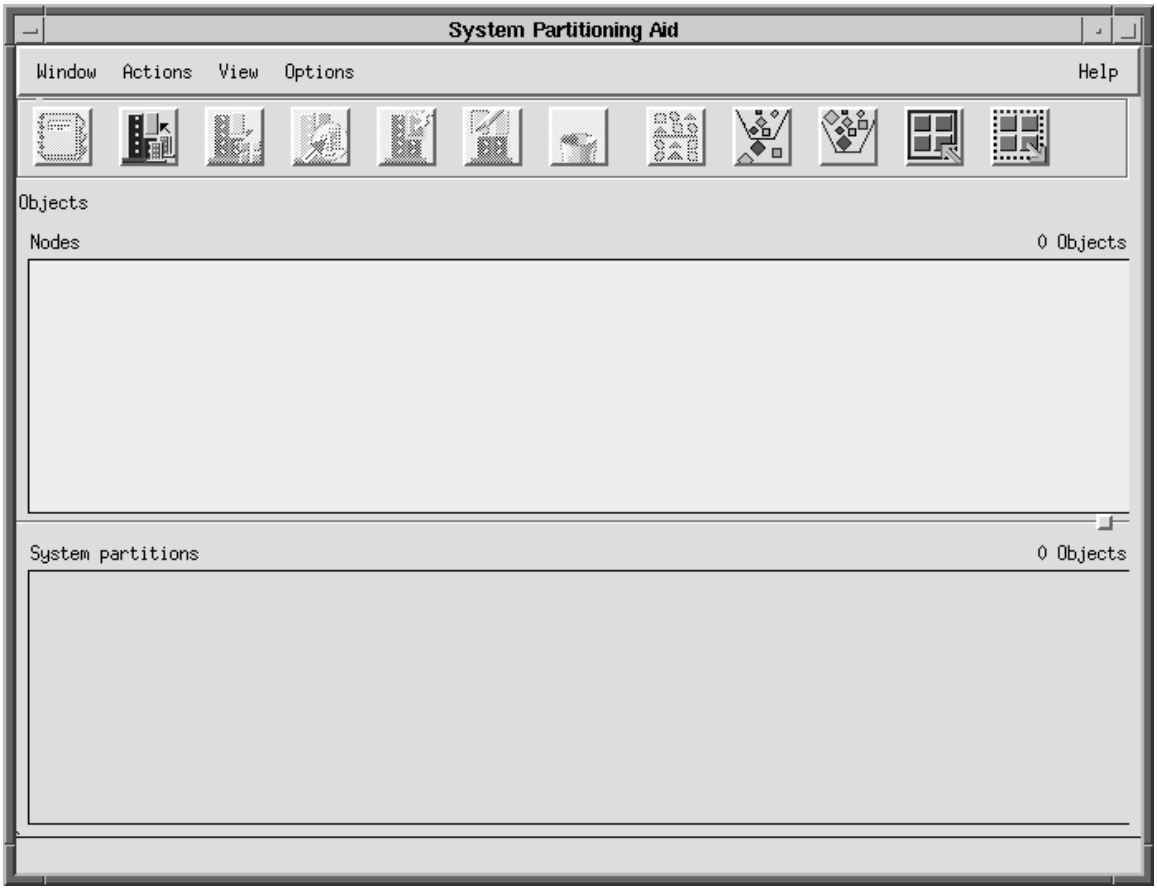



*Figure 35. spsyspar's Main Window*

In Figure 35, 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.

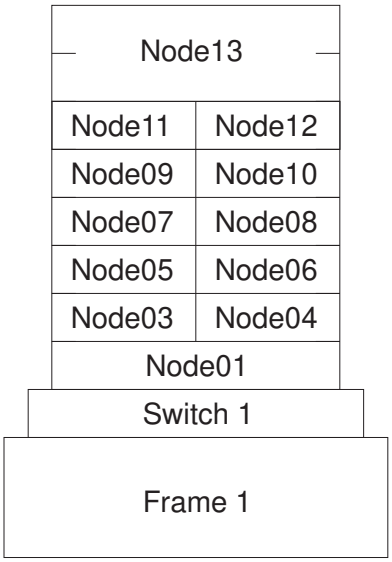



*Figure 36. Sample 1-Frame System (1 wide, 10 thin, and 1 high nodes)*

For example, assume you invoked **spsyspar** on the control workstation for the 1-frame system pictured in Figure 36, 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 37 on page 182. 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.

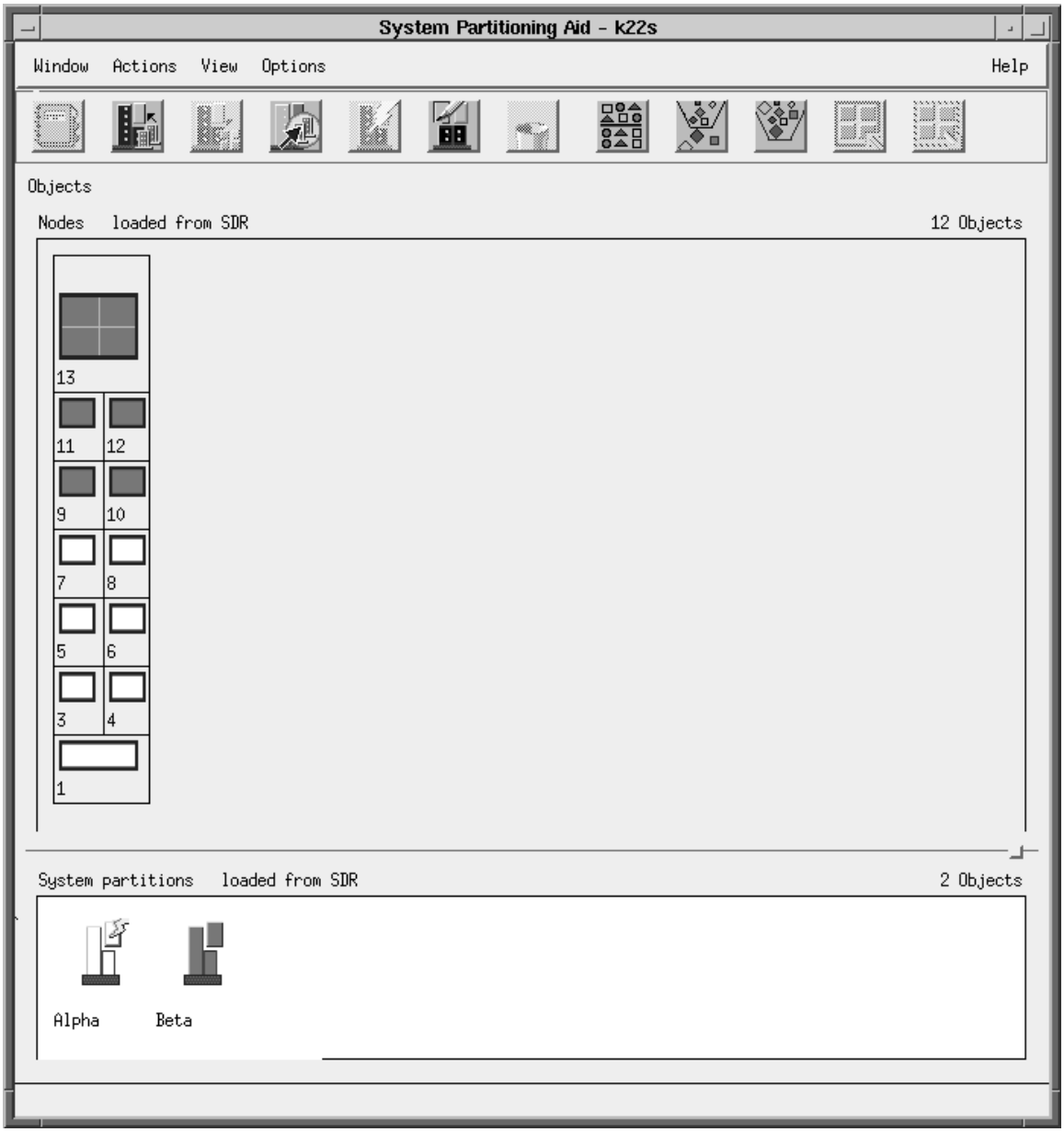



*Figure 37. Main Window for Sample System*

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 37, 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 partition" pane of Figure 37, you'll see that the Alpha partition is marked with a "lightening bolt". This signifies that Alpha is the *active partition*. Hence, any partition-specific activity, such as assignment of nodes, would be directed at partition Alpha. In addition, the brighter colored System partitions pane is the pane of "focus". This affects the choices available from the Tool Bar and the Pull Down Menu - items not applicable for the current focus are grayed out.

## 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/or system partitions previously selected. Actions which are not available appear grayed-out. For example, if you clicked on node 8 in the Nodes pane, and 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 38.

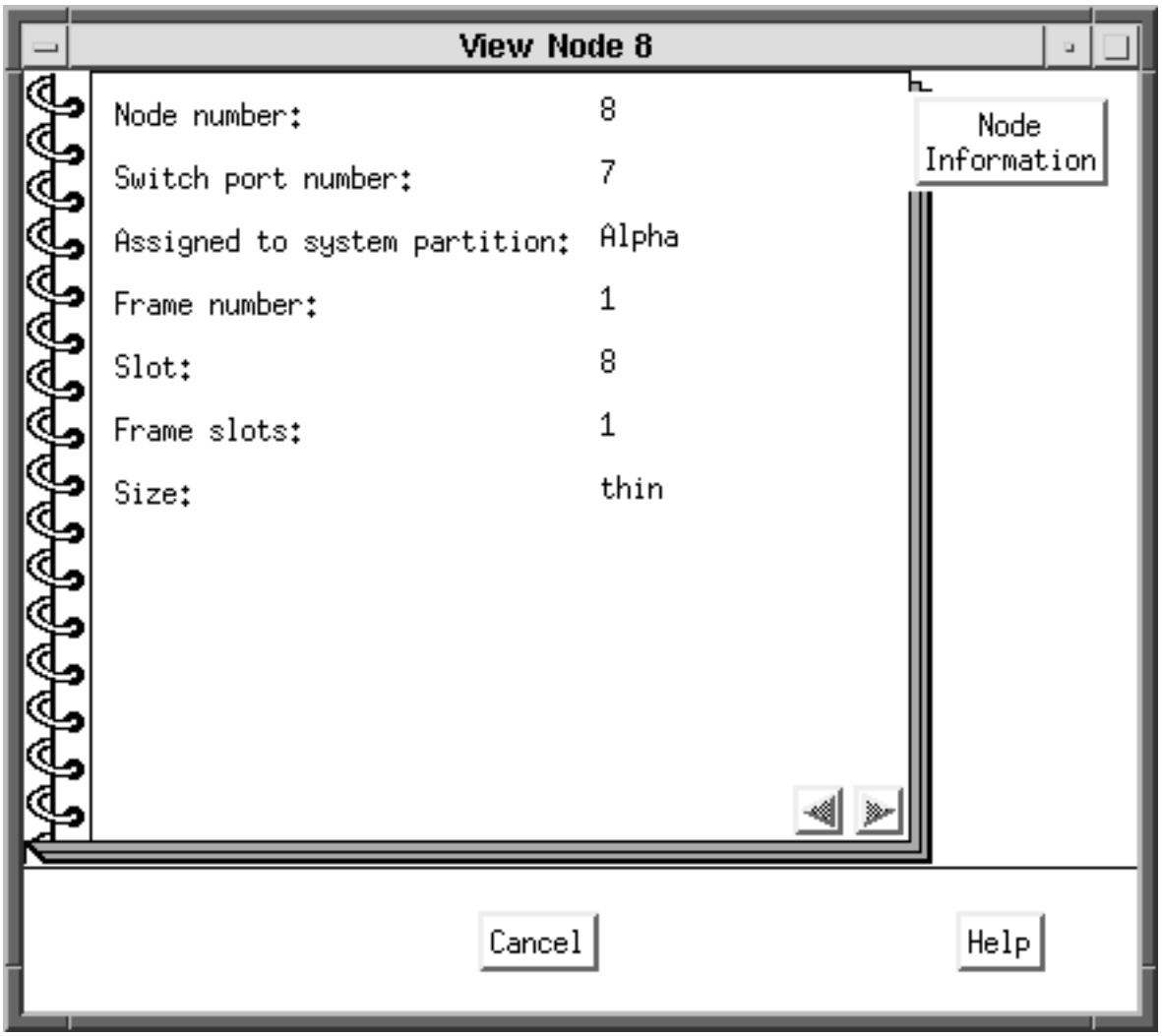



*Figure 38. Notebook for Node 8 of Sample System*

If you instead click on partition Alpha in the system partitions pane, and 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 39 on page 184 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 may not satisfy that criteria. Configurations created via the System Partitioning Aid are evaluated for correctness and performance. Hence, provision is made for the "Chip Allocation" and "Performance" pages of the system partition notebook to record such data for a user-created layout.

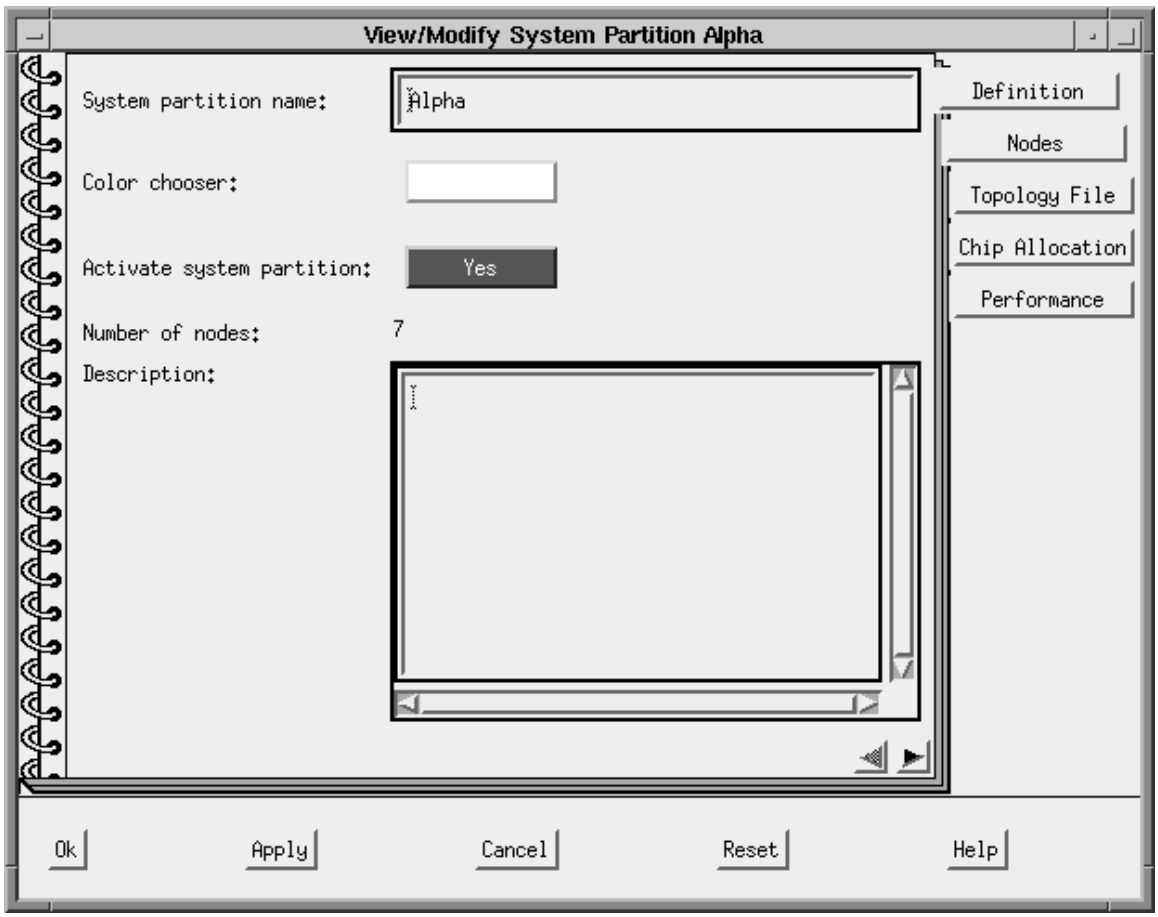



*Figure 39. Notebook for Partition Alpha of Sample System*

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 40 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 40, configuration 8_8 has been expanded showing there are three layouts available under this configuration.

If you click on a layout and press 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 which matches the real system, the user has a choice of seeing nodes pictured as defined in the SDR. The default, and the only possibility if the SDR is unavailable, is to show only thin nodes with all slots populated, since **spsyspar** cannot know the correct node types to show, and so depicts all nodes as thin.

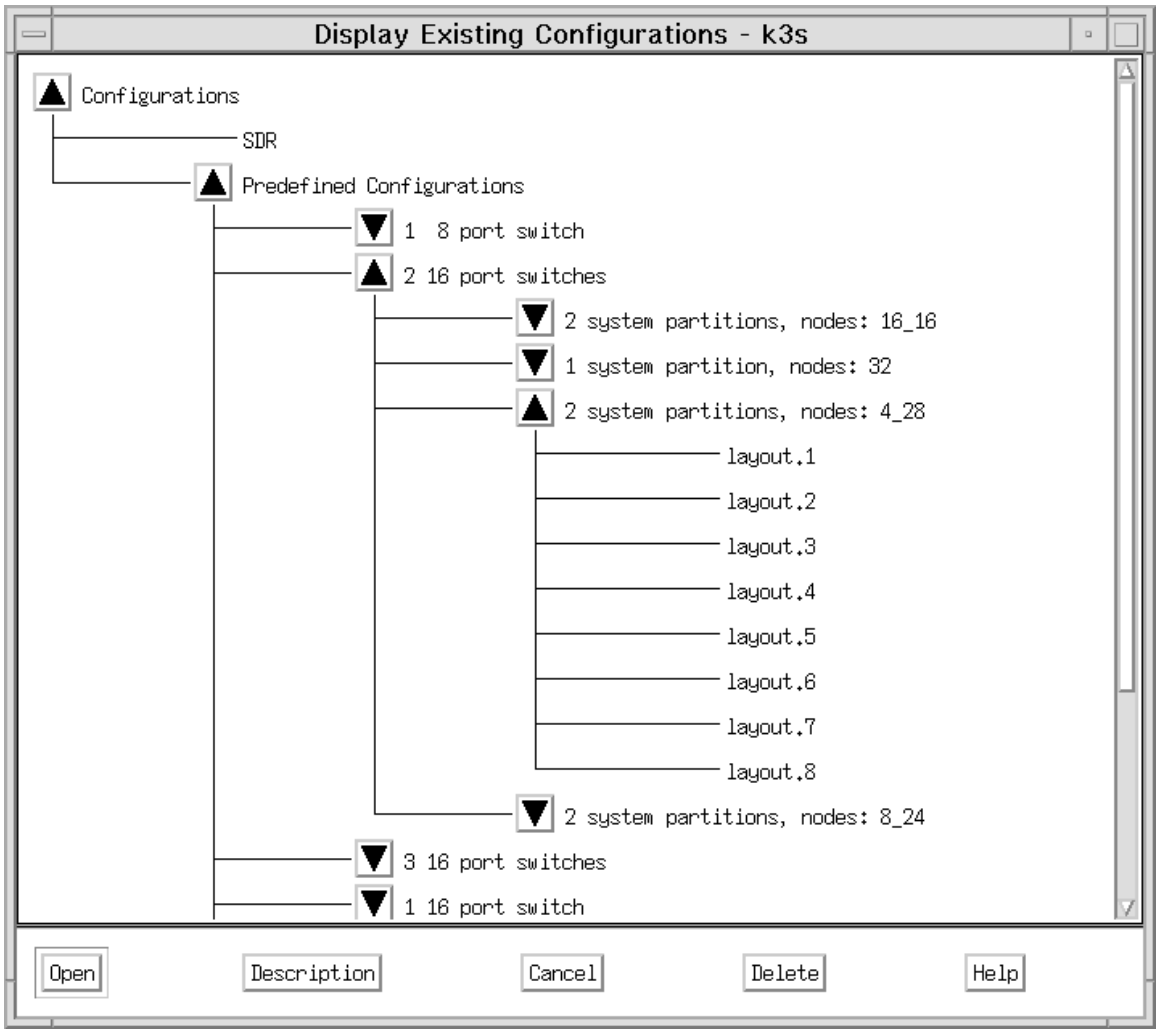



*Figure 40. Alpha Notebook for Sample System*

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 37 on page 182.

### 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 41 on page 187, and the nodes have sorted in descending order.

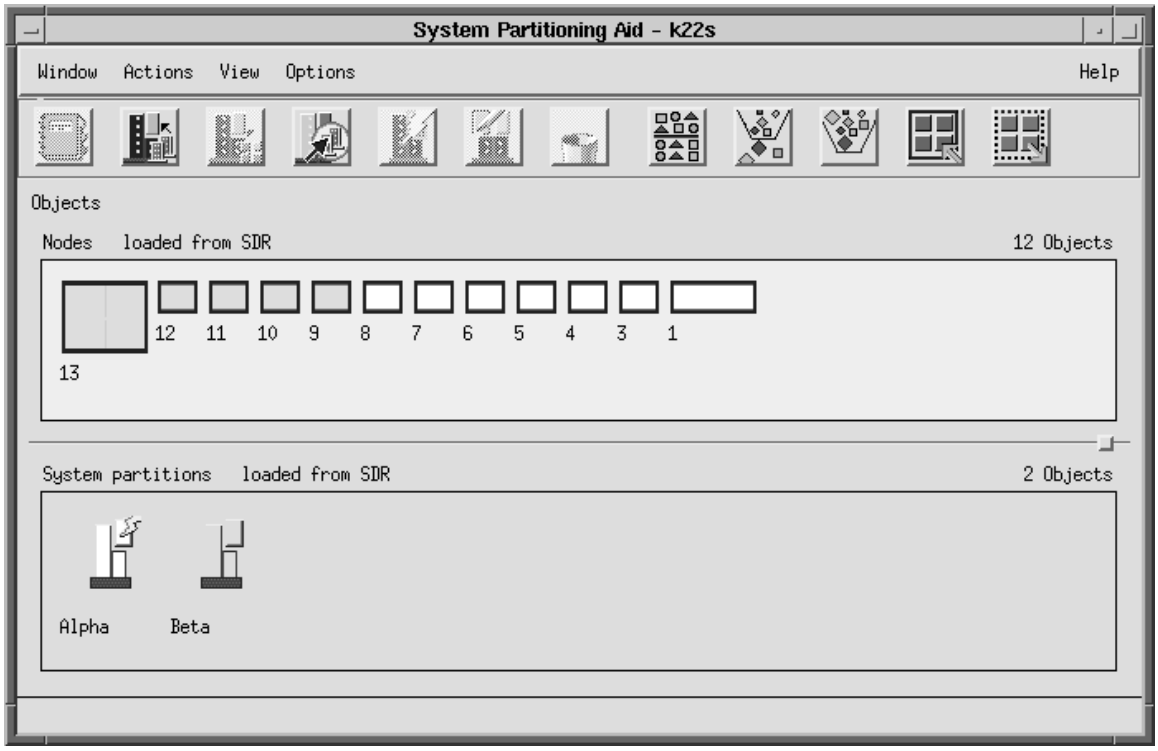



*Figure 41. Descending sort in Nodes pane (Icon View)*

## 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 42 on page 188 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.

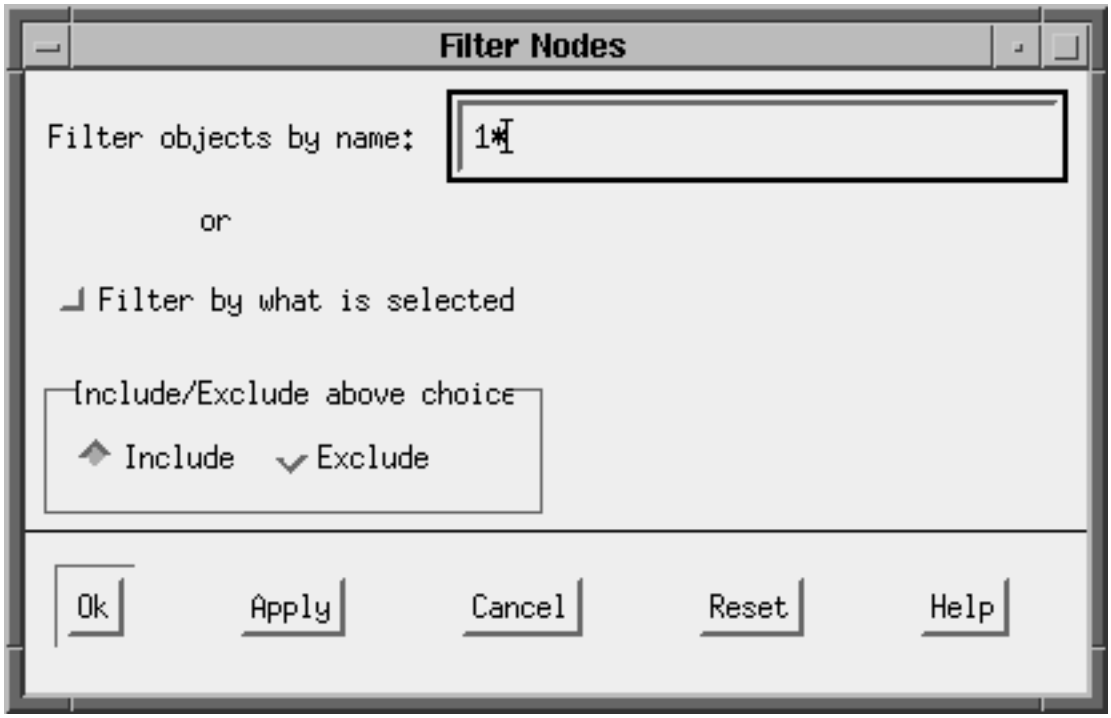



*Figure 42. Filter menu with "1*" filter specified for Nodes pane*

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, refer to "SP 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 43 on page 190 and the file shown is shipped with PSSP in **ssp.top** as "inpfile.template" in the directory **/spdata/sys1/syspar_configs/bin**.

Recall the Sample System of Figure 36 on page 181, and the Alpha and Beta partitions of Figure 37 on page 182. An input file for **sysparaid** which specifies that layout is the file "my_part_in" presented in Figure 44 on page 190.

```
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 identfied 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 43. 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 44. my_part_in file.*

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, metrics files; |
| **syspar.2.Beta** | directory for Beta - node list, topology, snapshot, metrics files. |

## Example 3 of Chapter 5

The picture of the 3-frame system discussed in Chapter 5 is reproduced in Figure 45 on page 192 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 switches, and various node sizes” on page 111:

```
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, and so 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.

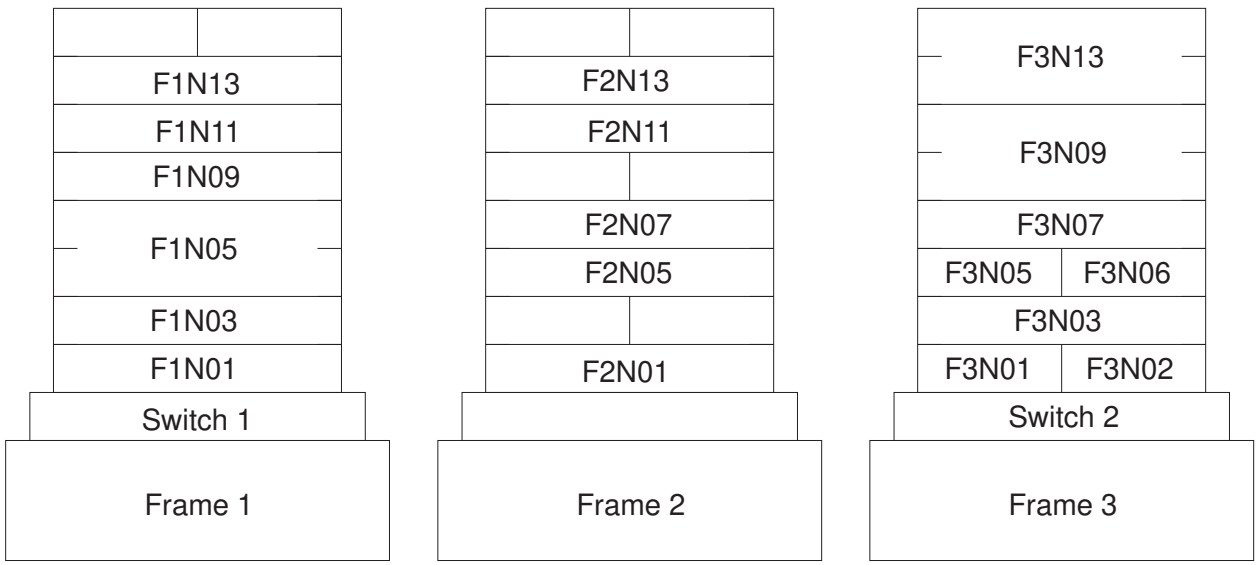



*Figure 45. Three frames with 2 switches*

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 46 on page 193. 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.

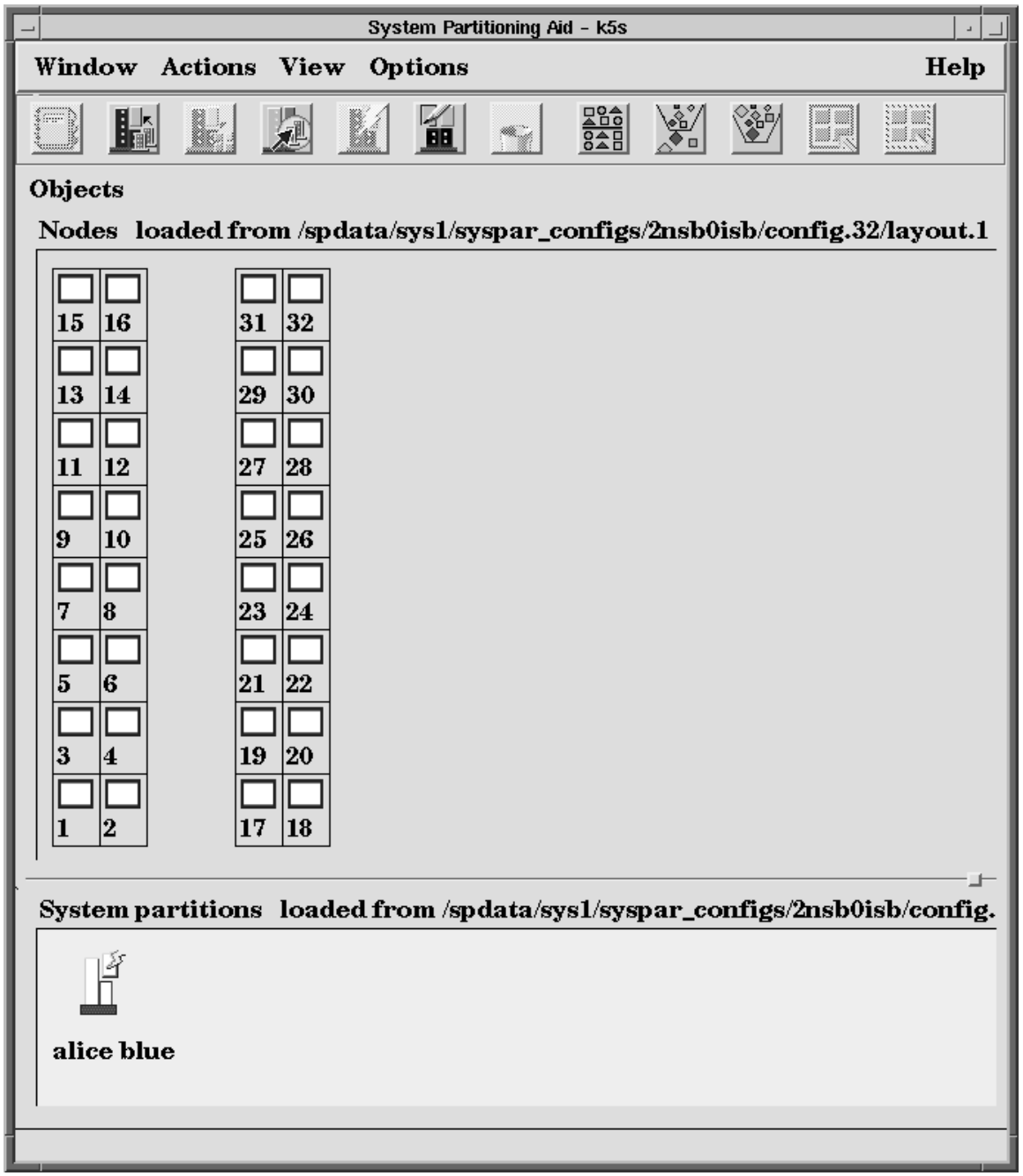



*Figure 46. Main Window for Example 3 of Chapter 5*

Your objective for system partitions is to divide the system pictured in Figure 46 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 47 on page 195:

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 and 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 on node 12, and 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, and 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.

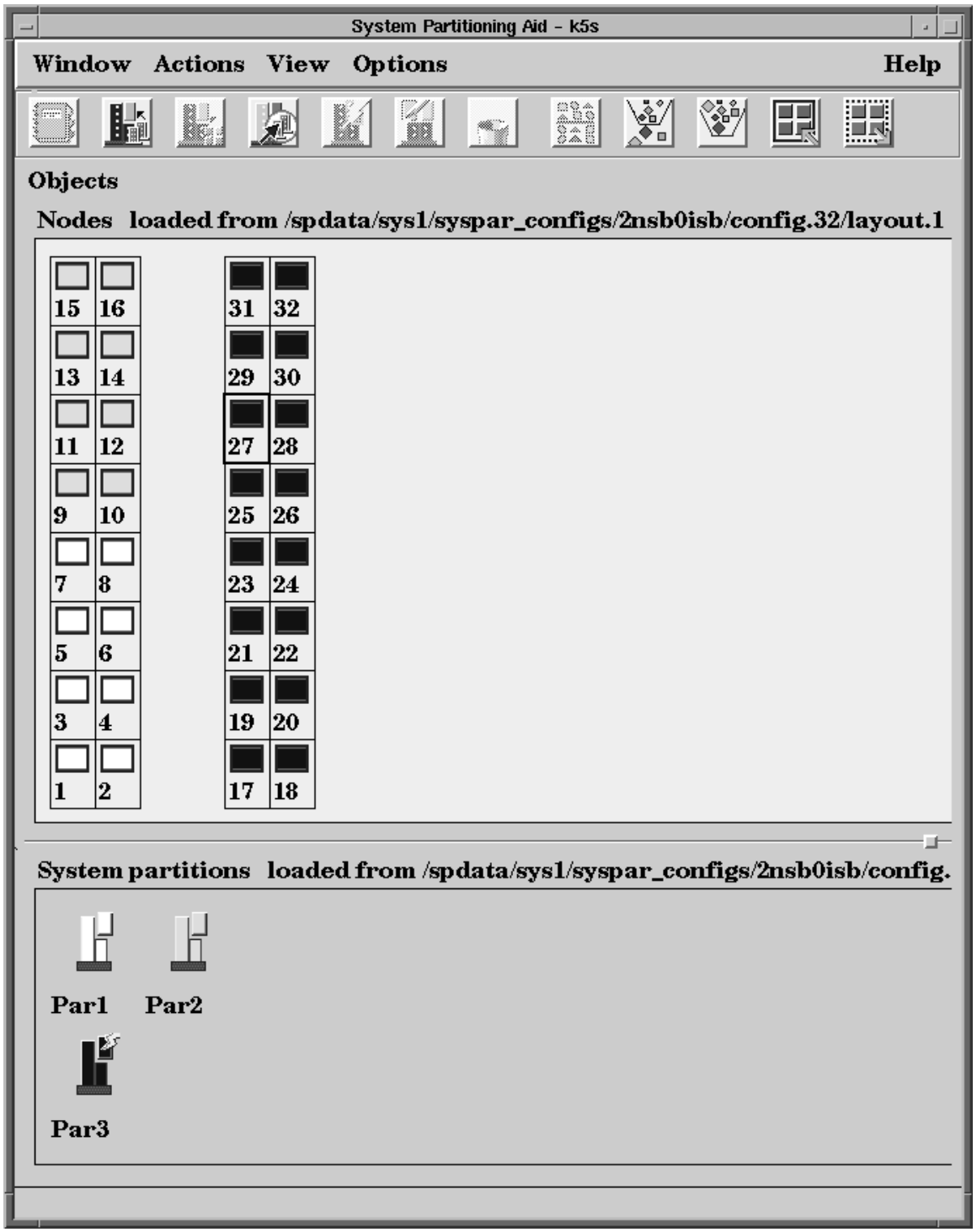



*Figure 47. System Partitioning for Example 3 of Chapter 5*

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 48 on page 196. For the real system, Nodes 5, 25 and 29 will be high. Figure 48 on page 196 looks good in this respect. However, Nodes 6 and 8 distort our perception of Node 5.

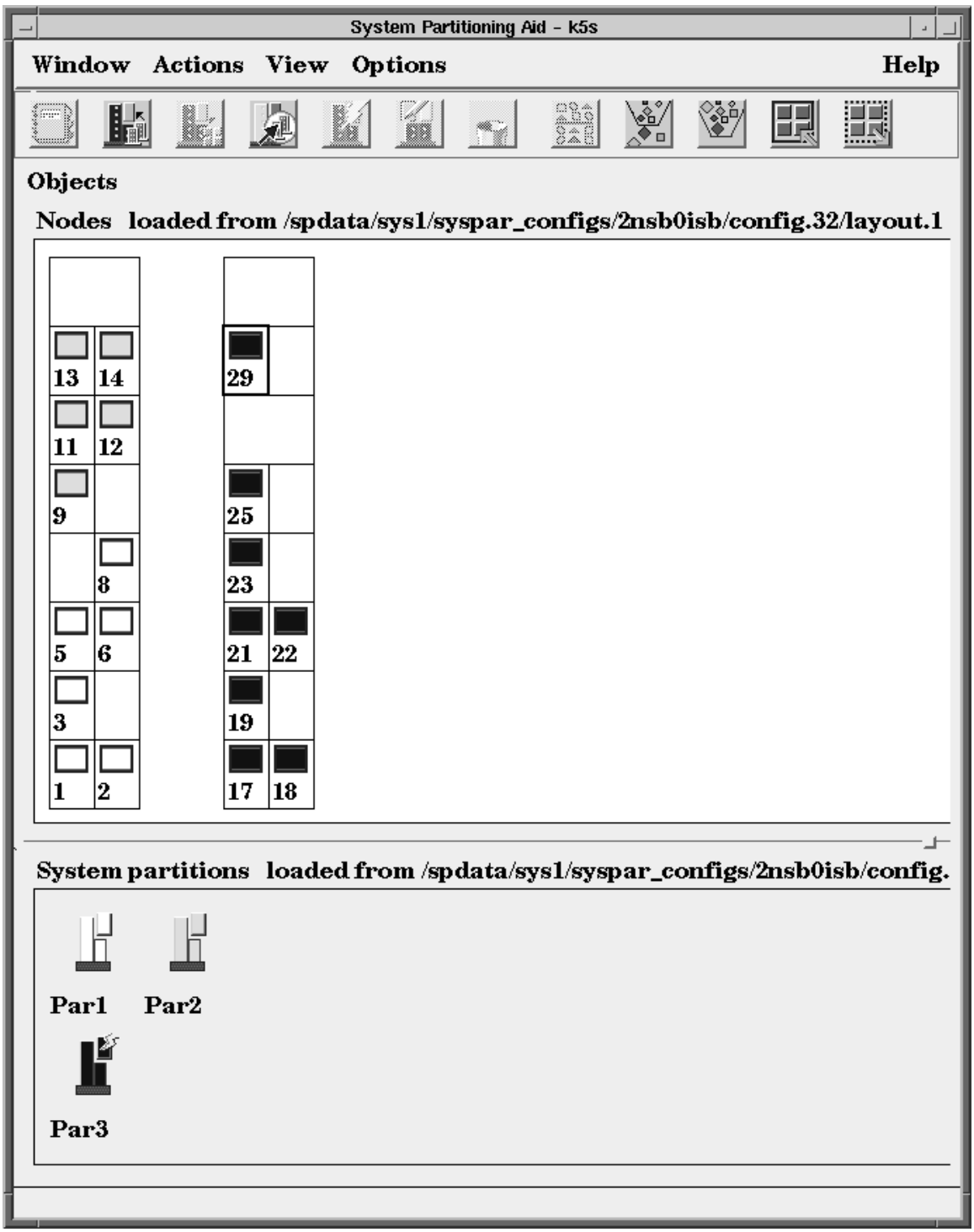



*Figure 48. System Partitioning for Example 3 of Chapter 5*

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 49 on page 197. 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".

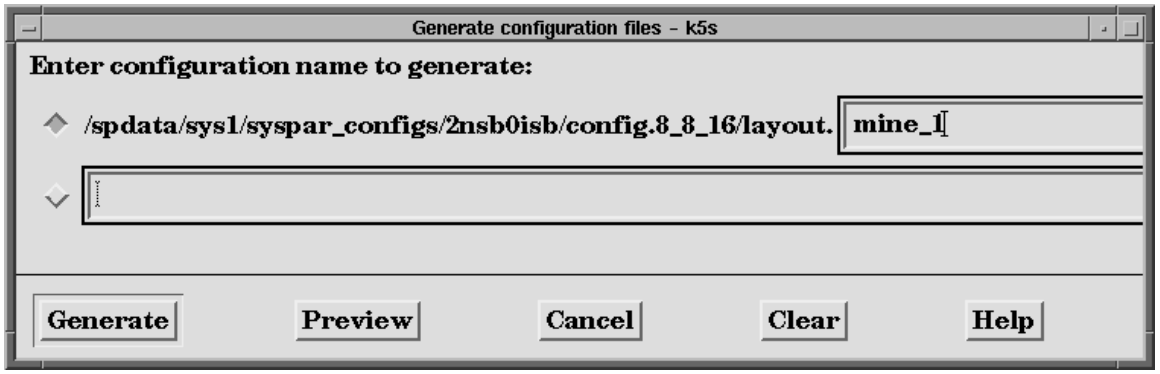



*Figure 49. Dialog box for specifying name of new layout*

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**.

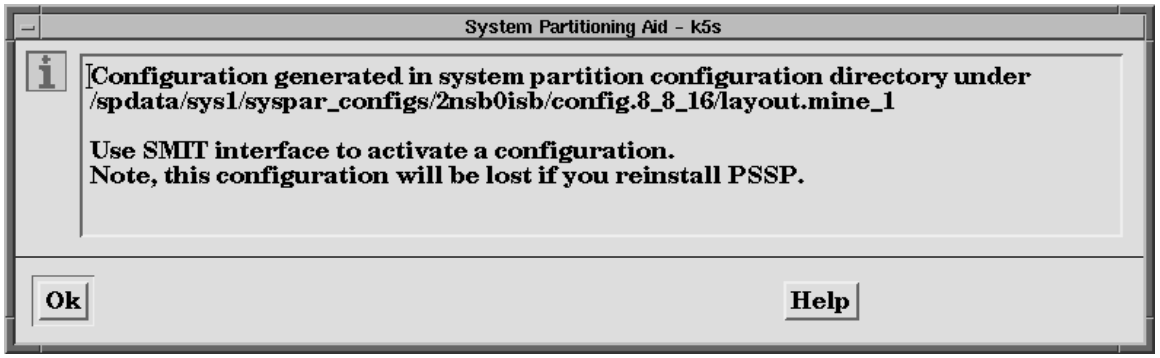



*Figure 50. Message issued when new layout has been saved*

## 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 51 on page 198. (**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 51. CLI input file from "spsyspar"*

If you use the CLI directly, you can use an input file similar to that in Figure 51, 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 52 on page 199 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 52. 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 data is available for viewing in the GUI as the "Chip Allocation" page of Par1's notebook. (First icon.) This GUI presentation is produced in Figure 53 on page 200. Par1 is completely contained in Frames 1 and 2 and so only uses Switch 1, denoted NSB 1 (Node Switch Board 1) in spa.snapshot. 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 at the layout.mine_1 directory level.

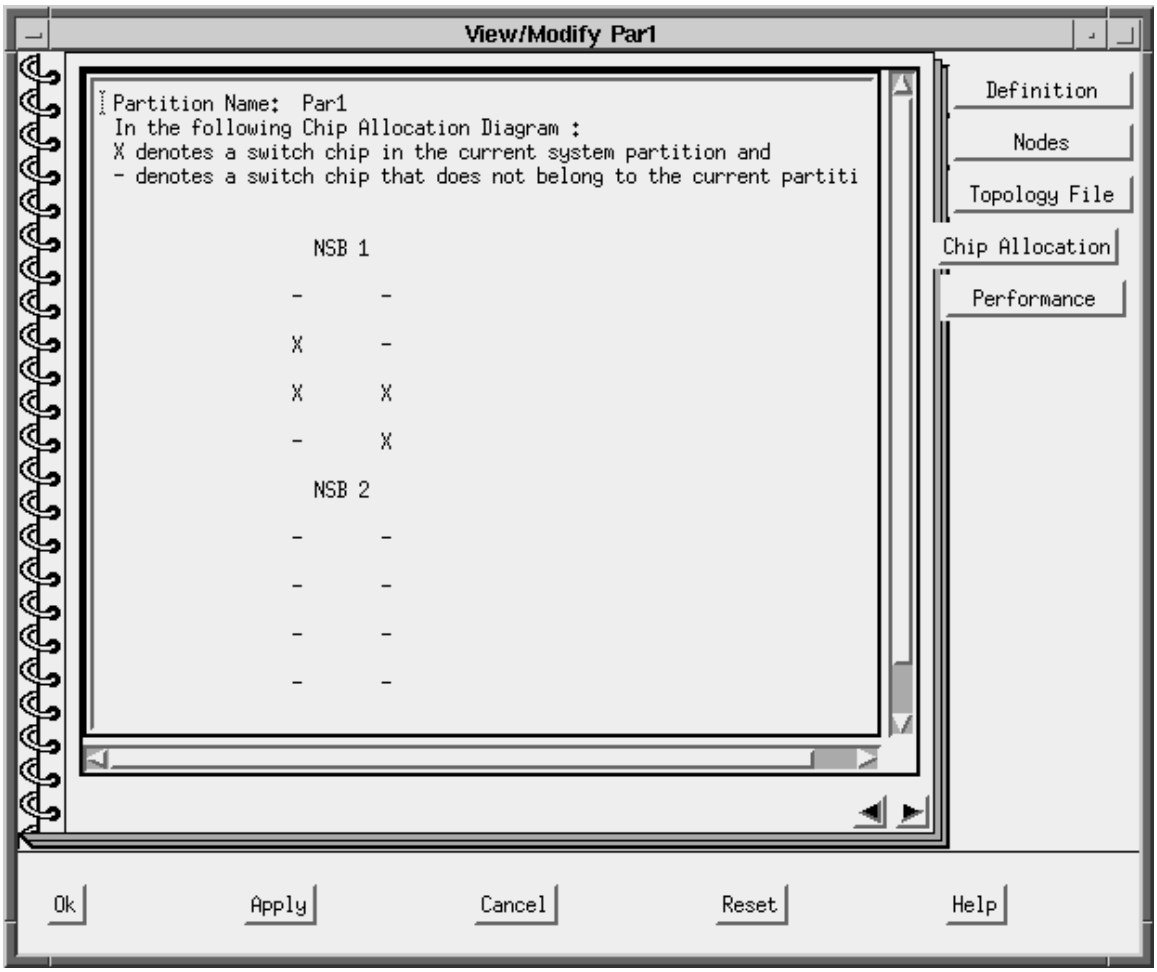



*Figure 53. Switch chips allocated to system partition Par1*

The spa.metrics data is available in the GUI on the "Performance" page of Par1's notebook. This GUI presentation is given in Figure 54 on page 201. Chips 5 and 6 are the node chips of Figure 53. 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.

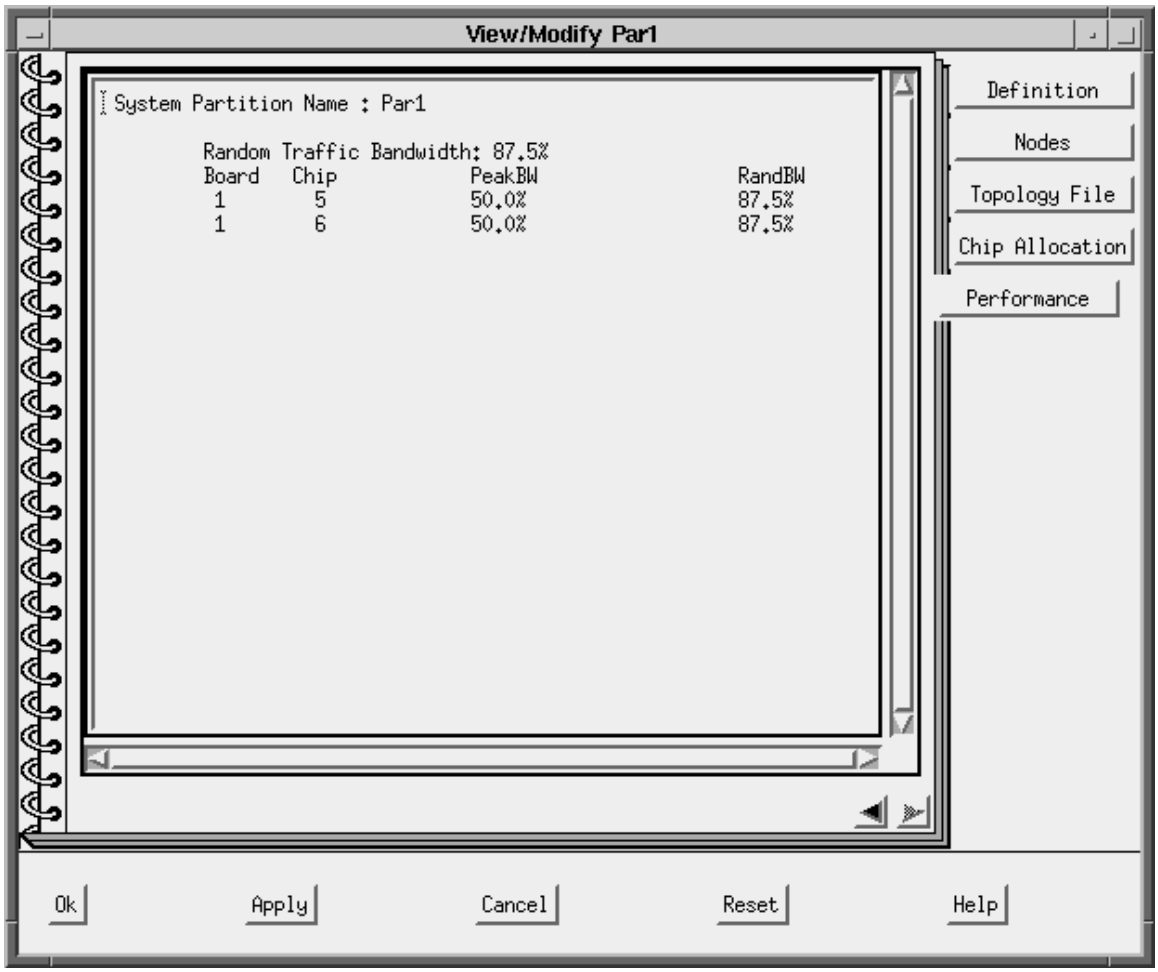



*Figure 54. Performance numbers for system partition Par1*

# Appendix B. System Partitioning

This appendix contains a description for each of the system partitioning layouts, ordered by system size, that IBM provides.

## 8 Switch Port System

### Layout for 4_4 Partition of 8 Switch Port System with a 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 a 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

© Copyright IBM Corp. 1997, 1998

### 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 40. List of SP Planning Worksheets*


| Number | Name | Page |
|---|---|---|
| 1 | SP Preliminary Application List | 228 |
| 2 | IBM Program Products | 228 |
| 3 | External Disk Storage Needs | 229 |
| 4 | SP Planning | 229 |
| 5, 6 | SP Node Layout Diagrams (several copies) | 230 |
| 7 | SP Hardware Configuration by Node | 231 |
| 8a, 8b | SP Node Network Configuration | 232 |
| 9 | Switch Configuration | 234 |
| 10a, 10b | Supported Adapters | 234 |
| 11 | SP System Image Worksheet (SPIMG) | 237 |
| 12 | PSSP 2.4 File sets | 238 |
| 13 | SP Control Workstation Image | 239 |
| 14 | Select a Time Zone | 240 |
| 15 | SP Control Workstation Network | 241 |
| 16 | SP Site Environment | 242 |
| 17 | SP Authentication Worksheets | 243 |


Make photocopies of these worksheets as required. Instructions for using the worksheets are contained in Chapter 2, “Defining the System that Fits Your Needs” on page 11, in Chapter 3, “Defining the Configuration that Fits Your Needs” on page 55, and in Chapter 6, “Planning for Security” on page 117.

© Copyright IBM Corp. 1997, 1998

*Table 41. Preliminary List of Applications*

| SP Preliminary List of Applications - Worksheet 1 | | |
|---|---|---|
| **Application** | **Parallel √** | **Need Switch √** |
| | √ n ? | √ n ? |
| | √ n ? | √ n ? |
| | √ n ? | √ n ? |
| | √ n ? | √ n ? |
| | √ n ? | √ n ? |
| | √ n ? | √ n ? |
| | √ n ? | √ n ? |
| | √ n ? | √ n ? |
| Use √ if you want this application, n if you do not, ? if you do not know if you want this application. | | |

*Table 42. IBM Program Products to Order*

| IBM Program Products - Worksheet 2 | | | |
|---|---|---|---|
| **Order** | **Program Product** | **Program Number** | **Level** |
| | IBM C for AIX | 5675-423 | 3.1 |
| | IBM C Set ++ for AIX | 5765-421 | 3.1 |
| | IBM Parallel System Support Programs for AIX (PSSP) | 5765-529 | 2.4 |
| | IBM LoadLeveler | 5765-145 | 1.3 |
| | IBM Client Input Output/Sockets (CLIO/S) | 5648-129 | 2.2 |
| | IBM Parallel Environment for AIX | 5765-543 | 2.3 |
| | | 5765-543 | 2.2 |
| | IBM Parallel Optimization Subroutine Library | 5765-392 | 1.3 |
| | IBM Parallel Engineering and Scientific Subroutine Library | 5765-C41 | 2.1 |
| | | 5765-B50, 5765-422 | 1.2 |
| | | 5765-422 | 1.1 |
| | IBM Parallel I/O File System for AIX | 5765-297 | 1.2 |
| | IBM High Availability Cluster Multi-Processing / Enhanced Scalability for AIX | 5765-A86 | 4.2 |
| | IBM High Availability Cluster Multi-Processing for AIX | | |
| | IBM Recoverable Virtual Shared Disk | 5765-646 | 2.1 |
| | | 5765-444 | 1.2 |
| | NetTAPE | 5765-637 | 1.2 |
| | NetTAPE Tape Library Connection | 5765-643 | 1.2 |
| | General Parallel File System for AIX | 5765-B95 | 1.1 |
| | Interactive Network Dispatcher for AIX | 5765-B67 | 1.2 |
| | | | |
| **Note:** Add other AIX program products or compilers that you expect to use. | | | |

*Table 43. External Disk Storage Needs*

| **External Disk Storage - Worksheet 3** | | |
|---|---|---|
| **Amount** | **External Disk Subsystem** | **Disk Space (MB)** |
| | 7133 (4.5 to 72GB) | |
| | 7134 (4.5 to 72GB) | |
| | 7135 (4.5 to 135GB) | |
| | 7137 (4 to 33GB) | |

*Table 44. Overall System Information*

| **SP Planning - Worksheet 4** | | |
|---|---|---|
| **Customer Name:**<br>**Customer Number:**<br>**Customer Contact:**<br>**IBM Contact:** | **Date:**<br><br>**Phone:**<br>**Phone:** | |
| *Complete the following by entering quantities to order:* | | |
| **Frames** | **Nodes** | **Nodes** |
| 500 (short): | 120 MHz Thin: | 200 MHz High: |
| 550 (tall): | 160 MHz Thin: | 135 MHz Wide: |
| | 332 MHz Thin: | 332 MHz Wide: |
| **SP Switches** | | |
| 8-port: | 16-port: | |
| SP Switch Router: | SP Switch Router Adapter: | |
| **External Storage Units:** | *Type* | *Quantity* |
| | | |
| | | |
| | | |
| **Network Media Cards:** | *Type* | *Quantity* |
| | | |
| | | |
| | | |
| *Fill in the remainder of this chart after you place your order.* | | |
| | *System Number* | *Purchase Order Number* |
| RS/6000 SP: | | |
| Control Workstation: | | |
| Peripherals: | | |
| | | |
| | | |
| | | |

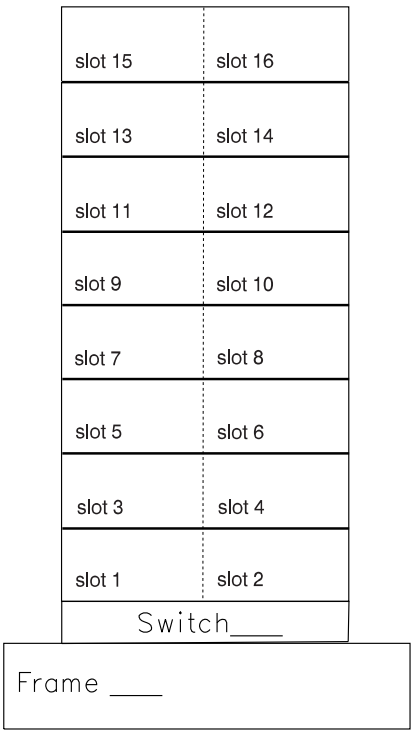



*Figure 55. SP Node Layout Worksheet For One Frame.*

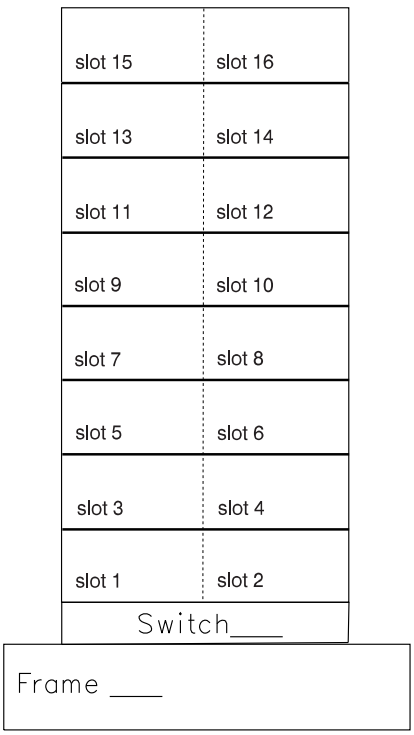



*Figure 56. Extra SP Node Layout Worksheet For One Frame*

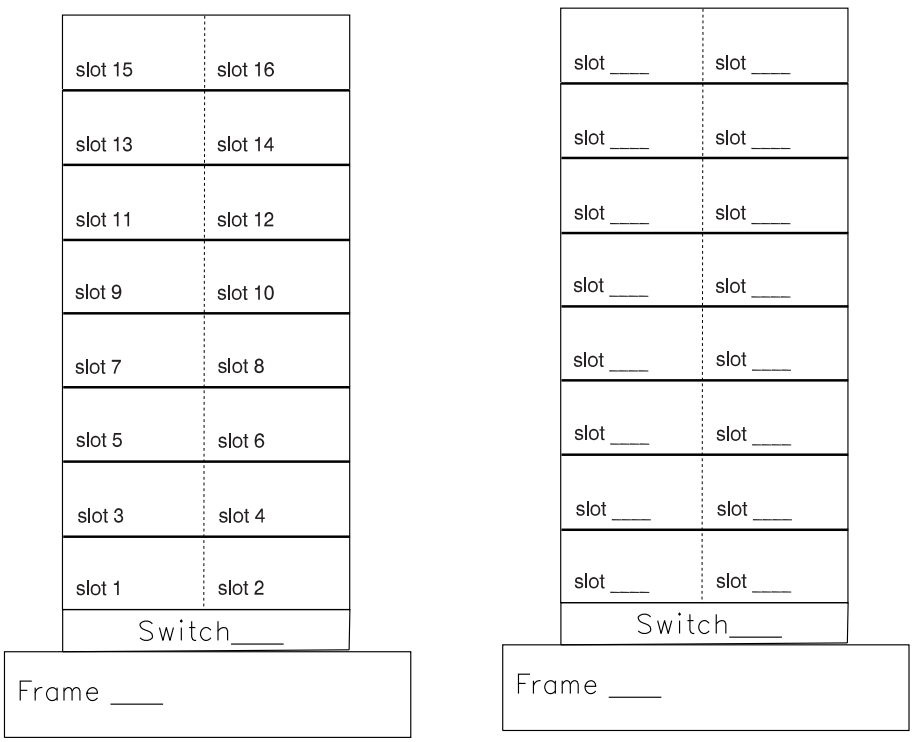



*Figure 57. Node Layout Worksheet For Two Frames*

*Table 45. Hardware Configuration By Node*

| **SP Hardware Configuration by Node - Worksheet 7** | | | | | | |
|---|---|---|---|---|---|---|
| **Frame Number _____ Switch Number _____** | | | | | | |
| **Slot Number** | **Node Number** | **Node Type** | **Processor Memory** | **Internal Disk** | **L2 Cache** | **Adapters** |
| Slot 1 | | | | | | |
| Slot 2 | | | | | | |
| Slot 3 | | | | | | |
| Slot 4 | | | | | | |
| Slot 5 | | | | | | |
| Slot 6 | | | | | | |
| Slot 7 | | | | | | |
| Slot 8 | | | | | | |
| Slot 9 | | | | | | |
| Slot 10 | | | | | | |
| Slot 11 | | | | | | |
| Slot 12 | | | | | | |
| Slot 13 | | | | | | |
| Slot 14 | | | | | | |
| Slot 15 | | | | | | |
| Slot 16 | | | | | | |

*Table 46. SP Ethernet Node Network Configuration*

**SP Ethernet Node Network Configuration - Worksheet 8A**

**Company Name**_____
**Date**_____
**Frame Number**_____
**Token Ring Speed**_____

| **Slot** | **SP Ethernet** *en0 adapters)* **Netmask**_____ | | **Default Route** |
|---|---|---|---|
| | **Hostname (note 1)** | **IP Address** | |
| 1 | | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| 9 | | | |
| 10 | | | |
| 11 | | | |
| 12 | | | |
| 13 | | | |
| 14 | | | |
| 15 | | | |
| 16 | | | |

**Notes:**

1. AIX is case sensitive. Use lower case for the hostname and addresses.
2. Wide nodes occupy two frame slots and use the *odd-numbered* slot number.
3. High nodes occupy four frame slots (2 drawers) and use the lowest odd-numbered slot number.

*Table 47. SP Additional Adapters Node Network Configuration*

**SP Additional Adapters Node Network Configuration - Worksheet 8B**

**Company Name**_____
**Date**_____
**Frame Number**____
**Token Ring Speed**___

| Slot | Additional Adapter Netmask_____ | | | Default Route |
|---|---|---|---|---|
| | Adapter Name | Hostname (note 1) | IP Address | |
| 1 | | | | |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| 9 | | | | |
| 10 | | | | |
| 11 | | | | |
| 12 | | | | |
| 13 | | | | |
| 14 | | | | |
| 15 | | | | |
| 16 | | | | |

**Notes:**

1. AIX is case sensitive. Use lower case for the hostname and addresses.
2. Wide nodes occupy two frame slots and use the *odd-numbered* slot number.
3. High nodes occupy four frame slots (2 drawers) and use the lowest odd-numbered slot number.

*Table 48. Switch Configuration Worksheet*

| **Switch Configuration -Worksheet 9** | | | |
|---|---|---|---|
| **Frame Number ____ Switch Number ____ Netmask _____** | | | |
| Slot Number | Switch Node Number | Hostname | IP Address |
| Slot 1 | | | |
| Slot 2 | | | |
| Slot 3 | | | |
| Slot 4 | | | |
| Slot 5 | | | |
| Slot 6 | | | |
| Slot 7 | | | |
| Slot 8 | | | |
| Slot 9 | | | |
| Slot 10 | | | |
| Slot 11 | | | |
| Slot 12 | | | |
| Slot 13 | | | |
| Slot 14 | | | |
| Slot 15 | | | |
| Slot 16 | | | |

*Table 49. PCI Adapters Supported on 332 MHz Nodes*

| **PCI Adapters Supported on 332 MHz Nodes - Worksheet 10A** | | | | | | |
|---|---|---|---|---|---|---|
| √ | **PCI Adapter Name** | **Feature Code** | **Max Per Wide Node** | **Max Per Thin Node** | **Number of Slots Required** | **AIX 4.2.1/4.3** |
| | FDDI SK-NET LP SAS | 2741 | 4 | 2 | 1 | yes |
| | FDDI SK-NET LP AS | 2742 | 4 | 2 | 1 | yes |
| | FDDI SK-NET UP SAS | 2743 | 4 | 2 | 1 | yes |
| | WAN RS232 128-port | 2944 | 7 | 2 | 1 | yes |
| | Ethernet 10/100 MB | 2968 | 4 | 2 | 1 | yes |
| | Token Ring Auto Lanstream | 2979 | 8 | 2 | 1 | yes |
| | Ethernet 10 MB BNC | 2985 | 8 | 2 | 1 | yes |
| | Ethernet 10 MB AUI | 2987 | 8 | 2 | 1 | yes |
| | ATM 155 MMF | 2988 | 4 | 2 | 1 | yes |
| | SCSI-2 Single-Ended | 6208 | 9 | 2 | 1 | yes |
| | SCSI-2 Differential | 6209 | 9 | 2 | 1 | yes |
| | SSA RAID EL | 6215 | 6 | 2 | 1 | yes |
| | SSA Fast-Write Cache Option | 6222 | Mounts on F/C 6215 | | 0 | yes |

*Table 50 (Page 1 of 2). MCA Adapters Supported on all but 332 MHz Nodes*

| MCA Adapters Supported on all but 332 MHz Nodes - Worksheet 10B | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| √ | **Adapter** | **Feature code** | **Wide node [1] quantity per node** | **Thin node [2] quantity per node** | **High node [3] quantity per node** | **Number of slots Required** | **AIX 4.1** | **AIX 4.2** |
| | Internal Ethernet | Standard | N/A | 1 | N/A | 0 | yes | yes |
| | FCS Dwtr | 1902 [7/8/11] | 0 - 2 | 0 -1 | N/A | 0 | yes | yes |
| | FCS 1GB | 1904 [8/11] | N/A | N/A | N/A | 1 | 4.1.4 | no |
| | NetW TA 256 | 2402 | 0 - 7 | 0 - 4 | 0 - 4 | 1 | no | yes |
| | NetW TA 2048 | 2403 | 0 - 7 | 0 - 4 | 0 - 4 | 1 | no | yes |
| | SCSI-2 Ext I/O | 2410 | 0 - 7 | 0 - 4 | N/A | 1 | 4.1.4 | yes |
| | SCSI Turbo | 2412 | 0 - 7 | 0 - 4 | 0 -14 | 1 | 4.1.3 | yes |
| | SCSI F/W DIF | 2415 | 0 - 7 | 0 - 4 | 1 -14 | 1 | 4.1.1 | yes |
| | SCSI F/W DIF | 2416 | 0 - 7 | 0 - 4 | 0 - 14 | 1 | 4.1.1 | yes |
| | SCSI EXT I/O | 2420 | 0 - 7 | 0 - 4 | N/A | 1 | 4.1.4 | yes |
| | 4 port Multi Comm | 2700 | 0 - 7 | 0 - 3 | 0 - 8 | 1 | 4.1.1 | yes |
| | FDDI D/R | 2723 [4] | 0 - 3 | 0 - 2 | 0 - 8 | 1 | 4.1.1 | yes |
| | FDDI S/R | 2724 | 0 - 6 | 0 - 2 | 0 - 8 | 1 | 4.1.1 | yes |
| | HIPPI[5/6] | 2735 | 0 - 1 | N/A | 0 - 2[6] | 5[5] | 4.1.4 | yes |
| | ESCON Chan Em. | 2754 | 0 - 2 | 0 - 1 | 0 - 4 | 2 | 4.1.4 | yes |
| | BMCA | 2755 | 0 - 2 | 0 - 2 | 0 - 2 | 1 | 4.1.4 | yes |
| | ESCON CNTRL | 2756 | 0 - 2 | 0 - 1 | 0 - 4 | 2 | 4.1.4 | yes |
| | 8-port async 232 | 2930[9] | 0 - 7 | 0 - 4 | 0 - 14 | 1 | 4.1.1 | yes |
| | 8-port async 422 | 2940[9] | 0 - 7 | 0 - 4 | 0 - 14 | 1 | 4.1.1 | yes |
| | X.25 inter co-p | 2960 | 0 - 7 | 0 - 4 | 0 - 8 | 1 | 4.1.3 | yes |
| | Token Ring | 2970 | 0 - 7 | 0 - 4 | 0 - 12 | 1 | 4.1.1 | yes |
| | Token Ring | 2972 | 0 - 7 | 0 - 3 | 0 - 12 | 1 | 4.1.1 | yes |
| | Ethernet | 2980 | | 0 - 3 | 1 - 8 | 1 | 4.1.1 | yes |
| | ATM 100 | 2984 | 0 - 2 | 0 - 2 | 0 - 2 | 1 | 4.1.4 | yes |
| | ATM 155 | 2989 [8] | 0 - 4 | 0 - 2 | 0 - 4 | 1 | 4.1.4 | yes |
| | Ether TP | 2992 | 0 - 7[13] | 0 - 3 | N/A | 1 | 4.1.4 | yes |
| | Ether BNC | 2993 | 0 - 7[13] | 0 - 3 | N/A | 1 | 4.1.4 | yes |
| | Ether BNC | 2994 | x | x | — | 1 | 4.1 | — |
| | SPSw (TB3) | 4020 | 0 - 1 | 0 - 1 | 0 - 1 | 1 | 4.1.1 | yes |
| | Enet 10baseT | 4224 | 0 -8 | 0 - 4 | 0 - 15 | 0 | 4.1.1 | yes |
| | HI-P Subsys | 6212 | 0 - 4 | 0 - 2 | 0 - 8[10] | 1 | 4.1.1 | yes |
| | SSA | 6214 | 0 - 4 | 0 - 2 | 0 - 8[10] | 1 | 4.1.4 | yes |
| | SSA | 6216 [8] | 0 - 4 | 0 - 2 | 0 - 8[10] | 1 | 4.1.4 | yes |
| | SSA 4RD | 6217 | 0 - 4 | 0 - 2 | 0 - 8 | 1 | 4.1.5 | yes |
| | SSA RAID EL | 6219 | 0 - 4 | 0 - 2 | 0 - 8 | 1 | 4.1.5 | yes |

*Table 50 (Page 2 of 2). MCA Adapters Supported on all but 332 MHz Nodes*

**MCA Adapters Supported on all but 332 MHz Nodes - Worksheet 10B**

| √ | Adapter | Feature code | Wide node [1] quantity per node | Thin node [2] quantity per node | High node [3] quantity per node | Number of slots Required | AIX 4.1 | AIX 4.2 |
|---|---|---|---|---|---|---|---|---|
| | Fast Write Cache Option | 6222 | Mounts on FC 6219 | | | | | |
| | Digital Truck | 6305 | 0 - 6 | 0 - 3 | 0 - 2 | 1 | 4.1.1 | yes |
| | Prtmstr 1MB | 7006[12] | 0 - 7 | 0 - 4 | 0 - 8 | 1 | 4.1.1 | yes |
| | 128 prt Cntrl | 8128 | 0 - 7 | 0 - 7 | 0 - 7 | 1 | 4.1.1 | yes |

**Note:**

[1]: There are a total of 7 MCA slots available per wide node.
[2]: There are a total of 4 MCS slots available per thin node.
[3]:There are a total of 16 MCA slots per high node.
[4]: FDDI D/R adapters (F/C 2723) have a mandatory prerequisite of FDDI S/R adapters (F/C 2724)
[5]: The HIPPI feature uses 3 physical MCA slots and requires a total of 5 MCA slots to satisfy power and thermal requirements.
[6]: HIPPI cannot be populated across the 2 micro channel bus on high nodes.
[7]: FCS Daughter card F/C 1902 does not require a micro channel slot.
[8]: These adapters are not supported on any 62MHz node.
[9]: This adapter has a co-requisite of 2995 feature cable.
[10]:The SSA Adapters in a high node are limited to a total count of 8 in any combination
[11]:1902, 1904, 1906 FCS adapters are not supported in the P2SC nodes which are the 135MHz wide nodes (F/C 2007) , and the 120MHz thin nodes (F/C 2008), nor are they supported on the SMP high nodes (F/C 2006 ).
[12]:7006 portmaster card requires the selection of 7042, 7044, 7046, or 7048.
[13]:The maximum of 2992 and 2993 in any combination is 8.

**Note:**

- If you order the SP Switch, you must order F/C 4020 for every node.
- You cannot have a combination of High Performance Switch and SP Switch adapters.

| *Table 51. 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 wish to install rootvg** | |
| **Check here if you want only the SPIMG minimal image ____** | |
| **IBM Licensed Products** | |
| | AIX |
| | PSSP |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| **Additional AIX Software** | |
| | |
| | |
| | |
| | |
| | |
| **Other Applications** | |
| | |
| | |
| | |
| | |
| | |
| | |

*Table 52. File Set List for PSSP 2.4*

| PSSP 2.4 File Sets - Worksheet 12 | | |
|---|---|---|
| System Image Name _____ | | |
| √ | **File Set** | **Description** |
| | ssp.st | Application programming interface for loading, unloading, and querying the job switch resource table. |
| | ssp.ha | Availability subsystems which include heartbeat, Group Services, and Event Management. |
| | ssp.perlpkg | Perl4 and Perl5 |
| | ssp.pman | Problem management |
| | ssp.clients | SP Authenticated Client Commands |
| | ssp.basic | SP System Support Package |
| | ssp.css | SP Communication Subsystem Package (only if switch installed) |
| | ssp.sysman | Optional System Management Programs |
| | ssp.sysctl | SP Sysctl Package |
| | ssp.authent | SP Authentication Server |
| | ssp.public | Public Code Compressed Tarfiles |
| | ssp.docs | PostScript, man pages, and HTML files for PSSP documentation |
| | ssp.gui | SP System Monitor Graphical User Interface and SP Perspectives |
| | ssp.jm | SP Resource Manager Package |
| | ssp.top | SP System Partition Support |
| | ssp.csd.vsd | IBM Virtual Shared Disk Package |
| | ssp.csd.cmi | SP Centralized Management Interface Package for IBM Virtual Shared Disk |
| | ssp.csd.hsd | IBM Virtual Shared Disk Data Striping package |
| | ssp.csd.gui | Perspectives for IBM Virtual Shared Disk |
| | ssp.csd.sysctl | IBM Virtual Shared Disk Usability Improvement |
| | ssp.hacws | High Availability Control Workstation Support |
| | ssp.ptpegui | PTPE graphical user interface |
| | ssp.topsvcs | Topology services |
| | spimg | Contains a single file with the mksysb image of a minimal AIX 4.2 system. |
| | ssp.top.gui | Graphical user interface for System Partitioning Aid |
| | ssp.spmgr | Extension Node SNMP Manager Support |
| | ssp.ucode | SP supervisor microcode package |
| **Note:** For information on whether these file sets are installed on the control workstation and the node, refer to chapter 2 of the *Installation and Migration Guide* | | |

*Table 53. Control Workstation Image Worksheet*

| **SP Control Workstation - Worksheet 13** | | |
|---|---|---|
| **SP 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 16MB | |
| | Multiport Serial Adapters | |
| | 8 mm tape drive | |
| | CD-ROM | |
| **IBM Licensed Products:** | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| **Other Applications:** | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Table 54. Time Zones*

| Select a Time Zone - Worksheet 14 | | | | |
|---|---|---|---|---|
| **Select using a √** | **Time Zone** | **Select using a √** | **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 55. Control Workstation Network Worksheet*

**SP Control Workstation Network - Worksheet 15**

**Name:**
**Date:**
**System Name:**
**Control Workstation Name:**

| Frame Hardware Control Connections (RS-232) | | | | Control Workstation Network Connections (note 2) | | | |
|---|---|---|---|---|---|---|---|
| **Frame Number** | **Serial Port for RS-233 Control Line (note 1)** | **tty Device** | | **Adapter** | **Hostname** | **IP Address** | **Netmask** |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

**Notes:**

1. Use the SMIT **tty** menu or the **mkdev** command to configure these ports.
2. Use the SMIT **mkinet** menu or the **mkdev** command to configure your SP Ethernet connection.

*Table 56. Site Environment Worksheet*

| SP Site Environment - Worksheet 16 | | | |
|---|---|---|---|
| **Name** _____ | | **Date** _____ | |
| **System Name** _____ | | **Control Workstation Name** _____ | |
| **SMIT Dialog Field Name** [a] | **Site Attribute** [b] | **Default Value** | **Your Choice** |
| Default Network Install Image | install_image | bos.obj.ssp.421 | |
| Remove Install Image After Installs | remove_image | false | |
| NTP Installation | ntp_config | consensus | |
| NTP Server Hostname | ntp_server | | |
| NTP Version | ntp_version | 3 | |
| Amd Configuration | amd_config | true | |
| Print Management Configuration | print_config | false | |
| Print System Secure Mode Login Name | print_id | | |
| User Administration Interface | usermgmt_config | true | |
| Password File Server Hostname | passwd_file_loc | control workstation hostname | |
| Password File Location | passwd_file | /etc/passwd | |
| Home Directory Server Hostname | homedir_server | control workstation hostname | |
| Home Directory Path | homedir_path | /home/<control workstation> | |
| File Collection Management | filecoll_config | true | |
| Home Collection Daemon uid | supman_uid | 102 | |
| Home 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 | |
| Control Workstation LPP source name | cw_lppsource_name | default | |

**a.** This is the name that appears on the SMIT dialog. **b.** This is the attribute name to use on the **spsitenv** command.

| *Table 57. PSSP Or Other Kerberos Authentication Servers* | | | | |
|---|---|---|---|---|
| **Authentication - Worksheet 17** | | | | |
| | **Hostname (long)** | **Default realm** | **Control workstation** | **SP Authentication?** |
| Primary Server | | | | |
| Secondary Servers | | | | |
| | | | | |
| | | | | |
| Client Systems | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

**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.

**SP Authentication**

`y` This workstation will run an SP authentication server.

`n` This workstation has a different MIT Kerberos 4 server installed.

This option is not applicable to client systems. You must install **ssp.authent** on all workstations for which you enter **y**.

| Table 58. Local Realm Information, PSSP Authentication Server | | |
|---|---|---|
| **Local realm name** | | **Master password** |
| **Administrative principal** | .admin | Password |
| **other principals** | name | password |
| | name | password |
| | name | password |
| | name | password |
| | name | password |
| | name | password |
| | name | password |
| | name | password |

| Table 59. AFS Authentication Servers | |
|---|---|
| **administrative principal** | |
| **Password** | |
| **Directory containing CellServDB, ThisCell files** | |
| **Directory containing kas command** | |

**local realm** The name of your local realm.

If blank, the local realm is the default realm you entered for the primary server.

**administrative principal** The name you will use as the primary administrator of the authentication database.

**password** The master password of the primary authentication server using SP authentication. Once you have written this password on the chart, be sure to keep the chart in a secure environment.

# Glossary

## Terms and Abbreviations

This glossary includes terms and definitions from:

- The *IBM Dictionary of Computing*, New York: McGraw-Hill, 1994.
- The *American National Standard Dictionary for Information Systems*, ANSI X3.172-1990, copyright 1990 by the American National Standards Institute (ANSI). Copies can be purchased from the American National Standards Institute, 1430 Broadway, New York, New York 10018. Definitions are identified by the symbol (A) after the definition.
- The *ANSI/EIA Standard - 440A: Fiber Optic Terminology*, copyright 1989 by the Electronics Industries Association (EIA). Copies can be purchased from the Electronic Industries Association, 2001 Pennsylvania Avenue N.W., Washington, D.C. 20006. Definitions are identified by the symbol (E) after the definition.
- The *Information Technology Vocabulary* developed by Subcommittee 1, Joint Technical Committee 1, of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC JTC1/SC1). Definitions of published parts of this vocabulary are identified by the symbol (I) after the definition; definitions taken from draft international standards, committee drafts, and working papers being developed by ISO/IEC JTC1/SC1 are identified by the symbol (T) after the definition, indicating that final agreement has not yet been reached among the participating National Bodies of SC1.

The following cross-references are used in this glossary:

> **Contrast with.** This refers to a term that has an opposed or substantively different meaning.
> **See.** This refers the reader to multiple-word terms in which this term appears.
> **See also.** This refers the reader to terms that have a related, but not synonymous, meaning.
> **Synonym for.** This indicates that the term has the same meaning as a preferred term, which is defined in the glossary.

This section contains some of the terms that are commonly used in the SP publications.

IBM is grateful to the American National Standards Institute (ANSI) for permission to reprint its definitions from the American National Standard *Vocabulary for Information Processing* (Copyright 1970 by American National Standards Institute, Incorporated), which was prepared by Subcommittee X3K5 on Terminology and Glossary of the American National Standards Committee X3. ANSI definitions are preceded by an asterisk (*).

Other definitions in this glossary are taken from *IBM Vocabulary for Data Processing, Telecommunications, and Office Systems* (SC20-1699) and *IBM DATABASE 2 Application Programming Guide for TSO Users* (SC26-4081).

© Copyright IBM Corp. 1997, 1998

# A

**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.

**Amd**. Berkeley Software Distribution automount daemon.

**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 or used by an application program.

**ARP**. Address Resolution Protocol.

**ATM**. Asynchronous Transfer Mode. (See *TURBOWAYS 100 ATM Adapter*.)

**authentication**. The process of validating the identity of a user or server.

**authorization**. The process of obtaining permission to perform specific actions.

# 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.

**BMCA**. Block Multiplexer Channel Adapter. The block multiplexer channel connection allows the RS/6000 to communicate directly with a host System/370 or System/390; the host operating system views the system unit as a control unit.

# C

**call home function**. The ability of a system to call the IBM support center and open a PMR to have a repair scheduled.

**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.

**Client Input/Output Sockets (CLIO/S)**. An IBM software package that enables high-speed data and tape access between SP systems, AIX systems, and ES/9000 mainframes.

**CLIO/S**. Client Input/Output Sockets.

**CMI**. Centralized Management Interface provides a series of SMIT menus and dialogues used for defining and querying the SP system configuration.

**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.

# 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**. An action that takes place in the operation of the system software.

**expect**. Programmed dialogue with interactive programs.

# 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 fall back, 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.

**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/or concentrators.

**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.

**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 server**. A centrally located computer that acts as a storehouse of data and applications for numerous users of a local area network.

**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/6000**. AIX High Availability Cluster Multi-Processing/6000.

**HACWS**. High Availability Control Workstation function, based on HACMP/6000, provides for a backup control workstation for the SP system.

**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/6000**. An IBM facility to cluster nodes or components to provide high availability by eliminating single points of failure.

**High Performance Switch**. An IBM multi-stage packet switch for high-performance communication between processor nodes.

**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.

**HSD**. The data striping device for the Virtual Shared Disk. The device driver lets application programs stripe data across physical disks in multiple Virtual Shared Disks, thus reducing I/O bottlenecks and hot spots.

# I

**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.

**Intermediate Switch Board**. Switches mounted in the High Performance Switch expansion frame.

**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.

**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.

**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.

# 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.

**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.

# 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.

**Motif**. The graphical user interface for OSF, incorporating the X-Window System. Also called OSF/Motif.

**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.

**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**. Network Installation Management is provided with AIX 4.1 to install AIX on the nodes.

**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.

**Node Switch Board**. Switches mounted in frames that contain nodes.

**NSB**. Node Switch Board.

**NTP**. Network Time Protocol.

# P

**parallel environment**. A system environment where message passing or SP resource manager services are used by they application.

**Parallel Environment**. A licensed IBM program used for message passing applications on the SP or RS/6000 platforms.

**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.

**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.

**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.

**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.

**PMR**. Problem Management Report.

**Problem Management Report**. The number in the IBM support mechanism that represents a service incident with a customer.

**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 High Performance Switch that initializes, provides diagnosis and recovery services, and performs other operations to the switch network. (3) In IBM Virtual Shared Disk function, the node at which the logical volume is actually local and that acts as server node to I/O requests from other nodes.

# R

**RAID**. Redundant array of independent disks.

**remote host**. *See foreign host*.

**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.

**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.

**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.

**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.

**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 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.

**SP Switch**. Newest version of the IBM High Performance Switch.

**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.

**SUP**. Software Update Protocol.

**Sysctl**. Secure System Command Execution Tool. An authenticated client/server system for running commands remotely and in parallel.

**System Administrator**. The user who is responsible for setting up, modifying, and maintaining the SP system.

**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.

**syslog**. A BSD logging system used to collect and manage other subsystem's logging data.

**system partition**. A group of nodes that act as a logical SP system. All nodes attached to the same switch chip must be in the same system partition.

# 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.

**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.

**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.

**Uninteruptable Power Supply (UPS)**. A UPS can supply electricity to a device to keep it running when main power is interrupted or is unreliable.

**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.

**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.

**user**. Anyone who requires the services of a computing system.

**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**. 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.

# Index


## A

ABC Corporation, used in examples 11
about this book xiii
accounting choices 62
acct_master 62
adapters 40
adapters supported 236
adapters, network connectivity 6
AFS
  authentication servers, choosing 126
AIX 7
AIX 4.2, new function 7
AIX 4.3, new function 8
AIX level selection 22
application planning worksheet 228
applications, preliminary list 17
audience of this book xiv
authentication
  administrative principals 119
  AFS servers, choosing 126
  choosing a configuration 119
  configurations 119
  creating configuration files 124
  deciding on realms 123
  planning checklists 125
  principals of 117
  selecting options to install 125
  servers 70
  setting up configurations 119
  SP authentication services 117
  worksheet 244
automount daemon 59
availability requirements 29

## B

backup control workstation 93
boot/install server 64, 150
boot/install server for AIX 3.2.5 107

## C

choosing a switch
  SP Switch, benefits 13
Client Input Output/Sockets 19
Client Input Output/Sockets (CLIO/S) 143
CLIO/S 170
commands
  dsh 129
configuration planning 55
connectivity adapters, network 6
connectivity, network 23
control workstation 6
  configuration decisions 92
  disk mirroring 92
  failure scenario 91
  function with High Availability Control Workstation 91
  hardware requirements 46
  maintenance with High Availability Control Workstation 92
  minimum hardware requirements 49
  planning for 87
  planning for a backup 93
  planning for High Availability Control Workstation 91
  planning site environment 55
  reliability 92
  requirements 46
  single point of failure 90
  worksheet 240
control workstation network workstation 240
control workstation system images worksheet 240
control workstation worksheet 241

## D

data access across system partitions 100, 106
decisions to make 11
default (persistent) system partitions 99
defining the system 11
directory structure, system partitions 113
disk mirroring 92
disk space
  installation image requirements 72
  lppsource 71
  system programs 26
  users' home directories 26
disk storage 25

## E

EMEA Service Planning 134
environment variable
  SP_NAME 107
error messages, finding and using 131
estimate the installation image requirements 72
ethernet 74
expansion frames 150
extension node 31
extension node support 165
external disk storage needs worksheet 229


© Copyright IBM Corp. 1997, 1998

external disk storage worksheet 28

## F

fault tolerance definition 87
filecoll-config 62
finding and using error messages 131
frame
  frame description 5
  frame numbers 81
  non-switched expansion frames 81
frame supervisor changes 94
future expansion, network install server 68

## G

General Parallel File System 169
General Parallel File System (GPFS) 143
General Parallel File System for AIX 21
GPFS 21
Group Services API 165

## H

HACWS 29
hard disk choices for nodes 70
hardware configuration by node worksheet 231
hardware overview 4
hardware requirements
  control workstation 46, 49
help
  getting from IBM 131
High Availability Cluster Multi-Processing 20, 165
High Availability Control Workstation
  control workstation maintenance 92
  description 29
  failure scenario with High Availability Control Workstation 91
  minimum requirements 50
  new function 29
  no loss of control workstation function 91
  ordering 89
  planning 91
  resource manager limitation 93
  system stability 92
  time services considerations 57
  worksheet 93
High Availability Control Workstation changes to the control workstation
  frame supervisor changes 94
high availability definition 87
high node 4, 31
high performance switch 6, 13, 75
  worksheet 233
home directory server planning 70
homedir_path 61
homedir_server 61
host name 37

## I

IBM C for AIX, V3, requirement 18
IBM parallel system support programs for AIX (PSSP) 7
IBM Program Products worksheet 228
IBM, getting help from 131
installation image requirements 72
installation worksheets
  network install image choices 56
  time service choices 56
installation, planning for 55
installp image requirements 72
IP address assignments to nodes 79
IP addresses 74, 75, 107
IP over the high performance switch 75

## K

kernel-to-kernel interface 135

## L

large scale configurations, network install server 68
listing your applications 17
LoadLeveler 19, 140, 170
location of customer data 69
lpp source directory name choices 63
lppsource
  disk space requirements 71

## M

management of system partitions 106
manual pages for public code xxi
messages, finding and using 131
migration
  planning 157
migration and coexistence limitations 163
  CLIO/S 170
  extension node support 165
  General Parallel File System 169
  Group Services API 165
  High Availability Cluster Multi-Processing 165
  LoadLeveler 170
  NetTAPE 170
  Parallel Environment 169
  perfagent 171
  PIOFS 170
  Recoverable Virtual Shared Disk 167
  switch support 164
  Virtual Shared Disk 166

minimum requirements, High Availability Control Workstation 50
Motif 87
multiple frame systems, network install server 66
multiple production environments 100

# N

NetTAPE 19, 142, 170
NetTAPE Tape Library Connection 19
NetTAPE TLC 19
network connectivity 23
network connectivity adapters 6
network install image choices 56
network install image choices, worksheet 56
network install server planning
  future expansion 68
  large scale configurations 68
  multiple frame systems 66
  single frame systems 65
network planning 74
Network Tape Access and Control System 19
network time protocol (NTP) 57
networking considerations for partitioning 107
new function
  High Availability Control Workstation 29
node
  determining how many nodes needed 30
  high nodes 32
  node configuration worksheet 38
  node hard disk choices 70
  node layout worksheet instructions 34
  node numbering 79, 83
  node slot 109
  thin nodes 32
  wide nodes 32
node layout worksheet for one frame 229
node layout worksheet for two frames 230
nodes, processor 4
non-disruptive management 100
non-switched expansion frames 80, 81, 151
numbering nodes 79
numbering switch nodes 79

# O

ordering the High Availability Control Workstation 89
other installp image requirements 72
overall system view of a High Availability Control Workstation 87
overview
  hardware 4
  software 6

# P

parallel computing 12
Parallel Engineering and Scientific Subroutine Library 20
parallel environment 19, 169
Parallel Environment for AIX 138
Parallel ESSL 20
Parallel I/O File System for AIX 20
Parallel Optimization Subroutine Library 20
Parallel System Support Programs for AIX (PSSP), IBM 7
partitions
  benefits 100
  change management 100
  data access 106
  default system partition 99
  description 99
  multiple production environments 100
  networking considerations 107
  single point of control 106
  switchless systems 104
  System Partitioning Aid 105
  understanding the switch board 102
passwd_file 61
PE 19
perfagent 171
Performance Toolbox Parallel Extension (PTPE) 139
Performance Toolbox Parallel Extensions for AIX 21
PIOFS 20, 142, 170
planning
  for High Availability Control Workstation 91
  for installation and configuration 55
  migration 157
  network 74
  partitions 106
  questions to ask 11
  server 64
  site environment 55
planning for
  CLIO/S 143
  GPFS 143
  High Availability Cluster Multi-Processing (HACMP) 140
  High Availability Cluster Multi-Processing Enhanced Scalability (HACMP/ES) 140
  LoadLeveler 140
  NetTAPE 142
  Performance Toolbox Parallel Extension (PTPE) 139
  PIOFS 142
planning for security 117
power independence 92
preloaded SP or default version 14
prerequisite knowledge for this book xiv

priced feature
Performance Toolbox Parallel Extensions for AIX 21
print management choices 60
problem management record (PMR) 132
problem resolution
EMEA Service Planning 134
Service Director for RS/6000 132
processor nodes 4
programs, related IBM 18
PSSP 7
PSSP 1.2 and 2.1 components list worksheet 238
PSSP 2.4, new function 8
PSSP print subsystem 60
PTPE 21

# Q

Question
1. why do you need an SP? 11
10. what do you need for your control workstation? 46
2. Do you want a preloaded SP or the default version? 14
3. what related IBM program products do you need? 18
4. what levels of AIX do you need? 22
5. what type of network connectivity do you need? 23
6. what are your disk storage requirements? 25
7. what are your reliability and availability requirements? 29
8. how many nodes do you need? 30
9. defining your system images 43
questions for planning decisions 11

# R

Recoverable Virtual Shared Disk 20, 135, 167
reference rate of customer data 69
related program products 18
related programs
Client Input Output/Sockets 19
General Parallel File System for AIX 21
High Availability Cluster Multi-Processing 20
LoadLeveler 19
NetTAPE Tape Library Connection 19
Network Tape Access and Control System 19
Parallel Engineering and Scientific Subroutine Library 20
parallel environment 19
Parallel I/O File System for AIX 20
Parallel Optimization Subroutine Library 20
Recoverable Virtual Shared Disk 20
reliability requirements 29
requirements
availability 29
control workstation requirements 46
IBM C for AIX, V3 18
reliability 29
resource manager
High Availability Control Workstation limitation 93

# S

SDR and system partitions 107
security
planning 117
sending problem data to IBM 131
server planning 64
server planning, home directory 70
servers, authentication 70
Service Director for RS/6000 132
single frame systems, network install server 65
single point of control with system partitions 106
single point of failure 90
site environment choices 56, 57, 59, 60, 61, 62, 63
site environment planning 55
site environment worksheet 241
slot numbers 79
SMIT 55
software migration 157
software overview 6
SP specific logs
table 130
SP Switch 5
SP Switch, benefits 13
SP system planning worksheet 229
SP_NAME environment variable 107
spacct_actnode_thresh 62
spacct_enable 62
spacct_exclusive_enable 62
spchuser command 61
home attribute 61
spmkuser command 61
home attribute 61
spsitenv 55
supfiesrv_port 62
supman_uid 62
switch
description of 5
switch, incompatibility 13
switch node numbering 79, 83
switch numbers 81
switch port numbering 83
switch support 164
switch, high performance 6, 13
switch, SP 5, 13
switchless systems
system partitions partition 104

system definition 11
system file management choices 62
system images 43, 46
system partitions 29
  benefits 100
  boot/install server requirements 64
  boot/install servers for AIX 3.2.5 107
  change management 100
  data access 100, 106
  default (persistent) system partitions 99
  directory structure 113
  management of a system 106
  multiple production environments 100
  overview 99
  switchless systems 104
  the SDR 107
system stability, High Availability Control Workstation 92

# T

thin node 30
time service choices 56, 57
time services considerations and High Availability Control Workstation 57
trademarks xi
tuning considerations 64

# U

understanding accounting 62
understanding lppsource name 63
understanding node hard disk choices 70
understanding user account management choices 60
unterruptable power supply 92
user account management choices 60, 61
user directory mounting choices 59
usermgmt-config 61
uses for an SP 11
uses for system partitions 100

# V

Virtual Shared Disk 135, 166
  kernel-to-kernel interface 135

# W

wide node 31
worksheet
  completing for High Availability Control Workstation 93
  external disk storage 28
worksheet entries
  Automount choices 58
  Automount choices, worksheet 58
worksheets
  adapters supported 236
  application planning 228
  authentication 244
  authentication planning 127
  control workstation network 240
  control workstation system images 240
  copying 11
  external disk storage needs 229
  hardware configuration by node 231
  High Performance Switch Configuration 233
  IBM Program Products 228
  node layout for one frame 229
  node layout for two frames 230
  node layout instructions 34
  PSSP 1.2 and 2.1 components list 238
  site environment 241
  SP Control Workstation Worksheet 241
  SP Node Configuration Worksheet 38
  SP system planning 229
workstation, control 6

# X

X-Windows 87

# Communicating Your Comments to IBM

IBM RS/6000 SP
Planning Volume 2, Control Workstation and
Software Environment

Publication No. GA22-7281-02

If you especially like or dislike anything about this book, please use one of the methods listed below to send your comments to IBM. Whichever method you choose, make sure you send your name, address, and telephone number if you would like a reply.

Feel free to comment on specific errors or omissions, accuracy, organization, subject matter, or completeness of this book. However, the comments you send should pertain to only the information in this manual and the way in which the information is presented. To request additional publications, or to ask questions or make comments about the functions of IBM products or systems, you should talk to your IBM representative or to your IBM authorized remarketer.

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.

If you are mailing an RCF from a country other than the United States, you can give the RCF to the local IBM branch office or IBM representative for postage-paid mailing.

- If you prefer to send comments by mail, use the RCF at the back of this book.
- If you prefer to send comments by FAX, use this number:
  - FAX: (International Access Code)+1+914+432-9405
- If you prefer to send comments electronically, use this network ID:
  - IBMLink: (United States customers only): KGNVMC(MHVRCFS)
  - IBM Mail Exchange: USIB6TC9 at IBMMAIL
  - Internet e-mail: mhvrcfs@vnet.ibm.com
  - World Wide Web: http://www.s390.ibm.com/os390

Make sure to include the following in your note:

- Title and publication number of this book
- Page number or topic to which your comment applies

Optionally, if you include your telephone number, we will be able to respond to your comments by phone.

# Reader's Comments — We'd Like to Hear from You

**IBM RS/6000 SP**
**Planning Volume 2, Control Workstation and**
**Software Environment**

**Publication No. GA22-7281-02**

You may use this form to communicate your comments about this publication, its organization, or subject matter, with the understanding that IBM may use or distribute whatever information you supply in any way it believes appropriate without incurring any obligation to you. Your comments will be sent to the author's department for whatever review and action, if any, are deemed appropriate.

**Note:** Copies of IBM publications are not stocked at the location to which this form is addressed. Please direct any requests for copies of publications, or for assistance in using your IBM system, to your IBM representative or to the IBM branch office serving your locality.

Today's date: _____

What is your occupation?

Newsletter number of latest Technical Newsletter (if any) concerning this publication:

How did you use this publication?

| | | | |
|---|---|---|---|
| [ ] | As an introduction | [ ] | As a text (student) |
| [ ] | As a reference manual | [ ] | As a text (instructor) |
| [ ] | For another purpose (explain) | | |

Is there anything you especially like or dislike about the organization, presentation, or writing in this manual? Helpful comments include general usefulness of the book; possible additions, deletions, and clarifications; specific errors and omissions.

Page Number: Comment:

Name

Address

Company or Organization

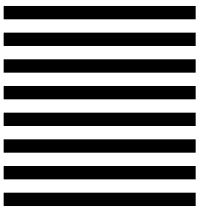
Phone No.

**Reader's Comments — We'd Like to Hear from You**
GA22-7281-02

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

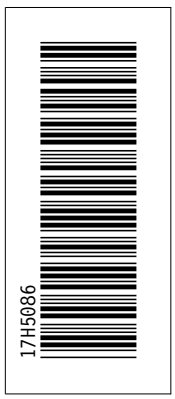
POSTAGE WILL BE PAID BY ADDRESSEE

IBM Corporation
Department 55JA, Mail Station P384
522 South Road
Poughkeepsie NY 12601-5400

Fold and Tape **Please do not staple** Fold and Tape



Cut or Fold
Along Line

Part Number: 17H5086
Program Number: 5765–529

Printed in the United States of America
on recycled paper containing 10%
recovered post-consumer fiber.