IBM Parallel System Support
Programs for AIX

Administration Guide

GC23-3897-04

IBM Parallel System Support
Programs for AIX

Administration Guide

GC23-3897-04

Note!

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

| **Fourth Edition (October 1997)**

| This is a major revision of GC23-3897-03.

| This edition applies to Version 2 Release 3 of IBM Parallel System Support Programs for AIX (5765-529), which runs on 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.

| This edition also contains information about support for the SP Switch Router Adapter that was added as a PTF to PSSP 2.3.

Order publications through your IBM representative or the IBM branch office serving your locality. Publications are not stocked at the address given 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 & 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.rs60000.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. 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 . . . xv
Trademarks . . . xv
Publicly Available Software . . . xvi

About This Book . . . xvii
Who Should Use This Book . . . xviii
How This Book Is Organized . . . xviii
Typographic Conventions . . . xxi
Interface Instructions . . . xxii
Accessing Online Information . . . xxii
Obtaining Documentation . . . xxii
Related Publications . . . xxiii
RS/6000 SP Publications . . . xxiii
Parallel System Support Programs for AIX Publications . . . xxiii
IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Publication . . . xxiii
Performance Monitor Publication . . . xxiii
General Parallel File System for AIX Publication . . . xxiii
LoadLeveler Publications . . . xxiv
Parallel Environment for AIX Publications . . . xxiv
PVMe for AIX Publications . . . xxiv
Client Input Output/Sockets (CLIO/S) Publications . . . xxiv
Parallel I/O File System Publications . . . xxiv
Network Tape Access and Control System for AIX (NetTAPE) Publications xxiv
Other IBM Publications . . . xxv
International Technical Support Organization Publications (Red Books) . . . xxv
AIX and RS/6000 Publications . . . xxv
Service . . . xxvi
Network Queueing System/MVS (NQS/MVS) . . . xxvi
Network Connectivity . . . xxvi
Adapters . . . xxvi
Non-IBM Publications . . . xxvii
Parallel Computing . . . xxvii
Tcl . . . xxvii
Ascend GRF . . . xxviii
Manual Pages for Public Code . . . xxviii

Chapter 1. Introduction . . . 1
New Features in the IBM Parallel System Support Programs for AIX Version 2 Release 3 . . . 1

Chapter 2. RS/6000 SP Overview . . . 5
Hardware Overview . . . 5
Software Overview . . . 9

Routine System Administration Tasks . . . 17

Chapter 3. Starting Up and Shutting Down the SP System . . . 19
Overview of System Startup and Shutdown . . . 19

© Copyright IBM Corp. 1995, 1997

Controlling the Sequence of Startup and Shutdown Operations 19
Starting the SP System 20
Shutting Down the SP System 22
Checking the Status of Startup and Shutdown Requests 25
Defining Startup and Shutdown Sequence Files 26
Creating a Special Subsystem 29

Chapter 4. Parallel Management Commands 31
Using dsh to Execute Parallel Management Commands 31
The hostlist Command 33
List of Parallel Management Commands 34

Chapter 5. Managing File Collections 41
Understanding File Collections 41
Reporting File Collection Information 49
Verifying File Collections Using scan 51
Updating Files in a File Collection 52
Adding and Deleting Files in a File Collection 54
Building a File Collection 56
Installing a File Collection 59
Refusing Files in a File Collection 61
Modifying the File Collection Hierarchy 61
Removing a File Collection 63
supper list File Syntax, Keywords, and Operands 64

Chapter 6. Managing User Accounts 67
Understanding SP User Management 67
Adding User Accounts 70
Changing User Accounts 71
Removing User Accounts 72
Listing a User Account 72
Changing Passwords 72
Controlling User Login 75

Chapter 7. Managing SP Resources 83
Understanding the Resource Manager 83
Installing the Resource Manager 84
Defining the Resource Manager Configuration 86
Starting the Resource Manager 88
Reconfiguring the Resource Manager 94
Stopping the Resource Manager 95
Locating the Resource Manager 96
Querying the Resource Manager 96
Resource Manager Recovery 97

Chapter 8. Managing Time Synchronization 99
Managing NTP 99
Time Setting Options 101
Nodes and Control Workstations Out of Time Synchronization 101

Chapter 9. Managing the Automounter 103
Understanding the Automounter Implementation 103
Establishing an Automount Map for Users' Directories 108
Using Automount for Other File Systems 109

Customizing the SP Automounter Function 112
Coexistence Considerations 121
Migration Considerations 122

Chapter 10. Managing Print Service 125

Chapter 11. Managing Mail Service 127

Chapter 12. Accounting 141
Accounting by Node or by Class 141
Setting Up SP Accounting 143
Accounting Files 148

Chapter 13. Maintaining the crontabs File 153

Security on the SP System 155

Chapter 14. Security Features of the SP System 157
Understanding Authentication 157
Using SP Security Services 166
Administering SP Security Services 172
Using AFS Authentication Services 193

Chapter 15. Sysctl 195
Sysctl Components 195
Relationship to the SP Authentication Services 197
Overview 197
Installation and Configuration 200
Using Sysctl 206
Developing a Sysctl Application 208
Sysctl Scripts 212
Sysctl Reference Information 214

Managing Sets of Nodes 219

Chapter 16. Using a Switch 221
Switch Communication Modes 221
Primary Node Takeover 222
System Partitioning and the Switch 223
Extension Nodes and the SP Switch 224
Selecting Appropriate Switch Clocks 224
Selecting the Primary and Primary Backup Nodes 226
Managing the Switch Topology File 229
Starting the Switch 231
Determining Switch Connectivity 231
Fencing and Unfencing Nodes from the Switch 233
Preparing an SP Switch for Repartitioning 235
Monitoring Nodes on a Switch 235
Global Shutdowns and Reboots of Nodes with a Switch 236
NFS Mounts Over a Switch 237

Chapter 17. Managing System Partitions 239

Understanding System Partitioning 239
Partitioning the SP System 242
Displaying System Partition Configuration Information 256
Verifying System Partition Configuration 258
Managing System Partition-Sensitive Subsystems Using syspar_ctrl 259

Chapter 18. Managing Node Groups 263
Understanding Node Groups 263
Working with Node Groups 265
Using Node Groups with Commands 266

Chapter 19. Managing Extension Nodes 269
Understanding Extension Nodes 269
Specific Implementations of Extension Nodes and Extension Node Adapters 269
Transferring Extension Node Configuration Information 271
Adding Extension Nodes and Extension Node Adapters 271
Changing Extension Nodes and Extension Node Adapters 273
Removing Extension Nodes and Extension Node Adapters 274
Reconfiguring Extension Nodes and Extension Node Adapters 275
Resetting Extension Nodes 276
Listing Extension Node and Extension Node Adapter Information 277

System Monitoring 279

Chapter 20. Using SP Perspectives 281
Getting Started 281
Understanding the SP Perspectives Interface 283
Accessing Help 293
Performing Common Administration Tasks with SP Perspectives 294

Chapter 21. Using the SP System Monitor 297
Understanding the Hardware Monitor 298
Configuring the SP System Monitor 300
Customizing SP System Monitor GUI Default Resources 301
Using the Command Line Interfaces 301
Using the SP System Monitor GUI 302
Starting the SP System Monitor GUI 303
Displaying Frame, Node, and Switch Layouts 307
Monitoring and Controlling the Switch 308
Operating Frame Environment Controls 310
Operating the Front Panel Controls 313
Using Global Controls 320
Determining If Nodes Are Operating: All Node Summary 322
Determining the State of Network Adapters within the SP System 325
Setting Up the Logging Daemon on Separate Workstations 328
Setting Up the Hardware Monitor Clients on Separate Workstations 328

Availability and Problem Management 331

Chapter 22. Managing a High Availability Control Workstation Configuration 333
Understanding HACWS Function 334

Managing SP Subsystems for High Availability Control Workstation Function 342
Customizing HACMP/6000 348
Verifying Frame Supervisor Cables 350

Chapter 23. The Group Services Subsystem 353
Introducing Group Services 353
Group Services Components 354
Components on Which Group Services Depends 357
Group Services Configuration and Operation 358
Group Services Procedures 362

Chapter 24. The Event Management Subsystem 365
Introducing Event Management 365
Event Management Components 365
Components on Which Event Management Depends 375
Event Management Configuration and Operation 377
Event Management Procedures 382

Chapter 25. Using the Problem Management Subsystem 387
Understanding the Problem Management Daemon 388
Understanding the Problem Management Resource Monitor Daemon 394

Chapter 26. Managing SP System Events in a Network Environment . . 399
Understanding Network Management 400
Enabling Network Management for the SP System 401
Disabling Network Management for the SP System 406
Configuring the SP Proxy Agent 406
Configuring NetView for AIX for the SP System 407

Chapter 27. Managing Error Logs 409
Installing and Configuring Error Log Management 409
Managing the AIX Error Log Facility 410
Managing the BSD syslog Facility 413
Viewing Error Logs 415
Archiving Error Logs 415
Collecting System Data for IBM Service 416

System Performance 419

Chapter 28. Tuning the SP System 421
Performance Tuning and Monitoring 421
System Topology Considerations 424
SP System-Specific Tunable Recommendations 426
Tuning the Switch Device Driver Buffer Pools 433
Tuning Problems and Issues 438
Subsystem Tuning 443

The Communications Low-Level Application Programming Interface 451

Chapter 29. Understanding and Using the Communications Low-Level Application Programming Interface (LAPI) 453
Introducing the LAPI 453

Understanding the LAPI 454
Using the LAPI 459

Appendixes 493

Appendix A. Using an Alternate Boot System Image 495
Enabling Alternate Boot System Image Function 495
Using the Alternate Boot System Image Function 496
Setting the Boot List Manually 496
Using Alternate System Image on the Control Workstation 497

Appendix B. SP Daemons 499

Appendix C. Sample Files 503
afsclient.cust 503
block_usr_sample 506
bootptab.info File 508
ibmSPDepNode.my 509
ibmSPMIB.my 516
jmd_config 537
script.cust File 540
tuning.cust File 545
/usr/lib/X11/app-defaults/Spmon 547

Appendix D. System Monitor Variables 549
Frame Hardware Variables Common to All Frame Supervisors 549
Frame Hardware Variables Unique to Card Type 17 and 20 Frame Supervisor 550
Frame Hardware Variables Unique to SEPBU (Type 18) and HACWS SEPBU (Type 22) Frame Supervisors 551
Frame Hardware Variables Unique to HACWS (Type 20) and SEPBU HACWS (Type 22) Frame Supervisors 552
Frame Hardware Variables Unique to Card Type 17 and SEPBU (Type 18) Frame Superviors 553
Node Hardware Variables Common to All Node Supervisors 553
Node-Specific Hardware Variables for Card Types 33 and 65 Node Supervisor 555
Node-Specific Hardware Variables for Card Type 97 Node Supervisors . . . 557
Node-Specific Hardware Variables for Card Types 81 and 83 Node Supervisors 559
Node-Specific Hardware Variables For Card Type 161 Node Supervisor . . . 562
Common Switch Variables 563
High Performance Switch and High Performance Switch-LC8 Card Types 49 and 50 564
SP Switch Variables for Card Type 129 565
Basecode Mode Hardware Variables for the SP Switch Card Type 128, SEPBU Frame Card Type 16 and 604 High Node Type 16 568

Appendix E. Listing SP System Resource Variables 569

Appendix F. The System Data Repository 575
Understanding the System Data Repository 575
Accessing the System Data Repository 577
Updating the host_responds Class (the hrd Daemon) 583
Backing Up the SDR 584

SDR Shadow Files 584
System Data Repository Classes and Attributes 585

Appendix G. IP Address and Host Name Changes for SP Systems . . . 611
SDR Objects with IP Address or Host Names 611
SP System Files with IP Address and Host Names 612
Changing the IP Address and Host Names for PSSP 2.3 Systems 614

Glossary of Terms and Abbreviations 631

Index 639

Figures

1. Basic SP Configuration . . . 5
2. SP Default File Collection Hierarchy . . . 45
3. /var/sysman/sup Files and Directories . . . 46
4. sup.admin Master Files . . . 47
5. SP Hierarchy of an Offline Boot/install Server . . . 62
6. Example of Adding an SP User . . . 71
7. Example Resource Manager Configuration . . . 88
8. Option 1: Using DNS . . . 129
9. Option 2: Using NIS . . . 134
10. Option 3: Using /etc/hosts . . . 137
11. Accounting Template for crontabs . . . 148
12. An Overview of Kerberos Authentication . . . 161
13. Sysctl Elements . . . 196
14. The Change/Show Switch Clock Source Settings SMIT Menu . . . 225
15. Initial Value . . . 226
16. Eprimary Value . . . 226
17. Estart Value . . . 227
18. Initial Values . . . 227
19. Estart Values . . . 228
20. Backup Takeover Values . . . 228
21. Primary Takeover Values . . . 228
22. System Partition Configuration SMIT Panel . . . 244
23. Archive System Data Repository SMIT Panel . . . 246
24. System Partition Configuration List SMIT Panel . . . 247
25. Select System Partition Configuration SMIT Panel . . . 248
26. Display Information for Given Partition or Layout . . . 249
27. Display Information for Given Partition or Layout Output . . . 249
28. Select System Partition Layout SMIT Panel . . . 250
29. Select System Partition Layout Output SMIT Panel . . . 251
30. Partition Directory SMIT Panel . . . 252
31. Enter Customization Arguments for this System Partition SMIT Panel . . . 252
32. Verify System Partition Configuration SMIT Panel . . . 254
33. Apply System Partition Configuration SMIT panel . . . 256
34. List System Partition Database Information Output . . . 257
35. Verify System Partition Configuration SMIT Panel . . . 258
36. The Ascend GRF and SP Switch Router Adapter Extension Node Implementation . . . 270
37. The SP Perspectives GUI Launch Pad . . . 284
38. Elements of an SP Perspectives Primary Window . . . 286
39. An SP Perspectives GUI Primary Window . . . 287
40. An SP Perspectives Dialog Box . . . 290
41. An SP Perspectives Notebook . . . 291
42. An SP Perspectives GUI Message Box . . . 292
43. The SP System Monitor . . . 298
44. SP System Monitor GUI Main Window . . . 305
45. View Pulldown Menu . . . 306
46. SP Pulldown Menu . . . 306
47. SP System Monitor GUI Display Layouts Selection Window . . . 307
48. SP System Monitor GUI Switch Front Panel Layout Window . . . 309
49. Frame Environment Layout for Type 18 Frame Supervisor . . . 312

© Copyright IBM Corp. 1995, 1997

50. Frame Environment Layout for Type 22 Frame Supervisor 313
51. Node Front Panel Display 314
52. Power Off Menu 315
53. Power On Menu 316
54. Reset Menu 317
55. Shutdown Menu 318
56. Fence Menu 319
57. Global Controls Display 321
58. Node Variable Selection Window 323
59. 3-Digit Display 324
60. LAN Topology Display 326
61. LAN Status Display 327
62. An HACWS Configuration 335
63. Addressing in an HACWS Configuration 338
64. Frame Environment Layout Showing Active Control Workstation 348
65. HACWS Pre- and Post-Events 349
66. An Example of Problem Management Daemon Configuration 389
67. Simplified View of a Single Parent–Multiple Child Application 442
68. A LAPI Thread Model 462
69. Sample Output from PSSP SDRGetObjects commands 571
70. Using the Event Perspective to Get Resource Variable Information . . 573
71. Comparing Two SDR Classes 575
72. List RS/6000 SP Configuration Information 581

Tables

1. Specifying Nodes to Start with cstartup 21
2. Specifying Nodes to Shut Down with cshutdown. 23
3. Time Service Choices 99
4. Authorization Variables 211
5. Built-in Tcl Commands and Their Default Authorization Callbacks . . . 215
6. Built-in Extended Tcl Commands and Their Default Authorization Callbacks 215
7. ACL Processing Commands 216
8. Service Commands 216
9. Other Commands 217
10. Summary of System Partition-Sensitive Subsystems 261
11. Frame, Node, and Switch Layouts 308
12. Default Variable Colors 325
13. Graphical Element Colors 327
14. High Availability Control Workstation Configuration and Task Summary 341
15. Initial Values for AIX 4.1 and 4.2 433
16. Initial Values for AIX 3.2.5 433
17. TCP/IP(MVS) parameters/settings 447
18. LED Translation Values 555
19. Node-Specific Hardware Variables For Card Type 161 Node Supervisor 562
20. SDR Authorizations 577
21. Commands Invoked by SMIT Panels 581

© Copyright IBM Corp. 1995, 1997

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 independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
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:

AIX
ESCON
IBM
LoadLeveler
Micro Channel
RS/6000
RS/6000 SP

Microsoft, Windows, and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation.

© Copyright IBM Corp. 1995, 1997

PC Direct is a trademark of Ziff Communications Company and is used by IBM Corporation under license.

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

C-bus is a trademark of Corollary, Inc.

Java and HotJava are trademarks of Sun Microsystems, Inc.

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

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

The *IBM Parallel System Support Programs for AIX: Administration Guide* provides information to help you understand and perform SP system administration. It includes concepts and instructions pertaining to:

- Understanding IBM RS/6000 SP Systems (SP)
- Performing routine system administration tasks
- Making the system secure
- Managing sets of nodes
- Monitoring the system
- Providing for system availability and problem management
- Tuning the system
- Understanding and using the Communications Low-level Application Programming Interface (LAPI)

Additional task information for System Administrators is covered in separate manuals.

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

- The *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment* book helps you plan for the SP system. It tells you how to prepare for software installation and provides considerations for integrating it into your existing computing network.
- 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 for:
  - Installing and configuring the IBM Parallel System Support Programs (PSSP) for AIX
  - Migrating to PSSP Version 2 Release 3
  - Reconfiguring the system
  - Performing system maintenance
- 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 3. To find out what version of PSSP is running on your control workstation, enter the following:

```
SDRGetObjects SP code_version
```

In response, the system displays something similar to:

```
code_version
PSSP-2.3
```

If the response indicates **PSSP-2.3**, this book applies to the version of PSSP that is running on your system.

© Copyright IBM Corp. 1995, 1997

To find out what version of PSSP is running on the nodes on your system, enter the following:

```
splst_versions -G -t
```

In response, the system displays something similar to:

```
1 PSSP-2.3
2 PSSP-2.3
7 PSSP-2.3
8 PSSP-2.3
```

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 system administrators responsible for setting up and maintaining the SP system. It assumes the administrators have a working knowledge of AIX or UNIX and experience with network systems.

The System Administrators Guide of USENIX (SAGE), has developed a classification for skills required for system administration as published in "Job Descriptions for System Administrators." Administrators of small SP systems (one or two frames in a relatively simple environment) are expected to have level II skills (Junior System Administrator). Administrators of larger SP systems are expected to have level III skills (Intermediate/Advanced System Administrator).

Parts of this book can also be used by system operators or others that need to monitor the status of the hardware and software or interact with the power and reset controls.

How This Book Is Organized

SP system administration is a complex topic encompassing many different kinds of tasks. This book is divided into sections that cover information related to managing the day-to-day running of the SP system, providing system security, managing sets of nodes, monitoring and operating the system, providing for system availability and problem management, and tuning the system. There is also a chapter, aimed at programmers, on the Communications Low-level Application Programming Interface (LAPI).

- Chapter 1, "Introduction" on page 1 briefly describes the new and improved features of the SP System.
- Chapter 2, "RS/6000 SP Overview" on page 5 describes the hardware and software components of the SP system.

Routine System Administration Tasks

- Chapter 3, "Starting Up and Shutting Down the SP System" on page 19 shows how to use commands to bring up the SP system and shut it down.
- Chapter 4, "Parallel Management Commands" on page 31 describes commands useful for controlling several nodes in parallel.
- Chapter 5, "Managing File Collections" on page 41 describes procedures for installing, updating, and removing file collections.

- Chapter 6, “Managing User Accounts” on page 67 provides information on passwords, user access, and instructions on how to give your users access to the system.
- Chapter 7, “Managing SP Resources” on page 83 describes how to use the Resource Manager.
- Chapter 8, “Managing Time Synchronization” on page 99 discusses how time service is used in the SP system.
- Chapter 9, “Managing the Automounter” on page 103 discusses how to manage an automounter for mounting file systems.
- Chapter 10, “Managing Print Service” on page 125 explains that in PSSP Version 2 Release 3, the SP print management system has been removed.
- Chapter 11, “Managing Mail Service” on page 127 shows you how to integrate your mail system.
- Chapter 12, “Accounting” on page 141 describes how to run SP Accounting in your environment.
- Chapter 13, “Maintaining the crontabs File” on page 153 contains a few hints on running **cron** jobs.

Security on the SP System

- Chapter 14, “Security Features of the SP System” on page 157 describes how to set up, maintain, and use authentication services.
- Chapter 15, “Sysctl” on page 195 describes how to install, configure, and use the authenticated client/server system for running commands remotely and in parallel.

Managing Sets of Nodes

- Chapter 16, “Using a Switch” on page 221 discusses switch operations.
- Chapter 17, “Managing System Partitions” on page 239 describes the tasks necessary to partition (and repartition) your SP system.
- Chapter 18, “Managing Node Groups” on page 263 presents information for understanding and working with node groups.
- Chapter 19, “Managing Extension Nodes” on page 269 describes how to add, change, and delete extension nodes and adapters in the SP system.

System Monitoring

- Chapter 20, “Using SP Perspectives” on page 281 describes Scalable POWERparallel Perspectives for AIX (SP Perspectives) and provides details on starting the GUI, understanding the SP Perspectives GUI interface, and accessing help.
- Chapter 21, “Using the SP System Monitor” on page 297 describes how to use the SP System Monitor with both the command interface and the SP System Monitor Graphical User Interface (GUI) to monitor and control the hardware.

Availability and Problem Management

- Chapter 22, “Managing a High Availability Control Workstation Configuration” on page 333 provides information for understanding High Availabiltity Control Workstation (HACWS) function, performing tasks to ensure that subsystems fail over to the backup control workstation in the event the primary control

workstation fails, and customizing AIX High Availability Cluster Multi-Processing/6000 (HACMP/6000).

- Chapter 23, "The Group Services Subsystem" on page 353 provides information on managing the Group Services subsystem.
- Chapter 24, "The Event Management Subsystem" on page 365 provides information on managing the Event Management subsystem.
- Chapter 25, "Using the Problem Management Subsystem" on page 387 describes the primary components of the Problem Management subsystem and how to manage it.
- Chapter 26, "Managing SP System Events in a Network Environment" on page 399 describes how to issue SNMP traps in response to system events.
- Chapter 27, "Managing Error Logs" on page 409 describes the functions provided for managing the AIX Error Log facility, the BSD syslog, and SP-specific log files.

System Performance

- Chapter 28, "Tuning the SP System" on page 421 contains recommendations for fine-tuning your environment for better performance.

The Communications Low-Level Application Programming Interface

- Chapter 29, "Understanding and Using the Communications Low-Level Application Programming Interface (LAPI)" on page 453 introduces the concepts and use of the Communications Low-Level Programming Interface (LAPI).

Appendixes

- This book also includes appendixes containing supplementary information as well as a glossary and index.

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. • **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. • *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.
[ ]	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.
...	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**>.

Interface Instructions

Some sections of this book give step-by-step instructions for performing tasks with a graphical user interface. The instructions use a format that distinguishes between the user action and the system response.

User actions appear in uppercase bold type.

PRESS Cancel

Selections from a menu bar are indicated with an →.

SELECT SP → Topology

The system response to an action follows the ▲ symbol.

▲ The diagnostic display appears.

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 the following:

- Access to a common HTML document browser (such as Netscape, WebExplorer, or Mosaic).
- The location of the HTML index file provided with the **ssp.docs** file set. Contact your system administrator or installer for this location.

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

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: 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 entire SP software library.

IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Publication

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

Performance Monitor Publication

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

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

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.

PVMe for AIX Publications

- *IBM PVMe for AIX: Fact Sheet*, GC23-3907
- *IBM PVMe for AIX: User's Guide and Reference*, GC23-3884
- *IBM PVMe for AIX: Licensed Program Specification*, GC23-3885

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

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 General Concepts and Procedures for RS/6000*, GC23-2202
- *IBM AIX Version 4 Files Reference*, SC23-2512
- *IBM AIX Version 4 System Management Guide: Communications and Networks*, SC23-2526
- *IBM AIX Version 4.1 Installation Guide*, SC23-2550
- *IBM AIX Version 4.2 Installation Guide*, SC23-1924
- *IBM AIX Version 4 Commands Reference*, SBOF-1851 (all volumes)
- *IBM AIX Versions 3.2 and 4 Performance Tuning Guide*, SC23-2365
- *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 System Management Guide: Operating System and Devices*, SC23-2525
- *IBM AIX Version 4 General Programming Concepts: Writing and Debugging Programs*, SC23-2533
- *IBM AIX Version 4 Communications Programming Concepts*, SC23-2610
- *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
- *IBM Performance Toolbox 1.2 and 2 for AIX: Guide and Reference*, SC23-2625

Service

- *IBM AIX Version 3.2 Problem Solving Guide and Reference*, SC23-2204
- *IBM AIX Version 4 Problem Solving Guide and Reference*, SC23-2606
- *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 R40* (No order number — this book comes as part of the ship group with the SP hardware)

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

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 Models 010 and 011 SCSI Expansion Units: Installation and Service Guide*, SY33-0165
- *9334 Models 500 and 501 SCSI Expansion Units: 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

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*. Addison-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.

Ascend GRF

Order according to your Ascend GRF IP switch model:

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

You can order the Ascend GRF as the IBM 9077.

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 /**usr**/**lpp**/**ssp**/**public**.

Chapter 1. Introduction

RS/6000 SP (SP) offer an information processing solution that provides:

- Scalability

  Standard system configurations of two to 128 processor nodes are supported. You can expand your system over time to meet your changing computing requirements.

- Cost effectiveness

  Because the SP system is based on AIX and RS/6000 technology, thousands of existing RS/6000 applications can be run on the SP, eliminating the need for costly reprogramming.

  The scalability of the SP system ensures that hardware and software invested in now will continue to serve you over time.

- Parallel computing

  The SP system provides a strong platform for parallel application development and execution. System management applications provided in the IBM Parallel System Support Programs for AIX, the High Performance Switch, the SP Switch, and related IBM products such as IBM Parallel Environment for AIX enhance the parallel computing capabilities of the SP system.

- Support of open system standards

  The SP system accommodates a wide number and variety of I/O attachments and disk storage.

  Enablers are provided that permit a number of commercial applications to exploit the parallel capabilities of the SP system.

- Flexibility

  The IBM Parallel System Support Programs for AIX software allows you to run various combinations of serial, parallel, batch and interactive jobs concurrently.

- Simplified operation

  The control workstation serves as a single point of control for installing, managing, and monitoring the SP system.

These features and capabilities make the SP an outstanding system for technical computing in a variety of market segments including Engineering Analysis, Chemistry, and Seismic/Petroleum. At the same time, the SP system is an excellent choice for the parallel commercial market, supporting database query, business management, decision support, online transaction processing, and batch business applications.

New Features in the IBM Parallel System Support Programs for AIX Version 2 Release 3

Version 2 Release 3 of the IBM Parallel System Support Programs for AIX provides the following functional enhancements:

© Copyright IBM Corp. 1995, 1997

Automatic Installation of Initial Set of Network Tuning Parameters

The automatic installation on the nodes of an initial set of network tuning parameters is now provided. These initial network tuning parameters provide improved performance in a typical SP environment over the standard settings provided with base AIX.

In addition, three sets of network tuning parameters are provided that are designed to improve initial performance for the following environments:

- Commercial
- Development/interactive
- Engineering/scientific

Refer to “Changing Network Options on the Nodes” on page 422 for more information.

Automounter Support

The Amd automounter is removed from PSSP 2.3 and is replaced with support for the AIX Automounter that is part of the AIX Network Support Facilities of the Base Operating System Runtime.

In addition, customization scripts are provided so that you can replace some of the AIX automounter function with your own options or you can use another automounter altogether.

If you are running older versions of PSSP on some of the nodes in your SP system, those nodes will continue to use the Amd automounter.

Refer to Chapter 9, “Managing the Automounter” on page 103 for more information.

The Communications Low-Level Application Programming Interface (LAPI)

The Communications Low-level Application Programming Interface (LAPI) is a non-standard application programming interface designed to provide an efficient and flexible interface to the SP Switch.

Refer to Chapter 29, “Understanding and Using the Communications Low-Level Application Programming Interface (LAPI)” on page 453 for more information.

Extension Node Support

Extension nodes are non-standard nodes that extend the SP system's capabilities or scope, but can't be used in all the same ways as standard SP nodes.

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

The specific dependent node implementation supported in this release is the attaching of an Ascend GRF switched IP router with one or more SP Switch Router Adapters to the SP Switch in order to provide higher bandwidth communications. Support includes:

- SMIT panels and commands for adding, changing, deleting, and listing SDR information for the switched IP router and for the SP Switch Router Adapter.
- Modifications to the Communications Subsystem and SP Switch commands to accommodate the characteristics of a switched IP router.
- Simple Network Management Protocol (SNMP) support for transfer of switch-related configuration information from the SP system to the switched IP router.

Refer to Chapter 19, “Managing Extension Nodes” on page 269 for more information.

Migration and Coexistence Support

Support is provided for migrating to PSSP 2.3 running on AIX 4.2.1 from the following PSSP releases:

- PSSP 2.2 on AIX 4.2.0 or later
- PSSP 2.2 on AIX 4.1.4 or later
- PSSP 2.1 on AIX 4.1.3 or later
- PSSP 1.2 on AIX 3.2.5

Support is provided for the coexistence of PSSP 2.3 in a system partition with other PSSP releases in the following configurations:

- PSSP 1.2 and PSSP 2.2
- PSSP 2.1 and PSSP 2.2
- PSSP 1.2, PSSP 2.2, and PSSP 2.3
- PSSP 2.1, PSSP 2.2, and PSSP 2.3
- PSSP 1.2 and PSSP 2.3
- PSSP 2.1 and PSSP 2.3
- PSSP 2.2 and PSSP 2.3

For more information on migration refer to *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

IBM Virtual Shared Disk Enhancements

Enhancements to the IBM Virtual Shared Disk include:

- The ability to determine the size of an underlying Logical Volume on a server and to provide that information to applications
- The capability to change which node acts as the server node for a single volume group, rather than for all the volume groups.

Refer to *IBM Parallel System Support Programs for AIX: Managing Shared Disks* for more information.

Improvements to Management of Kerberos Database and Services

Function is provided to make it easier for security administrators to manage the Kerberos authentication database and servers. New function in this release includes:

- The **lskp** command for listing principals in the local Kerberos database according to various criteria
- The **rmkp** command for removing principals from the local Kerberos database

- The **chkp** command for changing the current expiration or the maximum ticket lifetime of principals in the local Kerberos database
- The **mkkp** command for creating principals in the local Kerberos database

In addition, the Kerberos daemons **kerberos**, **kadmind** and **kpropd** are now under the System Resource Controller and can be queried, stopped, and started with SRC commands.

For more information, refer to Chapter 14, “Security Features of the SP System” on page 157 and *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

SP Perspectives Performance and Usability Enhancements

SP Perspectives is a set of applications, each of which has a graphical user interface, that enables you to perform system management tasks for your SP system. In this release the following improvements have been made to SP Perspectives:

- Performance for start up and execution of functions
- Scalability with small/large icon sizes and vertical panes
- Indexing of online help for retrievability

Switch Clock API

The new Switch Clock API provides a device-independent interface to the High Performance Switch and the SP Switch. It supports the switch clock in providing a highly synchronous and precise measurement of relative time across nodes.

Removal of SP Print Management System

The SP Print Management System is removed from PSSP 2.3. We suggest that you use the Printing Systems Manager (PSM) for AIX.

If you are running older versions of PSSP on some of the nodes in your SP system, those nodes will continue to use the SP print management system.

New Adapter Support for the SP LAN

The hardware adapters 2992 and 2993 are now supported for SP LAN use.

Support for GPFS and HACMP ES

PSSP 2.3 supports the new IBM Licensed Program Products General Parallel File System (GPFS) and High Availability Cluster Multi-Processor Enhanced Scalability (HACMP ES).

Chapter 2. RS/6000 SP Overview

This chapter provides an overview of the hardware and software that compose RS/6000 SP (SP).

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.

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

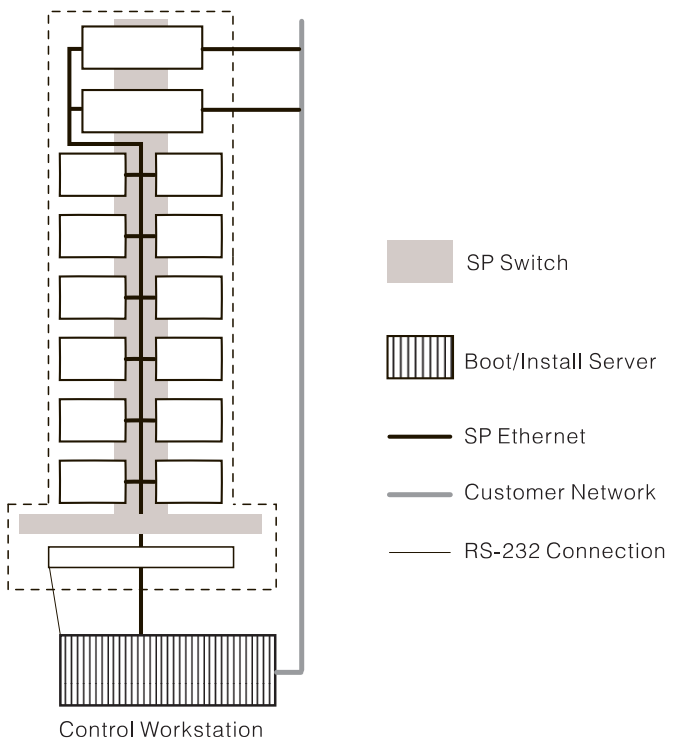


Figure 1. Basic SP Configuration

© Copyright IBM Corp. 1995, 1997

Processor Nodes

The SP system is scalable from two to 128 processor nodes that can be contained in multiple SP frames. A single frame can contain from two to 16 processor nodes.

The SP nodes are available in three types: thin nodes, wide nodes, and 604 High Nodes. Thin nodes are typically configured as compute nodes, while wide nodes are more often used as servers to provide high-bandwidth data access. 604 High Nodes provide advantages as data base servers or LAN consolidation servers. 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.

Each processor node includes a minimum of 64 megabytes of memory, one gigabyte of direct access storage devices (DASD), and an Ethernet adapter.

Thin Nodes

The SP thin node is functionally equivalent to a RS/6000 desktop system. Half the width of the wide node, it can be packaged 16 to a frame and can contain either a RS/6000 POWER or POWER2 processor. The POWER processor operates at 62.5 megahertz, and the POWER2 processor operates at 66.7 megahertz with 64 kilobytes of data cache, and 32 kilobytes of instruction cache. The thin processor node has four Micro Channel adapter (MCA) slots. It can have up to two SCSI disks packaged internally, and can have access to external storage devices attached through SCSI-2 adapter features. The Ethernet adapter for the SP Ethernet is integrated and does not use an MCA slot.

Wide Nodes

The SP wide node is functionally equivalent to a RS/6000 deskside system. It is a RS/6000 POWER2 processor that operates at 66.7 megahertz with 256 kilobytes of data cache, and 32 kilobytes of instruction cache. The wide node occupies twice the space of the thin node and can be packaged up to eight per frame. In addition to a maximum of four SCSI disks packaged internally, the wide node can have external storage devices attached through SCSI-2 adapters. It has seven MCA slots and also supports the High Performance Parallel Interface (HiPPI) adapter. The Ethernet adapter for the SP Ethernet uses one MCA slot.

604 High Nodes

The 604 High Node is a 2-, 4-, 6-, or 8-way symmetric multiprocessing system that uses PowerPC 604 processors that operate at 112 megahertz. The 604 High Node occupies two drawers (four slots) in a frame and can be packaged up to four per frame. It has two Micro Channel buses for I/O attachment, including attachment to the High Performance or SP Switch (however, only one switch adapter will be supported).

A unique feature of the 604 High Node is the SystemGuard collection of firmware programs that run on the bringup microprocessor (BUMP). The SystemGuard and BUMP provide service processor capability. They enable the operator to manage power supplies, check system hardware status, update various configuration parameters, investigate problems, and perform tests. Refer to *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* for more information on SystemGuard and the 604 High Node.

Extension Nodes

Extension nodes are non-standard nodes that extend the SP system's capabilities or scope, but can't be used in all the same ways as standard SP nodes.

A specific type of extension node is a dependent node. A dependent node depends on SP nodes for certain functions, but implements much of the switch-related protocol that standard nodes use on the SP Switch. Dependent nodes are attached to the SP Switch via a special dependent node adapter.

A physical dependent node, such as a switched IP router, may include multiple logical dependent nodes, one for each dependent node adapter it contains. If a dependent node contains more than one dependent node adapter, it can route data between SP systems or system partitions by connecting the adapters to different SP systems or system partitions.

A physical dependent node does not reside in a frame, but each of its dependent node adapters logically occupies the thin node slot that corresponds to its switch port, and that slot can't be occupied by another node.

Frames

An SP frame can house up to four 604 High Nodes or eight wide nodes or 16 thin nodes. 604 High Nodes and wide and thin nodes can be mixed together in a single frame.

The SP frame contains redundant power supplies; if one power supply fails, another takes over. The frame is also designed for concurrent maintenance; each processor node can be removed and repaired without interrupting operations on the other nodes.

An SP Switch-8 configuration can consist of either two short 49" frames with single phase power that supports two to eight nodes in up to two frames, or one 79" frame that supports up to eight nodes.

High Performance Switch

The High Performance Switch provides low latency, high-bandwidth communication between nodes, supplying a minimum of four paths between any pair of nodes. The High Performance Switch dramatically speeds 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). High Performance 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 or multi-user environments
- Error detection and retry

The powerful High Performance Switch Adapter-2 connects each SP node to the High Performance Switch subsystem. See Chapter 16, “Using a Switch” on page 221 for more information.

Scalable POWERparallel Switch

In addition to all the capabilities provided by the High Performance Switch described previously, the SP Switch provides:

- 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. See Chapter 16, “Using a Switch” on page 221 for more information.

The SP Switch Router Adapter connects Ascend GRF switched IP routers to the SP Switch subsystem. See “Specific Implementations of Extension Nodes and Extension Node Adapters” on page 269 for more information.

Control Workstation

The SP system requires a customer-supplied RS/6000 workstation with a color monitor. The control workstation serves as a point-of-control for managing and maintaining the SP processor nodes using the IBM Parallel System Support Programs for AIX. The 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. (See the section on control workstation planning in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* to ensure your workstation meets minimum requirements.)

The control workstation also acts as a boot/install server for other servers in the SP system. In addition, the control workstation can be set up as an authentication server. It can be the primary server, with the master database and administration service as well as the ticket-granting service. Or it can be set up as a secondary server, with a backup database and just the ticket-granting service. (For more information, see Chapter 14, “Security Features of the SP System” on page 157.)

The High Availability Control Workstation option provides for automatic failover and reintegration to a backup control workstation should the primary control workstation become unavailable.

Network Connectivity Adapters

Network connectivity for the SP system 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.

The SP Ethernet is a network that connects all nodes and the control workstation in the SP. A 50-foot Ethernet cable is provided with each frame to use in the wiring of this network. Additional optional adapters such as Ethernet, FDDI, and token-ring are automatically configured on each node. Other optional supported adapters can be individually configured on each node.

Software Overview

The SP system software infrastructure includes:

- AIX, the base operating system
- The IBM Parallel System Support Programs for AIX
- Other IBM system and application software products
- Independent software vendor products

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. Some of the AIX capabilities that can be exploited for system management on an SP system include:

- The Network Installation Management (NIM) environment provides the ability to install an AIX **mksysb** image over the network
- The Logical Volume Manager (LVM), which improves data management productivity, enables files to span multiple disk drives, and provides, through disk mirroring, a high availability option for critical data
- The System Management Interface Tool (SMIT), which provides a single consistent, expandable interface to system management commands
- The System Resource Controller (SRC), which simplifies the management of subsystems and daemons
- Security features, which include a shadow password file and comprehensive auditing facilities
- AIX device support, which allows devices to be added and deleted at any time without disrupting the system

AIX provides UNIX functionality and conformance to industry standards for open systems.

IBM Parallel System Support Programs for AIX (PSSP)

The PSSP software provides a single point of control for administrative tasks and helps increase productivity by letting administrators view, monitor, and control system operation.

Highlights

The PSSP software provides a comprehensive suite of applications for the installation, operation, management, and administration of the RS/6000 SP:

Installation: Installation of AIX 4.2.1 and later versions on nodes is provided for by the Network Installation Management (NIM) environment.

Configuration Management: Node configuration data is entered into the System Data Repository (SDR) using SMIT or line commands. The SDR provides storage and retrieval of system data across the control workstation, file servers, and nodes.

Operation:

- The Centralized Management Interface provides a SMIT-based interface for system management commands, as well as command line equivalents.
- The SP Perspectives GUI provides for monitoring and controlling hardware, creating and monitoring system events, managing IBM Virtual Shared Disks, generating and saving system partition configurations, and setting up performance monitoring hierarchies and archiving.
- The SP System Monitor and a command line interface provide for monitoring and controlling frame and node hardware.
- Tools are provided for orderly shutdown and restart of the system.

File Management: Groups of files can be defined as a file collection and changes to any files in that collection will be propagated to the appropriate nodes.

User Management:

- Commands are supplied to add, change, and delete users and passwords.
- Users have the same account, home directory, and environment across all nodes in the system.
- The **user.admin** file collection is provided for propagating user administration files to the nodes on the system.

Job Management: The Resource Manager server runs on one of the nodes and manages the execution of parallel jobs.

Accounting: Accounting support provides accounting record consolidation and parallel job accounting.

Change Management:

- System partitions can be used to apply and test new levels of software.
- Parallel management commands can be used to apply maintenance directly to nodes.
- File collections can be used to propagate changed files to nodes that have the file collection installed.
- For extensive service a new **mksysb** can be created and the nodes reinstalled using NIM.

Security: Authentication services, based on Kerberos IV, are provided for:

- The Sysctl remote command execution facility
- The SP System Monitor
- Versions of the **rsh** and **rcp** commands

Communication Subsystem Support: The Communication Subsystem supports the reliability and performance of the High Performance Switch and the SP Switch.

Note: PSSP supports both the SP Switch and the High Performance Switch, but the two switch networks are not compatible and cannot be mixed within an SP system.

Availability and Problem Management:

- Group Services provide a distributed coordination and synchronization service.
- Event Management matches information about the state of system resources with information about resource conditions that are of interest to client programs.

- Problem Management provides for pairing actions with system events.
- A single point of control for configuring, archiving, and maintaining various system logs on individual nodes is provided.
- Additional logs may be specified for use by applications.

Packaging

The IBM Parallel System Support Programs for AIX (Program Number 5765-529) are packaged as the following **installp** images:

- **pssp**
- **spimg**
- **ssp.csd**
- **ssp.ptpegui**

The following are optional separately priced features:

- **ssp.hacws**
- **ptpe**

pssp

The **installp** image for **pssp** contains the code to install, operate and manage your SP system. The options in this package include:

ssp.basic: The following support functions are included in this option:

- Installation and configuration
  (Refer to the *IBM RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment* and the *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment* guides to help you prepare for installing and configuring an SP system.)
- System operation and monitoring
  (See Chapter 21, “Using the SP System Monitor” on page 297.)
- Distributed shell
  (See Chapter 4, “Parallel Management Commands” on page 31.)
- Login control
  (See Chapter 6, “Managing User Accounts” on page 67.)
- System Data Repository (SDR)
  (See Appendix F, “The System Data Repository” on page 575.)

ssp.gui: This option contains the SP Perspectives GUI as well as the graphical user interface for the System Monitor. Refer to Chapter 20, “Using SP Perspectives” on page 281 and Chapter 21, “Using the SP System Monitor” on page 297 for more information.

ssp.sysman: This option contains the following functions:

- Time services
  (See Chapter 8, “Managing Time Synchronization” on page 99.)
- User management
  (See Chapter 6, “Managing User Accounts” on page 67.)
- Print support
  (See Chapter 10, “Managing Print Service” on page 125.)
- File collections

(See Chapter 5, “Managing File Collections” on page 41.)

- Accounting
  (See Chapter 12, “Accounting” on page 141.)
- Parallel management commands
  (See Chapter 4, “Parallel Management Commands” on page 31.)
- Error Log Management
  (See Chapter 27, “Managing Error Logs” on page 409.)

ssp.jm: This option contains the *Resource Manager* facility for allocating nodes to run parallel jobs, and for tracking parallel jobs across the SP system. It allows space sharing and time sharing of nodes as well as interactive and batch execution of parallel jobs. See Chapter 7, “Managing SP Resources” on page 83 for more information.

ssp.st: This option contains a library and commands that support an application programming interface for loading, unloading, and querying the status of job switch tables.

ssp.css: **Communications Subsystem Support** performs switch initialization during system bring-up, and automatic switch reconfiguration following switch detected errors. CSS also provides software error detection codes for all messages, and concurrent servicing of SP nodes and frames through message retry following a switch fault. This option also contains the switch clock API.Chapter 16, “Using a Switch” on page 221 contains more details.

ssp.spmgr: This option contains support for the transfer of switch-related information to dependent nodes attached to the SP Switch. The SP SNMP Manager resides on the control workstation. It sends configuration information to the SNMP Agent residing on the dependent node and receives information about dependent node state changes.

ssp.top: This option contains directories and files for supported system partitioning configurations, as well as the System Partitioning Aid, a tool for creating customized system partition configurations to suit your needs.

ssp.top includes:

- Subdirectories for SP system sizes
- Subdirectories for partition configurations for SP system sizes
- Layout directories for partition configurations
- A layout text file that describes the configuration
- Topology and nodelist files for the configuration
- The System Partitioning Aid

ssp.top.gui contains the graphical user interface for the System Partitioning Aid. Refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* and Chapter 17, “Managing System Partitions” on page 239 in this guide for more information.

ssp.topsvcs: This option contains the Topology Services/ES.

ssp.ha: This option contains subsystems for high availability, including heartbeat, Group Services, and Event Management.

ssp.pman: This option contains the Problem Management subsystem. The Problem Management subsystem provides configurable access to Event Management client and resource monitor function. It includes daemons, commands, utilities, and scripts. Refer to Chapter 25, “Using the Problem Management Subsystem” on page 387 and Chapter 26, “Managing SP System Events in a Network Environment” on page 399 for more information.

ssp.authent: This option contains the server (daemon) code that provides Kerberos ticket-granting services, and primary and secondary database administration. It also contains a number of privileged utility commands for manipulating the local Kerberos database. This option can be installed on the control workstation or on other RS/6000 workstations that are configured as Kerberos server systems. It is not installed on the SP nodes. Refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* and Chapter 14, “Security Features of the SP System” on page 157 in this guide for more information.

ssp.sysctl: *Sysctl* is an authenticated client/server system for running commands remotely and in parallel. Sysctl consists of a server daemon, commands, configuration files, and a client shell program. Refer to Chapter 15, “Sysctl” on page 195 for a full description.

ssp.clients: This option provides the nonprivileged (user) Kerberos commands, as well as client support for authenticated services, namely the SP System Monitor that runs on an SP control workstation and the **kshell** service of the SP authenticated remote commands (Kerberos-authenticated **rsh** and **rcp**). This option is required on the control workstation and is automatically installed on every SP node. It can also be installed on any other RS/6000 workstation on which you desire these capabilities. Refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* and Chapter 14, “Security Features of the SP System” on page 157 in this guide for more information.

ssp.perlpkg: This option contains Perl 4.036 and Perl 5.002 programs and libraries.

ssp.public: This option contains tar files of full public distributions of Perl (Practical Extraction and Report Language), NTP (Network Time Protocol), SUP (System Update Protocol), Tcl (Tool Command Language), TclX (Tool Command Language Extended), Tk (Tcl-based Tool Kit for X-windows), and Expect (Programmed dialogue with interactive programs).

ssp.docs: This option contains the PostScript, man page, and html files for PSSP documentation.

spimg

The **spimg** installp image contains a **mksysb**, or multiple **mksysb**s, of an AIX 4.2.1 or later system that you can use to install AIX 4.2.1 or a later version on the nodes in your SP system, after you have successfully installed and configured your control workstation. This minimal system is composed of the basic operating system, extensions, networking and basic utilities. It is provided as a base for getting your SP system up and running quickly. You should customize this image to meet the needs of your production environment. Instructions for customizing the minimal image are in the *IBM Parallel System Support Programs for AIX:*

Installation and Migration Guide. See the IBM Parallel System Support Programs for AIX (PSSP) *Memo to Licensees* for a description of the contents.

ssp.csd

This image provides the IBM Virtual Shared Disk and the IBM Virtual Shared Disk Data Striping Device (HSD) function for the SP system. The IBM Virtual Shared Disk function allows each node in the cluster to access a disk as if it were local to the node, even though the disk is not physically attached to all SP nodes. The HSD is a device driver that lets application programs stripe data across the physical disks in multiple IBM Virtual Shared Disks. The **csd** image is split into five install options:

ssp.csd.vsd IBM Virtual Shared Disk Device Driver, Methods and Commands.

ssp.csd.cmi IBM Virtual Shared Disk Configuration Management Support which includes the SMIT panels and the commands to interface with the System Data Repository (SDR).

ssp.csd.hsd IBM Virtual Shared Disk Data Striping Device. The **ssp.csd.vsd** and **ssp.csd.cmi** options must be installed on all SP nodes that are to be used as either IBM Virtual Shared Disk server or client nodes. The **ssp.csd.hsd** option must be installed on all nodes with IBM Virtual Shared Disks, if data striping will be used.

ssp.csd.sysctl IBM Virtual Shared Disk commands that use Sysctl function.

ssp.csd.gui The graphical user interface for IBM Virtual Shared Disks.

See *IBM Parallel System Support Programs for AIX: Managing Shared Disks* for details.

ssp.ptpegui

This option contains the graphical user interface for the Performance Toolbox Parallel Extensions for AIX.

ssp.hacws

The **ssp.hacws** install option supports High Availability Control Workstation (HACWS) function. HACWS provides automatic fail over and reintegration to a backup control workstation should the primary control workstation become unavailable.

ssp.hacws contains:

- Scripts to create a backup control workstation
- Error notification object samples
- Error log templates
- Object Database Manager updates
- Verification programs
- HACMP pre- and post-scripts
- Scripts to synchronize the backup and primary control workstations

Note

This is an optional separately priced feature.

ptpe

This image provides the Performance Toolbox Parallel Extensions for AIX, a performance monitor that provides access to performance information about SP system hardware and software. The **ptpe** image is split into three options:

ptpe.docs Performance Toolbox Parallel Extensions for AIX documentation, including a PostScript file and hypertext version (accessible with the **InfoExplorer** tool).

ptpe.gui The graphical user interface for the Performance Toolbox Parallel Extensions for AIX.

ptpe.program The Performance Toolbox Parallel Extensions for AIX program.

> **Note**
>
> This is an optional separately priced feature.

Other IBM Products

A large number of system and application software products developed by IBM can be run on the SP system. For example:

- IBM LoadLeveler
  - Distributed, network-wide job management
- IBM Parallel Environment for AIX
  - Parallel application development and execution
  - Message passing parallel task communications
- IBM PVMe for AIX
  - Message passing for parallel programs on distributed memory parallel computers
- IBM Parallel I/O File System
  - Data file system for parallel access to large files
- ADSM/6000
  - System and user data management
- PTX/6000
  - Performance monitoring
- NetView for AIX
  - Enterprise network management
- Trouble Ticket for AIX
  - Problem management
- PSF/6000
  - Printing support
- Job Scheduler for AIX
  - Schedules production batch workload in a distributed AIX environment

Independent Software Vendor Products

There is an aggressive program in place to enable and encourage independent software vendors to enable and port their applications to the SP system.

Routine System Administration Tasks

© Copyright IBM Corp. 1995, 1997

Chapter 3. Starting Up and Shutting Down the SP System

This chapter describes how to start up and shut down nodes on the SP system. It explains how to specify the sequence in which the nodes and subsystems in your system are started up and shut down. It also describes the tools that save you time and effort by operating on all or some of the nodes in your SP system with a single command. These tools include the **cstartup**, **cshutdown**, and **seqfile** commands, plus the files they use to control the sequence of startup or shutdown processing.

In addition, this chapter describes the interface that lets you coordinate the shutdown of a subsystem that operates across multiple nodes.

Overview of System Startup and Shutdown

You control the operation of the SP system with commands issued at the control workstation.

Use the **cstartup** command to power on SP nodes and boot up the operating system. You can also reset SP nodes with the **cstartup** command.

Use the **cshutdown** command to halt or reboot SP nodes.

Your SP may be a complex group of nodes, each dependent on other nodes for some resources. Startup and shutdown operations must occur in the proper sequence so that these dependencies are honored. As a simple example, a server node may have to be running before its client nodes are started up. A server node should not shut down before its clients have been shut down. The **seqfile** command defines sequencing relationships that result from using boot/install servers. You can define additional dependencies.

To provide best performance, **cstartup** and **cshutdown** can operate on multiple nodes concurrently.

Similar considerations apply to the shutdown of *special subsystems* (see “Creating a Special Subsystem” on page 29) that run on multiple nodes in the SP system. An example of a subsystem is a file system that spans several nodes within a system partition. Subsystems may need to do special processing to prepare for the shutdown of one or more nodes, or a set of subsystems may need to shut down in a specific order.

Controlling the Sequence of Startup and Shutdown Operations

You must define the sequencing relationships between your nodes. For example, boot/install servers must start up before their clients. Other programs on your SP system may create additional sequencing relationships.

The node sequence files, **/etc/cshutSeq** and **/etc/cstartSeq**, have lines that describe these dependencies. In these files, you define the groups of nodes and the sequence in which the groups are started up or shut down. If you don't have a node sequence file, some sequencing is performed automatically. The **seqfile** command helps you create these files by generating definitions for groups of nodes

© Copyright IBM Corp. 1995, 1997

and their sequencing order for dependencies relating to boot/install servers and their clients.

You can also create a subsystem sequence file, **/etc/subsysSeq**, which lists the subsystems, defines groups of subsystems, and defines relationships between subsystems. If you don't have a subsystem sequence file, no subsystems are notified of an impending node shutdown.

The sequence files, which reside on the control workstation, are:

/etc/cstartSeq Startup. The list defining groups of nodes and the order of startup processing.

/etc/cshutSeq Shutdown. The list defining groups of nodes and the order of shutdown processing.

/etc/subsysSeq Special subsystems. The list defining special subsystems, groups of special subsystems, and the order of shutdown processing.

The node sequence files are separate to allow for different startup and shutdown dependencies. However, if your startup and shutdown dependencies are mirror images, you can maintain a single file; create a symbolic link to provide both file names.

"Defining Startup and Shutdown Sequence Files" on page 26 describes the format of the sequence files and their interactions with command line flags.

Default Node Sequencing

The SP system can do some sequencing automatically during system startup or shutdown. The **seqfile** command shows you the default sequencing relationships.

If you don't create the **/etc/cstartSeq** file, the **cstartup** command uses a default sequence based on the System Data Repository (SDR). Any nodes defined in the SDR as boot/install servers are started up concurrently. After the servers are running, nodes defined as boot/install clients are started up. If you create an empty **/etc/cstartSeq** file, then all nodes are started up concurrently.

If you don't create an **/etc/cshutSeq** file, **cshutdown** uses a default sequence based on the SDR. If you create an empty **/etc/cshutSeq** file, then all nodes are shut down concurrently. If you don't have a **/etc/subsysSeq** file or the file is empty, no special subsystem shutdown occurs.

Starting the SP System

To start the SP system, login on the control workstation and issue the **kinit** command, specifying a principal name for which there is an entity in the **hardmon** acls file with control authorization for the frames to be started up. Then, use the **cstartup** command.

Note on 604 High Nodes

The 604 High Node takes a considerably longer time to power on than thin and wide nodes. It can take a 604 High Node 10 to 20 minutes to power on, depending on the number of processors in the node and the number of adapters. You may wish to control the startup and shutdown of 604 High Nodes manually, rather than with the **cstartup** and **cshutdown** commands.

The **cstartup** command powers on specified nodes that are not powered on, and may reset specified nodes that are already powered on. Those actions are equivalent to manually pressing the hardware switches. Once the hardware-equivalent actions are done, control is passed to **boot** and **init**.

Caution!

The **cstartup** command attempts to power on nodes that are powered off. This has safety implications if someone is working on the nodes. Proper precautions should be taken when using this command.

To restart a SP system that is already running, use **cshutdown -r -G**; this does *not* power on any nodes.

Note: If some nodes cannot be started in a specific sequence, or if some cannot be started, the system may still be usable.

The complete syntax of the **cstartup** command is described in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Starting or Resetting All Nodes

You can start or reset nodes with a single **cstartup** command, as shown in Table 1.

Table 1. Specifying Nodes to Start with cstartup

Command	Meaning
cstartup ALL	Starts all nodes in the current system partition.
cstartup *target_nodes*	Starts specific nodes in the current system partition. All nodes listed must be in the current system partition or the command will fail.
cstartup -G ALL	Starts all nodes in the physical SP system.
cstartup -G *target_nodes*	Starts specific nodes in the physical SP system, regardless of system partitioning.

If a node is powered off, this command powers it on. If the node is already powered on and is not running, this command resets it. The nodes are started in the sequence specified by /**etc**/**cstartSeq**. If this file doesn't exist, the nodes are started up in the default sequence.

To start all nodes in the physical SP system concurrently, ignoring the sequence file if it exists or bypassing the default sequencing if the file does not exist, use the **-E** flag:

```
cstartup -E -G ALL
```

Information about creating the /**etc**/**cstartSeq** file is in “Defining Startup and Shutdown Sequence Files” on page 26.

Starting or Resetting Selected Nodes

To start or reset a specific node within the current system partition, put its name on the **cstartup** command line. The following example starts the single node in the current system partition named **cs6**:

```
cstartup cs6
```

The command checks for dependencies before starting up the node. It uses the /**etc**/**cstartSeq** file, if one exists, or the default sequence rules.

If node **cs6** is already running, you must specify either **-z** or **-Z** to force the node to be reset. This example resets the node **cs6** in the current system partition without checking any sequencing dependencies that **cs6** has with other nodes:

```
cstartup -Z cs6
```

You can also identify nodes by node number or range of node numbers using the **-N** flag. The following command starts node numbers 4 through 12 in the current system partition.

```
cstartup -N 4-12
```

Shutting Down the SP System

To shut down the SP system, login on the control workstation and issue the **kinit** command, specifying a principal name for which there is an entry in the **hardmon** acls file with control authorization for the frames to be shutdown. Then, use the **cshutdown** command. The SP **cshutdown** command is similar to the workstation **shutdown** command, except it has these advantages over using **shutdown** to shut down each node of an SP:

- **cshutdown** provides a single point of control.

  Using one **cshutdown** command on the control workstation, you can shut down all or selected nodes.

- The sequencing of node shutdown and reboot is automated.

  You can use the /**etc**/**cshutSeq** file to control the order in which nodes are shut down, or you can let the system determine the order based on System Data Repository information about servers and clients.

- Special user-defined subsystems can be notified of impending node shutdown.

  The /**etc**/**subsysSeq** file lists these special subsystems and describes any sequencing relationships between them.

Note: The **cshutdown** command powers off the node after the **shutdown** process completes.

You can shutdown nodes with a single **cshutdown** command as shown in Table 2 on page 23.

Table 2. Specifying Nodes to Shut Down with cshutdown.

Command	Meaning
cshutdown ALL	Shuts down all nodes in the current system partition.
cshutdown *target_nodes*	Shuts down specific nodes in the current system partition. All nodes listed must be in the current system partition or the command will fail.
cshutdown -G ALL	Shuts down all nodes in the physical SP system.
cshutdown -G *target_nodes*	Shuts down specific nodes in the physical SP system, regardless of system partitioning.

The complete syntax of the **cshutdown** command is described in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*. Understanding how shutdown processing works may help you understand some of **cshutdown**'s options; this is described in "The Phases of Shutdown Processing" on page 24.

Halting Nodes

You can halt all nodes in the current system partition with one command, using:

```
cshutdown ALL
```

The **-h** (halt) flag is the default; you don't need to use it when you invoke **cshutdown**. The nodes are shut down in the sequence specified by **/etc/cshutSeq**. If this file doesn't exist, the nodes are shut down in a default sequence based on the SDR. If **/etc/subsysSeq** exists, those subsystem interfaces are invoked during shutdown.

Halting selected nodes is similar. To halt nodes **cs3**, **cs5**, and **cs7** in the current system partition, use the command:

```
cshutdown cs3 cs5 cs7
```

The command checks for dependencies before shutting down the nodes. If **/etc/cshutSeq** exists, those rules apply; otherwise, **cshutdown** applies default sequencing rules based on the SDR. If **/etc/subsysSeq** exists, those subsystem interfaces are notified of the shutdown of the target nodes.

Rebooting Nodes

Reboot all nodes in the current system partition by using the **cshutdown** command with the **-r** (reboot) flag:

```
cshutdown -r ALL
```

The target nodes are halted according to the sequencing rules in **/etc/cshutSeq**, and restarted according to the rules in **/etc/cstartSeq**. If these files don't exist, the default sequencing rules are determined at shutdown and startup.

You can control how the system starts during reboot by passing **cstartup** flags with the **-C** flag. For example, to pass the **-E** flag to tell **cstartup** to ignore the sequence file when rebooting current system partition nodes numbered 4 and 5, use this command:

```
cshutdown -r -C "-E" -N 4-5
```

To pass multiple startup flags, you must escape any space characters in the string. The double quotes in this example ("-E") do that, although they are not strictly required with a single flag.

Rebooting from Maintenance Mode

If you bring a node down to maintenance mode, you must ensure file system integrity before you reboot the node.

The **cshutdown** command, which runs from the control workstation, will be unable to **rsh** to the node to perform the node shutdown phase processing which includes synching of the file systems. We recommend that you issue the **sync** command three times in succession from the node console before running **cshutdown**. This is especially important if any files were created while the node was in maintenance mode.

To determine which nodes are in maintenance mode, run the **spmon -d** command and look for the combination of **nodePower** on and **hostResponds** no.

Notifying Users of an Impending Shutdown

You probably want to warn your users if nodes are going down, so they can wrap up any work in progress. By default, when you issue a **cshutdown** command, a message is sent to all users logged into the affected nodes, telling them how long they have before shutdown. You can customize the message as well as the time when the system shuts down.

To power off the entire SP system at 9 pm, issue the following command:

```
cshutdown -T 21:00 -M "All nodes will power off at 9 pm." -G ALL
```

To power off all nodes in the current system partition at 9 pm, issue the following command:

```
cshutdown -T 21:00 -M "All nodes will power off at 9 pm." ALL
```

Suppose you need to do an emergency reboot of the pair of nodes named **cs1** and **cs2** in the current system partition, but want to allow 10 minutes for jobs to complete. You can notify your users and schedule the reboot with this command:

```
cshutdown -r -T 10 \
          -M "Reboot on cs1&2 in 10 minutes--don't submit new jobs." cs1 cs2
```

The Phases of Shutdown Processing

cshutdown command processing has these phases:

1. Notifying users and terminating nonroot processes
2. Subsystem shutdown
3. Node shutdown
4. Restarting nodes, if requested

Notifying Users and Terminating Nonroot Processes

User processes that are not started by root and are running on the target nodes are sent a SIGTERM followed, 30 seconds later, by a SIGKILL. This gives user processes that handle SIGTERM a chance to do whatever cleanup is necessary.

Subsystem Shutdown

The subsystem shutdown phase tells special subsystems to terminate their own operations on the target nodes before the node phase shuts down the target nodes. Groups for which **/etc/subsysSeq** doesn't define a sequencing relationship are invoked concurrently. Subsystems not listed in **/etc/subsysSeq** are not notified of the pending system shutdown.

If **cshutdown** is issued without the **-Y** flag and a nonzero (failure) code is returned, you receive a prompt allowing you to continue, to quit, or to start a subshell so that you can check **/var/adm/SPlogs/cs/cshut.***MMDDhhmmss.pid* for messages and refer to the documentation for the subsystem. When you leave the subshell, you are prompted with the same choices.

The subsystem phase of shutdown waits for all subsystems to return before continuing with the node phase. If a subsystem does not return, the **cshutdown** command waits indefinitely. Use the **-W** flag to specify a time-out value.

Node Shutdown

The node phase operates on nodes specified on the **cshutdown** command line. The **/etc/cshutSeq** file controls the sequencing of the node phase operation.

Node shutdown has potentially two levels of coordinated concurrent operations:

- More than one group of nodes can be shut down concurrently.
- Within each one of those groups, more than one node can be shut down concurrently.

They are coordinated in that one group may wait for the completion of some other specific groups before starting to shut down.

In a halt operation, the node phase applies the standard workstation shutdown on the target nodes. The nodes are halted in the sequence specified in **/etc/cshutSeq**, if it exists, or in the default sequence. Nodes not listed in **/etc/cshutSeq** may be halted concurrently with any other nodes.

Restarting the Nodes, If Requested

If you request a system reboot with the **-r** flag, the node phase halts the nodes in the sequence specified in **/etc/cshutSeq** and then restarts the nodes consistent with the initialization sequence specified in **/etc/cstartSeq**. If these files do not exist, the default sequence is used. It does not power on any nodes that were powered off when the **cshutdown** command was issued. If you use the **-r** flag, the only nodes powered on during reboot are those powered off during shutdown.

Checking the Status of Startup and Shutdown Requests

Each time you invoke **cstartup** or **cshutdown**, the system creates a log of activities related to that command. The logs contain an entry for the initiation and completion of operations on each node. **cshutdown**'s log includes information about subsystem shutdown operations. These logs are in the directory **/var/adm/SPlogs/cs**. Two types of log files are created:

cstartup **/var/adm/SPlogs/cs/cstart.***MMDDhhmmss.pid*

cshutdown **/var/adm/SPlogs/cs/cshut.***MMDDhhmmss.pid*

Each log file's name contains the date and time that the command was issued, in the format *MMDDhhmmss.pid*, where:

MMDDhhmmss	The timestamp
pid	The process ID of the **cstartup** or **cshutdown** command

Defining Startup and Shutdown Sequence Files

Relationships between nodes must be defined in the /**etc**/**cstartSeq** and /**etc**/**cshutSeq** files. You define groups of nodes in these files. Nodes in the same group are started up or shut down concurrently. You also define the order in which groups of nodes are processed. Sequencing files can contain nodes from any and all system partitions; sequencing rules are applied without regard to system partitions. Groups without specified sequencing relationships are started up concurrently. Nodes not listed in /**etc**/**cstartSeq** may be started up concurrently with any other nodes. Nodes not listed in /**etc**/**cshutSeq** may be shut down concurrently with any other nodes.

The sequencing of special subsystems is handled similarly. Subsystems and groups of subsystems are defined in the /**etc**/**subsysSeq** file. Subsystems without specified sequencing relationships are shut down concurrently. Subsystems must be defined in this file. If they are not defined, they are not invoked during shutdown.

Planning for Concurrent Startup and Shutdown

It's important that you allow the **cstartup** and **cshutdown** commands to do as much work as possible concurrently. This ensures optimal performance when starting up and shutting down a system with many nodes and/or subsystems. There are potentially two levels of coordinated concurrent operations:

- More than one group of nodes or subsystems can be started up or shut down concurrently.
- Within each one of those groups, more than one node or subsystem can be started up or shut down concurrently.

They are coordinated in that you can specify that one group must wait for the completion of some other specific groups before being started up or shut down.

Format Rules for Sequence Files

A line in /**etc**/**cstartSeq**, /**etc**/**cshutSeq**, and /**etc**/**subsysSeq** can be one of the following:

- Comment line starting with a pound sign (#). A comment can also start within a line.
- Group definition line containing a group name followed by a colon (:), followed by one or more node names (/**etc**/**cstartSeq**, /**etc**/**cshutSeq**) or subsystem names (/**etc**/**subsysSeq**). To put multiple nodes or subsystems in a group, separate their names with commas (,) and, optionally, blanks.
  - In /**etc**/**cstartSeq** and /**etc**/**cshutSeq**, a node name is the en0 Ethernet hostname of the node. These are also known as **reliable_hostname** in the **Node** object class in the System Data Repository.

  - In /**etc**/**subsysSeq**, a group definition line is the arbitrary subsystem group name, followed by a colon (:), followed by the full path name of the command you want issued at shutdown.

  The group names in the file are arbitrary names and treated as tokens.

  Each node or subsystem can be in only one group. A group can consist of from 1 to all your nodes or subsystems.

- Sequence line containing a greater than symbol (>). Thus, in /**etc**/**cstartSeq**, `groupA > groupB` says to start groupA before starting groupB. And, in /**etc**/**cshutSeq**, `groupA > groupB` says to shut down groupB before shutting down groupA.

  A group cannot both precede and trail the same group. That is, the sequence `A > B > A` is not valid.

The lines can be in any order. Each node must be assigned to a group if you wish to control its sequencing.

A group or node name cannot contain a blank, comma, newline, tab character, greater than symbol (>), colon (:), parentheses (() or ()), or pound sign (#). Case is significant in node and group names.

Testing Your Sequence Files

You can verify that the /**etc**/**cstartSeq** file has the correct format without powering on or resetting any nodes by specifying:

```
cstartup -k ALL
```

The **-k** flag tells **cstartup** not to reset any nodes, just check the sequence file. If the sequence file is correct, you'll get the following message:

```
cstartup: No circular dependencies were detected
```

Verify that the /**etc**/**cshutSeq** and /**etc**/**subsysSeq** files have the correct format without shutting down any nodes by specifying:

```
cshutdown -kF ALL
```

The **-k** flag tells **cshutdown** not to shutdown any nodes, just check the sequence file. The **-F** flag tells **cshutdown** not to issue any messages to users.

Getting Started: Capturing the Default Sequences

An easy way to create your /**etc**/**cstartSeq** and /**etc**/**cshutSeq** files is to capture the default group definitions and sequencing relationships that are based on the boot/install server information in the System Data Repository. You can then add to or modify them as required for your installation.

Create the default /**etc**/**cstartSeq** file using this command:

```
seqfile -b > /etc/cstartSeq
```

The **-b** flag tells **seqfile** to include boot/install relationships, which are important only during system startup.

To create the default /**etc**/**cshutSeq** file, use this command:

```
seqfile > /etc/cshutSeq
```

After you've created the default files, you can edit them to reflect sequencing relationships based on other aspects of your installation's configuration.

Sample Startup File /etc/cstartSeq

Here is a simple file for sequencing startup:

```
ServerA > ClientsOfA
ServerA: node2
ClientsOfA: node3, node4, node5, node6
```

The line `ServerA: node2` defines node 2 as the group ServerA. The line `ClientsOfA: node3, node4, node5, node6` defines nodes 3 through 6 as a group that can be started up concurrently. The sequence line `ServerA > ClientsOfA` says that ServerA must be completely started up before starting up the nodes that are clients of server A (nodes 3 through 6).

This sample file is slightly more complex:

```
groupA > groupB
groupA > (groupF, groupE) > groupD > groupC
groupD > groupG
groupA: node0  #node0 is the node that node1-node7 depend on.
groupB: node1
groupC: node2
groupD: node4
groupE: node5
groupF: node6, node7
groupG: node3
```

The line `groupA > ( groupF, groupE ) > groupD > groupC` says to start up groupA. When groupA is up, concurrently start up groupF and groupE. When they are up, start up groupD. Finally, when groupD is up, start up groupC. The line `groupA > groupB` says that groupB depends on groupA, but may be started concurrently with groupE and groupF. No other groups have dependencies on groupB.

If your command line specifies a node but omits another that is listed in your **/etc/cstartSeq** or **/etc/cshutSeq** files as its leading node, the sequencing rules still apply unless overridden by the **-X** flag.

Sample Shutdown File /etc/cshutSeq

Here is one of the sample startup sequence files, interpreted now as a shutdown file:

```
ServerA > ClientsOfA
ServerA: node2
ClientsOfA: node3, node4, node5, node6
```

The line `ClientsOfA: node3, node4, node5, node6` defines nodes 3 through 6 as a group that can be shut down concurrently. The line `ServerA: node2` defines node2 as the group (of one node) called ServerA. The line `ServerA > ClientsOfA` says that the nodes that are clients of server A (nodes 3 through 6) can be shut down concurrently, but must be shut down before node2 in group ServerA.

Sample Subsystem File /etc/subsysSeq

A sample file for subsystem shutdown could be:

```
ssgroupA: /full/pathname/subsystemA
ssgroupB: /local/subsystemB, /local/subsystemC
ssgroupA > ssgroupB
```

The line `ssgroupA: /full/pathname/subsystemA` defines the group ssgroupA, consisting of the single subsystem invoked by the /full/pathname/subsystemA command. The line `ssgroupB: /local/subsystemB, /local/subsystemC` defines a group consisting of the two subsystems /local/subsystemB and /local/subsystemC.

The line `ssgroupA > ssgroupB` says to first notify ssgroupB before nodes are shut down. When both subsystems in ssgroupB have returned, then notify the single subsystem in ssgroupA.

Creating a Special Subsystem

Read this section if you write your own special subsystem and want the **cshutdown** command to notify the subsystem when nodes are being halted or rebooted. The SP system has no predefined special subsystems.

The special subsystem interface is a user-provided command that is invoked during the subsystem phase of SP system shutdown processing. The command is defined in the **/etc/subsysSeq** file and must reside on all nodes of the SP system; the **cshutdown** command invokes the subsystem interface on a randomly selected node in the current system partition.

The following sections define the interface to a special subsystem. This interface consists of:

- The inputs that the **cshutdown** command passes to your special subsystem
- The actions your subsystem must perform.

Note: The special subsystem interface is called *after* nonroot processes are killed. If you want your subsystem to take advantage of shutdown notification, some part of it must be running as root.

Input to the Special Subsystem Interface

The **cshutdown** command invokes the special subsystem interface with the following standard parameters:

-r or **-h** Indicates whether the **cshutdown** command is halting (**-h**) or restarting (**-r**) the target nodes.

-b *IPaddress* Specifies, in dotted decimal form, the IP address of the calling program, in this case, the control workstation.

-f /var/adm/SPlogs/cs/cshut.*MMDDhhmmss***.***pid*
Specifies the log file to which the subsystem interface program should append a record of status and activities.

target_nodes The list of node numbers of the target nodes specified on the **cshutdown** command line. The numbers are passed at the end of the argument list.

For example, if the **/etc/subsysSeq** file contains the following line:

```
ssgroupA: /full/pathname/subsystemA
```

and this **cshutdown** command is issued on the control workstation whose IP address is 129.40.64.70:

```
cshutdown -h -N 1 2 3
```

the **cshutdown** command issues the following command on a randomly-selected node during the subsystem phase of shutdown processing:

```
/full/pathname/subsystemA -h -b 129.40.64.70 \
                          -f /var/adm/SPlogs/cs/cshut.time.pid \
                          1 2 3
```

What Your Special Subsystem Must Do

In addition to any processing specific to the subsystem, the special subsystem must call the **csLogger** command to record its activities. Your program must call **csLogger** with the following required parameters:

-b *IPaddress*
: Specifies, in dotted decimal form, the IP address of the calling program, in this case, the control workstation.

-f /var/adm/SPlogs/cs/cshut.*MMDDhhmmss***.***pid*
: Specifies the log file to which the subsystem interface program should append a log of status and activities.

logString
: Specifies a string that will be appended to the specified log file. This is the last parameter.

Using the example started above, if the **cshutdown** command calls your program with this command:

```
/full/pathname/subsystemA -h -b 129.40.64.70 \
                          -f /var/adm/SPlogs/cs/cshut.time.pid \
                          1 2 3
```

then the **/full/pathname/subsystemA** command could make calls similar to these, to log the start and completion of subsystem shutdown:

```
csLogger -b 129.40.64.70 \
         -f /var/adm/SPlogs/cs/cshut.time.pid \
         subsystemA:  initiating shutdown at 15:43.55 on 1 Apr 94.

csLogger -b 129.40.64.70 \
         -f /var/adm/SPlogs/cs/cshut.time.pid \
         subsystemA:  shutdown complete at 15:46.23 on 1 Apr 94.
```

Chapter 4. Parallel Management Commands

This chapter describes commands provided with the SP system that are useful for performing common tasks on multiple nodes in parallel.*

These commands, along with the **sysctl** function and SP authenticated remote commands, provide for management of the SP System from a single point of control.

Using dsh to Execute Parallel Management Commands

You can use **dsh** from a single point of control to execute commands on:

- All nodes in the current system partition (**-a** option)
- All nodes in the physical SP system (**-aG** option)
- Selected nodes (**-w** option)

The target nodes need not be limited to those in your SP System. **dsh** will execute commands on any host in your network to which you can issue commands through the SP authenticated remote command **rsh** (see “Security Considerations” on page 33). Although **dsh** uses **rsh** to issue remote commands, **dsh** offers better performance than an **rsh** loop because the remote commands run concurrently.

Inputting to dsh

dsh provides two ways for you to specify input:

1. **dsh** reads lines from standard input and executes each line as a command via **rsh**. Specify commands in **rsh** syntax. Pipes, filters, output redirection, and paths can be used on the remote nodes in the same way they are used in **rsh**.

   The following **dsh** command targets three nodes, reads commands from stdin, and filters the output of the **ps** command remotely:

   ```
   $ dsh -w host1,host2,host3
   dsh> ps -ef "|" grep root
   ```

   Note: Any input line beginning with an exclamation point is passed directly to the shell on the local host.

2. **dsh** accepts commands specified on the command line. For example, to issue the **ps** command on three nodes and filter the results remotely, specify:

   ```
   dsh -w host1,host2,host3 ps -ef "|" grep root
   ```

* The parallel management commands described in this chapter were influenced by (but are not an implementation of) the tools described in Gropp, William and Ewing Lusk. 1994. Scalable UNIX tools on parallel processors. In *Proceedings of the Scalable High Performance Computing Conference*, 56-62. IEEE.

© Copyright IBM Corp. 1995, 1997

Targeting Nodes for dsh

dsh sends commands to a node set called the **working collective**. **dsh** assembles the working collective from the first existence of one of these sources:

1. A list of hostnames you specify on the command line with the **-a**, **-G**, **-w** flags.
2. The contents of a file you specify in the **WCOLL** environment variable.

If neither of these exist, an error has occurred and no commands are issued.

You can have commands run on the working collective concurrently or in sequence. Default processing is done concurrently, but you can specify a maximum number of nodes for concurrent execution to prevent system degradation. When this maximum number, or *fanout* is reached, results from outstanding **rsh**s are awaited before any further commands are run. These **rsh** results are displayed as soon as they return from the remotely executed commands.

This example specifies a maximum of 8 concurrent commands even though the working collective is defined as the entire system partition.

```
dsh -f 8 -a  cat /var/adm/SPlogs/sysman/"*"config.log"*" | pg
```

Note: Because **dsh** starts processes for each host in the working collective, there may be cases where a smaller fanout will result in less contention for resources on the local machine, and thus better overall performance, than the greater concurrency of a larger fanout.

Output from dsh

dsh displays returned information grouped by hostname. Stdout of the remotely executed commands goes to the stdout of **dsh**, and stderr of remotely executed commands goes to the stderr of **dsh**.

All **dsh** stderr and stdout lines are prefixed by the name of the host that sent them. You can format **dsh** stderr and stdout lines by piping them to **dshbak**, which strips off the hostnames and displays the lines grouped by host in alphabetic sequence, as shown in this example:

```
dsh -w host1,host2,host3 cat /etc/passwd 2>&1 | dshbak
```

Alternatively, **dshbak** can be specified with the **-c** option. This will cause any identical output from two or more nodes to be shown only once, with the hostnames displayed above the output. Remember, however, that the most efficient way to filter large amounts of output from parallel commands is to filter on the nodes before the output is returned to the workstation from which the parallel command was issued.

Error Handling

No special error recovery is provided for **dsh** command failure on remote hosts. If **dsh** finds that a node in the working collective is down, no further commands will be sent to that node unless you specify the **-c** (continue) flag on the command line. If hosts are down, the underlying **rsh** will time out in approximately 2.5 minutes.

Security Considerations

dsh uses the SP authenticated remote command **rsh**. This version of **rsh** attempts to authenticate users via Kerberos. If authentication fails, the AIX version of **rsh** is used instead. The authenticated version of **rsh** allows system administration without the requirement for the existence of a root level **.rhosts** file.

The hostlist Command

The **hostlist** command can be used to specify hostnames for parallel command execution via **dsh** or **sysctl**. The command writes hostnames to stdout based on various criteria:

- The contents of a **dsh** working collective file
- The contents of a POE hostname file
- All the nodes in the current system partition (**-a** option)
- All the nodes in the physical SP system (**-aG** option)
- Ranges of slots on particular frames
- A set of nodes to exclude
- A range of node numbers
- Explicitly listed hostnames

Before the names are written, there is an optional check of whether the nodes are responding.

The output from **hostlist** can be written to a file, which can then be used as a working collective file for **dsh**. Both **dsh** and **sysctl** can read hostnames from stdin, so the output of **hostlist** can be piped directly to **dsh** or **sysctl**.

Examples

To issue a command to the node in the first slot in each of the first 4 frames:

```
hostlist -s 1-4:1 | dsh -w - command
```

To run the program on the nodes on all slots for frame 1 and slots 1–3 for frame 3, and also on host **otherone**:

```
hostlist -n 1-16,33-35 -w otherone > /tmp/wcoll
WCOLL=/tmp/wcoll dsh program
```

To determine all the nodes in the SP system on which one is authenticated, and to bypass the nodes that are not responding and “badnode”, issue:

```
hostlist -av -e badnode | sysctl -c - -L whoami | dshbak -c
```

Note:

> The **-L** option on **sysctl** is necessary if the **dshbak** filter is used. It causes the output lines of **sysctl** to be formatted with the hostname preceding each line.

Output may look like this:

```
HOSTS===========================================================
host1             host2             host3             host4
host5             host6             host8
================================================================
joe.@A.B.C

HOSTS===========================================================
host7
================================================================
unknown
```

If the **-c** option is not specified, output will not be collapsed:

```
HOST: host1
===========
joe.@A.B.C

HOST: host2
===========
joe.@A.B.C
⋮
```

List of Parallel Management Commands

The SP system provides several parallel management commands. They all have a common syntax and are based on **dsh**, **sysctl**, **hostlist**, and **dshbak**. More complete descriptions of each is provided in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

The syntax of the parallel management commands is:

command_name remote_host_specification cmd_args_flags

where

command_name The name of a parallel command

remote_host_specification
Indicates which hosts the command should be executed on. Can be a range of slots, a list of hostnames, a set of flags to the **hostlist** command, or read from stdin.

cmd_args_flags The flags and arguments to the command

Where appropriate, the output from the parallel management commands is filtered as described above.

The following parallel management commands are based on **dsh**.

pexec

This application issues a command in parallel on the specified hosts. Any arguments to the commands are passed along with the command name. Results are filtered as described in “Output from dsh” on page 32. For example, to show users on nodes in slots 5–8:

```
pexec 5-8 who
```

pexscr

This application executes commands in parallel on different hosts, including the local host. Unlike **pexec**, the commands may differ on different hosts. Lines consisting of *hostname: command* are read from stdin. When the lines have been read, each command is executed on the corresponding host in parallel. Remote commands are executed via **rsh** and local commands are executed directly. For example, you could create the following file /**etc**/**doit** on r05n09:

```
r05n09: /usr/lpp/ssp/rcmd/bin/rcp /tmp/bozoid tserv11:/tmp/bozoid
r05n09: /usr/lpp/ssp/rcmd/bin/rcp /tmp/bozoid r05n01:/tmp/bozoid
r05n01: backup_script
r05n15: backup_script
```

Then, to copy the local file /**tmp**/**bozoid** to tserv11 and r05n01 and simultaneously start backup scripts on r05n01 and r05n15:

```
pexscr < /etc/doit
```

pcp

This command distributes a local file or directory to remote hosts in parallel. It uses the SP authenticated remote command **rcp**. To use this command, you must be authorized to issue commands of the form:

```
rcp localfile remotehost: remotefile
```

for each of the hosts to which the file is to be copied. To recursively copy /**etc**/**new**/ on the local host to /**etc**/**new1**/ on the host in the first slot of each of the first four frames:

```
pcp "-s 1-4:1" -r /etc/new/ /etc/new1/
```

p_cat

This application issues the **cat** command in parallel on the specified hosts. Any arguments to the command are passed along with the command name. Results are not filtered. For example,

```
p_cat "-w host1,host2,host3" /etc/file > localfile
```

copies the contents of the /**etc**/**file** files on each of the hosts to the **localfile** file on the local host. **localfile** would consist of unmodified lines from /**etc**/**file** on **host1**, on **host2**, and on **host3** concatenated together.

pls, pfind, prm, pps, pmv

These commands issue the **ls**, **find**, **rm**, **ps**, or **mv** commands in parallel on the specified hosts. Any arguments to the AIX commands are passed along with the command name. Results are filtered. For example,

```
pls "-av -e badnode" /
```

lists the contents of the root directory on all the responding hosts in the SP system with the exception of “badnode”. Note that the **-i** flags to **mv** and **rm** are not supported because **dsh** doesn't propagate the stdin of the remote hosts.

Examples: For example, a working collective could be set up with **hostlist**:

```
hostlist -av > /tmp/wcoll
export WCOLL=/tmp/wcoll

pps "" -ef
```

would list information about all the processes on the nodes in the working collective.

```
pfind "" /usr/lpp/ssp/bin/  -name "pfind" -print
```

may result in the following:

```
HOSTS ===========================================================================
r05n01              r05n03dx            r05n05              r05n07
r05n09              r05n10              r05n13              r05n14dx
r05n15              r05n16
 ===============================================================================
 /usr/lpp/ssp/bin/pfind
```

```
pmv "" /tmp/foo /tmp/newfoo
```

would rename /**tmp**/**foo** on all the working collective.

```
pls "" /tmp/\\*foo
```

could result in the following output:

```
HOSTS ===========================================================================
r05n01              r05n03dx            r05n05              r05n07
r05n09              r05n10              r05n13              r05n14dx
r05n15              r05n16
===============================================================================
/tmp/newfoo
```

Note that the shell metacharacter is escaped twice, once to protect it from the shell, and once to protect it when the **rsh**'s are done by **dsh**.

```
prm "" /tmp/newfoo
```

would remove /**tmp**/**newfoo** from all the nodes in the working collective.

ppred

This application issues a **ksh** test on each node in parallel and runs commands based on the results of the test. For example,

```
ppred 1-16 '-f /etc/foo' 'echo Here' 'echo Not here'
```

could result in output like the following:

```
HOSTS============================================================
host1             host2             host3             host4
host5             host6             host8
=================================================================
Here

HOSTS============================================================
host7             host9             host10            host11
host12            host13            host14            host15
host16
=================================================================
Not here
```

pfps

pfps is a parallel process find application. It takes arguments similar to those of the **find** command, but operates on processes instead of files. In addition, access control list-based permissions to **nice** or **kill** processes is provided. For example, to signal all the “daemond” processes in the SP system partition in parallel, a user that was in the /**etc/sysctl.pfps.acl** file on each node could issue the following:

```
pfps “-a” -tn sysctld -kill HUP
```

To list the **sysctl** daemon processes on nodes **r05n15** and **r05n16**, the following command:

```
pfps “-w r05n15,r05n16” -tn sysctld -print
```

produces this output:

```
HOSTS ===================================================================
r05n15
=========================================================================
USER       PID %CPU %MEM   SZ  RSS    TTY STAT    STIME  TIME COMMAND
root     11172  0.0  1.0  452  420      - S    11:54:47  0:00
/usr/lpp/ssp/bin/sysctld -l /var/adm/SPlogs/sysctl/sysctld.log
root     12497  0.0  1.0  300  360      - S    06:24:19  0:00
/usr/lpp/ssp/bin/sysctld -l /var/adm/SPlogs/sysctl/sysctld.log

HOSTS ===================================================================
r05n16
=========================================================================
USER       PID %CPU %MEM   SZ  RSS    TTY STAT    STIME  TIME COMMAND
root      7089  0.0  1.0  448  420      - S    11:54:48  0:00
/usr/lpp/ssp/bin/sysctld -l /var/adm/SPlogs/sysctl/sysctld.log
root      8423  0.0  1.0  296  360      - S    13:01:31  0:00
/usr/lpp/ssp/bin/sysctld -l /var/adm/SPlogs/sysctl/sysctld.log
root     10418  0.0  1.0  512  504      - S
```

pdf

pdf is a parallel command that performs function similar to **df**. It provides more information than the **df** command. To get filesystem information from all the nodes on the first slots of each of four frames:

```
hostlist -s 1:1-4 > /tmp/wcoll
export WCOLL=/tmp/wcoll
pdf “”
```

To find information on the /**var** filesystem on the nodes in the working collective:

```
pdf "" /var
Filesystem               Size-KB Used-KB Free-KB %Free  iUsed  iFree %iFree
==================       ======= ======= ======= =====  =====  ===== ======
HOST: r05n01
============
/var                        8192    2452    5740   71%    279   1769    87%

HOST: r05n03dx
==============
/var                        8192    1988    6204   76%    214   1834    90%

HOST: r05n05
============
/var                        8192    2080    6112   75%    223   1825    90%

HOST: r05n07
============
/var                        8192    2080    6112   75%    223   1825    90%

HOST: r05n09
============
/var                        8192    2088    6104   75%    223   1825    90%

HOST: r05n10
============
/var                        8192    2448    5744   71%    286   1762    87%

HOST: r05n13
============
/var                        8192    2220    5972   73%    229   1819    89%

HOST: r05n14dx
==============
/var                        8192    2012    6180   76%    221   1827     90%

HOST: r05n15
============
/var                        8192    2368    5824   72%    220   1828    90%

HOST: r05n16
============
/var                        8192    2100    6092   75%    225   1823    90%
```

pfck

pfck displays file system information in parallel based on various usage criteria. For example, to find information on all the file systems on the hosts in POE host list file **hostlist** that have used more than 100000 KB:

```
MP_HOSTFILE=./hostlist pfps "" -s 100000
```

To display the filesystems on hosts **r05n16** and **r05n15** that have more than 90% of their space used:

```
pfck "-w r05n15,r05n16" -pu 90
Filesystem                Size-KB Used-KB Free-KB %Free  iUsed  iFree %iFree
==========                ======= ======= ======= =====  =====  ===== ======
HOST: r05n15
============
/usr                       331776  327640    4136    2%  17673  66295    79%

HOST: r05n16
============
/usr                       331776  327640    4136    2%  17673  66295     79%
```

To find all the /**var** filesytems in the SP system that are 95% full:

```
pfck "" -pu 95 | egrep "(var|HOST)"

 Filesystem                Size-KB Used-KB Free-KB %Free  iUsed  iFree %iFree
 ==========                ======= ======= ======= =====  =====  ===== ======
 HOST: r05n10
 /var                       335872  331412    4460    2%  17819  66149    79%
 HOST: r05n13
 /var                       335872  331632    4240    2%  17820  66148    79%
 HOST: r05n14dx
 /var                       335872  331388    4484    2%  17817  66151    79%
```

Chapter 5. Managing File Collections

The SP system installs file collection technology by default, to simplify the task of maintaining duplicate files on multiple machines. In the delivered system, the files that are required on the control workstation, boot/install servers, and processor nodes belong to file collections. Grouping these files into file collections and using the provided tools to manage them allows you to easily maintain their consistency and accuracy on multiple nodes in the SP system.

This chapter explains the file collections delivered with the system. It shows you how to report on, update, and maintain the system file collections, as well as how to build file collections for your own purposes. It also explains the basic concepts you need to understand before working with file collections, such as what file collections are, how they differ from standard directories and files, and how they are used in the SP system.

The file collection technology provides you with the means to perform both beginning basic tasks and optional advanced tasks. For the basic tasks you should understand the fundamental concepts and be familiar with the delivered file collections. The more advanced tasks require an understanding of the master files, what they contain, and how they work.

Information is presented on the following basic file collection tasks:

- Understanding file collections
- Reporting file collection information
- Verifying file collections using **scan**
- Updating files in a file collection
- Adding and deleting files in a file collection

Information is presented on the following advanced file collection tasks:

- Building a file collection
- Installing a file collection
- Refusing files in a file collection
- Modifying the file collection hierarchy
- Removing a file collection

Understanding File Collections

A *file collection* is a set of files and directories that are duplicated on multiple machines in a network and managed by tools that simplify their control and maintenance. Included with the administrative software is a program called **supper**. The **supper** perl program (/**var**/**sysman**/**supper**) uses the Software Update Protocol (SUP) to manage the SP file collections and transfer them across the system.

© Copyright IBM Corp. 1995, 1997

Terminology

The following terms are used when defining file collections:

Resident A file collection that is installed in its true location and able to be served to other systems

Available A file collection that is not installed in its true location but able to be served to other systems

File collections have unique features:

- A file collection directory does not contain the actual files in the collection. Instead, it contains a set of master files to define the collection. Some master files contain rules to define which files can belong in the collection and others contain control mechanisms, such as time stamps and update locks. These files are explained in "Directory and Master Files" on page 46.
- You handle files in a collection with special procedures and **supper** commands rather than with the standard AIX file commands. The **supper** commands interpret the master files and use the information to install or update the actual files in a collection. You can issue these commands in either batch or interactive mode. The procedures and use of the **supper** commands are detailed throughout this chapter. You can find the command syntax and descriptions in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.
- File collections can be either *primary* or *secondary*. Primary file collections can contain other secondary file collections. When a primary collection that contains secondary collections is installed, the secondary collections are *available* but not *resident*. This means they can be served from this location but not actually executed or used there. "File Collection Types" on page 43 provides a further explanation of these concepts.
- File collections require special entries in the **/var/sysman/file.collections** and **/var/sysman/sup/.resident** files to define them to the **supper** program. They also require a symbolic link in the **/var/sysman/sup/lists** file pointing to their **list** master file. See "Building a File Collection" on page 56 for an explanation of these entries.
- File collections also require a unique, unused user ID for **supman**, the file collection daemon, along with a unique, unused port through which it can communicate. The default installation configures the user ID, *supman_uid*, to 102 and the port, *supfilesrv_port*, to 8431. You can change these values using SMIT or the **spsitenv** command.
- The file collection daemon, **supman**, requires read access permission to any files that you want managed by file collections. For example, if you want a security file such as **/etc/security/limits** managed, you must add the **supman** ID to the security group. This provides **supman** with read access to files that have security group permission and allows these files to be managed across the SP by file collections. You can add **supman** to the security group by adding **supman** to the security entry in the file **/etc/groups**.

File Collection Types

File collections can be either primary or secondary. A primary file collection can be a stand-alone collection or it can contain a secondary collection. Having a secondary collection allows you to keep a group of files available on a particular machine to serve to other systems without having those files installed.

For example, if you want to have one **.profile** on all nodes and another **.profile** on the control workstation, consider using the **power_system** collection delivered with the IBM Parallel System Support Programs for AIX. This is a primary collection that contains **node.root** as a secondary collection.

- Copy **.profile** to the /**share**/**power**/**system**/**3.2** directory on the control workstation.
- If you issue **supper install power_system** on the boot/install server, the **power_system** collection is installed in the /**share**/**power**/**system**/**3.2** directory. Because the **node.root** files are in that directory, they cannot be executed on that machine but are available to be served from there. In this case, **.profile** is installed as /**share**/**power**/**system**/**3.2**/**.profile**.
- If you issue **supper install node.root** on a processor node, the files in **node.root** collection are installed in the **root** directory and, therefore, can be executed. Here, /**share**/**power**/**system**/**3.2**/**.profile** is installed from the file collection as /**.profile** on the node.

Predefined File Collections

Looking at the file collections delivered with the SP system can help you to understand file collection concepts. You should become familiar with the delivered file collections before you begin to update or add to them.

The SP system provides four predefined file collections:

- **sup.admin**
- **user.admin**
- **power_system**
- **node.root**

You can display information about each collection on a particular machine using the **supper status** command. This reports whether or not the file collection is resident on that machine. To do this, log in as **root** or issue the command remotely. Enter:

```
/var/sysman/supper status
```

You can see a list of the files in a file collection by looking at the **scan** file. See "Verifying File Collections Using scan" on page 51 for instructions on how to do this. The **scan** file contains an entry for each file in the collection that could be installed on a boot/install server or processor node. Certain files in the collection may not exist on one particular system depending on the contents of the **refuse** file on that system. You can find more information about the **scan** and **refuse** files in "Directory and Master Files" on page 46.

SP File Collection Summary

The following table summarizes where these file collections are *resident* or *available* in the SP default configuration.

Control Workstation		Boot/Install Servers		Processor Nodes	
Resident	**Available**	**Resident**	**Available**	**Resident**	**Available**
	sup.admin user.admin power_system node.root	sup.admin user.admin node.root	sup.admin user.admin power_system node.root	sup.admin user.admin node.root	

sup.admin Collection

The **sup.admin** file collection is a primary collection that is *available* from the control workstation, is *resident* (that is, installed) and *available* on the boot/install servers, and *resident* on each processor node.

This file collection is important because it contains the files that define the other file collections. It also contains the file collection programs used to load and manage the collections. Of particular interest in this collection are:

- **/var/sysman/sup** which contains the directories and master files that define all the file collections in the system
- **/var/sysman/supper** which is the Perl code for the **supper** tool
- **/var/sysman/file.collections** which contains entries for each file collection.

user.admin Collection

The **user.admin** file collection is a primary collection that is *available* from the control workstation, *resident* and *available* on the boot/install servers and *resident* on each processor node. This collection is a good place to add other files related to user management, such as **/etc/environment**, **/etc/security/limits**, and **/etc/passwd**. If you want to keep user characteristics the same across nodes, you may want to add **/etc/security/user** to this collection also. However, you may **not** want to have user characteristics the same across nodes. For instance, you may want a user to be able to log in to some nodes, but not to others. In this case, do not add the **/etc/security/user** file to this collection.

power_system Collection

The **power_system** file collection is used for files that are system dependent. It is a primary collection that contains one secondary collection called the **node.root** collection. The **power_system** collection contains no files other than those in the **node.root** collection.

The **power_system** collection is *available* from the control workstation and *available* from the boot/install servers. When the **power_system** collection is installed on a boot/install server, the **node.root** file collection is *resident* in the **/share/power/system/3.2** directory and can be served from there.

node.root Collection

This is a secondary file collection under the **power_system** primary collection. The **node.root** collection is *available* from the control workstation, *resident* and *available* on the boot/install servers and *resident* on the processor nodes. It contains key files that are node-specific.

The **node.root** file collection is *available* on the control workstation and the boot/install servers under the **power_system** collection so that it can be served to all the nodes. We do not install **node.root** on the control workstation because the files in this collection might conflict with the control workstation's own **root** files.

Primary and secondary files are defined in the **file.collections** file (see “Step 5: Update the file.collections File” on page 58).

File Collections Organization

When you complete the SP installation and configuration process, the delivered file collections are organized in a hierarchy with the control workstation as the master server for all the collections.

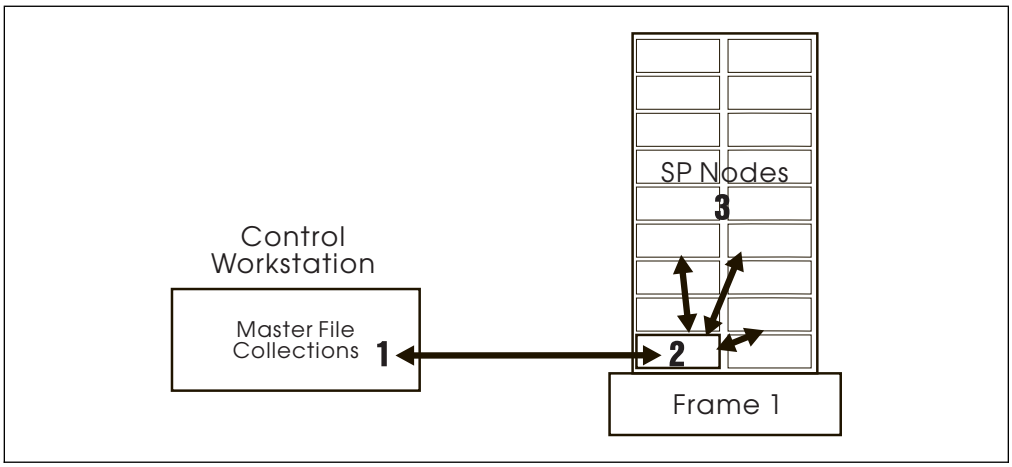


Figure 2. SP Default File Collection Hierarchy

The commands to update the file collections are set up in a **crontabs** file to run hourly in a staggered sequence. Figure 2 shows this organization. Files are updated on the control workstation in the master file collections (**1**). The boot/install servers use **supper update** to request the changes from the control workstation (**2**) and the nodes use **supper.update** to request the changes from the boot/install servers (**3**).

Understanding the File Collection Structure

Before you work with the delivered file collections or build a file collection of your own, you need to understand the directories, master files, what they contain, and how they work.

Directory and Master Files

The **/var/sysman/sup** directory contains master files for all the file collections. Figure 3 shows the structure for this directory.

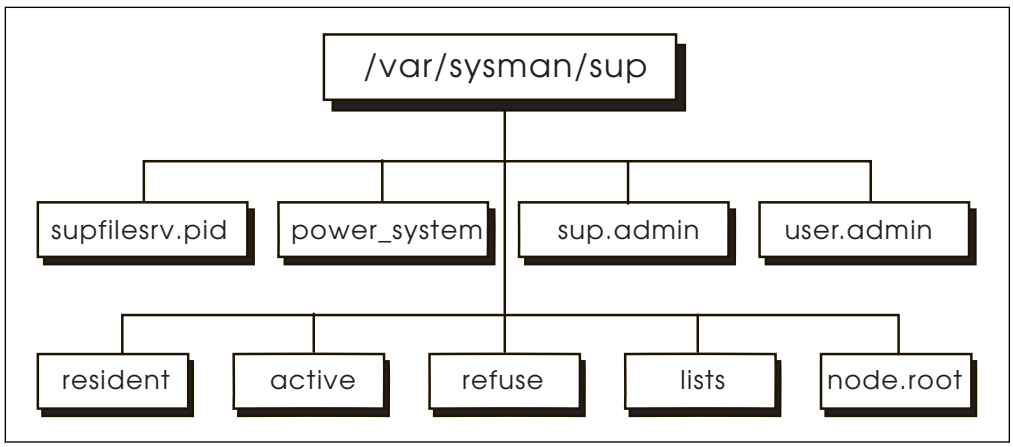


Figure 3. /var/sysman/sup Files and Directories

Following is an explanation of the files and directories.

.active A file identifying the active volume group. You can modify this file with the **supper activate** command. (This file is not on the control workstation.)

.resident Lists each file collection in the SP system. This file needs to be updated when you are adding a new file collection. (This file is not on the control workstation.)

lists A directory that contains links to the **list** files in each file collection.

node.root Directory of the master files in the **node.root** collection.

power_system Directory of the master files in the **power_system** collection.

refuse On a client system, this is a user-defined (you must define the file; it is not supplied) text file containing a list of files to exclude from all the file collections. This allows you to customize the file collections on each system.

You list the files to exclude by their fully qualified names, one per line. You can include directories, but you must also list each file in that directory you want excluded.

sup.admin Directory of the master files in the **sup.admin** collection.

supfilesrv.pid The process ID of the SUP **supfilesrv** process.

user.admin Directory of the master files in the **user.admin** collection.

Each individual file collection is composed of a directory and several master files. Figure 4 on page 47 shows the structure for the **sup.admin** file collection in the **/var/sysman/sup/sup.admin** directory. It includes a sample of the file contents. This example shows eight master files. Other file collections may have fewer master files.

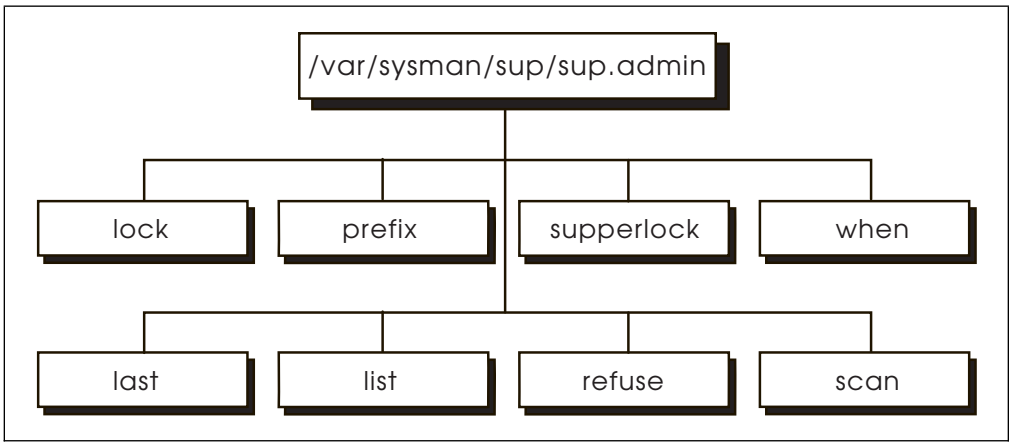


Figure 4. sup.admin Master Files

Following is an explanation of the master files shown in Figure 4.

last A system-maintained list of files and directories that have been updated. This helps identify files to be deleted on a local machine when they are no longer part of the collection.

The **supper files** command reads the **last** file to obtain a current list of files in the collection.

list A user-defined file containing special commands. These commands are interpreted by the **supper** scan process to define which files belong to the file collection relative to the base directory specified in the **prefix** file. You can find the specific format for these commands in "supper list File Syntax, Keywords, and Operands" on page 64

lock An empty file used by SUP to lock a collection and prevent more than one update of the same collection at the same time.

prefix A file containing the name of a directory to be used as the base directory for file references and the starting point for the **supper** scan process. This file is created by the system from data in the **file.collections** file.

refuse A system-created file that contains a list of all the files excluded during the update process. If there are no files for this collection listed in the **refuse** file in the **/var/sysman/sup** directory, the **refuse** file in this directory will have no entries.

scan A system-created file containing a list of files that comprise the collection and their permissions and time stamps. This file is the result of the **supper** scan process. It is optional but recommended for all collections and can cause maintenance problems if you do not keep it current.

supperlock Similar to **lock**, an empty file used by **supper** to lock a collection during updates.

when A system-maintained file containing the time of the last file collection update. This file protects against accidental update of a collection with an old version of the files. You can reset it using the **supper reset** command if you need to override this restriction.

How the Master Files Work

The **supper** master files play a key role during the install, update, and scan processes.

When you run the **supper scan** command for a file collection, the search begins at the point in the directory specified in the **prefix** file and traverses the directory tree to find all the files that meet the criteria defined in the **list** file. The output of this process is the **scan** file which lists the files in the collection. The **scan** file is optional but it documents the file collection contents and eliminates the need for the install and update processes to do the directory search.

The **supper install** and **update** commands use the **scan** file, if it is present, to identify the files to install or update. If a **scan** file is not present, **supper** performs the same directory search to identify the files as it performs when it creates the **scan** file. These commands also check the **refuse** file in **/var/sysman/sup** at each location and bypass any files it contains. As a result, these commands create a **refuse** file in the file collection's directory, listing the files that were bypassed.

During updates, the **lock** and **supperlock** files prevent update access by other SUP or **supper** commands. **supper** checks the update time in the **when** file against the actual file time stamps to verify that you are not updating with an old version of the files and, if so, does not update. It uses the **last** file for additional verification when deleting files that no longer exist in the file collection.

Master File Processing Order: Whether invoked with the **scan, update,** or **install** commands, **supper** performs the search process in the following manner:

prefix Reads this file first to obtain a directory from which it will start the search

list Reads this file next and processes its commands in the following order:

- **omit** Marks these files for omission
- **upgrade** Marks these files for upgrade when updating or installing
- **always** Overrides any files marked for omission

refuse Reads this file last when it is present on a client (boot/install server or processor node) and does not upgrade any of the specified files on that client.

Automatic Updates: The **supper update** commands are included in the **crontabs** file with the default set to run an update on each of the file collections hourly. You can modify the **crontabs** file to run the **supper update** more or less frequently. Refer to “Updating crontabs on the SP Nodes” on page 153 for more information.

Basic File Collection Tasks

In order to maintain the delivered file collections you will need to perform some basic tasks such as reporting on file collection information, updating files that belong to an existing file collection, and possibly adding or deleting files within an existing file collection. For example:

- To become familiar with the delivered collections you might list their name, access point, and whether or not they are resident. You might also want to verify the complete list of all the files in a collection.
- You might update files such as /**etc**/**profile** to change the user default profile.
- You might add files such as /**etc**/**environment** and have them managed by the file collection processes.

You can find basic file collection tasks on pages 41 through 54.

Advanced File Collection Tasks

Once you are familiar with file collection technology, you may want to build your own collections to manage other files on your system that you would like duplicated on multiple machines. You might have a group of files such as application data that you want installed on each node or local tools you want installed on each boot/install server.

The task of building and distributing your own collections is more advanced. It includes creating the directory and master files, adding entries to the **file.collections** and **lists** files, and possibly creating **scan** and **refuse** files. Once you complete these tasks and build your file collection, you can install it on the machines in your system. You can also install a file collection from a server other than its default server.

You can find advanced file collection tasks on pages 61 through 63.

Reporting File Collection Information

supper has subcommands that report information about file collections. These subcommands are useful for verifying information before performing file collection manipulations or for checking results at some point during a procedure.

The following table shows the **supper** subcommand to use, where you would run the command, and the data you can report. See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for additional command options.

Use:	On:	To Report This Data:
status	Control workstation, boot/install servers, and nodes	Name, resident status, and access point of all available file collections, plus the name and estimated size of their associated file systems
files	boot/install servers and nodes	All the resident files resulting from a **supper update** or **install** command
serve	Control workstation and boot/install servers	All the collections that can be served from your machine
where	Control workstation	Current boot/install servers for collections
when	boot/install servers and nodes	Last update time of all resident collections
diskinfo	Control workstation and boot/install servers	Available disk space and active volume for your machine

Use:	On:	To Report This Data:
log	boot/install servers and nodes	Summary of the current or most recent **supper** session
rlog	boot/install servers and nodes	Raw output of the current or most recent **supper** session

The following examples show the output produced by some of these commands.

From the Control Workstation:

```
[tserv11][/]> supper status

 Collection          Resident  Access Point          Filesystem           Size
==============================================================================
 node.root              -     /                        -                  -
 power_system           -     /share/power/system      -                  -
 sup.admin              -     /var/sysman              -                  -
 user.admin             -     /                        -                  -
==============================================================================
```

From a boot/install server:

```
[r05n01][/]> supper status

 Collection          Resident  Access Point          Filesystem           Size
==============================================================================
 node.root             Yes    /                        -                  -
 power_system          Yes    /share/power/system      -                  -
 sup.admin             Yes    /var/sysman              -                  -
 user.admin            Yes    /                        -                  -
==============================================================================
```

From a processor node:

```
[r05n10][/]> supper status

 Collection          Resident  Access Point          Filesystem           Size
==============================================================================
 node.root             Yes    /                        -                  -
 power_system           -     /share/power/system      -                  -
 sup.admin             Yes    /var/sysman              -                  -
 user.admin            Yes    /                        -                  -
==============================================================================
```

Example of a **supper where** command:

```
[r05n01][/var/sysman/sup/node.root]> supper where
supper:  Collection node.root would be updated from server r05cw.
supper:  Collection power_system would be updated from server r05cw.
supper:  Collection sup.admin would be updated from server r05cw.
supper:  Collection user.admin would be updated from server r05cw.
```

Verifying File Collections Using scan

The **scan** file provides an inventory of all the files and directories in the file collection. By looking at the **scan** file you can quickly see which files are in a collection. You can find the **scan** file in the file collection's directory in /**var**/**sysman**/**sup**.

Although **scan** files require maintenance whenever files are added, deleted, or changed in the master collection, they save processing time on larger file systems and are recommended.

Note

You must update the **scan** file each time you modify the file collection or you risk accidentally updating the collection from an outdated master file.

Example: To check for an existing **scan** file in the **user.admin** file collection, enter:

```
ls /var/sysman/sup/user.admin
```

To create a **scan** file for the **user.admin** file collection, enter:

```
/var/sysman/supper scan user.admin
```

Following is a sample of a **scan** file for the **user.admin** collection. It shows the name of each file preceded by three sets of numbers. The first set is the file permissions, the second set indicates when the collection was installed, and the third set indicates when the collection was last updated.

```
V2
240755 740845355 740845355 .
100644 740760294 740760294 .kshrc
100644 740760294 740760294 .profile
240755 740845315 740845315 etc
100644 740943550 740943550 etc/motd
240755 740760294 740760294 var
240755 740760294 740760294 var/sysman
240755 740760295 740760295 var/sysman/etc
100754 740760295 740760295 var/sysman/etc/kernswitch
100754 740760295 740760295 var/sysman/etc/ntpdate
100754 740760295 740760295 var/sysman/etc/sup
100754 740760295 740760295 var/sysman/etc/supfilesrv
100754 740760295 740760295 var/sysman/etc/supscan
100754 740760295 740760295 var/sysman/etc/xntpd
100754 740760295 740760295 var/sysman/etc/xntpdc
```

Updating Files in a File Collection

The SP file collections can be used to contain system control files that you will need or want to change. When you update files that are part of a file collection, you do not have to make the same update on each replication of the files. Instead, you make the update to the copy on the master server for that collection and run the **supper update** command on all systems you want updated. The **supper** program does the updates for you.

The **supper update** command is included in the **crontabs** file with the default set to run hourly. You can modify the **crontabs** file to run the **supper update** more or less frequently. Refer to Chapter 13, “Maintaining the crontabs File” on page 153 for more information.

Steps in this task:

Step 1. Make the changes to the files
Step 2. Run the **supper scan** command
Step 3. Run the **supper update** command

Step 1: Make the Changes to the Files

1. **Be sure you are working with the master files.**

   In a file collection there are many replications of the same file. It is possible for copies of a file to exist on the control workstation, on several boot/install servers, and on all the processor nodes.

   In the default system, the control workstation is the master server for all file collections. If you have not changed anything in this configuration, this is where you will always make updates.

   Note

   If you changed the default configuration, be aware of the hierarchy of servers in your system and be sure you are modifying the copy of the file that is on the master server for the collection.

   Consult “File Collections Organization” on page 45 to review file collection hierarchy.

2. Change the files as appropriate.

Example: In our default configuration, the master files for all file collections are on the control workstation.

Log in to the control workstation as **root** and copy a new version of /**.profile** into the /**share**/**power**/**system**/**3.2** directory.

Step 2: Run the supper scan Command

To keep the **scan** file current, you must run the **supper scan** command on the master collection every time you make a change.

Example: Run the **supper scan** command on the control workstation. Enter:

```
/var/sysman/supper scan power_system
```

Step 3: Run the supper update Command

You need to run the **supper update** command in every location that the collection is installed and where you want the change to be effective. For example, if you change the files on a server that is the master for eight processor nodes and you want each node to have the changes, you need to run the update on all eight nodes.

You can accomplish this update in several ways:

1. You can wait for the scheduled update. The **supper update** command is included in the **crontabs** file to be run on a scheduled basis.
2. You can issue the commands immediately, logging in to each machine or remotely issuing the commands with **rsh** or **rexec**.

Example:

1. Log in to each server requiring the update. The **.profile** file is in the **node.root** collection as a secondary collection available within the **power_system** collection. This means on the boot/install servers you update the **power_system** collection. Enter:

   ```
   /var/sysman/supper update power_system
   ```

2. Log in to each of the nodes requiring the update. On the nodes you update the **node.root** collection. Enter:

   ```
   /var/sysman/supper update node.root
   ```

3. Check the **supper** messages. **supper** returns the number of files updated, removed, and the number of errors produced as shown in this message which was returned after changing **.profile** and performing the update on a processor node.

   ```
   ================================================================
   [r02n07][/]> supper update node.root
   Updating collection node.root from server mrcs-r1.
   File Changes:  1 updated, 0 removed, 0 errors.
   ================================================================
   ```

   supper also writes summary messages to to the following file:
   /var/adm/SPlogs/filec/sup.<**month**>.<**day**>.<**year**>.<**hour**>.<**min**>
   Detailed information is written to the following file:
   /var/adm/SPlogs/filec/sup.<**month**>.<**day**>.<**year**>.<**hour**>.<**min**>**.r**
   File collection log files older than 48 hours are removed at midnight by a **crontab** entry. To see all the messages in the log, you have to display or print the **logs** file. You can check the **supper** log for messages from the last install or update using a **supper** command. The following example shows a **supper update**, followed by a **supper rlog** command:

```
[r05n01][/var/adm/SPlogs/filec]> supper update node.root
Updating collection node.root from server r05cw.
File Changes:  2 updated, 0 removed, 0 errors.


[r05n01][/var/adm/SPlogs/filec]> supper rlog
SUP 7.24 (4.3 BSD) for file /tmp/.sf13689 at May  6 12:01:57
SUP Upgrade of node.root at Fri May  6 12:01:57 1994
SUP Fileserver 7.12 (4.3 BSD) 26295 on r05cw.hpssl.kgn.ibm.com
SUP Locked collection node.root for exclusive access
SUP Requesting changes since Fri May  6 12:00:07 1994
SUP Receiving file .kshrc
SUP Receiving file .profile
SUP Upgrade of node.root completed at Fri May  6 12:01:58 1994
```

Adding and Deleting Files in a File Collection

Adding and deleting files in a file collection are a bit more complicated than adding and deleting files in a standard directory. The **prefix, list,** and **refuse** master files contain criteria that define which files are in a particular collection. The master files are read and processed by the **scan, install**, and **update** commands. When you add or delete files in a file collection, you need to review the contents of these master files. For more information about the master files and how they work, see "Building a File Collection" on page 56.

The examples in this section show how to add the **/etc/environment** file to the **node.root** collection.

Steps in this task:

Step 1. Add or delete files in the file system
Step 2. Run the **supper scan** command (optional)
Step 3. Run the **supper update** command

Step 1: Add or Delete Files in the File System

1. **Be sure you are working with the master files.**

   The same considerations you make to update files also apply here. Verify that you are on the machine that is the master server for the collection you are changing. Consult "File Collections Organization" on page 45 to review file collection hierarchy.

2. Add or delete files in the file system using standard AIX file commands.

   When adding files, you need to consider whether the files are in a primary or secondary collection, what the **prefix** and **list** files in that collection contain, and the position of the new files in the file tree. Remember, the **supper** commands begin their search starting at the point defined in the **prefix** file. All files encountered that pass the criteria defined in the **list** file are considered part of the file collection.

Examples:

1. If you wanted to add a file to the **node.root** collection which is a secondary collection under the **power_system** collection, you would copy the file to the

/share/power/system/3.2 directory. The **/var/sysman/sup/node.root/prefix** file shows that the directory search for this collection begins at **/share/power/system/3.2**. The **/var/sysman/sup/node.root/list** contains an entry (**upgrade .**) indicating that all files encountered during the search will be upgraded:

```
symlinkall
omit    ./usr
upgrade .
```

To add **/etc/environment** in our default configuration, log in to the control workstation and copy **/etc/environment** to the **/share/power/system/3.2** directory:

```
cp /etc/environment /share/power/system/3.2/etc/environment
```

2. If you wanted to add a file to the **user.admin** collection, you would proceed differently. The **/var/sysman/sup/user.admin/prefix** file shows that the directory search starts at the root of the file tree, therefore, the directory search would find the **/etc/environment** file. The **/var/sysman/sup/user.admin/list** file shows that only those files specifically listed will be upgraded instead of the entire directory.

   In this case, you do not need to copy the file. Instead, add a specific entry to the **list** file to upgrade it.

Note

To add a file to a collection, the file must be readable by others.

To delete a file from a file collection, you would either remove it on the master server or delete its entry in the **list** file and proceed with the other steps in this task.

For example, to remove **/etc/environment** from the **node.root** collection, log in to the control workstation and enter:

```
rm /share/power/system/3.2/etc/environment
```

Step 2: Run the supper scan Command (optional)

The **scan** is not a required file and not all file collections have one. Check the file collection directory to see if the collection you are updating has a **scan** file.

- If you have a **scan** file in this collection and you want to maintain it, you must run the **supper scan** command on the master collection every time you make a change. When a **scan** file is present, the **update** command reads it as an inventory of the files in the collection and does not do the directory search.

  You also have the option of deleting the **scan** file and allowing **supper** to process the file collection without it.

  Note

  If you keep the **scan** file and fail to run the **supper scan** command when you add files, the **scan** file will not be current and the new files will not be included in the collection.

- If you do not have a **scan** file in this collection, the **update** command will search the directory, apply the criteria in the master files and add the new file.

Example:

1. Run the **supper scan** command on the control workstation:

   ```
   /var/sysman/supper scan power_system
   ```

   This builds a new **scan** file for the collection.

2. Check the **scan** file to verify that it contains the additional /**etc**/**environment** file.

Step 3: Run the supper update Command

Run the **supper update** command everywhere you want this change to be effective. Refer to “Step 3: Run the supper update Command” on page 53.

Example:

1. Log in to each server requiring the new file. The /**etc**/**environment** file is in the **node.root** collection as a secondary collection available within the **power_system** collection. Enter:

   ```
   /var/sysman/supper update power_system
   ```

2. Log in to each of the nodes requiring the new file. On the nodes you update the **node.root** collection. Enter:

   ```
   /var/sysman/supper update node.root
   ```

3. Check the **supper** messages to verify the updates worked.

Building a File Collection

Once you become familiar with file collections and have worked with the delivered system collections, you may want to create your own. You can build a file collection for any group of files that you want to have identically replicated on nodes and servers in your system. These file collections must all reside on the control workstation or a boot/install server. Some good candidates for file collections are application data files or local tools.

The basic procedure to build a file collections is:

- Begin by logging in as **root** to the server where the files reside.
- Create the directory and create a set of master files by copying them from an existing file collection. The master files define the rules for including and excluding files in the collection.
- Optionally, run the **supper scan** command to search the file tree, apply the rules, and build a **scan** file.
- Set up a link in the /**var**/**sysman**/**sup**/**lists** for the new file collection's **list** file.
- Add an entry to the **file.collections** file to define the file to the **supper** program.

After you build your file collections, you can install them on other servers or nodes.

Steps in this task:

Step 1. Identify the files you want to collect
Step 2. Create a file collection directory
Step 3. Create the master files
Step 4. Add a link to the **lists** file
Step 5. Update the **file.collections** file
Step 6. Update the **.resident** file
Step 7. Build the **scan** file (optional)

This section provides examples showing you how to build a primary file collection for a set of local tools in the /**usr**/**local** directory for the control workstation and boot/install servers.

Related Information

Before you build a file collection of your own, be sure you understand the directory and master files, what they contain, and how they work. Refer to "Understanding File Collections" on page 41 if you need to review these concepts.

Step 1: Identify the Files You Want to Collect

The files you want to collect as a group must all reside on the same system from which you want to serve this collection and their permissions must allow them to be readable by everyone.

Example: In this example, the files will be served from the control workstation so they must all reside there in /**usr**/**local**.

Step 2: Create a File Collection Directory

Create a directory in /**var**/**sysman**/**sup** that will be the name of your file collection.

Example: This file collection will be called *tools*.

```
cd /var/sysman/sup
mkdir tools
chown bin tools
chgrp bin tools
```

Step 3: Create a list File

Create a **list** file to describe the rules for including and excluding files in that directory. Refer to "supper list File Syntax, Keywords, and Operands" on page 64 for details.

Example: The easiest way to create these files is to copy them from an existing file collection directory, delete those you do not need or the system creates, and modify them as required.

```
cp sup.admin/* tools
cd tools
rm when last scan
```

Modify the **list**, adding these commands:

```
symlinkall
omitany *~
omit ./usr/local/admin
upgrade ./usr/local
always ./usr/local/admin/control_file
execute /usr/local/refresh_fonts (./usr/local/fonts)
```

This instructs **supper** to:

- Copy links rather than resolve them
- Omit files that start with the backup notation
- Omit administrative tools
- Upgrade all files in the local tools directory
- Override an omit instruction and always update the control file
- Execute the script to refresh the font library whenever the /**usr**/**local**/**fonts** file is changed

Step 4: Add a Link to the lists File

The **lists** file in the /**var**/**sysman**/**sup** directory contains a symbolic link to the **list** file in each file collection. When you create a new file collection you must add a link to this file.

Example: Add a link pointing to the **list** file in the *tools* collection:

```
ln -s /var/sysman/sup/tools/list  /var/sysman/sup/lists/tools
```

Step 5: Update the file.collections File

Edit /**var**/**sysman**/**file.collections** using your text editor. Add the name of your new file collection as either a primary or secondary file.

Example: The *tools* collection is a primary collection and requires an entry similar to **server.root**. Edit /**var**/**sysman**/**file.collections** and add these lines to the end:

```
# tools - boot/install collection to manage local tools files on the file servers
primary tools - / - / EDO power no
```

The nine fields in this entry have specific meaning to **supper**:

Field 1 Defines the file as a primary or secondary collection

Field 2 Specifies the name of the file collection

Field 3 For a primary collection, this specifies the name of the file system associated with the collection. This file system will be created when the file collection is installed. For a secondary collection, this specifies the name of its primary collection. The notation (-) indicates no file system will be created or that it is a secondary collection.

Field 4 Specifies the name of the directory by which the files are normally accessed.

Field 5 If the collection has a file system, this specifies the file system size. The notation (-) indicates no file system is associated with this collection or that it is a secondary collection and the file system is associated with its primary collection.

Field 6 Specifies the prefix directory at which the **scan** process starts.

Field 7 Specifies SUP options. In our example,

- E allows **list** file **execute** statements
- D deletes files that no are longer in the collection.
- O indicates that all files (even those older than the last update) should be evaluated.

Field 8 Specifies the system architecture. **power** indicates RS/6000 architecture.

Field 9 Specifies if this file collection can be installed on a different architecture.

Step 6: Update the .resident File

The **.resident** file contains a list identifying all the SP file collections. Edit the **.resident** file on the control workstation or your master server and add your new collection.

Example: Add the tools file collection to the **/var/sysman/sup/.resident** file on the control workstation:

```
sup.admin 0
user.admin 0
power_system 0
tools 0
```

Step 7: Build the scan File

The **scan** file provides you with a list of files in the collection that you can use for verification and eliminates the need for **supper** to do a directory search on each update. If your directory tree is extensive, this can save processing time on large file systems

Note

You must keep the **scan** file current. When a **scan** file is present, the **update** command reads it as an inventory of the files in the collection and does not do the directory search. If you fail to create a new **scan** file when you add, modify, or delete files in the master collection, the file will not be current and **supper** will not upgrade the collection correctly.

Example: To create a **scan** file for the tools file collection, enter:

```
/var/sysman/supper scan tools
```

Installing a File Collection

During the initialization and customization process, the required SP file collections are first built on the control workstation and then installed on your boot/install servers and processor nodes. If you create your own file collections you have to install them yourself on each server or node.

To install a collection on a boot/install server or processor node, you must be logged in as root or issue remote commands to run **supper** there.

Steps in this task:

Step 1. Update the **sup.admin** file collection
Step 2. Run the **supper install** command
Step 3. Add the **supper update** to **crontabs**

Step 1: Update the sup.admin File Collection

The **sup.admin** file collection contains the files that identify and control all the file collections. It contains the **file.collections** and **.resident** files. Whenever you make changes to these files, you need to update the **sup.admin** collection to distribute these updates.

You can wait for the scheduled update in the **crontabs** file or issue the command yourself on every boot/install server and node that needs the new collection.

Example: Log in to each boot/install server and node or issue the command remotely and enter:

```
supper update sup.admin
```

Step 2: Run the supper install Command

Run **supper install** on each boot/install server or node that needs this collection. The **install** command assumes the dedicated server to be the default.

Example: In this example, the tools file collection will only be installed on the boot/install servers and not the processor nodes.

Log in to each boot/install server or issue the command remotely and enter:

```
/var/sysman/supper install tools
```

Step 3: Add the supper update to crontabs

You need to add the **supper update** command to the **crontabs** file on each server or node that has this collection. This insures that the new collection will be updated on schedule with the other file collections.

Example: To add the command so it runs hourly, edit **crontabs** and add this line:

```
0 * * * * /var/sysman/supper update tools 1>/dev/null 2>/dev/null
```

Until the update runs, the files will not be available at each location. If you need the files on these systems before the scheduled update, you can issue the update command on each server and processor node yourself.

Example: Log in to each boot/install server or issue the command remotely, and enter:

```
/var/sysman/supper update tools
```

Refusing Files in a File Collection

The **refuse** file allows you to customize the file collection at different locations. It is possible for you to create a file collection with one group of files and have different subsets of that group installed on the boot/install servers and the processor nodes.

You do this by creating a **refuse** file in the **/var/sysman/sup** directory on the boot/install server or processor node. This master file lists all the files in every file collection that you do not want installed or updated on that system.

Example: In this example, we want one boot/install server to have the full file collection including the test files and another to have the boot/install collection minus the test files. Log in to the boot/install server that should not have the test files and create a **refuse** file with the following entries:

```
/usr/local/draw.test
/usr/local/crunch.test
```

This instructs **supper** to never include these test files when installing or updating this file collection *at this location.*

Note

There are two kinds of **refuse** files. One is in the **/var/sysman/sup** directory and contains the files that will not be installed on that server or node. These **refuse** files are optional, are user-created, and can vary at each location. The other **refuse** file is in the file collection's directory and contains a list of all the files refused when the file collection is installed or updated. This **refuse** file is system-created and may be empty if you always install the entire collection.

Modifying the File Collection Hierarchy

The delivered file collections have a default hierarchy shown in Figure 2 on page 45. You can modify the hierarchy in several ways. One way is to make a given boot/install server function as the master server for a collection. This enables you to maintain a different set of files on a group of nodes that request their **supper** updates from that boot/install server. You do this by eliminating the logical path from the control workstation to the boot/install server for one or more boot/install collections. This is called taking the boot/install server *offline* and you do it with the **supper offline** command. Figure 5 on page 62 shows a boot/install server that is offline for the **power_system** collection.

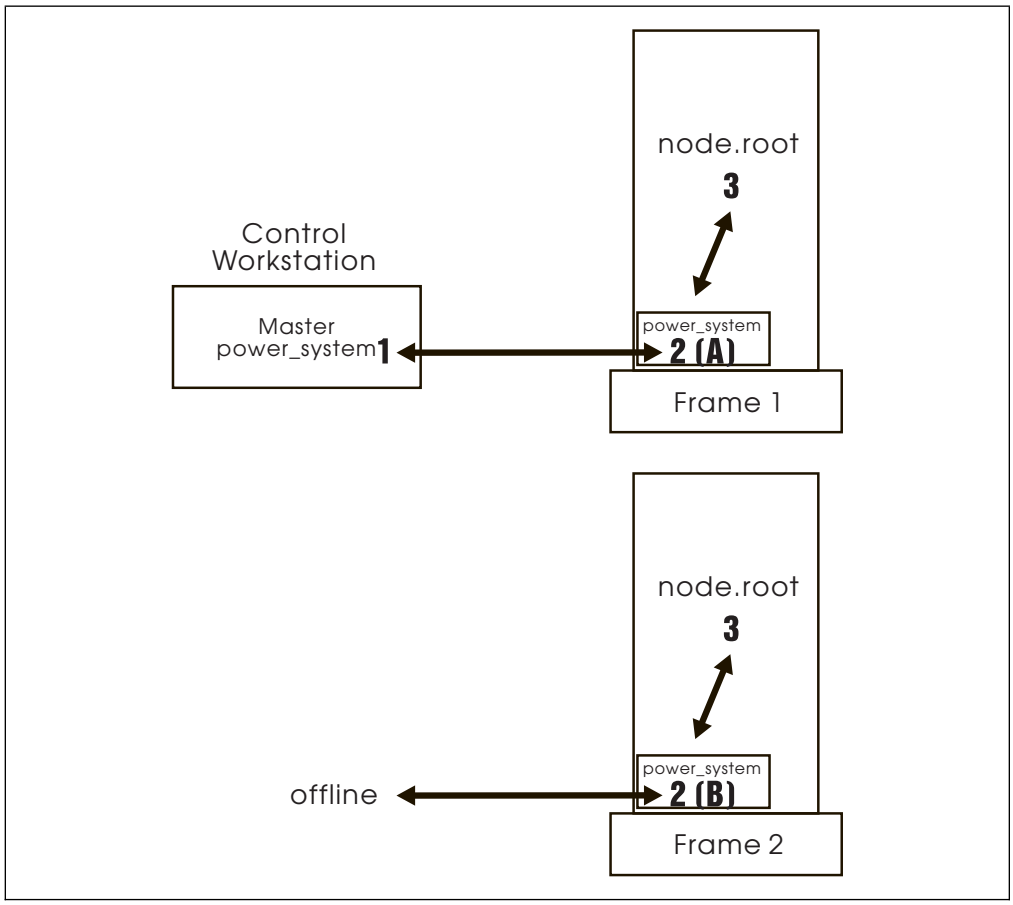


Figure 5. SP Hierarchy of an Offline Boot/install Server

In this case, Boot/install Server A and Frame 1 get any changes made on the control workstation; Boot/install Server B and Frame 2 do not. Only changes made directly to the files on Boot/install Server B affect the files on Frame 2.

Note

Whenever you change the hierarchy, you must be aware of the location of the master files for every file collection. In Figure 5, there are two master file collections for **power_system**. The master files for the **node.root** collection on Frame 2 reside on Boot/install Server B and the master files for the **node.root** collection on Frame 1 reside on the control workstation.

- If you want to change a file in the **node.root** collection for every node, you must make that change in two places, on Boot/install Server B and on the control workstation, and then run the **supper update** command on each node.
- If you want to change a file in the **node.root** collection for just one frame of nodes, you must make that change in the master file collection for that frame, either Boot/install Server B or the control workstation, and then run the **supper update** command on each node in that one frame.

The following example shows you the procedure and commands to change the default hierarchy as represented in Figure 2 on page 45. It shows you how to take the **power_system** collection offline, change a file in the **node.root** file collection, and distribute that update to the nodes on Frame 2.

Example: This example assumes that **.profile** is in the **node.root** file collection.

1. To take the boot/install server offline from the control workstation, log in to Boot/install Server B as **root** and enter:

   ```
   /var/sysman/supper offline power_system
   ```

2. On Boot/install Server B, modify the **.profile** in the /**share**/**power**/**system**/**3.2** directory.

   Remember, any changes made on the control workstation will not be updated on Boot/install Server B because it is offline. You must make the changes on Boot/install Server B.

3. Issue remote commands to update the **node.root** file collection on each of the nodes.

   ```
   /var/sysman/supper update node.root
   ```

4. Boot/install Server B no longer gets updated files from the control workstation. Unless you remove the **supper** command to update the **power_system** from the **crontabs** file on Boot/install Server B, the command will execute every hour and fail. Edit the **crontabs** file and remove:

   ```
   0 * * * * /var/sysman/supper update power_system 1>/dev/null 2>/dev/null
   ```

Removing a File Collection

> **Warning:**
> Never remove any of the file collections that come with the SP system!

The file collections that come with the SP system are required. If you remove them, the System Support Programs cannot run properly.

Removing a file collection does not delete it. It removes it from the place it was installed. To completely delete a file collection, you must remove it from every place it was installed. Be sure you do not need any files in a file collection before you remove it.

You can use the **supper scan** command to build a **scan** file for the file collection you are going to remove. The **scan** file shows you a list of all the files in a collection. You can check it to verify that you do not need any of these files before you remove the collection. See “Verifying File Collections Using scan” on page 51 for more information about the **scan** file.

Example: To build a **scan** file for verification, enter this command on the boot/install server where the *tools* collection is installed:

```
/var/sysman/supper scan tools
```

If you are sure you do not need any of the files, continue with the next command.

To remove the *tools* boot/install collection, on the same boot/install server enter:

```
/var/sysman/supper remove tools
```

This removes all the files in the collection only from that machine. If you also issue this command on the control workstation or the master server for the collection, you will no longer be able to propagate these files in your system.

---

supper list File Syntax, Keywords, and Operands

The following commands are used in the **supper list** file to define which files in a directory to include or exclude from a file collection. Refer to "Directory and Master Files" on page 46 for information about the **supper list**.

Syntax Rules

- Each command contains a keyword and a number of operands separated by spaces.
- The order of the commands does not matter and you can include blank lines between the commands.
- *filename* can be a single file, directory, or a list of files or directories separated by blanks. The names can include wild cards and metacharacters as used by **csh**, including *, ?, [....] and {...}.

Command Keywords and Operands

- **upgrade** *filename*

  The specified files or directories are included in the list of files to be upgraded unless they are also specified by an **omit** or **omitany** command or are in the **refuse** file. If a directory name is given, it recursively includes all subdirectories and files within that directory.

- **always** *filename*

  The specified files or directories are included in the list of files to be upgraded, regardless of the **omit** or **omitany** commands. They will not be included if they are in the **refuse** file.

- **omit** *filename*

  The specified files or directories will be excluded from the list of files to be upgraded. For example, by specifying **upgrade /usr/vision** and **omit /usr/vision/exp**, the generated list of files would include all subdirectories and files of **/usr/vision** except **/usr/vision/exp** and its subdirectories and files.

- **omitany** *pattern*

  The specified patterns are compared against the files specified in the **upgrade** entry. If a pattern matches, the file is omitted. The **omitany** command currently supports all wild card patterns except {...}. Also, the pattern must match the entire file name, so a leading */, or a trailing /*, may be necessary in the pattern.

- **backup** *filename*

The specified files are marked for backup. If they are upgraded, backup copies are created. Directories cannot be specified and no recursive file name construction is performed; you must specify the names of the specific files to be backed up before upgrading.

- **noaccount** *filename*

  The accounting information (permissions and time stamps) of the specified files is not preserved.

- **symlink** *filename*

  The specified files are to be treated as symbolic links. They are transferred as links and not followed. By default, symbolic links are followed. **symlinkall** transfers all symbolic links.

- **execute** *exec-command* (*filename*)

  The command specified is executed on the client process whenever any of the files listed in parentheses are upgraded. A special token, %s, may be specified in the command and is replaced by the name of the file that was upgraded. For example, if you use `execute ranlib %s (libc.a)`, whenever `libc.a` is upgraded, the client machine will execute `ranlib libc.a.`.

- **include** *listfile*

  The specified *listfile* is read at this point. This is useful when one collection subsumes other collections; the larger collection can simply specify the files for the smaller collections contained within it.

Related Information

The *IBM Parallel System Support Programs for AIX: Command and Technical Reference* presents the full **supper** command syntax with all the parameters and options. It also explains how to use **supper** in batch or interactive mode.

You can find additional information about **supper** and SUP in their manual pages.

Chapter 6. Managing User Accounts

The SP system supplies user administration commands, **spmkuser**, **spchuser**, **sprmuser**, and **splsuser**. This chapter shows you how to add, change, delete, or list user accounts using both the SP commands and the SMIT interface.

Task information included in this chapter includes:

- Understanding SP user management
- Adding user accounts
- Changing user accounts
- Listing a user account
- Removing user accounts
- Changing passwords
- Controlling user login

Understanding SP User Management

The SP system provides the ability for you to manage your user accounts by adding new users, deleting existing users, or changing the user account information from a single point of control. SP User Management ensures that users have the same account, home directory, and environment across all the nodes in the SP system. SP User Management is optional, and it is designed to fit into your existing user management system.

The objective of any user management system is to ensure that the user account files, namely **/etc/passwd**, **/etc/group**, **/etc/security/passwd**, and **/etc/security/group**, are consistent across all of your nodes. You can accomplish this by using NIS, by distributing these files using file collections, or by some other means.

If you use NIS, the SP system references the NIS maps that are located on the NIS master machine. The NIS maps provide each user with a single image of their user account information in the NIS domain. This information includes the user's login name, password, group information, default shell, and the name of the host where the home directory resides. If you use SP automounter support, the system also maintains automounter maps of mount points for the home directories. As the system administrator, you need to create and maintain the NIS environment on the NIS master machine in the **/var/yp** directory. The SP User Management commands do not interface with the NIS environment.

If you are not using NIS, the SP User Management commands use the **user.admin** file collection to propagate the user administration files to the nodes of the system. You may wish to use this file collection to keep other user management related files consistent across the nodes. For more information refer to the “user.admin Collection” on page 44.

In NIS, users can change their own passwords, as needed, while logged in any SP node. In non-NIS environments users must log into the host where the password file resides. The full pathname for the password file is set with the **passwd_file** parameter of the **spsitenv** command, or through SMIT panels during installation. Administrators can change passwords for any user account from the NIS master.

© Copyright IBM Corp. 1995, 1997

After you complete the installation and customization procedures, you establish an initial set of users from the existing user management files in your network. The SP system provides the ability for you to manage your user accounts by adding new users, deleting existing users, or changing the user account information.

SP User Management Options

Using the SP user management commands, you can add and delete users, change account information, and set defaults for your users' home directories. Specify the user management options you wish to use in your site environment during the installation process, or change them later, either through SMIT panels or by using the **spsitenv** command. The *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* contains detailed instructions for entering site environment information.

From SMIT

ENTER **smit enter_data**

▲ The Enter Database Information menu appears.

SELECT **Site Environment Information**

▲ The Site Environment Information menu appears.

The following options, displayed in the SMIT Panel for Site Environment Information, pertain to user management:

User Administration Interface
Whether or not you want to use SP User Management to add, delete, and change user account information.

Password File
The path of your password file. The user commands add, modify, and delete entries in this file. If you modify the default path of **/etc/passwd** and you are using NIS, ensure that you also modify your NIS makefile to build the password map from the new password file.

Password File Server Hostname
The hostname of the machine where your master password file resides. The initial value is the control workstation. The master password file location cannot be one of the nodes in the SP system.

Home Directory Server Hostname
The name or IP address of the machine where the user home directories reside.

You can specify a default host for users' home directories. If you use the SP automounter support the user management commands will use this host when building automounter maps to mount and link the home directories. If you do not specify a default, the initial value of the control workstation is assumed.

You can override the value in Hostname Home Directory Server when adding or modifying a user account with the **spmkuser** and **spchuser** commands.

If the users' home directories are served by a machine within the SP system and you follow the authentication setup described in Chapter 14, “Security Features of the SP System” on page 157, then you need not

do anything more to have the SP user management commands run correctly.

However, if the home directory server is outside the SP system, there are two cases:

1. The machine that serves the home directory is in the same authentication realm as the SP system. In this case, the administrator's principal name must be added to the **.klogin** file of that machine.
2. The machine that serves the home directory is outside the realm and nonauthenticated. In this case, add a line to the **.rhosts** of the machine with the following data:
   - The hostname where the commands run
   - The user name **root**

   In this case, when the SP user management commands are run, informational messages are displayed to show that authenticated commands could not be run. However, the user management task will be performed.

Home Directory Path

This is where you specify the default path of users' home directories. The default is /**home**/*control_workstation_name*/.

Enter another value to set another path as the default for your site. You can override the default path with the **home** attribute, **spmkuser**, and **spchuser** commands.

Any host that is serving user directories must have its /**etc**/**exports** file modified to grant access to all hosts that need to mount these directories. Export the file systems using **exportfs -a** after modifying the /**etc**/**exports** file.

SP User Management Commands

The SP User Management commands reside in the /**usr**/**lpp**/**ssp**/**bin directory** on the control workstation. If you wish to use these commands in an NIS environment and your control workstation is not your NIS master, you must mount these commands on the NIS master where you will be executing them. (Your NIS master must be an RS/6000.)

You can add, change, delete, or list user accounts using either the SP commands or, optionally, the SMIT interface. For example, you can add new users to the SP system in any of the following ways:

- Enter the **spmkuser** command with the appropriate attributes and values.
- Enter **smit spmkuser**.
- Enter **smit**, select *Security & Users*, and then select *RS/6000 SP Users*.
- Enter **smit**, select *RS/6000 SP System Administration*, and then select *RS/6000 SP Users*.

The examples in this section show the first two methods. You can find the complete syntax for the commands in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Adding User Accounts

To add new users to the SP system with the SMIT interface:

TYPE **smit spmkuser**

▲ The Add a User window appears.

Only the *User name* is required. You can accept the defaults or leave the other fields blank. The defaults for primary group, secondary groups, and the initial program are in **/usr/lpp/ssp/bin/spmkuser.default**. The default for the home directory is stored in the System Data Repository. The following table shows how the fields in this window relate to the parameters in the **spmkuser** command.

SMIT Field	Command Parameter	Description
User name	Name	User login name
User ID	id	ID of the user
LOGIN user?	login	Indicates whether the user can login to the system with the **login** command. This option does not change the **/etc/security/user** file but instead alters the user's password field in **/etc/security/passwd**.
Primary GROUP	pgrp	Principle group of the user
Secondary GROUPS	groups	The groups to which the user belongs
HOME Directory	home	Host name of the file server where the home directory resides and the full path name of the directory separated by a colon
Initial PROGRAM	shell	Shell program run when the user logs in
User Information	gecos	General information about the user

When you add a new user, the system generates a random password for the user and stores it in **/usr/lpp/ssp/config/admin/newpass.log**. The **root** user has read and write permission to this file. It is the administrator's responsibility to communicate this password to the new user and periodically delete the contents of this file.

Example

The following example shows how to enter the information for a user named **joe** using SMIT:

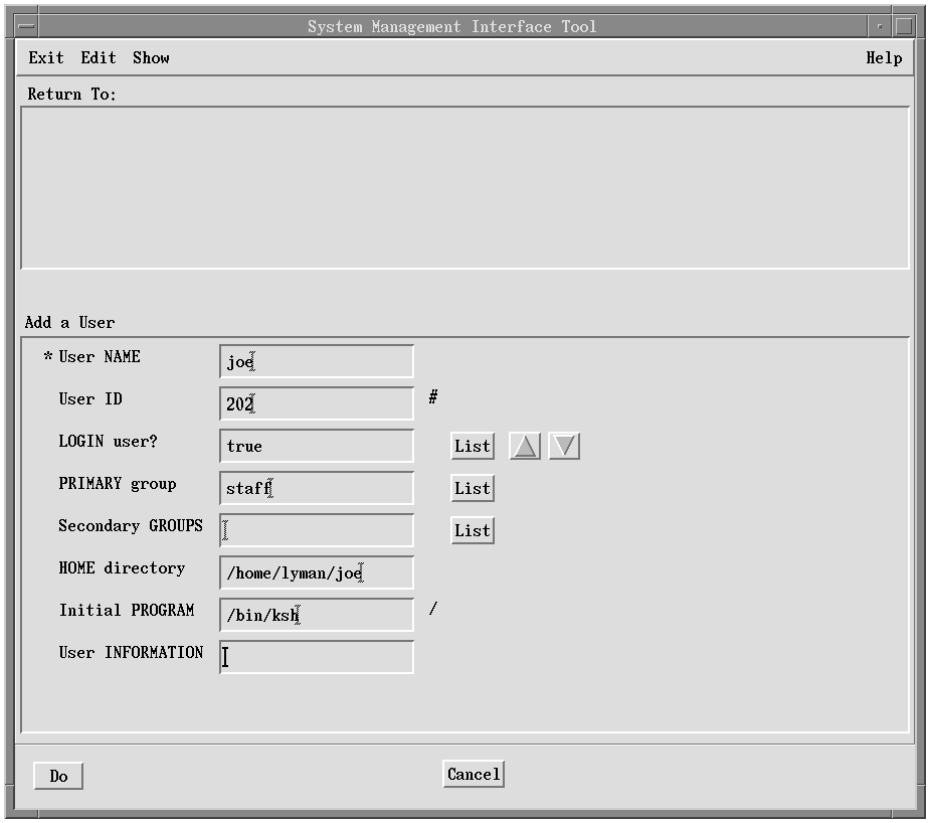


Figure 6. Example of Adding an SP User

TYPE The appropriate information in each field

PRESS **Do** to store the data.

To store the same information using the **spmkuser** command, enter:

```
spmkuser id=202 pgrp=staff home=lyman:/home/lyman shell=/bin/ksh joe
```

Changing User Accounts

To change user information with the SMIT interface:

TYPE **smit spchuser**

TYPE The name of the user whose information you want to change

▲ The Change/Show Characteristics window appears

To change information, type the new values in the appropriate fields. The only field you cannot change is *User name*. You can scroll to see the information in the home directory field.

Example

To change information using SMIT:

TYPE The new values

PRESS **Do** to store the data.

To change the home directory for **joe** using the **spchuser** command, enter:

```
spchuser home=lyman:/home/lyman/joe joe
```

Removing User Accounts

To remove an existing user account from the SP system with the SMIT interface:

TYPE **smit sprmuser**

TYPE The name of the user whose information you want to remove

▲ The Remove a User window appears.

You can choose to remove the user's home directory and the authentication data when /**etc**/**security**/**passwd** is utilized.

Example

To remove the **joe** user account with the **sprmuser** command but leave the authentication data and home directory intact, enter:

```
sprmuser joe
```

Listing a User Account

You can display the information for a user with the SMIT interface using **smit spchuser**.

TYPE **smit spchuser**

TYPE The name of the user whose information you want to view

▲ The SP Change/Show Characteristics window appears.

Be sure to cancel after viewing the information.

You can also list user information using the **splsuser** command. For example to list information for a user named **ken** in colon-separated records, enter:

```
splsuser -c ken
```

Changing Passwords

Your method of account management determines which steps are necessary to change passwords. The procedure for NIS is different than for a system with file collections. If neither NIS nor file collections is used, you will probably use **rdist**.

Changing Passwords on an NIS System

NIS users can modify their passwords using either the **yppasswd** or **passwd** command. They must run these commands within the NIS domain that serves their user account.

Note

If you maintain an NIS password file outside of an NIS environment, you should continue your current process for changing passwords and disable the **yppasswdd** daemon on the NIS master:

1. Enter **smit chmaster**.
2. Set *START the yppasswdd daemon?* to **no**.
3. Set *CHANGE server now, at system restart or both* to **both**.

As a result, whenever users try to change passwords, they receive a message stating that the master is not running.

Changing Passwords for a File Collection System

If you use file collections rather than NIS, you must change the password in the master password file. Although this file may not be maintained on the control workstation, any changes to passwords must be made known to the control workstation.

1. Change password on host where master password file resides.
2. If the master password server is not the control workstation, copy the files **/etc/passwd** and **/etc/security/passwd** from the password server machine to the control workstation.
3. New password will be propagated throughout SP system within one hour.

Restricting the Use of the Control Workstation for General Users

If you use file collections on your system, users must be able to log in to the machine designated as the password file server to change their passwords. If you use the control workstation for the password server, you must allow users to log in to it to change their passwords. However, you may wish to have only administrative tasks performed on the control workstation, and therefore, do not want general users to be able to do anything other than change their passwords.

A script is provided that allows full login on the control workstation to the users you designate and restricts full login to all others. Users not designated can only log in to change their passwords and then are logged out. The script can be incorporated into the profile that is run during every login to the control workstation.

The script is located in **/usr/lpp/ssp/config/admin/cw_restrict_login**. To use the script, take the following steps:

1. Edit **/usr/lpp/ssp/config/admin/cw_allowed**

   The **cw_allowed** file defines to the **cw_restrict_login** script which user names are allowed to fully log in to the control workstation. **cw_allowed** is shipped containing three administrators' user names to show how the file should be formatted. These sample names should be removed from the file.

Start each user name in the left most column and place one user name per line. Place no comments in the file. You do not need to list the root user in the file; the script is written to allow the root user to fully log in to the control workstation.

2. Integrate the restricted login script into the control workstation's login process by adding the following lines to the beginning (or at the most appropriate place) of the control workstation's **/etc/profile** file:

```
# Allow general users to login to control workstation to only change
# their password then log them out.
/usr/lpp/ssp/config/admin/cw_restrict_login
```

   If you are using the AIX Common Destktop Environment (CDE) on the control workstation, you will also need to make a link from the restricted login script for CDE to the appropriate CDE directory:

```
ln -s /usr/lpp/ssp/config/admin/cde_cw_restrict_login \
/etc/dt/config/Xsession.d/cde_cw_restrict_login
```

3. Verify that you have set up correctly:
   a. Log in to the control workstation as root and make sure you can fully log in.
   b. Log in as a user that should be allowed (listed in **cw_allowed**) to log in to the control workstation. Make sure this user can fully log in.
   c. Log in as a user that is not allowed (not listed in **cw_allowed**) to fully log in to the control workstation. Make sure this user can change his password and then is logged out.

At any time you wish to remove this restrictive login, you can do so simply by commenting out (adding a # at the beginning of the script execution line) or by deleting the lines you added from the control workstation's **/etc/profile** file.

If you are using CDE and wish to remove its restrictive login, simply remove the **cde_cw_restrict_login** link you created in the **/etc/dt/config/Xsession.d** directory.

Changing Passwords on Systems without NIS or File Collection

If neither NIS nor file collection is used, you must distribute these updated password files across the SP system after you change the password.

1. **/etc/passwd**
2. **/etc/security/passwd**

Changing Passwords when SP User Administration Interface is Configured

If you have configured the SP USER ADMINISTRATION INTERFACE to **true** and you are not using NIS, the following base AIX commands are linked to SP commands on the nodes:

- **/bin/chfn** → **/usr/lpp/ssp/config/sp_chfn**
- **/bin/chsh** → **/usr/lpp/ssp/config/sp_chsh**
- **/bin/passwd** → **/usr/lpp/ssp/config/sp_passwd**

The original AIX files have been saved in their original location with extension **.aix** added. For example, the original AIX **passwd** command is in **/bin/passwd.aix**.

The linked **sp_** commands inform users what machine to log in to when, or if, they need to change password, gecos, or shell information.

Important

Under certain circumstances, the user is prompted to change passwords during the login process. If this occurs, the user should log in to the machine serving the password file and make the change from there. If this is not done, the password change will be only local to the node; in addition, if File Collections is being used, the password change will be only temporary. To avoid this problem, communicate to users what machine to log in to when a password change is required during the login process. We recommend placing this information in the message of the day (MOTD).

Controlling User Login

Login Control is used to dynamically prevent interactive login of users on a node basis. Preventing interactive login of users on nodes running parallel jobs is desirable for performance purposes. Login Control is intended for use with the Resource Manager. Once you have restricted access for users on a node using Login Control, the Resource Manager uses Login Control to unrestrict interactive access to those users running a parallel job scheduled through the Resource Manager. Once the job has completed, the Resource Manager uses Login Control to return the user to the original state.

If you have correctly configured the Resource Manager to use Login Control, it can automatically control user access to your SP nodes on a per job basis.

Login Control can be used to control interactive user access on nodes that have not been assigned to the Resource Manager for running jobs. You may want to use Login Control to temporarily restrict all but a few users on a node for a specific purpose.

You can issue commands to temporarily restrict and unrestrict users' interactive access to nodes running parallel jobs. However, this function should be used with caution because it can interfere with the coordination between the Resource Manager and Login Control processes.

Understanding Login Control

User login completes successfully under AIX after checking attributes in the **/etc/security/user** file. The attributes are *login* and *rlogin*. If these values are set to `true`, the user is allowed to log in. Conversely, if these values are set to `false`, the user cannot log in. Login Control updates this file dynamically for a user or several users when a request is received.

The **spacs_cntrl** command takes four state keywords that manage user access to a parallel node:

1. **block**
2. **unblock**
3. **allow**
4. **deny**

Block and Unblock

These keywords are used by a System Administrator to dynamically restrict and unrestrict access to a node for a user or a group of users. These keywords are intended to set the user to a known state (either restricted or unrestricted) in the **/etc/security/user** file and in the **spacs_data** file. The block keyword can be used to initially restrict user access to a node before Resource Manager submits user jobs. These keywords should not be used on a node while Resource Manager is submitting jobs to the node; otherwise, the user state may be corrupted.

Allow and Deny

These keywords are used by a job submission system (such as the Resource Manager) on a transactional basis to update the user state. Multiple usage of **spacs_cntrl** with one of these keywords for a user results in the request state being updated in **spacs_data**.

A user can be in one of four states when these keywords are used:

1. Allowed interactive login
2. Denied interactive login
3. Allowed login after a deny request
4. Denied login after an allow request

State 3 can occur when other jobs by this user are still running. A deny request removes one allow request: in this case, the first allow from the first job.

State 4 can occur when more than one deny request is issued for a user. A deny request removes one allow request.

A special file, **spacs_data**, is created when a user access change request is made more than once by a job submissions system using the keywords **allow** or **deny**. It can be said that this file holds the state of requests for a user. After the initial allow or deny request has been processed in **/etc/security/user**, the next allow or deny request sets a request count for the user. The same number of subsequent opposite requests will check the count, decrement it and remove the user, if present, from **spacs_data**. The use of the block and unblock keywords by the System Administrator updates **spacs_data**, clearing all multiple job submission system allow/deny requests for a user before changing the state in **/etc/security/user**. These keywords are intended to set the user to a known state (either restricted or unrestricted) in the **/etc/security/user** file and in the **spacs_data** file. The **spacs_data** file is owned by root and is located in the **/var/spacs** directory.

Login Control recognizes requests made with the **spacs_cntrl** command. (See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for syntax.) This command is run on each parallel node and can be used with **dsh** for this purpose.

Login Control uses a lock on the **spacs_cntrl.lock** file to prevent more than one instance of Login Control from updating the files. This file is created the first time the **spacs_cntrl** command is executed. This file also resides in the **/var/spacs** directory.

Types of Access Disallowed

When **spacs_cntrl** block or deny is used, all of the following types of interactive access are disallowed:

- **login**
- **rlogin**
- AIX **rsh**
- AIX **rcp**
- AIX **rexec**
- SP **rsh**
- SP **rcp**

Disallowing ftp on a Node

As system administrator, you can disallow users from using **ftp** on a node by placing users in the **/etc/ftpusers** file. This file can be kept in a file collection and distributed to the appropriate nodes. The **ftp** daemon can also be disabled to prevent all ftp access.

You can also use the **ruser** command to manage **ftp** usage by a user on a node:

```
ruser -a -f username
```

Using this command with **dsh** to the appropriate nodes allows manipulation of the **/etc/ftpusers** file on a per user basis.

Using the Login Control Sample File

A sample Perl script called **block_usr_sample** is located in the **/usr/lpp/ssp/samples** directory. This script creates a file of user IDs, **/tmp/usr.input**, that can be used with **spacs_cntrl -f** to restrict user access. It serves as a model for formatting other files for use with the **spacs_cntrl** command:

```
userid1
userid2
userid3
```

Once you are familiar with Login Control, **spacs_cntrl** can be run within this script by removing the comments surrounding the command. A copy of this file is located in Appendix C, “Sample Files” on page 503.

Before using this script you must edit the **threshold UID ($uidstart)** in the script to be large enough so that system IDs and SP system IDs are not restricted. These IDs include **bin**, **admin**, **lpd**, **supman**, and **prtid**. The default threshold UID is set at 125. The **block_usr_sample** script can be added to the **script.cust** file or to the node's **inittab** file.

Adding the script to run from the **script.cust** file will create the **/tmp/user.input** file and run **spacs_cntrl block** when you install the node to initially restrict interactive access for all users. Users would remain in this state across a reboot unless specifically unrestricted by the System Administrator or by the job submission system. New users that are added to the system would not be included until the **block_usr_sample** script was run again at some later time.

Adding the **block_usr_sample** script to the **inittab** file will run this script at node boot time. This will include any new users on the node but may also increase the time to boot the node. This script must be added **AFTER** the **rc.sp** script in the **inittab** file. **rc.sp** removes the login control lock file, **spacs_cntrl.lock**, if present,

when booting a node to ensure login control will run if a node is rebooted after a crash.

Using Login Control

There are many ways you can use Login Control on your SP system. You can restrict all or a subset of users on a parallel node to prevent the users from gaining interactive access. When the Resource Manager is not configured to use login control, the restriction remains in effect until you specifically use the unblock keyword to unrestrict access for the restricted users. Any jobs scheduled to the nodes controlled by the Resource Manager will return an error for those users whose access is restricted.

This enables you to set aside a parallel node for the use of select users for interactive and parallel jobs. You may list the users on the command line or use a file containing the user names in a column. Note that this file must reside on the parallel node in order to be accessible to the login control command.

Here are a couple of suggestions to get you started. Whichever you decide to use, ensure that the Resource Manager is configured correctly. You can use the **locate_jm** command to identify the primary and backup Resource Manager servers. The **jm_status** command gets information about resource pools and nodes allocated for Resource Manager jobs. See “Using the Resource Manager” on page 87.

Restricting a User on a Single Node

To restrict user Betty on a particular node, on that node issue:

```
spacs_cntrl block betty
```

Restricting Users on Multiple Nodes

1. Run the **block_usr_sample** file after adjusting threshold UID.
2. Send file to the node(s). Note this example would require **rsh** priviledges on the nodes.

   ```
   dsh -a rcp  root@mynode:/tmp/usr.input /tmp/usr.input
   ```

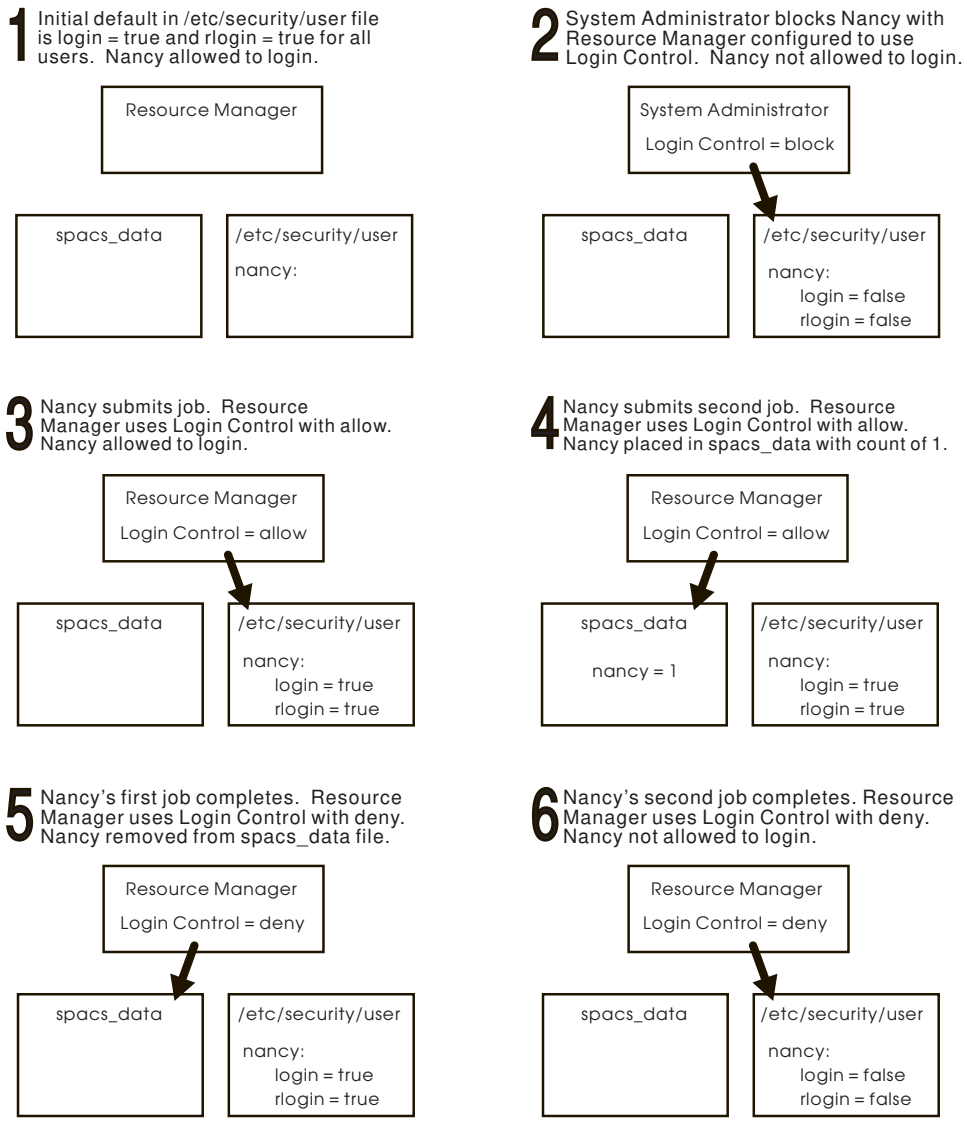
3. Issue login command to restrict interactive access for users to all the nodes.

   ```
   dsh -a spacs_cntrl -f /tmp/usr.input block
   ```

Login Control and the Resource Manager

The Resource Manager is used to submit parallel jobs to the SP System. When the Resource Manager has control of a node for scheduling parallel jobs, it must be able to unrestrict a user's access on the system in order to run the user's parallel job. Note that during the running of a parallel job, that user can also use the node interactively. Once the user is unrestricted, the AIX **login**, **rlogin**, and **telnet** executables cannot distinguish between the Resource Manager scheduling the job using the user's ID and the user logging into a node.

The following example shows how the Resource Manager and Login Control interact. For this example, the System Administrator has restricted access for users on this node before allowing the Resource Manager to run.



Using Login Control with the Resource Manager

You may configure the Resource Manager to use Login Control in order to dedicate a node to running parallel jobs. Once it has control of the node for scheduling jobs, the Resource Manager will use Login Control to allow a user access in order to schedule that user's parallel job. Once the job is finished, the user will be placed back in the original state.

Since the System Administrator's keywords, block and unblock, clear all multiple allow/deny requests in **spacs_data**, you should not use it on the nodes under the Resource Manager's control. You can initially block users on a node and then start the Resource Manager to handle the access for parallel jobs. Conversely, the Resource Manager should be quiesced before running **spacs_cntrl** again on those nodes it controls.

The only exception is blocking a new user on a parallel node. This should be done before allowing the user to submit parallel jobs to avoid conflict.

You can configure the Resource Manager to use Login Control whether or not you choose to initially block users on a parallel node.

Maintaining Login Control

Minimal maintenance is required after the Resource Manager is configured to use Login Control on nodes running parallel jobs. However, the following information is important for Login Control to function properly.

Since the **/etc/security/user** file is used by Login Control for state information, and by the AIX system for restricting and unrestricting login for users, this file is machine dependent and should not be overwritten. The **/etc/security/user** file should not be distributed through file collections or any other mechanism throughout the SP system.

The **spacs_cntrl** command may generate a large log file, depending on which flags you specify. The default for trimming the log file is 400 lines and is done every night through a **cron** job. If you wish to save the log for some reason, you can rename the file and a new log file will be created.

Usage Notes

You should not use the Login Control command, **spacs_cntrl**, while the Resource Manager is using Login Control. This is because the System Administrator keywords, **block** and **unblock**, return the user to a known state and remove status kept in the **spacs_data** file. The Resource Manager has no way to tell this has been done and may leave the user in an unpredictable state for running jobs.

You must issue the **spacs_cntrl** command to block new users added to a node if you have previously blocked all users on that node. Users denied login will see the standard message, “You have entered an invalid login name or password.”

The Login Control utility will not prevent **LoadLeveler** from running jobs submitted by blocked users because **LoadLeveler** logs in as **root** and then switches to the user. Root is never blocked on a node.

Each time the log is opened, a datestamp is written. This may or may not be followed by messages depending on the options chosen on the command line. The following **spacs.log** example shows the output when using the **-d** and **-l** flags on the **spacs_cntrl** command line.

Sample Log

```
"Tue Apr 26 10:11:47 EDT 1994"
spacs_cntrl:  Command Invocation: spacs_cntrl -d -l deny nancy.
spacs_cntrl:  CPFILE called for /etc/security/user and /etc/security/user.17008.
spacs_cntrl:  ACS_FILE/lock called for /var/spacs/spacs_data.lck.
spacs_cntrl:  List of users in deny list: .
spacs_cntrl:  List of users in allow list: ,mroz.
spacs_cntrl:  Deny processing.
spacs_cntrl:  List of users in send list: nancy.
spacs_cntrl:  MUSRFILE called.
spacs_user:  Called with parameters:  /usr/lpp/ssp/bin/spacs_user -l -d deny nancy.
spacs_user:  Attribute for user nancy changed to requested value.
spacs_cntrl:  List of users in return list: nancy:0.
spacs_cntrl:  CPFILE called for /var/spacs/spacs_data and /var/spacs/spacs_data.17008.
spacs_cntrl:  ACS_FILE/unlock called for /var/spacs/spacs_data.lck.
"Tue Apr 26 10:11:52 EDT 1994"
spacs_cntrl:  Command Invocation: spacs_cntrl -d -l deny pepper.
spacs_cntrl:  CPFILE called for /etc/security/user and /etc/security/user.20102.
```

```
spacs_cntrl:  ACS_FILE/lock called for /var/spacs/spacs_data.lck.
spacs_cntrl:  List of users in deny list: .
spacs_cntrl:  List of users in allow list: ,mroz.
spacs_cntrl:  Deny processing.
spacs_cntrl:  List of users in send list: pepper.
spacs_cntrl:  MUSRFILE called.
spacs_user:  Called with parameters:  /usr/lpp/ssp/bin/spacs_user -l -d deny pepper.
spacs_user:  Attribute for user pepper changed to requested value.
spacs_cntrl:  List of users in return list: pepper:0.
spacs_cntrl:  CPFILE called for /var/spacs/spacs_data and /var/spacs/spacs_data.20102.
spacs_cntrl:  ACS_FILE/unlock called for /var/spacs/spacs_data.lck.
```

Chapter 7. Managing SP Resources

This chapter describes the Resource Manager, its administration, and use. Task information is presented for:

- Understanding the Resource Manager
- Installing the Resource Manager
- Defining the Resource Manager configuration
- Starting the Resource Manager
- Reconfiguring the Resource Manager
- Stopping the Resource Manager

For information on Accounting related to Resource Manager, refer to “Accounting for Exclusive Use of Node Jobs” on page 142.

For many of the tasks involved in managing the Resource Manager you have the option of using either a command or the SMIT graphical user interface. Note that when you are using SMIT there is a list option (PF4) available for dialog screens and that help screens are available for all Resource Manager menus and fields.

Understanding the Resource Manager

The Resource Manager is a server that runs on one of the SP nodes. It prevents parallel jobs from interfering with each other and reports job related node information.

When applications such as the Parallel Operating Environment (POE) and LoadLeveler make a request, the Resource Manager identifies a group of nodes that meet the requested criteria. In doing so it:

- Allows POE applications to request dedicated or shared use of nodes or adapters on nodes. LoadLeveler always requests shared use of nodes unless the High Performance Switch adapter is used in user space mode, in which case it requests shared use of the node but dedicated use of the adapter. Requesting dedicated use of an adapter stops other jobs from using the adapter specified by the job (one of Ethernet, FDDI, FCS, Token Ring, the High Performance Switch used in user space mode and the High Performance Switch used in IP mode). It does not stop the node from being used by another job using a different adapter. This means that if a job requests and is allocated shared use of nodes A and B but dedicated use of the High Performance Switch-Adapter2's user space part on those nodes, then no other job will be allocated the High Performance Switch-Adapter2's user space part on A and B for as long as the first job is running. However, another job needing shared use of nodes A and B but dedicated or shared use of the High Performance Switch-Adapter's IP part on those nodes, will be allocated those nodes since the request does not conflict with the first job's resource criteria.
- Can select either the **en0** or **en1** adapter for Ethernet job application requests during initial configuration. The adapter type is considered to be static data and the Resource Manager must be stopped and restarted to change this information.
- Allows applications to be run interactively or submitted via a batch scheduler (LoadLeveler only).

© Copyright IBM Corp. 1995, 1997

- Controls accounting for dedicated node usage.
- Can create CSS authentication tables on identified nodes so that the authorized CSS user space job can run without interference from other jobs.
- Allows you to lock out users from nodes unless they request access via the Resource Manager. Refer to “Controlling User Login” on page 75 for more information.

A single point of control is provided for starting, stopping, and configuring the Resource Manager.

When the Resource Manager server is started, it automatically tries to start a backup on another node. If the primary Resource Manager server fails, this backup becomes the primary server and starts a backup on another node. Jobs will continue to run unaffected.

When the Resource Manager selects nodes for use by an application, it might have to execute certain functions (such as accounting, login control, or CSS authentication tables creation) on the selected nodes. To do so, it contacts the **inetd** daemons on those nodes which in turn spawn a Resource Manager daemon to perform the task.

A flat UID space is required for all clients connecting to the Resource Manager.

There are two Resource Manager log files:

- **/var/adm/SPlogs/jm/jmd_out** contains general information about the operation of Resource Manager daemons
- **/var/adm/SPlogs/jm/jmd_err** contains errors encountered by Resource Manager daemons

System Partitioning

Resource Manager function, as described above, is supported *within individual system partitions*. The Resource Manager operates within the scope of each system partition environment. Each partition may have its own Resource Manager server, but no functionality crosses system partition boundaries. The Resource Manager gets the correct node and adapter information for each system partition from the SDR automatically. Each system partition has its own configuraton database named on the control workstation. If the system is not partitioned, the *syspar_name* is the short host name of the control workstation and can be found by issuing **spget_syspar -n**.

Installing the Resource Manager

If you are not using the SP for parallel processing, you do not need to install the Resource Manager.

To install the Resource Manager, use **installp** to install **ssp.jm** of the IBM Parallel System Support Programs for AIX. **ssp.jm** may have been installed on the control workstation during system installation as described in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*. (To check if **ssp.jm** has already been installed on the control workstation, enter **lslpp -m | grep ssp.jm**) The **ssp.jm** fileset depends on the installation of **ssp.basic** and

ssp.clients. The **libjm_client.a** library, **jm_client.shr.o** loadable object, and the **jm_status** command are now installed via the **ssp.clients** fileset. The **ssp.clients** fileset should be installed on client machines when access to the Resource Manager shared library and the **jm_status** command is desired on machines off of the SP system.

If you choose to selectively install the Resource Manager, you must install it on the control workstation and every node of your SP system. Refer to the chapter on software maintenance in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

When you install the Resource Manager, the /**etc**/**services**, /**etc**/**inetd.conf**, and /**etc**/**inittab** files are updated. Backup copies of the original files are stored in /**etc**/**services.bak.jm**, /**etc**/**inetd.conf.bak.jm**, and /**etc**/**inittab.bak.jm**, respectively.

- Two lines identifying Resource Manager ports are added to /**etc**/**services**:

  ```
  jmd             4086/tcp
  ijmd            4173/tcp
  ```

  These port numbers may be changed to any other unused port number as long as it is consistent on all the nodes of your SP system. Also, any other machines needing services from the Resource Manager on your SP must update their /**etc**/**services** file to contain this jmd port number.

- The following entry is added to /**etc**/**inetd.conf**:

  ```
  ijmd    stream  tcp      nowait  root    /usr/bin/jmd jmd
  ```

- The following entry is added to /**etc**/**inittab**:

  ```
  jm_setupcss:2:boot:/usr/lpp/ssp/bin/jm_setupcss
  ```

You can verify that all the Resource Manager executables and required links have been properly installed on the control workstation and SP nodes, either from SMIT panels or by using the the verification script, **jm_install_verify**.

Using SMIT:

TYPE	**smit SP_verify**
	▲ The Installation/Configuration Menu appears.
SELECT	Resource Manager Installation
PRESS	Enter

Using the Command Line: This command must be executed from the control workstation as root and resides in /**usr**/**lpp**/**ssp**/**bin**.

Enter the command as follows:

```
jm_install_verify
```

Optional flags

-l *filename*	Specifies log file name (Default: /var/adm/SPlogs/jm_install_verify.log).
-q	Suppress messages.
-h	Print Usage information

See the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* for more information and examples.

Defining the Resource Manager Configuration

Begin by classifying the nodes on the SP system as either:

- Nodes not used for parallel jobs (**serial nodes**)

or

- Nodes used for parallel jobs (**parallel nodes**).

Serial Nodes

The Resource Manager does not allocate serial nodes; however, you may wish to record:

- Where the batch serial jobs run (**batch serial nodes**)
- Where the users are allowed to log into (**interactive serial nodes**)
- Where either of the above will be allowed (**general serial nodes**)

These serial nodes form the serial pool which is identified by a *pool_id* of -1 and an associated description string.

A node defined to the Resource Manager as a member of the serial pool cannot also be defined as a member of any parallel pools described below.

Parallel Nodes

You should group parallel nodes into parallel pools, since parallel applications indicate their resource requirements by specifying the pools from which nodes are to be allocated.

Note: 604 High Nodes cannot be configured as part of a parallel pool. Therefore, the Resource Manager will not allocate these nodes for parallel jobs.

There can be multiple parallel pools. Any number of nodes may form a pool, and any meaningful group of characteristics can be the basis for defining a pool. You might form one pool from all wide nodes and another from thin.

Each node can belong to only one pool. Each parallel pool is identified by a unique pool_id, greater than -1, and an associated description string. Each node in the parallel pool should now be identified as being able to run:

- Parallel jobs submitted through LoadLeveler (member of the **batch subpool**)
- Parallel jobs started interactively (member of the **interactive subpool**)
- Either of the above (member of the **general subpool**)

If you identify all nodes in a pool as general nodes (belonging to the general subpool), you may experience delayed batch processing due to heavy interactive use. For this reason it may be preferable to set aside some of the nodes in the pool for batch use only. Similarly, interactive users may be denied access to nodes due to heavy batch load, so you may choose to set aside some of the nodes in the pool for interactive use only.

When the Resource Manager allocates nodes for interactive use, it searches through the interactive nodes (members of the interactive subpool) of the specified pool first, and then the general nodes, to satisfy the user's request. Similarly, when the Resource Manager allocates nodes for batch use, it searches through the batch

nodes (members of the batch subpool) of the specified pool first, followed by the general nodes.

Refer to “Step 3: Configure the Resource Manager” on page 89 to specify the pool and subpools for your nodes.

Using the Resource Manager

You can use the Resource Manager for both interactive and batch support.

Interactive Support

Consider the user who is logged onto one of the serial interactive nodes, and is using POE to run a parallel job interactively on the SP system. The POE partition manager conveys the job's node requirements to the Resource Manager. The Resource Manager, in turn, allocates the appropriate parallel nodes from the parallel pools. Refer to *IBM Parallel Environment for AIX: Operation and Use*, SC28-1979 and SC28-1980, for more information about using the POE.

Batch Support

If users want to submit parallel jobs to LoadLeveler, the System Administrator must perform additional LoadLeveler configuration for interaction with the Resource Manager as described in *Using and Administering LoadLeveler*, SC23-3989.

Example Configuration

Figure 7 on page 88 shows a possible configuration of a SP system with 16 nodes. This system has three parallel pools. The pool with id=0 has six nodes: nodes 1 and 2 assigned to the general subpool, nodes 3 and 4 assigned to the batch subpool and nodes 5 and 6 assigned to the interactive subpool. The pool with id=1 has seven nodes: nodes 7, 8 and 9 assigned to the general subpool, nodes 10 and 11 assigned to the batch subpool and nodes 12 and 13 assigned to the interactive subpool. The pool with id=2 has three nodes, 14, 15, 16, all assigned to the general subpool.

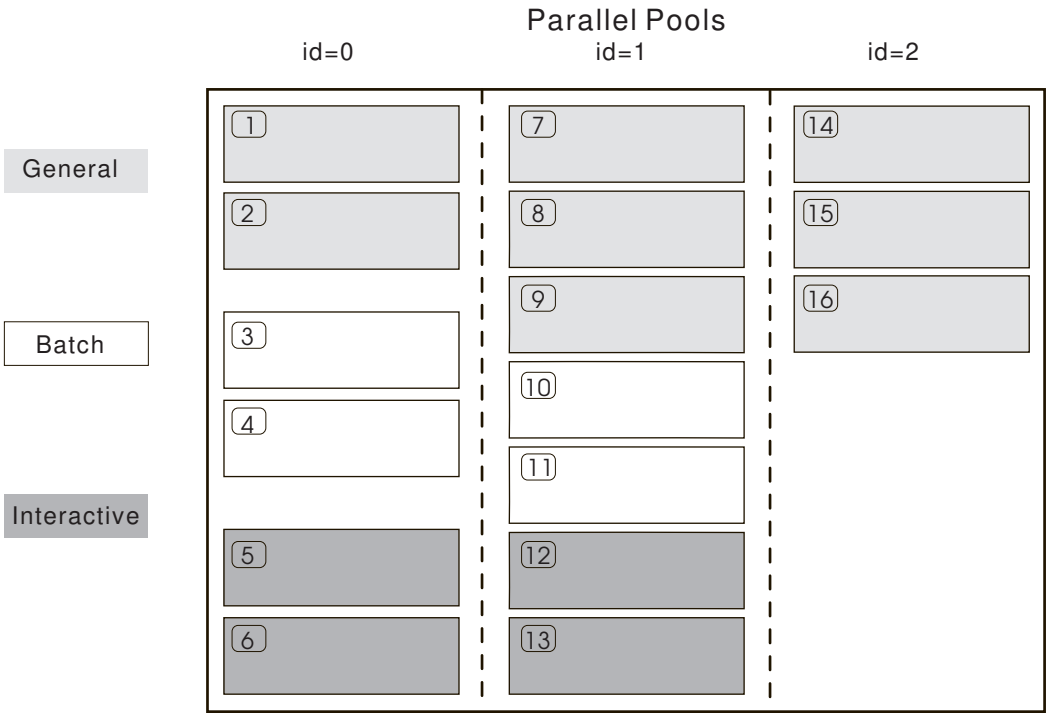


Figure 7. Example Resource Manager Configuration

Given this configuration, if an interactive job requested two nodes from Pool 1, it would receive nodes 12 and 13, if they were available. If an interactive job requested 4 nodes from pool 0, it would receive nodes 1, 2, 5, and 6. If a batch job requested 2 nodes from pool 2, it would receive two out of nodes 14,15,16. If a batch job requested 5 nodes from pool 1, it would receive nodes 7, 8, 9, 10 and 11.

Starting the Resource Manager

Follow these steps to start the Resource Manager within a system partition on your SP system.

Step 1: Run Login Access Control

If you wish to block users from parallel nodes unless they request access via the Resource Manager, then run the **spacs_cntrl** command as discussed in "Controlling User Login" on page 75. (See **spacs_cntrl** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for command syntax.)

Step 2: Ensure SDR Is Running

Run the verification test that checks for correct operation of the System Data Repository. You can run this verification test from SMIT panels or from the command line.

Using SMIT:

TYPE **smit SP_verify**

▲ The Installation/Configuration Menu appears.

SELECT System Data Repository

PRESS Enter

Using the Command Line:

Enter:

```
SDR_test
```

Refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* if verification test fails.

Step 3: Configure the Resource Manager

The Resource Manager keeps all general configuration data in an ASCII file named **/etc/jmd_config**. The system partition data is differentiated by adding a suffix to this file which corresponds to the current system partition environment (this environment can be determined by issuing **spget_syspar -n**). To define the Resource Manager configuration, edit the **/etc/jmd_config**.*syspar_name* file, which represents the current system partition environment, and assign your SP nodes to pools and subpools as described in "Defining the Resource Manager Configuration" on page 86. If you edit the **/etc/jmd_config**.*syspar_name* file manually, you must make sure the system partition named is correct (it should match the output of the **spget_syspar -n** command).Appendix C, "Sample Files" on page 503 shows a sample **jmd_config** file.

Using SMIT: As an alternative to editing the **/etc/jmd_config**.*syspar_name* file, you can use SMIT menus (available on the control workstation only) to define Resource Manager configuration. If you edit the **/etc/jmd_config** file using the SMIT interface, the correctly named **/etc/jmd_config**.*syspar_name* file is created for you. Once you have installed **ssp.jm** on all the nodes, you can configure the Resource Manager at the same time you use SMIT to configure the rest of the SP system. Detailed information on every field in the SMIT panels can be obtained via Help.

Note: Any data defined via SMIT is recorded in the **/etc/jmd_config**.*syspar_name* file. Therefore, the **jmd_config**.*syspar_name* file should always be consistent with the current configuration data.

To *initially* configure the Resource Manager:

TYPE **smit**

▲ The System Management menu appears.

SELECT RS/6000 SP System Management

▲ The RS/6000 SP System Management menu appears.

SELECT RS/6000 SP Resource Manager

▲ The RS/6000 SP Resource Manager menu appears.

The fastpath invocation for the RS/6000 SP Resource Manager menu is:

```
smit RM_options
```

```
                RS/6000 SP Resource Manager

□ Change/Show Configuration Data
□ Start the Resource Manager
□ Reconfigure the Resource Manager
□ Stop the Resource Manager
```

SELECT Change/Show Configuration Data

▲ The Configuration Data menu appears.

The fastpath invocation for the Configuration Data menu is:

```
smit RM_configure
```

```
                  Configuration Data

□ Change/Show Resource Manager Server List
□ Change/Show Access Control
□ Change/Show En Adapter
□ List All Pool Data
□ Add a Pool
□ Delete a Pool
□ Change/Show a Pool
```

SELECT Change/Show Resource Manager Server List

▲ The Change/Show Resource Manager Server List menu appears.

The fastpath invocation for the Change/Show Resource Manager Server List menu is:

```
smit server_list
```

```
                Change/Show Resource Manager Server List

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

     1st Node in Server List              [none]   + required
     2nd Node in Server List              [none]   +
     Additional Nodes in Server List      [none]     +
     Update Configuration Now ?           [no]       +
```

- The dialog fields show the current values found within the SDR. The initial configuration would show values of **none**.
- Enter the node names or use the list option to select the node names desired for each field.
- Enter **no** for `Update Configuration Now?`

SELECT Change/Show Access Control

▲ The Change/Show Access Control menu appears.

The fastpath invocation for the Change/Show Access Control menu is:

```
smit access_control
```

```
                        Change/Show Access Control

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

     Access Control                [no]          +
     Update Configuration Now ?    [no]          +
```

- The default value of the `Access Control Field` will be **no**.
- Use the list to select the value **yes**.
- Enter **no** for `Update Configuration Now?`

SELECT Change/Show En Adapter

▲ The Change/Show En Adapter menu appears.

The fastpath invocation for the Change/Show En Adapter menu is:

```
smit en_adapter
```

```
                        Change/Show En Adapter

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

     En Adapter                    [0]           +
     Update Configuration Now ?    [no]          +
```

- The default value of the `En Adapter` field will be **0** for **en0**.
- Use the list to select **en0** or **en1** (if it exists).
- Enter **no** for `Update Configuration Now?`
- Update will take affect only during a **jm_start** and not a **jm_config**.

SELECT Add a Pool (for each pool to be defined)

▲ The Add a Pool menu appears.

The fastpath invocation for the Add a Pool menu is :

```
smit add_pool
```

```
                          Add a Pool

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

□ Pool ID                          []
□ Description/Attribute            []
□ Delete nodes if previously defined ?      [yes] +
□ Batch Nodes                  [none]      +
□ Interactive Nodes            [none]      +
□ General Nodes                [none]      +
□ Update Configuration Now ? [no]      +
```

- The data fields initialize to blank or **none**.
- Enter the *pool_id* and *attribute* string.
- Enter the node names for the subpools or use the list option to select from valid nodes on the system.
- Enter **no** for `Update Configuration Now?`

(After using SMIT to define the initial configuration data, see the "Using SMIT" section of Step 4 to start the Resource Manager.)

Step 4: Start the Resource Manager

The **jm_start** command must be executed by root on the control workstation and operates within the scope of the current system partition environment.

Enter:

```
jm_start {-i}
```

The **jm_start** command reads the data in the /**etc/jmd_config**.*syspar_name* file into the SDR and starts the primary Resource Manager server on one of the nodes indicated. This server then starts up a backup on the next available node in the list of server candidates.

jm_start returns a message stating which **jmd_config** file (for which system partition) was used by the command.

If **jm_start** successfully starts the primary Resource Manager server, it returns a message indicating its location. If a backup is not successfully started, an additional message is issued and the primary Resource Manager server continues to run, attempting periodically to start the backup. If you do not have a backup running, and the primary Resource Manager server dies for any reason, there will be no recovery and the Resource Manager server must be restarted manually. Refer to "Resource Manager Recovery" on page 97 for more information.

Use the **-i** flag only if the **jm_start** command cannot determine whether the Resource Manager server is already running. The **-i** flag does not check if the Resource Manager server is already running, therefore it is best not to use this option unless you are sure the Resource Manager server is not running on any node.

Using SMIT: To start the Resource Manager using SMIT, select the Start the Resource Manager option from the RS/6000 SP Resource Manager menu.

The fastpath invocation for the RS/6000 SP Resource Manager menu is:

```
smit RM_options
```

```
                  RS/6000 SP Resource Manager

□ Change/Show Configuration Data
□ Start the Resource Manager
□ Reconfigure the Resource Manager
□ Stop the Resource Manager
```

Step 5. Verify that the Resource Manager Server Is Ready to Accept Requests

You can verify that the Resource Manager server is ready to accept requests from clients either from SMIT panels or by using the the verification script, **jm_verify**.

Using SMIT:

TYPE	**smit SP_verify**
	▲ The Installation/Configuration Menu appears.
SELECT	Resource Manager Configuration
PRESS	Enter

Using the Command Line: You can use the verification script, **jm_verify**, to verify that the Resource Manager server is ready to accept requests. This command resides in **/usr/lpp/ssp/bin** and can be run on any node by any user and operates within the scope of the current system partition environment.

Enter the command as follows:

```
jm_verify
```

Optional flags

-l *filename*	Specifies log filename (Default: **/var/adm/SPlogs/jm_verify.log**).
-q	Suppress messages
-h	Print Usage information

See the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* for more information and examples.

Reconfiguring the Resource Manager

At some point you may decide to redefine your subpools and assign different nodes for batch, interactive, or general use, or change the list of candidate nodes for running the Resource Manager server, or change the setting of login access control. You must change the **jmd_config**.*syspar_name* file to reflect the new configuration and issue the **jm_config** command. (**jm_config** will return a message stating the name of the **jmd_config** file (for which system partition) the command is using.)

If you decide to enable (or disable) SP Exclusive Use Accounting, thereby modifying the **spacct_excluse_enable** attribute in the SP object of the SDR, when the Resource Manager is already running, be sure to reconfigure the Resource Manager by issuing **jm_config**. If you modify any SDR attributes that affect Node or Adapter objects, such as adding or deleting a frame, when the Resource Manager is already running, you must stop the Resource Manager and restart it in order to reread the SDR data.

Step 1: Edit the Resource Manager Configuration File or Update Configuration Information

Edit the /**etc/jmd_config**.*syspar_name* file to specify which nodes belong to the batch, interactive, and general subpools of your serial and parallel hardware pools. You can also change the list of nodes to be used for running the Resource Manager server, and the setting for Login Control.

Using SMIT: If you use the SMIT to interface to reconfigure the Resource Manager, the new and correctly named /**etc/jmd_config**.*syspar_name* file is created for you. To *change* the current Resource Manager pool configuration using SMIT, use the fastpath invocation for the Change/Show a Pool option:

```
smit configure_pool
```

▲ The Change/Show a Pool menu appears.

```
                        Change/Show a Pool

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

     Pool ID                          1
     Description/Attribute                [widenode_userspace]
     Delete nodes if previously defined ?      [yes]    +
     Batch Nodes                  [none]      +
     Interactive Nodes            [r07n01 r07n03]
     General Nodes                    [none]      +
     Update Configuration Now ?       [no]  +
```

- Select a previously defined pool and view the current set of defined data.
- Enter the node names for the subpools or use the list option to select from valid nodes on the system.
- Enter **yes** for `Update Configuration Now?`

- If you do not wish to update configuration at this point, enter **no** for `Update Configuration Now?` At a later time you can use the Reconfigure the Resource Manager option on the RS/6000 SP Resource Manager menu to reconfigure according to the values you have just entered in the fields.

Step 2: Run jm_config

This command, which must be run from root on the control workstation, and operates within the scope of the current system partition environment, reads configuration data from the **jmd_config**.*syspar_name* file into the SDR. It checks the SDR for the location of the primary Resource Manager and then sends the Resource Manager a message telling it to reread the configuration data from the SDR. The new configuration takes effect with the next client request.

Enter:

```
jm_config
```

jm_config returns a message stating the name of the **jmd_config** file (for which system partition) the command is using.

Using SMIT: To reconfigure the Resource Manager using SMIT, select the Reconfigure the Resource Manager option from the RS/6000 SP Resource Manager menu. The fastpath invocation for the RS/6000 SP Resource Manager menu is:

```
smit RM_options
```

```
                    RS/6000 SP Resource Manager

□ Change/Show Configuration Data
□ Start the Resource Manager
□ Reconfigure the Resource Manager
□ Stop the Resource Manager
```

The currently executing clients will not be affected by the change in configuration. The server recognizes the new configuration for new clients making a connection with the Resource Manager server.

Stopping the Resource Manager

To stop the primary and backup Resource Manager servers, use the **jm_stop** command. This command must be run by root on either the control workstation or on any of its associated nodes and operates within the scope of the current system partition environment.

Enter:

```
jm_stop {-q}
```

The q flag is optional. If not specified, the primary and backup Resource Manager server exit immediately.

The system can be quiesced in a less disruptive fashion by specifying the **-q** option, which tells the Resource Manager server to first wait for all existing jobs to

terminate before exiting. The **jm_stop -q** command will return immediately. Any subsequent Resource Manager job requests, as well as **jm_status** and **jm_config** commands, will fail. When all jobs have finished, the primary and backup Resource Manager server will exit. Issue **locate_jm** to determine if the Resource Manager servers have exited. Anytime after **jm_stop -q** has been issued, you can issue **jm_stop** (without **-q**) to cause the Resource Manager servers to exit immediately.

Using SMIT: The fastpath invocation for the Stop the Resource Manager menu is:

```
smit RM_stop
```

```
                  Stop the Resource Manager

□ Stop the Resource Manager immediately
□ Wait for existing clients to exit, then stop the Resource Manager
```

Locating the Resource Manager

To identify the nodes on which the primary and backup Resource Manager servers are running, use the **locate_jm** command. This command must be executed by root and has no parameters. It operates within the scope of the current system partition environment.

Querying the Resource Manager

Use the **jm_status** command to get information about defined pools or about jobs running on nodes allocated by the Resource Manager. **jm_status** can be executed by any user, on any workstation.

Enter the command as follows:

```
jm_status {-n control_workstation_name} {-P|-p pool_id|-j} {-v}
```

1. The **-n** flag indicates the name, or alias name, of the SP control workstation. This flag is optional. If you choose not to specify the **-n** flag, you can set the `SP_NAME` environment variable to indicate the control workstation name. If neither is specified, the control workstation name is obtained from the local **/etc/SDR_dest_info** file.
2. One, and only one, of the **-P**, **-p**, or **-j** flags must be specified.
   - The **-P** flag requests information about all the pools in the SP system.
   - The **-p** flag requests information about the pool specified by **pool_id**. A **pool_id** value of -2 requests information about all the parallel pools. A **pool_id** value of -3 requests information about all pools and is equivalent to specifying **-P**.
   - The **-j** flag requests information about all the jobs running on the system.
3. The **-v** (verbose) flag is always accepted, but no special action is taken when it is specified together with **-j**. If **-P** or **-p** is specified:
   - Without the **-v** flag, only information about pool IDs, pool attributes and member nodes is displayed.

- With the **-v** flag, additional information is displayed about jobs running in the specified pools.

Resource Manager Recovery

When the Resource Manager is started, the primary server automatically tries to start a backup server on another node, identified in the configuration file. If it is unable to do so, it will retry at regular intervals. Once the backup server has been started, the primary server keeps an open IP socket connection with it, and polls it periodically to check that the backup server process is alive. In addition, it checks the SDR to see if the backup server node is alive, since a node crash will not cause any message to appear on the IP socket. If it is determined at any point in the primary server's operation that the backup server process has died, the primary server will attempt to start another backup. The primary server records node allocation in the SDR.

The backup server periodically polls its IP socket connection with the primary server to check that the primary server process is alive. For the same reasons as described above, it also checks the SDR to see if the primary server node is alive. If it is determined that the primary server has died, the backup takes over as the new primary server by reading configuration and job data from the SDR. It waits a period of three minutes for all existing clients to validate their existence by reconnecting to it. During this time, the new primary server will accept only connection and reconnection requests from clients. After the three-minute interval, the job data in the SDR pertaining to invalid clients (those that did not reconnect) is removed. The backup server is now ready to accept all client requests and operate as the primary server. With this recovery scheme, jobs should be able to run unaffected despite the failure of the primary Resource Manager server.

The Resource Manager requires that the SDR be operational. If the control workstation is rebooted, or any other activity occurs that causes the SDR to be restarted, the Resource Manager will no longer be usable, and will eventually die. Under these circumstances the administrator should issue **jm_stop**. Once the SDR is again operational, issue **jm_start**.

If the primary and backup servers die at the same time, there is no recovery. Should this occur:

1. Quiesce the system (wait till all currently running parallel jobs terminate).
2. If the Resource Manager was configured with ACCESS_CONTROL=yes, run the **spacs_cntrl** command to reset user node access to its original value. The Resource Manager might have left it in an undesirable state upon failure.
3. Execute **jm_start** from the control workstation to restart the Resource Manager.

Chapter 8. Managing Time Synchronization

There are several ways you may currently be handling synchronizing time-of-day clocks on your control workstation and processor nodes. You may already be using Network Time Protocol (NTP) either locally or through the Internet, or you may be using some other time service software. It is also possible that you have no established method for synchronizing the system clocks in your computing environment.

The SP system gives you 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 other methods at your site. **The SP machines do not have system batteries. If you choose not to use NTP, you must have another way to manage clock synchronization.**

Managing NTP

There are three attributes to set for NTP. *ntp_version* defaults to a value of **3**, the version shipped with the SP System Support Programs. If your installation is using an earlier version of NTP, change this value. The other two attributes are described in Table 3.

Table 3. 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.
- *Reference **/usr/lpp/ssp/ssp.public.README** for information on Internet time servers.

When you choose any of the options that involve using NTP, the installation scripts determine the IP address of the host specified as **ntp_server** and creates the **/etc/ntp.conf** file on each of the control workstation, boot/install server, and processor nodes. This file indicates which hosts can provide time service to which other hosts using the designations of *server* and *peer*, where

server Indicates a host from which you can request time.

© Copyright IBM Corp. 1995, 1997

peer Indicates a host with which this host can synchronize time. either peer host can request time from the other.

The following example shows a sample /**etc**/**ntp.conf** file:

```
# Server ip_addr version <ntp version>
server 9.8.77.66 version 3      # server is host_one.conrad.ibm.com
server 123.45.678.9 version 3   # server is host_two.ursula.ibm.com
peer 9.8.77.55 version 3        # server is host_x.conrad.ibm.com
peer 9.8.77.44 version 3        # server is host_y.conrad.ibm.com
peer 9.8.77.33 version 3        # server is host_z.conrad.ibm.com
```

Assuming this is the /**etc**/**ntp.conf** file on *your_host*, the following is true:

- *your_host* can request time from *host_one*, *host_two*, *host_x*, *host_y*, and *host_z*
- *your_host* can provide time, on request, to *host_x*, *host_y*, and *host_z*
- *your_host* cannot provide time to *host_one* or *host_two*.

The following table shows which host addresses will be used for servers and peers in the /**etc**/**ntp.conf** files created on the control workstation, file servers, and processor nodes for each **ntp_config** choice.

ntp_config	Control Workstation	Boot/Install Servers	Processor Nodes
timemaster	SERVER: your time server PEER: all boot/install servers	SERVER: your time server PEER: control workstation PEER: all boot file servers	SERVER: your time server SERVER: control workstation SERVER: all boot file servers
Internet	SERVER: Internet server	SERVER: control workstation PEER: other boot file servers	SERVER: control workstation SERVER: all boot file servers
consensus	SERVER: 127.127.1.10*	SERVER: control workstation PEER: other boot file servers	SERVER: control workstation SERVER: all boot file servers
none	no file created	no file created	no file created
Note: * This address signifies local time service to NTP and makes the control workstation a stratum 10 time server allowing you to hook into a higher stratum Internet time server at a later date.			

Related Information: The manual pages for NTP provide additional information on how NTP synchronizes time across a network. They are available in /**usr**/**lpp**/**ssp**/**man**/**man8**. The man pages are xntpd and xntpdc.

For more information on the installation scripts, see the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

For directions on updating the site environment see the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

Time Setting Options

To change time zones, use the **chtz** *timezone_info* command. For example:

```
chtz EST5EDT
```

sets the time zone to Eastern Standard Time and 5 hours from Greenwich Mean Time and sets the machine to know about Daylight Savings Time.

Nodes and Control Workstations Out of Time Synchronization

Generally, time services will adjust clocks incrementally to keep in sync with a master. However, when time is off by a large amount, for example, an hour, it becomes difficult and time-consuming to incrementally synchronize with the time master.

To ensure that clocks stay in sync, set nodes and the control workstation to understand Daylight Savings Time. Refer to the **script.cust** sample file (“script.cust File” on page 540) for information on setting time zones.

Chapter 9. Managing the Automounter

This chapter provides information on the use of an automounter on an SP system to manage the mounting of user home directories and other directories. The following topics are discussed:

- Understanding the automounter implementation
- Establishing an automount map for users' directories
- Using automount for other file systems
- Customizing the SP automounter function
- Coexistence considerations
- Migration considerations

For more information on the AIX automount daemon that is shipped as part of NFS in the AIX Network Support Facilities of the Base Operating System (BOS) Runtime, refer to *IBM AIX Version 4 System Management Guide: Communications and Networks*.

Understanding the Automounter Implementation

An automounter is a facility used to manage the mounting activity of a file system. When you access a file or directory under automounter control, the automounter transparently mounts the required file system. When there has been no activity to that file system for some pre-determined amount of time, the automounter unmounts the file system.

Automounters are typically used with the Network File System (NFS). NFS is a distributed file system that allows you to access files and directories located on remote systems and treat those files and directories as if they were local. When performing its mounting activity for an NFS file system, the automounter uses the NFS mounting facilities. The automounter reduces the period of time that a file system is actively mounted, thereby minimizing local system hangs due to NFS server outages.

On an SP system, the automounter is optionally used to manage mounting of user home directories and other directories. When configured, an automounter daemon runs on each node and is started at the time the node is booted. It mounts directories on demand and unmounts them after a period of inactivity. The mounted directories can come from SP boot/install servers or any workstation or server in the network. If the directory to be mounted appears as a local directory on the node (as, for example, with an AFS directory), the automounter simply creates a symbolic link to that directory instead of attempting to mount it.

The automounter manages directories specifically defined in automounter map files. Typically, there is one map file for each file system to be controlled by the automounter. The map files contain entries for each directory supported within the file system and the specific mount information for that directory. If SP User Management services have been configured, the SP creates and maintains a map file to control user home directories in the /**u** file system.

© Copyright IBM Corp. 1995, 1997

Requesting SP Automounter Support

You can choose to allow the SP to use an automounter to provide remote access to users' home directories or you can rely on other methods of automounting or file system management established at your site. You specify your choice by setting the automounter configuration site environment variable **amd_config** when defining your site environment. Specify your choice for the **Automount Configuration** variable on the Site Environment Information SMIT panel or use the **spsitenv** command on the control workstation. If **amd_config** is **true** (the default setting), the SP does the necessary directory and file configuration, and starts the automounter daemon. If SP User Management services have also been configured (the **usermgmt_config** site environment variable is **true**), the SP adds, removes, and changes entries in the automounter map file for the /**u** file system to maintain user home directory information when SP users are added, removed, or changed.

If you set **amd_config** to **false** when you initially install the SP system, it will not configure or start the automounter daemon and it will not maintain user home directory information in automounter map files. You can change the initial setting of **amd_config** at a later time using SMIT or the **spsitenv** command. However, this change will not affect the running state of the automounter daemon until the next time you reboot your nodes or control workstation or take other actions to start or kill the daemon. Note that if **amd_config** is initially **false** and then later set to **true**, all user management changes that had occurred while **amd_config** was **false** will not be reflected in the automounter map file for user home directories and that information may need to be manually added. If you choose to unconfigure the automounter by changing **amd_config** from **true** to **false**, make sure that all user home directories, as well as other directories previously served by the automounter, are distributed and made available by some other means.

When **amd_config** is **true**, the SP system uses the AIX automount daemon that is shipped as part of NFS in the AIX Network Support Facilities of the Base Operating System (BOS) Runtime. *IBM AIX Version 4 System Management Guide: Communications and Networks* describes the operation and use of this automounter and the format of the map files. In addition to the map file for the /**u** file system maintained by the SP system, you may provide your own automount map files to enable the **automount** daemon to control similar access to directories in other file systems that you would like to have automatically mounted on demand.

If you choose to install and use a different automounter, the SP system allows you to provide a set of user customization scripts that the system will use to replace its functions for configuring and starting the automounter daemon and managing the map file for the user home directory information. **amd_config** must still be set to **true**.

If your SP system contains nodes that are running PSSP 2.2 or previous releases, those nodes use Amd, the BSD automounter. The **amd_config** site environment variable specifies your choice for SP automounter configuration for all nodes, including these older nodes running Amd. The control workstation still maintains the Amd map file, in addition to the Automount map file, to support those nodes running Amd.

Automount Map Files

The AIX **automount** daemon reads automount map files to determine which directories to support. Typically, there is one map file for each file system to be controlled by the automounter. The map file contains entries for each directory supported within the file system, the host name where the directory resides, and the specific mount information for that directory. The automount map files are kept in the **/etc/auto/maps** directory, by default. The list of all map files to be used by the **automount** daemon is specified in the master map file **/etc/auto.master**. This master map file contains entries for each file system to be controlled by the automounter, the name of the map file containing the directory information, and optional default mount information.

If automounter support and the SP User Management interface are both configured for your site environment, **/u** is automatically controlled by the automounter. The SP-managed automount map file for the **/u** file system is **/etc/auto/maps/auto.u**. This map file contains a stanza for each user added using SP User Administration (either the **spmkuser** command or SMIT panels), containing the server name and directory location of the user's home directory. Existing users must be added to the map file using the **mkamdent** command.

An entry in the automount map file has the following format:

```
key -mount_options server_name:mount_directory:sub_directory
```

where:

key	The directory of interest within the file system
mount_options	Options used during the mount operation
server_name	The name of the machine that contains the directory
mount_directory	The exported directory from that machine
sub_directory	An optional sub-path under the exported directory

A special substitution value of & can be used to substitute the *key* value in the entry. If the server name is the name of the local machine, the **automount** daemon will create a symbolic link to the mount directory instead of attempting to mount it.

Note: An entry is added to **auto.u** for the net install ID's home directory. This **netinst** user ID is required by AIX for a net install.

Important...

The **automount** daemon will mount file systems in the **/tmp_mnt** directory using the following naming convention:

/tmp_mnt/*filesystem*/*first-accessed-sub_directory*

where *filesystem* is the name of the file system under automount control and *first-accessed-sub_directory* is the name of the first sub-directory accessed within a mounted directory. A symbolic link is made from the directory the user accessed to the appropriate sub-directory within that mount point:

/filesystem/sub_directory→**/tmp_mnt/***filesystem*/ \
/first-accessed-sub_directory/sub_directory

where *sub_directory* is the name of the subpath currently being accessed by a user. This naming convention can be quite confusing because the physical directory path is dependent on some arbitrary name (that is, the first sub-directory accessed within a mounted directory). Because this name can change from one access period to another, you should not make any assumptions about this name or directly use it in any scripts or programs. Always access the directory through the automounter-supported path name. This naming convention may cause problems for C-shell users because the C-shell **pwd** built-in command returns the actual physical path and not the automounter-supported path, which is a symbolic link. The results of the C-shell command should not be stored for later use by the same program or another program or shared across a parallel execution environment; there is no guarantee the physical path name will exist.

The following example illustrates how an automount map entry is used by the **automount** daemon to provide access to a user's home directory.

Example

In the SP system, by default, the users' directories reside in /**home**/*hostname*, where *hostname* is the short name for the host system. The following stanza shows the entry in the automount map /**etc**/**auto**/**maps**/**auto.u** for user **jws** whose home directory resides on **luna**:

```
jws luna:/home/luna:&
```

In this case, if a user is on **luna** and changes directories to /**u**/**jws**, automount will create a symbolic link to the local directory:

```
/u/jws→/home/luna/jws
```

However, if the user is on any machine other than **luna** and changes directories to /**u**/**jws**, automount mounts **luna:**/**home**/**luna** in the local /**tmp_mnt** directory at /**tmp_mnt**/**u**/**jws**. Automount then creates a symbolic link to the **jws** sub-directory under that mount point:

```
/u/jws→/tmp_mnt/u/jws/jws
```

Because **jws** was the first user to access a sub-directory in **luna:**/**home**/**luna**, the **jws** sub-directory name appears twice in the physical path name. If a second user, **kcb**, also accesses a sub-directory in **luna:**/**home**/**luna**, automount creates the following symbolic link:

```
/u/kcb→/tmp_mnt/u/jws/kcb
```

Note the presence of **jws** in the physical path name. This is the name of the mount point the automounter used when mounting **luna:/home/luna**.

If a user's home directory resides in some other type of file system that already appears locally on the machine (for example, AFS), the server name should be specified as the local host. A map entry for a user with an AFS home directory would look like this:

```
tberry $HOST:/afs/kgn.ibm.com/usr2:&
```

In this case, when a user changes directories to /**u**/**tberry**, automount creates a symbolic link to the AFS directory:

```
/u/tberry→/afs/kgn.ibm.com/usr2/tberry
```

Note: The use of the *$HOST* substitution variable is specific to the SP invocation of the **automount** daemon. The AIX **automount** command allows you to specify substitution variables and values to be used in the map files. The SP invocation of the **automount** daemon includes the parameter **-D HOST**=‘*uname* **-n**’ which sets *$HOST* to the name of the local machine as returned by the **uname** command.

See *IBM AIX 4 System Management Guide: Communications and Networks* for more information on the **automount** daemon and map file format.

Automount Master Map Files

The SP invocation of the **automount** daemon uses the master map file /**etc**/**auto.master**. This master map file contains entries for each file system to be controlled by automount, the name of the map file containing the directory information, and optional default mount information. If you add additional file systems to be controlled by the **automount** daemon, you must edit this file and add an entry indicating the name of the file system and the full directory path of the associated map file. See “Using Automount for Other File Systems” on page 109.

If you wish to change the default mount options for mounting directories for the /**u** file system or any other file system controlled by the **automount** daemon, edit the /**etc**/**auto.master** file to add the mount options to the end of the appropriate file system entry. For example, to add mount options to the /**u** file system, your /**etc**/**auto.master** entry may be modified to appear as follows:

```
/u /etc/auto/maps/auto.u -rw,hard,retry=3,timeo=40,rsize=4096,wsize=4096
```

The master map file can also specify an NIS database as a map file. The Network Information Service (NIS) is a distributed database that allows you to maintain consistent configuration files throughout a network. By default, the SP invocation of the **automount** daemon turns off the use of the **auto.master** NIS database. If you are using NIS to maintain and distribute your automounter map files, you will need to edit the /**etc**/**auto.master** file to add as the first entry +**auto.master** which includes the NIS **auto.master** database.

See *IBM AIX 4 System Management Guide: Communications and Networks* for more information on the automount master map file format and on NIS.

File Collections and Automounter Files

If you have configured the SP to use file collections by setting the site environment variable **filecoll_config** to **true**, automount map files will be automatically distributed to all nodes. The automount map files are part of the **user.admin** file collection. The **/var/sysman/sup/user.admin/list** file contains entries to distribute the **/etc/auto.master** file, all files in the **/etc/auto/maps** directory, and all files in the **/etc/auto/cust** directory.

If you decide to replace some or all of the SP automounter function with your own customizations, you may need to edit the **/var/sysman/sup/user.admin/list** file to distribute your own files or to remove the currently distributed SP files. In the list file, there exists a comment mentioning that SP automounter configuration has been added. Do **NOT** delete or change this comment. You may change or remove any other entries in the file as necessary for your automounter installation. However, if this comment is removed or altered on the control workstation, the next time your system is booted or the **services_config** command is run, the SP configuration will add the automount entries back again. See "Customizing the SP Automounter Function" on page 112 for more details on providing your own automounter customizations.

If you are not using file collections on your SP, you must use some other local means of distributing the automount map file, master map files, and user customization scripts to all of the nodes on your system. The SP will not copy your map files to the nodes for you.

See Chapter 5, "Managing File Collections" on page 41 for more information on the file collection processes.

Automounter Diagnostics

The SP automounter configuration and execution functions write error messages to an error log file, **/var/adm/SPlogs/auto/auto.log**. Also, the **automount** daemon uses the daemon facility of the BSD syslog subsystem to record errors. The control workstation and nodes have been configured to write all **daemon.notice** and greater messages to **/var/adm/SPlogs/SPdaemon.log**. If you are having problems with the automounter, you may need to stop the daemon, check the diagnostic information recorded in these logs, and restart the **automount** daemon. For more information on error logging, refer to Chapter 27, "Managing Error Logs" on page 409. You can find a description of the automounter error log, messages, and instructions for stopping and starting the **automount** daemon in *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide*.

Establishing an Automount Map for Users' Directories

If you have configured your SP to include both automounter and user management support, the SP will create and maintain an automount map file for users' home directories. When you add a new SP user to your system, a map entry will be added to **/etc/auto/maps/auto.u** associating the user's **/u** home directory with the actual physical directory located either on the current machine or on a remote file server that can be accessed through a network connection using NFS. This map entry will be changed or removed when SP User Management functions are used to change or remove the user from the system.

If you have existing users that do not have entries in the **auto.u** map file, you can use the **mkamdent** command to add home directory entries for them. The **mkamdent** command allows you to specify a list of users to add to the **auto.u** map. The command builds a stanza for each user in the list using the SP site environment values for **homedir_server** and **homedir_path** as the defaults (these can be set using either SMIT panels or the **spsitenv** command). **homedir_server** is the default file server and **homedir_path** is the default base path for the home directory (not including the individual user's sub-directory). You can override these defaults with the **-s** flag. See *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on the **mkamdent** and **spsitenv** commands.

You must make sure that the /**u** directory is not a symbolic link to some other directory. This will prevent the **automount** daemon from starting successfully. /**u** must exist as a local directory on each processor.

If you are not using SP User Management services, but would still like to have the **automount** daemon control access to directories in the /**u** file system, you can create and maintain your own **auto.u** map file. See "Using Automount for Other File Systems" for details on how to add the map to the SP automounter support.

Changing User Home Directory Information

It may become necessary to change the user home directory information stored in the automounter map file. If, for example, the home directory is moved to a new server machine or to a different directory, the existing entry in the automounter map file will need to be updated with the new information. We recommend changing this information by using the **spchuser** command or the equivalent SMIT panels.

After updating the map file, you may need to wait up to 5 minutes with no access attempts to the user's home directory through the /**u** file system. This allows the automounter to time out any previous access attempts and unmount any existing mounts that may be referencing the old information. Do not attempt to force the unmount by manually issuing the **umount** command as that will put the automounter in an inconsistent state. You would then need to stop and restart the **automount** daemon to recover access to that user's home directory.

Using Automount for Other File Systems

If you have configured your SP with automounter support, you can use the **automount** daemon to provide automounting control for other file systems in your network. Each file system you want managed by automount requires an automount map file and an entry in the /**etc**/**auto.master** map file identifying the file system and its map file.

You must make sure the directory you are controlling with the **automount** daemon exists as a local directory on each processor and is not a symbolic link to another directory. Otherwise, the **automount** daemon will not start successfully.

The process to add additional file systems requires you to stop and start the **automount** daemon on all nodes. To do so, all activity with files and directories currently controlled by the **automount** daemon must be stopped. This includes all activity in users' home directories if the **automount** daemon is controlling the /**u** file

system. You should plan to add all additional file systems to the automounter at a time when this would not adversely affect your system schedule or disrupt users.

The examples in this section illustrate the process for adding additional file systems and show how to provide automount support for automounting the /**vol** file system.

The steps in this process are:

1. Create an automount map file
2. Update the master map file
3. Add the map to the file collections
4. Distribute the map to the boot/install servers and nodes
5. Stop and restart the **automount** daemon

Step 1: Create an Automount Map File

The easiest way to create an automount map file is to copy an existing one and modify it for the new file system. The map can reside in any directory you like. The map for the /**u** file system is /**etc**/**auto**/**maps**/**auto.u**. It may be convenient for you to keep all of your automount map files in this **maps** directory.

Example

1. On the control workstation, copy the existing **auto.u** map and name it **auto.vol**, keeping it in the same directory. Enter:

   ```
   cp /etc/auto/maps/auto.u /etc/auto/maps/auto.vol
   ```

2. Replace the map entries with the directory and host information for /**vol** as shown here:

   ```
   inst.images  images1:/inst.images
   techlib      filesrv1:/share/&
   local        $HOST:/afs/site.edu/share/local
   ```

This example results in the following mounts:

- If a user issues **cd** /**vol**/**inst.images**, **inst.images** is mounted from **images1** linking /**vol**/**inst.images**→/**tmp_mnt**/**vol**/**inst.images**
- If a user issues **cd** /**vol**/**techlib**, /**share**/**techlib** is mounted from **filesrv1** linking /**vol**/**techlib**→/**tmp_mnt**/**vol**/**techlib**
- If a user issues **cd** /**vol**/**local**, automount simply creates a symbolic link because the directory is accessible from the local host. The result is /**vol**/**local**→/**afs**/**site.edu**/**share**/**local**.

Step 2: Update the Master Map File

Add an entry to the automount master map file /**etc**/**auto.master** so that when the **automount** daemon is restarted it will also control the file system being added. The entry may contain default mount options for all the directories listed in the automount map file.

Example

Add the following entry to **/etc/auto.master**:

```
/vol /etc/auto/maps/auto.vol -ro,soft,rsize=4096,wsize=4096
```

This will tell the **automount** daemon to use the **/etc/auto/maps/auto.vol** file when controlling access to directories within the **/vol** file system. The specified mount options will be used for each directory mount.

Step 3: Add the Map to the File Collections

If you are using file collections to distribute your map files, add the new map so that it can be distributed and managed by the file collection programs. If you have placed your map file in the **/etc/auto/maps** directory, you do not need to do anything for this step. The **user.admin** file collection already distributes all files in the **/etc/auto/maps** directory. However, if you have not placed your map file in the **/etc/auto/maps** directory, you will need to add the map to a file collection that is resident on all the systems you require. The **user.admin** collection is a good choice because it is resident on the control workstation, boot/install servers, and all processor nodes. See “SP File Collection Summary” on page 44 for a complete list of file collections provided by the SP.

Example

Add **auto.vol** to the **user.admin** file collection so that it can be distributed to all boot/install servers and nodes. To do this, edit the file collection list file **/var/sysman/sup/user.admin/list** and add the following entries to request an upgrade of this new map and to refresh the daemon after the upgrade:

```
upgrade ./etc/auto/maps/auto.vol
execute /etc/amd/refresh_amd (./etc/auto/maps/auto.vol)
```

Note: This step is not necessary for this example since the map file resides in the **/etc/auto/maps** directory and will already be automatically distributed. The details are provided here to illustrate how to update the file collection if the map file resides in some other directory.

Step 4: Distribute the Map to the Boot/Install Servers and Nodes

If you are using file collections to distribute your map files, check to see if you have a **scan** file for the file collection containing the new map.

Example

If you are using the **user.admin** file collection to distribute your map, check if the file **/var/sysman/sup/user.admin/scan** exists on the control workstation or any boot/install servers. If the **scan** file does exist, you will need to update it. On each processer that has a **scan** file for the collection, enter:

```
supper scan user.admin
```

You can wait for the scheduled **supper update** to update the file collection as specified in the **crontabs** file on each processor node. Or, if you want to distribute the map immediately, use a remote command to run the **supper update** command first on each boot/install server and then on each node.

Example

To distribute the **user.admin** file collection with the new map file immediately, run the following command first on each boot/install server and then on each node:

```
supper update user.admin
```

Step 5: Stop and Restart the automount Daemon

In order for automount to recognize the update to the /**etc**/**auto.master** file and handle the new file system, the **automount** daemon must be stopped and restarted. The daemon must be stopped at a time when no users are accessing files or directories under automount control. Plan this step for a time that is not disruptive to your users.

The cleanest and safest way to stop and restart the automount daemon is to reboot all the systems where this new map will reside.

If it is not possible to reboot the systems, then do the following to stop the **automount** daemon:

1. Issue the **mount** command with no parameters to get the process ID of the **automount** daemon and to see if there are any active mounts on the /**tmp_mnt** directory. This is the directory the **automount** daemon uses as its mount point.
2. Stop all processes that are accessing files or directories controlled by the **automount** daemon.
3. Stop the **automount** daemon with the following command using the process ID returned by the mount command:

   ```
   kill -term process_id
   ```

   Note: It is important to issue the **kill -term** command (**kill -15**) and not **kill -kill** (**kill -9**) to allow the **automount** daemon to properly clean up its mounts and directories.

Restart the **automount** daemon by running the following command:

```
/etc/auto/startauto
```

Customizing the SP Automounter Function

The SP automounter function has been implemented so that you can optionally replace some or all of the SP automounter function with your own customization scripts. By placing executable files in the /**etc**/**auto**/**cust** directory, you can provide your own automounter function that will be executed in place of the SP automounter function. This is useful if you need to do things such as change the map file entry format for entries in the /**u** map file /**etc**/**auto**/**maps**/**auto.u**, change the default options for invoking the **automount** daemon, or completely replace the SP use of the **automount** daemon with some other automounter that you provide.

In order to use these customization scripts, the SP automounter support must be configured for your system (the **amd_config** site environment variable is set to **true** using either the **spsitenv** command or SMIT panels). During the execution of the SP automounter function, a check will be made to see if any of the following files exist and are executable. If so, that file will be invoked in place of the corresponding automounter function:

/etc/auto/cust/cfgauto.cust Configure the automounter directories and create default files as necessary

/etc/auto/cust/startauto.cust Start the automounter daemon

/etc/auto/cust/refauto.cust Refresh the automounter daemon

/etc/auto/cust/checkauto.cust Verify the automounter configuration and daemon execution

/etc/auto/cust/mkautoent.cust Add an entry to the automounter map file for user home directories

/etc/auto/cust/rmautoent.cust Remove an entry from the automounter map file for user home directories

/etc/auto/cust/lsautoent.cust List an entry from the automounter map file for user home directories

Each file is independent and any or all files may exist. The SP automounter function will be replaced only for those functions which have an executable customization script. Details on how each of these files will be used by the SP is provided in the following sections.

If the automounter function is to be replaced on all boot/install servers and processor nodes, the customization scripts must be distributed. If you have file collections configured for your system, these files will automatically be distributed through the **user.admin** file collection. If you are not using file collections, you must distribute these files to the boot/install servers and processor nodes using some other means.

cfgauto.cust Customization Script

The **/etc/auto/cust/cfgauto.cust** file is an optional customization script that you may provide to configure the automounter directories and create default files. If this file exists, and it is executable, it will be invoked by the **services_config** installation command during the system boot process to configure your automounter installation. You should provide this file when you require a different automounter configuration than that provided by the SP. We recommend that your **cfgauto.cust** script create all directories and default files that are required by the automounter started by your **/etc/auto/cust/startauto.cust** script.

The **services_config** process will create the SP automounter log file as **/var/adm/SPlogs/auto/auto.log**. It will then check if **cfgauto.cust** exists and is executable, and if so, will invoke it. If not, the SP automounter configuration will be invoked using the **/usr/lpp/ssp/install/bin/cfgauto** installation script which creates the **/etc/auto** and **/etc/auto/maps** directories, and creates default **/etc/auto.master** master map and **/etc/auto/maps/auto.u** map files as necessary.

There are no input parameters passed to the **cfgauto.cust** script. The return value should indicate the following:

0 Success

non-0 A configuration problem occurred. In this case, the **services_config** process will not start the automounter daemon, but will continue with its other configuration functions.

Example

The following script is an example of a script you might use to configure directories and default files for a processor to use the Amd automounter instead of the AIX **automount** daemon:

```
#!/usr/bin/ksh

# Create /etc/amd directory
if [[ ! -a /etc/amd ]]; then
    mkdir /etc/amd
    if (($? != 0)); then
        echo "cfgauto.cust:  Cannot create directory /etc/amd."
      exit 1
    fi
fi

# Create /etc/amd/amd-maps dir
if [[ ! -a /etc/amd/amd-maps ]]; then
    mkdir /etc/amd/amd-maps
    if (($? != 0)); then
        echo "cfgauto.cust:  Cannot create directory /etc/amd/amd-maps."
       exit 1
    fi
fi

# Create default amd.u file with /defaults entry and an
# entry for the netinst userid
if [[ ! -a /etc/amd/amd-maps/amd.u ]]; then
    echo "/defaults rfs:=/homes/filesrv1;type:=nfs;sublink:=${key};
opts:=rw,hard,retry=3, timeo=40,rsize=4096,wsize=4096" >>
/etc/amd/amd-maps/amd.u
    if (($? != 0)); then
        echo "cfgauto.cust:  Cannot create default file
        /etc/amd/amd-maps/amd.u."
       exit 1
    fi
    echo "\nnetinst type:=link;fs:=/home" >> /etc/amd/amd-maps/amd.u
fi

# Set up /u so that Amd can handle mounts for amd.u map:
# No symbolic links allowed
if [[ -L /u ]]; then
    rm /u
    mkdir /u
fi
exit 0
```

startauto.cust Customization Script

The **/etc/auto/cust/startauto.cust** file is an optional customization script that you can provide to start your automounter daemon. If this file exists, and it is executable, it will be invoked by the **services_config** installation command during the system boot process to start your automounter daemon. You should provide this file when you require changes to the SP invocation of the **automount** daemon, or when you wish to replace it with the invocation of a different automounter daemon that you have made available on your system.

The **services_config** process will invoke the automounter configuration, and if that completes successfully, a check will be made to see if **startauto.cust** exists and is executable. If so, it will be invoked to start the automounter daemon. If not, the **automount** daemon will be started using /**etc**/**auto**/**startauto**.

There are no input parameters passed to the **startauto.cust** script. The return value has no affect on the **services_config** execution.

Example

The following script is an example of a user customization script that you might use to start the Amd daemon instead of the AIX **automount** daemon. This script assumes that you have Amd installed on your system and that the executables reside in the /**etc**/**amd** directory.

```
#!/usr/bin/ksh

# Check if Amd daemon is running
if [[ -n ‘ps -ef | grep /etc/amd/amd | grep -v grep’ ]]; then
    echo “startauto.cust:  amd daemon is already running.”
    exit 0
fi

# Build amd input list using all amd.* map files in
# /etc/amd/amd-maps directory
set -A amdmaps $(ls /etc/amd/amd-maps/amd.*)
let i=0
while (( $i < ${#amdmaps[*]} )); do
  amd_parms=“$amd_parms /${amdmaps[$i]##/*/amd.}
${amdmaps[$i]}”
  let i=$i+1
done

# Start the daemon
nice --4 /etc/amd/amd -t 16.120 -x all -l /var/adm/SPlogs/auto/auto.log
$amd_parms
exit $?
```

refauto.cust Customization Script

The /**etc**/**auto**/**cust**/**refauto.cust** file is an optional customization script that you can provide to refresh your automounter daemon. If this file exists, and it is executable, it will be invoked after a file collection update has been made to the /**etc**/**auto**/**maps**/**auto.u** map file for user home directories. You should provide this file when your automounter requires special refresh activity.

If you have file collections configured for your SP system, the **supper update** process on a boot/install server or processor node will update the /**etc**/**auto**/**maps**/**auto.u** map file when changes have been made. If this map file is updated, the automounter refresh is invoked. The refresh process will check to see if the **refauto.cust** customization script exists and is executable. If so, it will be invoked to refresh the automounter daemon. If not, no activity will occur since the **automount** daemon does not require refreshing.

If you are using file collections to distribute your map files and your automounter requires refreshing for updates to map files other than /**etc**/**auto**/**maps**/**auto.u**, you will need to update the list file for the collection. The **user.admin** collection is used

to distribute all map files in the **/etc/auto/maps** directory. If you are also using **user.admin** to distribute your maps, you will need to edit the list file **/var/sysman/sup/user.admin/list** to contain an execute statement similar to the one for **auto.u**:

```
execute /etc/amd/refresh_amd (./etc/auto/maps/auto.u)
```

There are no input parameters passed to the **refauto.cust** script. The return value is noted by the **supper update** process.

Example

The following script is an example of a user customization script that you might use to refresh the Amd daemon. This script assumes that Amd is installed and that the executables reside in the **/etc/amd** directory.

```
#!/usr/bin/ksh

# If amd is running then refresh the maps with amq -f
if [[ -n 'ps -ef | grep /etc/amd/amd | egrep -v grep' ]]; then
  /etc/amd/amq -f
fi
```

checkauto.cust Customization Script

The **/etc/auto/cust/checkauto.cust** file is an optional customization script that you can provide to verify your automounter installation. If this file exists, and it is executable, it will be invoked from the **SYSMAN_test** command to verify your automounter installation. You should provide this file when you have provided your own automounter configuration using **/etc/auto/cust/cfgauto.cust** or started your own automounter daemon with **/etc/auto/cust/startauto.cust**. We recommend that your **checkauto.cust** checks all files and directories required by your automounter and that your automounter daemon is running.

The **SYSMAN_test** process will check if **/etc/auto/cust/checkauto.cust** exists and is executable. If so, it will invoke it. If not, it will invoke the SP automounter verification. This code checks that all required SP automounter directories and files exist and that the SP automounter daemon is running.

The **checkauto.cust** script will be invoked with the same syntax as the **SYSMAN_test** command:

```
checkauto.cust [-q | -v] -l log_file
```

The parameters are those supplied to the **SYSMAN_test** command with the following definitions:

-q Specifies quiet mode; error messages are only appended to the log file and not written to stdout.

-v Specifies verbose mode; all error and informational messages are witten to both the log file and stdout.

-l *logfile* Specifies the path name of the log file to which messages are appended.

If neither **-q** nor **-v** are specified, error messages are written to both the log file and stdout.

The return value from **checkauto.cust** execution should be:

0	Indicates success
non-0	Indicates the number of verification errors that were found

Example

The following script is an example of a user customization script that you might use to verify your Amd configuration and daemon execution. This script assumes that Amd is installed and that the executables reside in the **/etc/amd** directory.

```
#!/usr/bin/ksh

while getopts "qvl:" opt; do
  case $opt in
      q ) QUIET="-q" ;;
      v ) VERBOSE="-v" ;;
      l ) LOGFILE=$OPTARG ;;
  esac
done

let ERRORS=0
# Check for /etc/amd directory
if [[ ! -a /etc/amd ]]; then
    if [[ -z $QUIET ]]; then
        echo "checkauto.cust:  Directory /etc/amd does not exist."
    fi
    echo "checkauto.cust:  Directory /etc/amd does not exist." >> $LOGFILE
    let ERRORS=$ERRORS+1
fi

# Check for /etc/amd/amd-maps dir
if [[ ! -a /etc/amd/amd-maps ]]; then
    if [[ -z $QUIET ]]; then
        echo "checkauto.cust:  Directory /etc/amd/amd-maps does not exist."
    fi
    echo "checkauto.cust:  Directory /etc/amd/amd-maps does not exist." >> $LOGFILE
    let ERRORS=$ERRORS+1
fi

# Check for default amd.u file
if [[ ! -a /etc/amd/amd-maps/amd.u ]]; then
    if [[ -z $QUIET ]]; then
        echo "checkauto.cust:  Map file /etc/amd/amd-maps/amd.u does not exist."
    fi
    echo "checkauto.cust:  Map file /etc/amd/amd-maps/amd.u does not exist." >> $LOGFILE
    let ERRORS=$ERRORS+1
fi

# Check if /u is a symbolic link
if [[ -L /u ]]; then
    if [[ -z $QUIET ]]; then
       echo "checkauto.cust:  /u is a symbolic link.  Not allowed by Amd."
    fi
    echo "checkauto.cust:  /u is a symbolic link.  Not allowed by Amd." >> $LOGFILE
    let ERRORS=$ERRORS+1
fi

# Check if Amd daemon is running
if [[ -z 'ps -ef | grep amd | grep -v grep' ]]; then
    if [[ -z $QUIET ]]; then
        echo "checkauto.cust:  amd daemon is not running."
    fi
    echo "checkauto.cust:  amd daemon is not running." >> $LOGFILE
    let ERRORS=$ERRORS+1
fi
exit $ERRORS
```

mkautoent.cust Customization Script

The **/etc/auto/cust/mkautoent.cust** file is an optional customization script that you can provide to add a user home directory entry to your automounter map file. If this file exists, and it is executable, it will be invoked by the SP User Management commands **spmkuser**, **spchuser**, and **mkamdent** either directly or through SMIT panels. You should provide this file when you require a different map file or map file format than that generated by the SP User Management services for the **/etc/auto/maps/auto.u** map file. If you do provide this file, you will probably also need to provide **/etc/auto/cust/rmautoent.cust** and **/etc/auto/cust/lsautoent.cust**. These files remove and list the user's home directory entry in the map file.

When SP User Management services process user home directory additions or updates, a check is made to see if **mkautoent.cust** exists and is executable. If so, it will be invoked to process the home directory information for your automounter map file. If not, the automount map file **/etc/auto/maps/auto.u** will be updated with the new home directory entry.

The Parallel Operating Environment (POE) component in the AIX Parallel Environment contains the command **mpamddir** which is dependent on the format of the SP supported automount map file. If you have this support installed on your system and are using the command, you may need to provide an additional user exit. See *IBM Parallel Environment for AIX: Operation and Use* for more information.

The **mkautoent.cust** script is invoked with the following syntax:

```
mkautoent.cust user host[,host...] pathname
```

The parameters have the following definitions:

user	The user name that is to be used as a key to the new map entry.
host[,host...]	One or more host names separated by commas indicating the target host where the user's home directory resides.
pathname	Directory pathname of the user's home directory as it is defined on the host. The trailing user subdirectory has been removed from the pathname.

For example, the syntax for a user **jws** with a home directory on **luna** with the path name of **/home/luna/jws** would be:

```
mkautoent.cust jws luna /home/luna
```

The return value from **mkautoent.cust** execution should indicate:

0	Success
non-0	An error occured creating the map file entry. The following actions will occur: • The **mkamdent** command will return this value as its return code. • The **spmkuser** command creates the user before invoking this script, so the return code has no affect on the command. The user is still added.

- The **spchuser** command will terminate and not change the user information.

Example

The following script is an example of a user customization script that you might use to add an entry to your **/etc/amd/amd-maps/amd.u** map file.

```
#!/bin/ksh

# get input parms
USER=$1
HOSTS=$2
HOMEPATH=$3
AMDMAP=/etc/amd/amd-maps/amd.u
HOSTS=‘echo $HOSTS | sed ‘s/,/ /g’’

# Append the entry to the amd /u map
FIRST=“ ”
for HOST in $HOSTS; do
  if [[ -z $FIRST ]]; then
    echo “$USER  host==$HOST;type:=link;fs:=$HOMEPATH \\” >> $AMDMAP
    FIRST=$HOST
  else
    echo “host==$HOST;type:=link;fs:=$HOMEPATH \\” >> $AMDMAP
  fi
done
echo “host!=$FIRST;type:=nfs;rhost:=$FIRST;rfs:=$HOMEPATH” >> $AMDMAP
```

rmautoent.cust Customization Script

The **/etc/auto/cust/rmautoent.cust** file is an optional customization script that you can provide to remove a user home directory entry from your automounter map file. If this file exists, and it is executable, it will be invoked by the SP User Management commands **sprmuser** and **spchuser** either directly or through SMIT panels. You should provide this file when you have provided **/etc/auto/cust/mkautoent.cust** support for a different map file or map file format than that generated by the SP User Management services for the **/etc/auto/maps/auto.u** map file.

When SP User Management services process user home directory removals, a check is made to see if **rmautoent.cust** exists and is executable. If so, it will be invoked to remove the home directory information from your automounter map file. If not, the entry will be removed from the automount map file **/etc/auto/maps/auto.u**.

The **rmautoent.cust** script will be invoked with the following syntax:

```
rmautoent.cust user
```

The parameters have the following definitions:

user The user name that is used as the key of the map entry to be removed.

For example, the syntax for a user **jws** with a home directory on **luna** with the path name of **/home/luna/jws** would be:

```
rmautoent.cust jws
```

The return value from **rmautoent.cust** execution should indicate:

0 Success

non-0 An error occurred removing the map file entry. The following actions will occur:

- The **sprmuser** command removes the user before invoking this script, so the return code has no affect on the command. The user will still be removed.
- The **spchuser** command will terminate and not change the user information

Example

The following script is an example of a user customization script that you might use to remove an entry from your **/etc/amd/amd-maps/amd.u** map file.

```
#!/bin/ksh

# Get input parms
USER=$1
AMDMAP=/etc/amd/amd-maps/amd.u

# Delete the entry to the amd /u map
sed -e"/^$USER /,/[^\\]$/d" $AMDMAP > $AMDMAP.tmp
mv $AMDMAP.tmp $AMDMAP
exit 0
```

lsautoent.cust Customization Script

The **/etc/auto/cust/lsautoent.cust** file is an optional customization script that you can provide to list a user's home directory information stored in your automounter map file. If this file exists, and it is executable, it will be invoked by the SP User Management command **splsuser** either directly or through SMIT panels. You should provide this file when you have provided **/etc/auto/cust/mkautoent.cust** support for a different map file or map file format than that generated by the SP User Management services for the **/etc/auto/maps/auto.u** map file.

The **splsuser** command checks to see if **lsautoent.cust** exists and is executable. If so, it will be invoked to list the home directory information from your automounter map file. If not, the information stored in the automount map file **/etc/auto/maps/auto.u** will be listed.

The **lsautoent.cust** script will be invoked with the following syntax:

```
lsautoent.cust user
```

The parameter has the following definition:

user The user name that is used as the key of the map entry containing the home directory information to be returned.

The user home directory information should be echoed to stdout in the following format:

```
host[,host...] pathname
```

where:

host[,host...]	Lists one or more host names separated by commas indicating the target host where the user's home directory resides.
pathname	Lists the directory path name of the user's home directory as it was provided to the **mkautoent.cust** script with the trailing *user* subdirectory removed from the pathname.

For example, the syntax for a user **jws** with a home directory on **luna** with the path name of /**home**/**luna**/**jws** would be:

```
lsautoent.cust jws
```

and the return value written to stdout would be:

```
luna /home/luna
```

The return value from **lsautoent.cust** execution should indicate:

0	Success
non-0	An error occurred listing the map file entry. The **splsuser** command will list default user home directory information if the output from this script is null.

Example

The following script is an example of a user customization script that you might use to remove an entry from your /**etc**/**amd**/**amd-maps**/**amd.u** map file.

```
#!/bin/ksh
# Get input parms
USER=$1
AMDMAP=/etc/amd/amd-maps/amd.u

# Find the entry in the amd /u map and return the host and directory
hosts=‘cat $AMDMAP | sed “
/^$USER /,/[^\\]$/!d
/host==/!d
s/.*host==\(.[^; ]*\).*$/\1/g”’
hosts=‘echo $hosts | sed “s/\([^ ]*\)[ ][ ]*\(.[^ ]*\)/\1,
\2/”g’
home_dir=‘cat $AMDMAP | sed “
/^$USER /,/[^\\]$/!d
/rfs:=/!d
s/.*rfs:=\([^; ]*\)[^ ]*/\1/g”’

echo $hosts $home_dir

exit 0
```

Coexistence Considerations

If your SP system contains nodes that are running PSSP 2.2 or previous releases and **amd_config** is **true**, those nodes are using the public domain BSD automounter, Amd. In order to support mixed installations containing both these older nodes and newer nodes running PSSP 2.3 and future releases, the SP supports and maintains both Amd and automount directories and map files.

The **amd_config** site environment variable specifies whether you want SP automounter support for the entire SP system. When **amd_config** is **true**, the SP will configure and run the AIX **automount** daemon on the newer nodes containing PSSP 2.3 and future releases, and the BSD Amd daemon on the older nodes containing PSSP 2.2 and previous releases. The configuration process **services_config** is executed during system installation and during a processor reboot. On the control workstation and newer nodes this process will create both **/etc/auto** and **/etc/amd** directory structures. On the older nodes, it will continue to only create the **/etc/amd** directory structure.

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

If file collections are used to distribute map files to the boot/install servers and processor nodes, both the automount map files in **/etc/auto/maps** and the Amd map files in **/etc/amd/amd-maps** will be distributed to all nodes. The older nodes will simply ignore the automount map files and the newer nodes will simply ignore the Amd map files. If you have only newer nodes on your SP system and do not wish to have the Amd map files distributed, you can edit the **list** file for the **user.admin** file collection on the control workstation and remove the entries for the Amd maps. If you have a mixed system, and do not wish to have the Amd maps distributed to the newer nodes or the automount maps distributed to the older nodes, you can edit the **refuse** file for the **user.admin** file collection on the appropriate nodes. See Chapter 5, “Managing File Collections” on page 41 for more information on these file collection control files.

The SP automounter customization support is active only on the control workstation and the newer nodes. Although the **/etc/auto/cust** files may be distributed to the older nodes by file collections, they are ignored there.

Migration Considerations

SP systems containing PSSP 2.2 and previous releases use the public domain BSD automounter, Amd. When migrating one of these systems to PSSP 2.3 or future releases, the SP automounter support replaces the use of Amd with the **automount** daemon that is resident in AIX. Amd is no longer shipped as part of the PSSP release. All existing Amd map files that are to be supported by the new **automount** daemon must be converted.

During a migration install, existing Amd files in the **/etc/amd** directory structure will not be removed or altered. Only the **/etc/amd/refresh_amd** command will be replaced and a copy of the previous version of that command will be saved in the same directory.

The configuration process **services_config** is executed during system installation and during a processor reboot. If SP automounter support is configured (the **amd_config** site environment variable is **true**), this process will create the necessary **/etc/auto** directory structure and default files. If SP User Management services is also configured (the **usermgmt_config** site environment variable is also **true**), this process will create a default map file for user home directories

/etc/auto/maps/auto.u. When creating this map file, if an Amd map file **/etc/amd/amd-maps/amd.u** already exists, its contents will be used to generate the new automount map.

Converting the amd.u Map File

The **services_config** process invokes the installation command **mkautomap** to convert Amd map entries from **/etc/amd/amd-maps/amd.u** to automount map entries, placing the final results in **/etc/auto/ maps/auto.u**. Only map entries previously generated by SP User Management services will be recognized. If you have modified or added any entries in **amd.u**, the results may be unpredictable. If an Amd map entry cannot be properly interpreted, a message will be issued, and that entry will be ignored. Automount map entries will be generated for all recognized entries and will be written to the temporary file **/etc/auto/maps/auto.u.tmp**. If no errors occurred during the **mkautomap** processing, this temporary file will be renamed to **/etc/auto/maps/auto.u**. Also, the automount master map file **/etc/auto.master** will be updated with an entry for the **/u** file system referencing this new map file. Any default mount options that were listed in the **/defaults** entry in the Amd map will be added to the master map file.

If no errors occurred during the **mkautomap** process, **services_config** continues with the automount configuration and then starts the **automount** daemon. However, if any problems were encountered, the automount configuration and startup process halt, and **services_config** continues with its other processing. This allows you to evaluate the map conversion errors and decide how to handle them. All conversion errors are logged in the SP automounter log file **/var/adm/SPlogs/auto/auto.log** in addition to being displayed on the console.

If the entries that were not converted can be safely ignored and are no longer relevant to the operation of your SP system, you can simply rename the temporary file, edit the master map file, and run **services_config** again:

```
mv /etc/automaps/auto.u.tmp /etc/auto/maps/auto.u
echo '/u /etc/auto/maps/auto.u'>> /etc/auto.master
services_config
```

However, if the failing entries are needed, you will have to address each entry individually. You may need to refer to your existing BSD documentation to understand the function the Amd map entry provided. Then you will need to create an equivalent automount entry and add it to the **/etc/auto/maps/auto.u** map file. If the final automount entry follows the standard SP format for user home directories, you can add it to **auto.u** using the **mkamdent** command. Otherwise, you will need to manually edit the map file to add your entry.

See *IBM AIX Version 4 System Management Guide: Communications and Networks* for information on automount map file formats.

Converting Other Amd Map Files

The **services_config** process attempts to convert only your **/etc/amd/amd-maps/amd.u** map file. It is your responsibility to manually convert all other Amd map files that you need automount support for. You may need to refer to your existing BSD documentation to understand the function provided by your Amd map files. You will then need to create equivalent automount maps and add them to the SP support. See "Using Automount for Other File Systems" on page 109 for details on adding the maps to the SP automounter support. *IBM AIX*

Version 4 System Management Guide: Communications and Networks contains information on the AIX automount daemon and map files.

Note: The SP automounter support on the older nodes running PSSP 2.2 and previous releases automatically passes all map files in the **/etc/amd/amd-maps** directory as input to the Amd daemon. This function has changed for the newer nodes. The SP automount support uses the master map file **/etc/auto.master** to tell the **automount** daemon which directories to control. Therefore, you now have an additional step of adding an entry to the master map file for any additional directory you want automount to control. However, your map files can now reside anywhere you choose, and all map file naming restrictions have been removed.

Chapter 10. Managing Print Service

The SP Print Management System is removed from PSSP 2.3. That is, the SP Print Management System cannot be configured on nodes running PSSP 2.3. We recommend 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. The **print_config** routine running on the control workstation will configure the SP Print Management System on nodes running versions of PSSP earlier than PSSP 2.3.

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.

© Copyright IBM Corp. 1995, 1997

Chapter 11. Managing Mail Service

There are several ways you can integrate the SP system into your current AIX mail environment. This section provides instructions for three recommended options. Each option offers a solution that:

1. Avoids the maintenance of mail queues on each SP processor node.
2. Allows the SP users to send mail in a customary manner and receive it at their home directory host.
3. Requires the use of SP automounter support when you use a SP node or control workstation as the mail acceptor host. Refer to Chapter 9, "Managing the Automounter" on page 103 for more information.

These options differ depending on whether you are using DNS, NIS, or **/etc/hosts** for your host name resolution method.

Terminology

The following terms are used in the mail discussion:

Central mail hub	A central host that either accepts mail for all users or routes users' mail to their home directory hosts.
Home directory host	The host where a user's home directory resides and where the user reads mail.
Mail acceptor host	The host that serves as a central mail hub for SP nodes only. It must have an account for each user being serviced.
Mail exchanger	A host that accepts mail addressed to users on other hosts through the use of **MX** records and the **Cw** option in the **sendmail.cf** file.

Note: All procedures in this section pertain only to the RS/6000 architecture. If your environment includes systems with other architectures, the procedures may require modification.

Related Information

Refer to Chapter 5, "Managing File Collections" on page 41 for more information about the SP file collection procedures.

You can find further information about NIS, **sendmail**, and the TCP/IP Domain Name in *IBM AIX Communication Concepts and Procedures*.

You can find further information about the **/etc/aliases** file and the Domain Name Service records in *IBM AIX Version 3.2 for RS/6000 Files Reference*.

Tasks covered in this chapter include:

- Option 1: Integrating mail using DNS
- Option 2: Integrating mail using NIS
- Option 3: Integrating mail using **/etc/hosts**
- Distributing mail-related files

© Copyright IBM Corp. 1995, 1997

Option 1: Integrating Mail Using DNS

If you are using DNS for your external hostname resolution, you can configure a SP node, the control workstation, or a host outside the SP system to act as a mail hub for the processor nodes. You accomplish this by adding Address (**A**) and Mail Exchanger (**MX**) records to your domain data file on your primary name server. These records direct mail addressed to the SP nodes to the mail acceptor host. Mail for the control workstation can be rerouted in the same manner as mail for other SP nodes. As a result, mail sent through the network will be relayed, using **.forward** files, to the users' home directories or a site central mail hub. This alleviates the need to run the **sendmail** daemon and maintain mail queues on each node, however, the mail acceptor host **must have access** to each user's **.forward** file in order to route mail to their home directories.

When a SP node or control workstation is used as a mail acceptor host, SP automounter support is required to mount the user's home directories and access the **.forward** files. When a host outside the SP system is used a mail acceptor host, a **.forward** file must exist in each user's home directory on that host in order for mail to be relayed to its final destination.

You can also use DNS to route the mail to a central mail hub, although the use of a mail hub can have drawbacks as explained in each step.

Figure 8 on page 129 shows how LAN and local mail are routed in a DNS configuration, assuming SP automounter support is running.

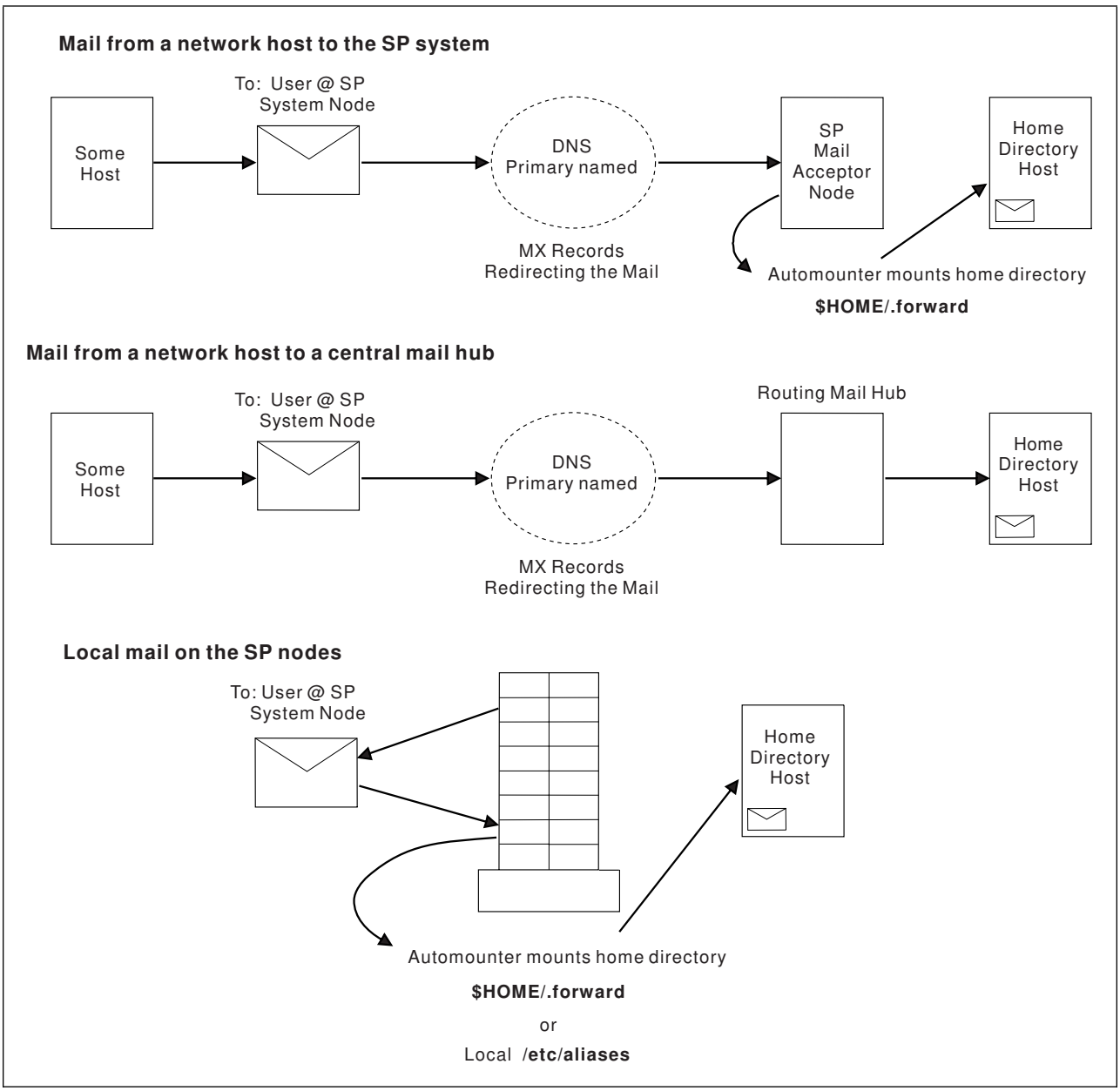


Figure 8. Option 1: Using DNS

Steps in this task:

Step 1. Modify the primary domain data file
Step 2. Modify **sendmail.cf**
Step 3. Add **.forward** files to home directories
Step 4. Turn off **sendmail** on the nodes

Step 1: Modify the Primary Domain Data File

Add Address (**A**) records and Mail Exchanger (**MX**) records to your primary domain data file. The Address records specify the address of the host serving as the mail acceptor. The Mail Exchanger records identify the control workstation and nodes whose mail is being forwarded, and the hosts to which the mail should be sent. For example, if SP node 1 serves as the mail acceptor host, then node 1 would receive all SP mail and forward it to the appropriate user home directories.

Note that you need to add **MX** records for the SP processor nodes and mail acceptor host to your **named.***domain***data** file.

Your primary domain data file resides on your primary DNS name server. The file is usually called **named.***domain***data**, where *domain* is the name of your DNS domain. For example, if your domain is called **rx7**, your domain data file would be **named.rx7data**.

Add the Records Manually

Follow this process to edit your primary domain data file and manually add the **A** and **MX** records:

1. Log in to your primary name server and edit the domain data file.
2. Add an **MX** record for each mail acceptor host and processor node, using this format:

   ```
   hostname    IN   MX    10  mail_exchanger
   ```

   where

   hostname The name of the host whose mail is being forwarded.
   mail_exchanger The name of the host to which the mail is redirected.

   All processor nodes use their mail acceptor host as their mail exchanger, and the mail acceptor node serves as its own mail exchanger. The following example shows node 1 serving as the mail acceptor host:

   ```
   spnode3.kgn.ibm.com.  IN     MX  10    spnode1.kgn.ibm.com.
   spnode2.kgn.ibm.com.  IN     MX  10    spnode1.kgn.ibm.com.
   spnode1.kgn.ibm.com.  IN     MX  10    spnode1.kgn.ibm.com.
   ```

   If you are using a central mail hub, use the mail hub host name as the mail exchanger for all nodes.
3. Add an **A** record for each mail acceptor host or mail hub (if one does not already exist) using the following format:

   ```
   hostname  IN   A     nnn.nnn.nnn.nnn
   ```

   For example:

   ```
   spnode1  IN   A    123.456.678.90
   ```

Step 2: Modify sendmail.cf

Change the /**etc**/**sendmail.cf** file in several locations, depending on your current mail configuration:

- **If you are forwarding all mail from the nodes to a mail acceptor host:**

  Add the **Cw** option at the proper location in the **sendmail.cf** file on the mail acceptor host and include the nodes it serves as aliases for that host. Follow

the format **Cw***MailacceptorHostname Node1Hostname Node2Hostname* as shown in this example where `spr1node1` serves four nodes.

```
Cwspr1node1 r1node2 r1node3 r1node4 r1node5
```

Maintenance Reminder

Whenever you expand the SP system to add new nodes, you must also add these nodes to the **sendmail.cf** file on their mail acceptor host. If file collection is used to manage the **sendmail.cf** file, note that the **sendmail.cf** file on the mail acceptor host is different from that on the other nodes in the SP system.

- **If you are forwarding all mail from the nodes to a central mail hub that routes mail:**

  Add the **Cw** option at the proper location in **sendmail.cf** on every host that will receive mail from the mail hub and include all the nodes in the SP system as aliases. If **sendmail.cf** does not include all of the processor nodes, sending a message can cause a loop. When someone sends a message to a user on the SP system, the message is intercepted and sent to the mail hub. The mail hub, in turn, sends the message to *user@homehost*. If the **sendmail** daemon on the home host does not accept the message, it forwards it to the SP system and starts the cycle again.

  Maintenance Reminder

  Whenever you expand the SP system and add new nodes, you must also add these nodes to the **sendmail.cf** file on every host that receives mail from the mail hub.

  If you have a large user base where people maintain their home directories on personal systems, this configuration may be difficult to administer and control. This might be a reason not to use a routing mail hub.

- **If you are forwarding all mail to a central mail hub that accepts mail for other hosts:**

  Add the **Cw** option at the proper location in **sendmail.cf** on the mail hub host and include the names of all the SP processor nodes and control workstation.

  Maintenance Reminder

  Whenever you expand the SP system and add new nodes, you must also add these nodes to the **sendmail.cf** file on this mail hub.

- **In all cases:**

  Enable the **OK MX** entry in **sendmail.cf** on every host in the network to allow the Mail Exchanger records to work. The default **sendmail.cf** comments this entry. Remove the comment as shown in this example:

  ```
  # The default is not to use a name server for resolving recipients.
  OK MX
  #
  ```

  Whenever you change **sendmail.cf**, you need to refresh the **sendmail** database on every host in your network, including the SP system. To do this, enter:

  ```
  /usr/lib/sendmail -bi
  /usr/lib/sendmail -bz
  ```

  Restart the **sendmail** daemon, if it is running.

Refer to “Distributing Mail-Related Files” on page 139 for suggestions on distributing the **sendmail.cf** file.

Step 3: Add .forward Files to Home Directories

Add a **.forward** file to each SP user's home directory to provide a destination for forwarding local mail on the processor nodes. When a node receives local mail, SP automounter support mounts the user's home directory. The **.forward** file redirects the mail to the user at the host name specified in that file, which is the user's home directory host.

Add a **.forward** file to the user's home directory on the mail acceptor host if this host is outside the SP system.

Step 4: Turn Off sendmail on the non-Mail Acceptor Nodes

In this configuration, the processor nodes do not need to receive incoming mail. You can save CPU cycles by turning off the **sendmail** daemon on the nodes. Even though the **sendmail** daemon is not running, it will be invoked whenever users send mail from any processor node.

Perform the following actions once on each processor node that does not serve as a mail acceptor host.

- Edit the **/etc/rc.tcpip** file and add a comment to the line that starts the **sendmail** daemon:

```
#
#start /usr/lib/sendmail "$src_running" "-bd -q-qpi"
#
```

- Stop **sendmail** on each node using one of these methods:
  - If your site runs the System Resource Controller **srcmstr** daemon, issue:

```
stopsrc -s sendmail
```

  - If your site does not run the System Resource Controller **srcmstr** daemon, issue:

```
kill -1 `cat /etc/sendmail.pid`
```

Option 2: Integrating Mail Using NIS

If you are using NIS for external host name resolution, you can use the NIS global **mail.aliases** map to route the mail to the users' home directories. A global **/etc/aliases** file on the NIS master provides mail addresses for all users in the NIS domain. Local **/etc/aliases** files on the processor nodes contain only a minimum set of local aliases and a reference to the global file on the NIS master.

You have different steps to follow depending on whether you created a new NIS domain specifically for the SP system or already had an established NIS domain in your environment.

If you have:	Procedure:
A new NIS domain for the SP	Perform Steps 1 through 5 using the control workstation as the NIS master
An established NIS domain that does not use **mail.aliases**	Perform Step 1 to create a partial **/etc/aliases** file of the SP users Add other users to this file Proceed with Steps 2 through 5
An established NIS domain that uses **mail.aliases**	Verify that all the SP users are in the **/etc/aliases** on your NIS master Proceed with Steps 3 through 5

Steps in this task:

Step 1. Modify **/etc/aliases** on the NIS master
Step 2. Create a global NIS map
Step 3. Create and distribute local **/etc/aliases**
Step 4. Modify **sendmail.cf**
Step 5. Turn off **sendmail** on the nodes

Figure 9 on page 134 shows how LAN and local mail are routed in an NIS configuration, assuming SP automounter support is running.

Step 1: Modify /etc/aliases on the NIS Master

The **/etc/aliases** file is part of the AIX Base Operating System. It is used to define system-wide aliases for user names or groups of users. When used with NIS, it propagates the NIS **aliases.map** from the NIS master to all clients, making aliases known throughout the NIS domain.

You must append alias records for all the SP users to the existing **/etc/aliases** file on your NIS master to create a global version of this file. You then use this global file to build an NIS map for mail aliases.

Add users to your **/etc/aliases** file by creating entries in the format *username: username@host*, where *host* specifies the node location of the user home directory being mounted to the SP nodes.

Here are sample entries in a **/etc/aliases** file:

```
fvtst1:             fvtst1@tserv5.hpssl.kgn.ibm.com
fvtst2:             fvtst2@tserv5.hpssl.kgn.ibm.com
fvtst3:             fvtst3@tserv5.hpssl.kgn.ibm.com
fvtst4:             fvtst4@tserv5.hpssl.kgn.ibm.com
saffycat:           saffycat@mroz.kgn.ibm.com
ken:                ken@tserv5.hpssl.kgn.ibm.com
wendy2:             wendy2@tserv11.hpssl.kgn.ibm.com
pepper:             pepper@mroz.kgn.ibm.com
```

If you are routing all users through a mail hub, replace the home directory hosts with the fully qualified host name of the mail hub.

Maintenance Reminder

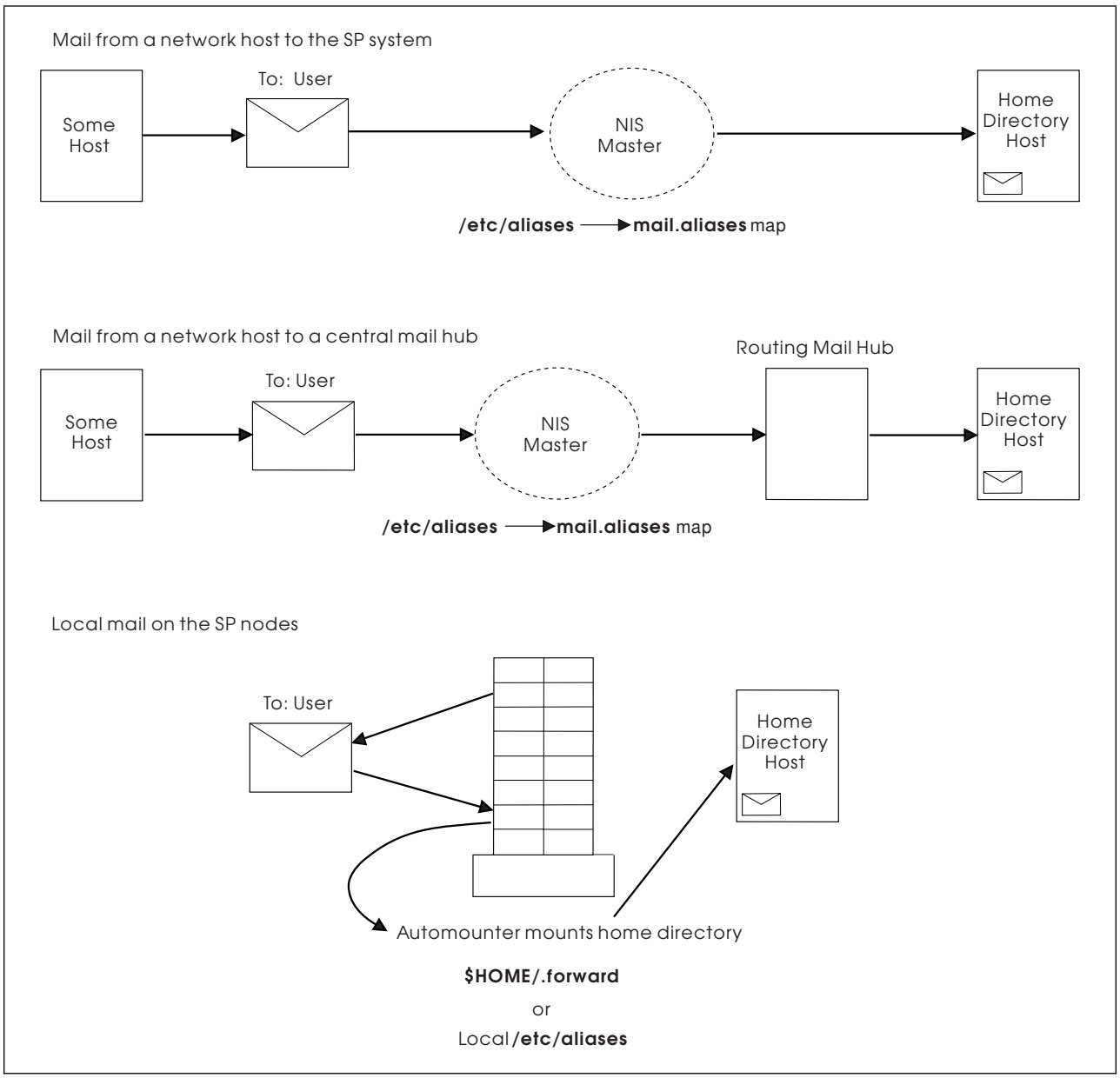


Figure 9. Option 2: Using NIS

Whenever you add a new user to your system, you must also add them to the master **/etc/aliases** file, distribute it, and refresh **sendmail** on every host.

In order for the mail services to resolve host addresses correctly, the **/etc/hosts** file must be kept up to date. Whenever you add or remove a host from your environment, you must also update this file and distribute it.

Step 2: Create a Global NIS Map

Create a global NIS map from the **/etc/aliases** file on the NIS master. When NIS is running, this map augments the **/etc/aliases** file by using the global **mail.aliases** file to provide the mail information for all the users in the NIS environment. If you did not have an established NIS domain before installing the SP system, the control workstation should be set up as your NIS master.

1. Log in to your NIS master server as **root**.

2. Enter:

```
cd /etc/yp
make PUSH=1 aliases
```

This creates the global map and propagates it to all the slaves and clients in the NIS domain.

Step 3: Modify and Distribute Local /etc/aliases

Modify the **/etc/aliases** file on every host in the NIS environment except the NIS master. Local **/etc/aliases** files must not have any entries for users specified in the global file or the local entries will override the global map. If necessary, you can maintain aliases in the local file for users that are not known globally. Each local **/etc/aliases** file requires an entry (NIS escape sequence) at the end that tells NIS to reference the global **mail.aliases** map. Follow this procedure:

1. Edit the **/etc/aliases** file on each node.
2. Remove all entries for users that are already specified in the global **mail.aliases** map.
3. Add the NIS escape sequence (+:) at the end of the file. This directs NIS to use the global **mail.aliases** map.

Step 4: Modify sendmail.cf

If you have a new NIS environment or an established NIS environment that does not use **mail.aliases**, make the following changes to the **sendmail.cf** on every AIX host in your NIS environment.

Or, if you have an established NIS environment and are using **mail.aliases**, make the following changes to the **sendmail.cf** file on the control workstation and processor nodes only.

1. Edit **/etc/sendmail.cf**.
2. Remove the comment on the option for using NIS mail aliases, as shown in this example:

```
# map name for NIS aliases
Opmail.aliases
#
```

3. Remove the comment on the option for future mail delivery. This allows the SP system to attempt to deliver mail at a later time if the NIS master is down. See the following example.

```
# If the I option is set, sendmail treats "connection refused" and "timed out"
# errors from the name server as non-fatal, leaving the message in the queue
# for future delivery attempts.
OI
#
```

Step 5: Turn Off sendmail on the Nodes

In this configuration, the processor nodes do not need to poll for incoming mail. You can save CPU cycles by turning off the **sendmail** daemon on the nodes. Refer to “Step 4: Turn Off sendmail on the non-Mail Acceptor Nodes” on page 132 for instructions on how to do this.

Option 3: Integrating Mail Using /etc/hosts

If you are using **/etc/hosts** for host name resolution, you can either use a local **/etc/aliases** file or **.forward** files to route mail to the users' home directories or a central mail hub.

Each RS/6000 host has a standard **/etc/aliases** file as part of the AIX Base Operating System. You can maintain this file as a master file of mail aliases, directing all network mail to users' home directory hosts or to a central mail hub, and distribute it to all the hosts in your network. You can use the file collection technology to maintain and distribute **/etc/aliases** throughout the SP system and use your standard method of file distribution for the other hosts in your network.

The other choice using the **/etc/hosts** file for host name resolution is to have a **.forward** file in every user's home directory pointing to the home directory host or a central mail hub. When mail is received, SP automounter support mounts the home directory on the SP system and the **.forward** file routes the mail to the specified location. This method requires each processor node to run the **sendmail** daemon which could impact their efficiency and requires more administrative maintenance.

Figure 10 on page 137 shows how LAN and local mail are routed in an **/etc/hosts** configuration, assuming SP automounter support is running.

If you want to use:	Perform:
local**/etc/aliases** files	Steps 1 through 3
.forward files	Steps 4 and 5

Steps in this task:

Step 1. Modify **/etc/aliases** on every host
Step 2. Update the **sendmail** database on every host
Step 3. Turn off **sendmail** on the nodes
Step 4. Add **.forward** files to home directories
Step 5. Run **sendmail** on each processor node

Step 1: Modify /etc/aliases on Every Host

Each RS/6000 host has a standard **/etc/aliases** file as part of the AIX Base Operating System. You can maintain this file as a master file of mail aliases, directing all network mail to users' home directory hosts or to a central mail hub, and distribute it to all the hosts in your network. You can use the file collection technology to maintain and distribute **/etc/aliases** throughout the SP system and use your standard method of file distribution for the other hosts in your network.

Add users to your **/etc/aliases** file by creating entries in the format *username: username@host*, where *host* specifies the node location of the user home directory being mounted to the SP nodes.

Here are sample entries in a **/etc/aliases** file:

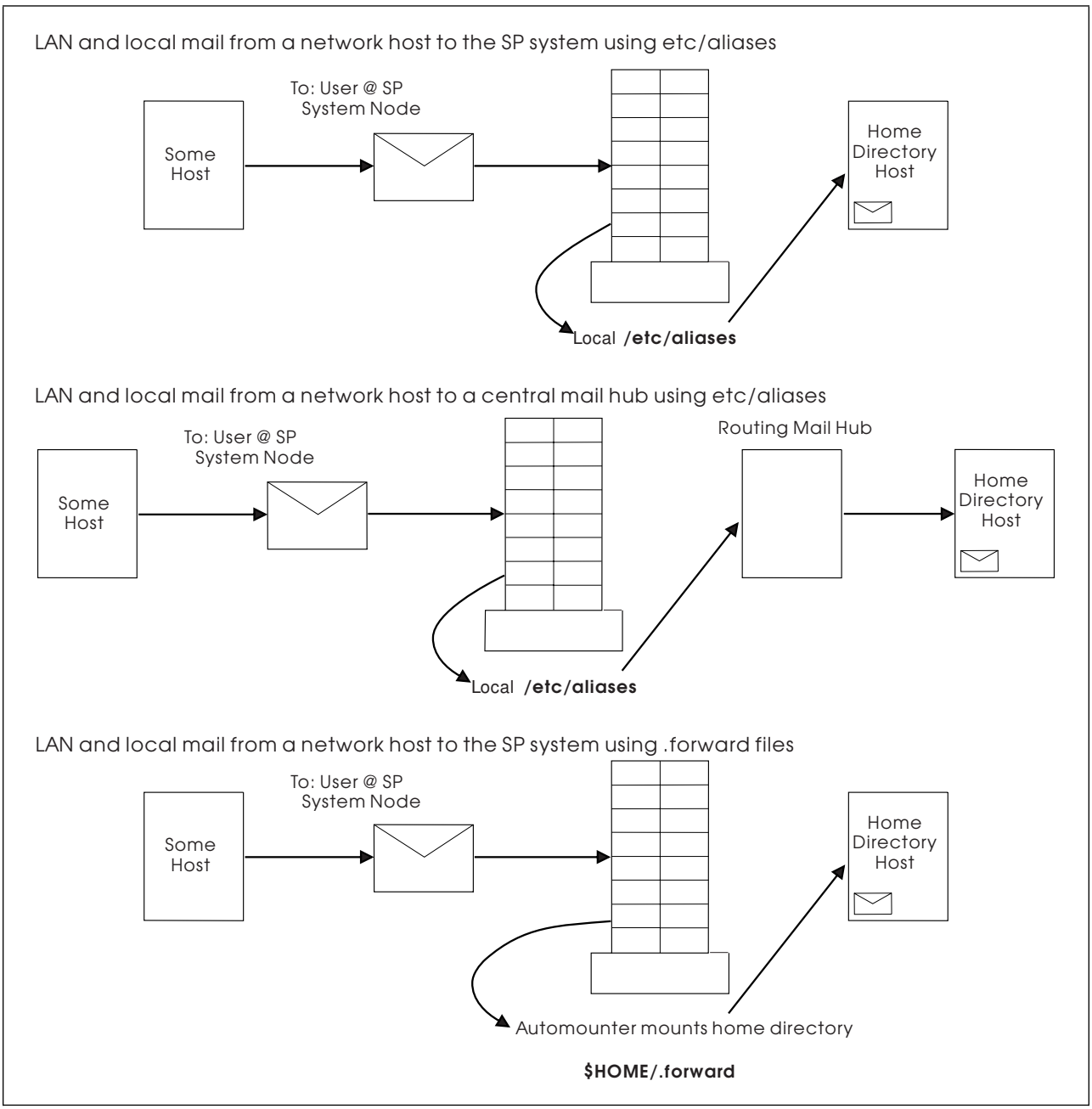


Figure 10. Option 3: Using /etc/hosts

```
fvtst1:          fvtst1@tserv5.hpssl.kgn.ibm.com
fvtst2:          fvtst2@tserv5.hpssl.kgn.ibm.com
fvtst3:          fvtst3@tserv5.hpssl.kgn.ibm.com
fvtst4:          fvtst4@tserv5.hpssl.kgn.ibm.com
saffycat:        saffycat@mroz.kgn.ibm.com
ken:             ken@tserv5.hpssl.kgn.ibm.com
wendy2:          wendy2@tserv11.hpssl.kgn.ibm.com
pepper:          pepper@mroz.kgn.ibm.com
```

Step 2: Update the sendmail Database on Every Host

AIX uses the **/etc/aliases** file to build the **sendmail** database. Whenever you change **/etc/aliases**, you need to refresh the **sendmail** database on every host in your network, including the SP system. To do this, enter:

```
/usr/lib/sendmail -bi
/usr/lib/sendmail -bz
```

Restart the **sendmail** daemon, if it is running.

Maintenance Reminder

Whenever you add a new user to your system, you must also add them to the master **/etc/aliases** file, distribute it, and refresh **sendmail** on every host.

In order for the mail services to resolve host addresses correctly, the **/etc/hosts** file must be kept up to date. Whenever you add or remove a host from your environment, you must also update this file and distribute it.

Step 3: Turn Off sendmail on the Nodes

In this configuration, the processor nodes do not need to poll for incoming mail. You can save CPU cycles by turning off the **sendmail** daemon on the nodes. Refer to "Step 4: Turn Off sendmail on the non-Mail Acceptor Nodes" on page 132 for instruction to do this.

If the SP users do not have home directories on the boot/install servers, these servers may not have incoming mail. In this case, you may also want to turn off the **sendmail** daemon on each of these servers.

Step 4: Add .forward Files to Home Directories

Request that each user adds a **.forward** file to their home directory pointing either to their home directory host or to a central mail hub. The file needs just one entry, as shown in this example:

```
marylou@hostofchoice
```

Step 5: Run sendmail on Each Processor Node

When you are using the **.forward** file configuration, the processor nodes need to poll for incoming mail and must run the **sendmail** daemon. The command to start **sendmail** is in the **rc.tcpip** file.

Most environments usually run the **sendmail** daemon by default. If this is not the case in your environment, follow these instructions to enable this daemon:

1. Edit **/etc/rc.tcpip** on the processor nodes.
2. Remove the comment from the following line:

   ```
   #
   start /usr/lib/sendmail "$src_running" "-bd -q{qpi}"
   #
   ```

3. If the **sendmail** daemon is running, refresh it by entering:

   ```
   /usr/lib/sendmail -bi
   /usr/lib/sendmail -bz
   ```

4. Reboot the nodes.

Distributing Mail-Related Files

To ease the administrative burden of maintaining mail services across the system, you can place mail-related files in a file collection for distribution to the processor nodes. The nodes must all have the same configuration for this to work. Be sure to review Chapter 5 and refer to the directions in "Adding and Deleting Files in a File Collection" on page 54.

sendmail.cf

To use DNS (option 1) to integrate the SP into your mail environment, you must modify the **/etc/sendmail.cf** file. Mail acceptor hosts have different **sendmail.cf** files from other nodes. Ensure that your file collection technology manages both kinds of **sendmail.cf** files.

To manage the **/etc/sendmail.cf** file using file collection technology:

1. Add **sendmail.cf** to the **node.root** file collection.

   ```
   cp /etc/sendmail.cf /share/power/system/3.2.3/etc/sendmail.cf
   ```

2. Add a command to refresh the **sendmail** subsystem to the **list** file for that collection.

   - If your site runs the System Resource Controller **srcmstr** daemon, add:

     ```
     execute /bin/refresh -s sendmail (./etc/sendmail.cf)
     ```

   - If your site does not run the System Resource Controller **srcmstr** daemon, add:

     ```
     execute /usr/lib/sendmail -bz (./etc/sendmail.cf)
     ```

/etc/aliases

To use **/etc/host** (option 3) to integrate the SP into your mail environment, you must modify the **/etc/aliases** file on each node. You can also add this file to the **node.root** file collection:

1. Add **aliases** to the **node.root** file collection.

   ```
   cp /etc/aliases /share/power/system/3.2.3/etc/aliases
   ```

2. Add commands to refresh the **sendmail** subsystem to the **list** file for that collection.

   - If your site runs the System Resource Controller **srcmstr** daemon, add:

     ```
     execute /usr/lib/sendmail -bi (./etc/aliases)
     execute /bin/refresh -s sendmail (./etc/aliases)
     ```

   - If your site does not run the System Resource Controller **srcmstr** daemon, add:

     ```
     execute /usr/lib/sendmail -bi (./etc/aliases)
     execute /usr/lib/sendmail -bz (./etc/aliases)
     ```

3. Restart the **sendmail** daemon if it is running on the nodes.

Chapter 12. Accounting

Your hardware resources are likely to be distributed unevenly across your SP System, with some nodes equipped for greater processing capacity and I/O than others. For this reason you may want to charge different usage fees for different nodes. You may also want to consolidate all accounting records, for the entire SP system, or for groups of nodes.

One way to make efficient use of the nodes in your SP system is to logically partition them so that each pool supports a different type of workload and is considered a separate management domain. The pooling of nodes according to configuration or workload type is the concept behind the *accounting class*. SP accounting allows you to logically groups SP nodes with similar operating characteristics into an accounting class for which the same charge fee applies.

You may also want to charge a different usage fee for compute-intensive jobs when they demand exclusive use of one or more nodes. SP System accounting provides a mechanism for treating these jobs separately as well.

Accounting by Node or by Class

An accounting class is a group of nodes for which accounting data is merged together, providing a single report for the nodes in that class. Individual nodes within a class may be enabled or disabled for accounting.

The **acct_class_id**, an attribute of each node object in the System Data Repository, is an arbitrary string you define and specify for all nodes for which you wish to group and merge accounting data.

Default values eliminate the need to define each attribute for each node. The default value of **acct_class_id** for each node is **default** Therefore, all nodes initially belong to one accounting class called **default**.

The initial value of the **acct_enable** attribute for each node is also **default**. All nodes with **acct_enable=default** are enabled or disabled for accounting based on a system-wide accounting enabled attribute (**spacct_enable**) you define for your site environment. When you set **spacct_enable** to **true**, it enables accounting on all nodes for which the individual accounting enabled attribute (**acct_enable**) is set to **default**. Likewise, when you set **spacct_enable** to **false** (the initial value at installation), accounting is disabled on all nodes that have **acct_enable=default**. Thus, accounting can be enabled or disabled for the entire SP system with individual nodes treated as exceptions.

How SP Accounting Differs from Standard AIX

Standard AIX accounting uses a script called **runacct**. This script is normally run every night by **cron** to provide daily usage reports. It also keeps a running total of usage over some fiscal period (usually a month). SP accounting splits the **runacct** function into two parts. The first part is a script that is run on each SP node every night, just as **runacct** does. This script is named **nrunacct**. The second part is a script that executes on an **accounting_master** node each night, after the **nrunacct** scripts are run. This script is named **crunacct** and it consolidates the **nrunacct** information from each node. The scripts **cmonacct** and **cprdaily** are SP versions

© Copyright IBM Corp. 1995, 1997

of standard AIX programs, modified to handle accounting class and exclusive use accounting.

nrunacct

The **nrunacct** script provides the function that merges raw data from the process, login, print, fee and disk subsystems. It also provides new support to account for the exclusive use of a node by a Resource Manager request.

nrunacct also creates an SP version of the **loginlog** file using the **var/adm/acct/sum/loginlog** files for the specified accounting cycle. Only the most recent **loginlog***yyyymmdd* file from each node is pertinent.

nrunacct provides a user exit for any site-specific accounting procedures in the **/var/adm/nsiteacct** file.

crunacct

The **crunacct** script provides the function that produces daily summary reports and accumulates data for the fiscal period, using the merged data from each node.

SP accounting merges total accounting data by user ID and login name, just as the standard AIX accounting does. This means that all SP nodes must have identical user IDs and login names. In other words, all SP nodes must be members of the same user name space, whether it is implemented through NIS, the SP User Management tools, or through other means.

cmonacct

An SP version of **monacct**, **cmonacct** creates summary files for monthly or fiscal reports.

cprdaily

An SP version of **prdaily**, **cprdaily** creates daily report files of accounting activity and cleans up files no longer needed.

Accounting for Exclusive Use of Node Jobs

While SP nodes are expected to execute exclusive parallel jobs, nonexclusive parallel jobs, and nonparallel jobs during a single accounting cycle, normal operation often results in the exclusive use of a node when executing a compute-intensive parallel job. A single job that makes such great demands on the compute resources of a node where no other jobs can share that node's resources is said to be an *exclusive use of a node* job. Such jobs can be handled separately for accounting purposes.

Standard AIX accounting produces a record for each process as the process completes. These records are later merged by user ID and login name. A user who executes one or more types of jobs on the same node has all process accounting records for that node merged together, resulting in the loss of those processes that had exclusive use of the node. It is important to track the exclusive use of a node by a user because the cost is likely to be different than the cost of shared use.

SP accounting addresses exclusive-use jobs as follows: Before starting a job that has exclusive use of a node, the Resource Manager executes a start job executable program under the real user ID of the job. When the job is finished the

Resource Manager executes an end job command, again under the real user ID of the job. These commands consist solely of an **exit()** system call in their **main()** procedure. Their purpose is to generate process accounting records for the user ID of the job that can then be used to *bracket* all of the normal process accounting records of the job. In other words, all of the process accounting records generated by an exclusive job can be identified and excluded from the other process accounting records generated on the node.

The process accounting records for the start job and end job programs are used to calculate the time the job had exclusive use of the node. The time is then charged to the user ID of the job by converting it to chargefee units. Use of the chargefee mechanism eliminates the need to generate additional accounting reports.

Accounting records for the start job and end job programs are identified by the names of the commands as found in the records. These start and stop programs are given the nonstandard names of **SJSJSJSJ** and **EJEJEJEJ** to avoid conflict with any other standard job names.

An attribute of the SP object in the SDR, **spacct_excluse_enable**, determines whether the start job program and end job program are executed for exclusive-use jobs. A node-level SDR attribute, **acct_excluse_enable**, determines whether accounting is activated to process start and end job records, and to generate chargefee records for jobs making exclusive use of that node.

Setting Up SP Accounting

IBM General Concepts and Procedures for RS/6000, GC23-2202, describes a procedure for setting up an accounting system that involves performing the following steps for each AIX system:

- Creating directories
- Creating and updating files
- Enabling accounting
- Adding accounting procedures to **crontabs**

In a parallel environment, where there may be hundreds of nodes, it is essential that these tasks be automated.

The following procedure helps you to tailor your accounting system by defining accounting class identifiers and accounting enabled attributes on a SP level or on an individual node basis, if necessary. Accounting class identifiers need only be specified when you want more than one accounting class. You can also specify which nodes are to have accounting enabled, however, if all nodes are to be enabled, then only SP Accounting Enabled need be set to `true` and the **acct_enable** attribute can be left to default on each Node object in the SDR.

In addition, this procedure shows you how to specify exclusive use accounting, and how to specify an Accounting Master other than the default control workstation. It also shows how to set an SP Accounting Active Node Threshold value other than the default value of 80.

You can define or change these values after installation by using SMIT or the **spacctnd** and **spsitenv** commands.

Accounting and System Partitioning

If you do not partition your SP system or if you run with a single system partition, you can assign any node or the control workstation as the accounting master.

If you configure multiple system partitions on your SP system the following restrictions apply:

- If all of your system partitions are IBM Parallel System Support Programs for AIX (PSSP) Version 2 Release 1 (running on AIX 4.1) system partitions, you can assign any node in any system partition or the control workstation as the accounting master.
- If you have configured a single PSSP Version 1 Release 2 (running on AIX 3.2.5) system partition, the accounting master can be any node in that PSSP V1R2 system partition or the control workstation. Under this configuration you cannot select a node in a PSSP V2R1 system partition to be the accounting master.
- If you have configured two or more PSSP V1R2 (running on AIX 3.2.5) system partitions, the accounting master must be the control workstation.

Step 1: Enter System Accounting Information

Use SMIT or the **spacctnd** and **spsitenv** commands to define how accounting is to be done on your SP System. This information is read by the installation scripts, which enter specific details about your site's configuration into the System Data Repository (SDR).

Note

If you are setting up accounting and modifying Node objects in the SDR on a SP System that is already running the Resource Manager, you must stop and restart the Resource Manager.

If you are only changing the SP object in the SDR on a SP System that is already running Resource Manager (to enable exclusive use accounting, for example), then you must reconfigure the Resource Manager using the **jm_config** command.

Refer to “Reconfiguring the Resource Manager” on page 94.

Using SMIT:

TYPE	**smit enter_data**
	▲ The Enter Database Information Window appears.
SELECT	Site Environment Information
	▲ The Site Environment Information menu appears.
TYPE	The data in the fields, as needed.

Using spsitenv:

The following example enables accounting as the default for all nodes and specifies 2 as the accounting master. It also enables exclusive use accounting.

```
spsitenv acct_master=2 spacct_enable=true spacct_excluse_enable=true
```

Accounting information on this panel includes:

SP Accounting Enabled
: Use this field to specify whether by default, accounting is enabled or disabled on all nodes that have an accounting enabled attribute of **default**. A value of **true** indicates that accounting is enabled on all nodes with an accounting enabled attribute of **default**. A value of **false** indicates that accounting is disabled on all nodes that have an accounting enabled attribute of **default**. The default value for this attribute is **false**.

  This sets the **spacct_enable** attribute in the SP object of the SDR. You can specify this attribute by using the **spsitenv** command instead of SMIT.

SP Accounting Active Node Threshold
: Use this field to specify the minimum percentage of nodes for which accounting data must be present. At the time when the new **crunacct** script runs, data must be available from at least this percentage of nodes in order for processing to continue. The default value for this attribute is **80**.

  This sets the **spacct_actnode_thresh** attribute in the SP object of the SDR. You can specify this attribute when using the **spsitenv** command instead of SMIT.

SP Exclusive Use Accounting Enabled
: Use this field to specify whether to activate start and end job records, and thus chargefee records for jobs requesting the exclusive use of a node. The default value for this attribute is **false**.

  This field sets the **spacct_excluse_enable** attribute in the SP object of the SDR. You can specify this attribute when using the **spsitenv** command instead of SMIT.

  A value of **true** specifies that exclusive use accounting is enabled on all nodes where the **acct_excluse_enable** attribute is set to `true`. You must check each node's **acct_excluse_enable** attribute for account processing. A value of **false** (the default) specifies that exclusive use accounting is disabled on the node.

Accounting Master
: Use this field to specify which node is to act as the accounting master. This is the node that executes **crunacct**. The default value for this attribute is **0** (the control workstation) but you can specify any SP node.

  This sets the **acct_master** attribute in the SP object of the SDR. You can specify this attribute when using the **spsitenv** command instead of SMIT.

Step 2: Enter Node Accounting Information

Use SMIT or the **spacctnd** command to further define SP accounting at the node level.

Using SMIT:

TYPE **smit node_data**

▲ The Node Database Information menu appears.

SELECT Accounting Information

▲ The Accounting Information menu appears.

TYPE The data in the fields, as needed.

Accounting information on this panel includes:

Start Frame
: Enter the frame number that corresponds to the nodes required for this operation. Specify a value between 1 and 64 inclusive.

Start Slot
: Enter the slot number of first node to be used for this operation. Specify a value between 1 and 16 inclusive. Remember that a wide node counts as two slots.

Number of Nodes
: Number of nodes to be used for this operation. The node information is added for successive nodes within a frame. If the count of nodes causes the nodes in a frame to be exhausted the operation continues in the next sequential frame. Specify a value between 1 and 1024 inclusive.

Accounting Class Identifier
: Use this field to change the accounting class. All nodes with the same string value constitute a class for purposes of grouping and merging accounting data. All nodes will have an initial accounting class identifier of **default**.

  This sets the **acct_class_id** attribute in the Node object of the SDR. You can specify this attribute when using the **spacctnd** command instead of SMIT.

  You only need to specify a value for this field when you want to change the accounting class.

Accounting Enabled
: Use this field to indicate whether accounting is enabled or disabled for a node. A value of **true** indicates that accounting is enabled for the node. A value of **false** indicates that accounting is disabled for the node. All nodes have an initial value of **default** assigned to this attribute. A value of **default** indicates that the node is enabled for accounting based on your entry in the SP Accounting Enabled field in the SMIT Site Environment panel in Step 1.

  This sets the **acct_enable** attribute in the Node object of the SDR. You can specify this attribute when using the **spacctnd** command instead of SMIT.

Exclusive Use Accounting Enabled

Use this type field to indicate whether accounting start and end job records, and thus chargefee records, are processed for jobs having exclusive use of the node. A value of **true** indicates that exclusive use accounting is enabled and start and end job records will be generated. A value of **false** indicates that exclusive use accounting is not enabled and start and end job records will not be generated. All nodes have an initial value of **false** assigned to this attribute.

This sets the **acct_excluse_enable** attribute in the Node object of the SDR. You can specify this attribute when using the **spacctnd** command instead of SMIT.

Accounting Job Charge Value

This value is used to determine the number of charge fee units to charge a user for exclusive use of the node and its value is in units of *seconds per charge fee unit*. All nodes have an initial value of **1.0** assigned to this attribute.

This sets the **acct_job_charge** attribute in the Node object of the SDR. You can specify this attribute when using the **spacctnd** command instead of SMIT.

Using spacctnd::

The following example adds accounting information for a system with one frame and 16 nodes. Accounting and exclusive use accounting is to be enabled for each node and 60 seconds of exclusive use by a user is to constitute one charge fee unit.

```
spacctnd -e true -j 60.0 -x true 1 1 16
```

Additional Accounting Setup Tasks

Depending on your environment, you may also wish to perform these tasks:

- Defining the file systems for which disk accounting is to be done.
- Defining the print queues for which printer accounting is to be done.
- Defining how the all print jobs owned by the print manager id (**prtid**) in Secure Mode are to be handled for accounting purposes.
- Disk accounting may be enabled for each file system as it is created. Alternatively, if many file systems already exist, then disk accounting may be turned on or off by executing the **chfs -t'yes'** <**filesystem**> command on the nodes.

Accounting Installation

The accounting installation phase is performed automatically when you select the Accounting Support option in **ssp.sysman** during the **installp** process. (Refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.) This accounting installation process is also performed when you update accounting using **spsitenv**.

The accounting installation process:

1. Creates all necessary directories and files on the nodes for which accounting is enabled and on the **acct_master** node.

2. Adds the **startup** command to the **/etc/rc** file on the nodes for which accounting is enabled.
3. Creates the **/var/adm/acct/nite/jobcharge** file on each node for which accounting is enabled. This file contains the job charge value previously defined for the node.
4. The **holidays** file (**/etc/acct/holidays**) is placed in the **user.admin** file collection and is available on all nodes. Its source is the control workstation. If you do not use file collections, propagate this file, using whatever means is available in your environment, to all nodes in the SP System.

   Note: You may have to update the **holidays** file to reflect the correct information for the current year.
5. Update the **crontabs** file to schedule the accounting processes. See "Accounting Template for crontabs."
6. Initialize the login and process account data files. Enter:

   ```
   /usr/sbin/acct/nulladm /var/adm/wtmp /var/adm/pacct
   ```

Accounting Template for crontabs

A default **crontab** template file, describing when each accounting command should be run, is shipped as **/usr/lpp/ssp/config/cron_template** and distributed via a file collection. If you wish to run your accounting system on a different schedule, make the appropriate changes to this template. When booting, each node for which accounting is enabled, ensures that its **crontab** file is updated using this template. If the template does not exist, it is assumed that you use file collections to distribute **crontab** and nothing is done.

The **crontab** template file, shown below, contains a partial **crontab** entry. Each line begins with a keyword specifying a command used by the accounting process.

```
# This is the crontab template file. It is read by acct_config
# to update the root crontab file with accounting-related entries.

DODISK   0 1 * * 4
CHKPACCT 5 * * * *
NRUNACCT 0 2 * * 1-6
CRUNACCT 0 4 * * 1-6
CMONACCT 15 5 1 * *
```

Figure 11. Accounting Template for crontabs

DODISK and **CHKPACCT** are standard **crontab** keywords. **NRUNACCT**, **CRUNACCT**, and **CMONACCT** are variations of standard AIX accounting commands discussed earlier in this section. Refer to Chapter 13, "Maintaining the crontabs File" on page 153 for more information on **crontabs**.

Accounting Files

All accounting data, report, and summary files reside in the **/var/adm** directory. The accounting data files belong to members of the *adm* group, and all active data files (such as **wtmp** and **pacct**) reside in the adm home directory, **/var/adm**.

Note

It is very important that the System Administrator monitor the capacity of the **/var** filesystem when exclusive use accounting is enabled. It has been found that the Resource Manager may create large **/var/adm/pacct** files, and that accounting will save the **excl_Spacct** files in the **/var/adm/acct/nite** directory. This may require the administator to periodically increase the **/var** filesystem size or remove unnecessary **excl_Spacct** files. The System Administrator may edit the **/usr/lpp/ssp/bin/nrunacct** script in the CLEANUP section to automatically remove **excl_SPacct** files.

Data Files

Data files are:

/var/adm/acct/nite/dacct
Contains disk total accounting records, created by the **dodisk** command.

/var/adm/fee
Contains output from the **chargefee** command, in ASCII **tacct** records.

/var/adm/pacct
Active process accounting file.

/var/adm/wtmp
Active login accounting file.

/var/adm/Spacct*i.yyyymmdd*
Contains process accounting files for *yyyymmdd* during the execution of the **nrunacct** command.

Report and Summary Files

Report and summary files reside in the **/var/adm/acct** subdirectory. You must create the following subdirectories before the accounting system is enabled.

/var/adm/acct/nite
Contains files that **nrunacct** and **crunacct** command use in daily accounting.

/var/adm/acct/sum
Contains the cumulative summary files that **nrunacct** and **crunacct** update daily.

/var/adm/cacct
Contains files that the **crunacct** command uses in daily accounting.

nrunacct Command Files

The following report and summary files, produced by the **nrunacct** command, are of particular interest.

/var/adm/nsiteacct
Contains site-specific accounting procedures for node accounting.

/var/adm/acct/nite/lineuse*yyyymmdd*
Contains usage statistics for each terminal line on the system. This report is especially useful for detecting bad lines. If the ratio between the number of logouts and logins exceeds 3 to 1, it is possible that a line is failing.

/var/adm/acct/nite/daytacct*yyyymmdd*
Total accounting file for the previous day.

/var/adm/acct/sum/daycms*yyyymmdd*
Contains the daily command summary.

/var/adm/acct/sum/loginlog*yyyymmdd*
Contains a record of the last time each user ID was used.

/var/adm/acct/nite/reboots*yyyymmdd*
Contains beginning and ending dates from the **wtmp** file and a listing of system restarts.

crunacct Command Files

/var/adm/csiteacct
Contains site-specific accounting procedures for cluster accounting.

These files are copied to the **acct_master** node during **crunacct** execution.

/var/adm/cacct/*node***/nite/lineuse***yyyymmdd*
/var/adm/cacct/*node***/nite/daytacct***yyyymmdd*
/var/adm/cacct/*node***/nite/reboots***yyyymmdd*
/var/adm/cacct/*node***/sum/daycms***yyyymmdd*
/var/adm/cacct/*node***/sum/loginlog***yyyymmdd*

The *node* level files are merged to become:

/var/adm/cacct/sum/*class***/tacct***yyyymmdd*.
/var/adm/cacct/sum/*class***/daycms***yyyymmdd*.
which in turn update cumulative summary files:
/var/adm/cacct/sum/*class***/tacct** file..
/var/adm/cacct/sum/*class***/cms** file..

Files in the /var/adm/acct/nite Directory

active
Used by the **nrunacct** command to record progress and to print warning and error messages. The file **active.***yyyymmdd* is a copy of the active file made by **nrunacct** after it detects an error.

jobcharge
Contains a floating point number used to determine the number of charge fee units to charge a user for exclusive use of the node and its value is in units of *seconds per charge fee unit*.

jobrecs
Contains *start job* and *end job* records used for accounting on jobs that make exclusive use of nodes.

ptacct.*yyyymmdd*
Contains summary version of **pacct** files.

ctmp
Contains connect session records during **nrunacct**

accterr
Contains standard error messages produced during the execution of the **nrunacct** command.

lastdate
Indicates the last day the **nrunacct** command executed, in **yyyymmdd** format.

lock
Used to control serial use of the **nrunacct** command.

statefile*yyyymmdd*
Used to record the current state during execution of the **nrunacct** command.

wtmperror
Contains **wtmpfix** error messages.

owtmp.
Contains the previous day's **wtmp** file.

Files in the /var/adm/cacct/ Directory

active
Used by the **crunacct** command to record progress and to print warning and error messages. The file **active.***yyyymmdd* is a copy of the active file made by **crunacct** after it detects an error.

lastdate
Indicates the last day the **crunacct** command executed, in **yyyymmdd** format.

lock
Used to control serial use of the **crunacct** command.

statefile*yyyymmdd*
Used to record the current state during execution of the **crunacct** command.

lastcycle
Indicates the date of the last successfully completed **crunacct** cycle.

fiscal_periods
Defines the start date of each fiscal period.

Files in the /var/adm/cacct/sum Directory

<**class**>/**cms**
Daily command summary.

loginlog
File listing current login records.

rprt*yyyymmdd*
Saved output of the **cprdaily** routine executed in **crunacct**.

<**class**>/**tacct**
Process account summaries.

Files in the /var/adm/cacct/fiscal Directory

/var/adm/cacct/fiscal/*class***/cms***mm*
Total class summary file for the month, in binary format.

/var/adm/cacct/fiscal/*class***/tacct***mm*
Total accounting file for month, in binary format.

/var/adm/cacct/fiscal/fiscrpt*mm*
Report similar to that of the **prdaily** command for month, in binary format.

Chapter 13. Maintaining the crontabs File

The SP system uses the **crontabs** file to periodically update file collections and clean up log files. The installation process appends entries to the **crontabs** files in **/var/spool/cron/crontabs/root** on the control workstation and each of the nodes. The files contain different entries depending on their location.

- Control workstation
  - Merge accounting information
  - Clean up the logs
- Processor nodes
  - Update the **sup.admin, user.admin** and **node.root** file collections
  - Merge accounting information
  - Clean up the logs

The following example shows these entries on a control workstation:

```
 0 * * * * /usr/lpp/ssp/bin/dev/null 2>/dev/null
 0 0 * * * /usr/lpp/ssp/bin/cleanup.logs.ws
0 4 * * 1-6/usr/lpp/ssp/bin/crunacct 2>/var/adm/cacct/nite/accterr
15 5 1 * * /usr/lpp/ssp/bin/cmonacct
```

The default is to run these processes hourly, in a staggered sequence. You can customize this for your environment to make any of these processes run more or less frequently. You can also add your own entries to these files for other processes you want automated.

Updating crontabs on the SP Nodes

If you want the same **crontab** file on all SP nodes, one way to update it across the whole SP System is:

1. Get the current **crontab** from one of the nodes:

   ```
   rsh hostname crontab -l > /tmp/mycrontab
   ```

2. Edit the file:

   ```
   vi /tmp/mycrontab
   ```

3. Propagate the changed **crontabs** file to all SP nodes:

   ```
   dsh -a rcp root@mynode:/tmp/mycrontab /tmp/mycrontab
   dsh -a "crontab < /tmp/mycrontab"
   ```

If you change the default configuration to modify the hierarchy of file collection servers, you need to modify the **crontabs** files to reflect the new relationship. See “Modifying the File Collection Hierarchy” on page 61 for more information. If you create your own file collections, you will also need to add them to the **supper update** commands in the **crontabs** file to keep them on schedule with the other upgrades.

© Copyright IBM Corp. 1995, 1997

Security on the SP System

© Copyright IBM Corp. 1995, 1997

Chapter 14. Security Features of the SP System

This chapter includes information on the security services provided with the IBM RS/6000 SP (SP). The security services are part of the IBM Parallel System Support Programs for AIX (PSSP) software that runs on the SP.

- "Understanding Authentication" provides an introduction to general security concepts and terminology, as well as an overview of the way authentication services are provided.
- "Using SP Security Services" on page 166 describes the tasks involved in running authenticated client-server applications supplied with the SP system.
- "Administering SP Security Services" on page 172 explains the administration of the security services, including the tasks of setting up authentication services on an SP system and on other workstations, adding principals to the authentication database, changing principals' passwords and other attributes, managing server key files, backing up the authentication database, and using secondary authentication servers.
- "Using AFS Authentication Services" on page 193 provides an overview of how AFS can provide an alternative to using PSSP authentication servers.

For more information about installing and configuring authentication services, refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*. For information about diagnosing authentication problems, refer to the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide*.

Understanding Authentication

Administration of an SP system from a single point-of-control requires the use of administrative applications that are based upon the client-server programming model. Typically, a client-side command-line or GUI interface is used to perform an administrative task on one or more network-attached nodes or workstations, in parallel or serially. Where privileged applications such as these communicate across networks, security is compromised unless a reliable means is used to restrict access to authorized users. Accomplishing this authorization first requires authentication, which is the validation of the identities of client and server. Authentication is also the foundation for providing other security features such as accountability, discretionary access controls, and least privilege.

An SP system implementation, for example, can consist of SP nodes with Internet connections, a control workstation, and connections to other clustered workstations. Any of these system elements can be vulnerable to a security compromise. All networked services, as well as clients who may need to provide sensitive data to services, must be able to restrict access to those services according to a meaningful security policy. Such access restrictions can be successful only by the reliable authentication of the identities of clients and servers.

The PSSP software includes two types of administrative programs that benefit from these security features:

- Applications that manage the hardware resources of the SP system

© Copyright IBM Corp. 1995, 1997

- Applications that provide secure remote execution of SP administration commands. These include support for delegation of root user tasks to other users without distributing root's password.

SP authentication is based on Version 4 of the Kerberos authentication service developed at MIT. It was named after the mythical three-headed dog that guards the entrance to the underworld. The significance of the name is that a so-called “trusted third party”, namely a Kerberos authentication server, provides credentials that allow the other two parties (client and server) to reliably determine each other's identity.

Terminology

To understand the SP security services and how they are used and managed, you should be familiar with the following terms:

ACL Access Control List. A list that defines who has permission to access certain services; that is, for whom a server may perform certain tasks.

AFS A distributed file system that uses Kerberos authentication and includes authentication services that can be used by an SP system.

Authentication The process of validating the identity of either a user of a service or the service itself.

Authentication Database
A set of files containing the names and authentication information of all principals within a realm. An authentication realm has one primary database and may have multiple secondary databases. Secondary databases are backup copies of the primary database and may be provided to improve performance or availability.

Authenticator An authentication protocol string created each time authentication occurs and sent with the ticket to the server. It contains a time-stamp encrypted in the session key that can reliably show that the authentication request actually came from the client identified in the ticket.

Authorization The process of obtaining permission to access resources or perform tasks. In SP security services, authorization is based on the principal identifier.

Cell An independently administered collection of file server and client machines running AFS. An AFS cell is equivalent to a Kerberos realm.

Client The user of the shared services of a server. This term may refer to a program or to the person who invoked it. For authentication purposes it is a principal identifier registered in the authentication database.

Credentials A protocol message, or part thereof, containing a ticket and an authenticator supplied by a client and used by a server to verify the client's identity. The message can contain additional information used by the server to verify its identity to the client.

Data Encryption Standard (DES)
The secret-key (also known as “private-key”) encryption algorithm that is used by Kerberos.

Domain A collection of systems over which an administrator exercises control.

Instance A qualifier for a principal name. For services, an instance represents a particular occurrence of the server. For users, an instance allows a single user to assume additional (or alternate) roles with different authority.

Kerberos A service for authenticating users in a network environment.

Key An 8-byte value derived by encrypting a password or generating a random number, used to encrypt protocol strings used for authentication. Principal's private keys are stored in the authentication database. Session keys are contained (encrypted) in tickets and other protocol strings.

Kerberos Master Key
The key derived from the Kerberos Master Password supplied initially by the administrator when the primary SP authentication server is created. This key is saved in the **/.k** file, from whence the Kerberos daemons read it (in lieu of prompting for a password). It can also be read by certain database utility commands for the same purpose.

Kerberos Master Password
The password the administrator supplies when initializing the primary authentication server.

Principal An entity whose identifier and key are maintained in the authentication database. A principal can represent a user or a service. The format of a principal identifier is *name.instance@realm*, where *name* is either a user's login-name or the name by which a server identifies itself to Kerberos.

Realm A domain which shares an authentication database and servers. There is a single name-space for principal name/instance pairs within a realm. A realm is also a logical collection of clients and servers registered in the database.

Server A functional unit (usually a daemon program and child processes) running on a particular host that provides shared services to users over a network.

Service The name of a principal whose identity is assumed by a server for purposes of authentication. Multiple servers can use the same service name.

Server Key File (Also referred to as a **srvtab**) A file containing the names and private keys of the local instances of services. It is accessible only to processes that run under the UID of the user owning the server daemons. On an SP system all services run as root.

Session Key A temporary key, a random 64-bit value, supplied by an authentication server to clients and servers, that is used to encrypt parts of authentication protocol messages. Its lifetime is the same as the ticket with which it is created.

Ticket An encrypted protocol message used to pass the identity of a user from a client to a server. Tickets are created by the Kerberos authentication server and cached in disk files on the client's system.

Ticket-granting-ticket
The initial ticket obtained by a user. It is used by client programs to obtain additional tickets for authentication with application services.

Kerberos Authentication

SP authentication services use the authentication database and the server to:

- Initially authenticate the identity of the user
- Provide the information through which a server can authenticate the identity of a client (user) in a distributed environment

The authentication services provided in the SP System are based on MIT Kerberos version 4. Kerberos servers function as a trusted third party to authenticate the identities of users and distribute encryption keys to clients and servers, known as *principals*. Every user and every service instance on a particular host is a principal in the Kerberos database. Each Kerberos principal has an identifier, a password, an expiration date, and a maximum ticket lifetime registered in the authentication database. Kerberos does not store the clear-text password, but a 64-bit binary value derived from it using a one-way hash. Database information is also encrypted in the Kerberos master key, known only to the database administrator. Passwords are never sent across the network in clear text-form, either for user authentication or during database administration.

Overview of Kerberos Interaction with Client and Server

On the authentication server system there are two daemons that access the database. The **kadmind** daemon provides network access for database administration via the **kpasswd** and **kadmin** commands. The **kerberos** daemon is the actual authentication server. It is sometimes referred to as the “Key Distribution Center (KDC)” because of its role in distributing session keys. The KDC actually consists of two separate functions: the Authentication Service, which generates ticket-granting-tickets based on authentication using the client user's private key; and the Ticket Granting Service (TGS), which generates service tickets, based on authentication using a ticket-granting-ticket.

The underlying mechanism for accomplishing the authentication is a ticket scheme, consisting of the following steps, illustrated in Figure 12 on page 161.

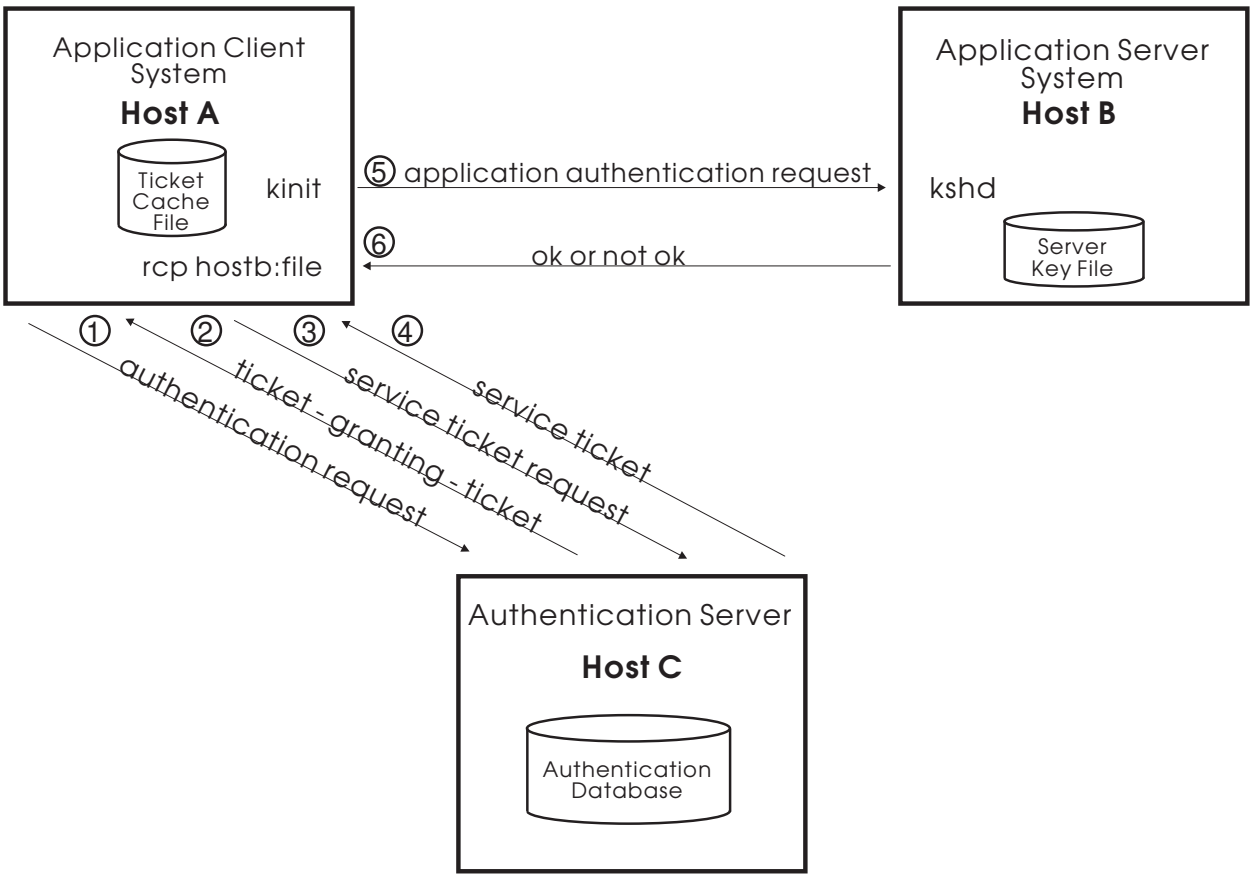


Figure 12. An Overview of Kerberos Authentication

1. The client process, usually as the result of a user invoking the **kinit** command, sends a message to the Authentication Service containing its principal identifier and the identifier of the Ticket Granting Service (a predefined Kerberos principal). This message is not encrypted. The Authentication Service generates a random session key to be shared by the client and the TGS and creates a ticket for the client to use with the Ticket Granting Service. It contains the client identifier and the session key, as well as fields defining the ticket's valid lifetime. The ticket is encrypted in the private key of the Ticket Granting Service. A return message is built, containing the key, the client's identifier, and the ticket-granting-ticket. The Authentication Service encrypts the message in the private key of the user, a one-way hash of the password.
2. The client program reads the user's password and applies the one-way hash necessary to determine the private key. If the password is correct, the key can be used to successfuly decrypt the returned message block containing the session key, the client identifier, and the ticket. When successful, the ticket and the session key are stored in the current ticket cache file.
3. The user invokes an application's client interface, which needs to perform authentication on its first network communication with its server. It creates an authenticator, containing the client identifier and a current time-stamp, and encrypts the authenticator using the sesion key returned with the ticket-granting-ticket. It sends to the Ticket Granting Service a message consisting of the authenticator, the ticket-granting-ticket, and the principal identifier of the service instance which it wants to use. The TGS reads its private key from the authentication database and uses it to decrypt the ticket,

obtaining a copy of the sesssion key and the identifier of the client to whom the ticket-granting-ticket was issued. Using the session key, it decrypts the authenticator. The TGS verifies that the client's identity is the same, and that the authenticator's time stamp indicates that it was very recently created.

4. Like the Authentication Service, the TGS generates a session key, this one to be shared by the client and the application server. It also generates another ticket, called a service ticket, encrypted in the private key of the application service principal. The ticket is returned to the client in a message encrypted with the old session key (shared by the client and the TGS).
5. The client program uses the original session key to decrypt the message. It extracts the new session key and ticket, and saves them in the ticket cache file. When the client program sends its first application protocol message to the application server, it includes the service ticket and a new authenticator, encrypted in the new session key. This session key was also placed in the ticket by the TGS, so the server can retrieve it when it decrypts the service ticket. In order to do so, the application server obtains its private key from the server key file, **/etc/krb-srvtab** on the server system.
6. The authentication by the application server is performed in the same way as by the TGS. The message returned to the client simply indicates success or failure, and an error message number if failure.

The scenario outlined above assumes no unexpired service ticket was found in the ticket cache file. The ticket-granting-ticket can be used to get as many different tickets as are needed, without requiring the user to reenter the password.

A client's service ticket can be used to re-authenticate as often as the client and server require. Typically, clients using TCP/IP authenticate once on connection to the server. Users of connectionless protocol should authenticate each message exchange, though less rigorous schemes can be used following an initial authentication, by using the shared session key to encrypt a small piece of known data.

Kerberos and DES

The SP authentication services use the Data Encryption Standard (DES) algorithm. The DES algorithm is used only as an internal mechanism for authenticating the identities of clients and servers for SP services. In order to comply with U.S. export regulations on the DES algorithm, SP authentication services provide a subset of the full MIT Kerberos version 4 services. No interface is provided for using the encryption routines directly. No interface is provided to allow the encryption of any portion of service messages that are not directly connected with the authentication of the client and server identities.

For additional information, refer to the **Kerberos** man page in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Types of Principals

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

User Principals for SP System Management

An implementation of the SP system must have at least one user principal defined. This user is the authentication database administrator, who must be defined first so that other principals can be added later. When AFS authentication servers are being used, the AFS administrator ID already exists when the SP authentication services are initialized. When PSSP authentication servers are being used, one of the steps included in setting up the authentication services is the creation of a principal whose identifier includes the **admin** instance. It is recommended, but not essential, that the first authentication administrator also be the root user. Various installation tasks performed by root, or other users with UID 0, require the Kerberos authority to add service principals to the authentication database. At sometime prior to invoking the **setup_server** command to complete setup of the control workstation and boot server nodes, a user with UID 0 must be defined as a Kerberos database administrator with authority to add other principals. This is because the service principals used by Kerberos-authenticated applications are all defined automatically by **setup_server** during installation and customization. Apart from this installation requirement, you may use other user principal identities to perform system administration (and authentication administration) tasks.

Service Principals Used by PSSP Components

Two service names are used by the Kerberos-authenticated applications in an SP system:

1. **hardmon**, used by the System Monitor daemon on the control workstation and by logging daemons
2. **rcmd**, used by **sysctl** and the authenticated **rsh** and **rcp** commands

By convention, a Kerberos Version 4 server runs as a Kerberos principal whose identifier is formed from the service name and the short host name of the node on which the server runs, converted to lowercase. Different principal identifiers thus indicate the target network address of the different providers of each service. The authentication support in the SP system expands the Kerberos implementation to support workstations with multiple network interfaces as servers. This provides instances for all network interfaces rather than just the standard host name. Kerberos works with the resolved network names, rather than any aliases you may use locally on your hosts.

The **hardmon** daemon runs only on the control workstation. The SP logging daemon, **splogd**, can run on other IBM RS/6000 workstations. Therefore, for each (short) network interface name on these workstations, a service principal is created with the name **hardmon** and the network name as the instance.

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

Authentication Realms

The authentication realm is the set of systems that is served by a single logical authentication database. A realm can include any combination of one or more SP systems, IBM RS/6000 workstations, and other systems/workstations. The authentication database for a realm is administered as a single entity: principals are defined across the systems within a realm and principal identifiers must be unique within a realm.

There are options for defining authentication realms within a network of systems:

- Each system may be its own realm
- A single realm may include all of the systems in a network
- Systems may be grouped into realms with multiple realms occurring within a network

Generally, systems that are administered together, or by the same set of administrators, should be grouped into a single authentication realm. For example, one or more SP systems and their associated workstations could form a single authentication realm. Within an SP system, all SP nodes are automatically added to the control workstation's local realm during installation.

If multiple realms are configured within a network, the realm names should be unique. During setup of the SP authentication services, the name of the realm defaults to the local system's network domain name converted to uppercase. If a different realm name is going to be used, the local system's realm name must be supplied by the administrator in the **/etc/krb.conf** file. For information on the format of this file, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

If you are using AFS authentication servers, the entire SP system must be in the same cell. The authentication realm known to the SP authentication services is the local AFS cell. The realm name is set to the AFS cell name converted to uppercase.

Although a realm is served by a single logical authentication database, multiple copies of the database may exist on different workstations within a realm. One must be the primary database in a realm; others are secondary servers.

Understanding Tickets

Kerberos tickets are encrypted byte strings returned to clients by an authentication server. Tickets contain the client principal's identifier, a time-stamp and ticket lifetime used to determine the expiration time, the session key, and other information, all encrypted in the private key of the service instance with whom the client wishes to authenticate. In the case of a ticket-granting-ticket, such as that created by the **kinit** command, that service is the Ticket Granting Service. Other tickets are application service tickets, encrypted in the private key of one particular instance of an application service. The client cannot understand the key; it simply forwards it to the server with an authenticator in an internet protocol packet, using either UDP or TCP socket communications.

Ticket Cache Files

The caching of tickets takes place automatically in disk files on the client system that is the workstation or SP node on which the client user is logged in. These files are created with UNIX permissions set so they can be accessed only by the owner, for they contain sensitive information. Any process using the tickets can authenticate as the owner of the file.

A user can have tickets cached in any number of files simultaneously, and may have reason to do so. If you have multiple logins, run background processes, or use a shared UID, you must do so to avoid destruction of tickets still in use by other processes. This is accomplished by setting the KRBTKFILE environment variable to the path name of a unique ticket cache file prior to using authentication services for each login session or background process. If you do not set KRBTKFILE, then your tickets will be cached in your default file: /**tmp**/**tkt***uid*.

The first ticket stored in the file is always the ticket-granting-ticket. The file is created by the **kinit**, **ksrvtgt**, or **rcmdtgt** command; and when the TGT is stored, all previously issued tickets in the file are deleted. The file contains, in addition to the TGT, any number of other tickets for application services. There will be one for each service instance with which you have communicated using that particular current ticket cache file. They are obtained by the client commands for the various Kerberos-authenticated services, such as **rcp** and **rsh** (and indirectly by SP system administration tools that invoke them).

Kerberos does not automatically delete tickets after they expire. Users must either explicitly destroy their ticket cache files or replace them by re-authenticating (by obtaining a fresh TGT). Tickets are never removed individually from the ticket cache file. The **kdestroy** command eliminates the entire current ticket cache file. Ticket cache files other than the current file are untouched, however.

Ticket-lifetimes

The tickets that you obtain when you issue the **kinit** command are valid only for a limited time. The same limit applies also to service tickets obtained by client programs that you execute. In most circumstances, you probably will not be concerned about how this limit is determined and applied. However, when you run **kinit**, you have the option of requesting a ticket with a specific lifetime. The maximum lifetime you may specify, which is also the default value, was assigned when your principal was registered in the authentication database, or it was changed subsequently by the database administrator. The minimum ticket lifetime is five minutes; the maximum lifetime can be set to one of several discrete values from five minutes to thirty days. The ticket lifetimes of service tickets are always the same as the lifetime of the ticket-granting-ticket used to obtain them. It is the TGT held in the current ticket cache file owned by the user who invokes the application client command.

The external representation of the ticket lifetime is rather inconsistent among the various user and administrator commands. Some commands represent the one-byte ticket lifetime value as a decimal integer between 0 and 255. This is the representation used by the administrator when creating or modifying a principal using the **kdb_edit** command. It is also the format of the ticket life field displayed by the **get** subcommand of **kadmin**. The output of the **klist** command shows the actual ticket creation and expiration as standard date and time stamps. If you specify **-l** to request a a specific ticket lifetime when you run **kinit**, you are prompted to enter the value as a number of minutes. The value you enter is then

rounded up to the next higher discrete lifetime interval or the maximum ticket lifetime allowed for your principal, whichever is less, and then mapped to a one-byte value between 1 and 191.

The following table shows the ticket lifetimes that correspond to the encoding used internally by Kerberos. The values on the left of each column are the one-byte encodings; on the right, the lifetime interval in days, hours, and minutes.

Ticket lifetimes from 0 to 128 represent multiples of 5 minutes, for example:

```
0-1     00:05    12     01:00    24     02:00    36     03:00    48     04:00
 60     05:00    72     06:00    84     07:00    96     08:00   108     09:00
120     10:00   125     10:25   126     10:30   127     10:35   128     10:40
```

Lifetimes from 129 to 255 represent intervals of from 11:24 to 30 days:

```
129     11:24   130     12:11   131     13:02   132     13:56   133     14:54
134     15:55   135     17:01   136     18:12   137     19:28   138     20:48
139     22:15   140     23:47   141  1d+01:26   142  1d+03:11   143  1d+05:04
144  1d+07:05   145  1d+09:14   146  1d+11:32   147  1d+13:59   148  1d+16:37
149  1d+19:25   150  1d+22:26   151  2d+01:38   152  2d+05:04   153  2d+08:44
154  2d+12:40   155  2d+16:51   156  2d+21:21   157  3d+02:08   158  3d+07:16
159  3d+12:45   160  3d+18:36   161  4d+00:52   162  4d+07:34   163  4d+14:44
164  4d+22:23   165  5d+06:35   166  5d+15:20   167  6d+00:41   168  6d+10:42
169  6d+21:23   170  7d+08:50   171  7d+21:03   172  8d+10:08   173  9d+00:06
174  9d+15:03   175 10d+07:01   176 11d+00:06   177 11d+18:22   178 12d+13:53
179 13d+10:46   180 14d+09:05   181 15d+08:56   182 16d+10:27   183 17d+13:44
184 18d+18:53   185 20d+02:04   186 21d+11:24   187 22d+23:02   188 24d+13:08
189 26d+05:52   190 28d+01:26   191+ 30d
```

Using SP Security Services

Managing Your Authentication Credentials

Users who are going to use the Kerberos-authenticated applications that are provided with the SP system, as system administrators or ordinary users, need to understand how to perform a few simple tasks to manage their authentication environment. These tasks involve:

- Obtaining Kerberos credentials
- Examining Kerberos credentials
- Deleting credentials that are no longer needed
- Changing their Kerberos passwords

When your realm is set up to use PSSP authentication servers, the PATH variable for these users should include **/usr/kerberos/bin**. When AFS authentication servers are used, the PATH variable should include **/usr/afsws/bin:/usr/kerberos/bin**.

Obtaining Authentication Credentials

The first step in using the SP authentication services is to login to the authentication system. Logging in to the SP authentication services is how you obtain a ticket-granting ticket. The ticket-granting-ticket permits the client programs you invoke, such as **sysctl**, to obtain service tickets they can use to certify your identity to the servers with which they communicate.

Logging in to SP authentication services is a separate step from logging in to your AIX system; the Kerberos principal name space is separate from the AIX user name space. The administrator may assign principal names and passwords that are the same as AIX login names and passwords.

With PSSP Authentication: Use the **kinit** command to login to the SP authentication services. You can either enter the **kinit** command interactively as needed, or you could add a **kinit** command to your login script (**.login** or **.profile** file). If you make infrequent use of the Kerberos-authenticated services, you might prefer the first option. Another alternative for frequent users of Kerberos authenticated management tools is to have the System Administrator set up Kerberos authentication as an authentication method to be invoked automatically at AIX login. For more information on this option, see "Setting Up a User Account to Obtain Kerberos Credentials at AIX Login" on page 176.

The easiset way to use the **kinit** command is to specify your Kerberos principal identifier as the only argument with no options.

kinit *principal-identifier*

For the principal identifier, specify your Kerberos user name, followed by a period and the instance, if you have more than one Kerberos identity and the instance you want to use is not null. Append an "at" sign (@) and a realm name if the realm you are logging into is not the local realm. Since the entire SP system is in a single realm, you will probably never need to specify a realm. Examples of this form of **kinit** are:

```
kinit joeuser
kinit root.admin
kinit ben@XYZ.ABC.COM
```

You might also choose to invoke **kinit** without arguments. It responds by prompting for your Kerberos name, but only for the name portion of the fully qualified identifier.

```
kinit
Kerberos name: ben
```

A common error is to omit the identifier on the command line and try to enter both name and instance in response to the prompt for the name. The result is:

```
kinit
Kerberos name: root.admin
2502-003 Bad Kerberos name format
```

Instead, you can enter the instance separately by specifying the **-i** flag on the command line and responding to the prompt:

```
Kerberos instance:
```

Similarly, the realm qualifier can only be entered as part of the command-line identifier unless you use the optional **-r** flag.

The other option that you might want to use, **-l**, allows you to request a ticket with a lifetime shorter than the maximum allowed for your principal by the database administrator. If you specify this flag, **kinit** prompts you:

```
Kerberos ticket lifetime (minutes):
```

After you enter the **kinit** command and respond to these optional prompts, the command contacts the authentication server to obtain a ticket. If no authentication server can be contacted, you will be notified of a time-out and the command will fail. Reasons for failure could be a missing or invalid configuration file or a networking problem.

When a ticket is obtained, **kinit** prompts you for your Kerberos password. If you enter it correctly, **kinit** saves the ticket-granting-ticket in the ticket file specified by your KRBTKFILE environment variable. If this variable is undefined, your ticket is stored in the file /**tmp**/**tkt***uid*, where *uid* specifies your AIX user identification number.

When the ticket-granting-ticket expires, subsequent attempts to be authenticated to use a Kerberos-supported service will fail. If the ticket-granting-ticket has expired, you must use **kinit** again to obtain a new ticket. There is no automatic renewal of tickets upon their expiration.

When you use **kinit**, it is valid for all processes running with your AIX user ID. It is not necessary to do a **kinit** for each window you open, as long as you have no background processes that require authentication. Note, however, that if you run authenticated applications as root in background processes or login under a shared user ID, you can potentially destroy the tickets of someone who logged in prior to you using the shared ID. In those cases you should use the KRBTKFILE environment variable to specify multiple ticket cache files. If you are uncertain of the current ticket cache file, use the **klist** command to display the name and content of the file.

Refer to the **kinit** and **klist** commands in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

With AFS Authentication: In most situations users of your SP system will be able to use the same **kinit** command described above without regard to which type of authentication servers are being used. **kinit** can always be used whenever a user is logged into an SP node or the control workstation. When the user is logged into another IBM RS/6000 workstation which has been set up as an authenticated services client system, **kinit** can be used if the user has the SP_NAME environment variable set to the network name of the control workstation in the local realm.

Users who do not have the SP_NAME variable set, or who wish to use more familiar AFS commands, should use the **klog.krb** command. The following example shows the format of this command:

```
klog.krb -pr root.admin -li 48:00
```

When the **admin** instance is included in the principal identifier, **klog.krb** responds with a warning message:

```
Non-null instance (admin) may cause strange behavior.
```

You can ignore this message, since these principals are not used with the AFS file system.

As with **kinit**, a lifetime can be specified, but in a more reasonable *hh:mm* format.

Displaying Your Credentials

With PSSP Authentication: Use the **klist** command to display a list of currently held authentication tickets and their expiration time. The output would appear like the following, for example, if the user **ben** enters the **klist** command after using the **kinit** command followed by any of the remote command services on four other hosts:

```
Ticket file:    /tmp/tkt12763
Principal:  ben@XYZ.ABC.COM

  Issued           Expires          Principal
Aug 12 16:26:11  Sep 11 16:26:11  krbtgt.XYZ.ABC.COM@XYZ.ABC.COM
Aug 12 16:45:15  Sep 11 16:45:15  rcmd.spcwkst@XYZ.ABC.COM
Aug 15 09:02:42  Sep 14 09:02:42  rcmd.spnode3@XYZ.ABC.COM
Aug 15 09:02:43  Sep 14 09:02:43  rcmd.spnode5@XYZ.ABC.COM
Aug 15 09:02:43  Sep 14 09:02:43  rcmd.spnode6@XYZ.ABC.COM
```

Refer to the **klist** command in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

With AFS Authentication: You may also use the **klist** command to display tickets obtained from an AFS authentication server. The format of the output is the same as shown above, except that a service ticket for the “afs” service will appear, if you used the AFS **klog.krb** command to identify yourself. This additional service ticket is not used by SP authentication services. As an AFS user, you may prefer to use the **tokens.krb** command to display your tickets. Its output will appear like:

```
Tokens held by the Cache Manager:

User's (AFS ID 12763) tokens for afs@afs.abc.com [Expires May 31 03:24]
User ben's tokens for krbtgt.AFS.ABC.COM@afs.abc.com [Expires Jun 25 19:49]
User ben's tokens for rcmd.wksta2@afs.abc.com [Expires Jun 25 19:50]
   --End of list--
```

Notice that **tokens.klog** shows the AFS cell name in place of the Kerberos realm name.

Deleting Your Credentials

With PSSP Authentication: Use the **kdestroy** command to destroy Kerberos tickets in the current ticket cache file. Depending on your local security policy, you may want to destroy your tickets whenever you end your AIX login session. **kdestroy** overwrites the ticket cache file with nulls before removing the file, so storage containing credentials cannot be reused. When you enter the **kdestroy** command, the system will respond:

```
Tickets destroyed.
```

If the ticket cache file does not exist, you will get the response:

```
No tickets destroyed.
```

Refer to the **kdestroy** command in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

With AFS Authentication: You may also use the **kdestroy** command to delete tickets obtained from an AFS authentication server and stored in a ticket cache file. The responses are the same as shown above. **kdestroy** will not remove the token

held by the AFS Cache Manager, however, that was also obtained if you used **klog.krb** instead of **kinit** to obtain your initial ticket.

The AFS **unlog** command will remove your tokens held by the AFS Cache Manager, but will not remove tickets from the Kerberos ticket cache file.

Changing Your Authentication Password

With PSSP Authentication: To change your password, use the **kpasswd** command:

1. Enter the **kpasswd** command:
2. At the prompt, enter your old password
3. At the prompt, enter your new password
4. At the prompt, reenter your new password

For example:

```
kpasswd
Old password for ben:
New Password for ben:
Verifying, please re-enter New Password for ben:
Password changed.
```

With AFS Authentication: To change a password for a principal in AFS:

1. Enter the AFS **kpasswd** command. Be sure you have the directory for the AFS commands, /**usr**/**afsws**/**bin** ahead of the directory path for the Kerberos command, /**usr**/**kerberos**/**bin** in your search order, or specify the full pathname:
2. At the prompt, enter your old AFS password
3. At the prompt, enter your new password
4. At the prompt, reenter your new password to verify

For example:

```
kpasswd
Changing password for 'ben' in cell 'abc.com'
Old password:
New password (RETURN to abort):
Retype new password:
Password changed.
```

SP Authenticated Services

The SP System provides authentication services for:

- The SP **rsh** and **rcp** commands
- The Sysctl remote command execution facility
- The SP System Monitor

Separate chapters describe the configuration and use of the System Monitor and the Sysctl remote command execution facility.

Authenticated Remote Commands on the SP System

SP authentication services provide authenticated versions of **rsh** and **rcp**. The use of the authenticated versions of **rsh** and **rcp** removes the requirement for a root level **.rhosts** file.

Kerberos version 4 authentication services must be installed on both the source host issuing and the target host servicing the authenticated remote commands. Refer to "Setting Up PSSP Authentication" on page 173 for more information.

To use the **rsh** and **rcp** authenticated services, a user or client must be included in the authentication database and must first issue the **kinit** command.

If you cannot be authenticated or if authentication services fail, messages are issued and the SP authenticated remote commands pass the arguments to the corresponding AIX remote commands.

Remote Command Authorization

The root user's **.klogin** file contains a list of principals that are authorized to invoke processes as the root user with the SP authenticated remote commands (**rsh**, **rcp**). A **.klogin** file is created in the root user's home directory rather than a root **.rhosts** file. No root **.rhosts** file is used by the SP system management installation and configuration programs. The **.klogin** file contains service principal entries to allow the system management installation and configuration programs to authenticate without requiring login as an administrative user principal. The service tickets allow these programs running on the SP nodes to communicate with their boot/install servers and the control workstation without forcing the root user to log in at the nodes and perform a **kinit** as administrative user during the installation process and/or administrative operations.

Although the **.klogin** file may contain more entries, the following are required for the installation and operation of the SP system authenticated remote commands and **sysctl** to perform properly. These entries are created during installation on the control workstation and each SP node:

- An entry for the administrative user
- An entry for each host in the SP System

in the form

`rcmd.`*`host name@realm`*

where *host name* is the short host name.

A principal name for the root ID is not automatically added to the **.klogin** file or the authentication database. Refer to **.klogin** file in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

Adding Principals and Distributing the root .klogin File: The root **.klogin** file on the control workstation is distributed to an SP node when it is rebooted. Note that you should only add principals to the **.klogin** file on the control workstation; don't delete the principals that already exist in the file.

If you wish to add principals to the root **.klogin** file and distribute the new file without rebooting, you have several options:

- **dsh**

You can use **dsh** to copy the **.klogin** file from the control workstation to each node in the SP System.

- **pcp**

  You can use the parallel copy command, **pcp**, to copy the **.klogin** file from the control workstation to each node in parallel.

- File collections

  You can use file collections to manage the **.klogin** file. The appropriate collection to place the file in would be **node.admin**.

File collection information can be found in Chapter 5, “Managing File Collections” on page 41. Refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for information on **dsh** and **pcp**.

Administering SP Security Services

Protecting Your Authentication Database

An important consideration when you install and configure your SP system is the security of the workstations on which you will be running Kerberos authentication servers. Because it is the repository of all principals' secret keys, you need to consider steps to protect it from break-in or compromise of the master password. The following are steps you should consider:

- Authentication server systems should be located in physically secure areas, with access limited to administrators authorized to perform tasks related to their maintenance.
- Ordinary users, including SP system administrators who are not authorized as security administrators, should not have user IDs on these systems. Remote access to them via **telnet**, **rlogin**, and **ftp** should be disabled.
- AIX system auditing should be enabled on these systems to record security events.
- An appropriate password selection and aging policy should be implemented for the master password, and it should be changed regularly.
- Because compromise of the master password would expose all private keys stored in the database, a recovery plan should be available that would provide for changing all passwords, replacement of all server key files, and destruction of all outstanding tickets.

The root User and Authentication Administration Roles

When you set up your authentication realm and install and configure your SP system, the roles of system installer and administrator and security administrator become intertwined. This is necessitated by the fact that the PSSP installation and customization scripts are designed to automatically recognize the service principals that have to be created for the **hardmon** and **rcmd** services. The **setup_server** script creates service principals whenever it discovers a new network interface defined on the control workstation for which it has no Kerberos principal defined. At these times, the script must be running in an environment where the invoking user has authentication credentials as a database administrator. If the script needs to add a principal but finds that you have not got a ticket as an **admin** principal, it will

fail. If this happens, simply get the required ticket and issue the command or invoke it from SMIT again.

Other installation and customization tasks invoke the **setup_server** script internally. Any of these tasks which involve adding or renaming of adapters or host names affect the Kerberos principals, and will also require a ticket as a database administrator. Examples of installation scripts used to perform these tasks are **spadaptrs**, **sphostnam**, **spbootins**, and **spethernt**.

Other administration tasks that simply use the remote command facilities of **dsh**, **rsh**, or **rcp** require normal user principal's credentials, since they do not involve changes to the authenication database.

Setting Up PSSP Authentication

To provide authentication services on your SP System, you must choose from the following three authentication implementations.

1. The authentication services provided with the SP System, based on MIT Kerberos version 4. SP authentication services are designed to provide an easily-installable and configurable Kerberos for the SP System.
2. An existing implementation of MIT Kerberos version 4, provided it is compatible with the SP Kerberos-authenticated services.
3. AFS version 3.3 or later.

Once you have installed and initialized authentication on your SP system, you might want to add other IBM RS/6000 workstations to your authentication realm. Two reasons are:

1. If you are not using AFS authentication servers, you may want the added reliability of one or more secondary SP authentication servers. Providing secondary servers is particularly important if your SP system has to provide high availability with no single points of failure.
2. You or other administrators may also want to install the SP authenticated services on office workstations. This will allow system administrators to manage the SP system, without having to login to the control workstation or the SP nodes directly.

When you use the SP authentication server support in the **ssp.authent** option, you may choose to have more than one server system. If you configure your local authentication realm with more than one server, you must designate one as the primary server. All others are secondary servers. Only the primary server has the administrative daemon, **kadmind**, that manages the content of the authentication database. The databases used by the secondary authentication servers are copies of the primary database and are updated periodically from the primary authentication server.

Additional servers can provide for greater reliability, since the authentication service will try all servers listed in the configuration file before failing a service request. If there are network problems, or if your primary server system fails, authentication requests can be handled as long as one of the configured servers is active and accessible. There may also be performance reasons to configure a secondary server. Only when there is an authentication server running on the control workstation can the High Performance Switch be used for authentication protocol traffic to and from the SP nodes. In configurations where you are integrating the

SP system into an existing authentication realm, where the primary authentication server is already on another workstation, you can make this possible by setting up a secondary server on the control workstation. In this way, for example, you could have multiple SP systems in the same authentication realm, and give all SP nodes access to an authentication server across the switch.

The addition of principals and changing of passwords takes place in the primary database, and is propagated to secondary databases through these periodic updates only. Secondary databases are maintained by a special daemon called **kpropd**, that runs only on secondary server workstations and receives the database content in encrypted form from a program called **kprop** that runs on the primary server workstation. The SP authentication services include a script that you can schedule for execution in the root **crontab** file to keep your secondary authentication databases up-to-date.

Configuration Files

There are several configuration files in the **/etc** directory that may be used in setting up your SP system's authentication realm:

- **/etc/inetd.conf**
- **/etc/services**
- **/etc/krb.conf**
- **/etc/krb.realms**

inetd.conf: The **inetd.conf** file contains information used by the internet routing daemon **inetd** to route incoming requests for service to one of a large number of dynamically started daemons, named in the file. When the SP athentication services are installed on your systems, the file is updated to route “kshell” service requests to the Kerberos-authenticated remote command daemon named **kshd**. You should not modify this information, but could possibly have to resolve conflicts with locally installed services with conflicting names.

/etc/services: The **/etc/services** file maps the names of network services to the well-known ports that they use to receive requests from their clients. This file may already contain entries relating to authentication services, since the file shipped with AIX 4.1 contains entries used by DCE. In addition to the DCE entries, many other reserved port names were added in this version of AIX, including an entry for the Kerberos 5 port, 88. The service name for this port is given as “kerberos”, which is also the name used by the standard MIT Kerberos 4 service. The port number usually assigned to the Kerberos 4 service is 750. In order to be consistent with and interoperate with AIX 3.2.5 systems running PSSP 1.2 with authentication services based on Kerberos 4, it was necessary to use the name “kerberos4”. It is not necessary for you to create an entry in the file for “kerberos4”, because the default port of 750/udp will be used if no “kerberos4” entry is found. If you are not using AFS Version 3.4 authentication servers, you should only have to modify **/etc/services** if you are using some other service that uses one or more of the traditionally used (but not formally reserved) Kerberos 4 ports. They are:

Service name: kerberos4 Port: 750/udp

Service name: kerberos_admin Port: 751/tcp

Service name: krb_prop Port: 754/tcp

You will also have to modify /**etc**/**services** if you are using AFS Version 3.4 authentication servers. The **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 assigned to port 88 in the file distributed with the base operating system. MIT Kerberos Version 4, on which PSSP authentication services are based, uses a default port number of 750. PSSP authentication commands use the service name "kerberos4" to avoid this conflict with the Kerberos 5 service name, which is also used by DCE. For PSSP authentication commands to communicate with an AFS 3.4 **kaserver**, you must do either of the following:

1. Stop the **kaserver**, redefine the udp port number for the "kerberos" service to 750 on the server system, then restart the **kaserver**.
2. 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.

/etc/krb.conf: The /**etc**/**krb.conf** file contains the name of the local realm for your SP system and identifies the host names of all of the authentication servers for all Kerberos realms known to the local realm. For more information on the content of the file, refer to **krb.conf** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

If you configure your realm to use AFS authentication servers, this file is built for you automatically from the corresponding AFS configuration file **CellServDB**. When you use PSSP authentication servers, you create this file or one is generated by default by the **setup_authent** script that must run on each workstation after you install the PSSP files. You must provide the file if the workstation is a client workstation or a secondary authentication server. If it is the primary authentication server, you need to provide the file only if the system's host name contains no domain portion, following conventional network naming rules.

By convention, Kerberos uses the domain portion of a network interface name, converted to upper case, as the default realm to which the host belongs. If you want to define your own local realm name that does not follow convention, or if the host name of your primary server host has no domain portion, you must supply a file with your choice of local realm name in the first line. When using AFS authentication, the local AFS cell name, contained in the **ThisCell** file, converted to uppercase, becomes the local Kerberos realm name.

/etc/krb.realms: The /**etc**/**krb.realms** file is a file that maps network interface (host) names to realms. When Kerberos needs to determine what realm a network interface is assigned to, it looks in this file. If there is no entry for the host name or for the domain part of the host name, the default is a realm name equal to the domain portion of the host name, converted to upper case. Any network interface names on systems using SP authenticated services which have a domain portion which is not the same as the local realm name, except for case, must have an entry in this file.

Entries for all network interfaces on the SP nodes and on the control workstation that require this mapping in /**etc**/**krb.realms** are added automatically by either **setup_authent** initially or by the **setup_server** script. The file is kept identical on the control workstation and the nodes by the node installation customization process. There may be cases involving other work stations, however, where you will have to add entries to the file on one or more systems.

When you set up a secondary authentication server or set up a system as an authentication client system, you are instructed to copy the configuration files from the primary server. If the new system requires mapping in **krb.realms**, the **setup_authent** script will create the entry in the local copy of the file. You will have to manually add the entry for the workstation you just configured into the control workstation's file and into the files on any other client and server workstations already installed.

Setting Up a User Account to Obtain Kerberos Credentials at AIX Login

As System Administrator, you can make using the RS/6000 SP authenticated services easier for nonroot users by setting up their user accounts so that Kerberos credentials are obtained whenever they login to the system. To do this, you must:

1. Add a stanza to the **/etc/security/login.cfg** file that names a program that will perform the required **kinit** processing on the user's behalf. The name of the method (shown in the following example as KERBEROS) can be anything you want. The rest of the stanza should be entered exactly as shown:

   ```
   KERBEROS:
      program = "/usr/bin/ksh /usr/kerberos/etc/Kinit"
   ```

2. Use SMIT to update either the **auth1** or **auth2** attribute in the user account for each user you want to use this capability. Add this method after the default **SYSTEM** method in the **auth1** attribute, if you want to require the user to obtain Kerberos credentials in order to successfully login to the system. Specify this method separately in the **auth2** attribute if you want to allow (but not require) the user to obtain Kerberos credentials at AIX login. The *Kerberos-principal* must be the user's AIX login name, optionally followed by a period and an instance (for example, **foo** or **foo.admin**). Note the placement of the comma (,) and the semicolon (;) in this rather unusual syntax.

   The following is an example of an auth1 attribute for AIX user **foo**. When **foo** logs in, the **kinit foo** command will be executed after the AIX password for principal **foo**. If **kinit** fails, **foo** is not logged into the system.

   ```
   auth1 = "SYSTEM,KERBEROS;foo"
   ```

   The following is an example of an **auth2** attribute for AIX user **bar**. When **bar** logs in, the **kinit bar.admin** command will be executed after the AIX password is validated. The user is prompted for the Kerberos password for **bar.admin**. In this case, if **kinit** fails, **bar** remains logged into the AIX system, but has no Kerberos credentials.

   ```
   auth2 = "KERBEROS;bar.admin"
   ```

Setting Up the Primary Authentication Server

The very first system that you set up, when you install your PSSP software, is your primary authentication server. That may be your control workstation; or it could be some other IBM RS/6000 workstation. You may want to have your servers entirely off of the SP system itself, to provide greater physical security for them. You may want to allow logins to the control workstation that are not appropriate for the authentication servers.

The **setup_authent** script is used to configure your authentication services. If you have installed the **ssp.authent** fileset on the workstation prior to running the script, it assumes that the system will be a PSSP authentication server. If you provide your own **krb.conf** file, the script will look for an entry for the local host name. If it

cannot find one, it allows you to configure the system without any server (as an authentication client system).

setup_authent creates the service principals for local service instances and creates the server key file for them. It updates the **krb.realms** file as needed, and creates the root user's **.klogin** file to authorize the initial administrator principal to use remote commands.

The daemons which provide the authentication server (the KDC), and the database administration server, are started by adding entries to **/etc/inittab**.

The procedure you must follow to set up the primary authentication server is as follows:

1. Install **ssp.clients** and **ssp.authent**.
2. Create a **/etc/krb.conf** file if you want or if your host names do not follow Kerberos convention.
3. Execute the **/usr/lpp/ssp/bin/setup_authent** program and follow the instructions in the prompts.

Setting Up Secondary Authentication Servers

The procedure you must follow to set up a secondary authentication server is as follows:

1. Install **ssp.clients** and **ssp.authent**.
2. Copy the **/etc/krb.conf** file from the primary authentication server to this secondary server system.
3. Add a line to the **/etc/krb.conf** file listing this system as a secondary server for the local realm.
4. Copy the **/etc/krb.realms** file from the primary server to the secondary server system.
5. Execute the **/usr/lpp/ssp/bin/setup_authent** program and follow the instructions when prompted.
6. If **setup_authent** created an entry for the local **/etc/krb.realms** file, copy the file to the other systems.
7. Add an entry for the new secondary server to the **/etc/krb.conf** file on the systems on which you had previously initialized authentication.
8. On the primary server, if this is the first secondary server, you should create a root **cron** entry that invokes the script **/usr/kerberos/etc/push-kprop** to periodically propagate database changes.

These steps are the same as those you would have followed to set up your SP control workstation as a secondary authentication server. For a more detailed description of how to do this, see *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

Setting Up Authentication Client Systems

The procedure you must follow to set up an authentication client system is as follows:

1. Install **ssp.clients**.
2. Copy the **/etc/krb.conf** file from the primary authentication server to this system.
3. Copy the **/etc/krb.realms** file from the primary server to this system.

   Note: If the new workstation is outside the realm of the primary server, the new workstation needs to be added to the primary server's **/etc/krb.realms** file first, before copying the **/etc/krb.realms** over to the new workstation. Otherwise, the next step, which runs the **setup_authent** program, will fail on the new client workstation.
4. Execute the **/usr/lpp/ssp/bin/setup_authent** program and follow the instructions when prompted.
5. If **setup_authent** created an entry for the local **/etc/krb.realms** file, copy the file to the other systems.

These steps are the same as those you would have followed to set up your SP control workstation as an authentication client system. For a more detailed description of how to do this, see *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

Directory Pathnames for SP Authentication Services

The following are some of the directory paths that are used with authentication services.

/usr/lpp/ssp/bin
: Contains **setup_authent**

/usr/lpp/ssp/kerberos/bin
: Contains commands for users of authentication services and authentication database administrators.

/usr/kerberos/bin
: Contains a symbolic link for **/usr/lpp/ssp/kerberos/bin**

/usr/lpp/ssp/kerberos/etc
: Contains authentication daemons and commands used by root to maintain a local SP authentication database.

/usr/kerberos/etc
: Contains a symbolic link for **/usr/lpp/ssp/kerberos/etc**

/var/kerberos/database
: Contains the SP authentication database on the server

/var/adm/SPlogs/kerberos
: Contains error logs for the authentication servers

/etc Contains the files **krb.conf**, **krb.realms**, and **krb-srvtab**

/usr/lpp/ssp/rcmd/bin
: Contains the Kerberos-authenticated versions of **rsh** and **rcp**.

/usr/vice/etc Contains AFS cell information and utilities

/usr/afsws/etc Contains AFS executables

Managing the Authentication Database

Backing up the Authentication Database

An important task that the administrator of a PSSP authentication database should perform is to keep a current back-up of the database contents. The database utility commands that are available to the root user on the primary authentication database system include functions you can use for this purpose. If you have secondary authentication servers and follow the steps outlined for setting them up, you will automatically create a backup of the database every time the **cron** file entry runs that propogates the database to the secondary servers.

The **push-kprop** script, invoked daily or at some other interval from the root user's **cron** file, creates a backup file in ASCII format called **slavesave**. This file is created in the database directory, where it is accessible only to the root user. If you want to create a backup file without using **push-kprop** (for example, if you do not have secondary servers and want to avoid the mail messages **push-kprop** sends to root), just add a **cron** entry to invoke the following command instead:

```
kdb_util dump /var/kerberos/database/slavesave
```

The result will be a backup file with the same pathname that can be used with the **kdb_util load** command to recover from a lost or corrupted database, or viewed or printed to examine the database contents.

Changing the Kerberos Master Password

Changing the master password can enhance database security somewhat, but should not be substituted for adequate physical security. Changing the master password re-encrypts the database in the new key. It does not, however, change any of the Kerberos principals' passwords (keys) themselves. Therefore, it offers limited protection unless all (possibly compromised) passwords are changed also.

Changing the master key is one of the capabilities of the **kdb_util** command. See the command description in the reference manual. Note that the various **kdb_*** commands may only be issued by the root user, on a system configured as a PSSP authentication server. An idiosyncrasy of these commands is that they prompt only once for the new password (unlike normal password-changing protocol).

After you change the password, be sure to issue the **kstash** command to store the new key in the /**.k** file. You'll also have to kill and respawn the Kerberos daemons on the server (**kerberos** and **kadmind**). If you have secondary servers, you will have to force re-propagation of the database and respawning of the backup servers also.

The sequence that should be executed on the primary server is:

1. Login to Kerberos as the admin principal used to set up authentication. For example:

   ```
   /usr/kerberos/bin/kinit root.admin
   ```

2. Change the password (you will be prompted for old and new).

   ```
   /usr/kerberos/etc/kdb_util new_master_key /var/kerberos/database/newdb.$$
   /usr/kerberos/etc/kdb_util load /var/kerberos/database/newdb.$$
   ```

3. Replace the /**.k** file (another prompt for the new password).

```
/usr/kerberos/etc/kstash
```

4. Kill and respawn the primary authentication server daemons.

```
kill 'ps -e | egrep 'kerberos|kadmind' | cut -d'-' -f1'
```

5. If secondary servers exist, propagate the new database, copy the new master key cache to the secondary servers, and kill and respawn the secondary server daemons.

```
if [[-s /var/kerberos/database/slavelist]]
then
  /usr/kerberos/etc/push-kprop
 cat /var/kerberos/database/slavelist | while read slave
 do
    /usr/lpp/ssp/rcmd/bin/rcp /.k $slave:/
   /usr/lpp/ssp/rcmd/bin/rsh $slave \
        kill \ ps -e \| grep kerberos \| cut -d'-' -f1\
  done
fi
```

Managing Authentication Principals

Listing Principals in the Authentication Database

With PSSP Authentication: Use the **lskp** command to list all or some principals in the local authentication database. Information displayed for each principal includes the name and instance, the maximum ticket lifetime, the key version number, and the expiration date (local time on the server).

lskp allows you to select principals to list in several different ways:

- The entire authentication database
- Individual entries by name
- Entries with a specific name
- Entries with a specific instance
- Client (user) principals
- Service principals
- Principals pre-defined by Kerberos

The following example lists all principals who are Kerberos administrators plus all instances of the **rcmd** service principal, used by the SP remote commands and the Sysctl facility:

```
#lskp .admin rcmd.

mary.admin tkt-life: 30d key-vers: 3 expires: 2000-12-31 23:59
root.admin tkt-life: 30d key-vers: 5 expires: 2004-05-31 23:59
rcmd.cwsta tkt-life: Unlimited key-vers: 1 expires: 2037-12-31 23:59
rcmd.node1 tkt-life: Unlimited key-vers: 1 expires: 2037-12-31 23:59
rcmd.node2 tkt-life: Unlimited key-vers: 1 expires: 2037-12-31 23:59
rcmd.node3 tkt-life: Unlimited key-vers: 1 expires: 2037-12-31 23:59

#
```

When logged into a system that is a Kerberos authentication server, the root user can run the **lskp** command directly. Additionally, any users who are Kerberos

database administrators listed in the **/var/kerberos/database/admin_acl.get** file can invoke this command remotely through a Sysctl procedure of the same name. The administrator need not be logged in on the server host to run **lskp** through Sysctl, but must have a Kerberos ticket for that **admin** principal (*name*.**admin**).

In the following example, **control2** is the Kerberos server host name. The user has previously obtained a ticket as *name*.**admin**, for whom an entry exists in the **admin_acl.get** file:

```
$sysctl -h control2 lskp default

default tkt-life: 30d key-vers: 1 expires: 2037-12-31 23:59

$
```

For more information, refer to **lskp** in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

With AFS Authentication: To list the authentication database information, an AFS administrator uses the **kas list** command. It lists all entries in the database. For example (other entries have similar format):

```
$kas list -l -admin_user afsadm

Administrator's (afsadm) Password:
User data for afsadm (ADMIN)
   key (0) cksum is 1479418564, last cpw: no date
   entry never expires. Max ticket lifetime 720.00 hours.
   last mod on Fri Jun 28 09:12:56 1996 by afsadm
User data for joseph
   key (0) cksum is 1852746310, last cpw: Mon Mar 25 14:45:00 1996
   entry never expires. Max ticket lifetime 720.00 hours.
   last mod on Fri Jun 28 09:15:09 1996 by afsadm
...

$
```

Displaying Information about Principals

With PSSP Authentication: There are two ways for a PSSP authentication database administrator to examine information about principals. The **kadmin** command offers any administrator listed in the **admin_acl.get** file access from any network-connected system running SP authentication services. The root user on the primary authentication database system can also directly examine the database content by viewing the database backup file in **/var/kerberos/database/slavesave**.

Using **kadmin**, for example:

```
kadmin
Welcome to the Kerberos Administration Program, version 2
Type "help" if you need it.
admin: get ben
Admin password:
Info in Database for ben.:
   Max Life: 255   Exp Date: Fri Dec 31 23:59:00 1999

   Attribs:  00  key: 0 0
admin: quit
Cleaning up and exiting
```

Using the database backup file:

```
grep "^rcmd node3sw" /var/kerberos/database/slavesave
rcmd node3sw 255 1 2 0 a49bf286 d45c6560 200001010459 199503301502 root admin
```

The fields in each entry are:

- The principal's name
- The principal's instance
- The maximum ticket lifetime
- The key version of the master key used to encrypt the entry.
- The key version of the principal's secret key.
- The two halves of the encrypted private key.
- The expiration date (in Universal Coordinated Time) as YYYYMMDDHHMM
- The date the entry was created or modified
- The name and instance of the administrator who made the last update

With AFS Authentication: To display the database information for a user of AFS, an AFS administrator uses the **kas examine** command. For example:

```
kas ex -na ben -admin_user afsadm
Administrator's (afsadm) Password:

User data for ben
  key (0) cksum is 1004957312, last cpw: Mon May 22 10:12:50 1995
  password will never expire.
  An unlimited number of unsuccessful authentications is permitted.
  entry never expires.  Max ticket lifetime 25.00 hours.
  last mod on Mon May 22 10:12:50 1995 by afsadm
```

Adding Principals and Assigning Intitial Passwords

With PSSP Authentication: To add a principal and set the initial password in the authentication database, a database administrator with an **admin** principal that is listed in the **admin_acl.add** file, can use **kadmin** or **add_principal**.

To use **kadmin**:

1. Enter the **kadmin** command:

   ```
   kadmin
   ```

   A welcome message and explanation of how to ask for help are displayed.

2. Enter the **add_new_key** or **ank** subcommand:

   ank *name[.instance]*

   The only required argument for the subcommand is the principal's name.

3. At the prompt, enter your **admin** password

4. At the prompt, enter the principal's new password
5. At the prompt, reenter the principal's new password

For more information, refer to **kadmin** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

add_principal operates like **kadmin**, but allows an arbitrarily large number of principals to be added at one time. Whereas **kadmin** always prompts for your administrative password, **add_principal** will use tickets you have already obtained as *login-name*.**admin**. This makes **add_principal** suitable for use in scripts or background processes where a password cannot be provided interactively.

1. Create an ASCII file containing a list of principals and their initial passwords. Make sure it is readable only by you.
2. Enter the **add_principal** command.

   ```
   add_principal filename
   ```

3. Immediately remove the file containing the initial passwords.

The following script fragment shows one way to add a user to the database without interactively editing a file, and keeping the initial password on disk for a minimal amount of time:

```
#( OMASK=$(umask); umask 077; INITIALPW=changeme$RANDOM; \
   print "johndoe $INITIALPW">/tmp/Addp$$; \
   /usr/kerberos/bin/add_principal /tmp/Addp$$; \
   /bin/rm /tmp/Addp$$; umask $OMASK; \
   print "johndoe's initial password is $INITIALPW"; )
johndoe's initial password is changeme28745
#
```

On the system which is the primary authentication server, the root user can use the **kdb_edit** command to add principals. The steps are:

1. Enter the command, which operates interactively, prompting for requests.

   ```
   kdb_edit -n
   ```

2. Reply to the prompt for a principal name by entering the user name of the new principal and then the instance.

   ```
   Principal name: joeuser
   Instance:
   ```

3. Reply **y** to confirm that you want to add the pricipal.
4. Enter the initial password, when prompted.
5. Reenter the password, when prompted.
6. Accept the default expiration date by pressing Enter or type an alternate date in the format shown by the prompt.
7. Accept the default maximum ticket lifetime value or enter another.
8. Press enter when prompted for attributes.
9. The command will respond Edit OK and begin prompting for another principal to add. Repeat the same sequence or press enter to terminate the command.

For a complete example and more information, refer to **kdb_edit** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

You can use the **mkkp** command to add Kerberos principals to the authentication database instead of **kdb_edit**.

When logged into a system that is a Kerberos authentication server, the root user can run the **mkkp** command directly. Additionally, any users who are Kerberos database administrators listed in the /**var**/**kerberos**/**database**/**admin_acl.add** file can invoke this command remotely through a Sysctl procedure of the same name. The administrator need not be logged in on the server host to run **mkkp** through Sysctl, but must have a Kerberos ticket for that **admin** principal (*name*.**admin**).

Like **kdb_edit** (and unlike the **kadmin** command), **mkkp** can set attributes other than the password. Because it is not interactive, however, **mkkp** cannot set an initial password. To complete that step, so users can **kinit** as the new principals, an administrator should use the **cpw** sub-command of **kadmin**.

In the following example, these steps are used to add two new principals and then assign passwords to enable Kerberos logins. **control2** is the Kerberos server host name. The user has previously obtained a ticket as *name*.**admin**, for whom an entry exists in the **admin_acl.add** file:

```
$sysctl -h control2 mkkp -e 1998-06-30 jack jill

$kadmin -m

Welcome to the Kerberos Administration Program, version 2
Type "help" if you need it.
admin: cpw jack
Admin password:
New password for jack:
Verifying, please re-enter New password for jack:
Password changed for jack.
admin: cpw jill
New password for jill:
Verifying, please re-enter New password for jill:
Password changed for jill.

admin: quit

Cleaning up and exiting.

$
```

For more information, refer to **mkkp** in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

With AFS Authentication: To add a principal and set a password in an AFS authentication database:

1. Enter the **kas** command:

   ```
   kas create -name user [-cell cellname]
   ```

2. At the prompt, enter your AFS password
3. At the prompt, enter the principal's new password
4. At the prompt, reenter the principal's new password to verify

Adding an Authentication Administrator

With PSSP Authentication: To perform the tasks involved in administering SP authentication services, the system administrator must be authorized to do so. The administrator must be a user with an **admin** instance whose name appears on one of the authentication administration access control lists (ACLs).

For example, to perform authentication administration, the administrator **gail** must have a principal **gail.admin** in the authentication database. In addition, **gail.admin** must be in the access control list file for the task she wants to perform. There are three ACL files:

/var/kerberos/database/admin_acl.add
: The list of principals authorized to add entries to the authentication database

/var/kerberos/database/admin_acl.get
: The list of principals authorized to retrieve entries from the authentication database

/var/kerberos/database/admin_acl.mod
: The list of principals authorized to modify entries in the authentication database

Adding an authentication database administrator requires two steps:

1. Adding the principal with the **admin** instance, as described above.
2. Adding the principal identifier to one or more of the ACL files, which must be done by the root user on the primary database system.

With AFS Authentication: AFS administration does not use an **admin** instance or ACL files. You, an AFS administrator, can authorize another AFS user to perform administrative tasks by assigning it the **admin** attribute:

1. Follow the steps described above to add the user, if not already defined.
2. Enter the **kas** command:

   ```
   kas setfields -name user -flags admin [-cell cell]
   ```

3. At the prompt, enter your AFS password

Changing a Principal's Password and other Attributes

With PSSP Authentication: To change a password for a principal in the authentication database, a PSSP authentication database administrator can use either the **kpasswd** command or the **kadmin** program's **change_password** subcommand. These commands can be issued from any system running SP authentication services and do not require a prior **kinit**.

To use the **kpasswd** command:

1. Enter the **kpasswd** command with the name of the principal whose password is being changed:

   ```
   kpasswd -n name
   ```

2. At the prompt, enter the old password
3. At the prompt, enter the new password
4. At the prompt, reenter the new password

To use the **kadmin** program:

1. Enter the **kadmin** command:

   ```
   kadmin
   ```

   A welcome message and explanation of how to ask for help are displayed.
2. Enter the **change_password** or **cpw** subcommand with the name of the principal whose password is being changed:

   ```
   cpw name
   ```

   The only required argument for the subcommand is the principal's name.
3. At the prompt, enter your **admin** password
4. At the prompt, enter the principal's new password
5. At the prompt, reenter the principal's new password

For more information, refer to **kpasswd** and **kadmin** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

To change your own **admin** instance password, you can use either the **kpasswd** command or the **kadmin** program's **change_admin_password** subcommand.

To use the **kpasswd** command:

1. Enter the **kpasswd** command with your **admin** instance name:

   ```
   kpasswd -n name.admin
   ```
2. At the prompt, enter your old **admin** password
3. At the prompt, enter your new **admin** password
4. At the prompt, reenter your new **admin** password

To use the **kadmin** program:

1. Enter the **kadmin** command:

   ```
   kadmin
   ```

   A welcome message and explanation of how to ask for help are displayed.
2. Enter the **change_admin_password** or **cap** subcommand:

   ```
   cap
   ```
3. At the prompt, enter your old **admin** password
4. At the prompt, enter your new **admin** password
5. At the prompt, reenter your new **admin** password

For more information, refer to **kpasswd** and **kadmin** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

In addition to changing the password, you may want to change either the expiration date of the principal or its maximum ticket lifetime, though these are not so likely to be necessary. To do so, the root user on the primary authentication database system must use the **kdb_edit** command, just as when adding new principals locally. Instead of not finding the specified principal, the command finds it already exists, and prompts for changes to all its attributes, starting with the password, followed by the expiration date and maximum ticket lifetime.

The following example shows the use of the **kdb_edit** command to change a user principal's maximum ticket lifetime from the default value (30 days) to approximately 7 days:

```
#/usr/kerberos/etc/kdb_edit -n
Opening database...
Previous or default values are in [brackets] ,
enter return to leave the same, or new value.

Principal name: johndoe
Instance:

Principal: johndoe, Instance: , kdc_key_ver: 1
Change password [n] ?
Expiration date (enter yyyy-mm-dd) [ 1996-09-01 ] ?
Max ticket lifetime [ 255 ] ? 171
Attributes [ 0 ] ?
Edit O.K.
Principal name:
#
```

You can use the **chkp** command to change the maximum ticket lifetime and expiration date for Kerberos principals in the authentication database.

When logged into a system that is a Kerberos authentication server, the root user can run the **chkp** command directly. Additionally, any users who are Kerberos database administrators listed in the **/var/kerberos/database/admin_acl.mod** file can invoke this command remotely through a Sysctl procedure of the same name. The administrator need not be logged in on the server host to run **chkp** through Sysctl, but must have a Kerberos ticket for that **admin** principal (*name*.**admin**).

In the following example, **chkp** is used by root to set the expiration date to the end of the year 2000 and the lifetime to approximately two weeks for several existing principals:

```
#chkp -e 2000-12-31 -l 181 joeuser joe.admin harry.admin user12

#
```

In the following example, the maximum ticket lifetime for new users (those added subsequently) is set to approximately a week. **control2** is the Kerberos server host name. The user has previously obtained a ticket as *name*.**admin**, for whom an entry exists in the **admin_acl.mod** file:

```
$sysctl -h control2

Sysctl (Version 1.1) on control2.abc.com
sysctl> chkp -l 170 default
sysctl> lskp default
default tkt-life: 7d+08:50 key-vers: 1 expires: 2037-12-31 23:59
sysctl> quit

$
```

For more information, refer to **chkp** in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

With AFS Authentication: To change the attributes of AFS user principals, the administrator uses the **kas setfields** command. For example, to change the maximum ticket lifetime to two days:

```
kas setfields -name ben -lifetime 48:00 -admin_user afsadm
administrator's (afsadm) Password:
```

Deleting Principals

With PSSP Authentication: You can use the **rmkp** command to remove Kerberos principals from the authentication database.

When logged into a system that is a Kerberos authentication server, the root user can run the **rmkp** command directly. Additionally, any users who are Kerberos database administrators listed in the **/var/kerberos/database/admin_acl.add** file can invoke this command remotely through a Sysctl procedure of the same name. The administrator need not be logged in on the server host to run **rmkp** through Sysctl, but must have a Kerberos ticket for that **admin** principal (*name*.**admin**).

Normally, this command requires you to confirm your intention to remove each principal as it is processed. An option is provided (**-n**) to allow the command to suppress prompting and run non-interactively. Running **rmkp** through Sysctl requires non-interactive execution; and doing so forces that option to be applied. When running **rmkp** interactively, the root user can select multiple principals by name or instance using the operand notation *name.* or *.instance* as for the **lskp** command.

Removed entries are preserved in an ASCII text file, in the format produced by the **kdb_util dump** command. If a principal must be recovered after having been removed, the line in that file containing its entry can be appended to a current database backup file (see “Backing up the Authentication Database” on page 179) to be restored using **kdb_util load**. The following example shows how **rmkp** could be used by root to remove obsolete entries after the host name on the control workstation has been changed from **cwksta** to **cwksta2**:

```
#rmkp -v .cwksta

Confirm removal of principal hardmon.cwksta? (y or n): y
hardmon.cwksta was removed
Confirm removal of principal rcmd.cwksta? (y or n): y
rcmd.cwksta was removed
Removed entries were saved in /var/kerberos/database/rmkp.save.3036

#
```

In the following example **rmkp** is used to remove two principals. **control2** is the Kerberos server host name. The user has previously obtained a ticket as *name*.**admin**, for whom an entry exists in the **admin_acl.add** file:

```
$sysctl -h control2 rmkp -v pauline frank

pauline was removed
frank was removed
Removed entries were saved in /var/kerberos/database/rmkp.save.18365

$
```

For more information, refer to **rmkp** in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

With AFS Authentication: An AFS administrator can delete principals directly by using the **kas delete** command. For example:

```
kas delete -name ben -admin_user afsadm
Administrator's (afsadm) Password:
```

Managing Service Keys

Management of principals centers primarily on changing passwords in accordance with local security policy. For user principals this task is straightforward and can be done using simple tools. For service principals, password maintenance requires more than periodic changes to the authentication database. It also requires creation and distribution of files containing encrypted password information to the application server hosts.

Creation of Service Key Files

During setup for a node or control workstation, the keys for service principals are stored in the authentication database (for use by the authentication server) and on the node corresponding to the instance in the file **/etc/krb-srvtab** (for use by the service itself). Therefore, every node in the SP System contains a file, **/etc/krb-srvtab**, that contains the keys for the services that are provided on that node. On the control workstation, the **hardmon** and **rcmd** service principals are in the file. On the nodes, the **rcmd** service principals are in the file.

The local server key files on the control workstation and on other workstations are created by the **setup_authent** command when authentication is first set up on the system.

With PSSP Authentication: The **setup_server** script creates server key files for the SP nodes. When the control workstation is a PSSP authentication server, this occurs for each node when the script is run on the node's boot server. **setup_server** creates the files in the **tftpboot** directory, readable only by their owner **nobody**. When network boot occurs during install or customization, the file is transferred to the node using **tftp**. The node uses the file to obtain tickets and invokes authenticated **rsh** to execute a script on the boot server that deletes the file.

With AFS Authentication or No Server on the Control Workstation: The **setup_server** script creates server key files for the SP node, when the script is run on the control workstation for the first time. **setup_server** creates the files in the **tftpboot** directory, readable only by their owner **root**. The file is copied to the node's boot server and ownership is changed to **nobody**, when the script runs on the boot server. When network boot occurs during install or customization, the file is transferred to the node using **tftp**. The node uses the file to obtain tickets and invokes authenticated **rsh** to execute a script on the boot server that deletes the file (gives it back to root, if the boot server is the control workstation). In this case the copy on the control workstation is never deleted.

Examining Service Key Information

There are two programs that you can use to examine the content of a server key file on the local system, **ksrvutil** and **klist**: To view the local system's server key file in **/etc/krb-srvtab**, enter **ksrvutil list**. The output will look like:

```
Version    Principal
   1     rcmd.cwktr@XYZ.ABC.COM
   1     hardmon.cwktr@XYZ.ABC.COM
   1     rcmd.cwkfddi@XYZ.ABC.COM
   1     hardmon.cwkfddi@XYZ.ABC.COM
   1     rcmd.cwken0@XYZ.ABC.COM
   1     hardmon.cwken0@XYZ.ABC.COM
```

The same file, viewed by entering **klist -srvtab** appears as:

```
Server key file:   /etc/krb-srvtab
Service         Instance        Realm      Key Version
-----------------------------------------------------
rcmd            cwktr           XYZ.ABC.COM 1
hardmon         cwktr           XYZ.ABC.COM 1
rcmd            cwkfddi         XYZ.ABC.COM 1
hardmon         cwkfddi         XYZ.ABC.COM 1
rcmd            cwken0          XYZ.ABC.COM 1
hardmon         cwken0          XYZ.ABC.COM 1
```

To examine the server key file in **/tftpboot** directory for a node whose host name is **sp2node8**, you would enter **ksrvutil list -f /tftpboot/sp2node8-new-srvtab** and the output might look like:

```
1     rcmd.n08tr0@XYZ.ABC.COM
1     rcmd.n08tr1@XYZ.ABC.COM
1     rcmd.n08fddi@XYZ.ABC.COM
1     rcmd.n08css0@XYZ.ABC.COM
1     rcmd.n08en0@XYZ.ABC.COM
```

Changing Service Keys

Changing the service keys involves two responsibilities:

- Changing the service keys in accordance with local system policy
- Ensuring that the service keys contained in the authentication database and in the server key files on all workstations and SP nodes always match.

As security administrator, you decide how frequently service keys need to be changed on your system. You may decide to set the service keys at setup time and never change them or you may decide to change them every six months. You should certainly change the service keys if you have reason to believe system security has been compromised.

Use the **ksrvutil** command to change service keys. The **ksrvutil** command changes all of the service keys in the local service key file. The keys for all of the services provided on a single node, or the control workstation, are changed in a single call to **ksrvutil**. The keys are set to randomly generated values. The method of running **ksrvutil** to change the service keys for a node varies according to the authentication configuration of the system:

Primary Authentication Server on the Control Workstation: The **ksrvutil** command should be run on each node where the keys need to be changed. To change all of the service keys for a node to randomly generated values, enter:

```
ksrvutil change
```

The keys are updated in the node's /**etc**/**krb-srvtab** file and in the authentication database on the control workstation. To issue the command, you must be root. The **dsh** command can be used to issue the **ksrvutil** command on multiple nodes.

The **ksrvutil** command should also be used to change service keys on the control workstation, regardless of the location of the primary authentication database.

All Other Authentication Configurations: If the primary authentication database does not reside on the SP system's control workstation, the **ksrvutil** command should *never* be used on the nodes of the SP System. If the primary authentication database for the local realm resides in some location other than the SP system's control workstation (even if a secondary authentication database is on the control workstation) or if the system is using AFS-based authentication services, the system maintains a copy of each node's /**etc**/**krb-srvtab** file on the control workstation. These files are in the directory /**tftpboot** with a filename of *nodename*-**new-srvtab**. Changes to the **srvtab** file should be made only to the copy on the control workstation in /**tftpboot**. The **ksrvutil** command is used to change the service keys.

To change all of the service keys for a node:

1. Ensure that no users are using authenticated services on the target node
2. On the control workstation, issue:

   ```
   ksrvutil -f nodename-new-srvtab change
   ```

3. Copy the *nodename*-**new-srvtab** file from the control workstation to the /**etc**/**krb-srvtab** file on the target node. To do this, you can use **ftp** or, as an alternative, you can customize the node using the **spbootins** command, followed by a **netboot** on the node.

To change the service keys for a node, you must have read/write permission to the /**tftpboot**/*nodename*-**new-srvtab** file and have write access to the /**etc**/**krb-srvtab** file on the node.

When you change the service keys for a node, or for the control workstation, any service tickets that have been previously issued to users for those service instances are no longer valid. An attempt to use such a ticket to authenticate the user's identity to the server will fail. A message will be issued that the server could not decode the ticket authenticator. In such a case, the user must issue a **kinit** command to obtain a new ticket-granting ticket and then reissue the service request.

Refer to **ksrvutil** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information. Also, refer to the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* for information on recreating server key files.

Distributing Service Key Files to Client Systems

Normal distribution of server key files to SP nodes occurs as described using **tftp**. This entails some risk, but it is reduced by not making the files world-readable, as is usually the case for files transferred in that way. Probably the only secure method is to log into the destination system, use **kinit** to get a ticket, and then invoke Kerberos-authenticated **rcp** to pull the file from the boot server. That program cannot be used to push the file in the other direction because the remote

command daemon on the target system needs a valid server key file to perform authentication.

Another means of copying the file that offers some security is the use of the standard AIX **rcp** command, requiring the creation, temporarily, of a root **.rhosts** file.

Less secure yet, and probably unacceptably so, are **ftp**. Unless it is done over a secure (internal) network interface.

Stopping and Restarting PSSP Kerberos Daemons

If a Kerberos V4 daemon fails, it records an error condition in its log file and stops responding to client requests. Under these circumstances, you would need to restart it, after correcting the problem which caused the failure. There can be other circumstances, such as application of certain maintenance, where you might also have to stop and then restart one or more of the daemons that provide Kerberos services. Each host that is an authentication server has two Kerberos daemons that are started when the system is booted. The primary authentication server runs the **kerberos** daemon to provide ticket-granting services and the **kadmind** daemon to provide authentication database administration. Each secondary authentication server host also runs **kerberos**, and also the **kpropd** daemon, which is the "catcher" for database updates propagated from the primary server, usually via a cron job.

Authentication Daemons (Primary or Secondary Servers)

To stop the Kerberos authentication server, enter the command:

```
stopsrc -s kerberos
```

To restart the Kerberos authentication server, enter the command:

```
startsrc -s kerberos
```

Administration Daemons (Primary Servers Only)

To stop the Kerberos database administration server, enter the command:

```
stopsrc -s kadmind
```

To restart the Kerberos database administration server, enter the command:

```
startsrc -s kadmind
```

Database Propagation Daemons (Secondary Servers Only)

To stop the Kerberos database propagation server, enter the command:

```
stopsrc -s kpropd
```

To restart the Kerberos database propagation server, enter the command:

```
startsrc -s kpropd
```

Using AFS Authentication Services

AFS is a distributed file system developed by Transarc Corporation. AFS includes an authentication implementation that is based on Kerberos version 4. Therefore, AFS authentication servers can be used in place of SP authentication servers to provide credentials for SP principals.

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

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

Administering AFS Authentication Services

The primary tasks the system administrator performs in administering AFS authentication services are:

- Adding principals (users and services) and setting passwords in the authentication database
- Changing passwords for principals
- Managing the authentication keys for the set of services that use authentication (refer to “Managing Service Keys” on page 189 for more information)

The AFS system has its own administration commands and utilities for the administrator user. These commands and utilities provide for authentication, protection of files, file manipulation, and backup services. The following commands and utilities are relevant for administering authentication services. Refer to the AFS documentation for more information.

afsd	Daemon used to connect AFS clients to AFS server
cacheinfo	Specifies the directory to place the AFS directory on an AFS client
CellServDB	Specifies the name of AFS cell and location of AFS server
kas	Administrator interface to work with AFS authentication
klog.krb	User interface to obtain a Kerberos ticket in addition to AFS tokens
pts	Administrator interface to work with AFS protection services
token.krb	Allows the user to list the current tokens and Kerberos tickets

The following SP authenticaton services commands work in an AFS authentication system:

- **kinit** (The SP_NAME environment variable must be set in order to use this command on a client workstation, when using AFS authentication)
- **klist**
- **kdestroy**
- **ksrvutil -afs change**
- **ksrvtgt**
- **rcmdtgt**
- **setup_authent**

The SP authentication services provide only the interfaces from the SP System to work in an AFS authentication system. The SP authentication services do not supply any of the AFS executables.

The AFS system provides multiple daemons that are activated on AFS server machines and support authentication tasks:

bosserver Basic OverSeer server monitors the AFS server processes on the AFS server machine

kaserver The AFS authentication server

ptserver Used to create system entries in the authentication database

buserver Maintains a backup authentication database on the authentication server

The authentication services work primarily with the **kaserver** and the **ptserver**. The **kaserver** allows the administrator to execute various commands to administer the authentication database. The **admin** user can enter **kas help** to view the commands available. The **admin** user needs to provide a password when using the **kas** command.

The **ptserver** protects and maintains system entries that are stored in the authentication database. The **admin** user can enter **pts help** to view the commands available. The **admin** user needs to provide a password when using the **pts** command. The admin user needs to create an AFS token (**klog.krb**) when working with the **pts** command.

Refer to AFS documentation for information on installing and administering the AFS system.

Chapter 15. Sysctl

Sysctl is an authenticated client/server system for running commands remotely and in parallel. It provides:

- **Least privilege capability**

  Root authority can be dynamically provided to non-root users based on their authenticated identity, the task they are trying to perform, access control lists, and any other relevant criteria. The root password need not be given out to as many people, and thus is kept secure.

- **Distributed execution**

  Sysctl applications can be executed on remote hosts with full authentication and authorization. It is essentially a secure, easy-to-program remote command execution mechanism for AIX commands, scripts and programs.

- **Parallel execution**

  Sysctl applications can be efficiently executed in parallel on many hosts.

- **Programmability**

  Sysctl applications can be coded as scripts for ease of implementation.

For example, consider a systems management script that contains commands that are only executable by root. If implemented as a Sysctl application, its security, usability, and performance are enhanced. Least privilege capability lets one of a select group of administrators run the script on each node in the SP without having access to the root password on any node in the system. With distributed execution, the administrator does not have to log in to each node to run the script. With parallel execution, the script can be run on all the nodes simultaneously with one command invocation. With programmability, the task can be automated or simplified with scripting.

Sysctl Components

Sysctl contains the following components:

- A server daemon (**sysctld**) that runs on all nodes and the control workstation.

  This server daemon has root privileges. It executes commands as root on behalf of authorized clients.

- Commands built into the server.

  These do the actual system administration work. Each command is paired with an authorization-checking script (callback) that has access to authenticated information about the client attempting to run the command. These commands include IBM-provided commands and typically also include installation-written ones.

- Configuration files that control aspects of the server operation as well as extend the command set available on a given node.

- Access control lists that list authorized users.

- A client program (**sysctl**) that offers command-line and interactive interfaces for communicating with **sysctld** servers.

© Copyright IBM Corp. 1995, 1997

Figure 13 on page 196 illustrates the way in which the Sysctl elements work together.

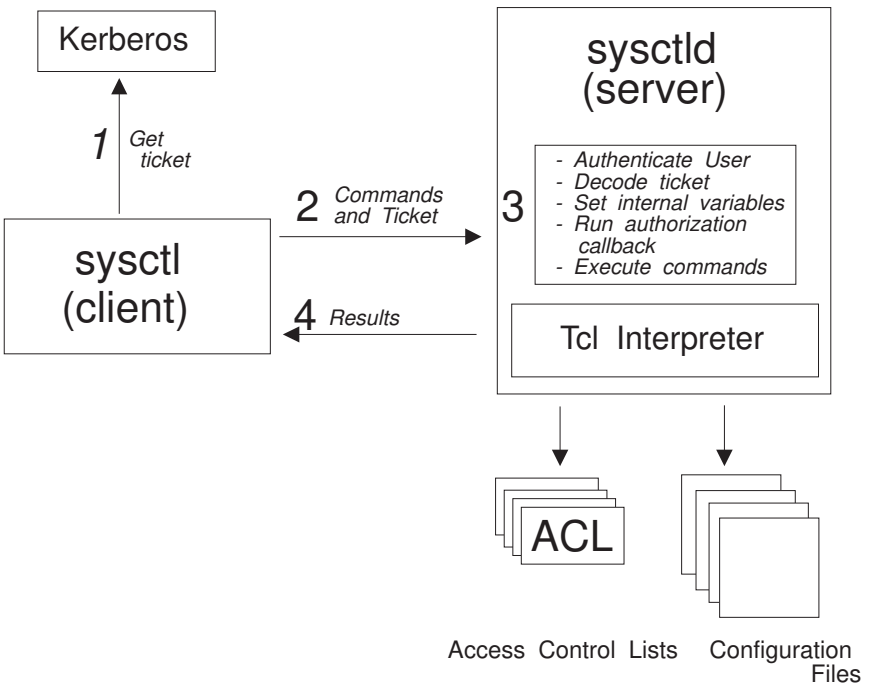


Figure 13. Sysctl Elements

The interpretation of a Sysctl command starts when the user issues a Sysctl client command. This can be embedded in a script or issued directly from the command line. It is of the following form:

```
sysctl target_nodes_specification command args
```

Note that more than one node can be targeted. The command is the name of a Sysctl server command to be run as root on each of the targeted servers, if the client user is authorized.

1. The Sysctl client code obtains authentication information about the command issuer from SP Authentication Services.
2. The Sysctl client code sends the authentication information, along with the command string, to each of the specified Sysctl servers in parallel.
3. Each server then performs the following steps:
   a. The authentication information is decoded and the authenticated information about the client is obtained.
   b. Internal variables, such as the username and host of the client are set. These variables are available to the server routines that will check the authorization of the client (the authorization callbacks) as well as the Sysctl commands to be run.
   c. The authorization is checked via an authorization callback. This can be a supplied callback, or an installation-written one. This callback may check an ACL. If the client passes the authentication check, the next step is done.
   d. The server commands requested are executed as root, on behalf of the client.

4. **stdout** and **stderr** are sent back to the client. Output from each server is displayed with labeling indicating its source.

Relationship to the SP Authentication Services

Sysctl uses the SP Authentication Services. Client identity is authenticated across the network to the server using these services. Users of Sysctl typically need to identify themselves to the SP Authentication Services via the **kinit** command.

See Chapter 14, "Security Features of the SP System" on page 157 for more information on using Authentication Services.

Overview

Architecture

A Sysctl server provides remote access to a set of commands for authenticated and authorized clients. Since the Sysctl server is running with UID 0, these commands are executed with root privilege. These commands are embedded in an interpreter within the server. Sysctl uses the embeddable Tool Command Language (Tcl) as the foundation for its built-in command interpreter. The commands are of the following types:

- Base Tcl commands.

  These are the commands making up the Tcl language. They include control commands (such as **if**), input/output commands (such as **puts**), arithmetic commands (such as **expr**) and commands allowing the execution of other commands, including external AIX commands and scripts (**exec**).

- Built-in Sysctl commands.

  These are provided for you. They include commands to configure and manage Sysctl, as well as some systems management applications. See "Built-in Sysctl Commands" on page 216.

- Your applications.

The Tcl interpreter that Sysctl uses is built from base Tcl and Extended Tcl. Refer to the **ssp.public.README** file for more information on Tcl as supplied with the SP.

Using an embedded language, particularly Tcl, provides the following advantages:

- The command language interface is ideal for administrative tasks.
- The Sysctl command set is extended by writing Tcl scripts (instead of C code). These new Tcl scripts are embedded in the server's command set.
- Users can write simple Tcl wrappers for Perl and shell scripts and then add the scripts to the server command set.
- The Sysctl authentication mechanism gives the administrator complete control over the set of commands available to a given set of users. (see Chapter 14, "Security Features of the SP System" on page 157).

Security

The management tools used to administer a large site need to be secure to prevent outside intrusion and to prevent accidental disasters, such as inadvertently deleting critical data.

Sysctl security has the following components:

- Client authentication
- Authorization callbacks
- Access Control Lists

Client Authentication

The SP Authentication Service provides trusted third-party authentication of users. When a client issues a Sysctl command, a ticket identifying the client is sent along with the command. The server uses the ticket to verify the identity of the client.

Authorization Callbacks

The Sysctl authorization mechanism is based on callbacks. A callback is a Tcl routine paired with a Sysctl command that determines whether the client is authorized to run the command. The server invokes the authorization callback for a command just before executing the command. If the authorization callback returns a Tcl error, the command is not executed. When the callback returns a normal Tcl result, the command is executed.

With the callback mechanism, you can customize the server's authorization model to fit your needs. You can use the built-in authorization callbacks or write new ones, as needed.

Built-In Authorization Callbacks: These authorization callbacks are provided:

NONE No special authorization is required to run a command. Even clients with unauthenticated identity (clients not identified to the SP Authentication Services) can run commands with this callback.

AUTH Any client who has been successfully authenticated by the SP Authentication Service can run commands with this callback.

ACL This callback checks an ACL file for an authenticated client. If the client is listed in the ACL, the callback succeeds. The pathname of an ACL file can be specified as an argument. If no pathname is specified, the client's principal must be listed in the global Sysctl ACL file, /**etc**/**sysctl.acl**.

SYSTEM This level is used for commands that will only be run when authorization callbacks are bypassed (see below). This authorization callback always returns a Tcl error. A client cannot issue commands with this callback from the command line or the Sysctl prompt.

Bypassing Authorization Callbacks: In some situations, the server bypasses the authorization callbacks. Authorization callbacks are bypassed while the server:

- Reads configuration files (that is, when starting up)
- Executes an authorization callback
- Executes the body of a procedure that a user is authorized to run

Access Control Lists (ACLs)

An ACL is a text file that can be used to grant authority to issue particular commands to particular users. Sysctl provides ACL-based authorization. See "Sysctl Files" for more information.

Sysctl Executables

Sysctl Daemon (sysctld)

The Sysctl server daemon, **sysctld**, processes all Sysctl client requests for the node on which it runs. There is a **sysctld** daemon running on each node of the SP system, as well as on the control workstation. **sysctld** runs as **root** and executes Tcl procedures for authenticated and authorized clients who may be local or remote users. See the **sysctld** man page for more information.

Client (sysctl)

The Sysctl client shell program, **sysctl**, offers a command line interface for communicating with **sysctld**. Together, **sysctl** and **sysctld** provide monitoring and execution abilities needed to remotely manage SP nodes. **sysctl** connects to a remote node's **sysctld** using TCP/IP, passes keywords and commands to the server, and writes any output returned to **stdout** and **stderr**. See the **sysctl** man page for more information.

Sysctl Files

Sysctl Configuration File

A configuration file is read in and interpreted each time the Sysctl server is started or restarted. When Sysctl is installed, each SP node will have a provided configuration file. The default pathname for this file is /**etc**/**sysctl.conf**. Alternate configuration files can be used. See the **sysctld** man page for more information.

The configuration file modifies the state of the server by executing one or more configuration commands. The configuration commands create read-only variables, register procedures, include other configuration files, and create classes of commands.

See "Installation and Configuration" on page 200 or the **sysctl.conf** man page for more information on configuring your system for Sysctl.

Access Control Lists (ACLs)

An ACL is a text file that lists authorized clients. (These files should be inaccessible to change by other than trusted administrators.)

There are two type of ACLs:

- The global Sysctl ACL file /**etc**/**sysctl.acl**

  This identifies a server's **trusted** users. Trusted users are authorized to run any Sysctl or Tcl command, except for those with the SYSTEM authorization level. Trusted users have privileges equal to those of a root user. This ACL identifies those clients with the ACL authorization level.

  The supplied /**etc**/**sysctl.acl** file contains commented lines that are samples. Copy the sample lines and insert your own principal identifiers.

- Installation-created ACLs

You can create ACLs for specific reasons. For example, you may designate a set of users that can issue a particular Sysctl command with an ACL for that command. The authorization callback for the command could check if the client was in the ACL.

Each line in an ACL contains either a Kerberos name (principal identifier) or the name of another ACL file. The principals in the included ACL file are added to those listed explicitly in the ACL file in determining the set of authorized clients. For example, an ACL file stored in one system can identify a core set of system administrators, followed by the name of another ACL file that is local to a node. Authorization for each SP node can be customized, while allowing the core administrators access to every machine.

The following is a sample ACL (note that the first line in an ACL file always begins with the magic cookie **#acl#**):

```
#acl#
_PRINCIPAL sarah.admin@TESTCELL.IBM.COM
_PRINCIPAL andrew@TESTCELL.IBM.COM
_ACL_FILE /usr/admin/acls/global_sysctl.acl
```

This ACL identifies the Kerberos names of two trusted users, **sarah.admin** and **andrew**. Both of these users are trusted only if their tickets come from realm **TESTCELL.IBM.COM**. The ACL also names another file, **/usr/admin/acls/global_sysctl.acl**, that contains additional ACL entries.

Note: You may want to use file collections to maintain configuration and ACL files. See Chapter 5, “Managing File Collections” on page 41.

Logfile

/var/adm/SPlogs/sysctl/sysctld.log is a logging file to which the Sysctl daemon (**sysctld**) writes. It logs the invocation of each Sysctl command. This is helpful for debugging. See the **sysctld** man page for information on specifying logfile pathnames.

Installation and Configuration

Installation

Sysctl is contained in the installation option file **ssp.sysctl**. This is installed when you install the SP system management software. See the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* for more information.

Configuring the System

Listed below is the basic procedure for configuring the system for Sysctl. Most of the steps are performed for you, either by the installation process or when you run **setup_authent** (both processes are described in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*).

Note: Some of these steps require administrative access to the authentication database. See Chapter 14, “Security Features of the SP System” on page 157 for details.

Configuration Steps Done for You

When you install **ssp.sysctl** the following steps happen automatically:

- The **/etc/sysctl.conf** file, which contains configuration information for the **sysctld** daemon, is installed on each node.
- The following, which describes the IP port and protocol used for that port, is added to **/etc/services**:

  ```
  sysctl    6680/tcp   sysctld
  ```

- **/etc/inittab** is updated to automatically start the **sysctld** daemon. **init**, as root, starts **sysctld** as follows:

  ```
  sysctld:2:once:/usr/bin/startsrc -s sysctld
  ```

Configuration Steps You Need to Perform

Here are the configuration steps that you need to perform:

- Add any trusted users to the global Sysctl ACL (**/etc/sysctl.acl**).

  When the server is operational, you can manipulate Sysctl ACLs through the Sysctl built-in commands described in Table 7 on page 216.

- Define in the **/etc/sysctl.conf** file any new commands you have created and distribute it to each node either manually or via file collections.
- Optionally, create any new ACLs and distribute them either manually or via file collections to each node.
- If you have modified the configuration file, the **sysctld** daemons must be restarted. See "Developing a Sysctl Application" on page 208 for more information.

Debugging Configuration Problems

At this point, your system should be configured for Sysctl. After running **kinit**, you should be able to do Sysctl tasks, such as:

```
sysctl whoami
```

where, if your principal name is **andrew@kingston.ibm.com**, the system response looks like this:

```
andrew@kingston.ibm.com
```

If the authentication principals and server key files are set up properly, the server echoes your authenticated identity. If it returns with **unknown** or another error message, some portion of your configuration is incorrect. See the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide*.

If a command does not work, run **sysctld** with debugging and security auditing turned on. First, you need to kill the local server and prevent its automatic respawning (such as with **/usr/bin/stopsrc -s sysctld**). Then you can start the daemon with:

```
sysctld -sd -l /var/adm/yourdebuglogfilename
```

The **-s** option writes a security audit trace to the log file as the server executes each user request. The **-d** option writes debugging information to the server log

file. It is useful to run **tail -f** (to view the debugging information as it's written to the log file) on the log file.

Once you have issued **kinit** and issued any Sysctl commands, you can issue **klist** to display your current tickets. The **klist** command displays the principal name used with **kinit** and displays the list of service tickets currently active (with an expiration date) for the user.

After issuing **klist**, do the following:

1. Check the principal name. If this is incorrect, you need to issue **kinit** and enter the correct principal name.
2. Check the list of service tickets. The list of service tickets should contain one ticket for each **rcmd** service for which a Sysctl command was issued.
   - If the **rcmd** tickets do not appear, there is a problem with the principal definition set up in the authentication database.
   - If the **rcmd** service tickets appear, but issuing **sysctl whoami** returns `unknown`, it is likely that the SP Authentication Service is improperly configured. See the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* for more information.

Sysctl Configuration Commands

Configuration commands create read-only variables, register procedures, include other configuration files, and create classes of procedures. Configuration commands are interpreted when a server is started or restarted, and are contained in configuration files. The configuration commands are:

- **include**
- **create var**
- **create proc**
- **create class**
- **set**
- **load**
- Any other valid Sysctl command

include

The **include** command specifies additional configuration files for the server to read. This makes it easier to manage many external command definitions by letting them be split among a hierarchical set of files. For example:

```
include $buildTop/samples/sysctl/pdfpfck.cmds
```

causes the Sysctl server to read the contents of the specified file at initialization time.

The syntax of **include** is:

```
include filename
```

create var

The **create var** command defines a read-only variable in the Sysctl server. Variables defined in the configuration file using the standard Tcl **set** command exist only while the configuration files are being read.

The syntax of **create var** is:

```
create var variable value [authcallback]
```

where:

variable A read-only variable to which you assign a value.

value The value of the variable

authcallback
An authorization callback for the command. This can be **NONE, AUTH, ACL** or **SYSTEM** (to assign it one of the built-in authorization levels), or the name of another previously-registered procedure.

The following example defines a variable (**STARTTIME**) that contains the time you started up the server:

```
create var STARTTIME [exec /bin/date] NONE
```

create proc

The **create proc** command defines external Sysctl procedures.

The syntax of **create proc** is:

```
create proc name args authcallback body
```

where:

name The name of the Sysctl procedure you are creating.

args A list of argument names to the procedure.

authcallback
An authorization specifier for the command. This can be **NONE, AUTH, ACL** or **SYSTEM** (to assign it one of the built-in authorization levels), or the name of a previously-registered procedure.

body A Tcl expression that forms the body of the new procedure.

See “Developing a Sysctl Application” on page 208 for an example of **create proc**.

create class

The **create class** command defines classes of commands within the server. Command classes provide a way to organize commands into logical groups for clarity.

The syntax of **create class** is:

```
create class name file [authcallback]
```

where

name The name that you assign to the command class.

This name constitutes a **tag** that is prefixed to all commands defined in the command file.

For example, if a class named **test** is created and the class file defines a procedure named **help**, the procedure is created in the interpreters as **test:help**.

file The pathname of the file containing the definition of the class objects.

authcallback
An authorization specifier for the command. This can be **NONE, AUTH, ACL** or **SYSTEM** (to assign it one of the built-in authorization levels), or the name of a previously-registered procedure.

set

The **set** statement sets the value of the server variables ACL, LOG or KEY, or sets values in the Tcl env() array. The default values for the Sysctl server's ACL file, log file and Kerberos key file can be overriden by assigning values to the ACL, LOG and KEY variables in the configuration file. For example, the following line overrides the default value for the log file name:

```
set LOG /var/sysctld.logfile
```

The values assigned to the ACL, LOG and KEY variables are overriden by the optional command line arguments **-a**, **.-l** and **-k** on the **sysctld** command line. Environment variables (such as the default PATH) can also be set within the configuration file by assigning values to the env() array. For example:

```
set env(PATH) /usr/bin:/bin:/usr/etc:/etc
```

sets the PATH environment variable for the Sysctl server. The env() array is assigned an authorization callback of SYSTEM which prevents its modification from outside the Sysctl server by a request sent by a Sysctl client.

load

The **load** command dynamically loads the shared library at **lib_path** into memory. If the **init_proc** parameter is given, it is used as the library's initialization procedure. Otherwise the name of the initialization function is derived from the library name as follows:

```
library name    init function
============    =============
libxxx.sl       xxx_Init()
libxxx.a        xxx_Init()
```

The Sysctl server exports an API that the library uses to define commands, variables, authorization callbacks and interpreter deletion callbacks. See the **load** help page for details.

Other Sysctl Commands

You can include other commands in the Sysctl configuration file. These commands are executed by the server when it is started (or restarted).

Authorization Tasks

Authorization Callbacks

Sysctl provides an extremely flexible authentication mechanism to allow you to restrict the use of registered Sysctl procedures to any easily defined subset of users. Authorization callbacks are the means by which this authentication is implemented. As previously discussed, Sysctl includes four supplied authorization callbacks which provide the following authorization levels:

- NONE
- AUTH
- ACL
- SYSTEM

Adding New Authorization Callbacks

As an administrator you can develop a procedure with an authorization level that does not fit one of the supplied authorization callbacks. To do this, write your own authorization callback. Here is an example of a custom authorization callback:

```
create proc scdate {} SYSTEM {exec /bin/date}


create proc RCRAUTH {cmdName} SYSTEM {
    global SCUSER
    global SCPRINCIPAL
    if {$SCUSER != "rcr"} {
        svclog "Denying \"$cmdName\" access for user $SCPRINCIPAL"
        error "Authorization denied!"
    } else {
        svclog "Authorizing \"$cmdName\" access for user $SCPRINCIPAL"
        return "Authorization OK!"
    }
}

create proc rcrdate {} {RCRAUTH rcrdate} {scdate}
```

This example contains the definition of three procedures:

- **scdate** is a procedure that the SYSTEM authorization callback is attached to. The SYSTEM callback prevents **scdate** from being directly executed from outside of the server (that is, a user cannot successfully issue **scdate** from the Sysctl client). The **scdate** procedure can, however, be successfully executed from within another procedure which has been previously registered with the Server (that is, included in the Sysctl configuration file), since authorization callbacks are bypassed during this execution.
- **RCRAUTH** is a procedure that functions as an authorization callback. It accepts a command name as its single argument (used for logging purposes only). Its function is to return a normal result if the basename of the user is **rcr** and to return an error result in all other cases.
- **rcrdate** is a procedure that invokes the **scdate** procedure. Before control is given to the **scdate** procedure, the RCRAUTH authorization callback is invoked. If the base name of the user is **rcr**, then the callback returns a normal result and the **scdate** command is allowed to be invoked. If the callback returns an error result, then the **rcrdate** procedure is terminated.

ACL Files in Authorization Callbacks

The supplied **ACL** authorization callback is used to check for the remote user's principal name in an ACL file. If this callback is used without any parameters then the ACL file named by the ACL variable (normally /**etc**/**sysctl.acl**) is checked. If a file name is supplied as a variable to the ACL callback, then that file is checked for the user's principal name.

This allows an authorization callback to reference an alternate ACL file in performing the authorization check:

```
create proc myproc {} {ACL /etc/myproc.acl} {
        ....
        ....
        ....
}
```

The above example defines the procedure **myproc**, which takes no parameters, and has an authorization callback of **ACL** /**etc**/**myproc.acl**. This callback returns a normal result if the remote user's principal name appears in the file /**etc**/**myproc.acl** and an error return if it does not.

ACL files can be created and deleted and their contents can be manipulated and checked by using the Sysctl commands **aclcreate**, **acldestroy**, **acladd**, **acldelete**, **acllist**, **aclcheck** and **aclrecreate** (see the Sysctl help information for these commands).

Using Sysctl

Prerequisite Checklist

Before using Sysctl, the following is required:

- You have a principal in the local Kerberos realm and have run the **kinit** command to authenticate your identity to the system.
- The configuration is complete (see “Configuring the System” on page 200).
- The server daemon (**sysctld**) is running on all nodes.

Using Sysctl in Interactive Mode

Sysctl has an interactive mode. Note that only one node can be communicated with at a time in interactive mode.

To enter interactive mode, enter the **sysctl** client command without a command argument. For example, to use Sysctl in interactive mode on the local node, enter:

```
sysctl
```

If you issued the command from **diane.kgn.ibm.com**, you get the following prompt:

```
Sysctl (Version 1.1) on diane.kgn.ibm.com
sysctl>
```

At this point, you can enter Sysctl commands.

You can also specify some of the Sysctl flags when going into interactive mode. For example, to use Sysctl in interactive mode for node **redhook.kgn.ibm.com**, enter:

```
sysctl -h redhook.kgn.ibm.com
```

Sysctl issues this response:

```
Sysctl (Version 1.1) on redhook.kgn.ibm.com
sysctl>
```

In this example, if you enter **svcversion** and **pdf**, the interaction is as follows, where your actions are in **bold**:

```
sysctl> svcversion
1.1
sysctl> pdf
redhook.kgn.ibm.com: ======================================================
Filesystem                Size-KB Used-KB Free-KB %Free  iUsed  iFree %iFree
/                            8192    5164    3028   37%    815   1233    61%
/var                        16384    4176   12208   75%    343   3753    92%
/usr                       552960  467888   85072   16%  23203 116061    84%
/tmp                        49152   23860   25292   52%    235  12053    99%
/home                       65536   63556    1980    4%   5375  11009    68%
sysctl> quit
```

To exit the interactive mode, enter one of the following:

```
exit
quit
<Ctrl-d>
```

Using Sysctl in Noninteractive Mode

Sending Commands to a Local Server

If no target nodes are specified on the Sysctl command, the command is sent to the server on the local node. For example, to send the **whoami** command to the local server, enter:

```
sysctl whoami
```

Directing Commands to Remote Hosts

To direct a command to remote host **sivle**, use **-h** and enter:

```
sysctl -h sivle sys:info
```

The system's response:

```
sivle.kgn.ibm.com power 3.2 rs_aix32
```

To direct a command to remote hosts **sivle**, and **yelserp**, enter:

```
sysctl -h sivle -h yelserp sys:info
```

The system's response:

```
>> sivle
sivle.kgn.ibm.com power 3.2 rs_aix32
<<
>> yelserp
yelserp.kgn.ibm.com power 3.2 rs_aix32
<<
```

Sending Commands to a Collection of Nodes

Multiple nodes can be targeted with a single Sysctl command. The collection of nodes can be specified in a file. Alternatively, the Sysctl command will read node names from **stdin**. In the following example, the file /**tmp**/**node-file** contains node names **r05n13.hpssl.kgn.ibm.com**, **r05n15.hpssl.kgn.ibm.com**, and **r05n09.hpssl.kgn.ibm.com**.

To direct the **pdf** command to the /**tmp**/**node-file** collection, enter:

```
sysctl -c /tmp/node-file pdf
```

The response in this example is the following:

```
>> r05n13
r05n13.hpssl.kgn.ibm------------------------------------------------------------
Filesystem                Size-KB Used-KB Free-KB %Free  iUsed  iFree %iFree
/                            4096    3372     724   18%    792    232    23%
/var                         8192    6320    1872   23%    297   1751    86%
/usr                       335872  331632    4240    2%  17820  66148    79%
/tmp                         8192     520    7672   94%     67   1981    97%
/home                        4096     168    3928   96%     18   1006    99%
<<
>> r05n15
r05n15.hpssl.kgn.ibm------------------------------------------------------------
Filesystem                Size-KB Used-KB Free-KB %Free  iUsed  iFree %iFree
/                            4096    3364     732   18%    796    228    23%
/var                         8192    3028    5164   64%    281   1767    87%
/usr                       335872  331412    4460    2%  17825  66143    79%
/tmp                         8192     464    7728   95%     57   1991    98%
/home                        4096     168    3928   96%     18   1006    99%
<<
>> r05n09
r05n09.hpssl.kgn.ibm------------------------------------------------------------
Filesystem                Size-KB Used-KB Free-KB %Free  iUsed  iFree %iFree
/                            4096    3340     756   19%    790    234    23%
/var                         8192    3336    4856   60%    285   1763    87%
/usr                       331776  331012     764    1%  17742  66226    79%
/tmp                         8192    1212    6980   86%     58   1990    98%
/home                        4096     164    3932   96%     17   1007    99%
<<
```

Developing a Sysctl Application

Creating a Sysctl application consists of the following basic steps:

1. Examine what you are trying to do and divide it among client/server lines. The server side runs as **root**. The client side typically does not require root authority, but requires that the user has run **kinit**.

2. Define the client/server interface. This consists of Sysctl command names and their arguments, return values, and authorizations.
3. Code the server commands. This consists of Tcl code that performs your task, as well as callback authorizations if you do not use one of the supplied ones. Note that the Tcl code may be simple (such as a wrapper around an AIX command or script) or elaborate (a detailed Tcl program).
4. Code the client side. This typically gathers any necessary information and then issues the **sysctl** command to cause the servers to perform the desired action. This can be a GUI, Perl or shell script. Note that client side code may not always be necessary. In some cases, issuing the **sysctl** command directly from the command line is sufficient.
5. If the authorization scheme requires changes to existing ACL files or new ACL files, these changes must be made, and the changed files distributed.
6. Register the new server command with each node's **sysctld** daemon by restarting the daemons with updated configuration files.

Here is an example of how to write a Sysctl application that allows a user to execute a script including some commands requiring root execution privileges on all the nodes in a SP System in parallel. The user is not required to know any of the root passwords, but must be listed in an ACL specifically set up to authorize use of the script. A copy of this ACL resides on each node. The script containing the commands executable only by root is called **/etc/root_script**. Assume that **/etc/root_script** accepts a single argument, and that this script resides on each node.

The server side of the application does the root part of the job, which is to issue the **/etc/root_script** command with the correct arguments on behalf of authorized users. It first needs to determine if the client is authorized based on an authorization policy. As mentioned above, authorization is based on an access control list. There are no return values, but an error message is displayed when the user is not authorized to run the command. **stdout** and **stderr** are returned to the client.

The server command is coded as follows and resides in **/etc/sysctl/apps/adminscript**:

```
create proc adminscript {argument} AUTH {
        global SCUSER
        if [ aclcheck -f /etc/sysctl.adminscript.acl $SCUSER ] {
                exec /etc/root_script $argument
                return
        }
        puts "Not authorized to run adminscript"
}
```

The first line creates a new Sysctl proc called **adminscript**. One argument is accepted, which will be passed unchanged to **/etc/root_script**. This proc uses the provided AUTH authorization callback, which means that only authenticated clients can execute the command.

The second line issues the Tcl **global** command to provide the content of the SCUSER variable to other Sysctl commands. (This is needed because SCUSER is passed as an argument to **aclcheck** later.) The SCUSER variable is the

authenticated username of the client, and is one of the variables set by the server and available to procedures and callbacks. (See “Authorization Variables” on page 211 for a list of all these variables.)

The third line is Tcl code that uses the built-in **aclcheck** command to see if the user is in the ACL (/**etc**/**sysctl.adminscript.acl**) set up for this command. If the user is authorized, the program issues the /**etc**/**root_script** command with the supplied argument and then returns. If the user is not authorized, the program issues an error message. The program returns any output (information from /**etc**/**root_script** command or the error message) to the client user's terminal.

Note that in this example, further authorization checking is done in the body of the procedure after the authorization callback has completed. This is to demonstrate the use of the SCUSER variable provided by the server to procedures. Alternatively, this check could have been made part of the authorization callback itself.

The client side of the application needs to find out the necessary information to provide to the servers. It does not require **root** authority to be issued. The information in this case includes the single argument to be passed through to /**etc**/**root_script**. The client also needs to determine the hostnames of all the nodes in the SP System, so that the command can be sent to and run in parallel on all the nodes.

The client side is coded as follows in /**usr**/**bin**/**doadminscript**:

```
#!/usr/lpp/ssp/perl/bin/perl

# Run /etc/root_script on all the SP nodes in parallel by invoking
# the adminscript sysctl command on all the nodes. Invoke via:
# doadminscript arg_to_root_script

$#ARGV > 1 && die "too many arguments";

exec "/usr/lpp/ssp/bin/hostlist -av |
      /usr/bin/sysctl -c - adminscript $ARGV[0]";
```

The first line checks for extraneous arguments. The **exec** statement uses the SP **hostlist** command to write all the currently responding hostnames in the SP System to **stdout**. The single argument to **adminscript** is passed along to the Sysctl **adminscript** server command.

A new ACL file needs to be put on each node in the system. The file /**etc**/**sysctl.adminscript.acl** looks like this:

```
#acl#
# These are the users that can issue the adminscript command on this
# node:
_PRINCIPAL root@HPSSL.KGN.IBM.COM
_PRINCIPAL newbie@HPSSL.KGN.IBM.COM
_PRINCIPAL guru@HPSSL.KGN.IBM.COM
```

Note that each node can authorize different users to run the new command. In this case, assume that the ACL will be the same on all nodes. To distribute the ACL file to all the responding nodes, an authorized user enters the following (see the **pcp** man page for details):

```
pcp "-av" /etc/sysctl.adminscript.acl
```

In addition, the **/etc/sysctl.conf** file needs to be updated to include the new command. If the server code is to reside in **/etc/sysctl/apps** directory, the following lines can be added to **/etc/sysctl.conf**:

```
# Add the adminscript application
include /etc/sysctl/apps/adminscript
```

The updated **sysctl.conf** file and the actual server code need to be distributed:

```
pcp "-av" /etc/sysctl.conf
pcp "-av" /etc/sysctl/apps/adminscript
```

Now the servers need to be restarted to include the updated **sysctl.conf** file. Note that the following command requires ACL authorization:

```
hostlist -av | sysctl -c - svcrestart
```

Now, any of the users in **/etc/sysctl.adminscript.acl** (**root**, **newbie**, or **guru**) can run **/etc/root_script** with the argument **mikef** on all the nodes simultaneously by entering the following commands:

```
kinit (to get authentication information)

doadminscript mikef
```

Authorization Variables

These variables are set by the server based on information on the client provided by SP Authentication Services. Authorization callbacks and Tcl scripts have access to these variables. Table 4 lists the authorization variables.

Table 4. Authorization Variables

Variable	Description
SCACL	Indicates if the client has been authenticated and is a trusted user. Value is **1** if the client has been authenticated and is listed on the server's ACL of trusted users, **0** otherwise.
SCHOST	The name of the host SP node from which the Sysctl request was issued.
SCINSTANCE	The instance of the user or NULL if unauthenticated.
SCLHOST	The host name of the local server.
SCLREALM	The local realm of the server.
SCMODE	The communication mode between the client and server. May be WAIT, NOWAIT, or SOCKET.
SCPRINCIPAL	The authenticated identity of the command issuer, in the form user.instance@realm. If the client has not been authenticated, the value is NULL.
SCREALM	The realm of the user or NULL if unauthenticated.
SCSCRIPT	The commands sent from the client to the server. If a script was passed, the contents of the script is shown.
SCUSER	The base name of the user. NULL if unauthenticated.

Sysctl Scripts

Replaying a Sysctl Script

A Sysctl script consists of a set of Sysctl commands that can be executed with a single Sysctl client invocation. Sending a Sysctl script to a server or set of servers is called *replaying* the script on those servers.

When a script is replayed, each Sysctl command is run sequentially by the servers. Authorization is done on a command by command basis as the script runs. The script stops execution when any command fails to be executed due to an authorization failure.

To run a Sysctl script, send it to the Sysctl server by specifying **-r** (replay) on the **sysctl** command:

```
sysctl -h node -r script-filename
```

The following is a simple script, **/examples/sysctl/script1**:

```
echo I am $SCLHOST
echo You are $SCPRINCIPAL
```

Here is the command that user **sarah** issues to run the script on SP node **r05n11**:

```
sysctl -h r05n11 -r /examples/sysctl/script1
```

sarah receives the following output:

```
I am r05n11.testcell.ibm.com
You are sarah@testcell.ibm.com
```

You can also have Sysctl read a script from **stdin** by typing a dash (-) after the **-r** flag, instead of a file name. For example, to run three commands for node **tappan.kgn.ibm.com**, enter the following sequence of commands:

```
sysctl -h tappan.kgn.ibm.com -r -
echo $SCLHOST
pdf
svcversion
<Ctrl-D>
```

You will not get any responses until you press <**Ctrl-D**>. In this example, the response is the following:

```
tappan.kgn.ibm.com
tappan.kgn.ibm.com: ============================================================
Filesystem                Size-KB Used-KB Free-KB %Free  iUsed  iFree %iFree
/                            8192    5164    3028   37%    815   1233    61%
/var                        16384    4184   12200   75%    343   3753    92%
/usr                       552960  467888   85072   16%  23203 116061    84%
/tmp                        49152   23868   25284   52%    239  12049    99%
/home                       65536   63556    1980    4%   5375  11009    68%
1.1
```

In addition, Sysctl scripts can be written as self-contained programs. Do this by beginning the script with (flags other than **-r** can be specified this way as well):

```
#!/usr/lpp/ssp/bin/sysctl -r
```

Enter the name of the script to run it.

Consider the following example, /**etc**/**fscheck**. This script checks filesystem usage and reports on filesystems that have usage higher than that specified in an argument. The **-L** flag causes output to be formatted as a set of lines, each preceded by the hostname from which it came:

```
$ cat /etc/fscheck
#!/usr/lpp/ssp/bin/sysctl -Lr
#
# Check filesystem usage - high-water pct mark is passed as a parameter
#
foreach fs [listfs] {
     statfs $fs sbuf
     set pct [expr $sbuf(used).0/$sbuf(size).0*100]
if {$pct > $argv(1)} {
          set pct [format "%.2f" $pct]
          echo "$fs ==> ${pct}% Used, $sbuf(free) KB Free"
     }
}
$ cat /tmp/hostlist
harlem
badger
$ fscheck -c /tmp/hostlist 80
harlem: / ==> 84.29% Used, 1538 KB Free
badger: /usr ==> 98.35% Used, 3656 KB Free
$
```

You can override any parameters specified in the script by invoking it explicitly. For example, when you issue:

```
sysctl -r /etc/fscheck
```

Sysctl ignores the **#!** line in the script. Since the **-L** option is not interpreted in this case, the output would be:

```
$ sysctl -c /tmp/hostlist /etc/fscheck 80
<<
>> harlem
/ ==> 84.29% Used, 1538 KB Free
<<
>> badger
/usr ==> 98.35% Used, 3656 KB Free
<<
```

Specifying a Host Target in a Script

Sysctl allows you to target a particular host within a script. The target host is specified in the body of the script via use of the **#host#** keyword. If the script contains a line beginning with **#!**, the **#host#** line must immediately follow it. Otherwise, **#host#** must be the first line in the script.

If you use the **-h** option to specify an SP node name when you invoke **sysctl**, any target host specified in the script is ignored. Here is a script, /**examples**/**sysctl**/**script3**, targeted to host **issc3**:

```
#!/usr/bin/sysctl -r
#host#issc3
echo I am $SCLHOST
```

Using the **-h** option to run the script overrides the host specified within the script. For example, if you specify:

```
sysctl -h r05n11 -r /examples/sysctl/script3
```

the script runs on **r05n11** instead of **issc3**.

Passing Parameters to a Script

When you issue the **sysctl** command with the **-r** parameter, any command line arguments following the script name are treated as parameters to be passed to the script. Here is a sample script, /**examples**/**sysctl**/**script6**:

```
#!/usr/bin/sysctl -r
#
echo Number of parameters is $argc
for {set i 0} {$i < $argc} {incr i 1} {
   echo Parameter #$i is \"$argv($i)\"
}
```

When you invoke the script with parameters **arg1** and **arg2**:

```
/examples/sysctl/script6 arg1 arg2
```

The output is the following:

```
Number of parameters is 3
Parameter #0 is script6
Parameter #1 is "arg1"
Parameter #2 is "arg2"
```

In this example, two parameters were passed to the script. The parameters are stored in the **argv** array, and the number of parameters is stored in variable **argc**, just as in the standard Tcl **exec** function. Because **argv[0]** always contains the name of the script being run, the value of **argc** is one greater than the number of arguments passed to the script. That is why the first line of output says that there are three parameters, when only two were passed to the script.

Sysctl Reference Information

Online Help Facility

To access online help on built-in Sysctl commands, enter:

```
sysctl help  command-name
```

For example, to access help on the **setauth** command, enter:

```
sysctl help setauth
```

Tcl Commands

Table 5 shows the commands available in the Tcl language interpreter and the Sysctl base authorizations required to use them.

Table 5. Built-in Tcl Commands and Their Default Authorization Callbacks

append (AUTH)	array (AUTH)	break (AUTH)	case (AUTH)
catch (AUTH)	cd (ACL)	close (ACL)	concat (AUTH)
continue (AUTH)	eof (ACL)	error (AUTH)	eval (ACL)
exec (ACL)	exit (AUTH)	expr (AUTH)	file (ACL)
flush (ACL)	for (ACL)	foreach (ACL)	format (AUTH)
gets (ACL)	glob (ACL)	global (AUTH)	history (AUTH)
if (AUTH)	incr (AUTH)	info (ACL)	join (AUTH)
lappend (AUTH)	lindex (AUTH)	linsert (AUTH)	list (AUTH)
llength (AUTH)	lrange (AUTH)	lreplace (AUTH)	lsearch (AUTH)
lsort (AUTH)	open (ACL)	pid (ACL)	proc (SYSTEM)
puts (ACL)	pwd (ACL)	read (ACL)	regexp (AUTH)
regsub (AUTH)	rename (SYSTEM)	return (NONE)	scan (AUTH)
seek (ACL)	set (AUTH)	source (SYSTEM)	split (AUTH)
string (AUTH)	switch (AUTH)	tell (ACL)	time (ACL)
trace (ACL)	unset (AUTH)	uplevel (SYSTEM)	upvar (ACL)
while (ACL)			

Extended Tcl Commands

Table 6 shows the commands available in the TclX language interpreter and the Sysctl base authorizations required to use them.

Table 6. Built-in Extended Tcl Commands and Their Default Authorization Callbacks

alarm (ACL)	bsearch (ACL)	catclose (ACL)	catgets (ACL)
catopen (ACL)	cequal (AUTH)	cexpand (AUTH)	chgrp (ACL)
chmod (ACL)	chown (ACL)	chroot (SYSTEM)	cindex (AUTH)
clength (AUTH)	cmdtrace (ACL)	commandloop (ACL)	convertclock (AUTH)
copyfile (ACL)	crange (AUTH)	csubstr (AUTH)	ctoken (AUTH)
ctype (AUTH)	dup (ACL)	echo (AUTH)	execl (ACL)
fcntl (ACL)	flock (ACL)	fmtclock (AUTH)	fork (ACL)
frename (ACL)	fstat (ACL)	funlock (ACL)	getclock (AUTH)
id (ACL)	infox (ACL)	keyldel (AUTH)	keylget (AUTH)
keylkeys (AUTH)	keylset (AUTH)	kill (ACL)	lassign (AUTH)
lempty (ACL)	lgets (ACL)	link (ACL)	lmatch (AUTH)
loop (ACL)	lvarcat (AUTH)	lvarpop (AUTH)	lvarpush (AUTH)
max (ACL)	min (ACL)	mkdir (ACL)	nice (ACL)
pipe (ACL)	profile (ACL)	random (AUTH)	readdir (ACL)
replicate (AUTH)	rmdir (ACL)	scancontext (ACL)	scanfile (ACL)
scanmatch (ACL)	select (ACL)	server_open (ACL)	signal (ACL)
sleep (ACL)	sync (ACL)	system (ACL)	times (ACL)
translit (AUTH)	umask (ACL)	unlink (ACL)	wait (ACL)

Built-in Sysctl Commands

The following Sysctl commands are provided:

Table 7. ACL Processing Commands

ACL Processing Commands	Default Auth	Description
acladd	ACL	Adds entries to an ACL file. If a filename is not specified, the entry is added to the server's ACL file.
aclcheck	ACL	Checks to see if a principal is in an ACL file. Returns **1** if in ACL, **0** if not in ACL. If no filename is specified, Sysctl looks in the server's ACL file.
aclcreate	ACL	Creates a new ACL file, inserting the principals specified. The **-f** parameter must be supplied.
acldelete	ACL	Deletes an ACL from a file.
acldestroy	ACL	Erases an ACL file.
acllist	ACL	Lists the ACLs in a file. If no file is specified, the server's ACL file is listed.
aclrecreate	ACL	Similar to **aclcreate**, but removes the target ACL file if it already exists before creating it again.

Table 8. Service Commands

Service Commands	Default Auth	Description
svcconnect	AUTH	Determines the connection authorization policy for the **sysctld** server.
svcdetach	ACL	Detaches the current process from the calling client; that is, breaks the connection between server and client. Use to start a background daemon.
svclog *string*	ACL	Writes the *string* to the server log file.
svclogevent	SYSTEM	Writes a line to the server log file indicating a user has connected to the server.
svcpid	ACL	Prints the process ID of the **sysctld** server daemon.
svcredirect *handle*	ACL	Redirects the **stdout** and **stderr** for the current session to a new file.
svcrestart	ACL	Restarts the server, causing configuration files to be reread and interpreters to be reinitialized. Use to activate changes to Sysctl servers.
svcversion	NONE	Returns the Sysctl version number.

Table 9. Other Commands

Other Commands	Default Auth	Description
ACL	AUTH	A supplied authorization callback.
AUTH	AUTH	A supplied authorization callback.
NONE	AUTH	A supplied authorization callback.
SYSTEM	AUTH	A supplied authorization callback.
checkauth	ACL	Determines if you are authorized to access an object.
confadd	ACL	Adds items to a server's configuration file.
confdelete	ACL	Removes items from a server's configuration file.
create	SYSTEM	Creates new objects in the server.
getauth	ACL	Displays the authorization callbacks for an object.
include	SYSTEM	Includes additional configuration files.
listfs	ACL	Lists file system names.
load	SYSTEM	Dynamically load a shared library into the server and call an initialization routine within the library.
quit	AUTH	Ends the current session.
safeargs *arg-list*	ACL	Checks arguments for illegal shell characters. Does not execute the command.
safeexec *commands ...*	ACL	Runs the Tcl exec or **system** command, checking arguments for illegal shell characters and preventing execution if any are found.
safesystem	ACL	Runs the Tcl exec or **system** command, checking arguments for illegal shell characters and preventing execution if any are found.
setauth	SYSTEM	Sets the authorization callback for an object.
statfs	ACL	Queries the status of all local file systems.
unload	SYSTEM	Unloads a previously loaded shared library.
whoami	NONE	Echoes the authenticated identity of the issuer or **unknown** if the user is not authenticated.

Managing Sets of Nodes

© Copyright IBM Corp. 1995, 1997

Chapter 16. Using a Switch

The optional Scalable POWERparallel Switch (SP Switch) and the High Performance Switch offered with the SP system provides high bandwidth, low latency, location-independent communication between nodes. The Scalable POWERparallel Switch adapter is the SP Switch Adapter. The High Performance Switch adapter is the High Performance Switch Adapter-2. This chapter presents information on:

- Switch communication modes
- Primary node takeover
- System partitioning and the switch
- Extension nodes and the switch
- Selecting switch clocks
- Selecting the primary and primary backup node
- Managing the switch topology file
- Starting the switch
- Fencing and unfencing nodes from the switch
- Preparing for system repartitioning
- Monitoring nodes on a switch
- Global shutdowns and reboots of nodes with a switch
- NFS mounts over a switch

Usage Note

Throughout this chapter, use of the term SP Switch also applies to the SP Switch-8 unless specifically stated otherwise. Differences will be noticed in discussions of system partitioning, switch node numbering, and cabling.

Switch Communication Modes

Two communication protocols are offered for the adapter:

1. AIX sockets using standard IP protocol
2. Dedicated user space access (via message passing libraries) and multiuser environments (via IP)

Both modes can be used concurrently, as if the adapter was actually two separate adapters.

Switch Communication Using IP (TCP/IP and UDP/IP)

The switch adapters are configured for IP by default during the installation process. You can assign specific IP network addresses statically during installation or dynamically using ARP (Address Resolution Protocol). When you assign the IP address statically, there are restrictions on mapping of nodes to low order addresses (see the information on switch node numbering in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*). When you assign the IP address range dynamically using ARP, there are no restrictions.

Any application can communicate over the switch by opening a standard socket and specifying appropriate IP addresses. Additional system facilities based on IP, such as NFS, and licensed programs such as AFS, can also be configured to use

© Copyright IBM Corp. 1995, 1997

the switch. With the possible exception of applications that depend on network-specific functions (LAN broadcasts, for example), most applications work over the switch without modification.

Notes:

1. The **iptrace** command supports both switch adapters and provides interface-level packet tracing for incoming and outgoing packets.
2. The standard AIX commands **ifconfig** and **arp** have been modified to support the switch adapters. For further information, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.
3. Static routes associated with the switch will not be automatically reconfigured during node reboot. The routes for the switch are associated with the default interface (**en0**) upon reboot because the switch is not available during the phase of system initialization when the static routes are reconfigured. If you want the routes added automatically, we recommend that you put them in a script and have the script called as the last entry in **/etc/inittab**.

Switch Communication Using User Space Message Passing

The user space message passing mode supports parallel applications requiring maximum bandwidth and minimum latency. This mode is used by PVMe and the Parallel Environment for AIX to provide their exported message passing interfaces. In this mode, only a single user-space process may access the adapter, and the nodes to be used must be allocated to the parallel application via the Resource Manager. The Resource Manager should group nodes into named resource pools to facilitate node scheduling.

Both the SP Switch adapter and the High Performance Switch Adapter-2 provide multiplexing for and automatic availability of both communication modes. Although the user space message passing mode may be used by only one process per node, TCP/IP and UDP/IP messages can be processed concurrently (for example, NFS I/O).

Primary Node Takeover

Error recovery on the SP Switch is done locally instead of globally as on the High Performance Switch. On the High Performance Switch, a fault on a link would cause the entire fabric to be interrupted while recovery was performed. On the SP Switch, detection and fault isolation is done at the link level while normal traffic flows across the rest of the fabric. Detected faults are forwarded to the active primary node for analysis and handling. When the primary node completes assessing the fault, the remaining nodes on the fabric are nondisruptively informed of status changes.

The primary node initializes the switch fabric and processes switch operations and errors reported from the switch fabric. The primary node is no longer a single point of failure on an SP system containing an SP Switch.

A new node type has been developed called the primary backup node. The primary backup node passively listens for activity from the primary node. When the primary backup node detects that it has not been contacted by the primary node for approximately 7 minutes, it assumes the role of the primary node. This takeover

involves nondisruptively reinitializing the switch fabric, selecting another primary backup, and updating the System Data Repository (SDR).

The primary node also watches over the primary backup node. If the primary node detects that the primary backup node can no longer be contacted on the switch fabric, it selects a new primary backup node. The criteria to select a new primary backup node follows:

1. First, select a node on a switch assembly other than the switch assembly to which the primary node is attached.
2. Second, if no other switch assembly exists, select a node attached to a switch chip other than the one to which the primary node is attached.
3. If no other switch chip exists, select any available node on the switch chip to which the primary node is attached.

During the period of time between the failure or loss of the primary node and the takeover by the primary backup node, the switch fabric continues to function. Secondary failures are detected during the reinitialization of the switch fabric.

System Partitioning and the Switch

As discussed in Chapter 17, “Managing System Partitions” on page 239, the SP system can be organized into groups of nodes for various purposes. System partitioning has the following implications for the switch:

- Commands related to the switch:
  - **Eannotator** - Operates only within the current system partition
  - **Eclock** - Always operates globally (across system partitions)
  - **Eduration** - Operates only within the current system partition
  - **Efence** - Can operate globally (**-G**) or only within the current system partition
  - **Eprimary** - Operates only within the current system partition
  - **Equiesce** - Operates only within the current system paritition
  - **Estart** - Operates only within the current system partition
  - **Etopology** - Operates only within the current system partition
  - **Eunfence** - Can operate globally (**-G**) or only within the current system partition
  - **Eunpartition** - Operates only within the current system partition
- Switch faults and message traffic are contained *within a system partition*.
- The subset of the switch defined by the topology file for the system partition is viewed as a logical switch.
- There is no connectivity over the switch *between system partitions*. (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. Use the explicit host routes on the gateway node or enable ARP on all system partitions and redefine the IP addresses within a system partition as a different subnetwork.)

- There remains one switch clock source, regardless of system partitioning.
- There is one primary node per system partition.

Extension Nodes and the SP Switch

Extension nodes are non-standard nodes that extend the SP system's capabilities or scope, but can't be used in all the same ways as standard SP nodes.

A specific type of extension node is a dependent node. Dependent nodes are attached to the SP Switch via a special dependent node adapter.

A dependent node implements much of the switch-related protocol that standard nodes use on the SP Switch, but it depends on standard SP nodes for certain functions. A dependent node can use the SP Switch as an IP network, but it relies upon a standard SP node to generate its switch route table. (The communication subsystem does not send dependent nodes switch topology database updates, but rather generates and sends appropriate switch route tables.)

SNMP support provides for the communication of switch-related configuration information between the SP system and the dependent node.

A dependent node cannot be a primary or primary backup node for the SP Switch.

The **CSS_test**, **Efence**, and **Eunfence** commands operate on dependent nodes in the same way they operate on standard nodes. The **Eprimary** command returns an error message if a dependent node is specified for a primary or primary backup node for the SP Switch. The **Eannotator** command notes dependent nodes in the annotated topology file.

For more information on extension and dependent nodes, refer to Chapter 19, “Managing Extension Nodes” on page 269.

Selecting Appropriate Switch Clocks

One of two switches in the system provides the master switch clock to the rest of the system. If the current master clock fails, or is removed from the configuration, the alternate master clock must be selected and the switch reinitialized.

The **Eclock** command establishes a master clock source after the system is powered up or when an alternate must be selected. It can set the appropriate clock input for every switch in the system or for a single switch after power on. After **Eclock** completes, the **Estart** command must be run.

Refer to the *IBM RS/6000 SP: Maintenance Information Volume 1* for information on removing and restoring the High Performance Switch resources.

Warning: Running **Eclock** after the switch is started may cause nodes to lose their clock source. Refer to **Eclock** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

Selecting the Switch Clock Source

To set the switch clock source using SMIT:

TYPE **smit**

▲ The System Management menu appears

SELECT RS/6000 SP Systems Management

▲ The RS/6000 SP Systems Management menu appears.

SELECT RS/6000 SP Cluster Management

▲ The RS/6000 SP Cluster Management menu appears.

SELECT Perform Switch Operations

▲ The Perform Switch Operations menu appears.

SELECT Change/Show Switch Clock Source Settings

▲ The Change/Show Switch Clock Source Settings menu appears.

At this point, you can select options:

```
Change/Show Switch Clock Source Settings

□ List Switch Clock Source Settings
□ Initialize Switch Clock Source Settings
□ Initialize Switch Clock Source Settings for a Switch
□ Restore the Switch Clock Source Settings
□ Select Alternate Switch Clock Source Settings
□ Create an Eclock Topology File
□ Automatic selection of Switch Clock Settings
```

Figure 14. The Change/Show Switch Clock Source Settings SMIT Menu

When you choose a switch clock source setting option, you are prompted for a topology file name, and then the **Eclock** command is invoked.

The fastpath invocation for this menu is:

```
smit clock_src
```

Switch Clock Reset

The switch clock is synchronous across all nodes active on a switch. Its value is set to zero when the primary node powers on, and is later distributed by **Estart** to all nodes on or joining the switch. The switch clock will increase monotonically unless it is reset by one of the following events:

- The switch is powered off.
- An oscillator failure occurs on the switch board designated as the master clock source. (The switch designated as the master clock source provides the clocking signal for all other switch boards within the configuration.)
- The **Eclock** command is issued.

- The **Estart** command is issued while the oncoming primary node (or the primary node on High Performance Switch systems) is not active on the switch (for example, after the primary node is powered on).
- SP Switch Only:
  - The **Estart** command is issued after **rc.switch** or reboot has run on the primary node, but before the primary backup node has taken over. For futher information, see "Primary Node Takeover" on page 222.

Note: We recommend that a new primary node that is active on the switch be selected before the old primary node is powered off. This will prevent the switch clock from being reset to zero or to an inderminate value on High Performance Switch systems. For more information on selecting a primary node, refer to **Eprimary** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Selecting the Primary and Primary Backup Nodes

High Performance Switch Systems:

On High Performance Switch systems, select one standard node to be the switch primary node. This node will be used as the primary node on the next **Estart**.

If the primary node is unavailable (powered-down, rebooted, and so forth), run **Eprimary** to designate a new primary. (This should be followed by **Estart** to restart the High Performance Switch. On a primary node failure, the High Performance Switch and the "switchResponds" System Monitor display are in an inconsistent state until **Eprimary** and **Estart** are run.)

Note: Whenever a new switch primary is assigned, the old primary must be rebooted or powered off.

The following scenario describes how to select a primary node on a High Performance Switch system.

The initial value for the primary node defaults to 1 upon powering up the system. This value is placed in the SDR, but there is no active primary node until **Estart** is executed.

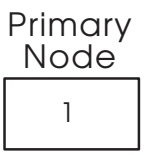


Figure 15. Initial Value

You can change the primary node value in the SDR by executing **Eprimary**. The specified node value is placed in the SDR, but there is still no active primary node.

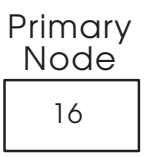


Figure 16. Eprimary Value

When **Estart** is executed, the primary node specified in the SDR becomes the active primary node.

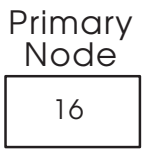


Figure 17. Estart Value

SP Switch Systems:

On SP Switch systems, select two standard nodes, one to be used as the primary node and the other to be used as the primary backup node. Note that a dependent node cannot be a primary or primary backup node for the SP Switch. Until the next **Estart** is issued, these nodes are reflected as the oncoming primary and oncoming primary backup nodes (**Eprimary**). Once **Estart** completes, the primary field is updated based on the oncoming primary field and the backup field is updated based on the oncoming backup field.

The criteria to select a new primary backup node follows:

1. First, select a node on a switch assembly other than the switch assembly to which the primary node is attached.
2. Second, if no other switch assembly exists, select a node attached to a switch chip other than the one to which the primary node is attached.
3. If no other switch chip exists, select any available node on the switch chip to which the primary node is attached.

The primary and primary backup fields are updated automatically when primary node and primary backup node failures occur. Therefore, the primary and primary backup fields reflect the state of the current system.

The following scenario describes how to select a primary backup node on an SP Switch system.

Initially, the primary node and backup primary node fields are blank. In a 16-node system, the oncoming primary node field has the default value of 1 and the oncoming primary backup has the default value of 16.

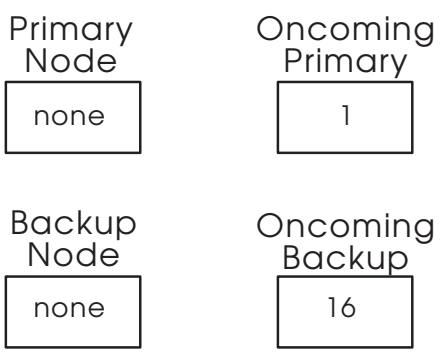


Figure 18. Initial Values

When **Estart** is executed, the node specified in the oncoming primary field becomes the primary node. The node specified in the oncoming primary backup becomes the primary backup.

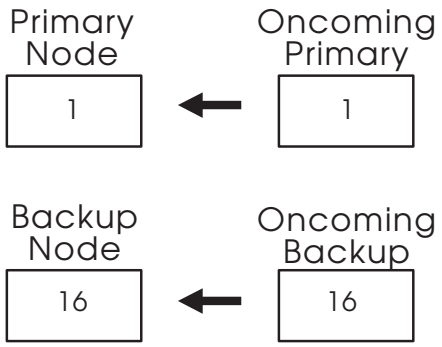


Figure 19. Estart Values

If the primary backup node fails, the primary node automatically selects a replacement.

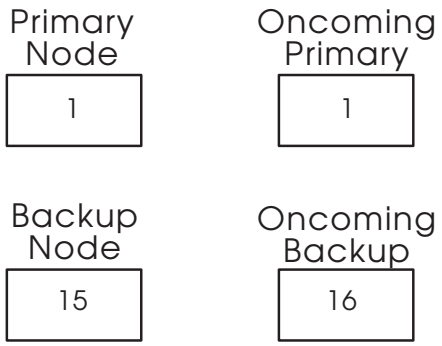


Figure 20. Backup Takeover Values

If the primary node fails, the primary backup node automatically becomes the primary node and a new primary backup is selected.

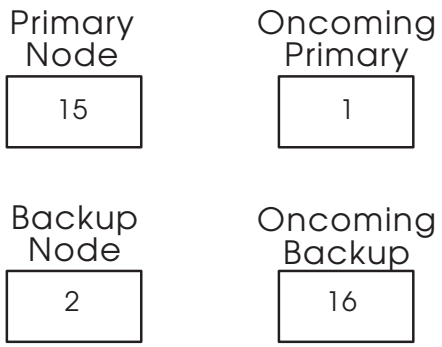


Figure 21. Primary Takeover Values

In summary, primary and primary backup fields reflect the current state of the system and the oncoming fields are not applicable until the next invocation of the **Estart** command.

To select the primary and primary backup nodes using SMIT:

TYPE	**smit**
	▲ The System Management menu appears
SELECT	RS/6000 SP Systems Management
	▲ The RS/6000 SP Systems Management menu appears.
SELECT	RS/6000 SP Cluster Management
	▲ The RS/6000 SP Cluster Management menu appears.
SELECT	Perform Switch Operations
	▲ The Perform Switch Operations menu appears.
SELECT	Set Primary/Primary backup Node
	▲ The Set Primary/Primary backup Node menu appears.

Using the Command Line: On High Performance Switch systems, to select a switch primary node by IP address, use the command:

```
Eprimary 129.33.34.1
```

On SP Switch systems, to select an oncoming switch primary node and an oncoming switch primary backup node by IP address, use the command:

```
Eprimary 129.33.34.1 -backup 129.33.34.56
```

Managing the Switch Topology File

The switch topology file describes the wiring configuration for the switch; it contains node-to-switch or switch-to-switch cable information. You choose a switch topology file based on the number of **switches** in your SP system. (For more information on choosing a switch topology file, refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.) You can annotate connection labels onto the selected switch topology file with the **Eannotator** command. Once you have selected and annotated the switch topology file, you store it in the System Data Repository (SDR) with the **Etopology** command. When you start the switch, the switch initialization code reads the switch topology file stored in the SDR. (Refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on the **Eannotator** and **Etopology** commands.)

An **expected.top** file in the /**etc**/**SP** directory of the primary node overrides the topology file specified in the SDR. (This is used when servicing or debugging the switch. Refer to the *IBM RS/6000 SP: Maintenance Information Volume 2* for more information.)

Annotating a Switch Topology File

Before a switch topology file is stored in the SDR, it should be annotated with **Eannotator**. **Eannotator** will update the switch topology file's connection labels with their current physical locations. If the **-O yes** flag is specified, the topology file is stored in the SDR. Running **Eannotator** makes the switch hardware easier to debug because the switch diagnostics information is printed based on physical locations. The base topology files only contain logical locations. **Eannotator** is not a necessary step for switch operations. On SP Switch systems, however, **Eannotator** is required to update jack information for switch diagnostics. For more

details on **Eannotator**, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Using SMIT:

TYPE **smit**

▲ The System Management menu appears

SELECT RS/6000 SP Systems Management

▲ The RS/6000 SP Systems Management menu appears.

SELECT RS/6000 SP Cluster Management

▲ The RS/6000 SP Cluster Management menu appears.

SELECT Perform Switch Operations

▲ The Perform Switch Operations menu appears.

SELECT Topology File Annotator

▲ The Topology File Annotator menu appears.

At this menu:

- Input the name of the topology file to be annotated
- Input the name to which you want the annotated topology file stored
- Enter **yes** to storing the topology file in the SDR

Using the Command Line: To annotate a switch topology file, use the command:

```
Eannotator -F input_file -f output_file -O yes
```

For example, to annotate a topology file for a two-switch or maximum 32-node system:

```
Eannotator -F expected.top.2nsb.0isb.0 -f expected.top.2nsb.0isb.0 -O yes
```

Storing the Switch Topology File in the SDR

Use **Etopology** to store the topology file in the SDR in the current system partition:

Using SMIT:

TYPE **smit**

▲ The System Management menu appears

SELECT RS/6000 SP Systems Management

▲ The RS/6000 SP Systems Management menu appears.

SELECT RS/6000 SP Cluster Management

▲ The RS/6000 SP Cluster Management menu appears.

SELECT Perform Switch Operations

▲ The Perform Switch Operations menu appears.

SELECT Fetch/Store Topology File

▲ The Fetch/Store Topology File menu appears.

Input the topology file for your configuration.

Using the Command Line: To store a topology file, use the **Etopology** command:

```
Etopology topology_filename
```

For example, to store a topology file for a two-switch or 32-node maximum configuration:

```
Etopology expected.top.2nsb.0isb.0
```

Starting the Switch

Once all the processor nodes are running, you can start the switch.

1. Check that all processor nodes are running (hostResponds indicator). Refer to "Determining If Nodes Are Operating: All Node Summary" on page 322 for a discussion of the ways you can use the System Monitor to do this.
2. Start the switch. On the control workstation, enter:

   ```
   Estart
   ```

   (**fault_service_Worm_RTG**, the fault-handling daemon for the High Performance Switch, and **fault_service_Worm_RTG_SP**, the fault-handling daemon for the SP Switch, that check, initialize, and prepare the switch for operation, must be running on the primary node.)

   Note: Before you do an **Estart**, use the **rc.switch** command to start the daemon.

   When **Estart** finishes, it sends a message similar to the following example verifying that it completed:

   ```
   Switch initialization started on tserv10.hpssl.kgn.ibm.com
   Initialization successful for 16 nodes
   Switch initialization completed
   ```

   If the switch initialization process encounters any problems, other messages may appear. If this happens, refer to diagnosis information in the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* and resolve any errors. See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on **Estart**.

Determining Switch Connectivity

You can monitor switch connectivity to the nodes by using SP Perspectives. For specific information on how to do this, view the online help available from the SP Perspectives application.

You can also monitor switch connectivity to the nodes by using the "All Node Summary" for the **switchResponds** SP System Monitor variable. (See "Determining If Nodes Are Operating: All Node Summary" on page 322 for additional information.)

On SP Switch systems, the node's status is represented by the following three colors within the "All Node Summary":

green Indicates that a node is active on the switch.

red Indicates that a node is not active on the switch for one of the following reasons:

- It was “fenced” without the **-autojoin** option
- It faulted and was removed from the switch
- The node's adapter failed to configure properly
- The node's SP Switch adapter was recently reset by the **/usr/lpp/ssp/css/rc.switch** script

To determine the reason why the error occurred, query the **switch_responds** object in the SDR as follows:

```
SDRGetObjects switch_responds
```

If the **adapter_config_status** attribute that is displayed is anything other than **css_ready**, the node's adapter failed to configure for the reason indicated.

If all the attributes displayed are zero (except for the **adapter_config_status = css_ready**), your SP Switch adapter was recently reset. A node in this state will join the switch at the next initialization, by either an explicit issuance of the **Estart** command or a recovery initialization driven by the switch manager within the Communication Subsystem.

If only the **isolated** attribute is set to 1, the node was removed from the switch because it was either “fenced” without the **-autojoin** option by the system administrator or because it faulted.

To determine if a node was “fenced” by the system administrator, examine the **/var/adm/SPlogs/css/Ecommands.log** log of switch-related commands on the control workstation for a recent entry of an **Efence** command specifying this node. If no entry references the node in question, refer to the “Diagnosing Switch Problems” chapter in *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* to determine why the node faulted.

yellow Indicates that a node is not active on switch because it was “fenced” with the **-autojoin** option.

To determine if a node was “fenced” with the **-autojoin** option by the system administrator, examine the **/var/adm/SPlogs/css/Ecommands.log** log of switch-related commands on the control workstation for a recent entry of an **Efence** command specifying the node in question.

For High Performance Switch systems, the node's status is represented by the following three colors within the “All Node Summary”:

green Indicates that a node is active on the switch.

red Indicates that a node is not active on the switch. Refer to the “Diagnosing Switch Problems” chapter in *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* to determine why this node is not active on the switch.

yellow Indicates that a node is not active on switch because it was "fenced."

To determine if a node was "fenced" by the system administrator, examine the **/var/adm/SPlogs/css/Ecommands.log** log of switch-related commands on the control workstation for a recent entry of an **Efence** command specifying the node in question.

Fencing and Unfencing Nodes from the Switch

The node fencing function provides for isolating nodes that will be rebooted or powered-off from the switch in order to prevent such nodes from impacting other nodes. The capability of allowing the node to rejoin the switch once the node is again operational is also provided.

Notes:

1. The primary node and primary backup node for an SP Switch cannot be fenced.
2. The primary node cannot be fenced on High Performance Switch systems.
3. If you specify the **-autojoin** option, a node cannot return via **Eunfence** to the switch network if a switch initialization (**Estart**) or switch fault occurred while the node was fenced. If this condition occurs, you must perform another switch initialization to return the node to the High Performance Switch. For the SP Switch, a node fenced with the **-autojoin** option will rejoin the switch on a switch fault, node reboot, or **Estart**.
4. A node cannot return to the High Performance Switch network via **Eunfence** if a switch initialization (**Estart**) or switch fault occurred while the node was isolated. If this condition occurs, you must perform another switch initialization to return the node to the High Performance Switch. For the SP Switch, a fenced node can only be returned to the switch via **Eunfence** regardless of whether a fault or **Estart** has occurred.

Using SMIT:

TYPE **smit**

▲ The System Management menu appears

SELECT RS/6000 SP Systems Management

▲ The RS/6000 SP Systems Management menu appears.

SELECT RS/6000 SP Cluster Management

▲ The RS/6000 SP Cluster Management menu appears.

SELECT Perform Switch Operations

▲ The Perform Switch Operations menu appears.

SELECT Fence / Unfence SP Nodes

▲ The Fence / Unfence SP Nodes menu appears.

Step 1: Fence the Node from the Switch

On the control workstation, use the **Efence** command to fence nodes. For example:

```
Efence node1 node2
```

When you issue the **Efence** command, you can use **-autojoin** to optionally specify if the nodes should rejoin the switch once they are again operational. For example:

```
Efence -autojoin node1 node2
```

A return code will indicate success or failure. In addition, a message will be displayed to indicate success or failure.

Step 2: Deactivate the Node

Enter the appropriate command to reset, power off, or shut down the node.

Step 3: Return the Node to the Switch

If you did not specify that the node automatically rejoin the switch in "Step 1: Fence the Node from the Switch," use the **Eunfence** command to return the node to the switch network. For example:

```
Eunfence node1 node2
```

A return code will indicate success or failure. In addition, a message will be displayed to indicate success or failure.

Setting the Run Phase of the High Performance Switch

Node isolation requests are queued until they can be serviced. Such requests can be serviced only at the end of the current "run phase" of the High Performance Switch. Therefore, the duration of the run phase determines how quickly the system responds to node isolation requests. Use the **Eduration** command to set the run phase for the switch in days, hours, or minutes. For example:

```
Eduration [days days] [hours hours] [minutes minutes] [-h]
```

Optionally, you can use a SMIT panel to set the run phase duration.

Using SMIT

TYPE	**smit**
	▲ The System Management menu appears
SELECT	RS/6000 SP Systems Management
	▲ The RS/6000 SP Systems Management menu appears.
SELECT	RS/6000 SP Cluster Management
	▲ The RS/6000 SP Cluster Management menu appears.
SELECT	Perform Switch Operations
	▲ The Perform Switch Operations menu appears.
SELECT	Change/Show Switch Run Phase Duration
	▲ The Change/Show Switch Run Phase Duration menu appears.
SELECT	Set Switch Run Phase Duration
	▲The Set Switch Run Phase Duration menu appears.

See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on the **Efence**, **Eunfence**, and **Eduration** commands.

Preparing an SP Switch for Repartitioning

Use the **Eunpartition** command to prepare the current system partition for repartitioning. For example:

```
Eunpartition
```

Optionally, you can use a SMIT panel to prepare a switch for repartitioning.

Using SMIT

TYPE	**smit**
	▲ The System Management menu appears
SELECT	RS/6000 SP Systems Management
	▲ The RS/6000 SP Systems Management menu appears.
SELECT	RS/6000 SP Cluster Management
	▲ The RS/6000 SP Cluster Management menu appears.
SELECT	Perform Switch Operations
	▲ The Perform Switch Operations menu appears.
SELECT	Prepare a switch network for repartitioning
	▲ The Prepare a switch network for repartitioning menu appears.

See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on the **Eunpartition** command.

Monitoring Nodes on a Switch

You can monitor and attempt to maximize the availability of nodes on a switch with the **Emonitor** daemon. **Emonitor** is controlled by the System Resource Controller (SRC). One instance of the daemon exists for each system partition and is named **Emonitor**.*partition_name*. A system-wide configuration file, /**etc**/**SP**/**Emonitor.cfg** lists all node numbers (one per line) on the system to be monitored.

Emonitor is invoked with **Estart -m**. Once invoked, it is SRC controlled and so will restart if halted abnormally. To end monitoring, run /**usr**/**lpp**/**ssp**/**bin**/**emonctrl -s** to stop the daemon in the system partition.

For more information on **Emonitor**, refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Global Shutdowns and Reboots of Nodes with a Switch

The **Equiesce** command disables switch error recovery and primary node takeover. It is used to shut off the normal error actions when global activities are performed. For example, when nodes are shut down or rebooted, they are fenced from the switch.

It is possible that the primary node may fence all the other nodes, including the primary backup node, if it is not the first node to shut down during a global shutdown/reboot of the entire system. Primary node takeover can also occur if the primary node is shut down and the primary backup node remains up. Issuing the **Equiesce** command before the shutdown will prevent these situations from occurring.

The **Equiesce** command causes the primary and primary backup nodes to shut down their recovery actions. Data still flows over the switch, but no faults are serviced and primary node takeover is disabled. Only the **Eannotator**, **Eclock**, **Eprimary**, **Estart**, and **Etopology** commands are functional after the **Equiesce** command is issued.

Estart must be issued when the global activity is complete to reestablish switch recovery and primary node takeover.

Use the **Equiesce** command to disable switch error recovery and primary node takeover. For example:

```
Equiesce
```

Optionally, you can use a SMIT panel to quiesce the SP Switch.

Using SMIT

TYPE	**smit**
	▲ The System Management menu appears.
SELECT	RS/6000 SP Systems Management
	▲ The RS/6000 SP Systems Management menu appears.
SELECT	RS/6000 SP Cluster Management
	▲ The RS/6000 SP Cluster Management menu appears.
SELECT	Perform Switch Operations
	▲ The Perform Switch Operations menu appears.
SELECT	Switch Quiesce
	▲ The **Equiesce** command is run.

See *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on the **Equiesce** command.

NFS Mounts Over a Switch

We strongly discourage NFS mounts over the switch in the root (/) directory. If a mount is necessary in the root (/) directory, place the NFS mount elsewhere in the file system with a symbolic link into the root (/) directory. For example:

```
mount -o intr,hard server:/myfs /otherdir/myfs
ln -s /otherdir/myfs /myfs
```

Chapter 17. Managing System Partitions

System partitioning is a method for organizing the SP system into non-overlapping groups of nodes for various purposes such as testing new software and creating multiple production environments. In this chapter we present information on:

- Understanding system partitioning
- Partitioning the system
- Displaying system partition configuration information
- Verifying system partition configuration
- Managing system partition-sensitive subsystems using the **syspar_ctrl** command

You will also find information on system partitioning in the *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment* and the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*. For information on accounting and system partitioning, refer to "Accounting and System Partitioning" on page 144.

In this chapter we presume that you have installed the IBM Parallel System Support Programs for AIX (PSSP) software. In doing so you have automatically created a single default system partition which consists of all the nodes in your SP system. If you now wish to select an alternate system partition configuration, the information in this chapter will guide you in selecting, customizing, and applying a system partition configuration.

Understanding System Partitioning

Again, system partitioning is a method for organizing an SP system into non-overlapping groups of nodes that can be dedicated to specific purposes. A group of such nodes (not including the control workstation) is a *system partition*. For example, you may wish to configure a system partition to use for non-disruptive testing when you migrate to a new release of software. Or you may wish to configure system partitions for multiple isolated production environments so that the workload in one environment will not interfere with the workload in another environment. Isolation of production environments is achieved by partitioning the switch in such a way that communication paths for different system partitions do not cross each other.

Note that dependent nodes can be part of a system partition along with standard nodes.

System partitions appear to most subsystems and for most user tasks as logical SP systems. It is important to understand that from an administrative point-of-view, each system partition is a logical SP system within one administrative domain.

On the control workstation, the administrator is in an environment for *one system partition at a time*. The SP_NAME environment variable identifies the system partition to subsystems. (If this environment variable is not set, the system partition is defined by the `primary:` stanza in the **/etc/SDR_dest_info** file.) Most tasks performed on the control workstation that get information from the SDR will get the information for *that particular system partition*.

© Copyright IBM Corp. 1995, 1997

In managing multiple system partitions it is helpful to open a window for each system partition. You can set and export the SP_NAME environment variable in each window and set up the window title bar or shell prompt with the system partition name. The following script is an example:

```
sysparenv:
# !/bin/ksh
  for i in 'splst_syspars'
  do
     syspar='host $i | cut -f 1 -d"."'
     echo "Opening the $syspar partition environment"
     sleep 2
     export SP_NAME=$syspar
     aixterm -T "Work Environment for CWS 'hostname -s' - View: $syspar" -ls -sb &
  done
  exit

.profile addition:
# Added for syspar environment setup
  if [ "'env | grep SP_NAME | cut -d= -f1'" = SP_NAME ]
     then
        PS1="['hostname -s'<p>"$SP_NAME] ['$PWD]> '
     else
        PS1="['hostname -s']["'$PWD]> '
  fi
  export ENV
```

As a user, you can check what system partition you're in with the command:

```
spget_syspar -n
```

System partitioning can be achieved either by applying one of a fixed set of supplied configurations (refer to the *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment* for specific information on supplied configurations) or by creating a different system partition configuration and applying it. The **ssp.top** option in the **pssp** install package contains a directory structure with all relevant files for partitioning the system with supplied system partition configurations. In addition, the **ssp.top** option contains the System Partitioning Aid, a tool for creating customized system partitions to suit your needs. This option must be installed on the control workstation if you are partitioning your system into more than one system partition.

SP Switch Usage Note

If you are repartitioning your system and it has an SP Switch, you must run **Eunpartition**. Failure to do this will produce unpredictable **Estart** results in the new system partitions. The **Estart** command results can range from only the old system partition portion of the new system partition being accessible on the switch, to a complete **Estart** timeout failure. Various recovery scenarios are available. A guaranteed recovery can be accomplished by:

1. Issuing **Eclock** to reset the switch
2. Rebooting all the nodes or issuing a **css_restart_node** on all nodes
3. Issuing **SDRChangeAttrValues switch_responds isolated=0** from the global view of the system monitor (or in each system partition)
4. Issuing **Estart** in each of the system partitions

This recovery scenario is disruptive to all system partitions, even those unaffected by the repartition.

Before You Define System Partitions

In the beginning, after you have installed the PSSP software, but before you have configured system partitions, there is one system partition corresponding to the name and IP address of the control workstation returned by the **hostname** command. This is the *default* or *persistent* system partition. It always exists. When you configure more than one system partition, one of them must be defined as the default system partition with this name and IP address.

When configuring more than one system partition, you must define alias IP addresses and names for all but the default system partition. To configure an alias IP address, you use the **ifconfig** command with the alias parameter to establish an additional IP address for the interface which the control workstation's **hostname** maps to.

Note: The alias must be defined on the same subnet as the interface that the control workstation **hostname** maps to.

You should also add alias names that map to these alias IP addresses through your name resolution mechanism. The alias IP addresses should be added to **rc.net** so that they are automatically configured when the control workstation reboots.

The steps for defining alias IP addresses and names are:

1. In the Part II - Traditional Configuration section of the /**etc**/**rc.net** file, a template is provided for changing **inet0** to an alias. Follow the template and edit the file to define alias IP addresses and names. (After editing the /**etc**/**rc.net** file, ensure that you have execute permission on the file or the control workstation will not boot correctly.) For example:

   ```
   /usr/sbin/ifconfig tr0 alias 129.40.127.101 netmask 255.255.255.0 up >>$LOGFILE 2>&1
   ```

2. Execute the /**etc**/**rc.net** script. Alternatively, you could enter the **ifconfig** command directly. For example:

   ```
   ifconfig tr0 alias 129.40.127.101 netmask 255.255.255.0 up
   ```

3. Validate that the alias is defined correctly (that is, that the same interface is defined for the alias and the control workstation host name). For example, to list the IP addresses and aliases on your system:

```
netstat -in
tr0   129.40.127   129.40.127.43
tr0   129.40.127   129.40.127.101
```

Note: When you add an alias for the control workstation for the purpose of defining a name for a new system partition, you must run the **setup_server** command on the control workstation to properly define the new **rcmd** principal associated with the new host name alias.

Note: To delete the alias we defined in the previous example, the command is:

```
/usr/sbin/ifconfig tr0 delete 129.40.127.101
```

Remember to comment out or delete any entries you made to define aliases in the **/etc/rc.net** file.

In the next section we discuss how you select, customize, and apply a system partition configuration.

Partitioning the SP System

You partition your SP system by choosing a configuration for your size system. The steps involved in partitioning an SP system are:

1. Check if the desired system partition configuration is supplied.
2. If the desired system partition configuration is not supplied, use the System Partitioning Aid to generate and save a system partition configuration.
3. Archive the System Data Repository.
4. Select a system partition configuration and layout to apply.
5. Customize (name and specify attributes) the system partitions.

   Note: If you are repartitioning your system, and your system contains an SP Switch, you must run **Eunpartition** in each of the partitions that are going to change. This prepares the switch in those partitions for repartitioning.
6. Shut down all nodes in changing system partitions.
7. Run the **setup_server** command on the control workstation.
8. Apply the configuration.
9. Reboot all nodes in changed system partitions.

We refer to the above steps as *partitioning* the SP system. However, if you wish to change or add system partitions, that is, to *repartition*, the steps are the same. For example, if you wish to modify an existing system partition or create a new partition for a new production environment, you would follow these steps to repartition the system.

How to Start

If you know that the desired system partition configuration exists, start with "Step 3: Archive the System Data Repository" on page 245; otherwise, start with "Step 1: Check if the Desired System Partition Configuration is Supplied" on page 244.

For partitioning or repartitioning the system, you have the option of using SMIT panels or commands. SMIT panels can be accessed through the **smit** command or through the SP Perspectives GUI. (For more information, and further examples, of the commands used in this chapter, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.)

Note

Some of the commands used in system partitioning (**spdisplay_config**, **splstdata -p**, **spapply_config -v**, for example) produce lengthy output. It is helpful to log the output of such commands. For example, to execute the **spapply_config** command while simultaneously logging a copy of the output to the file /**tmp**/**syspar.log**, enter:

```
spapply_config -v config.4_4_8/layout.3 2>&1 | tee /tmp/syspar.log
```

Using SMIT:

TYPE **smit**

▲ The System Management menu appears.

SELECT RS/6000 SP System Management

▲ The RS/6000 SP System Management menu appears.

SELECT RS/6000 SP Configuration Database Management

▲ The RS/6000 SP Configuration Database Management menu appears.

SELECT Enter Database Information

▲ The Enter Database Information menu appears.

SELECT System Partition Configuration

▲ The System Partition Configuration menu appears.

Using the SP Perspectives GUI to Access SMIT:

TYPE **perspectives &**

▲ The SP Perspectives Launch Pad appears.

SELECT smit config_data on CWS

▲ The RS/6000 SP Configuration Database Management menu appears in a separate AIX window.

SELECT Enter Database Information

▲ The Enter Database Information menu appears.

SELECT System Partition Configuration

▲ The System Partition Configuration menu appears.

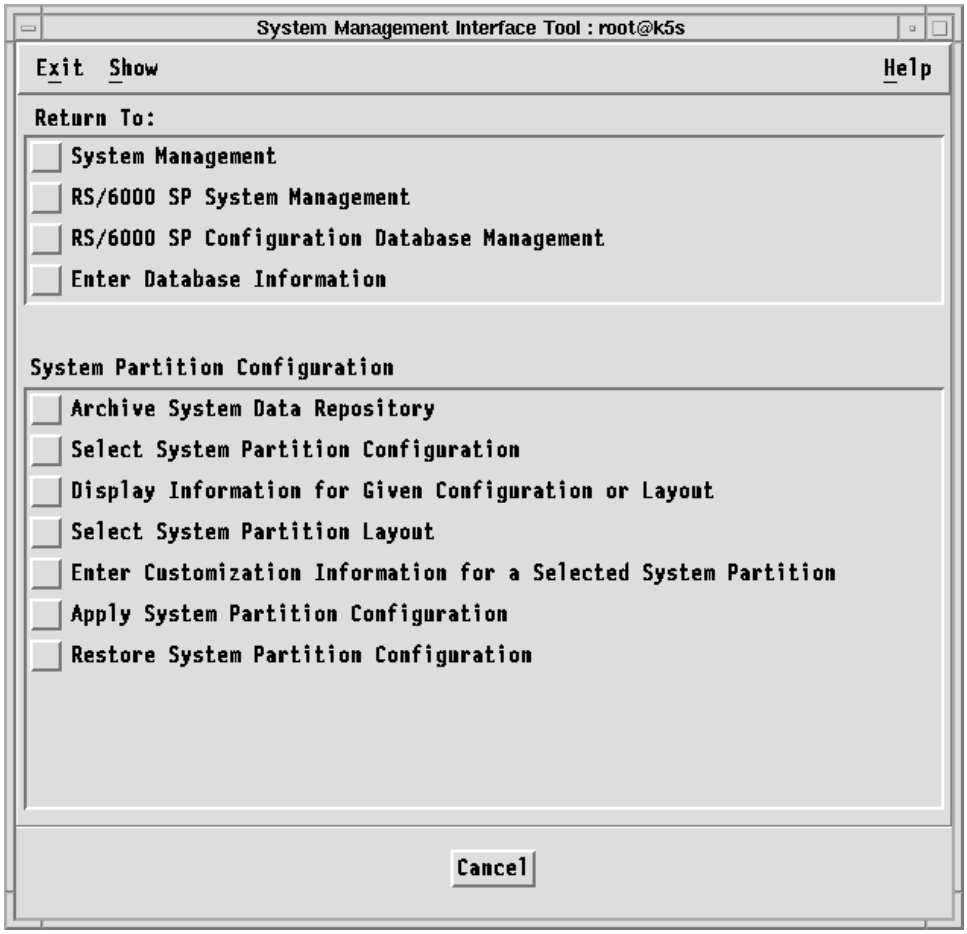


Figure 22. System Partition Configuration SMIT Panel

Step 1: Check if the Desired System Partition Configuration is Supplied

Use the System Partitioning Aid to check if the desired system partition configuration is supplied. The System Partitioning Aid can be invoked from the command line or from the SP Perspectives GUI. In this section, we will provide examples of using the command line interface. (For information on the SP Perspectives GUI to the System Partitioning Aid, invoke the Aid with the **spsyspar** command and view the online help that is available.) For additional information on using the System Partitioning Aid, refer to *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment*.

To check if the desired system partition configuration is supplied:

1. Enter information for your desired system partition configuration into a file. The template for this file is **/spdata/sys1/syspar_configs/bin/inpfile.template**.
2. Once you have created the input file, enter:

   ```
   sysparaid input_file
   ```

3. If the desired system partition configuration is found, layout directory information is displayed. Go to “Step 3: Archive the System Data Repository” on page 245.
4. If the desired system partition configuration is not found, a message is displayed indicating so and stating whether the desired system partition configuration is possible or not possible. If the system cannot be partitioned as desired, an error message is output indicating the reason for the failure. If the desired system partition configuration is possible, a snapshot of the switch chips in the system and performance data for each system partition is output to a set of files.
   a. If the system cannot be partitioned as desired, analyze the error message and create another input file specifying an alternative system partition configuration and rerun the **sysparaid** command.
   b. If the desired system partition configuration is possible, review the performance data that is output and, if it is acceptable, proceed to “Step 2: Generate and Save the Desired System Partition Configuration.”

For more information on the **sysparaid** command, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Step 2: Generate and Save the Desired System Partition Configuration

Once you have determined that the system partition configuration you desire is possible, generate and save it:

```
sysparaid -s layout_name input_file
```

The newly generated system partition configuration layout will be saved to the **spdata/sys1/syspar_configs** directory.

The fastpath invocation for the System Partition Configuration menu is:

```
smit syspar
```

Step 3: Archive the System Data Repository

Always archive the System Data Repository before partitioning (or repartitioning) your SP system. If you change your mind once you have committed a system partition configuration, or if applying a system partition configuration fails, you should use the archived SDR to recreate the previous system partition configuration. (To do this, select the Restore System Partition Configuration option from the SMIT menu.)

Using SMIT: At the System Partition Configuration menu:

SELECT Archive System Data Repository

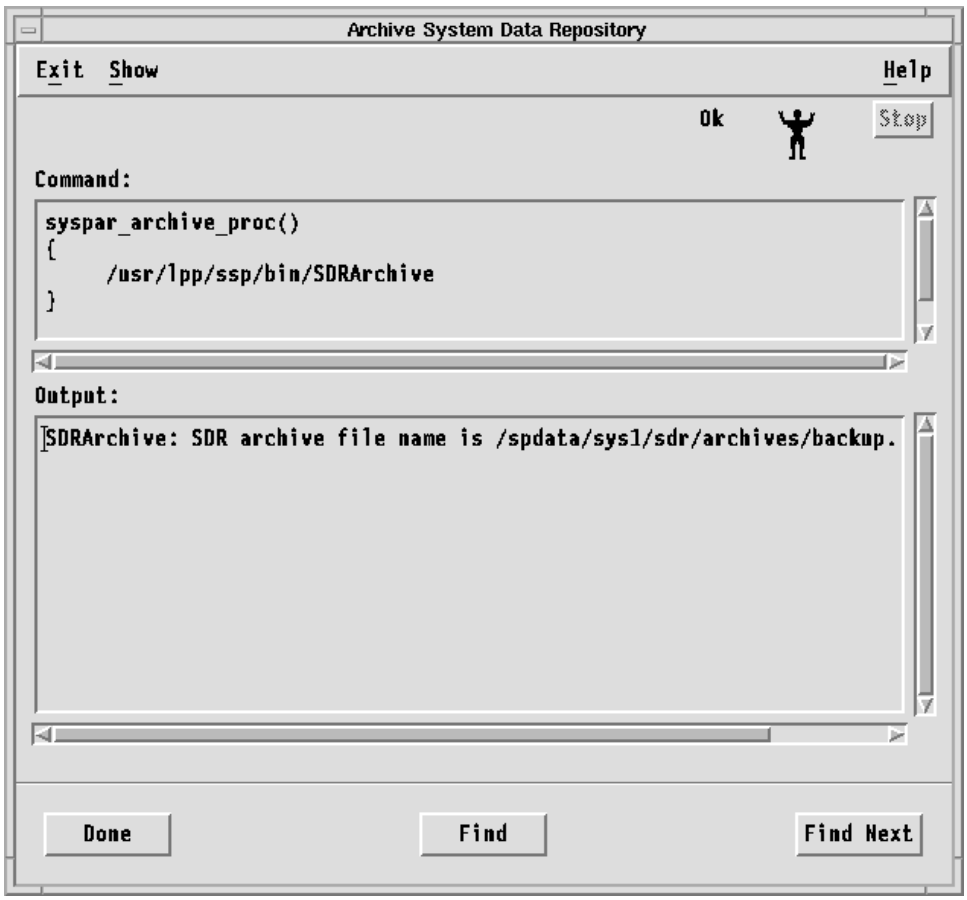


Figure 23. Archive System Data Repository SMIT Panel

Using the command line: Enter:

```
SDRArchive
```

Note: When you archive the SDR from either SMIT or by using the **SDRArchive** command, the archive produced is in **tar** format. Issuing **tar -x** for this archive does not restore the system partition configuration. To restore the system partition configuration, always use the SMIT Restore System Partition Configuration option.

Step 4: Select a System Partition Configuration to Apply

For detailed information on partition schemes for various size SP systems, refer to the chapter "Planning for System Partitions" in *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment*.

To select a system partition configuration to apply:

1. Select a configuration from a list of supported system partition configurations for your SP system
2. Display layout information for the selected system partition configuration
3. Select a system partition layout

Select a System Partition Configuration

To select a supported system partition configuration for your SP system,

Using SMIT: At the System Partition Configuration menu:

SELECT Select System Partition Configuration

▲ A list of system partition configurations appears. (If you have created layouts using the System Partitioning Aid, such layouts will also be listed on the SMIT panel.)

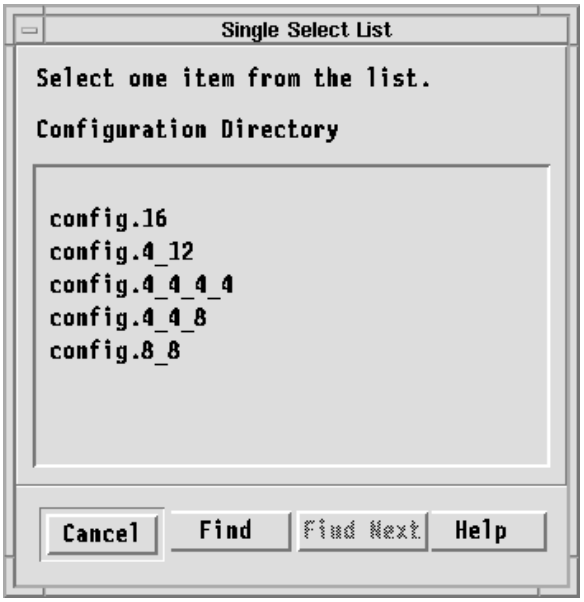


Figure 24. System Partition Configuration List SMIT Panel

SELECT A system partition configuration

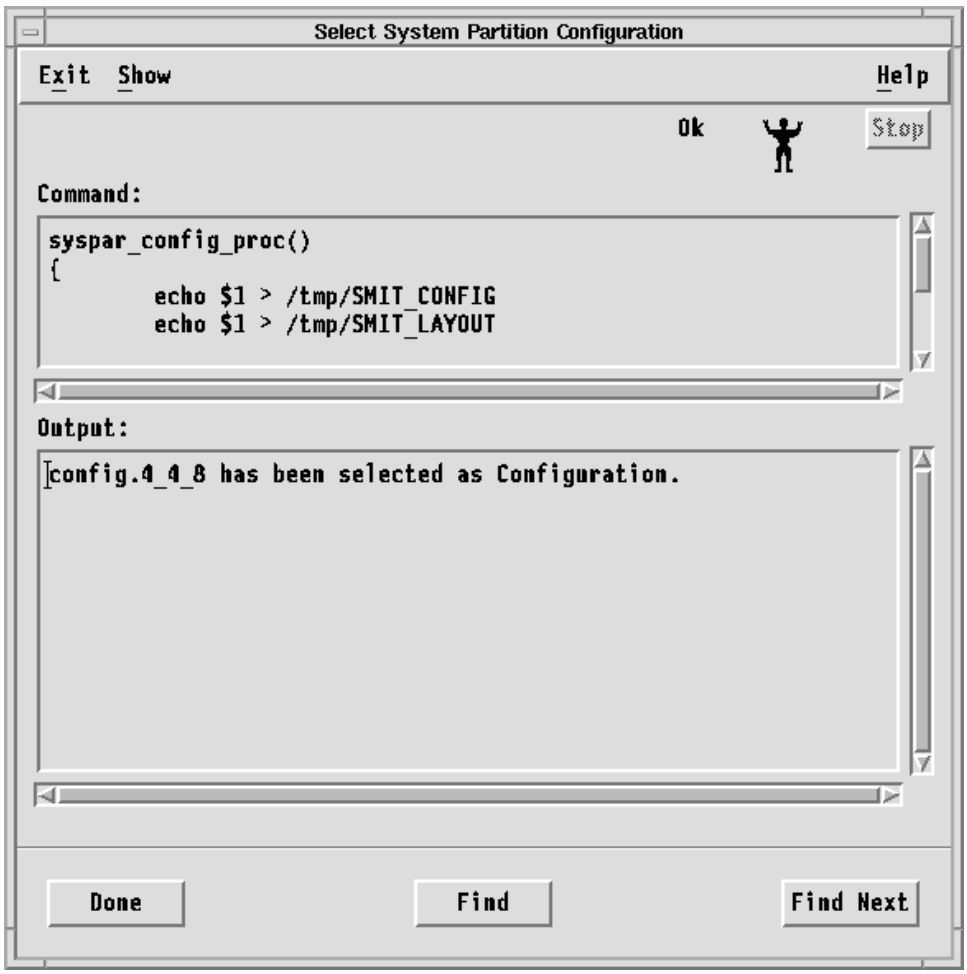


Figure 25. Select System Partition Configuration SMIT Panel

In this example we have selected a system partition configuration for our 16-node system of three system partitions consisting of four nodes, four nodes, and eight nodes respectively.

Using the command line: To display all supported system partition configurations for your SP system, enter:

```
spdisplay_config
```

Display Information for the Selected Configuration

To display information about the supported layouts for the system configuration you have selected,

Using SMIT: At the System Partition Configuration menu:

SELECT Display Information for Given Configuration or Layout

▲ The Display Information for Given Configuration or Layout menu appears

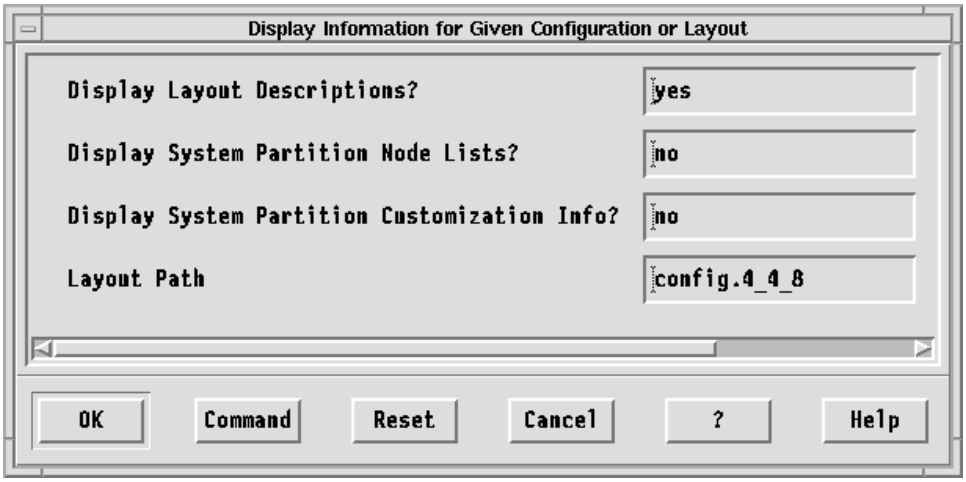


Figure 26. Display Information for Given Partition or Layout

SELECT OK

▲ System partition layout information for the chosen system partition configuration appears.

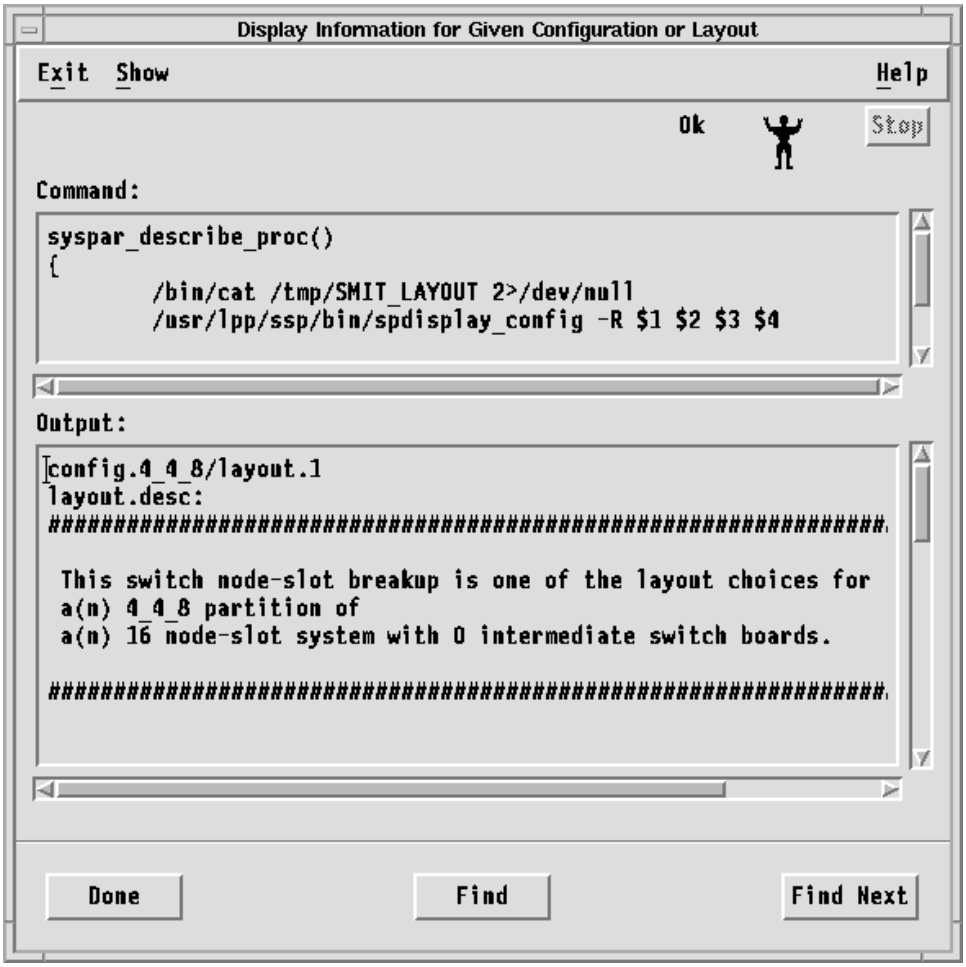


Figure 27. Display Information for Given Partition or Layout Output

(This example shows just the first screen of the layout information displayed for our selected 4-4-8 system partition configuration.)

Using the command line: To display information for the supported layouts for the 4-4-8 system partition configuration we've chosen, enter:

```
spdisplay_config -R -d config.4_4_8
```

Select a System Partition Layout

Based on the layout information you have displayed for the selected system partition configuration, select a layout,

Using SMIT: At the System Partition Configuration menu:

SELECT Select System Partition Layout

▲ A list of system partition layouts for the chosen system partition configuration appears. (If you have created layouts using the System Partitioning Aid, such layouts will also be listed on the SMIT panel.)

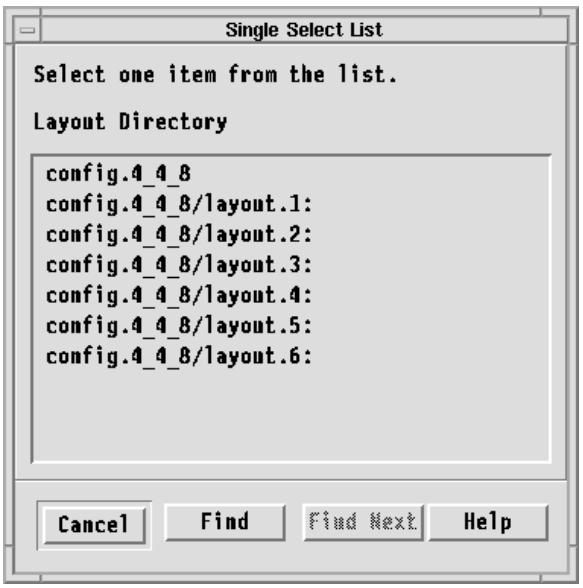


Figure 28. Select System Partition Layout SMIT Panel

SELECT A system partition layout

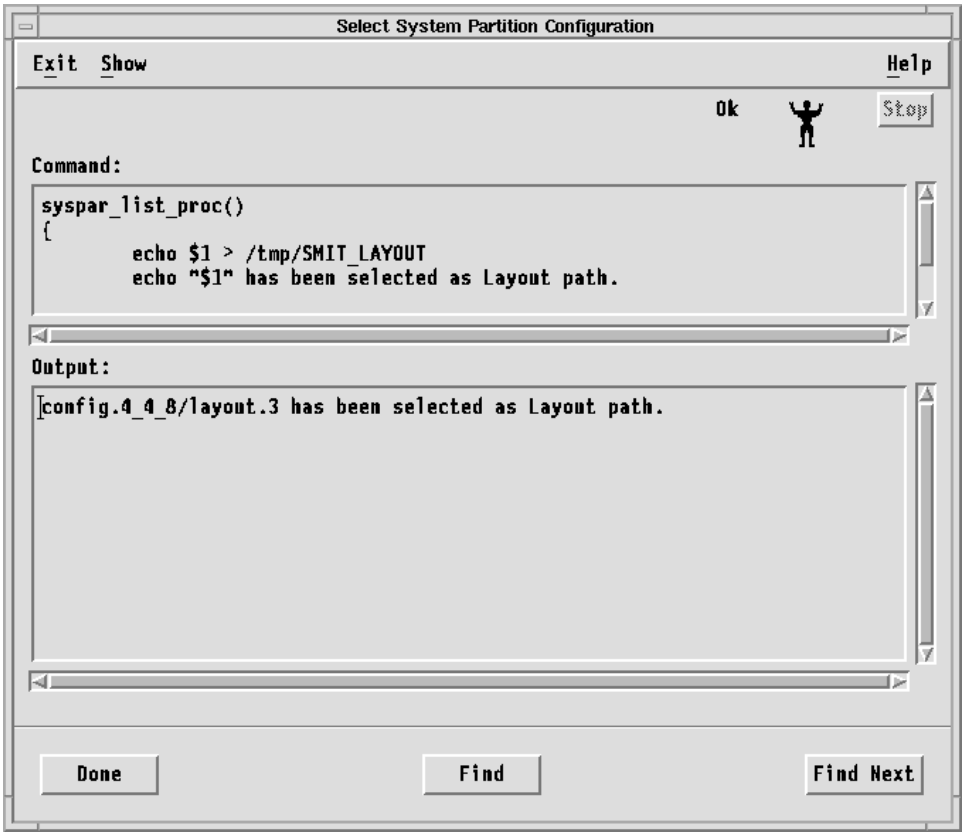


Figure 29. Select System Partition Layout Output SMIT Panel

In this example, we have chosen layout 3 for our selected 4-4-8 system partition configuration.

Step 5: Customize a System Partition

You customize a system partition by specifying information for:

- System partition name or IP address
- PSSP code level
- Partition default installation image (optional)
- Primary node (optional)

The customization information is retained in the **custom** file for the partition. Because the customization information is retained in the **custom** file, the customization step would not have to be repeated when doing a reconfiguration if the same system partition is being used and the customization information does not need to be changed. It is imperative that all new IP addresses and host names being used will be resolvable on your SP system.

Using SMIT: At the System Partition Configuration menu:

SELECT Enter Customization Information for a Selected System Partition

▲ A list of the system partitions for the chosen configuration and layout appears.

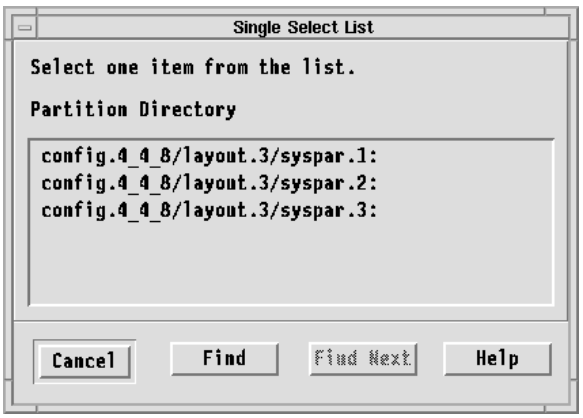


Figure 30. Partition Directory SMIT Panel

SELECT A system partition and enter customization arguments

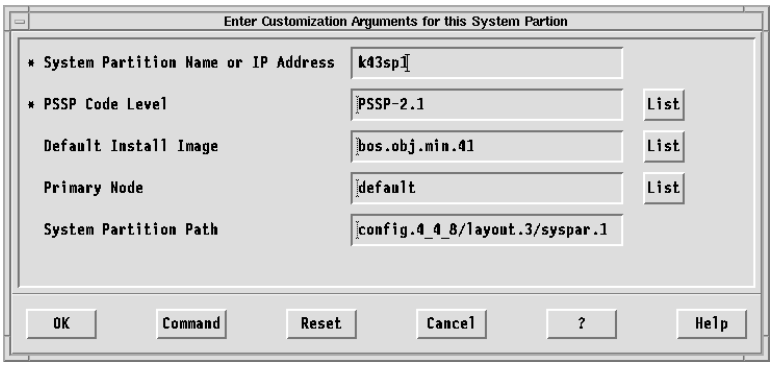


Figure 31. Enter Customization Arguments for this System Partition SMIT Panel

Note: If you have already customized this system partition this panel displays those values. Change the values as desired.

The customization arguments and their meanings are:

System Partition Name or IP Address
The name of the system partition is either the real host name (in the case of the default system partition) or the alias host name of the control workstation. The IP address of the system partition is either the real IP address (in the case of the default system partition) or the alias IP address of the control workstation.

PSSP Code Level
The IBM Parallel System Support Programs for AIX (PSSP) version and release, **PSSP-1.2** or **PSSP-2.1** or **PSSP-2.2** or **PSSP-2.3**. For system partitions that contain mixed levels of PSSP, this should be set to the minimum, or earliest, level installed in this system partition.

Default Install Image
The name of the **mksysb** image that is used to install a node in this system partition (if the node itself doesn't have an installation image value). For system partitions that contain mixed levels of PSSP, this default install image should be associated with the PSSP code level specified above.

Primary Node The node on the switch that initializes, recovers, provides diagnosis, and performs other operations to the switch network. The default is the first node in the **nodelist** file.

Backup Primary Node
Available only on SP Switch systems, the node designated as the oncoming switch primary node.

System Partition Path
The full path name for the configuration, layout, and system partition name.

Using the command line: To list the system partitions for our chosen 4-4-8 system partition configuration, layout 3, enter:

```
spdisplay_config config.4_4_8/layout.3
```

To enter customization information (as shown in our example SMIT panel) for system partition 1 of our chosen 4-4-8 system partition configuration, layout 3, enter:

```
spcustomize_syspar -n k43sp1 -l PSSP-2.1 -d bos.obj.min.41\
-e default config.4_4_8/layout.3/syspar.1
```

Repeat this step, using the appropriate values, for each system partition in the configuration.

Step 6: Shut Down All Nodes in Changing Partitions

When you are partitioning or repartitioning the system, system partitions are considered the same if all of the following are identical:

- Node list files
- Topology files
- IBM Parallel System Support Programs for AIX code level
- Default install image
- IP address
- Partition name

If any of the above items are not identical, the system partition is being changed and nodes in that partition should be shut down before applying the system partition configuration. If your system contains an SP Switch, you must invoke **Eunpartition** in each system partition that is being changed. See the **Eunpartition** command in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

Note: Modifying only the primary node will not result in a changed system partition. If you want to change the primary node, use **Eprimary** followed by **Estart**. See the **Eprimary** command in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

You can use the Verify option on the Apply System Partition Configuration SMIT panel to determine what nodes are changing. The output for the Verify option will list changing system partitions as well as any nodes in those system partitions that are not shut down.

Using SMIT: At the System Partition Configuration menu:

SELECT Apply System Partition Configuration

▲ The previously selected system partition configuration and layout appears

SELECT The system partition configuration and layout

▲ The Apply System Partition Configuration menu appears.

SELECT Verify only. from the list of System Partition Apply Options

SELECT OK

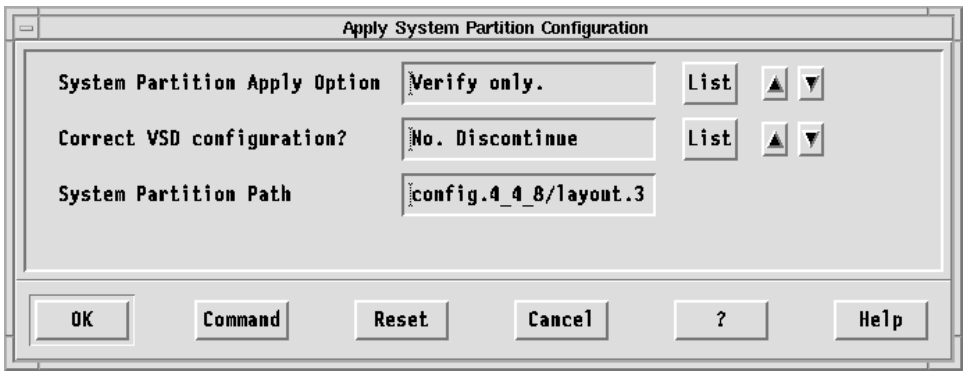


Figure 32. Verify System Partition Configuration SMIT Panel

The Apply System Partition Configuration arguments and their meanings are:

System Partition Apply Option
The options are provided to either only verify the system partition configuration or to actually apply the system partition configuration.

Correct VSD configuration?
The option to correct or discontinue when nonfatal errors are encountered in the IBM Virtual Shared Disk subsystem in the application of the specified system partition configuration.

System Partition Path
The full path name for the configuration, layout, and system partition name.

Using the command line: Enter:

```
spapply_config -v config.4_4_8/layout.3
```

Step 7: Run the set_server Command on the Control Workstation

If you have added an alias for the control workstation for the purpose of defining a name for a new system partition (see “Before You Define System Partitions” on page 241), you must run the **setup_server** command on the control workstation to properly define the new **rcmd** principal associated with the new host name alias.

Step 8: Apply a System Partition Configuration

When you apply or commit a system configuration, the following occurs:

- The existence of all required files (**nodelist**, **topology**, **custom**) for each system partition in the specified layout directory is verified. In addition, the contents of the **custom** file is checked for completeness.
- Any new system partitions are created on the control workstation.

- The system partition object for each system partition is created in the SDR.
- Affected node objects in the SDR (**Node**, **Adapter**, **host_responds**, **switch_responds**) and affected extension node objects in the SDR (**DependentNode** and **DependentAdapter**) are moved to the newly created partition.
- New **sdrd** daemons are created and started.
- New system partition-sensitive subsystems are created and started via the **syspar_ctrl** command.
- The **Syspar_map** information in the SDR is set for each node slot in the partition.
- The **verparvsd** command is invoked to verify the IBM Virtual Shared Disk data for system partitioning. The command extracts all IBM Virtual Shared Disk data from nodes involved in the system partitioning and writes SDR commands to update the IBM Virtual Shared Disk data to an output file. The output file reloads the IBM Virtual Shared Disk SDR data into the correct new partitions. (Note that the **verparvsd** command can also be used prior to applying a system partition configuration to test the desired layout for the existing IBM Virtual Shared Disk configuration defined in the SDR. See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more details.)
- Old unused system partitions on the control workstation, along with their **sdrd** daemons, are deleted and old system partition-sensitive subsystems are deleted.
- Updated SDR information is checked for accuracy against the information in the supplied system partition configuration files.

To apply a system partition configuration,

Using SMIT: At the System Partition Configuration menu:

SELECT	Apply System Partition Configuration
	▲ The previously selected system partition configuration and layout appears
SELECT	The system partition configuration and layout
	▲ The Apply System Partition Configuration menu appears.
SELECT	Apply this config.
SELECT	OK

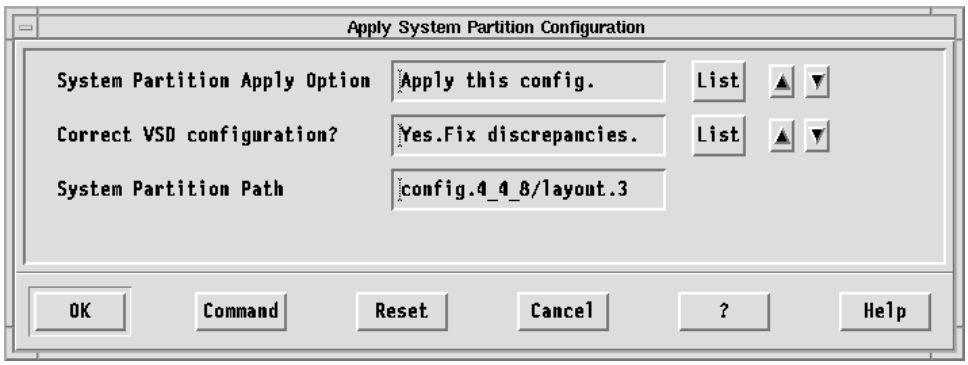


Figure 33. Apply System Partition Configuration SMIT panel

Using the command line: To apply our chosen 4-4-8 system partition configuration, layout 3, enter:

```
spapply_config -F config.4_4_8/layout.3
```

Important

If a failure occurs at this step, your system is put into an inconsistent state. Use the **sprestore_config** command or Restore System Partition Configuration SMIT menu option to restore the backed-up SDR and appropriate system partition-sensitive subsystems.

Note: Due to system partitioning changes, your SP_NAME environment variable may no longer be set to a valid system partition name. To get a list of valid system partition names, enter the **splst_syspars -n** command. Then verify that your SP_NAME environment variable is either unset or set to one of the partition names in the list.

Step 9: Reboot All Nodes in Changed Partitions

Reboot any nodes that were shut down in "Step 6: Shut Down All Nodes in Changing Partitions" on page 253.

Displaying System Partition Configuration Information

You can display the following information for the system partitions in your current configuration:

- A list of the system partitions in your current configuration
- For each system partition in your current configuration:
  - System partition name
  - IP address
  - Install image
  - Layout directory name
  - IBM Parallel System Support Programs for AIX code version

Using SMIT:

TYPE **smit**

▲ The System Management menu appears.

SELECT RS/6000 SP System Management

▲ The RS/6000 SP System Management menu appears.

SELECT RS/6000 SP Configuration Database Management

▲ The RS/6000 SP Configuration Database Management menu appears.

SELECT List Database Information

▲ The List Database Information menu appears.

SELECT List System Partition Database Information

▲ The list of system partitions along with information for each of the system partitions appear.

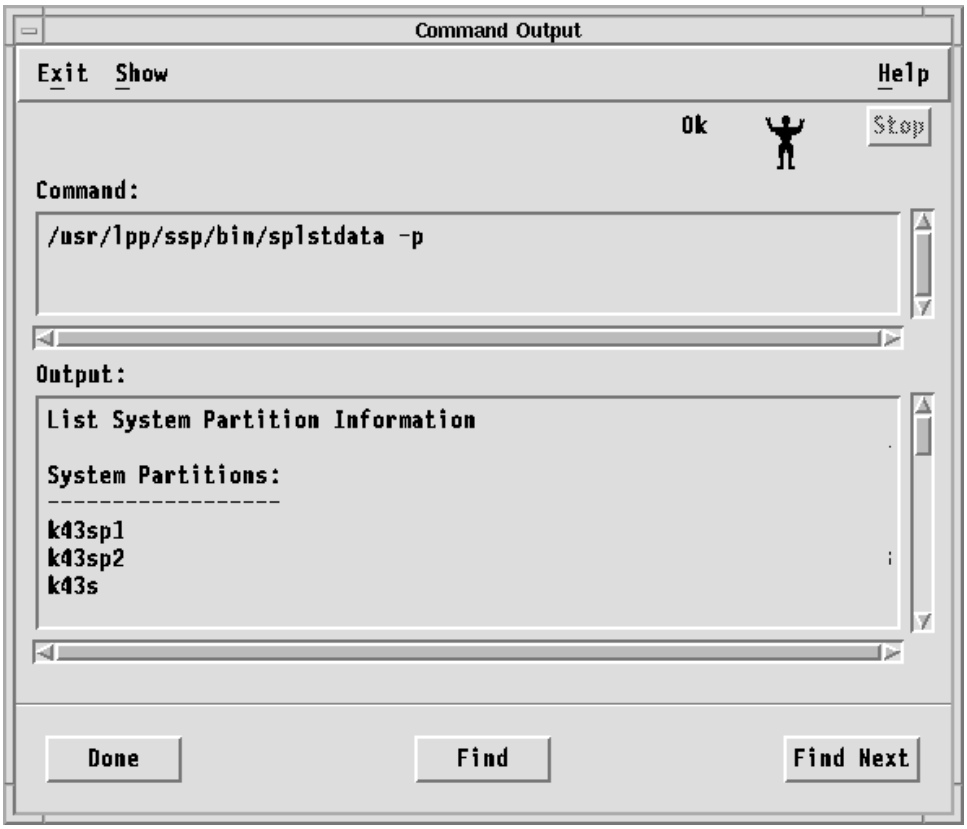


Figure 34. List System Partition Database Information Output

(This example shows just the first screen of the system partition configuration information for the system partition configuration we have applied.)

The fastpath invocation for the List System Partition Configuration Information menu is:

```
smit list_syspar
```

Using the command line: To display information for the system partitions in your current configuration, enter:

```
splstdata -p
```

Verifying System Partition Configuration

You can verify that the system partition information in the SDR is logically consistent.

Using SMIT

TYPE **smit**

▲ The System Management menu appears.

SELECT RS/6000 SP System Management

▲ The RS/6000 SP System Management menu appears.

SELECT RS/6000 SP Installation/Configuration Verification

▲ The RS/6000 SP Installation/Configuration Verification menu appears.

SELECT System Partition Configuration

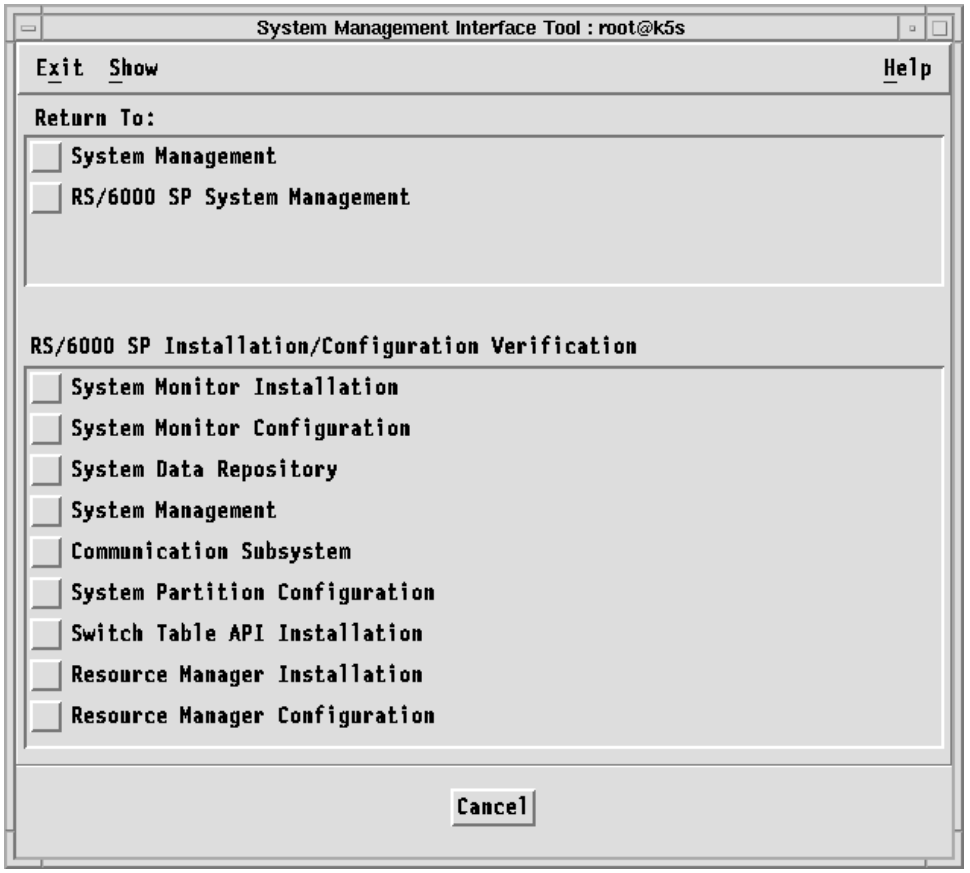


Figure 35. Verify System Partition Configuration SMIT Panel

The fastpath invocation for the Verify System Partition Configuration SMIT menu is:

```
smit SP_verify
```

Using the command line: Enter:

```
spverify_config
```

If the **spverify_config** command runs successfully, the SMIT status indicator will display **OK**.

Managing System Partition-Sensitive Subsystems Using syspar_ctrl

Some of the subsystems that run in the SP environment operate within the domain of a system partition, rather than the domain of the system as a whole. These subsystems are said to be **system partition-sensitive**.

To help you manage these subsystems, PSSP includes the Syspar Controller, which operates through the **syspar_ctrl** command. With this command you can add, start, stop, refresh, and delete system partition-sensitive subsystems, as well as control other aspects of their operation, such as tracing.

The **syspar_ctrl** command provides a single interface to the control scripts for the system partition-sensitive subsystems, which perform the real work of configuring and operating the subsystems.

Listing the System Partition-Sensitive Subsystems

To list the system partition-sensitive subsystems that are managed by the Syspar Controller and their control scripts, enter:

```
syspar_ctrl -E
```

In response, the system displays information similar to the following:

```
Syspar Controller managed subsystems and control scripts:
  hats        /usr/lpp/ssp/bin/hatsctrl
  hb          /usr/lpp/ssp/bin/hbctrl
  hags        /usr/lpp/ssp/bin/hagsctrl
  haem        /usr/lpp/ssp/bin/haemctrl
  hr          /usr/lpp/ssp/bin/hrctrl
  pman        /usr/lpp/ssp/bin/pmanctrl
  emon        /usr/lpp/ssp/bin/emonctrl
  sp_configd  /usr/lpp/ssp/bin/sp_configdctrl
  emcond      /usr/lpp/ssp/bin/emconditionctrl
```

The order in which the subsystems are listed (from top to bottom) is the order in which the subsystems are added or started by the **syspar_ctrl** command. The command deletes or stops the subsystems in the reverse order (from bottom to top). Other optionally installed subsystems may be listed. Refer to the documentation for those subsystems for more information.

The system partition-sensitive subsystems provide the following services:

- **hats** – the Topology Services subsystem, which provides information to other PSSP subsystems about the state of the nodes and adapters on the SP system.
- **hb** – the Heartbeat subsystem, which communicates with several PSSP subsystems as part of providing information about the state of the nodes and adapters on the SP system.
- **hags** – the Group Services subsystem, which provides a general purpose facility for coordinating and monitoring changes to the state of a client program that is running on a set of nodes in an SP system.

- **haem** – the Event Management subsystem, which compares information about the state of system resources provided by resource monitors with requests for information by client programs, and creates events that notify the clients when a match occurs
- **hr** – the Host_Responds subsystem, which provides information about the state of the nodes and adapters on the SP system.
- **pman** – the Problem Management subsystem, which allows the installation to be notified when events related to problem management occur
- **emon** – the Emonitor subsystem, which monitors the status of nodes in a switch network, bringing a node back into the switch network when necessary. This subsystem is not started by the **syspar_ctrl** command. It is started using the **Estart -m** command.
- **sp_configd** – the SNMP Support portion of the Problem management subsystem, which creates and sends SNMP traps when selected types of errors are recorded.
- **emcond** – the Event Management Conditions subsystem loads the SDR with pre-defined events or conditions of interest. These conditions are used by the SP Perspectives GUI.

Understanding System Partition-Sensitive Subsystems

Except where noted, the PSSP system partition-sensitive subsystems operate under the control of the System Resource Controller (SRC) and have associated SRC group and subsystem names. Each subsystem has a control script. Except where noted, each subsystem is associated with one or more daemons.

Except where noted, for each system partition, one instance of a subsystem executes on the control workstation and one instance executes on each of the system partition's nodes.

On the control workstation, there are multiple instances of each subsystem, one for each system partition that is defined. Accordingly, the subsystem names on the control workstation have the system partition name appended to them.

On a node, only one instance of a subsystem executes. Hence, the system partition name is not required to differentiate instances of a subsystem.

Several of the subsystems that require configuration have an associated script that configures the daemon before starting it.

Table 10 on page 261 summarizes key information about the system partition-sensitive subsystems. It uses the following notation:

- *syspar_name* is the name of a system partition.
- CWS indicates the control workstation

For details about the operation of each subsystem, see the man page for its control script. For information about the SRC, see *AIX Version 4 General Programming Concepts: Writing and Debugging Programs*.

Table 10. Summary of System Partition-Sensitive Subsystems				
Subsystem	*SRC Group*	*Control Script*	*Daemons*	*Subsystem Names*
Topology Services	**hats**	**hatsctrl**	**hatsd**	CWS: **hats.***syspar_name* Nodes: **hats**
Heartbeat	**hb**	**hbctrl**	**hbd**	CWS: **hb.***syspar_name* Nodes: **hb**
Group Services	**hags**	**hagsctrl**	**hagsd** **hagsglsmd**	CWS: **hags.***syspar_name* Nodes: **hags** CWS: **hagsglsm.***syspar_name* Nodes: **hagsglsm**
Event Management	**haem**	**haemctrl**	**haemd**	CWS: **haem.***syspar_name* Nodes: **haem**
Host_ Responds	**hr**	**hrctrl**	**hrd**	CWS: **hr.***syspar_name* Nodes: None
Problem Management (Event Management)	**pman**	**pmanctrl**	**pmand** **pmanrmd**	CWS: **pman.***syspar_name* Nodes: **pman** CWS: **pmanrm.***syspar_name* Nodes: **pmanrm**
Emonitor	**emon**	**emonctrl**	**Emonitor**	CWS: **Emonitor.***syspar_name* (started only by **Estart -m**) Nodes: None
Problem Management (SNMP Support)	None	**sp_configdctrl**	**sp_configd**	CWS: **sp_configd** (only one instance per system) Nodes: **sp_configd**
Event Management Conditions	None	**emconditionctrl**	None	None

Checking the Status of System Partition-Sensitive Subsystems

Normally, you do not have to do anything to manage the system partition-sensitive subsystems. In general, they are all added and started during installation and use standard mechanisms to restart themselves when they are not explicitly stopped.

To make sure that all of the system partition-sensitive subsystems for a system partition are running, login to the control workstation and use the **lssrc -a** command, grepping for the system partition name. Issue another command to check the **sp_configd** subsystem, which does not have the system partition name appended to it.

For example, if the system partition is named **k22sp1**, enter:

```
lssrc -a | grep k22sp1
lssrc -a | grep sp_configd
```

In response, you should receive output that indicates that all of the system-partition subsystems except Emonitor are active, similar to the following:

```
hats.k22sp1      hats             34126   active
hb.k22sp1        hb               48664   active
hags.k22sp1      hags             34446   active
hagsglsm.k22sp1  hags             34464   active
haem.k22sp1      haem             26088   active
hr.k22sp1        hr               16566   active
pman.k22sp1      pman             32806   active
pmanrm.k22sp1    pman             40496   active
Emonitor.k22sp1  emon                     inoperative
sdr.k22sp1       sdr              33912   active

sp_configd                        39568   active
```

Note that, although its subsystem appears in the output for the system partition, the SDR is not managed by the Syspar Controller.

Also, the Emonitor subsystem is not started by the **syspar_ctrl** command. Therefore, its status is inoperative until it is started by the **Estart -m** command.

Chapter 18. Managing Node Groups

Node grouping is a persistently saved naming facility for sets of nodes: nodes can be grouped together so that they can be acted on as an individual entity. Many system management commands can accept a node group as an argument, causing the action to be performed on each node in the node group. Node groups can be created and managed through a command line interface, SMIT panels, and the SP Perspectives GUI.

This chapter provides information on:

- Understanding node groups
- Working with node groups
- Using node groups with commands

Understanding Node Groups

A node group is a set of nodes whose members may be individual nodes and other node groups. Individual nodes that belong to a node group are *member nodes*; node groups that belong to a node group are *member node groups*. Node groups containing only nodes are *flat node groups*; node groups containing node groups are *hierarchical node groups*. Any node group can be resolved into a flat version that contains only valid member nodes.

Dependent nodes cannot be part of a node group.

Commands, except those that directly manage the node groups, issued against a node group have the same effect as issuing the commands against the set of nodes that represent the flat version of the node group.

Node Group Information in the SDR

Node groups are stored persistently in the SDR. A node group in the SDR is made up of two components: a name and a definition list. A node group name must start with an alphabetic character (A-Z, a-z), but can then contain a sequence of alphanumeric characters (A-Z, a-z, 0-9), a period (.) or an underbar (_). (Thus named, node groups can be easily differentiated from node objects, whose names are a string of numeric characters representing the node number.) A node group name may not contain spaces. The definition list of a node group is made up of a list of nodes and node groups that belong to the node group directly, (not through member node groups).

Node Group Modes

Some node groups exist within a system partition (*partition-bound* node groups or *local* node groups) and some node groups ignore system partition boundaries (*global* node groups). Partition-bound or local node groups can be accessed only within the context of the system partition they belong to. Global node groups are not bound to any system partition, but can be acted on only by commands that accept the **-G** flag or when using the SP Perspectives GUI with a Global view of the SP system. Partition-bound and global refer to the *mode* of the node group.

© Copyright IBM Corp. 1995, 1997

Partition-bound node groups (local node groups) are stored in the partitioned SDR class **NodeGroup**. Global node groups are stored in the system SDR class **SysNodeGroup**. Valid member nodes of a partition-bound node group must be node numbers that are defined in that particular system partition. Valid member nodes of a global node group are node numbers that are defined in that particular SP system (the union of all nodes in each system partition). Any node number or node group can be added to the definition list of a node group, but only valid member nodes or valid member node groups will be used when resolving (flattening) the node group.

Note: The global name space and the partition name spaces are mutually exclusive for node groups.

Most SP system end users work within system partitions and, therefore, will use partition-bound node groups.

System Administrators, however, will often find it necessary, and advantageous, to work with global node groups. The use of global node groups allows a System Administrator to perform certain tasks, such as installing a daemon on specific nodes or powering off and reconfiguring all nodes on which a given application is running, without regard to system partitions.

Partition-bound node groups and global node groups are different, but they both function to provide a naming facility for sets of nodes.

Considerations for Creating and Resolving Node Groups

When creating a node group or adding to an existing node group's definition list, you can use the command line interface or SMIT panels to add any node number or node group. The member node group does not need to have been defined at the time it is added to another node group's definition list. The member node does not have to be a valid node number for a system partition or even the SP system at the time it is added to another node group's definition list. In this way, you can create node groups to accommodate future growth or changes, such as adding nodes or repartitioning your SP system.

Only users that have the appropriate authority to modify objects in the SDR can create, modify or delete node groups. Any user that has the appropriate authority to read objects from the SDR can reference existing node groups. Any user can modify or reference node groups created by other users, provided they have the appropriate SDR authority. For more information on the authority needed to access the SDR, refer to “Authorization” on page 577.

When a node group is resolved into its flat version, only valid member nodes will appear in the flat version and only valid member node groups will be resolved recursively to produce the flat version. If a node group is a partition-bound node group, when it is flattened all of its member nodes and member node groups must be defined in the same system partition; if they are not, they will not be used to produce the flat version of this node group. If the node group is a global node group, when it is flattened all of its member nodes must be node numbers for the system and member node groups must also be defined as global node groups. Member nodes or member node groups that do not correspond to entities in persistent storage will be ignored when producing a flat version of the node group. A node group's flat version can be empty.

The following is a simple example of how a partition-bound node group will be resolved into a flat list of node numbers.

```
system partition P1 contains nodes 1, 2, 5, 6, 9, 10, 13, 14
system partition P1 contains node groups A, B, C
Node Group A name = A; definition list = 1, 2, B, C, D
Node Group B name = B; definition list = 1, 2, 3, 4, 5, 6
Node Group C name = C; definition list = 13, 14
flat(A) = nodes 1, 2, 5, 6, 13, 14
flat(B) = nodes 1, 2, 5, 6
flat(C) = nodes 13, 14
```

Note: Nodes 3,4 are not in flat(A) or flat(B) because they are not part of partition P1.

Note: Node group D has no affect on the flat(A) because node group D is not defined in partition P1.

Working with Node Groups

A command line interface, SMIT panels, and the SP Perspectives GUI are provided for working with node groups. For a complete description of commands, refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference*. For information on using the SP Perspectives GUI to create and manage node groups, refer to the online help.

The following commands are provided for creating, modifying, and deleting node groups:

ngcreate Creates and optionally populates a named node group.

ngnew Creates new node groups in persistent storage.

ngaddto Adds specified nodes and node groups to named node group.

ngdelfrom Removes specified nodes and node groups from named node group.

ngdelete Removes node groups from persistent storage.

ngclean Removes nodes and node groups that do not currently exist from specified node groups or every node group in the current system partition.

The following commands are provided to display node group information to standard output:

ngresolve
Displays a list of nodes in the node group.

nglist Writes a list of all persistent node groups.

ngfind Writes a list of all node groups containing the supplied node or node group.

To access SMIT panels (using the fastpath invocation) for creating, modifying, and deleting node groups:

TYPE **smit nodegroups**

▲ The Node Group Informaton menu appears.

SELECT Partition Node Groups OR System Node Groups

▲ The Partition Node Groups or System Node Groups panel appears.

To access SMIT panels (using the fastpath invocation) for examining existing node groups:

TYPE **smit list_nodegroup**

▲ The List Node Group Database Information menu appears.

SELECT Examine Partition Node Groups OR Examine System Node Groups

▲ The Examine Partition Node Groups or Examine System Node Groups panel appears.

To access SMIT panels (using the fastpath invocation) for deleting node groups:

TYPE **smit delete_nodegroup**

▲ The Delete Node Group Information menu appears.

SELECT Delete Partition Node Groups OR Delete System Node Groups

▲ The Delete Partition Node Groups or Delete System Node Groups panel appears.

Using Node Groups with Commands

The following commands will accept a node group as an argument. The node group may be supplied as an operand, if this is what is most consistent with the existing syntax. Some commands use the **-N** flag to indicate a node group. Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for actual syntax for each command.

- **cshutdown**
- **cstartup**
- **dsh** (can be used to run a command on every node in a flattened node group)
- **Efence**
- **Eunfence**
- **hostlist**
- **pmandef**
- p-commands

  The following commands all pass arguments to **hostlist**. **hostlist** accepts a node group as an argument, so these commands get node group support through **hostlist**:
  - **p_cat**
  - **pcp**
  - **pdf**
  - **pexec**
  - **pexscr**
  - **pfck**
  - **pfind**
  - **pfps**
  - **pls**
  - **pmv**
  - **ppred**
  - **pps**

  - **prm**
- **spacctnd**
- **spadaptrs**
- **spbootins**
- **spdeladap**
- **spdelnode**
- **spethernt**
- **sphostnam**
- **sphrdwrad**
- **splst_versions**
- **splstdata**

Chapter 19. Managing Extension Nodes

This chapter provides the following information:

- Understanding extension nodes and extension node adapters
- Specific implementations of extension nodes and extension node adapters
- Transferring configuration information for extension nodes adapters
- Adding extension nodes and extension node adapters
- Changing extension nodes and extension node adapters
- Removing extension nodes and extension node adapters
- Reconfiguring extension nodes and extension node adapters
- Resetting extension nodes and extension node adapters
- Listing extension node and extension node adapter information

For switch-related information on extension nodes, refer to "Extension Nodes and the SP Switch" on page 224.

Understanding Extension Nodes

Extension nodes are non-standard nodes that extend the SP system's capabilities or scope, but can't be used in all the same ways as standard SP nodes. Extension nodes are attached to the SP Switch via a special **extension node adapter**.

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

SMIT panels and commands are provided for adding configuration information about extension nodes to the SDR. Once this information has been entered into the SDR, a Simple Network Management Protocol (SNMP) Manager daemon that runs on the control workstation (the SP SNMP Manager) allows the transfer of extension node configuration information to the SNMP Agent running on the extension node.

Specific Implementations of Extension Nodes and Extension Node Adapters

An Ascend GRF switched IP router (GRF) can be connected to the SP Switch via the SP Switch Router Adapter. The SP Switch Router Adapter provides a high performance, 100 MB/s, full duplex interface between the SP Switch and the Ascend switched IP router.

When the SP Switch Router Adapter is installed in a GRF, it allows the switched IP router to be used as a networking gateway for the SP. The GRF may be populated with additional adapters for standard network interfaces, including Ethernet 10/100BaseT, FDDI, ATM OC3c, SONET OC3c, ATM OC12c, HIPPI and HSSI. More than one SP Switch Router Adapter may be installed and connected to the same SP system or system partition or to other SP systems. When multiple SP

© Copyright IBM Corp. 1995, 1997

Switch Router Adapters are installed and connected to more than one SP system or system partition, they may 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 *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment*. A 10 meter SP Switch cable and a 10 meter ground strap are provided for connecting the SP Switch Router Adapter, located in the GRF chassis, to the SP Switch.

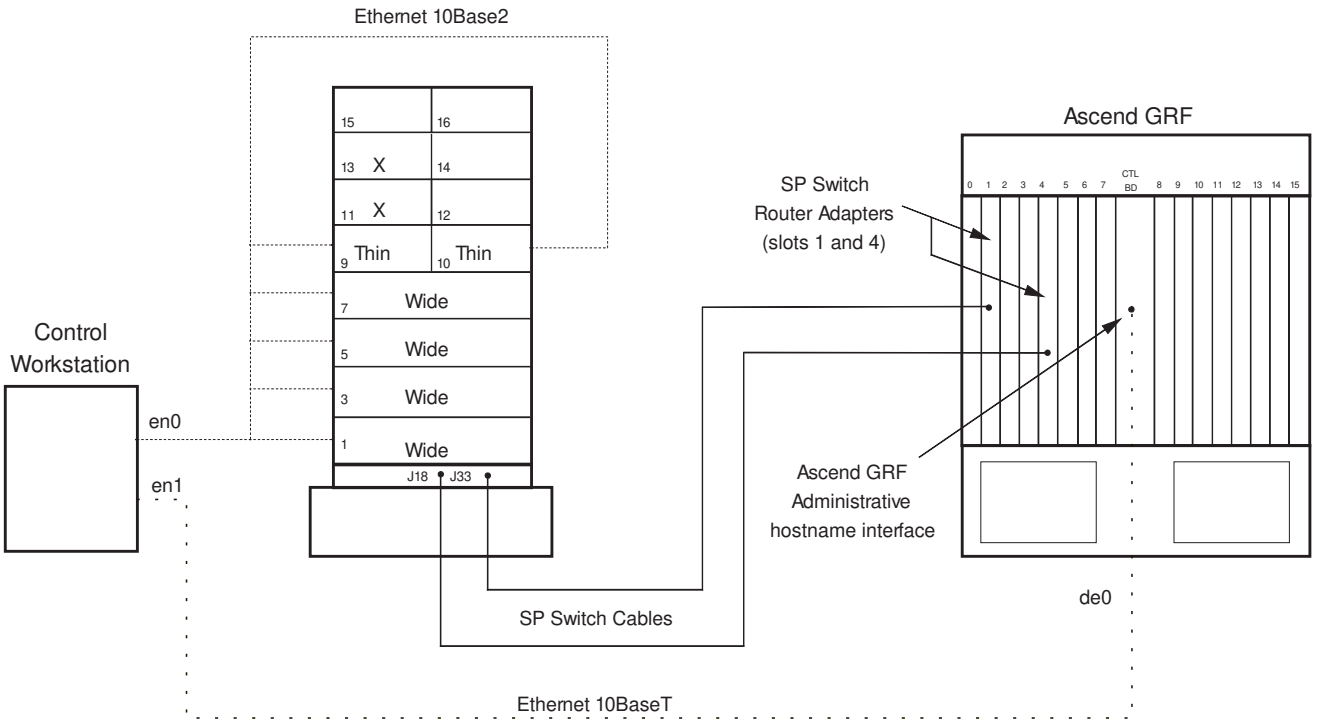


Figure 36. The Ascend GRF and SP Switch Router Adapter Extension Node Implementation

Extension Node Attribute Values

When you add information to the SDR for an Ascend GRF and the SP Switch Router Adapter (see "Adding Extension Nodes and Extension Node Adapters" on page 271), specify the following extension node attributes as described:

Administrative Hostname The fully-qualified host name of the administrative Ethernet network on the Ascend GRF.

Extension Node Identifier The slot number on the GRF where the SP Switch Router Adapter is located.

SNMP Agent Hostname The fully-qualified host name of the administrative Ethernet network on the Ascend GRF.

For more information on configuring the GRF and SP Switch Router Adapter as extension nodes for the SP system, see:

- *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment*

- *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*
- *SP Switch Router Adapter Guide*

Transferring Extension Node Configuration Information

Once you have used SMIT panels or commands to enter extension node information into the SDR, the SP SNMP Manager running on the control workstation allows the transfer of extension node configuration information to the SNMP Agent running on the extension node.

Refer to “Understanding Network Management” on page 400 for a high-level overview of Simple Network Management Protocol (SNMP).

The SP SNMP Manager daemon, **spmgrd**, runs on the control workstation and treats each extension node as a managed node, with a corresponding SNMP Agent. The SP SNMP Manager waits for trap messages from extension node SNMP Agents reporting state changes or requesting configuration information and sends configuration information to the SNMP Agents when requested. For more information on **spmgrd**, refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

An SNMP Agent may be responsible for one or several extension nodes attached to the SP Switch. The SNMP Agent maintains configuration information for each of its extension nodes in the SNMP Management Information Base (MIB). When an SNMP Agent needs configuration information for one of its extension nodes, it sends a trap message to the SP SNMP Manager responsible for that extension node. The SP SNMP Manager extracts the requested information from the SDR and sends it to the SNMP Agent.

See “ibmSPDepNode.my” on page 509 for a copy of the MIB for dependent nodes.

Adding Extension Nodes and Extension Node Adapters

You can add information for extension nodes and extension node adapters into the SDR by using either the SMIT or command line interface.

From SMIT

1. Access the Extension Node Database Information menu:

   ENTER **smit enter_extdata**

   ▲ The Extension Node Database Information menu appears.

2. Add information to the SDR for extension nodes:

   SELECT **Enter Extension Node Information**

   ▲ The Enter Extension Node Information menu appears.

   a. Enter values for the following extension node attributes:

   Administrative Hostname The fully qualified host name associated with the extension node's network interface on its system's administrative network. This name must resolve to an IP address. *Required*.

SNMP Community Name The Simple Network Management Protocol (SNMP) community name for the authentication field in SNMP messages exchanged by the extension node's SNMP Agent and the SP SNMP Manager. This value can be 1 to 255 ASCII characters.

Extension Node Identifier The string assigned to the extension node in its system's administrative environment. This string uniquely identifies the extension node to its system and can be 1 to 255 ASCII characters. *Required*. (See "Specific Implementations of Extension Nodes and Extension Node Adapters" on page 269 for more information.)

SNMP Agent Hostname Host name that resolves to an IP address that is the source or destination of SNMP packets exchanged with the control workstation via the SP administrative network. *Required*.

Reconfigure the extension node? yes or **no**. **yes** notifies the SNMP Agent on the extension node that extension node information has been added or changed. This notification causes the SNMP Agent on the extension node to request configuration information from the SP SNMP Manager on the control workstation.

Node Number The node number that corresponds to the valid unused switch port assigned to the extension node. *Required*. See *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment* for information on switch node numbering.

b. Select **OK**.

3. Return to the Extension Node Database Information menu.
4. Add information to the SDR for extension node adapters:

   SELECT **Enter Extension Node Adapter Information**

   ▲ The Enter Extension Node Adapter Information menu appears.

   a. Enter values for the following extension node adapter attributes:

   Network Address The IP network address for the extension node adapter.

   Network Netmask The network mask for IP addresses on the attached network on which the extension node adapter resides.

Reconfigure the extension node? yes or **no**. **yes** notifies the SNMP Agent on the extension node that extension node information has been added or changed. This notification causes the SNMP Agent on the extension node to request configuration information from the SP SNMP Manager on the control workstation.

Node Number The node number that corresponds to the valid unused switch port assigned to the extension node. *Required*. See *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment* for information on switch node numbering.

b. Select **OK**.

From the Command Line:

1. Use the **endefnode** command to add extension node information to the SDR.
2. Use the **endefadapter** command to add extension node adapter information to the SDR.

Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for information on these commands.

If you are adding an extension node and extension node adapter after the SP Switch has already been started, you must run **Estart**.

Changing Extension Nodes and Extension Node Adapters

You can change information for extension nodes and extension node adapters that you've already defined in the SDR by using either the SMIT or command line interface.

From SMIT:

1. Access the Extension Node Database Information menu:

   ENTER **smit enter_extdata**

   ▲ The Extension Node Database Information menu appears.

2. Change information in the SDR for extension nodes:

   SELECT **Enter Extension Node Information**

   ▲ The Enter Extension Node Information menu appears.

   a. Enter new values for the extension node attributes you wish to change. (Refer to “Adding Extension Nodes and Extension Node Adapters” on page 271 for a description of these values.)

   Note: You cannot change the node number of the extension node. You must remove the extension node number from the SDR (by using the SMIT menus described in “Removing Extension Nodes and Extension Node Adapters” on page 274 or the **enrmnode** command) and add it back with a new node number (by using the SMIT menus described in “Adding

Extension Nodes and Extension Node Adapters" on page 271 or the **endefnode** command).

b. Select **OK**.

3. Return to the Extension Node Database Information menu.
4. Change information in the SDR for extension node adapters:

   SELECT **Enter Extension Node Adapter Information**

   ▲ The Enter Extension Node Adapter Information menu appears.

   a. Enter new values for the extension node adapter attributes you wish to change. (Refer to "Adding Extension Nodes and Extension Node Adapters" on page 271 for a description of these values.)

      Note: You cannot change the node number of the extension node. You must remove the extension node number from the SDR (by using the SMIT menus described in "Removing Extension Nodes and Extension Node Adapters" or the **enrmnode** command) and add it back with a new node number (by using the SMIT menus described in "Adding Extension Nodes and Extension Node Adapters" on page 271 or the **endefnode** command).

   b. Select **OK**

From the Command Line:

1. Use the **endefnode** command with the appropriate flags and operand to change the definition for an extension node in the SDR.
2. Use the **endefadapter** command with the appropriate flags and operand to change the definition for an extension node adapter in the SDR.

Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for details on these commands.

You can use the **splstnodes** and **splstadapters** commands to verify that the changes you've specified have been made.

Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for details on these commands. Also, see "Listing Extension Node and Extension Node Adapter Information" on page 277.

Removing Extension Nodes and Extension Node Adapters

You can remove information for extension nodes and extension node adapters from the SDR by using either the SMIT or command line interface.

Note: You must **reset** the extension node (using either **enadmin -a reset** or **enrmnode -r**) *before* removing the extension node adapter or extension node.

From SMIT

1. Access the Delete Database Information menu:

   ENTER **smit delete_data**

   ▲ The Delete Database Information menu appears.

2. Delete extension node information:

SELECT **Delete Extension Node Information**

▲ The Delete Extension Node Information menu appears.

a. Enter values to delete extension node information:

Reset the Extension Node?	**yes** or **no**. **yes** notifies the SNMP Agent on the extension node to remove the node from the switch network.
Node Number	The number ID assigned to the extension node.

b. Select **OK**.

3. Return to the Delete Database Information menu.
4. Delete extension node adapter information:

SELECT **Delete Extension Node Adapter Information**

▲ The Delete Extension Node Adapter Information menu appears.

a. Enter value to delete extension node adapter information:

Node Number	The number ID assigned to the extension node.

b. Select **OK**.

From the Command Line:

1. Use the **enrmnode** command to remove extension node information from the SDR.
2. Use the **enrmadapter** command to remove extension node adapter information from the SDR.

Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for information on these commands.

Reconfiguring Extension Nodes and Extension Node Adapters

Once extension node and extension node adapter information has been changed in the SDR, the extension node SNMP Agent must be notified.

From SMIT: Use the SMIT panels for adding or changing extension node information as described in "Adding Extension Nodes and Extension Node Adapters" on page 271 and "Changing Extension Nodes and Extension Node Adapters" on page 273. Specify **yes** for **Reconfigure the extension node?**.

SMIT panels are also provided that invoke the **enadmin** command.

1. Access the RS/6000 Cluster Management menu:

ENTER **smit cluster_mgmt**

▲ The RS/6000 Cluster Management menu appears.

2. Run the **enadmin** command:

SELECT **Run enadmin Command**

▲ The Extension Node Management menu appears.

a. Enter values to manage the extension node:

Action to be performed on the extension node — **reconfigure** or **reset**.

Node Number — The number ID assigned to the extension node.

b. Select **OK**.

From the Command Line:

1. Use the **endefnode -r** command to reconfigure extension nodes.
2. Use the **endefadapter -r** command to reconfigure extension node adapters.
3. As an alternative to using **endefnode -r** and **endefadapter -r**, you can use the **enadmin** command to notify the SNMP Agent managing the extension node to request updated configuration information from the SP SNMP Manager (the **spmgrd** daemon) running on the control workstation.

Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for details on these commands.

For more general information on reconfiguring extension nodes and extension node adapters, see "Transferring Extension Node Configuration Information" on page 271.

Resetting Extension Nodes

Resetting an extension node notifies the SNMP Agent managing the extension node to remove the extension node from the switch network.

Note: You must **reset** the extension node (using either **enadmin -a reset** or **enrmnode -r**) *before* removing the extension node adapter or extension node.

From SMIT: Use the SMIT panels for removing extension node information as described in "Removing Extension Nodes and Extension Node Adapters" on page 274. Specify **yes** for **Reset the extension node?**.

You can also use the SMIT panels for running the **enadmin** command as described in "Reconfiguring Extension Nodes and Extension Node Adapters" on page 275. Specify **reset** for **Action to be performed on the extension node**.

From the Command Line: Use the **enrmnode -r** or **enadmin -a reset** commands to notify the SNMP Agent managing the extension node to remove the node from the switch network.

Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for details on these commands.

Listing Extension Node and Extension Node Adapter Information

You can display information for extension nodes and extension node adapters from the SDR by using either the SMIT or command line interface.

From SMIT

1. Access the List Database Information menu:

   ENTER **smit list_data**

   ▲ The List Database Information menu appears.

2. List extension node database information:

   SELECT **List Extension Node Database Information**

   ▲ The List Extension Node Database Information menu appears.

   a. Enter values to list extension node database information:

   Allow Nodes Outside Current System Partition? yes or **no**.

Node Type	**all** specifies the query for both standard and dependent nodes. **standard** or **dependent** restricts the query to those types of nodes.
Node Group	The name of the node group to which the query is restricted. Only the nodes in the named node group will be queried. *Not supported for extension nodes.*
Sort Attribute	The name of a valid attribute of one of the node classes in the SDR. (See "System Data Repository Classes and Attributes" on page 585 for more information.) Listing output will be sorted by this attribute.
Query Attribute	The name and value (specified as *attribute*==*value*) of a valid attribute of one of the node classes in the SDR. (See "System Data Repository Classes and Attributes" on page 585 for more information.) Node information will be listed by the specified attribute. (The default is **node_number**, in which case information is listed by node number.)
Output Attribute	The attributes you want displayed.

   b. Select **OK**.

3. Return to the List Database Information menu.

4. List extension node adapter database information:

   SELECT **List Extension Node Adapters Database Information**

   ▲ The List Extension Node Adapters Database Information menu appears.

a. Enter values to list extension node adapter information:

   Allow Adapters Outside Current System Partition? yes or **no**.

Node Type	**all** specifies the query for both standard and dependent nodes. **standard** or **dependent** restricts the query to those types of nodes.
Query Attribute	The name and value (specified as *attribute*==*value*) of a valid attribute of one of the node classes in the SDR. (See "System Data Repository Classes and Attributes" on page 585 for more information.) Node information will be listed by the specified attribute. (The default is **node_number**, in which case information is listed by node number.)
Output Attribute	The attributes you want displayed.

b. Select **OK**.

From the Command Line:

1. Use the **splstnodes** command to list information for standard and extension nodes.
2. Use the **splstadapters** command to list information for standard node adapters and extension node adapters.

Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for information on these commands.

System Monitoring

© Copyright IBM Corp. 1995, 1997

Chapter 20. Using SP Perspectives

Scalable POWERparallel Perspectives for AIX (SP Perspectives) is a set of applications, each of which has a graphical user interface (GUI), that enables you to perform system management tasks for your SP system through direct manipulation of system objects represented by icons.

You simply select an SP system object (a managed SP system resource such as a frame, node, or switch) by clicking on it with a mouse and then select an action to perform on that object from the menu or tool bar. Use this pattern of selecting an object and then selecting an action to perform on that object to accomplish numerous system management tasks with SP Perspectives.

You can use SP Perspectives to perform the following system management tasks:

- Monitor and control hardware
- Create and monitor system events
- Define and manage IBM Virtual Shared Disks
- Generate and save system partition configurations
- Set up performance monitoring hierarchies and archiving

SP Perspectives also provides access to system management tools such as SMIT panels and AIX 4.1 Visual System Management (VSM) windows. In addition, you can create your own applications and make them accessible from the SP Perspectives GUI Launch Pad.

Because SP Perspectives provides access to a wider range of system management tasks, and because SP Perspectives will eventually replace the SP System Monitor, we recommend its use over the SP System Monitor.

This chapter presents overview information for SP Perspectives; it covers the following topics:

- Getting started
- Understanding the SP Perspectives interface
- Accessing help for SP Perspectives

Information for the tasks you can perform using the various SP Perspectives applications is available through an extensive online help system (see "Accessing Help" on page 293) and is not repeated in this chapter.

Getting Started

This section provides information on authorizing users for SP Perspectives and starting SP Perspectives.

© Copyright IBM Corp. 1995, 1997

Authorizing Users for SP Perspectives

No special authorization is required to use the application GUIs available from SP Perspectives. However, the applications themselves may require special authorization.

In order to be able to control the SP system hardware or open a serial link from the Hardware perspective, an SP Perspectives client needs the appropriate **hardmon** ACLs (access control lists) set up on the control workstation in **/spdata/sys1/spmon/hmacls** (refer to the step on initializing authentication services in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*). Then the client must have a Kerberos token that has the desired **hardmon** authority in order to control the SP system hardware.

There are five commands that root can issue on the control workstation that a non-root user can be given access to:

- **cshutdown**
- **cstartup**
- **Efence**
- **Eunfence**
- **Estart**

These commands must be run by root on the control workstation unless **sysctl** acls have been set up on the control workstation for the principal that the SP Perspectives client is using. Also, in order to be able to issue these commands through **sysctl**, the principal must have authority for **hardmon** for the frames that the command is issued for. To set up the **sysctl** acl for these commands, edit the file **/etc/sysctl.rootcmds.acl** and add the principal that you want to be authorized to issue these commands.

To create, modify, or delete node groups from the Hardware perspective, you must be root on the control workstation.

No special authorization is required to access the Event perspective. However, to use the Problem Management subsystem from SP Perspectives (available under the Event perspective), your Kerberos principal must be added to the **/etc/sysctl.pman.acl** file on the control workstation and nodes. Add the principal to the file on the control workstation and copy it to the nodes by issuing:

```
pcp -a /etc/sysctl.pman.acl /etc
```

If you wish, you can use another way to distribute the ACL file, such as file collections. You may also wish to be more selective about what nodes will be authorized for Problem Management. For more information on security in Problem Management, refer to “Security Considerations” on page 390.

To work with the IBM Virtual Shared Disk and System Performance Tool perspectives, the same authorization as is described for node groups is required.

Starting SP Perspectives

To start SP Perspectives running as a background process:

1. Enter **perspectives &** (presuming that you have placed **/usr/lpp/ssp/bin** in your PATH) at the command line to bring up the Launch Pad.
2. Double-click on the icon for the application you want to run. Initially an hourglass icon appears, followed by the SP Perspectives splash screen, to let

you know processing is occurring. Then, the primary window (also known as the “perspective”) for the application will open.

Alternatively:

1. Enter **perspectives &** (presuming that you have placed **/usr/lpp/ssp/bin** in your PATH) at the command line to bring up the Launch Pad.
2. Click on the icon for the application you want to run.
3. Click on the **Actions** menu bar item.
4. Click on **Launch**. Initially an hourglass icon appears, followed by the SP Perspectives splash screen, to let you know processing is occurring. Then, the primary window (or “perspective”) for the application will open.

To start SP Perspectives applications (running as background processes) without using the Launch Pad, enter the following commands at the command line:

sphardware & Manage/Control Hardware

spevent & Manage Events

spvsd & Manage IBM Virtual Shared Disks and Data Striping Devices

spsyspar & Define System Partitions

spperfmon & Setup/Configure Performance Monitor

You can also start SP Perspectives applications from the Common Desktop Environment (CDE). (Double-click on the SP Perspectives icon in the CDE Application Manager window.)

For a general overview of how to perform common SP system administration tasks with SP Perspectives, refer to “Performing Common Administration Tasks with SP Perspectives” on page 294.

Understanding the SP Perspectives Interface

SP Perspectives application GUIs consist of the following graphical elements:

- Launch Pad
- Primary windows (also known as “perspectives”)
- Dialog boxes
- Notebooks
- Message boxes

The Launch Pad

The Launch Pad starts applications associated with managing an SP system. It also provides options for customizing the appearance of the Launch Pad.

System management applications provided by SP Perspectives are represented as icons on the Launch Pad. Other system management tools, such as SMIT panels and VSM windows, are also accessible from the Launch Pad. In addition, you can add your own applications to the Launch Pad.

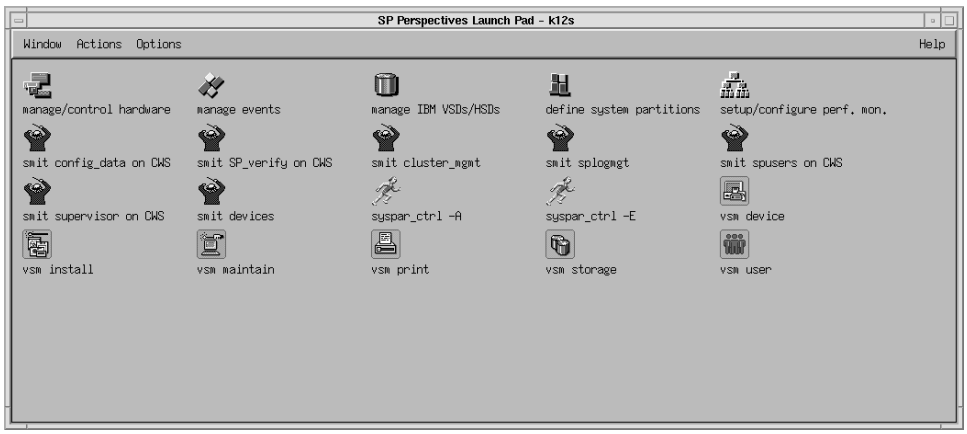


Figure 37. The SP Perspectives GUI Launch Pad

The Launch Pad window contains the following elements:

- The title bar presents the name of the window and an identifier for the control workstation: **SP Perspectives Launch Pad** - *controlworkstation_name*
- The menu bar contains menu items that provide for various actions:
  - **Window**

    Close provides for closing the Launch Pad window. Closing the Launch Pad window does not affect application primary windows that are open.
  - **Actions**
    - **Launch** starts the selected application (represented by an icon)
  - **Options** (actions that operate on the presentation or appearance of the Launch Pad only)
    - **Customize Applications** extends the Launch Pad window to include a panel that provides for adding, deleting, or modifying applications on the Launch Pad. You specify a unique application name, a one-line application description, executable file name, and icon file name (a pixmap in xmp format) and select whether you want to add, delete, or modify the icon representing the application. The Launch Pad is updated immediately.
    - **Details View** presents the launch area in a way that displays icons representing applications, along with the application name, application description, and executable path.
    - **Load Preferences...** presents a dialog box that provides for loading files of preferences. You specify the system or user file name from which to load the preferences and select aspects of preferences to load. Aspects of preferences include foreground and background colors, type font, window position and size, and applications.
    - **Save Preferences...** presents a dialog box that provides for saving files of preferences. You specify the system or user file name to save the preferences to, and select aspects of preferences to save. Aspects of preferences include foreground and background colors, type font, window position and size, and applications.
    - **Set Colors...** presents a dialog box that provides for setting colors. You choose to set the colors for the background and foreground and

select the colors from sliders, a palette, or a list. A sample of the colors you select is dynamically displayed in the dialog box.

- **Set Font...** presents a dialog box that provides for specifying typeface style and size. You select a typeface family and point size and specify bold or italic. A sample of fonts you select is dynamically displayed in the dialog box.

– **Help**

The **Help** button provides options for information about windows, actions (tasks), and objects. (See "The Menu Bar Help Button" on page 293 for more information.)

Primary Windows (Perspectives)

The primary window for each SP Perspectives application is called a "perspective" because it provides a unique view of your SP system. The Hardware perspective, for example, allows you to view the system as a set of hardware objects (for example, nodes, frames, switches). The Event perspective allows you to view the system in terms of conditions of interest related to system resources.

Primary windows contain work areas known as "panes". Panes contain icons representing system objects (a managed SP system resource such as a node, frame, or switch).

- To perform actions on the *display* of objects in panes:
  1. Click on the pane to give it focus.
  2. Select **View** from the menu bar.
  3. Select the desired action from the pulldown menu. (Alternatively, some of the **View** options are available from the tool bar.)
- To perform actions on the *objects themselves*:
  1. Select one or more objects.
     – To select multiple objects that are next to each other in a pane, click on the icon of the first desired object, hold the left mouse button down, and move the cursor over the icons of other desired objects in the pane.
     – To select multiple objects that are not next to each other in a pane, hold the **Ctrl** key on your keyboard down while you click on each desired object in the pane.
  2. Select **Actions** from the menu bar.
  3. Select the name of the pane from the pulldown menu.
  4. Select the desired action from the slide-right menu. (Alternatively, some of the **Actions** options are available from the tool bar.)

Each perspective consists of the same basic format as illustrated in Figure 38 on page 286.

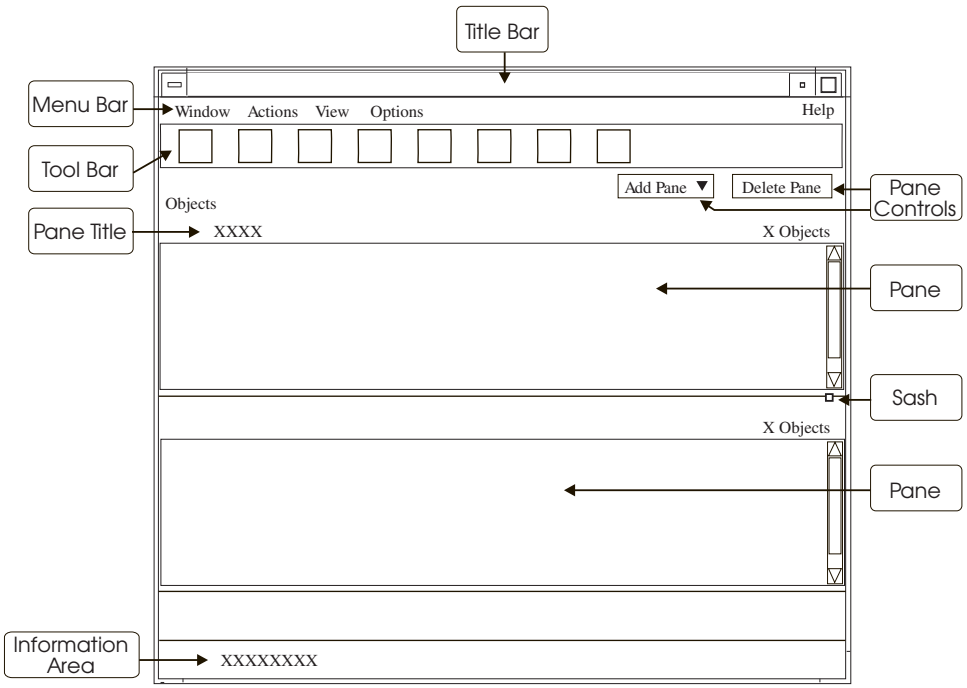


Figure 38. Elements of an SP Perspectives Primary Window

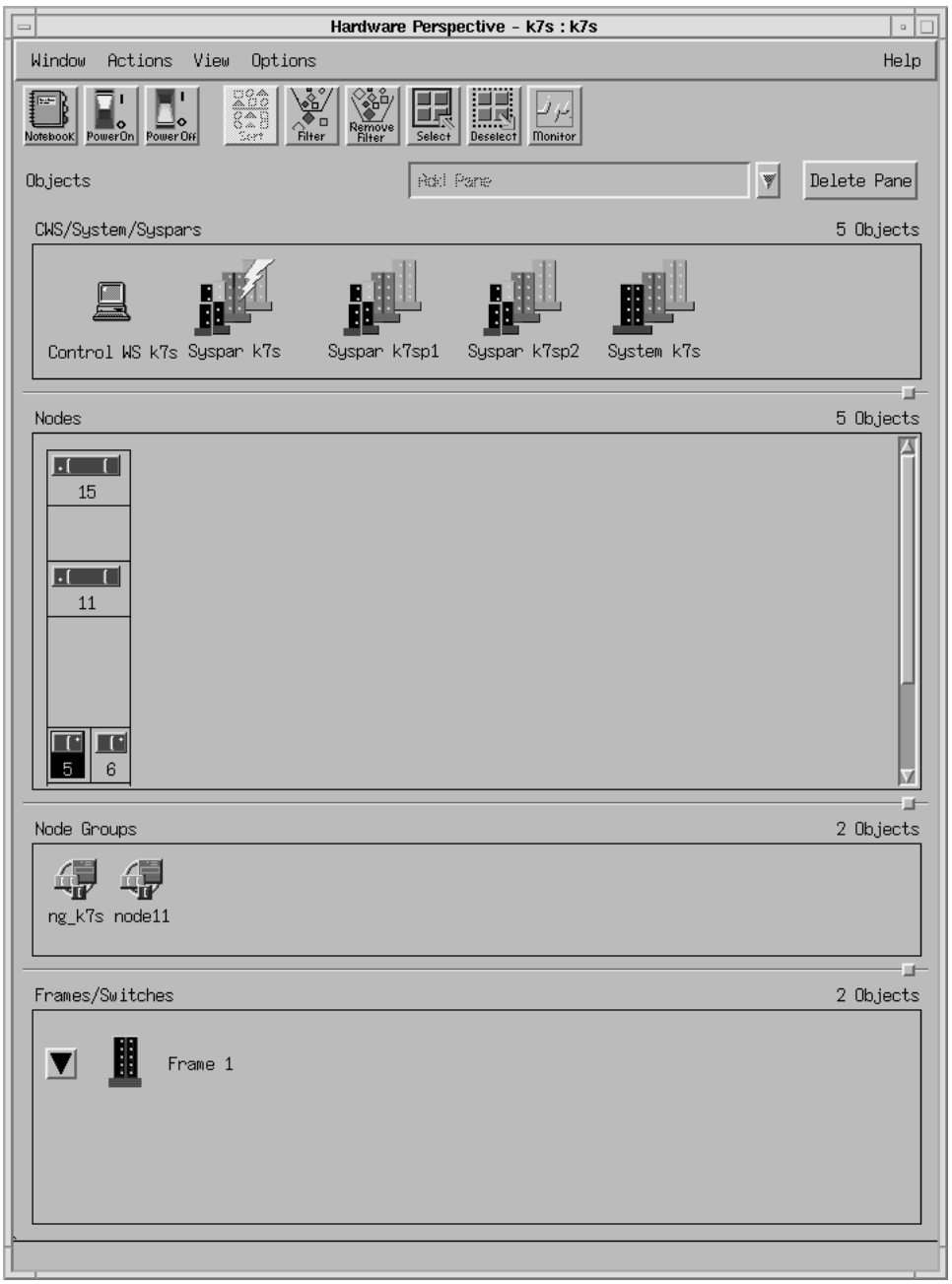


Figure 39. An SP Perspectives GUI Primary Window

The elements of this format are:

- The title bar displays the name of the application, the name of the control workstation running SP Perspectives, and an identifier for the system. If the selected object is the **System** icon, the system is identified as “Global”. If the selected object is a **Syspar** icon, the system is identified by the system partition name. The format for title bar information is:

  *application_name - controlworkstation_name:***Global** or *partition_name*

- The menu bar contains menu items that provide for various actions:
  - **Window** (actions that operate on a window as a whole)

Close provides for closing the window. Closing a primary window also closes any subordinate windows that have been opened for that perspective.

– **Actions** (actions that operate on SP system objects)

  Actions provides a list of panes against which actions are performed. For each pane in the list there is a slide-right menu that lists the actions available for objects within the pane. Actions that are unavailable are grayed out. The application determines what **Action** options are available.

– **View** (actions that operate on the display of SP system objects in a selected pane)

  The following **View** actions are common to most of the SP Perspectives applications. The application determines what **View** options are available.

  - **Vertical/Horizontal Panes** provides for displaying panes vertically or horizontally.
  - **Hide/Display Info Area** provides for making the Information Area viewable or not viewable.
  - **Sort...** presents a dialog box that provides for sorting objects by name in either ascending or descending order.
  - **Filter...** presents a dialog box that provides for specifying which objects are displayed in a pane. You can include or exclude objects by object name or by selecting objects.
  - **Remove Filter** undoes a **Filter** action and redisplays all the objects in a pane.
  - **Select All** selects all objects in the selected pane.
  - **Deselect All** deselects all objects in the selected pane.
  - **Monitor...** provides a dynamic display of selected object status.
  - **Object Label...** sets the text label for objects in the currently selected pane.
  - **Enable/Disable Frame Display** provides a view of node objects as nodes in frames or as individual node icons.
  - **Large/Small Icon Display** provides for displaying nodes as large or small icons. This option is available for only the node pane. The small icon display is not preserved when the nodes pane is deleted and added back.
  - **Enable/Disable Tree View** displays objects in the selected pane in a 'tree' structure or free form.
  - **Expand Tree** extends the 'tree' structure to display 'children' of all the objects in the currently selected pane.
  - **Collapse Tree** collapses the expanded 'tree' structure of all the objects in the currently selected pane.

– **Options** (actions that operate on the presentation or appearance of individual primary windows)

  The **Option** actions are available for all applications.

- **Load Preferences...** presents a dialog box that provides for loading files of preferences. You specify the system or user file name from which to load the preferences and select aspects of preferences to load. Aspects of preferences include foreground and background colors, type font, and window position and size.
- **Save Preferences...** presents a dialog box that provides for saving files of preferences. You specify the system or user file name to save the preferences to, and select aspects of preferences to save. Aspects of preferences include foreground and background colors, type font, and window position and size.
- **Set Colors...** presents a dialog box that provides for setting colors. You choose to set the colors for the background and foreground and select the colors from sliders, a palette, or a list. A sample of the colors you select is dynamically displayed in the dialog box.
- **Set Font...** presents a dialog box that provides for specifying typeface style and size. You select a typeface family and point size and specify bold or italic. A sample of fonts you select is dynamically displayed in the dialog box.

– **Help**

The **Help** button provides options for information about windows, areas, actions (tasks), and objects.

- The tool bar displays icons for often-used actions against selected objects and against the display of objects. These actions are a subset of those available from the menu bar.
- Two pane controls enable you to add and delete panes.
- Panes are sections of the window work area that contain icons that represent system objects. System objects in a pane can be manipulated using actions available in the menu bar or tool bar.
- A sash (a small square and a separator line) provides for controlling the relative size of the panes above and below it.
- The Information Area, the bottom line of the window, displays information about the object or area at the current cursor location. This display can be toggled on and off from the **View** menu.

Dialog Boxes

When you use an SP Perspectives primary window to perform certain actions, a dialog box is presented. The dialog box solicits or displays additional information needed to complete the action. The specific information or actions contained in the dialog box depend on the application.

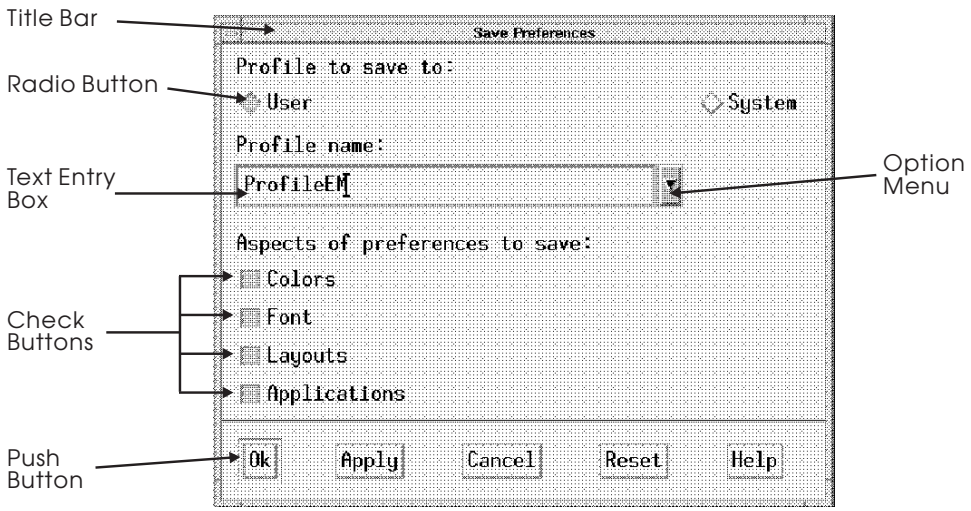


Figure 40. An SP Perspectives Dialog Box

SP Perspectives dialog boxes may contain the following elements (not every dialog box contains all of these elements):

- The title bar displays the name of the action being performed, the object name or type, and, optionally, the system partition name.
- Text entry boxes provide for entering or viewing text information.
- Check buttons provide for selecting one or more choices from a list of options that are not mutually exclusive.
- Radio buttons provide for making a single choice from a list of mutually exclusive options.
- Option menus provide for entering information and selecting a single choice from a list of mutually exclusive options.
- Push buttons provide for selecting an action:
  - **OK** performs the action based on information entered or selected in the dialog box and closes the dialog box.
  - **Apply** performs any changes made to information entered or selected in the dialog box and leaves the dialog box open.
  - **Cancel** resets any changes made to information in the dialog box to the last applied values and closes the dialog box.
  - **Reset** returns any changes made to information in the dialog box to the last applied values and leaves the dialog box open.
  - **Help** displays help about the purpose of and items in the dialog box.

Notebooks

An SP Perspectives notebook is a type of dialog box that is in a notebook format. Each page in the notebook contains attribute information (defining or operational characteristics) about system objects. Like attributes are grouped on the same page and “Tabs” provide a visual identification of the category of attributes on the page. The specific attribute information contained on a notebook page depends on what system object is selected (for example, the hardware or a system event).

Notebooks are provided as a mechanism for viewing or altering object attributes.

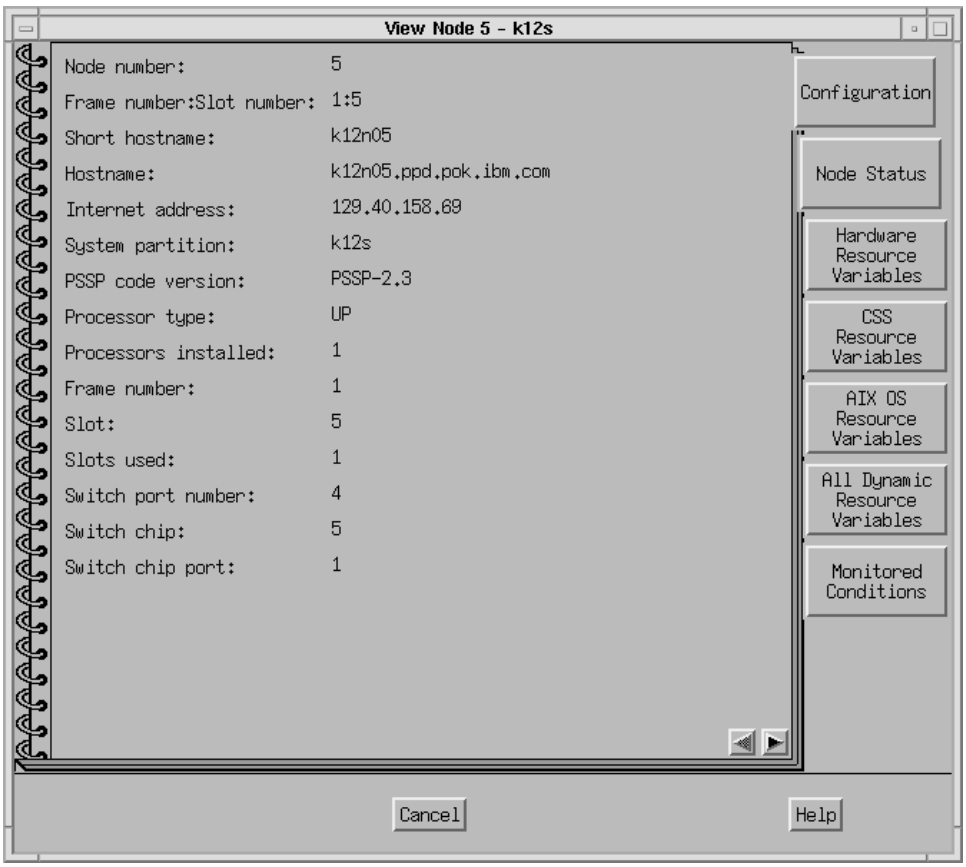


Figure 41. An SP Perspectives Notebook

SP Perspectives notebooks may contain the following elements (not every notebook contains all of these elements):

- The title bar displays the action (view, modify), the object name or type, and, optionally, the system partition name.
- Text entry boxes provide for entering text information for object attributes that can be modified.
- Static text entry fields provide for displaying object attribute information that cannot be modified (view only).
- Check buttons provide for selecting one or more choices from a list of options that are not mutually exclusive.
- Radio buttons provide for making a single choice from a list of mutually exclusive options.
- Option menus provide for entering information and selecting a single choice from a list of mutually exclusive options.
- Push buttons provide for selecting an action:
  - **OK** updates attribute information based on data entered in the notebook page and closes the notebook.
  - **Apply** performs any changes made to information entered in the notebook and leaves the notebook open.
  - **Cancel** resets any changes made to information in the notebook to the last applied values and closes the notebook.

- **Reset** returns any changes made to information in the notebook to the last applied values and leaves the notebook open.
- **Help** displays help about the purpose of items in the notebook.

Message Boxes

A message box is a type of dialog box used to convey a message to you. There are four types of messages:

- Informational
- Error
- Fatal
- Warning

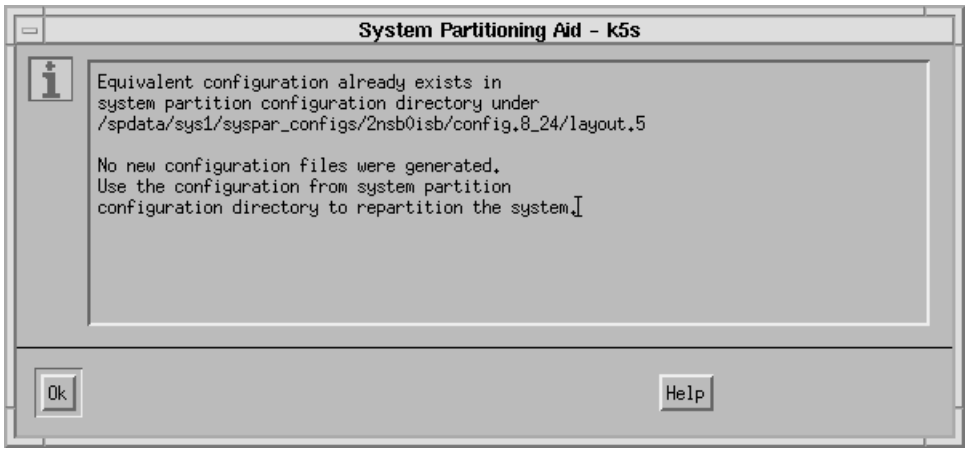


Figure 42. An SP Perspectives GUI Message Box

Elements common to SP Perspectives message boxes are:

- The title bar displays the name of the perspective.
- A symbol represents the message type:
  - Informational message

  - Error message / Fatal message

  - Warning message

- The text of the message.
- Push buttons provide for selecting an action, depending on the type of message:

- Informational, Error, Fatal
  - **OK** closes the message box.
  - **Help** displays extended information on the contents of the message and provides a way to view internal diagnostics information.
- Warning
  - **Enter**/**OK** confirms the action.
  - **Cancel** stops the action.
  - **Help** displays extended information on the contents of the message and provides a way to view internal diagnostics information.

Accessing Help

You can access help for SP Perspectives from:

- The menu bar **Help** button
- The Information Area
- The dialog box, notebook, and message box **Help** buttons

The Menu Bar Help Button

The menu bar **Help** button contains items that provide information about windows, actions (tasks), and objects. To display the menu items, click on **Help**.

- **Getting Started** provides information on starting SP Perspectives.
- **Overview** activates a standalone help volume that contains overview and task-oriented information for SP Perspectives.
- **Contents** displays a list of hyper-linked help topics for SP Perspectives.
- **Tasks** accesses task-oriented help information for the perspective that you are in when you click on **Tasks**.
- **Using Help** provides information on how to access help for SP Perspectives.
- **About SP Perspectives** displays the name and version of the application.
- **Give Feedback to IBM** provides information about how to contact the SP Perspectives team with your comments.

The Information Area

The Information Area is present on primary windows and displays information about the object or area at the current cursor location. This display can be toggled on and off from the **View** menu.

The Dialog Box, Notebook, and Message Box Help Buttons

The dialog box and notebook **Help** buttons display information about the purpose of and items in the dialog box or notebook. The message box **Help** button displays extended information on the contents of the message and provides a way to view internal diagnostics information.

Performing Common Administration Tasks with SP Perspectives

The following table provides an overview of how to perform common SP system administration tasks using SP Perspectives. **SELECT** indicates clicking on an icon or a menu choice; **PRESS** indicates a button press; → indicates a menu item sequence.

Task	Perspectives
Bringing up the Launch Pad	**TYPE** perspectives
Launching applications	Double click on application icon in Launch Pad or 1. **SELECT** application icon 2. **SELECT** Actions → Launch
Closing the Launch Pad / Closing a perspective	**SELECT** Window→ Close
Powering off one node	From hardware perspective menu bar 1. **SELECT** node to be powered off 2. **SELECT** Actions → Nodes → Power Off... 3. **SELECT** power off options 4. **PRESS** Apply or From hardware perspective tool bar 1. **SELECT** node to be powered off 2. **PRESS** Power off icon from Tool bar 3. **SELECT** power off options 4. **PRESS** Apply
Powering off more than one node	From hardware perspective menu bar 1. **SELECT** nodes to be powered off 2. **SELECT** Actions → Nodes → Power Off... 3. **SELECT** power off options 4. **PRESS** Apply or From hardware perspective menu bar with Node Groups 1. **SELECT** node group to be powered off 2. **SELECT** Actions → Node Group → Power Off... 3. **SELECT** power off options 4. **PRESS** Apply or From hardware perspective tool bar 1. **SELECT** nodes to be powered off 2. **PRESS** Power off icon from Tool bar 3. **SELECT** power off options 4. **PRESS** Apply

Task	Perspectives
Powering on one node	From hardware perspective menu bar 1. **SELECT** node to be powered on 2. **SELECT** Actions → Nodes → Power On... 3. **SELECT** power on options 4. **PRESS** Apply or From hardware perspective tool bar 1. **SELECT** node to be powered on 2. **PRESS** Power on icon from Tool bar 3. **SELECT** power on options 4. **PRESS** Apply
Powering on more than one node	From hardware perspective menu bar 1. **SELECT** nodes to be powered on 2. **SELECT** Actions → Nodes → Power On... 3. **SELECT** power on options 4. **PRESS** Apply or From hardware perspective menu bar with Node Groups 1. **SELECT** node groups to be powered on 2. **SELECT** Actions → Node Groups → Power On... 3. **SELECT** power on options 4. **PRESS** Apply or From hardware perspective tool bar 1. **SELECT** nodes to be powered on 2. **PRESS** Power on icon from Tool bar 3. **SELECT** power on options 4. **PRESS** Apply
Verifying node status for many nodes: hostResponds, powerLED, switchResponds, and 3 digit display	From hardware perspective menu bar 1. Click on the Nodes pane 2. **SELECT** View→ Monitor 3. **SELECT** hostResponds 4. **PRESS** Apply Repeat these steps for individual viewing of powerLED and switchResponds status, or select all three. Check to see if any node icons do not turn green. If a red X is showing, then there is a problem with that node. Three digit display is always displayed in it's own window. Bring up the 3 digit display: 1. Click on the Nodes pane 2. **SELECT** Actions→ Nodes→ 3 digit display

Task	Perspectives
Verifying node status for one node: hostResponds, powerLED, switchResponds, and 3 digit display	From hardware perspective node notebook 1. Double click on a node 2. **SELECT** Node Status tab The Node Status page does not display the value of the 3 digit display for the node. To bring up the 3 digit display: 1. Click on the Nodes pane 2. **SELECT** Actions→ Nodes→ 3 digit display
Viewing node information	1. Double click on a node 2. **SELECT** Tab corresponding to desired page
Changing to the system (global) view	From hardware perspective menu bar 1. **SELECT** System icon 2. **SELECT** Actions→CWS/System/Syspars→Set Current System Partition
Changing to a system partition (syspar) view	From hardware perspective menu bar 1. **SELECT** Syspar icon 2. **SELECT** Actions→CWS/System/Syspars→Set Current System Partition

Chapter 21. Using the SP System Monitor

The SP System Monitor provides a way for authorized operators and administrators to operate and monitor the system hardware. Because the SP Perspectives GUI provides access to a wider range of system management tasks, we recommend its use over the SP System Monitor. For more information on the SP Perspectives GUI, refer to Chapter 20, “Using SP Perspectives” on page 281.

There are four main parts to the SP System Monitor:

Hardware Monitor A set of commands and a daemon used to monitor and control the SP hardware platform.

SP System Monitor GUI A graphical user interface to the Hardware Monitor that provides a point-and-click method to view the system status and interact with the frame, node, and switch controls. The SP System Monitor GUI provides the option of viewing an individual system partition or viewing the system globally (all system partitions).

Command interface The **spmon**, **hmmon**, **hmcmds**, **s1term**, and **nodecond** commands with their flags and parameters, provide a way to monitor and control the system at an individual system partition level or at a global (all system partitions) level. These can be used from an ASCII terminal, such as those used on a dial-up line, from scripts, or from another terminal.

Logging daemon The **splogd** logging daemon logs SP hardware state changes, reports SP hardware errors, and provides a state change alert.

See Figure 43 on page 298 for an overview of the SP System Monitor.

This chapter provides conceptual information and procedures for configuring and using the SP System Monitor. It also shows you how to operate the system controls and monitor the system activity using the SP System Monitor GUI.

© Copyright IBM Corp. 1995, 1997

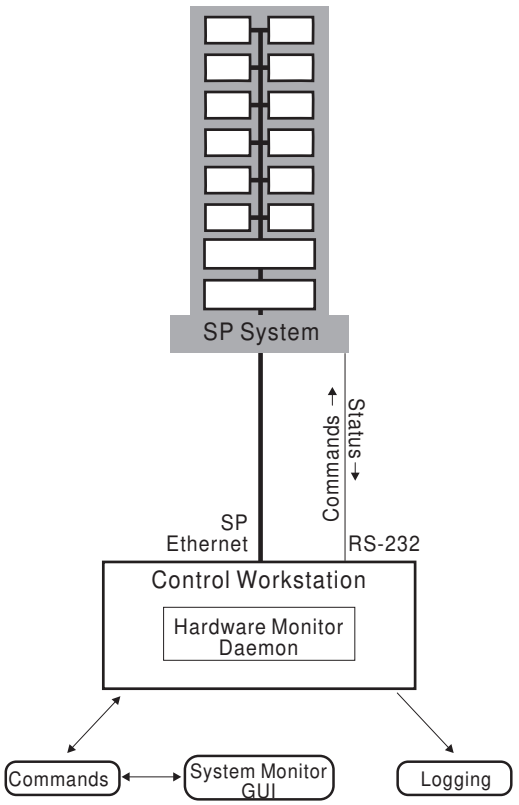


Figure 43. The SP System Monitor

Understanding the Hardware Monitor

The Hardware Monitor consists of a daemon, named **hardmon**, and a set of client commands. The daemon executes on the control workstation and, using the RS-232 lines to each frame, polls the frames for the state of the hardware within the frame. The daemon also sends hardware-level commands to each frame to change the state of the hardware within the frame in some way. The client commands **spmon** and **hmmon** interact with the daemon to obtain the current or changing state of the hardware. The client commands **spmon** and **hmcmds** also interact with the daemon to change the state of (that is, control) the hardware. The SP System Monitor Graphical User Interface issues the **spmon** command internally to control and monitor the hardware. The **splogd** logging daemon is also a client of the hardware monitor for logging purposes.

The **hardmon** daemon reports variables representing the state of the frame, node and switch hardware. These state variables are made available through the **spmon** and **hmmon** client commands and through the SP System Monitor GUI. You can find a list of these variables in Appendix D, “System Monitor Variables” on page 549. This list provides variable names, display attributes and descriptions for all variables available in the SP System Monitor. The **hmmon** command, with the appropriate flag, also provides a list of variable names and descriptions. You will see these variable names in the SP System Monitor log. They also serve as parameters to the **spmon** and **hmmon** commands.

Frame Hardware State

The **hardmon** daemon reports variables to *monitor* the frame hardware status. This information includes:

- Switch failure
- Power LEDs
- Environment LEDs
- Temperature
- Voltage
- State of the power supply

The **hardmon** daemon allows the administrator to *control* the frame hardware state. This information includes:

- Power on/off

Node Hardware State

The **hardmon** daemon reports variables to *monitor* the node hardware status. This information includes:

- Three-digit display
- Power and Environment LEDs
- Temperature
- Voltage
- Fan failure

Note: Some environment information, such as fan speed and voltage, is not available for a 604 High Node.

The **hardmon** daemon allows the administrator to *control* the node hardware state. This information includes:

- Power on/off
- Key Mode Switch
- Reset button

Switch Hardware State

The **hardmon** daemon reports variables to *monitor* the switch hardware status. This information includes:

- Temperature
- Voltage
- Fan failure
- Power and Environment LEDs

The **hardmon** daemon allows the administrator to *control* the switch hardware state. This information includes:

- Power on/off
- Change multiplexor clock setting

Note: On SP Switch systems, you must use the **Eclock** command to change the multiplexor clock setting.

Authority

There are four classes of authority for the SP System Monitor:

1. Monitor authority permits users to display system statistics.
2. Control authority permits users to display system statistics and control the hardware.
3. S1 authority permits read and/or write access to a node's serial port.

4. Administrative authority permits a user to issue administrative commands to **hardmon**.

Configuring the SP System Monitor

The system administration tasks for configuring the SP System Monitor include authorizing users and configuring the error log.

Step 1: Authorize Users for the SP System Monitor

The SP System Monitor Access Control Lists (ACLs) are found in **/spdata/sys1/spmon/hmacls** on the control workstation. Edit this file if you wish to add users for your system. The **/spdata/sys1/spmon/hmacls** file is initially set up giving all levels of authority (including administrator) to the same user that is defined as the primary authentication services administrator by the **setup_authent** command (refer to Chapter 14, "Security Features of the SP System" on page 157 for more information). The **hardmon** principal is initially set up with monitor authority (for use by **splogd**).

The fields for each entry in the **/spdata/sys1/spmon/hmacls** file are

object name permissions

where:

object A frame number or hostname (where the **hardmon** daemon is running)

name A Kerberos principal name and optional instance

permissions

- **v** Virtual Front Operator Panel (VFOP). This gives permission to issue commands to the hardware.
- **s** Serial link. This gives permission to read and write to a serial port.
- **m** Monitor. This gives permission to receive state changes.
- **a** Administrative. This gives authority to control **hardmon**.

Invoke the **hmadm setacls** command after the ACL configuration file has been modified to update the hardware monitor daemon's internal ACL tables.

Refer to Chapter 14, "Security Features of the SP System" on page 157 for more information on security considerations.

Step 2: Configure the SP System Monitor Error Log

When the hardware supervisors indicate a warning or shutdown condition, the SP System Monitor writes a message using the AIX **syslog** facility and the AIX error log facility. For example, when the hardware supervisors determine that a fan has failed, the SP System Monitor writes a precise message into the log file that includes the time, node, type of error, variable name, and, in some cases, associated values.

The installation process creates the default syslog file **/var/adm/SPlogs/SPdaemon.log** on the control workstation. You may want to configure your system to send the system log information to other locations. For

example, you may want to send the **SPdaemon.log** messages to another workstation for convenience. You can do this using the *@hostname* parameter in the /**etc**/**syslog.conf** file. For more details, refer to *IBM AIX Version 4.1 Files Reference*. The facility name for the SP System Monitor is **daemon**.

Customizing SP System Monitor GUI Default Resources

1. The default resources are in /**usr**/**lib**/**X11**/**app-defaults**/**Spmon**
2. To customize, add the contents of /**usr**/**lib**/**X11**/**app-defaults**/**Spmon** to the **.Xdefaults** file in your $HOME directory. **Note:** If you have any comments in the **.Xdefaults** file, add a : to the comment line.

Refer to "/usr/lib/X11/app-defaults/Spmon" on page 547 for a copy of this file.

Using the Command Line Interfaces

The SP system includes command line interfaces that allow you to perform the SP System Monitor functions without using a graphical user interface. With the **spmon** and **hmmon** commands you can query the value of a state variable or monitor that variable for a change in state. With the **spmon** and **hmcmds** commands you can control the nodes, frames and switches. The **hmmon** and **hmcmds** commands provide equivalent function to the **spmon** command but provide a more compact interface than **spmon** that can be appropriate in scripts. The **spmon** command is provided for compatibility with earlier SP software releases.

By default, these commands operate at a system partition level. That is, they work only on nodes *in the current system partition*. (To determine the current partition, issue the **spget_syspar** command.)

Note: Hardware Monitor queries or commands using a wildcard character do not return an error code if nodes outside of the current system partition are specified by the wildcard. Nodes outside of the current partition are treated as if they don't exist.

Each of the commands has a **-G** flag that removes system partition boundaries, allowing the command to work on any hardware in the SP system. The **-G** option is *always* needed to monitor or control frame or switch hardware. Frames and switches are considered to be outside all system partitions.

Before using any SP System Monitor command your identity must be authenticated. Use the **kinit** command to login to the SP authentication services. If your site uses an alternative to the SP authentication services, you must use authentication procedures appropriate to such an alternative. (Refer to Chapter 14, "Security Features of the SP System" on page 157 for more information on SP authentication services.)

See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for a complete description of all SP System Monitor commands.

Using the SP System Monitor GUI

The SP System Monitor GUI is a graphical user interface that allows you to view hardware objects at either a system partition level or across system partitions and operate the frame, node, and switch controls. Using a point-and-click method, you can select from menus and icons to display information in a concise visual format. The SP System Monitor GUI windows use different colors to indicate variations in status. The windows are updated as the state of the hardware changes.

SP System Monitor GUI Windows

The SP System Monitor GUI offers several ways for you to obtain monitoring information. You can get a high level overview of the system activity and status or just focus on one frame or slot (which may contain a switch or processing node). You have a similar choice when operating the controls and can initiate actions across the entire system, across one frame, or on just one slot.

Following is a list of the SP System Monitor GUI windows and an explanation of the information they display. The order in which you approach system monitoring depends on the task you are performing. Once you are logged in, you can start with any of these windows and move freely between them.

Topology Displays variables representing SP LANs (local area networks) connecting SP processor nodes. It uses colors to indicate the status of the adapters connecting the nodes to the LANs.

All Node Summary Display
Displays the state of one variable for slots 1–16. Allows you to monitor status of items such as the three-digit display or power LEDs for each slot in a single partition or across all partitions (Global view).

Display Layouts Displays windows for a particular frame, node, or switch. Allows you to monitor the status of several variables and interact with the controls.

Global Controls Provides access to the front panel controls of nodes. Allows you to issue actions on several nodes at once, within a single partition or across all partitions (Global view).

Dynamic Configuration

If a node is not created in the All Node Summary, Topology, or in the Global Controls display, or if a node or frame does not appear in a Display Layouts selection list, then there are two possible explanations:

- The variable is not valid for your hardware type. Not all variables are valid for all three types of nodes. (This only occurs on the All Node Summary Display or Topology display.)
- Some nodes were not included in the SP System Monitor configuration. Any SP frame that is not connected through RS-232 when the **hardmon** daemon is initialized will not be recognized in the system monitor configuration and its nodes will not be selectable. Once the RS-232 is connected and responds to the **hardmon** daemon poll, it will be added to the SP System Monitor configuration.

If this is the case, the All Node Summary, Topology, and Global Controls windows must be closed and reopened for the entire configuration to be displayed. Display Layouts will become selectable for frames and nodes once they are included in the SP System Monitor configuration.

If all frames don't come up in the GUI, you may not be authorized for them. Refer to “Step 1: Authorize Users for the SP System Monitor” on page 300 for information.

Any active GUI sessions should be terminated before deleting a system partition. Failure to do so may result in an SDR connection failure within the GUI logic or in obsolete windows containing a view of the deleted system partition.

Window Elements

The SP System Monitor GUI windows contain elements common to many graphical user interfaces, such as menu bars, scroll bars, and radio buttons. If you are not familiar with these elements, refer to *IBM Getting Started: Using IBM RS/6000* for an explanation of how to manage the windows and interact with the common interface elements.

Interface Instructions

The tasks in the section are presented with step-by-step instructions. The instructions use a format that distinguishes between the user action and the system response.

User actions appear in uppercase bold type.

PRESS Cancel

Selections from a menu bar are indicated with an →.

SELECT SP → Display Layouts

The system response to an action follows the ▲ symbol.

▲ The diagnostic display appears.

Closing Windows to End the SP System Monitor GUI Session

In addition to the window menu button in the upper-left corner (provided by the window manager), every SP System Monitor GUI window has either a push button or a menu bar choice to close the window.

Starting the SP System Monitor GUI

Follow these steps to start the SP System Monitor GUI after logging in to the control workstation.

Note

If you are using a system other than the control workstation as the X-server, you need to export the SP System Monitor GUI. The following example shows you how to do it in **ksh**.

On the control workstation, set the DISPLAY to your server. Enter:

```
export DISPLAY=displayhost:0
```

displayhost is the name of the X-server.

On the X-server, enter:

```
xhost +controlhost
```

controlhost is the name of control workstation.

Step 1: Get Authentication to Use the Hardware Monitor.

If you have not already been authenticated to use the hardware monitor server, use the **kinit** command to login to the SP authentication system.

Refer to Chapter 14, “Security Features of the SP System” on page 157 for more information on authentication.

Step 2: Start the SP System Monitor GUI

From the control workstation:

TYPE **/usr/lpp/ssp/bin/spmon -g**

Note: For a Global view (all system partitions), type **/usr/lpp/ssp/bin/spmon -G -g**

▲ The SP System Monitor GUI main window (in this example, for system K43s, View for partition K43sp2) appears.

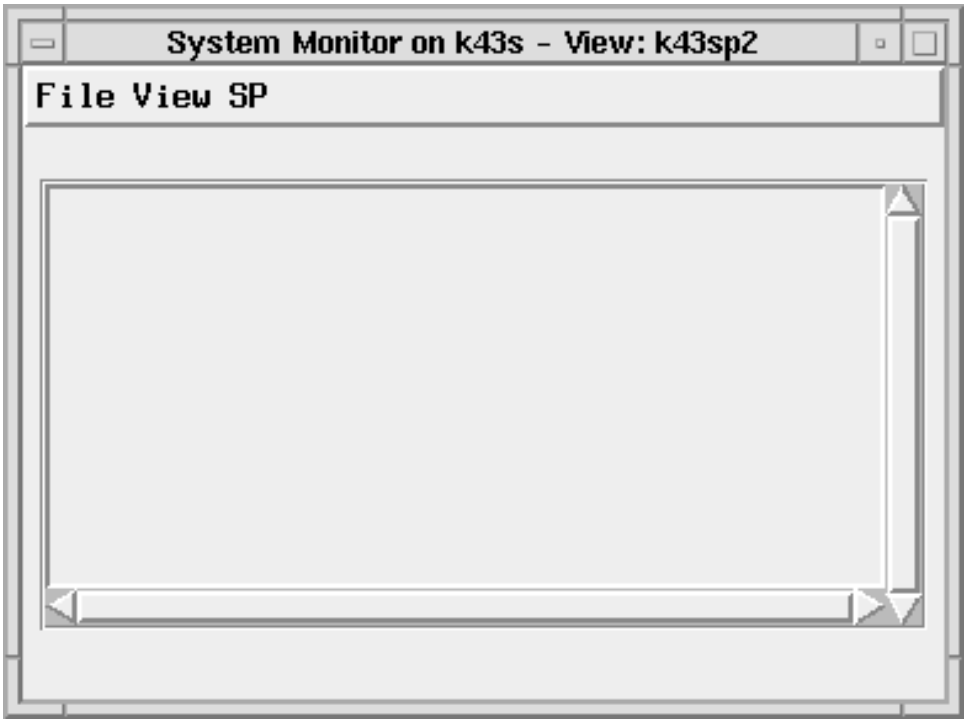


Figure 44. SP System Monitor GUI Main Window

The main window contains:

- A title that includes your control workstation name and the default view (either for an individual system partition or for a Global view) the SP System Monitor will display. If you have High Availability Control Workstation (HACWS) function installed, the title also includes the word 'primary' or 'backup' to indicate which control workstation is active.
- A menu bar with three choices, **File** (that allows you to exit your SP System Monitor GUI session), **View** (that allows you to specify an individual system partition or Global view), and **SP** (a pulldown menu that displays other window selections).
- A window area that echoes commands issued internally by the SP System Monitor GUI.

Step 3: Select View

SELECT **View**,

▲ A pulldown menu appears, shown in Figure 45 on page 306.

The pulldown menu lists the Global (All System Partitions) option and individual system partition names. The view option selected serves as the default view of the system.

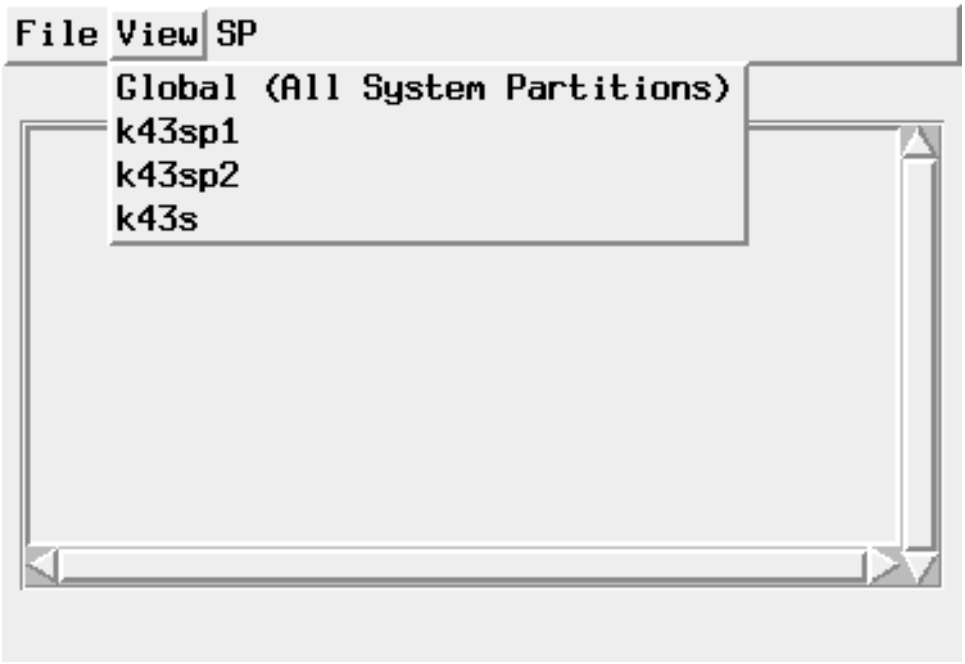


Figure 45. View Pulldown Menu

Step 4: Select SP

SELECT **SP**,

▲ A pulldown menu appears, shown in Figure 46.

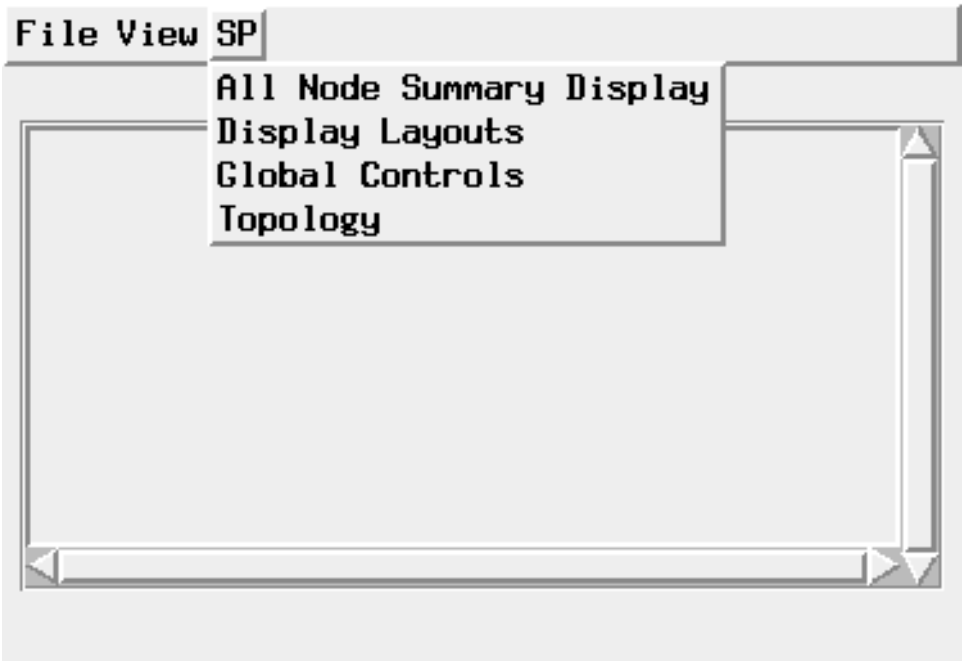


Figure 46. SP Pulldown Menu

Displaying Frame, Node, and Switch Layouts

Start at the menu bar on the SP System Monitor GUI to access the Display Layouts window.

From the SP System Monitor GUI Main Window:

SELECT SP → Display Layouts

▲ The display layout selection window appears.

SELECT View

▲ A pulldown menu lists defined system partitions and 'Global' (all system partitions) selections (in a multipartitioned system; if a system is not partitioned, 'View' won't appear on the menu bar)

SELECT The desired view of the system

SELECT The number of the frame or the number of the frame and slot in the option menu at the top of the screen

▲ The layout lists become active and you can select choices valid for the specified frame/node type.

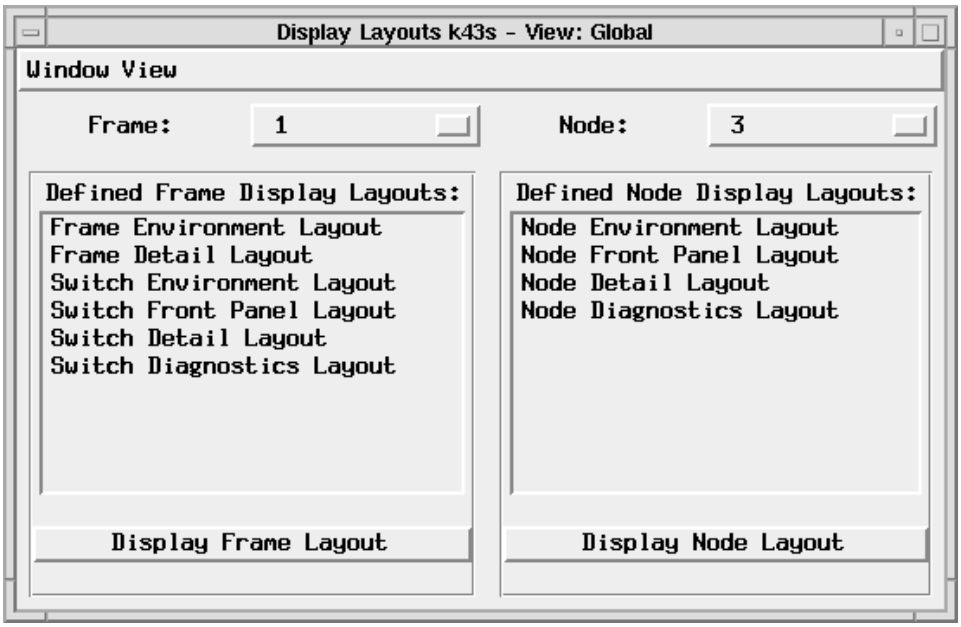


Figure 47. SP System Monitor GUI Display Layouts Selection Window

The Display Layouts window lists all the available layouts for the selected frame and node. Each layout is a window and displays a grouping of variables and controls. Layout windows provide you with detailed information about the hardware and, if you have the authorization, allow you to operate the system controls.

Note: On SP systems with more than one system partition, you must use the 'Global' view of the system to control frames and switches.

To select and display a layout:

SELECT The layout of your choice from one of the lists

PRESS Display Frame Layout or Display Node Layout

▲ The layout window of your choice appears.

Table 11 summarizes the information available on the layouts that come with the SP system.

Defined Layouts

Table 11. Frame, Node, and Switch Layouts

Select:	**To get:**
Frame Environment	Information about frame power and communications Control for the frame power supplies
Frame Detail	Information on type, code version, etc.
Switch Environment	Temperature and voltage indicators
Switch Front Panel	Power, environment, and communications indicators, etc.
Switch Detail	Information on type, code version, etc.
Switch Diagnostics	Information about the cause of switch failures
Node Environment	Temperature and voltage status
Node Front Panel	Three-digit display, power, environment, and communications indicators Controls for power, key mode switch, reset button, fencing, and opening tty
Node Detail	Information on type, code version, etc.
Node Diagnostics	Information about the causes of node failures (The 604 High Node does not have a Node Diagnostics layout.)

Monitoring and Controlling the Switch

You can use the Switch Front Panel layout to view the switch indicators and monitor and control the switch.

Note: On SP systems with more than one system partition, you must use the 'Global' view of the system to control the switch. A 'Global' view is required to control switches residing in slots 1–16.

On a Frame with Nodes

At the Display Layouts window:

SELECT The number of the frame that the switch is on from the option menu

SELECT Switch Front Panel Layout

PRESS Display Frame Layout

▲ The Switch Panel for that frame appears.

On an All Switch Frame

The available layouts for an all switch frame will appear on the right side of the menu where node selections are displayed.

SELECT The number of the frame

SELECT The number of the slot with the switch you want to display

SELECT Switch Front Panel Layout

PRESS Display Node Layout

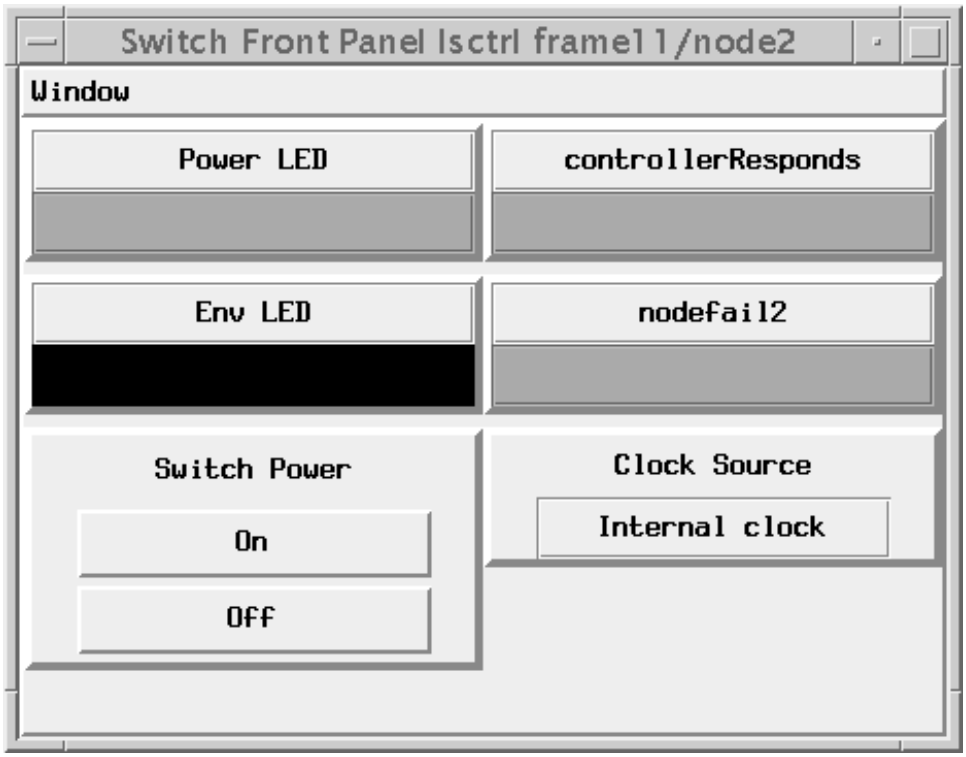


Figure 48. SP System Monitor GUI Switch Front Panel Layout Window

This window displays the following indicators:

Power LED	Status of the switch power light. Green (running), gray (off but power is available), or black (power not available).
Env LED	Status of the switch environment light. Black (environment light is off, no indication of failure), gray (environment light is flashing a failure), or yellow (environment light is on indicating an environment warning).
nodeFail*X*	*X* is the slot where the switch is. Indicates if the switch supervisor is responding to requests from the frame supervisor. Green (responding) or red (not responding).
controllerResponds	Indicates if the frame supervisor is responding to requests from the control workstation. Green (responding) or red (not responding).

Switch Power	Power controls for the switch.
Clock Source	Reflects the current clock source selected.

Notes:

1. We recommend terminating all applications that use the switch before powering off the switch.
2. You must use the **Eclock** command to change the clock source. See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for details on command syntax.

Operating Frame Environment Controls

The Frame Environment Layout window monitors the status of your system frame and allows you to turn the power on or off for the entire frame.

Note: On SP systems with more than one system partition, you must use the 'Global' view of the system to control frames. The number of indicators varies according to your frame supervisor. The Type 18/22 frame supervisors (Scalable Electrical Power Base Unit) have more indicators than the Type 17/20 (Power Distribution Unit) frame supervisors.

Warning

Powering off the frame power supply turns off the nodes and switch. Perform **Shutdown** before turning off the power supplies. See "Using Global Controls" on page 320.

Invoking the Frame Environment Layout

At the SP System Monitor GUI main window,

SELECT	SP → Display Layouts
	▲ The Display Layouts window appears. This window is shown with "Displaying Frame, Node, and Switch Layouts" on page 307.
SELECT	The number of the frame you want to see
SELECT	Frame Environment Layout
PRESS	Display Frame Layout
	▲ The Environment window for that frame appears.

Frame Environment Indicators and Controls

controllerResponds	Indicates if the frame supervisor is responding to requests from the control workstation. Green (responding) or red (not responding).
rs232 Active	Status of the RS-232 communication with the control workstation. Yellow (RS-232 activity) or black (no RS-232 activity) or gray (RS-232 error).

Bus Error	Status of communication with the nodes over the supervisor bus. Gray (no nodes responding) or yellow (1 or more nodes responding) or black (all nodes responding).
DC OK	Status of the DC power. Green (DC OK) or gray (DC Failure).

Type 17 (Power Distribution Unit) Frame Supervisor Indicators

AC OK	Status of the AC power. Green (AC OK) or gray (AC Failure).
Power_DC_x	Indicators for the four frame power supplies A–D. Green (OK) or red (Failure).
On/Off	Power control for frame power supplies A–D. **Warning**: When powering off a main PCI frame, other frames may also be powered off. See *IBM RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment*, SC23-3869, for more information on PCI.

Type 18 (Scalable Electrical Power Base Unit) Frame Supervisor Indicators

Although as many as four power supply modules may be installed in the Scalable Electrical Power Base Unit, all frame power is controlled with a single setting.

Modules Present	Power supply modules installed. Green (at least one power module is present) or gray (no power modules present).
ModulePowerX	Status of power supply modules installed. Green (OK) or red (Failure).
On/Off	Single power control for entire frame. **Warning**: When powering off a main PCI frame, other frames may also be powered off. See *IBM RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment*, SC23-3869, for more information on PCI.

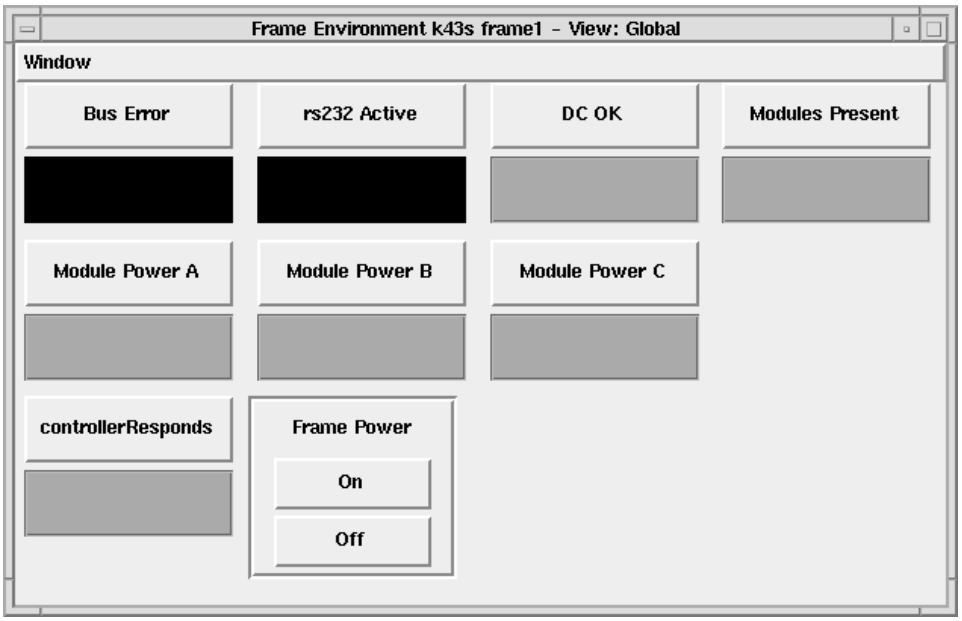


Figure 49. Frame Environment Layout for Type 18 Frame Supervisor

Type 20 Frame Supervisor Indicators

The Type 20 frame supervisor has the same indicators as the Type 17 (except for rs232 Active), along with the rs232 Active1, rs232 Active2, and controllerTail X as described for Type 22.

Type 22 Frame Supervisor Indicators

The Type 22 frame supervisor has the same indicators as the Type 18 (except for rs232 Active), along with:

rs232 Active1	Status of the RS-232 Active1 communication with the control workstation. Yellow (RS-232 activity) or black (no RS-232 activity) or gray (RS-232 error).
rs232 Active2	Status of the RS-232 Active2 communication with the control workstation. Yellow (RS-232 activity) or black (no RS-232 activity) or gray (RS-232 error).
controllerTail *X*	The monitor is connected to the frame supervisor via S1 or S2. For whichever controller tail is used, green (active) or red (inactive).

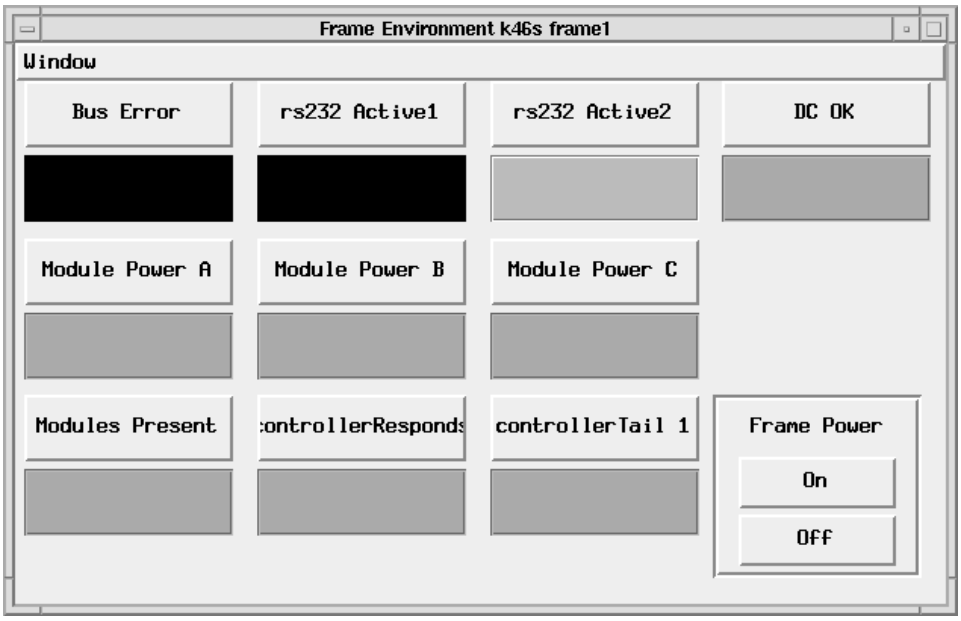


Figure 50. Frame Environment Layout for Type 22 Frame Supervisor

Operating the Front Panel Controls

If a processor node is having problems, you may want to shut it down, reset it, or prepare it for servicing. The node Front Panel window offers you the same controls that you would have on the RS/6000 front control panel.

You can also use Front Panel controls to isolate nodes from the switch, in order to avoid a switch fault, if you want to shut down or reset nodes.

Invoking the Node Front Panel Layout

At the SP System Monitor GUI main window,

SELECT SP → Display Layouts

▲ The Display Layouts window appears. This window is shown with “Displaying Frame, Node, and Switch Layouts” on page 307.

SELECT The number of the frame you want to see

SELECT The number of the node within the frame

SELECT Node Front Panel Layout

PRESS Display Node Layout

▲ The Front Panel window for that node appears.

Note: If you do not have a switch installed on your system, Fence and Unfence will not appear on the Node Front Panel Display.

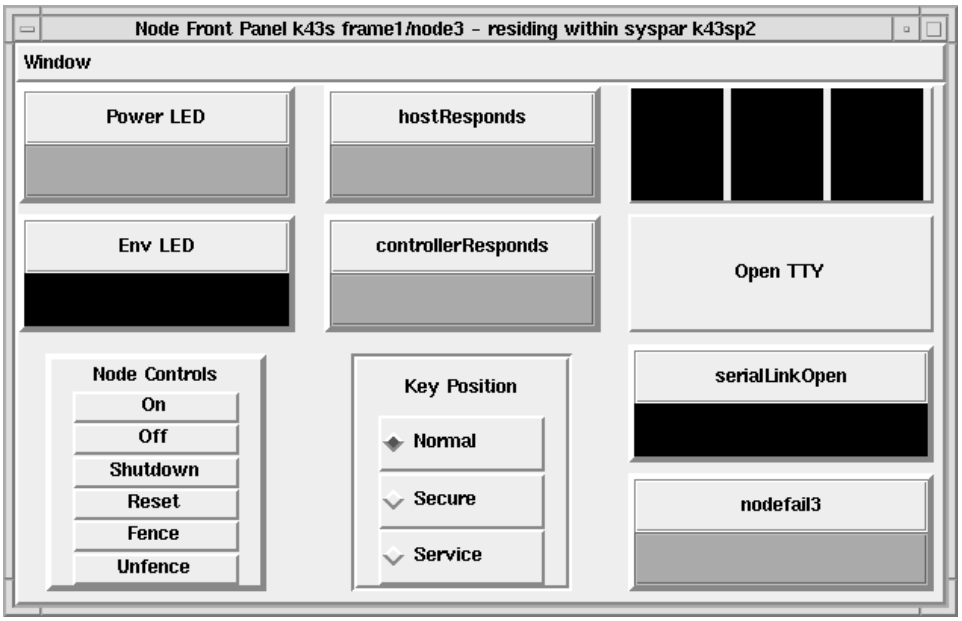


Figure 51. Node Front Panel Display

Node Front Panel Indicators

Power LED	Status of the node power light. Green (running), gray (off, but power is available), or black (power not available).
Env LED	Status of the node environment light. Black (environment light is off, no indication of failure), gray (environment light is flashing a failure), or yellow (environment light is on, indicating an environment warning).
controllerResponds	Indicates if the node supervisor is responding to requests. Green (responding) or red (not responding). Green (on) under normal operating conditions.
hostResponds	Status of communication between the processor node and the control workstation. Green (on) indicating node is up.
Three-digit display	Shows three-digit codes or black if nothing to display.
serialLinkOpen	Status of the serial port 1 (**tty** connection). Yellow (open) or black (closed).
nodeFail*X*	*X* is the slot where the node is. Indicates if the node supervisor is responding to requests from the frame supervisor. Green (responding) or red (not responding).

Node Front Panel Controls

On/Off/Reset	Power controls for the node, with Shutdown and Restart options.
Shutdown	System shutdown without powering off.
Fence	Fences the node from the switch
Unfence	Returns the node to the switch
Normal/Secure/Service	Key mode switch controls for the node
Open TTY	Allows you to open an **aixterm** connection to the node

Turning Node Power Off or On

Note: Powering on the 604 High Node takes a longer time than powering on a thin or wide node. It may take 15 minutes to power on a 604 High Node.

Be aware that turning off the power ends all processes running on the node and data may be lost. Perform a shutdown, either to individual nodes or globally to the SP system, before turning off power. Shutdown can be selected from the Front Panel and Global Controls display. (Refer to "Operating the Front Panel Controls" on page 313 or refer to "Using Global Controls" on page 320 for instructions on accessing these displays.)

Chapter 3, "Starting Up and Shutting Down the SP System" on page 19 describes how the **cshutdown** and **cstartup** commands work.

From the Node Front Panel:

CHECK The color of the power indicator.

- If the indicator is green, the power is on.
- If the indicator is gray, the power is available but not on.
- If the indicator is black, the power is unavailable to the node.

Turning the Power Off:

PRESS Off

▲ The Power Off menu is displayed.

Note: If you do not have a switch installed on your system, Fence and Enable AutoJoin will not appear on the Power Off Menu.

Figure 52. Power Off Menu

PRESS Options:

Fence
: Fences the node from the switch

Enable AutoJoin (can be selected only in conjunction with Fence)
: Enables node to automatically rejoin the switch when the node has restarted

Power Off
: Turns off node power.

Shutdown → Off
: Halts node, ignoring **cshutdown** dependencies, and turns off power

Cluster Shutdown → Off
: Halts node and turns off power only if **cshutdown** dependencies are met

▲ The power indicator changes to gray.

Turning the Power On:

PRESS On

▲ The Power On menu is displayed.

Figure 53. Power On Menu

PRESS One of the two options:

Power On
: Turns on power to nodes.

Cluster Power On
: Turns on power to nodes only if **cstartup** dependencies are met

▲ The power indicator changes to green and a three-digit display shows various codes while the node goes through its system start-up routine. When this completes without error, the three-digit display remains black, the hostResponds indicator turns green, and the node is ready (up).

Note: If you set *bootp_response* to `install` the powering process may take 20 minutes or more (because it is installing AIX and **pssp** on all of the nodes). If you set *bootp_response* to `customize` it will go faster. You can check the current value of *bootp_response* by querying the node class object in the SDR. Enter:

```
splstdata -b
```

See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for details on command syntax. Also, see the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* for more information on the *bootp_response*.

If the system stops longer than three minutes with a three-digit code displayed, refer to information on diagnosing boot problems in the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* and the *IBM RS/6000 Problem Solving Guide* or the *IBM RS/6000 SP: Maintenance Information Volume 1* to find out the meaning of the code and respond accordingly.

Resetting Nodes

The reset function allows you to boot the node when its key mode switch is in the Normal or Service position and to read error codes and diagnostic messages after a flashing **888** is displayed in the three-digit display. Be aware that resetting a node while it is operating can also result in corrupted or lost data.

From the Node Front Panel:

PRESS Reset

▲ The Reset menu is displayed.

Note: If you do not have a switch installed on your system, Fence and Enable AutoJoin will not appear on the Reset Menu.

Figure 54. Reset Menu

PRESS Options:

Fence
Fences node from the switch

Enable AutoJoin (can be selected only in conjunction with Fence)
Enables node to automatically rejoin the switch when the node has restarted

Reset
Reboots the nodes

Shutdown → Restart
Halts and reboots the nodes, ignoring **cstartup** dependencies

Cluster Shutdown → Restart
Halts and reboots the nodes obeying **cstartup** dependencies

▲ The three-digit display shows various codes while the node goes through its start-up routine. When this completes without error the three-digit display remains black.

If the system stops longer than three minutes with a three-digit code displayed, refer to information on diagnosing boot problems in the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* and the *IBM RS/6000 Problem Solving Guide* to find out the meaning of the code and respond accordingly.

Shutting Down Nodes

The Shutdown and Cluster Shutdown options allow you to halt nodes. Be sure you have terminated all applications running on the node.

From the Node Front Panel:

PRESS Shutdown

▲ The Shutdown menu is displayed.

Note: If you do not have a switch installed on your system, Fence and Enable AutoJoin will not appear on the Shutdown Menu.

Figure 55. Shutdown Menu

PRESS Options:

Fence
Isolates node from High Performance Switch

Enable AutoJoin (can be selected only in conjunction with Fence)
Enables node to automatically rejoin the High Performance Switch when the node has restarted

Shutdown
Halts the nodes, ignoring **cshutdown** dependencies and turns off power

Cluster Shutdown
: Halts the nodes and turns off power, obeying **cshutdown** dependencies

Fencing Nodes from the Switch

To use the Node Front Panel Display to fence nodes from the switch, take the following steps:

1. Select Fence

   Note: If you do not have a switch installed on your system, the Fence option will not appear on the Node Front Panel Display.

2. Select Fence

Figure 56. Fence Menu

Unfencing Nodes from the Switch

To use the Node Front Panel Display to unfence nodes from the switch, select Unfence.

Note: If you do not have a switch installed on your system, the Unfence option will not appear on the Node Front Panel Display.

Changing the Key Mode Switch Position

The key mode switch establishes the initial program load (boot) path and has three possible positions:

Normal
: The most common setting used for attended operation.

Secure
: Used for unattended operation to prevent unauthorized reset. The reset function is not active when the key mode switch is in this position.

Service
: Used for attended operation when hardware and software service is conducted. The reset function is active in this position.

  Note: To successfully boot in service mode, you need to open a **tty** console connecting to the control workstation.

From the Node Front Panel:

CHECK
: The key mode switch buttons to verify the current setting

PRESS
: The appropriate button for the setting you want.

Opening a tty Connection

You can open a **tty** connection to display a system console for system debugging. This connection opens /**dev**/**tty0** on the node.

From the Node Front Panel:

PRESS Open TTY

▲ A system console window appears.

Note

When using this **tty** connection, the appropriate terminal type is **aixterm.**

This uses the RS-232 and supervisor bus connections and competes with node controller traffic. It is not appropriate for general node access. Limit its use to debugging purposes only.

If you use more than one **tty** connection per frame, the buffer may be overrun resulting in lost data.

Using Global Controls

The level (within a system partition or across system partitions) at which commands issued from the Global Controls panel operate is dependent on the View of the system you've selected.

You can issue a command to operate the controls on several nodes at once. You can turn the node power on or off, reset nodes, or change their key mode switch positions. You can select the nodes by frame or individually, on a node to node basis. Any switches occupying slots 1–16 in a frame will be colored gray. Use the All Node Summary ("Step 1: Display the All Node Summary" on page 322) to display changing LED status. Refer to "Node Front Panel Controls" on page 315, which describes how each of these options works, with the exception of Network Boot, or to "Monitoring and Controlling the Switch" on page 308 as appropriate.

You can issue commands from the Global Controls window to isolate nodes from the switch. Refer to "Fencing Nodes from the Switch" on page 322 for the steps to perform node isolation.

You can invoke a network boot from the Global Controls display panel. Network Boot causes booting over the SP Ethernet LAN, allowing installation, customization, or maintenance. You might find this useful after powering down a node, when the bootlist goes back to disk.

Refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* for more information.

Accessing the Global Controls Display:

Note

These choices can have wide ranging effects. Be cautious when using the global controls.

At the SP System Monitor GUI main window:

SELECT SP → Global Controls

▲ The Global Controls panel appears with all the slots deselected.

SELECT The slots to control

SELECT The action you want to perform

Choose one of the available actions such as On, Off, or Reset.

PRESS Do Command

▲ The SP System Monitor performs this action on all the selected slots.

Note: If you do not have a switch installed on your system, Fence and Unfence will not appear on the Global Controls Display menu.

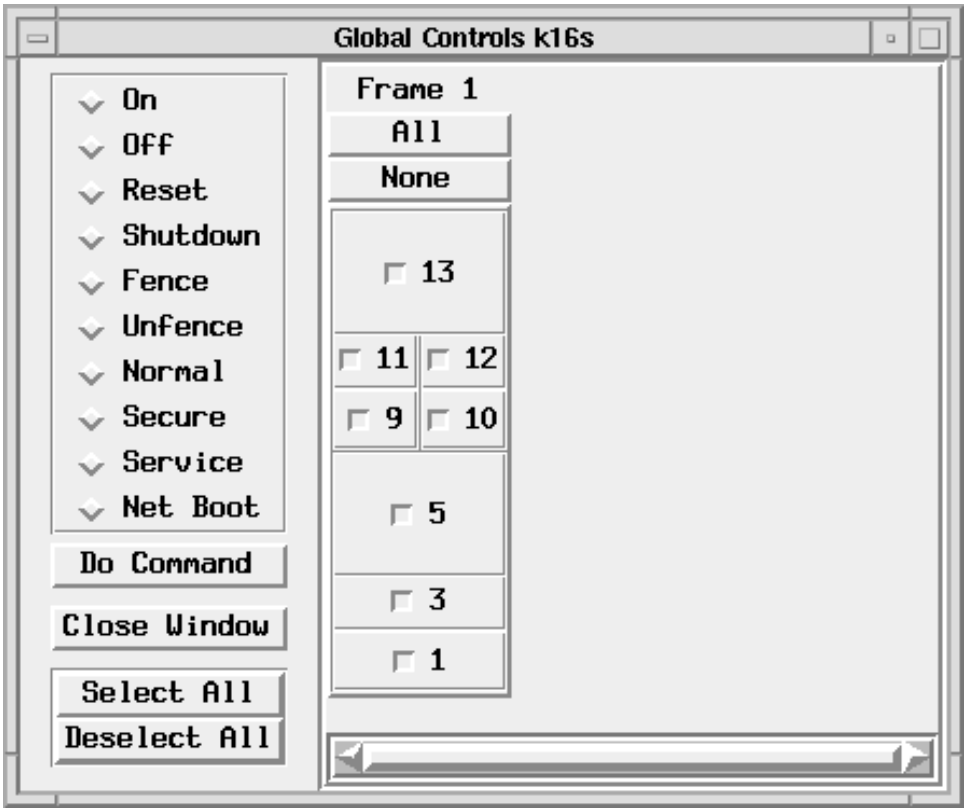


Figure 57. Global Controls Display

Frame 1 in this display contains two wide nodes, two 604 High Nodes, and four thin nodes.

Note: If a system is multipartitioned, a 'View' selection is available below the 'Deselect All' option; if a system is not partitioned, 'View' won't appear. If 'View' is present and selected, a new Global Controls Display is opened for the selected view.

Fencing Nodes from the Switch

To use the Global Controls Display to isolate nodes from the switch, take the following steps:

1. Select nodes
2. Select Off, Shutdown or Reset
3. Select Fence and, optionally, AutoJoin (Note that if you are doing Fence from Off, Shutdown, or Reset, and the Fence fails, the Off, Shutdown, or Reset will not execute.)
4. Select the type of Power Off, Reset, or Shutdown
5. If Enable AutoJoin was not selected, select Unfence to return nodes to the switch once nodes have restarted.

For additional information on isolating nodes from the switch, refer to Chapter 16, "Using a Switch" on page 221.

Determining If Nodes Are Operating: All Node Summary

With the SP System Monitor GUI you can monitor the system to determine, at a glance, if processor nodes are operating. Switches in slots 1–16 will also appear in an All Nodes Summary display, if you are in a Global view.

The All Node Summary provides you with a way to globally monitor the system and determine if nodes are operating. The View pulldown menu provides the option of viewing an individual system partition or viewing the system globally (all system partitions). From the All Node Summary window, you can select an SP System Monitor variable and display its values for nodes. Certain variables, such as *hostResponds*, the three-digit display, power, and environment lights, can provide you with system status.

Not all variables are valid for all node types. When a variable is not supported by the target node, that variable is blanked out. Refer to Appendix D, "System Monitor Variables" on page 549 to determine which variables your node type supports.

Step 1: Display the All Node Summary

At the SP System Monitor GUI main window:

SELECT SP → All Node Summary Display

▲ The selection window for node variables appears.

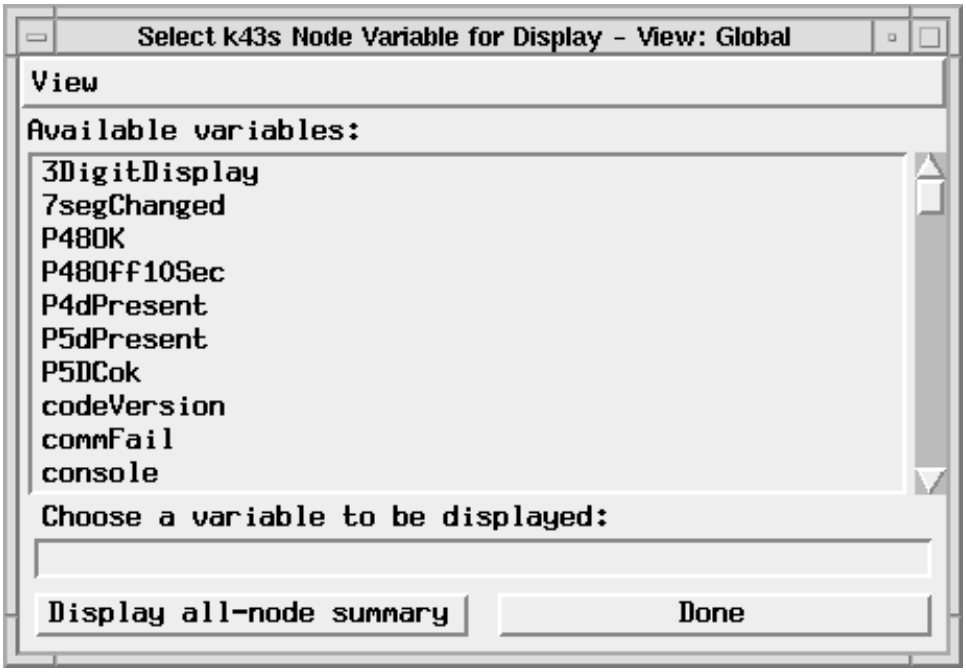


Figure 58. Node Variable Selection Window

This window contains:

- View pulldown menu that allows you to select an individual system partition or Global view of the system (in a multipartitioned system; if a system is not partitioned, 'View' won't appear on the menu bar)
- Scrolling list, in alphabetical order, of SP System Monitor variables that pertain to node hardware
- Entry field where you can type the name of a SP System Monitor variable
- Two push buttons: one to display the All Node Summary, one to close the window.

Step 2: Display Node Variables

At the Node Variables display:

SELECT 3DigitDisplay from the list of variables

PRESS Display all-node summary

▲ The three-digit display summary appears.

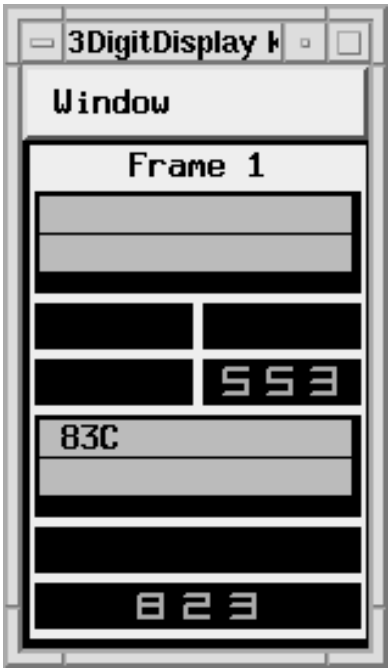


Figure 59. 3-Digit Display

This 3-Digit Display show a frame with four thin nodes, two wide nodes, and two 604 High Nodes. Each node shows the LED information. While thin and wide nodes support only 3-digit LED information, the 604 High Node supports a 16 character, 2 line LCD display.

Displaying hostResponds, powerLED, and envLED gives you an at-a-glance view of the status of your system. Use the scroll bar to find these other variables and repeat the selection process.

Every All Node Summary display has the same structure. The window title bar identifies the variable. The node format follows the physical organization of nodes within the frame. This means that, for a particular frame, the bottom left node in the window corresponds to the bottom left node within that frame.

Step 3: Check the System Status

Together, these summaries give you a logical view of the front panel indicators on all your nodes. Table 12 on page 325 explains how to interpret these summaries:

Note

The variable `console` has control functions. Clicking on this variable executes the function if you have serial link authority.

Table 12. Default Variable Colors

Variable Name	Front Panel Indicator	Color and Meaning
3DigitDisplay	Three-digit display	Any number: Can be an informational message, a warning, or failure. Check the codes in the *IBM RS/6000 Problem Solving Guide*. Pressing the left mouse button in a frame causes the slot numbers to appear in the node areas. Pressing the right mouse button in a frame causes the node number (relative to the whole SP system) to be displayed. These numbers are displayed only while the mouse button is pressed. Black: If nothing is displayed and hostResponds is green, this indicates the node is running (up)
hostResponds	hostResponds light	Green: The host node is responding. Red: The host node is not responding.
powerLED	Power light	Green: Power light is on. Node is running. Gray: Power light is flashing. Node is off but power is available. Black: Power light is off. Node is not running. Check the power switch on the front panel.
envLED	Environment light	Black: Environment light is off. No indication of failure. Gray: Environment light is flashing a failure. Yellow: Environment light is on. Indicates an environment warning.
Key Mode Switch	Key position	Green: normal Yellow: secure Red: service

Determining the State of Network Adapters within the SP System

The Topology windows provide you with a way to globally determine which network adapters are configured for TCP/IP within each SP processor node and determine if the adapters are operating. The window allows you to select a variable associated with each configured LAN in the SP system and display its values for all the processor nodes.

Not all LANs can be configured for all processor nodes. When a LAN is not configured within a target node, that variable is blanked out. Determination of which LANs are configured for each target node is based on network adapter information contained within the System Data Repository combined with communications with the SNMP (Simple Network Management Protocol) agent daemons running in each processor node. Until communication with an agent daemon in a target node is completed, the state of the variable is reflected as unknown for the target node.

Step 1: Display the Topology

At the SP System Monitor GUI main window:

SELECT SP → Topology

▲ The Topology display appears.

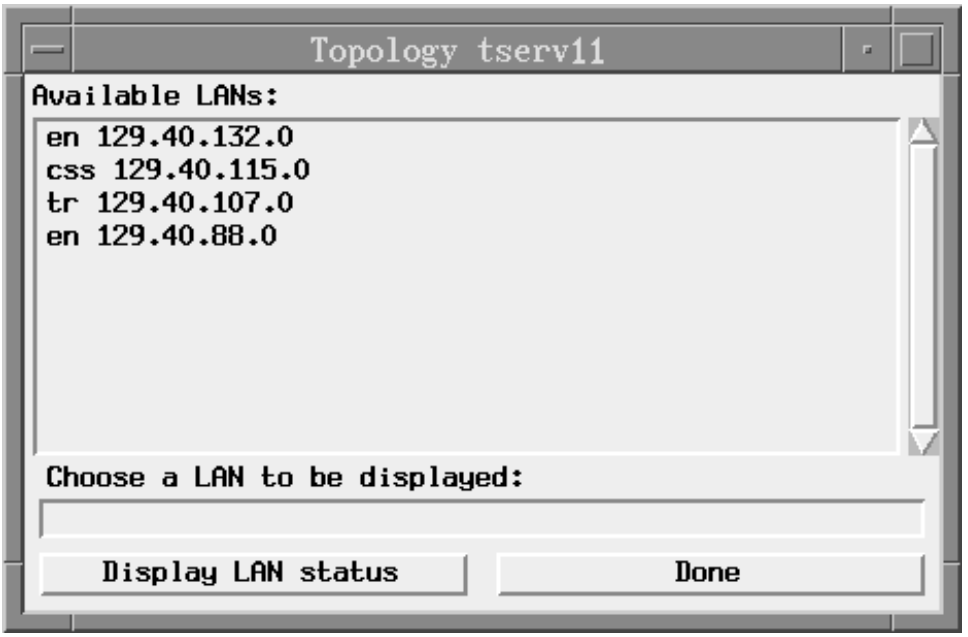


Figure 60. LAN Topology Display

This window contains:

- A scrolling list of SP System Monitor LAN variables that pertain to configured LANs.
- The entry field where you can type the name of a SP System Monitor LAN variable
- Two push buttons: one to display the LAN status, one to close the window.

Step 2: Display the LAN Status

At the Topology display:

SELECT A LAN ID

PRESS Display LAN Status

▲ The LAN Status display summary appears.

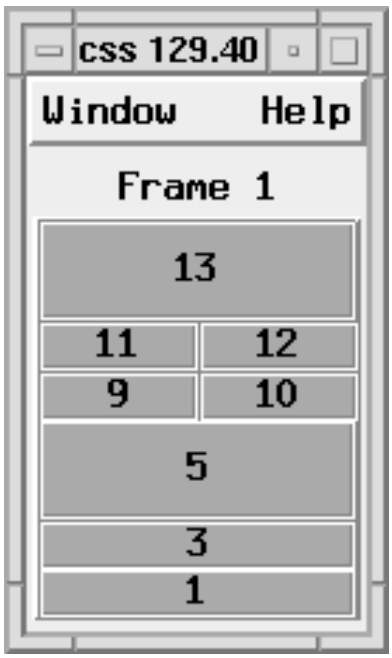


Figure 61. LAN Status Display

This display shows that wide nodes residing in slots 1 and 3, the 604 High Nodes residing in slots 5 and 13, and the thin nodes residing in slots 9, 10, 11, and 12 are configured for the chosen LAN.

Graphical Element Colors

The Topology window displays use color to show the state of adapters and communications with the SNMP agents running in the processor nodes. The colors are orange, yellow, red, and green.

Table 13. Graphical Element Colors

Color	Adapter Status	Meaning
Orange	Initial status	Communication with the SNMP agent running in the target node has not been successful.
Yellow	Adapter exists: state unknown	Communication with the SNMP agent indicates an adapter connecting to the LAN being displayed exists within the target node. The state of the adapter is being determined or is not yet defined on the processor node.
Red	Adapter is not operational	Communication with the SNMP agent running in the target node indicates the adapter is configured within the target node but is not operational.
Green	Adapter is operational	Communication with the SNMP agent running in the target node indicates the adapter is configured within the target node and is operational.

Note: A pulldown help menu that contains a summary of the colors and their meaning is available within each Topology window.

Communication with the SNMP agent running in each processor node occurs serially. After communication with all of the processors completes, it is reinitiated after a delay of 1 minute. When communication with the agent running within a target node is unsuccessful, the target node is placed on an “unresponsive list”. Communication with each node on the “unresponsive” list is attempted every 5 minutes.

LAN configuration changes occurring while the LAN window is open is not reflected until Topology is reentered from the main window.

Setting Up the Logging Daemon on Separate Workstations

Step 1: Install the **ssp.clients** option on your workstation. Refer to the appropriate steps in the chapter on installing and configuring the SP system in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

Step 2: Edit the /**spdata**/**sys1**/**spmon**/**hwevents** file to configure the logging options you want. Refer to **splogd** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

Step 3: If you have set up **hwevents** to do error logging (SP_ERROR_LOG keyword), run /**usr**/**lpp**/**ssp**/**bin**/**setup_logd**. (Refer to **setup_logd** in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.) If you are installing the logging daemon only to call your own executable when a state change is reported, just add an entry to **inittab** or start **splogd** in the background:

```
splogd -b
```

Step 4: If you have run /**usr**/**lpp**/**ssp**/**bin**/**setup_logd** or added **splogd** to the **inittab** file in Step 3, then you must add the following to the /**etc**/**environment** file:

```
SP_NAME=control_workstation
```

where *control_workstation* is the hostname of the control workstation. Otherwise, the environment variable SP_NAME must be defined to the hostname of the control workstation prior to executing **splogd**.

Step 5: For each frame in the SP system, add the following line to the /**spdata**/**sys1**/**spmon**/**hmacls** file on the control workstation:

frame_number **hardmon.***workstation_name* **m**

where *frame_number* is the number of the frame and *workstation_name* is the short hostname, as returned by **hostname -s**, of the workstation where you have installed the logging daemon.

Setting Up the Hardware Monitor Clients on Separate Workstations

Step 1: Install the **ssp.clients** option on your workstation. Refer to the appropriate steps in the chapter on installing and configuring the SP system in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

Step 2: Prior to executing any Hardware Monitor client command, set the environment variable SP_NAME to the hostname of the control workstation. Before using any command you must be authenticated to use the Hardware Monitor. Use

the **kinit** command to login to the SP authentication system. If your site uses an alternative to the SP authentication services, you must use authentication procedures appropriate to such an alternative.

Availability and Problem Management

© Copyright IBM Corp. 1995, 1997

Chapter 22. Managing a High Availability Control Workstation Configuration

Notice

The High Availability Control Workstation hardware and software functions are separately priced features.

This chapter provides overview and task information for the High Availability Control Workstation (HACWS) function. The HACWS function provides for a backup control workstation for your SP system.

We presume that you have planned for, installed, and configured the HACWS function.

For information on planning for HACWS function, refer to *IBM RS/6000 SP: Planning, Volume 1, Hardware and Physical Environment* and *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment*.

We also suggest that you refer to the documentation for AIX High Availability Cluster Multi-Processing/6000 (HACMP/6000). HACMP/6000 is the foundation for HACWS function.

Information is provided in this chapter for the following tasks:

- Understanding HACWS function
- Modifying typical administrative tasks (maintaining the /**etc**/**passwd** database, setting up mail spools, setting up file collections, for example) so that subsystems will fail over to the backup control workstation in the event the primary control workstation becomes unavailable. We'll cover this information in the order of the related chapters in the *IBM Parallel System Support Programs for AIX: Administration Guide*:
  - Starting Up and Shutting Down the SP System
  - Parallel Management Commands
  - Managing File Collections
  - Managing User Accounts
  - Managing SP Resources
  - Managing Time Synchronization
  - Managing an Automounter
  - Managing Mail Service
  - Accounting
  - Security Features of the SP System
  - Sysctl
  - Using a Switch
  - Managing System Partitions
  - Managing Extension Nodes
  - Using the SP System Monitor or SP Perspectives
- Customizing HACMP/6000
- Verifying Frame Supervisor Cables

© Copyright IBM Corp. 1995, 1997

Understanding HACWS Function

The following definitions will help in understanding HACWS function.

Primary control workstation
The first control workstation that is installed with the SP system. The primary control workstation is a physical machine.

Backup control workstation
The standby backup control workstation that is installed during the HACWS installation. The backup control workstation is a physical machine.

Active control workstation
The control workstation that is currently running the SP control workstation applications (the SDR and **hardmon**, for example). This can be either the primary or backup control workstation.

Inactive control workstation
The control workstation that is not running the SP control workstation applications. This can be either the primary or backup control workstation.

The HACWS function improves the reliability, availability and serviceability of an SP system by providing for a backup control workstation. HACWS provides for extending the SP system configuration to include a second RS/6000 to function as a backup to the primary control workstation. Automatic failover and reintegration to the backup control workstation are provided should the primary control workstation fail or in the case of taking the primary control workstation down to perform hardware and software maintenance.

HACWS provides hardware and software features that remove a single point of failure for the SP system. HACWS does not, however, protect against double failures. You should not expect control workstation functions to work during the time that a control workstation is failing over. If a failure on the control workstation occurs while a control workstation function is being performed, the function needs to be restarted.

The following figure illustrates a typical HACWS configuration:

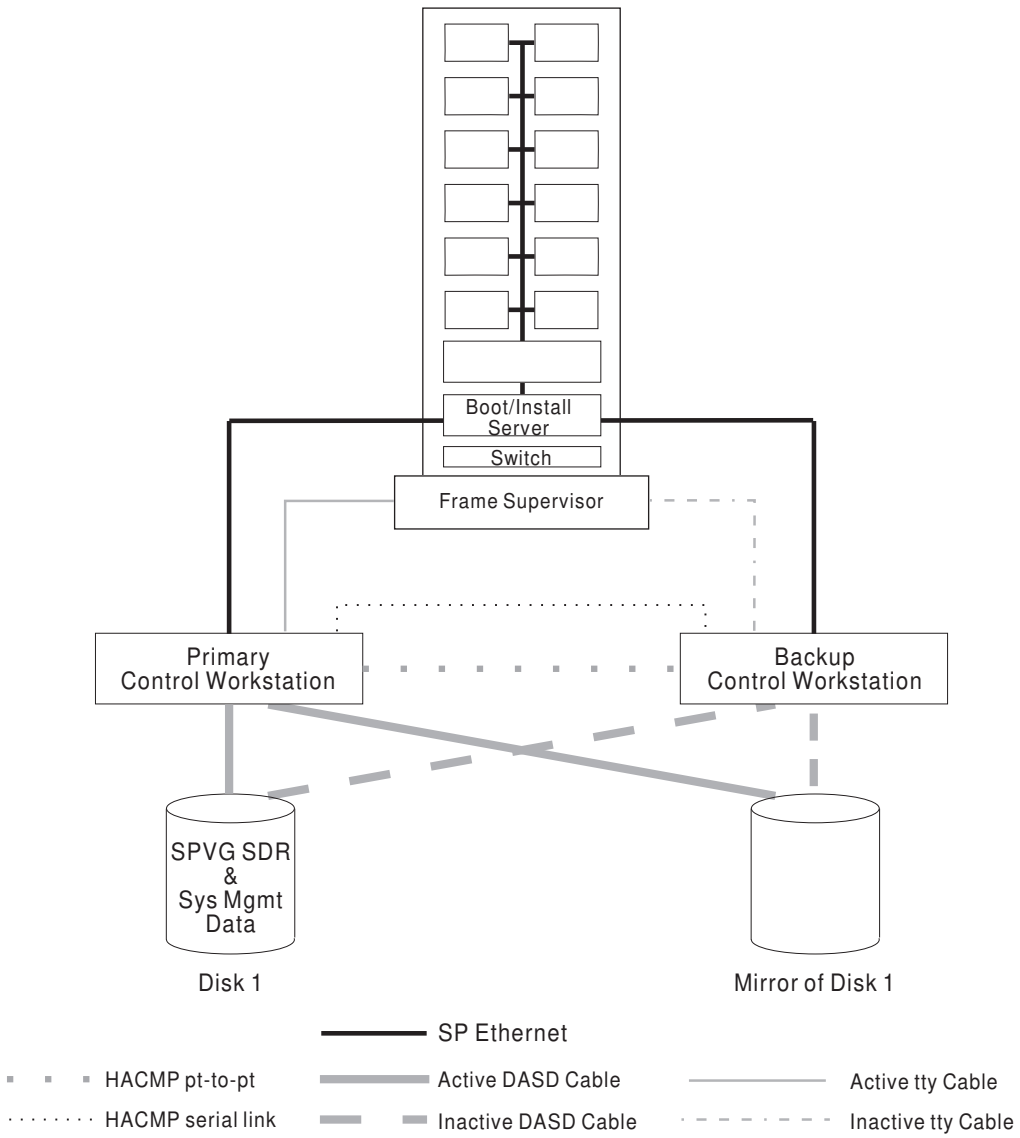


Figure 62. An HACWS Configuration

Hardware

HACWS is a two-node configuration: a primary control workstation with a backup control workstation. (Don't confuse the use of *node* in this context with an SP system node.) We highly recommend that the two RS/6000s be configured in exactly the same way with the same model hardware and I/O configurations. This is not required, but simplifies the management of the HACWS configuration. We also highly recommend that the primary and backup control workstations do not use the same power source.

External Fixed Disks and Disk Controllers

External fixed disks that provide only nonconcurrent access are used in an HACWS configuration. We strongly recommend that the external fixed disks be mirrored across two disk controllers. If the external fixed disks are not mirrored, the single point of failure has not been removed from the SP system. The fixed disk becomes the single point of failure: if the disk fails the SP system will be down until the disk

can be replaced. We also recommend that the two disk controllers be on different power sources.

Frame Supervisor Cards

HACWS requires a dual RS-232 frame supervisor card per frame with a connection from each control workstation to each SP frame. Hardware feature #1245 provides this new type of frame supervisor card with an RS-232 Y-cable that connects from a frame in the SP system to the control workstation. All frames in the SP system need this new frame supervisor card and the Y-cables in order for the HACWS software to be configured. The control workstation must run AIX 4.1 to support this new frame supervisor card. The frame supervisor connections from each frame must be connected to the exact same tty ports on both control workstations. If not cabled correctly, frame supervisor connections will not be activated by the hardware monitor.

Network Connections

One of the following network connections must exist between the primary and backup control workstations :

- A dedicated TCP/IP network link
- An RS-232 tty link
- The target mode SCSI across the external SCSI fixed disks

Adapters

Each control workstation requires the same number of connections to the SP Ethernet on the same LAN segments. Each SP Ethernet LAN segment must be cabled to the same Ethernet (**en***x*) adapter. Standby network adapters are optional in an HACWS configuration because there may not be enough adapter slots in the control workstation for a standby adapter on each service network. On a small one- or two-frame SP system the normally used control workstation has only two slots and does not have enough slots for standby adapters. In a large multi-frame SP system the number of slots required for frame supervisor and SP Ethernet LAN segments may not allow for standby adapters. However, the presence of a standby adapter may avoid the need to fail over to the inactive control workstation.

Software

The software foundation for HACWS is the separately orderable AIX High Availability Cluster Multi-Processing/6000 (HACMP/6000). Only the high availability subsystem is required. The concurrent Resource Manager of HACMP/6000 is not supported (all the SP data on the control workstation is in an AIX Journaled File System and, therefore, does not allow concurrent access).

You may use any level of HACMP/6000 4.2 or greater that is supported with the level of AIX that you are using. Refer to the appropriate HACMP/6000 documentation to determine what levels of HACMP are supported with the level of AIX that you are using.

Note: In an HACWS configuration, the nodes may be running any level of AIX and PSSP.

The primary and backup control workstation are configured in a two-node rotating configuration. The external fixed disks are configured in nonconcurrent access only. For information on HACMP/6000, refer to *HACMP/6000: Concepts and Facilities*.

Building on HACMP/6000, the **ssp.hacws** optional installation package that is part of the IBM Parallel System Support Programs for AIX provides:

- Scripts to create a backup control workstation
- Verification programs
- HACMP/6000 pre- and post- event scripts
- Scripts to synchronize the backup and primary control workstations

When a failure of the primary control workstation takes place there is a disruptive failover (that is, the failover is noticeable) of the control workstation to the backup control workstation. This failover:

1. Switches the external fixed disks
2. Reconfigures network adapters from boot addresses to service addresses to perform IP address takeover
3. Performs hardware address takeover
4. Remounts file systems
5. Restarts the control workstation applications
6. Resumes hardware monitoring
7. Allows clients to reconnect to obtain services or update control workstation data

The backup control workstation assumes the IP address (service address) and IP aliases of the primary control workstation, resulting in only one active control workstation at a time and allowing client applications to run without changes. See Figure 63 on page 338 for an illustration of addressing in an HACWS configuration.

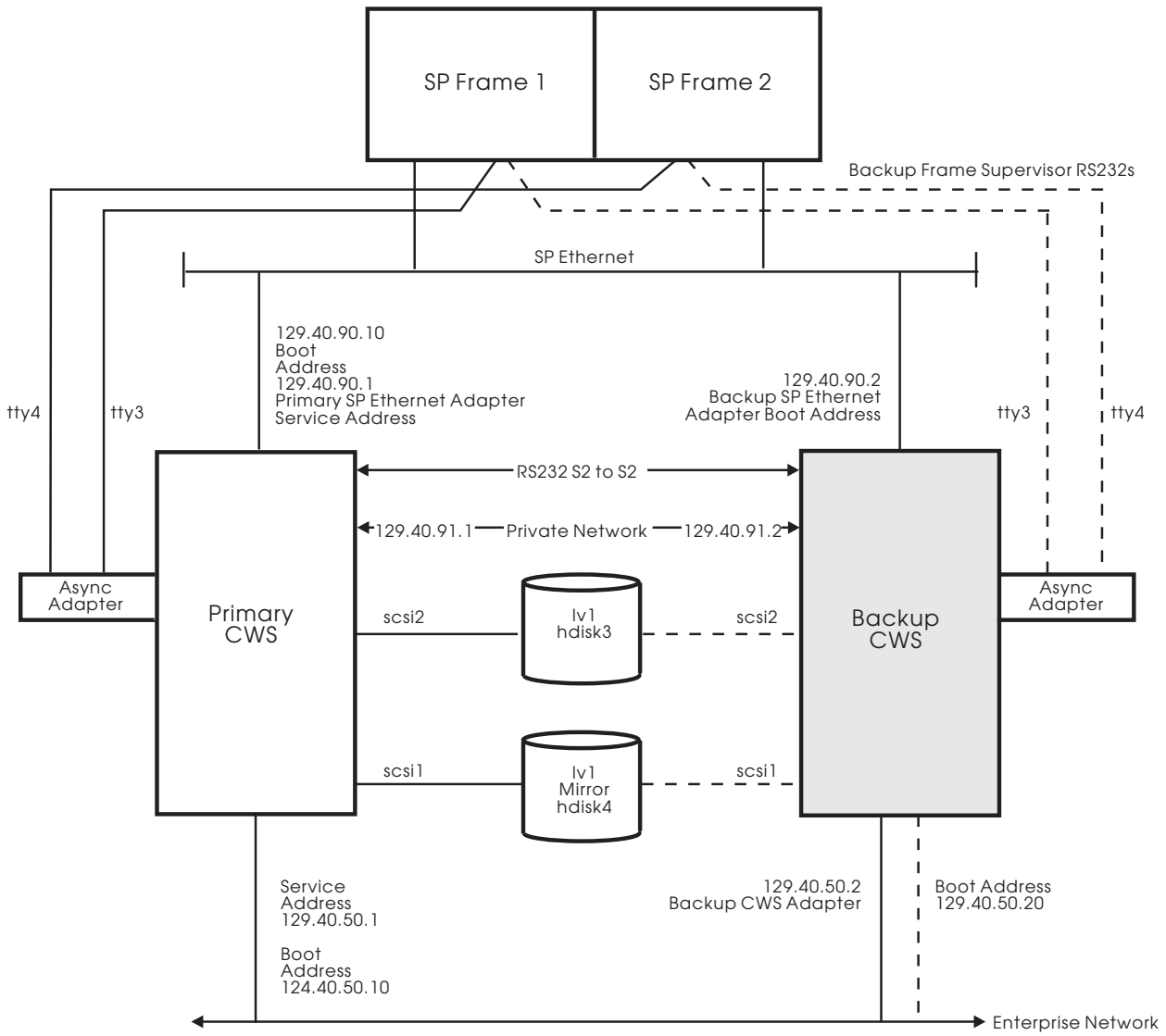


Figure 63. Addressing in an HACWS Configuration

Operating HACMP/6000 and the Control Workstation

This section describes some of the operational characteristics of the control workstation and HACMP/6000.

Booting the Control Workstations

In a two-node rotating HACMP/6000 configuration, if you are booting both the primary and backup control workstations at the same time, the first control workstation to have the cluster manager started will acquire the shared resources. In an HACWS scenario this will cause the control workstation IP addresses (service address) and the /**spdata**/**sys1** file system to be configured on that control workstation. It is important that the RS/6000 that is intended to be the active control workstation have the HACMP/6000 cluster manager started on it first. To help ensure that the correct control workstation acquires the shared resources we recommend that the cluster manager be started manually instead of being defined to start at system restart. This allows both control workstations to be booted in any order and any potential race conditions can be avoided during the boot process.

After both control workstations are booted the cluster manager can be started on the intended control workstation.

If the control workstation function is started on the unintended control workstation, you can perform the following sequence of events to make the intended control workstation active:

1. Boot the intended control workstation without starting the cluster manager.
2. Do a graceful shutdown without takeover on the incorrectly active control workstation.

   Using SMIT:

TYPE	**smit clstop**
	▲The Stop Cluster Services menu is displayed.
SELECT	now (for Stop now option)
SELECT	graceful (for Shutdown mode option)
HIT	Enter (to Do)

   Using the Command Line:

   Enter

   ```
   /usr/sbin/cluster/utilities/clstop -y -N -g
   ```

   Note: **Never** use **kill** to stop the HACMP/6000 cluster manager. Doing so will cause the node to fail.
3. Start the cluster manager on the intended control workstation.

   Using SMIT:

TYPE	**smit clstart**
	▲The Start Cluster Services menu is displayed.
SELECT	now (for Start now option)
SELECT	true (for Startup Cluster Information Daemon? option)
	Note: Don't use the **on system restart** option.
HIT	Enter (to Do)
4. Execute the **clstat** utility to make sure HACMP/6000 is up and running on the correct control workstation.
5. Start the cluster manager on the inactive control workstation.

Avoiding False Adapter Failures on Configurations Without Standby Adapters

If a control workstation is configured without a standby adapter on one of its service networks, and the other control workstation is powered off, the HACMP software may find it difficult to determine service adapter failure. This is because the HACMP cluster manager cannot use a standby adapter to force packet traffic over the service adapter to verify its operation. This shortcoming is less of an exposure if one or more of the following is true:

1. There are network devices which answer broadcast ICMP ECHO requests. This can be verified by pinging the broadcast address and determining the number of different IP addresses that respond.

2. The service adapter is under heavy use. In this instance the inbound packet count will continue to increase over the service adapter without stimulation from the cluster manager.

If neither of these cases is true, then HACMP may report that a service adapter has failed, even though there has not been a failure; it is just that there are not any other adapters to communicate with. (For example, this can occur when all the nodes are powered off.)

An enhancement to **netmon**, the network monitor portion of the HACMP cluster manager, is described below. It can be configured to allow more accurate determination of a service adapter failure. This function can be used in configurations that require a single service adapter per network.

A **netmon** configuration file, **/usr/sbin/cluster/netmon.cf**, will specify additional network addresses to which ICMP ECHO requests can be sent. The configuration file consists of one IP address or IP label per line. The maximum number of addresses used is five. All addresses specified after the fifth will be ignored. No comments are allowed in the file.

The following is an example of a **netmon.cf** configuration file:

```
180.146.181.119
steamer
chowder
180.146.181.121
mussel
```

This file must exist at cluster startup. The cluster software will scan the configuration file during its initialization phase. When **netmon** needs to stimulate the network to verify adapter function, it will send ICMP ECHO requests to each address. After sending the request to every address, **netmon** will check the inbound packet count before determining whether there is an adapter failure.

Conditions that Cause a Failover of the Control Workstation

The following conditions will cause a failover of a control workstation:

1. An intentional failover

   This occurs when the operator stops the cluster manager on a node with a shutdown mode of **graceful with takeover**.

2. A failure of the operating system on the active control workstation

   If the AIX operating system fails, the cluster manager on the inactive control workstation will detect this via missed heartbeats. The cluster manager then declares the crashed control workstation dead and performs a takeover of the shared resources.

3. A failure of an external disk adapter that is not backed up

   If HACWS is configured with only one external disk adapter per control workstation, you need to write an Error Notification Object for the failure of the disk adapter to do a graceful shutdown with takeover. The other control workstation will then take over the shared resources.

4. A service network failure occurs on the active control workstation and the inactive control workstation is available to take over

A failure of a service adapter on the active control workstation, without a standby adapter to take over, will cause some SP node subsystems to perceive that the control workstation has failed and the system will quickly become unstable. To prevent such system instability, a service network failure on the active control workstation is promoted to a node failure, causing a failover to the inactive control workstation.

Failures that Do Not Cause a Failover of the Control Workstation

The following conditions will not cause a failover of a control workstation:

1. A hung control workstation with network adapters that still heartbeat

   In the case where the application software hangs and the network adapters being monitored via HACMP/6000 still respond to heartbeats from the inactive control workstation, HACMP/6000 will not detect the hang and will not fail over.

2. A service network failure occurs on the active control workstation and the inactive control workstation is not available to take over

   If the inactive control workstation is not available to take over, then a service network failure on the active control workstation will not be promoted to a node failure because no failover can occur.

General Guidelines for When to Do Configuration Changes

We recommend that you do configuration changes on the SP system only when the primary control workstation is the active control workstation. Most configuration changes are allowed when the backup control workstation is the active control workstation. However, if changes are made when the backup control workstation is active, some configuration files will need to be updated on the primary control workstation. When changes are made on the primary control workstation, some configuration files will need to be updated on the backup control workstation. If configuration changes are done on the backup control workstation, the updates go the opposite direction (backup to primary instead of primary to backup). It is easier to manage the system when the updates go only in one direction.

You can decide to what extent you do configuration changes on the backup control workstation. Because only two machines are involved, the bidirectional updates may not be a problem.

The following table lists tasks that may be performed in an HACWS configuration and whether or not it is possible to update configuration files on the active backup control workstation. A 'No' under the 'Backup Control Workstation Active' column indicates that the file cannot be updated on the active backup control workstation. In such a case, you will have to update the file on the primary control workstation when it is again active.

Table 14 (Page 1 of 2). High Availability Control Workstation Configuration and Task Summary

Task	Primary Control Workstation Active	Backup Control Workstation Active
Update passwords	Yes	No
Add or change users	Yes	No
Change Kerberos keys	Yes	No

Table 14 (Page 2 of 2). High Availability Control Workstation Configuration and Task Summary

Task	Primary Control Workstation Active	Backup Control Workstation Active
Install a node	Yes	Yes
Change or add system partitions	Yes	Yes
Add nodes to the system	Yes	No
Hardware monitoring	Yes	Yes
Reboot nodes	Yes	Yes
Run diagnostics	Yes	Yes
Shut down and restart system	Yes	Yes
Run parallel jobs	Yes	Yes
Update file collections	Yes	Yes
Accounting	Yes	Yes
Change site environment information	Yes	No

Managing SP Subsystems for High Availability Control Workstation Function

In this section we present information on the administrative tasks you must perform to ensure that the various SP subsystems will fail over to the backup control workstation.

Starting Up and Shutting Down the SP System

The sequence files need to be kept in sync between the primary and backup control workstations in order for the **cshutdown** command to work on the backup control workstation. The following files need to be synchronized:

- **/etc/cstartSeq**
- **/etc/shutSeq**
- **/etc/subsysSeq**

Parallel Management Commands

You can use parallel management commands with the backup control workstation as a target, even when the backup control workstation is inactive, as long as the host name of the backup control workstation is listed in the working collective or host list.

Managing File Collections

The inactive control workstation acts as a file collection client (**supper** is run via **cron**). This keeps all the file collection files and other default file collections in sync between the primary and backup control workstations. If a configuration or file collection master change is done on the active control workstation, you should perform a **supper update** for the particular file collection. Note that you can do a **supper update** for all the file collections. If the update is not performed after the change and a failover occurs, the updates will not be on the new active control workstation and such updates will not be available to the SP nodes.

Managing User Accounts

We highly recommend that HACWS configurations use NIS for user management when large numbers of users need to be managed on the SP system. NIS enforces a master-slave relationship for password databases and is easier to manage than file collections for user management in an HACWS configuration.

Using NIS with HACWS

If you are using NIS in an HACWS configuration and the NIS master database is the primary or backup control workstation, there must be at least one slave database within the NIS domain. For high availability of an SP system, we recommend that all the nodes where users log in be NIS slave database nodes.

Using File Collections with HACWS

If you use file collections for user management, the **/etc/passwd** and associated files should be on the primary control workstation. Once a failover occurs and the backup control workstation is the active control workstation, the actual master **passwd** files are no longer available to the backup control workstation. For this reason changing your passwords, gecos information, and login shell is not allowed when the backup control workstation is active. (This is similar to the restriction with NIS when the NIS master is down.) The backup control workstation, when inactive, will act as a file collection client (similar to an SP node) and update its **passwd** files from the primary control workstation when it is active.

Controlling User Login

If you are using login control on your SP system, the **spacs_data** must be kept in sync between the primary and backup control workstations. The **block_usr_sample** script and any derivatives of it need to be kept in sync between the primary and backup control workstations.

Managing SP Resources

When a failover or reintegration occurs, you need to make a decision about the currently running parallel jobs on the system. You can either kill the running jobs (**suspend** won't work) and restart the Resource Manager or wait until all the running jobs complete and restart the Resource Manager. New parallel jobs cannot start until the Resource Manager is restarted.

Note: This restriction will be removed in a future release of the PSSP.

Managing Time Synchronization

Both control workstations must use the same source for time (for example, timemaster, Internet, or consensus). NTP should be running on both the inactive and active control workstations.

Managing an Automounter

Include the backup control workstation in the distribution of automounter configuration and map files.

Managing Mail Service

If the primary control workstation is the mail hub, you should use NFS to export the **/var/spool/mail** file system to the backup control workstation so that mail files for nodes are put in the same file system. When a failover occurs you should have the **/var/spool/mail** locally mounted on the backup control workstation and the primary control workstation. Use NFS to mount the file on the control workstation when it is the inactive control workstation. The **/var/spool/mail** file system must be an externally mounted file system.

Accounting

If accounting was enabled in the Site Environment Information SMIT panel and the node number **0** was selected as the Accounting Master, you need to make a directory change. Create a symbolic link from both the primary and backup control workstations to put the **cacct** directory in the **/spdata/sys1** file system.

Do the following on both the primary and backup control workstations after accounting has been enabled:

1. Enter:

   ```
   ln -s /spdata/sys1/cacct /var/adm/cacct
   ```

   This creates a symbolic link between the **cacct** directory so that the data in the **rootvg** is automatically copied into the external file system when **crunacct** is run via **cron**. This allows accounting reports over a period of time to have no indication of which control workstation was active.

   On an inactive control workstation the **crunacct** and **cmonacct** scripts will fail because the symbolic link for the **cacct** directory on the inactive control workstation will not have a target directory.

2. To keep those scripts from failing, change the invocation in the **crontab** file to run a wrapper shell script that determines if the control workstation is the inactive control workstation. If the control workstation is the inactive control workstation, the script should exit without running the **cmonacct** or **crunacct** scripts. The **lshacws** command should be used for this. Alternatively, you may choose to let the script fail. The following is a sample script:

   ```
   #/bin/ksh
   # This is the state of the control workstation taking
   # the string output of the lshacws command.
   # It takes the name either crunacct or cmonacct
   # as an argument.

   HACWS_STATE=‘lshacws‘

   # Check to see if we are on an INACTIVE control workstation.
   # If we are, exit quietly, if not, run the argument passed.

   if [[ $HACWS_STATE = 1  ||  $HACWS_STATE = 16 ]] ; then
           exit 0
   else

   # Go run the command
   $1
   fi
   exit 0
   ```

Security Features of the SP System

In an HACWS configuration the primary authentication service is **not** duplicated on the backup control workstation. If the primary control workstation is either a primary or secondary authentication server, then the backup control workstation must be a secondary authentication server. The backup control workstation cannot be a primary authentication server.

When the primary control workstation is down, and it is the primary authentication server, no updates to the authentication database can be made.

Kerberos Configuration Files

The backup control workstation needs to be added to the **.klogin** files on all the nodes and the control workstation. This **.klogin** file also needs to be kept in sync on the backup control workstation.

When control workstation services move back and forth between the two control workstations, the Kerberos service keys need to remain the same. To accomplish this, the following commands were issued on the backup control workstation when HACWS was initially configured:

```
$ /usr/lpp/ssp/rcmd/bin/rcp -p primary_name:/etc/krb-srvtab /etc/krb-srvtab.primary
$ cp -p /etc/krb-srvtab /etc/krb-srvtab.backup
$ cat /etc/krb-srvtab.primary >>/etc/krb-srvtab
```

Repeat this procedure whenever you change Kerberos service keys on either of the two control workstations.

Sysctl

When Sysctl is installed the configuration file /**etc**/**sysctl.conf** is created and updated on the backup control workstation. If you update this file or use a different configuration file, the files needs to be kept in sync between the primary and backup control workstation.

The Sysctl ACL files must also be kept in sync between the primary and backup control workstations. This file is shipped as /**etc**/**sysctl.acl**. If you add an ACL record to the /**etc**/**sysctl.acl** file, the file should be kept in sync between the primary and backup control workstations.

When the backup control workstation is the inactive control workstation, Sysctl commands can be used because Sysctl is installed and running on the inactive control workstation.

Using a Switch

The tasks you should perform to be sure the switch is connected to the backup control workstation when there is a failover depend on the type switch you have on your SP system.

The High Performance Switch

When a High Performance Switch fault occurs during a control workstation failover you should check to see if connectivity across the High Performance Switch is still working. If switch connectivity is not working, you should run **rc.switch** on all the SP nodes. When this double failure occurs (a very low probability) there may be updates to the control workstation for the switch subsystem that are not made. The switch subsystem will stop working until an **Estart** is invoked from the primary

switch node. There are instances when, depending on exactly when the control workstation initially failed and the backup took over as the control workstation, an **Estart** will not start the switch subsystem correctly. In such a case **rc.switch** needs to be run on all the SP nodes. You can try an **Estart** first and if it is not successful, run **rc.switch** on every SP node.

We recommend that you use an Error Notification Object for switch faults and control workstation failures to send mail or a message to alert you about such double failures. You can then determine if the switch subsystem is working correctly. Because the timeframe for this problem is very small, we do not recommend that you run **rc.switch** or **Estart** whenever a failover occurs.

The SP Switch

When an SP Switch event occurs during a control workstation failover, the communication subsystem continues to handle switch faults, primary node failures, and primary backup node failures. System data about the switch is inaccessible during the control workstation failover and can become out-of-date. When the backup control workstation assumes the duties of the active control workstation, it must determine if a double failure has occurred (a switch event that occurred during a control workstation failover).

Perform the following to determine if a switch event has occurred and to resynchronize the communication subsystem with the system data on the new control workstation:

1. To determine if a primary node takeover (failure) has occurred, check all nodes in each system partition for a file named **act.top.1** in the **/var/adm/SPlogs/css** directory that contain a timestamp during the control workstation failover. If you locate an **act.top.1** file containing a timestamp during the control workstation failover:
   a. Run **/usr/lpp/ssp/css/rc.switch** on all the nodes in each system partition that has a node with an **act.top.1** file.
   b. Run **Estart** for each system partition that has a node with an **act.top.1** file.
2. To determine if a primary backup node failed or a switch fault occurred, examine the **flt** file in the **/var/adm/SPlogs/css** directory on the primary node for entries with a timestamp during the control workstation failover. If you locate entries that contain a timestamp during the control workstation failover, run **Estart** for each system partition having entries on its primary node.

   We recommend that you use an Error Notification Object for switch faults, primary node takeovers, and control workstation failures to send mail or a message to alert you about such double failures. You can then examine the communication subsystem as described previously. Because the timeframe and probability for these double failures are very small, we do not recommend that you run **rc.switch** or **Estart** whenever a control workstation failover occurs.

Managing System Partitions

System partition configuration changes must be done from the active control workstation. Such changes may be done on either the primary or backup control workstation, depending on which one is active. We recommend that configuration changes be done from the primary control workstation, but it is not required.

When you configure HACWS you are instructed to put any IP address aliases that are required for partitioning or any other reason in a script called **/spdata/sys1/hacws/rc.syspar_aliases**. This script will be invoked by the HACWS event scripts at the correct times to enable system partitioning.

Managing Extension Nodes

The SP SNMP Manager that runs on the control workstation transfers dependent node configuration information to the SNMP Agent running on a dependent node. This SP SNMP Manager support will transfer to the active control workstation.

Using the SP System Monitor or SP Perspectives

If you have installed **ssp.clients** or **ssp.gui** on a RS/6000 other than the primary or backup control workstations, you need to give special consideration to monitoring the SP system from other than the active control workstation. If you execute the **spmon** or **perspectives** commands without logging into the active control workstation and the active control workstation does a failover, this will not be represented on the SP System Monitor or SP Perspectives GUI until you take another action. Once you take another action, the **spmon** or **perspectives** commands will exit. You will have to reexecute the **spmon** or **perspectives** commands and reconnect to the currently active control workstation. This occurs because when you are monitoring a TCP/IP connection and the host fails, the client side of that connection does not know about the failure of the connection until a packet is sent from the client side of the connection. Since the users of the **spmon** or **perspectives** commands who are off the control workstation should be well known, an Error Notification Object can be set up to send mail or a message to the list of well known users to inform them that the control workstation has performed a failover. Such a notification list for the users that monitor the SP system is a good idea in any event. After a failover of a control workstation (or any other major recovery action) an administrator should be notified. Several options for such notification exist:

- Error Notification Objects
- NetView for AIX control desk alerts
- Trouble Ticket for AIX problem incidents
- HACMP Notify script

Viewing which Control Workstation is Active

You can view which RS-232 tail is active in the Frame Environment Layout display of the SP System Monitor. There is an indicator for activity on both RS-232 connections from the frame and an indicator for which tail of the frame supervisor connection is active.

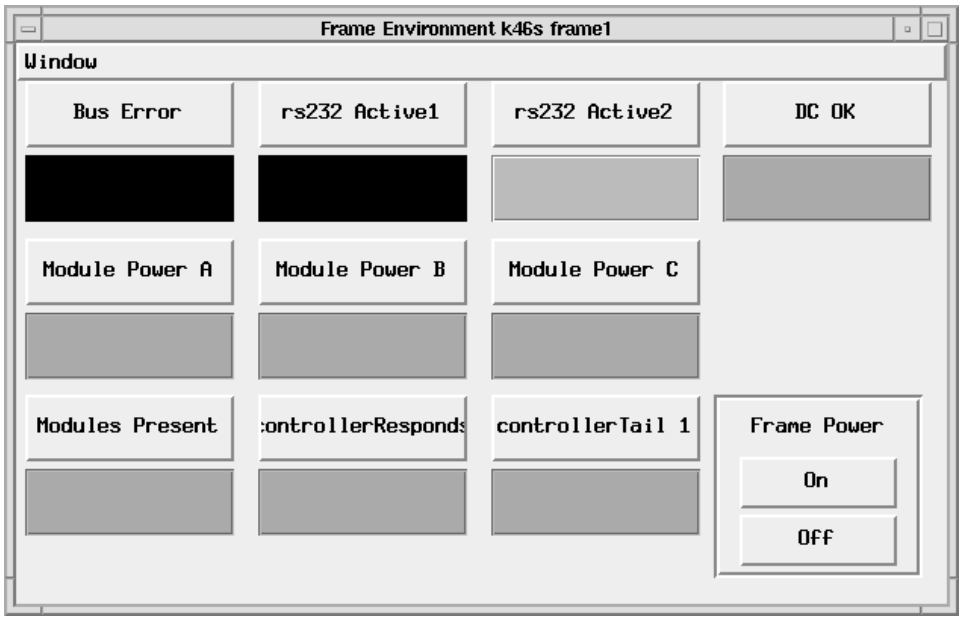


Figure 64. Frame Environment Layout Showing Active Control Workstation

rs232 Active1	Status of the RS-232 Active1 communication with the control workstation. Yellow (RS-232 activity) or black (no RS-232 activity) or gray (RS-232 error).
rs232 Active2	Status of the RS-232 Active2 communication with the control workstation. Yellow (RS-232 activity) or black (no RS-232 activity) or gray (RS-232 error).
controllerTail *X*	The monitor is connected to the frame supervisor via S1 or S2. For whichever controller tail is used, green (active) or red (inactive).

Customizing HACMP/6000

To customize HACMP/6000, follow the basic steps for configuring a two-node rotating configuration with nonconcurrent access in the *AIX High Availability Cluster Multi-Processing/6000 Installation Guide*.

Configuring Resources

When you get to the configuring resources step in the HACMP/6000 installation procedure, you need to define a resource group name. For an HACWS configuration, you **must** name the resource group **hacws_group1**. To define event scripts to handle failovers for HACWS configuration, the name of the resource group needs to be known. The HACWS configuration will not fail over if the **hacws_group1** resource group name is not used.

HACMP/6000 Event Processing

The HACWS software supplies HACMP/6000 pre- and post-event scripts for all HACMP/6000 events. These supplied pre- and post-event scripts are used to customize the HACMP/6000 configuration for HACWS. To permit you to add your own pre- and post-event scripts, HACWS software provides for pre- and post-event

scripts in each pre- and post-event script that is supplied by HACWS. The following flow describes when the pre- and post-event scripts are called.

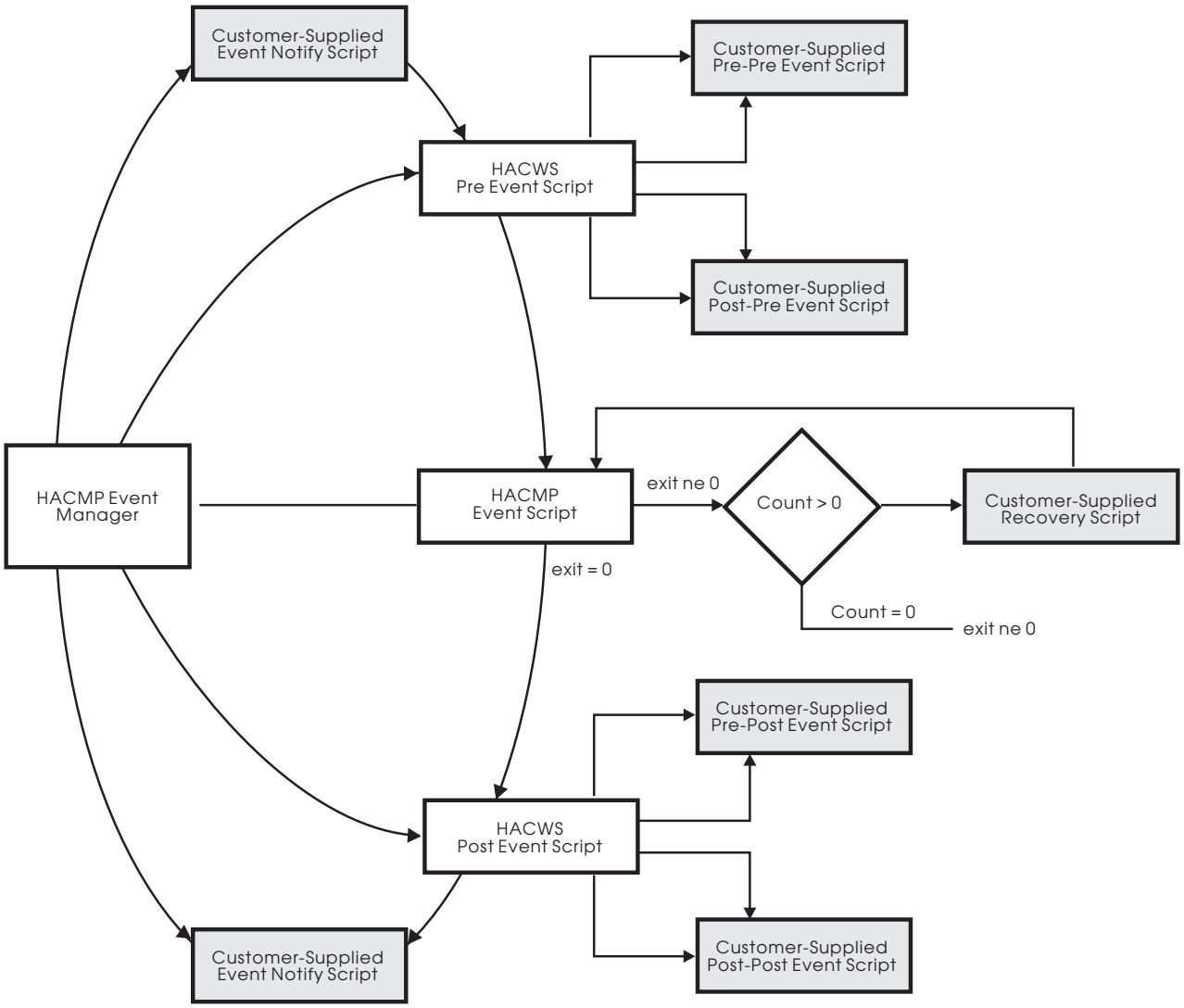


Figure 65. HACWS Pre- and Post-Events

The pre- and post-event scripts supplied with HACWS function are located in the **/usr/sbin/hacws** directory. These scripts are defined to HACMP/6000 with the **spcw_addevents** command. Do not modify these scripts. These scripts are generic and will call the scripts supplied with HACWS (found in **/usr/sbin/hacws/events**) on a per event basis. Do not modify the per event scripts.

You may provide additional pre- and post-event scripts to be called by the pre- and post-event scripts supplied with HACWS. The additional scripts you provide must be located in the **/var/adm/hacws/events** directory.

Name the scripts you supply following this naming convention:

event_name.**pre_pre_event**

event_name is the HACMP/6000 event name and **pre_pre_event** is the name for a pre-event script for an HACMP/6000 pre-event script. This naming convention extends to having **post_pre**, **pre_post** and **post_post** event scripts.

To enable the execution of event scripts, place the scripts in the proper directories and turn on the execution mode bits via the **chmod** command.

Verifying Frame Supervisor Cables

In an HACWS configuration the tty assignments for each frame **must** be identical on both the primary and backup control workstations. The new frame supervisor cards (Type 22 and 20) used for HACWS configurations have flash memory on the cards. This flash memory is used to write a unique ID so the frame can be identified. With previous cards the order in which the frames were cabled determined the frame numbers in the configuration. There was no method to check that the frame was actually still the same frame number.

If a frame number on the card does not match the frame number in the SDR, the hardware monitor will not bring the frame online. An error log entry is written to the AIX error log and BSD syslog indicating this. Note that in most cases of miscabling, two frames will not be brought online. An error log entry will be written for each frame. You have to correct the cabling to repair the situation.

When the **hardmon** daemon initializes it verifies that the frame supervisor cables match the information in the SDR. To check that the cables are correct from an inactive control workstation (there must be an active control workstation because the SDR information needs to be compared):

1. Enter **stopsrc -s hardmon** to ensure there is no hardmon daemon running on the inactive control workstation.
2. Issue the **hardmon -B** command from the inactive control workstation.

Any cable problems will be written to the AIX error log.

When an SP system is reconfigured and frames moved around and reassigned, new frame IDs need to be written to existing frames. Note that the system must be completely reinstalled so node numbers and system images are correct. This action should rarely, if ever, be taken.

In the following example we list the steps to change the frame numbers in the flash memory of the card. In this example frame number 4 is now frame number 6 and frame number 6 is now frame number 4. A complete reinstall of the SP system is also being done. Note that it would be far easier to make the cables match and not swap frames around. Swapping frames should be done only in extreme cases.

1. Determine the tty number of the cable that is connected to the frames.
2. Verify the SDR information about the frame.

   Ensure that the SDR information about the frame and tty number match how the frame is cabled. Use the **SDRGetObjects Frame** command to see which frame number matches with which tty number. Assuming that the SDR information is correct, proceed to the next step.

   If the tty number in the SDR is incorrect, use the **spframe** command to make frame number 6 use the correct tty special device name. For example, if you

want to make frame 6 use **/dev/tty6** (assuming no other frame is using **/dev/tty6**), enter **spframe 6 1 /dev/tty6**.

3. Set the frame ID in the flash memory.

   Enter **hcmds -G setid 4,6:0** to set the frame ID to the values in the SDR frame class for frames 4 and 6.

Chapter 23. The Group Services Subsystem

This chapter introduces you to the Group Services subsystem. It includes information about the subsystem's components, its configuration, other components on which it depends, and its operation. It also discusses the relationship of the Group Services subsystem to the other PSSP high availability subsystems. Finally, it describes a procedure you can use to check the status of the subsystem.

After reading this chapter, you will be able to manage the Group Services subsystem and, if necessary, perform problem determination.

Introducing Group Services

Group Services is a distributed subsystem of the IBM Parallel System Support Programs (PSSP) on the RS/6000 SP (SP). It is one of several subsystems in PSSP that provide a set of high availability services. Other subsystems that provide high availability services include the Event Management and Topology Services (heartbeat) subsystems.

The function of the Group Services subsystem is to provide other subsystems with a distributed coordination and synchronization service. These other subsystems that depend upon Group Services are called **client subsystems**. Each client subsystem forms one or more **groups** by having its processes connect to the Group Services subsystem and use the various Group Services interfaces. A process of a client subsystem is called a **GS client**. For example, Event Management is a Group Services client subsystem. The Event Manager daemon on each node ia a GS client.

A group consists of two pieces of information:

- The list of processes that have joined the group, called the **group membership list**
- A client-specified **group state value**.

Group Services guarantees that all processes that are joined to a group see the same values for the group information, and that they see all changes to the group information in the same order. In addition, the processes may initiate changes to the group information via **protocols** that are controlled by Group Services.

A GS client that has joined a group is called a **provider**. A GS client that wishes only to monitor a group, without being able to initiate changes in the group, is called a **subscriber**.

Once a GS client has initialized its connection to Group Services, it may join a group and become a provider. All other GS clients that have already joined the group (those that have already become providers) are told as part of a join protocol about the new providers that wish to join. The existing providers may either accept new joiners unconditionally (by establishing a one-phase join protocol) or vote on the protocol (by establishing an n-phase protocol). During a vote, they can choose to approve the protocol and accept the new providers into the group, or reject the protocol and refuse to allow the new providers to join.

© Copyright IBM Corp. 1995, 1997

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

Join and failure protocols are used to modify the group's membership list. Any provider in the group may also propose protocols to modify the group's state value. All protocols may be either unconditional (one-phase) protocols, which are automatically approved and not voted on, or conditional (n-phase) protocols, which are voted on by the providers.

During each phase of an n-phase protocol, each provider may take application-specific action and *must* vote to approve, reject, or continue the protocol. The protocol completes when it is either approved (the proposed changes become established in the group), or rejected (the proposed changes are dropped).

For more conceptual information about the Group Services subsystem, see *PSSP Group Services Programming Guide and Reference*.

Group Services Components

The Group Services subsystem consists of the following components:

Group Services daemon
The central component of the Group Services subsystem.

Group Services API (GSAPI)
The application programming interface that GS clients use to obtain the services of the Group Services subsystem.

Port numbers TCP/IP port numbers that the Group Services subsystem uses for communications. The Group Services subsystem also uses Unix domain sockets.

Control script A shell script that is used to add, start, stop, and delete the Group Services subsystem, which operates under System Resource Controller (SRC) control.

Files and directories
Various files and directories that are used by the Group Services subsystem to maintain run-time data.

The sections that follow contain more details about each of these components.

The Group Services Daemon (hagsd)

The Group Services daemon is contained in the executable file **/usr/lpp/ssp/bin/hagsd**. This daemon runs on each node of a system partition and on the control workstation. If there is more than one system partition, then multiple daemons run on the control workstation, one per system partition.

Note that the operational domain of the Group Services subsystem is a system partition. The control workstation is considered to be part of each domain. Unless otherwise stated, a reference to "the Group Services subsystem" is a reference to the Group Services subsystem in a single system partition.

A GS client communicates with a Group Services daemon that is running on the same node as the GS client. A GS client communicates with the Group Services daemon, through the GSAPI, using a Unix domain socket. Before a GS client registers with Group Services, it must set the **HA_SYSPAR_NAME** environment variable to the name of the system partition in which it is executing.

The Group Services API (GSAPI)

The Group Services Application Programming Interface (GSAPI) is a shared library that a GS client uses to obtain the services of the Group Services subsystem. This shared library is supplied in two versions: one for non-thread safe programs and one for thread-safe programs. These libraries are referenced by the following path names:

- **/usr/lib/libha_gs.a** (non-thread safe version)
- **/usr/lib/libha_gs_r.a** (thread-safe version)

These path names are actually symbolic links to **/usr/lpp/ssp/lib/libha_gs.a** and **/usr/lpp/ssp/lib/libha_gs_r.a**, respectively. The symbolic links are placed in **/usr/lib** for ease of use. For serviceability, the actual libraries are placed in the **/usr/lpp/ssp/lib** directory. These libraries are supplied as shared libraries, also for serviceability.

For details on the GSAPI, see *Group Services Programming Guide and Reference*.

To allow non-root users to use Group Services:

1. Create a group named **hagsuser**.
2. Add the desired user IDs to the **hagsuser** group.
3. Stop and restart **hags** (if it was running before you created the **hagsuser** group).

Users in the created **hagsuser** group can use Group Services (**libha_gs**).

Port Numbers and Sockets

The Group Services subsystem uses several types of communications:

- UDP port numbers for intra-domain communications, that is, communications between Group Services daemons within an operational domain.
- Unix domain sockets for communication between GS clients and the local Group Services daemon (via the GSAPI).

Intra-Domain Port Numbers

For communication between Group Services daemons within an operational domain, the Group Services subsystem uses a single UDP port number. This port number is recorded in the **Syspar_ports** SDR class. The class contains two attributes: **subsystem** and **port**. The class object whose **subsystem** attribute has a value of **hags** contains the port number for the Group Services subsystem. Note that the **Syspar_ports** class is a partitioned SDR class, that is, there is a set of objects for each system partition. When you configure the Group Services subsystem for a particular system partition, you use the SDR data that corresponds to that system partition.

The Group Services port number is stored in the SDR so that, when the Group Services subsystem is configured on each node, the port number is fetched from the SDR. This ensures that the same port number is used by all Group Services daemons in the same operational domain. Note that because a Group Services daemon from each operational domain runs on the control workstation, the Group Services subsystem in each system partition must have a unique port number.

This intra-domain port number is also set in the **/etc/services** file, using the service name **hags.***syspar_name*, where *syspar_name* is the name of the system partition. The **/etc/services** file is updated on all nodes in the system partition and on the control workstation. The Group Services daemon obtains the port number from the **/etc/services** file during initialization.

Unix Domain Sockets

Unix domain sockets are used for communication between GS clients and the local Group Services daemon (via the GSAPI). These are connection-oriented sockets. The following socket name is used (*syspar_name* is the name of the system partition):

/var/ha/soc/hagsdsocket.*syspar_name*
: Used by the GSAPI to connect to the Group Services daemon.

The Control Script (hagsctrl)

The Group Services control script is contained in the executable file **/usr/lpp/ssp/bin/hagsctrl**. This script is normally invoked by the **syspar_ctrl** script, which provides an interface to all of the system partition-sensitive subsystems.

If necessary, you can invoke **hagsctrl** directly from the command line. Note that before you invoke **hagsctrl**, you must ensure that the **SP_NAME** environment variable is set to the appropriate system partition name.

For more information about **hagsctrl** and **syspar_ctrl**, see *PSSP Command and Technical Reference*.

The purpose of the **hagsctrl** command is to add (configure) the Group Services subsystem to a system partition. It can also be used to remove the subsystem from a system partition, start the subsystem, stop it, and clean the subsystem from all system partitions.

For more information, see “Group Services Configuration” on page 358.

Files and Directories

The Group Services subsystem uses the following directories:

- **/var/ha/lck**, for lock files
- **/var/ha/log**, for log files
- **/var/ha/run**, for Group Services daemon current working directories
- **/var/ha/soc**, for socket files.

The /var/ha/lck Directory (Lock Files)

In the **/var/ha/lck** directory, the **hags.tid.***syspar_name* directory is used to ensure a single running instance of the Group Services daemon, and to establish an instance number for each invocation of the daemon. In the directory name, *syspar_name* is the name of the system partition. On the control workstation, there may be several instances of the Group Services daemon running, but they are in different operational domains.

The /var/ha/log Directory (Log Files)

The **/var/ha/log** directory contains trace output from the Group Services daemon.

On the nodes, the file is called **hags**_*nodenum_instnum.syspar_name*, where:

- *nodenum* is the node number on which the daemon is running
- *instnum* is the instance number of the daemon.
- *syspar_name* is the name of the system partition to which the node belongs.

On the control workstation and the nodes, the file **hags.default.***syspar_name.nodenum_instnum* contains trace output from the initial startup of the daemon.

The Group Services daemon limits the log size to a pre-established number of lines (by default, 5,000 lines). When the limit is reached, the daemon appends the string **.bak** to the name of the current log file and begins a new log. If a **.bak** version already exists, it is removed before the current log is renamed.

The /var/ha/run Directory (Daemon Working Files)

In the **/var/ha/run** directory, a directory called **hags.***syspar_name* is created, where *syspar_name* is the system partition name. This directory is the current working directory for the Group Services daemon. If the Group Services daemon abnormally terminates, the core dump file is placed in this directory. Whenever the Group Services daemon starts, it renames any core file to **core**_*nodenum.instnum*, where *nodenum* is the node number on which the daemon is running and *instnum* is the instance number of the previous instance of the daemon.

Components on Which Group Services Depends

The Group Services subsystem depends on the following components:

System Resource Controller (SRC)
: An AIX feature that can be used to define and control subsystems. The Group Services subsystem is called **hags** on SP nodes and is called **hags.***syspar_name* on the control workstation, where *syspar_name* is the name of a system partition. The subsystem name is used with the SRC commands (for example, **startsrc** and **lssrc**).

System Data Repository (SDR)
: A repository of SP system configuration information.

Topology Services
: A PSSP subsystem that is used to determine which nodes in a system can be reached (that is, are running) at any given time. It is often referred to as **heartbeat**. The Topology Services subsystem is SRC-controlled. It is called **hats** on

SP nodes and is called **hats.***syspar_name* on the control workstation, where *syspar_name* is the name of a system partition.

Group Services Configuration and Operation

The following sections describe how the components of the Group Services subsystem work together to provide group services. Included are discussions of Group Services:

- Configuration
- Daemon initialization and errors
- Operation

Group Services Configuration

The Group Services subsystem is installed as part of the installation of the PSSP product. The Group Services subsystem is contained in the **ssp.ha** fileset.

After the components are installed, the subsystem must be configured for operation. Group Services configuration is performed by the **hagsctrl** command, which is invoked by the **syspar_ctrl** command.

The **syspar_ctrl** command configures all of the system partition-sensitive subsystems. The person who installs PSSP issues the **syspar_ctrl** command during installation of the control workstation. The **syspar_ctrl** command is executed automatically on the nodes when the nodes are installed. The **syspar_ctrl** command is also executed automatically when system partitions are created or destroyed. For more information on using the **syspar_ctrl** command, see *PSSP Installation and Migration Guide* and *PSSP Command and Technical Reference*.

The **hagsctrl** command provides a number of functions for controlling the operation of the Group Services system. You can use it to:

- Add (configure) the Group Services subsystem
- Start the subsystem
- Stop the subsystem
- Delete (unconfigure) the subsystem
- "Clean" all Group Services subsystems
- Turn tracing of the Group Services daemon on or off

Except for the clean function, **hagsctrl** affects the Group Services subsystem in the current system partition, that is, the system partition that is specified by the **SP_NAME** environment variable.

Adding the Subsystem

If the **hagsctrl** command is running on the control workstation, the first step in the add function is to select a Group Services daemon communications port number and save it in the **Syspar_ports** SDR class. This port number is then placed in the /**etc**/**services** file. If the **hagsctrl** command is running on a node, the port number is fetched from the SDR and placed in the /**etc**/**services** file. Port numbers are selected from the range 10000 through 10100.

The second step is to add the Group Services daemon to the System Resource Controller (SRC) using the **mkssys** command. The system partition name is an argument to the **hagsd** program in the SRC subsystem specification.

The third step is to add an entry to the /**etc**/**inittab** file so that the Group Services daemon will be started during boot. However, if **hagsctrl** is running on a High Availability Control Workstation (HACWS), no entry is made in the /**etc**/**inittab** file. Instead, HACWS manages starting and stopping the Group Services daemon.

Note that if the **hagsctrl** add function terminates with an error, the command can be rerun after the problem is fixed. The command takes into account any steps that already completed successfully.

Starting and Stopping the Subsystem

The start and stop functions of the **hagsctrl** command simply run the **startsrc** and **stopsrc** commands, respectively. However, **hagsctrl** automatically specifies the subsystem argument to these SRC commands.

Deleting the Subsystem

The delete function of the **hagsctrl** command removes the subsystem from the SRC, removes the entry from /**etc**/**inittab**, and removes the Group Services daemon communications port number from /**etc**/**services**. It does **not** remove anything from the SDR, because the Group Services subsystem may still be configured on other nodes in the operational domain.

Cleaning the Subsystem

The clean function of the **hagsctrl** command performs the same function as the delete function, except in all system partitions. In addition, it removes the Group Services daemon remote client communications port number from the /**etc**/**services** file.

The clean function does **not** remove anything from the SDR. This function is provided to support restoring the system to a known state, where the known state is the (possibly restored copy of the) SDR database.

Tracing the Subsystem

The tracing function of the **hagsctrl** command is provided to supply additional problem determination information when it is requested by the IBM Support Center. Normally, tracing should **not** be turned on, because it may slightly degrade Group Services subsystem performance and can consume large amounts of disk space in the /**var** file system.

Group Services Daemon Initialization

Normally, the Group Services daemon is started by an entry in the **/etc/inittab** file, via the **startsrc** command. If necessary, the Group Services daemon can be started using the **hagsctrl** command or the **startsrc** command directly.

During its initialization, the Group Services daemon performs the following steps:

1. It gets the number of the node on which it is running using the **/usr/lpp/ssp/install/bin/node_number** command. Node 0 is the control workstation.
2. It tries to connect to the Topology Services subsystem. If the connection cannot be established because the Topology Services subsystem is not running, it is scheduled to be retried every 20 seconds. This continues until the connection to Topology Services is established. Until the connection is established, the Group Services daemon writes an AIX error log entry periodically and no clients may connect to the Group Services subsystem.
3. It performs actions that are necessary to become a daemon. This includes establishing communications with the SRC subsystem so that it can return status in response to SRC commands.
4. It establishes the Group Services domain, which is the set of nodes within the SP system partition in which a Group Services daemon is executing.

   At this point, one of the GS daemons establishes itself as the GS nameserver. For details, see "Establishing the GS Nameserver."

   Until the domain is established, no GS client requests to join or subscribe to groups are processed.
5. It enters the main control loop.

   In this loop, the Group Services daemon waits for requests from GS clients, messages from other Group Services daemons, messages from the Topology Services subsystem, and requests from the SRC for status.

Establishing the GS Nameserver

The Group Services subsystem must be able to keep track of the groups that its clients want to form. To do this, it establishes a GS nameserver within each domain (a domain is the set of running nodes within each SP system partition). The GS nameserver is responsible for keeping track of all client groups that are created in that domain.

To ensure that only one node becomes a GS nameserver, Group Services uses the following protocol:

1. When each daemon is connected to the Topology Services subsystem, it waits for Topology Services to tell it which nodes are currently running in this system partition.
2. Based on the input from Topology Services, each daemon finds the lowest-numbered running node in the domain. The daemon compares its own node number to the lowest-numbered node and performs one of the following:
   - If the daemon's node is the lowest-numbered node, the daemon waits for all other running nodes to nominate it as the GS nameserver.

- If the daemon's node is not the lowest-numbered node, it sends nomination messages to the lowest-numbered node periodically, initially every 5 seconds.

3. Once all running nodes have nominated the GS nameserver-to-be and a coronation timer (about 20 seconds) has expired, the nominee sends an insert message to the nodes. All nodes must acknowledge this message. When they do, the nominee becomes the established GS nameserver, and it sends a commit message to all of the nodes.
4. At this point, the Group Services domain is established, and requests by clients to join or subscribe to groups are processed.

Note that this description is in effect when all nodes are being booted simultaneously, such as at initial system power-on. It is often the case, however, that a Group Services daemon is already running on at least one node (for example, the control workstation) and is already established as the domain's GS nameserver. In that case, the GS nameserver waits only for Topology Services to identify the newly running nodes. The GS nameserver will then send the newly running nodes proclaim messages that direct the nodes to nominate it as nameserver. Once those nodes then nominate the GS nameserver, the GS nameserver simply executes one or more insert protocols to insert the newly-running nodes into the domain.

Group Services Initialization Errors

The Group Services subsystem creates AIX error log entries to indicate severe internal problems. For most of these, the best response is to contact the IBM Support Center.

However, if you get a message that there has been no heartbeat connection for some time, it indicates that the Topology Services subsystem is not running.

To check the status of the Topology Services subsystem, issue the **lssrc -g hats** command. If the response indicates that the Topology Services subsystem is inoperative, you will need to restart it using the **hatsctrl -s** command. If you are unable to restart it, call the IBM Support Center.

Group Services Daemon Operation

Normal operation of the Group Services subsystem requires no administrative intervention. The subsystem normally recovers from temporary failures, such as node failures or failures of Group Services daemons, automatically. However, there are some operational characteristics that may be of interest to administrators.

Due to AIX per-process file descriptor limits, the Group Services subsystem can support a maximum of approximately 2000 GS clients on each node.

The maximum number of nodes that can be contained within a domain is 2048, although Group Services can support node numbers in the range 0 to 65535.

The maximum number of groups to which a GS client may subscribe or that a GS client may join is equivalent to the largest value containable in a signed integer variable.

The maximum number of groups allowed within a domain is 65,535.

These limits are the theoretical maximum limits. In practice, the amount of memory available to the Group Services daemon and its clients will reduce the limits to smaller values.

Group Services Procedures

For the most part the Group Services subsystem runs itself without requiring administrator intervention. However, on occasion, you may need to check the status of the subsystem.

Displaying the Status of the Group Services Daemon

You can display the operational status of the Group Services daemon by issuing the **lssrc** command.

On the control workstation, enter:

```
lssrc -l -s hags.syspar_name
```

where *syspar_name* is the name of the system partition of interest.

On a node, enter:

```
lssrc -l -s hags
```

In response, the **lssrc** command writes the status information to standard output. The information includes:

- The information provided by the **lssrc -s hags** command (short form)
- The number of currently connected clients and their process IDs
- The status of the Group Services domain
- The node number on which the GS nameserver is running
- Statistics for client groups with providers or subscribers on this node.

Note that if the **lssrc** command times out, the Group Services daemon is probably unable to connect to the Topology Services subsystem. For more information, see "Group Services Initialization Errors" on page 361.

Here is a sample of the output from the **lssrc -l -s hags** command on a node for the **k21sp2** system partition:

```
Subsystem         Group            PID     Status
 hags             hags             11938   active
 4 locally-connected clients.  Their PIDs:
 21344 17000 18852 23486
 HA Group Services domain information:
 Domain established by node 9.
 Number of groups known locally: 3
                     Number of   Number of local
 Group name          providers   providers/subscribers
 cssMembership             5           1           0
 WomSchg_1                 5           1           1
 ha_em_peers               7           1           0
```

Here is a sample of the output of the **lssrc -l -s hags.k21sp2** command on the control workstation for the same domain:

```
Subsystem         Group            PID     Status
 hags.k21sp2      hags             24804   active
 1 locally-connected clients.  Their PIDs:
 44146
 HA Group Services domain information:
 Domain established by node 9.
 Number of groups known locally: 1
                    Number of   Number of local
 Group name         providers   providers/subscribers
 ha_em_peers              7           1           0
```

In this domain, the GS nameserver is on node 9 of the system partition.

If a GS nameserver has not yet been established, the status indicates that the domain is not established. Similarly, if the GS nameserver fails, the status shows that the domain is recovering. Both of these conditions should clear in a short time. If they do not and the Topology Services subsystem is active, call the IBM Support Center.

Chapter 24. The Event Management Subsystem

This chapter introduces you to the Event Management subsystem. It includes information about the subsystem's components, its configuration, other components on which it depends, and its operation. It also discusses the relationship of the Event Management subsystem to the other PSSP high availability subsystems. Finally, it contains procedures you can use to check the status of the subsystem, or to add or change some of the Event Management configuration data.

After reading this chapter, you will be able to manage the Event Management subsystem and, if necessary, perform problem determination.

Introducing Event Management

Event Management is a distributed subsystem of the IBM Parallel System Support Programs (PSSP) on the RS/6000 SP. It is one of several subsystems in PSSP that provide a set of high availability services. Other subsystems that provide high availability services include the Group Services and Topology Services (heartbeat) subsystems.

The function of the Event Management subsystem is to match information about the state of system resources with information about resource conditions that are of interest to client programs, which may include applications, subsystems, and other programs. In this chapter, these client programs are referred to as **EM clients**.

Resource states are represented by resource variables. Resource conditions are represented as expressions, called **predicates**, that have a syntax that is a subset of the expression syntax of the C programming language.

Resource monitors are programs that observe the state of specific system resources and transform this state into several resource variables. The resource monitors periodically pass these variables to the Event Manager daemon. The Event Manager daemon applies predicates, which have been specified by EM clients, to each resource variable. If the predicate is true, an event is generated and sent to the appropriate EM client. EM clients may also query the Event Manager daemon for the current values of resource variables.

Resource variables, resource monitors, and other related information are specified in several System Data Repository (SDR) object classes. Information stored in these SDR classes is then translated into a form that can be easily used by the Event Management subsystem.

For more conceptual information about the Event Management subsystem, see *PSSP Event Management Programming Guide and Reference*.

Event Management Components

The Event Management subsystem consists of the following components:

Event Manager daemon
: The central component of the Event Management subsystem.

© Copyright IBM Corp. 1995, 1997

Event Management API (EMAPI)
The application programming interface that EM clients use to obtain the services of the Event Management subsystem.

Resource Monitor API (RMAPI)
The application programming interface that resource monitors use to supply resource variables to the Event Manager daemon.

Resource monitors Programs that monitor the state of system resources. Several resource monitors are shipped with PSSP.

Port numbers TCP/IP port numbers that the Event Management subsystem uses for communications. The Event Management subsystem also uses Unix domain sockets.

Configuration data Information that defines resource variables, resource monitors, and related data to the Event Management subsystem. This information is stored in the SDR.

Configuration utility
A program that translates the Event Management configuration data that is stored in the SDR to a binary format that is used by the Event Manager daemon and the RMAPI.

Configuration database
The Event Management Configuration Database (EMCDB) is the output of the configuration utility.

Control script A shell script that is used to add, start, stop, and delete the Event Management subsystem, which operates under System Resource Controller (SRC) control.

Configuration load utility
A program that loads the default configuration data that is shipped with PSSP into the SDR.

Default configuration data
A file that is shipped with PSSP that contains the default configuration data for the Event Management subsystem.

Files and directories
Various files and directories that are used by the Event Management subsystem to maintain run-time data.

The sections that follow contain more details about each of these components.

The Event Manager Daemon (haemd)

The Event Manager daemon is contained in the executable file **/usr/lpp/ssp/bin/haemd**. This daemon runs on each node of a system partition and on the control workstation. If there is more than one system partition, then multiple daemons run on the control workstation, one per system partition.

Note that the operational domain of the Event Management subsystem is a system partition. The control workstation is considered to be part of each domain. Unless otherwise stated, a reference to "the Event Management subsystem" is a reference to the Event Management subsystem in a single system partition.

EM clients communicate with a single Event Manager daemon in the operational domain that contains the resources in which they are interested, as follows:

- If the operational domain contains the SP node on which the EM client is running, the client communicates with the local Event Manager daemon, through the EMAPI, using a Unix domain socket.
- If an EM client is running on the control workstation, it must specify, explicitly or implicitly, a system partition. The EM client then communicates with the Event Manager daemon on the control workstation, in that system partition, using a Unix domain socket.

An EM client that is running on a node that is outside of the operational domain that contains the resources in which it is interested uses a TCP socket to communicate with the Event Manager daemon on the control workstation that is in that operational domain. An EM client that is running on any RS/6000 is considered to be on a node outside of any operational domain. In both of these cases, the EM client must specify, explicitly or implicitly, the name of a system partition that is in the operational domain of interest. For an EM client to run on an RS/6000, the PSSP fileset **ssp.clients** must be installed on the RS/6000.

Once an EM client has established communications with an Event Manager daemon, the EM client has access to all of the resources in that operational domain.

Resource monitors communicate with the Event Manager daemon, through the RMAPI, using Unix domain sockets. Resource monitors are always on the same node as the Event Manager daemon. Note that the control workstation is node 0.

The Event Management API (EMAPI)

The Event Management Application Programming Interface (EMAPI) is a shared library that an EM client uses to obtain the services of the Event Management subsystem. This shared library is supplied in two versions: one for non-thread safe programs and one for thread-safe programs. These libraries are referenced by the following path names:

- **/usr/lib/libha_em.a** (non-thread safe version)
- **/usr/lib/libha_em_r.a** (thread-safe version)

These path names are actually symbolic links to **/usr/lpp/ssp/lib/libha_em.a** and **/usr/lpp/ssp/lib/libha_em_r.a**, respectively. The symbolic links are placed in **/usr/lib** for ease of use. For serviceability, the actual libraries are placed in the **/usr/lpp/ssp/lib** directory. These libraries are supplied as shared libraries, also for serviceability.

For details on the EMAPI, see *Event Management Programming Guide and Reference*.

The Resource Monitor API (RMAPI)

The Resource Monitor Application Programming Interface (RMAPI) is a shared library that a resource monitor uses to supply resource variables to the Event Manager daemon. This shared library is supplied only in a non-thread safe version. This library is referenced through the path name

- **/usr/lib/libha_rr.a**

which is a symbolic link to **/usr/lpp/ssp/lib/libha_rr.a**. The symbolic link is placed in **/usr/lib** for ease of use. For serviceability, the actual library is placed in the **/usr/lpp/ssp/lib** directory. This library is supplied as a shared library, also for serviceability.

For details on the RMAPI, see *Event Management Programming Guide and Reference*.

Resource Monitors

A resource monitor conforms to one of the following programming models:

- The resource monitor is a daemon. When necessary, it is started by the Event Manager daemon. The Event Manager daemon connects to the resource monitor to establish communications. The resource monitor has a connection type of **server**.
- The resource monitor logic is incorporated in a subsystem that manages the resources. The Event Manager daemon connects to the resource monitor to establish communications, but the Event Manager daemon does not start the resource monitor (because it is actually another subsystem). The resource monitor has a connection type of **server**.
- The resource monitor logic is implemented in a command. The command can be used by scripts to supply resource variables to the Event Manager daemon. A command-based resource monitor connects to the Event Manager daemon to establish communications. The resource monitor has a connection type of **client**.

In addition, the Event Manager daemon itself performs some resource monitoring function. This function is considered to be a resource monitor with a connection type of **internal**.

PSSP supplies the following resource monitors:

IBM.PSSP.harmld Supplies resource variables for the CSS, VSD, and LoadLeveler subsystems. This data is also furnished to the Performance Monitor subsystem. This is a daemon with a connection type of **server**.

IBM.PSSP.harmpd Supplies resource variables that represent the number of processes executing a particular program. These variables can be used to determine whether a particular system daemon is running. This is a daemon with a connection type of **server**.

IBM.PSSP.hmrmd Supplies resource variables that represent the hardware state of the SP system. The resource information is obtained from the PSSP hardware monitoring subsystem (**hardmon**). This is a daemon with a connection type of **server**.

IBM.PSSP.pmanrmd
Supplies resource variables provided by the PSSP Problem Management subsystem. This is a command-based resource monitor with a connection type of **client**.

There are also several internal resource monitors incorporated in the Event Manager daemon itself:

Membership	Supplies resource variables that represent the Host Membership and Adapter Membership states. The Event Manager daemon obtains this information directly from the Group Services subsystem by subscribing to the **HostMembership**, **enMembership**, and **cssMembership** system groups.
Response	Supplies resource variables that represent the information in the **host_responds** and **switch_responds** SDR classes. These resource variables are provided for compatibility with earlier releases of PSSP.
aixos	Supplies resource variables that represent AIX operating system resources.

Port Numbers and Sockets

The Event Management subsystem uses several types of communications:

- UDP port numbers for intra-domain communications, that is, communications between Event Manager daemons within an operational domain.
- TCP port numbers for remote communications, that is, communications between Event Manager daemons and EM clients that are running outside their operational domains.
- Unix domain sockets for communication between EM clients and the local Event Manager daemon (via the EMAPI) and between resource monitors and the Event Manager daemon (via the RMAPI).

Intra-Domain Port Numbers

For communication between Event Manager daemons within an operational domain, the Event Management subsystem uses a single UDP port number. This port number is recorded in the **Syspar_ports** SDR class. The class contains two attributes: **subsystem** and **port**. The class object whose **subsystem** attribute has a value of **haem** contains the port number for the Event Management subsystem. Note that the **Syspar_ports** class is a partitioned SDR class, that is, there is a set of objects for each system partition. When you configure the Event Management subsystem for a particular system partition, you use the SDR data that corresponds to that system partition.

The Event Management port number is stored in the SDR so that, when the Event Management subsystem is configured on each node, the port number is fetched from the SDR. This ensures that the same port number is used by all Event Manager daemons in the same operational domain. Note that because an Event Manager daemon from each operational domain runs on the control workstation, the Event Management subsystem in each system partition must have a unique port number.

This intra-domain port number is also set in the /**etc**/**services** file, using the service name **haem.***syspar_name*, where *syspar_name* is the name of the system partition. The /**etc**/**services** file is updated on all nodes in the system partition and on the control workstation. The Event Manager daemon obtains the port number from the /**etc**/**services** file during initialization.

Remote Port Numbers

The Event Management subsystem also uses a single TCP port number to accept communications from clients outside its operational domain. All Event Manager daemons that run on the control workstation use this port number to accept connections from EM clients outside their respective operational domains. Each daemon binds to a distinct IP address that is the IP address of its system partition.

The TCP port number is stored in the **SP_ports** SDR system class. The class contains the **daemon**, **hostname**, and **port** attributes. The class object whose **daemon** attribute has a value of **haemd** contains the port number for the Event Management subsystem(s). The **hostname** attribute is not used in this object.

This remote port number is also set in the /**etc**/**services** file on the control workstation, using the service name **haemd**.

Unix Domain Sockets

Unix domain sockets are used for communication:

- Between EM clients and the local Event Manager daemon (via the EMAPI)
- Between resource monitors and the Event Manager daemon (via the RMAPI).

These are connection-oriented sockets. The following socket names are used (*syspar_name* is the name of the system partition):

/**var**/**ha**/**soc**/**em.clsrv.***syspar_name*
: Used by the EMAPI to connect to the Event Manager daemon.

/**var**/**ha**/**soc**/**em.rmsrv.***syspar_name*
: Used by resource monitors with a connection type of **client** to connect to the Event Manager daemon.

/**var**/**ha**/**soc**/**em.RM***rmname***.***syspar_name*
: Used by the Event Manager daemon to connect to the resource monitor that is specified by *rmname*. This resource monitor has a connection type of **server**.

Configuration Data

Configuration information for the Event Management subsystem is stored in the SDR as objects in the following classes:

EM_Resource_Variable
: Contains the definitions of all resource variables. The resources with which these variables are associated are implied in the definitions.

EM_Instance_Vector
: Contains the definitions of all of the instance vectors that are associated with the resources that are specified in the **EM_Resource_Variable** class.

EM_Structured_Byte_String
: Contains the definitions of the structured fields for all resource variables that have a data type of structured byte string (SBS).

EM_Resource_Class
Contains the definitions for the resource classes with which each resource variable is associated.

EM_Resource_Monitor
Contains the definitions for all resource monitors.

For details about the Event Management SDR classes and attributes, see *Event Management Programming Guide and Reference*.

These SDR classes are partitioned classes. Therefore, each Event Management subsystem operational domain has its own configuration data. This permits a different set of resources to be monitored in each operational domain, as appropriate for the workload within the system partition.

The Event Management SDR objects are loaded from the default configuration data when the Event Management subsystem is added to a system partition using the control script.

There are several attributes that you, as the system administrator, may wish to change. They are:

- The **rvPredicate** attribute of the **EM_Resource_Variable** class

  Change this attribute to add or replace a default predicate for a resource variable.

- The **rcObservation_interval** attribute of the **EM_Resource_Class**.

  Change this attribute to increase or decrease the time between observations of a resource variable value by the Event Manager daemon.

For more information, see “Changing Configuration Data in the SDR” on page 384.

If resource monitors supplied by other IBM products or by third parties are installed on the SP system, you must load the configuration data supplied with the resource monitors into the SDR. For more information, see “Loading Non-PSSP Configuration Data into the SDR” on page 384.

The Configuration Utility (haemcfg)

The Event Management configuration utility is contained in the executable file **/usr/lpp/ssp/bin/haemcfg**. This utility is normally run on the control workstation in the environment of a system partition, that is, the value of the **SP_NAME** environment variable determines the partitioned SDR classes that are referenced. If **SP_NAME** is not set, the default system partition is assumed.

The **haemcfg** command converts the data stored in the Event Management SDR classes to a binary format called the Event Management Configuration Database (EMCDB). The EMCDB is placed in the staging file called **/spdata/sys1/ha/cfg/em.***syspar_name***.cdb**, where *syspar_name* is the name of the system partition in which the utility is run.

A version string that uniquely identifies the EMCDB is placed both in the EMCDB itself and in the **Syspar** partitioned SDR class. The version string is set as the value of the **haem_cdb_version** attribute in the object whose **syspar_name** attribute value matches the current system partition. The version string is used to

ensure that all of the Event Manager daemons and resource monitors that are running within the operational domain are using the same copy of the EMCDB.

If a previous copy of the **em.***syspar_name***.cdb** file exists, it is renamed to /**spdata**/**sys1**/**ha**/**cfg**/**em.***syspar_name***.cdb.***ssss,nnnn,v*, where *ssss,nnnn,v* is the version string of the previous copy.

The **haemcfg** command is automatically invoked by the control script when the Event Management subsystem is added to a system partition.

For details on the **haemcfg** command, see *PSSP Command and Technical Reference*.

The Configuration Database (EMCDB)

The Event Management Configuration Database (EMCDB) is produced by the **haemcfg** command from the information in the SDR. The format of the EMCDB is designed to permit quick loading of the database by the Event Manager daemon and the RMAPI. It also contains configuration data in an optimized format to minimize the amount of data that must be sent between Event Manager daemons and between an Event Manager daemon and its resource monitors.

When the SDR data is compiled, the EMCDB is placed in a staging file. When the Event Manager daemon on a node or the control workstation initializes, it automatically copies the EMCDB from the staging file to a run-time file on the node or the control workstation. The run-time file is called /**etc**/**ha**/**cfg**/**em.***syspar_name***.cdb**, where *syspar_name* is the name of the system partition.

For more information, see “Reading the EMCDB” on page 380.

The Control Script (haemctrl)

The Event Management control script is contained in the executable file /**usr**/**lpp**/**ssp**/**bin**/**haemctrl**. This script is normally invoked by the **syspar_ctrl** script, which provides an interface to all of the system partition-sensitive subsystems.

If necessary, you can invoke **haemctrl** directly from the command line. Note that before you invoke **haemctrl**, you must ensure that the **SP_NAME** environment variable is set to the appropriate system partition name.

For more information about **haemctrl** and **syspar_ctrl**, see *PSSP Command and Technical Reference*.

The purpose of the **haemctrl** command is to add (configure) the Event Management subsystem to a system partition. It can also be used to remove the subsystem from a system partition, start the subsystem, stop it, and clean the subsystem from all system partitions.

For more information, see “Event Management Configuration” on page 377.

The Configuration Load Utility (haemloadcfg)

The configuration load utility is contained in the executable file **/usr/lpp/ssp/install/bin/haemloadcfg**. The purpose of this program is to load the default configuration data into the SDR. Normally, it is invoked by the **haemctrl** script, but you can also use it to load additional configuration data into the SDR. Note that before you invoke **haemloadcfg**, you must ensure that the **SP_NAME** environment variable is set to the appropriate system partition name.

By default, this utility does not replace existing objects in the SDR. Input data is matched with existing objects based on key attributes. The key attributes for each of the Event Management SDR classes are:

SDR Class	Key Attributes
EM_Resource_Variable	**rvName**
EM_Instance_Vector	**ivResource_name**, **ivElement_name**
EM_Structured_Byte_String	**sbsVariable_name**, **sbsField_name**
EM_Resource_Class	**rcClass**
EM_Resource_Monitor	**rmName**

Note that the way in which the **haemloadcfg** command handles existing SDR objects is different from the way in which the **SDRCreateObjects** command handles them. The **SDRCreateObjects** command creates a new object as long as the attributes, taken as a group, are unique.

The **haemloadcfg** command is actually a shell script that invokes the **/usr/lpp/ssp/bin/loadsdr** program. The purpose of this program is to batch load any number of objects into SDR classes. However, the **loadsdr** command cannot load arbitrary SDR classes. It can load only the Event Management SDR classes.

The Default Configuration Data (haemloadlist)

The default Event Management configuration data for PSSP is supplied in the **/usr/lpp/ssp/install/config/haemloadlist** file. The data in this file is loaded into each system partition to which the Event Management subsystem is added by the **haemctrl** script, using the **haemloadcfg** command.

Files and Directories

The Event Management subsystem uses the following directories:

- **/var/ha/lck**, for lock files
- **/var/ha/log**, for log files
- **/var/ha/run**, for Event Manager daemon current working directories
- **/var/ha/soc**, for socket files
- **/etc/ha/cfg**, for run-time EMCDB and related files
- **/spdata/sys1/ha/cfg**, as the staging area for EMCDB files.

The /var/ha/lck Directory (Lock Files)

In the **/var/ha/lck** directory, the **em.RM***rmname***.***syspar_name* file is used to ensure a single running instance of a resource monitor. In the file name, *rmname* is the name of the resource monitor and *syspar_name* is the name of the system partition. On the control workstation, there may be several instances of the same resource monitor running, but they are in different operational domains.

The **em.RM***rmname***SHM** file is used by the RMAPI to create a shared memory key when the resource monitor attaches to the shared memory that is managed by the SPMI library.

The **em.haemd.***syspar_name* file is used to ensure a single running instance of an Event Manager daemon in an operational domain.

The /var/ha/log Directory (Log Files)

In the **/var/ha/log** directory, the **em.trace.***syspar_name* file contains trace output from the Event Manager daemon. In the file name, *syspar_name* is the name of the system partition.

The **em.msgtrace.***syspar_name* file contains message trace output from the Event Manager daemon.

The **em.default.***syspar_name* file contains any error messages from the Event Manager daemon that cannot be written to the AIX Error Log. Normally, all daemon error messages are written to the AIX Error Log.

In addition to error messages that cannot be written to the AIX Error Log, the **/var/ha/log/em.default.syspar_name** file also contains error messages that result from repetitive operational errors. Therefore, both the AIX Error Log and the **/var/ha/log/em.default.syspar** file must be examined when performing problem determination on the Event Management subsystem.

The size of the **/var/ha/log/em.default.syspar_name** file is examined every two minutes. If the size of the file exceeds 256K, the file is renamed to **/var/ha/log/em.default.syspar_name.last** and a new default file is created. No more than two copies of this file are kept: the “current” **em.default.syspar_name** file and the “last” file.

The /var/ha/run Directory (Daemon Working Files)

In the **/var/ha/run** directory, a directory called **haem.***syspar_name* is created, where *syspar_name* is the system partition name. This directory is the current working directory for the Event Manager daemon. If the Event Manager daemon abnormally terminates, the core dump file is placed in this directory. Whenever the Event Manager daemon starts, it renames any core file to **core.last**.

This directory also contains the working directories of any resource monitors that are started by the Event Manager daemon. Each directory has the name of its resource monitor.

Finally, this directory contains the **Rcache_local** and **Rcache_remote** directories. These directories contain the registration cache for local and remote client registration requests, respectively.

For each EM client that establishes communications with the Event Manager daemon, a cache subdirectory is created in the appropriate registration cache directory. This subdirectory has a name of the form *p,s,n,q,i*, where:

- *p* is the process ID of the EM client
- *s* is the seconds portion of the timestamp recorded when the EM client established its session with the Event Manager daemon
- *n* is the nanoseconds portion of the timestamp
- *q* is a sequence number for the session
- *i* is the IP address of the host where the EM client is running. For local clients, the IP address is 0.0.0.0.

Within each subdirectory are several files with numeric names. Each file contains a registration request. The file name is the event command group ID of the group of events within the request.

The /etc/ha/cfg Directory (Run-Time EMCDB and Related Files)

The **/etc/ha/cfg** directory contains the run-time EMCDB file for each system partition. The file is called **/etc/ha/cfg/em.***syspar_name***.cdb**, where *syspar_name* is the name of the system partition.

In addition, there is a file called **em.***syspar_name***.cdb_vers**, where *syspar_name* is the name of the system partition. This file is created by the Event Manager daemon and contains the version string of the EMCDB file used by the daemon. This file is also used by the RMAPI to ensure that it is using the same EMCDB as the Event Manager daemon.

Components on Which Event Management Depends

The Event Management subsystem depends on the following components:

SPMI The System Performance Monitor Interface (SPMI), a library that is included with Performance Aide for AIX, which is a component of the IBM Performance Toolbox for AIX 2.2 product. This library is used by the RMAPI and the Event Manager daemon.

Group Services Another PSSP subsystem. It provides high availability membership services for coordinating activities on multiple nodes that are used by the Event Management subsystem.

Host_Responds daemon
A daemon that supplies information to the Event Management subsystem for compatibility with earlier releases of PSSP.

System Data Repository (SDR)
A repository of SP system configuration information that includes the configuration information for the Event Management subsystem.

The sections that follow contain more details about each of these components.

The System Performance Monitor Interface (SPMI) Library

The System Performance Monitor Interface (SPMI) library resides at **/usr/lib/libSpmi.a**. The RMAPI uses the shared memory technology of this library to deliver resource variable values to the Event Manager daemon. In addition, the data that is placed in shared memory can also be used by the Performance Toolbox for AIX product and the Performance Toolbox Parallel Extensions (PTPE) optional feature of PSSP. Thus, a resource monitor supplies data not only for event management, but also for performance monitoring.

Also, the Event Manager daemon uses the SPMI to obtain AIX operating system statistics that are the source of the AIX operating system resource variables that are supplied by PSSP.

The Group Services Subsystem

The Group Services subsystem provides another set of high availability services. The Event Management subsystem primarily uses Group Services to monitor the state of each Event Manager daemon in the operational domain, as follows. Each Event Manager daemon in the domain joins a Group Services group called **ha_em_peers**. Group Services informs each Event Manager daemon, in a synchronized fashion, when a daemon has joined or left the group.

Associated with the **ha_em_peers** group is a group state that contains the version string of the EMCDB. Thus, when an Event Manager daemon joins the **ha_em_peers** group, it can determine the version of the EMCDB that the rest of the group is using.

The Event Manager daemon also subscribes to the **HostMembership**, **enMembership**, and **cssMembership** system groups. Instances of the **IBM.PSSP.Membership.Node.state** resource variable are derived from **HostMembership** group information. Instances of the **IBM.PSSP.Membership.LANAdapter.state** resource variable are derived from **enMembership** and **cssMembership** group information.

The Host_Responds Daemon

On the control workstation, the Event Manager daemon is a client of the Host_Responds daemon, which is configured and controlled by the **/usr/lpp/ssp/bin/hr** script. Instances of the **IBM.PSSP.Response.Host.state** resource variable are created from information from the Host_Responds daemon. This is the same information that the Host_Responds daemon places in the **host_responds** SDR class. You can use the **IBM.PSSP.Response.Host.state** resource variable to obtain the status of any node in the cluster, even if the node is running a level of PSSP that is earlier than PSSP 2.2.

The System Data Repository (SDR)

The Event Manager daemon uses the SDR primarily as a repository of configuration information. In addition, on the control workstation, the Event Manager daemon polls the SDR for information that it uses to create instances of the **IBM.PSSP.Response.Switch.state** resource variable. This information is taken from the **switch_responds** SDR class.

Event Management Configuration and Operation

The following sections describe how the components of the Event Management subsystem work together to provide event management services. Included are discussions of:

- Event Management configuration
- Event Manager daemon initialization
- Event Management operation.

Event Management Configuration

The Event Management subsystem is installed as part of the installation of the PSSP product. The Event Management subsystem is contained in the **ssp.ha** fileset. The EMAPI libraries are contained in the **ssp.clients** fileset.

After the components are installed, the subsystem must be configured for operation. Event Management configuration is performed by the **haemctrl** command, which is invoked by the **syspar_ctrl** command.

The **syspar_ctrl** command configures all of the system partition-sensitive subsystems. The person who installs PSSP issues the **syspar_ctrl** command during installation of the control workstation. The **syspar_ctrl** command is executed automatically on the nodes when the nodes are installed. The **syspar_ctrl** command is also executed automatically when system partitions are created or destroyed. For more information on using the **syspar_ctrl** command, see *PSSP Installation and Migration Guide* and *PSSP Command and Technical Reference*.

The **haemctrl** command provides a number of functions for controlling the operation of the Event Management system. You can use it to:

- Add (configure) the Event Management subsystem
- Start the subsystem
- Stop the subsystem
- Delete (unconfigure) the subsystem
- "Clean" all Event Management subsystems
- Turn tracing of the Event Manager daemon on or off

Except for the clean function, **haemctrl** affects the Event Management subsystem in the current system partition, that is, the system partition that is specified by the **SP_NAME** environment variable.

Adding the Subsystem

If the **haemctrl** command is running on the control workstation, the first step in the add function is to select an Event Manager daemon communications port number and save it in the **Syspar_ports** SDR class. This port number is then placed in the /**etc**/**services** file. If the **haemctrl** command is running on a node, the port number is fetched from the SDR and placed in the /**etc**/**services** file. Port numbers are selected from the range 10000 through 10100.

The second step is to add the Event Manager daemon to the System Resource Controller (SRC) using the **mkssys** command. On the control workstation, the IP address of the system partition is an argument to the **haemd** program in the SRC subsystem specification.

The third step is to add an entry to the /**etc**/**inittab** file so that the Event Manager daemon will be started during boot. However, if **haemctrl** is running on a High Availability Control Workstation (HACWS), no entry is made in the /**etc**/**inittab** file. Instead, HACWS manages starting and stopping the Event Manager daemon.

The remaining steps in the add function are performed only on the control workstation. The **haemloadcfg** program is run to load the default configuration data into the SDR. The **haemcfg** command is run to create the EMCDB and place it into the staging directory. Finally, if it is not already stored in the SDR, an Event Manager daemon remote client communications port number is selected from the range 10000 through 10100. This port number is then placed in the /**etc**/**services** file. This port number is used by all of the Event Manager daemons on the control workstation.

Note that if the **haemctrl** add function terminates with an error, the command can be rerun after the problem is fixed. The command takes into account any steps that already completed successfully.

Starting and Stopping the Subsystem

The start and stop functions of the **haemctrl** command simply run the **startsrc** and **stopsrc** commands, respectively. However, **haemctrl** automatically specifies the subsystem argument to these SRC commands.

Deleting the Subsystem

The delete function of the **haemctrl** command removes the subsystem from the SRC, removes the entry from /**etc**/**inittab**, and removes the Event Manager daemon communications port number from /**etc**/**services**. It does **not** remove anything from the SDR, because the Event Management subsystem may still be configured on other nodes in the operational domain.

Cleaning the Subsystem

The clean function of the **haemctrl** command performs the same function as the delete function, except in all system partitions. In addition, it removes the Event Manager daemon remote client communications port number from the /**etc**/**services** file.

The clean function does **not** remove anything from the SDR. This function is provided to support restoring the system to a known state, where the known state is the (possibly restored copy of the) SDR database.

Tracing the Subsystem

The tracing function of the **haemctrl** command is provided to supply additional problem determination information when it is requested by the IBM Support Center. Normally, tracing should **not** be turned on, because it may slightly degrade Event Management subsystem performance and can consume large amounts of disk space in the /**var** file system.

Event Manager Daemon Initialization

Normally, the Event Manager daemon is started by an entry in the **/etc/inittab** file, via the **startsrc** command. If necessary, the Event Manager daemon can be started using the **haemctrl** command or the **startsrc** command directly.

During its initialization, the Event Manager daemon performs the following steps:

1. It gets the number of the node on which it is running using the **/usr/lpp/ssp/install/bin/node_number** command. Node 0 is the control workstation.
2. It fetches the name of the system partition and the EMCDB version string from the **Syspar** SDR class. (Recall that one instance of the Event Manager daemon runs on the control workstation for each system partition to which the Event Management subsystem was added.) It also fetches the Event Manager daemon remote client communications port number from the **SP_ports** SDR class.
3. It performs actions that are necessary to become a daemon. This includes establishing communications with the SRC subsystem so that it can return status in response to SRC commands.
4. It removes from the registration cache all of the subdirectories for local EM clients that no longer exist. That is, if the process ID in the subdirectory name cannot be found, it removes the subdirectory.

   Note that subdirectories for remote clients cannot be removed automatically, because the Event Manager daemon cannot determine if remote processes still exist.
5. It tries to connect to the Group Services subsystem. If the connection cannot be established because the Group Services subsystem is not running, it is scheduled to be retried in 5 seconds. This continues until the connection to Group Services is established. Meanwhile, Event Manager daemon initialization continues.
6. It enters the main control loop.

   In this loop, the Event Manager daemon waits for requests from EM clients, messages from resource monitors and other Event Manager daemons, messages from the Group Services subsystem, and requests from the SRC for status. It also waits for internal signals that indicate a function that was previously scheduled should now be executed, for example, retrying a connection to Group Services.

   However, EM client requests, messages from resource monitors, and messages from other Event Manager daemons (called peers) are refused until the Event Manager daemon has successfully joined the daemon peer group (the **ha_em_peers** group) and has fetched the correct version of the EMCDB.

Joining the Peer Group

After the Event Manager daemon has successfully established a connection with the Group Services subsystem, it tries to join the daemon peer group, a Group Services group called **ha_em_peers**. If this is the first Event Manager daemon to come up in the operational domain, it establishes the peer group. Otherwise, the other daemons in the peer group either accept or reject the daemon's join request. If an existing peer group member is still recovering from a prior termination of the joining daemon, the join request is rejected. If its join request is rejected, the

daemon tries to join again in 15 seconds. This continues until the daemon's join request is accepted.

When it joins the daemon peer group, the Event Manager daemon examines the group state. The group state is the EMCDB version string.

If the group state is null, the joining daemon proposes that the group state be set to the version string that the daemon has fetched from the SDR. If several daemons try to join the group at about the same time, and the group state is null, then each daemon proposes the group state. When the group is formed, Group Services selects one of the proposals and sets the group state to it. Note that each daemon is proposing the EMCDB version string that it has fetched from the SDR. Unless the **haemcfg** command has been run at about the same time, the proposed version strings should be identical.

If the group state is not null when it is examined by the joining daemon, a group has already formed and the daemon does **not** propose a new group state.

After the daemon has successfully joined the peer group, it compares the EMCDB version string contained in the group state to the version string it fetched from the SDR. If they are different, the version that was fetched from the SDR is replaced by the version in the group state.

An Event Manager daemon is prevented from joining the peer group as long as any other Event Manager daemon, currently in the peer group, is non-responsive to "pings" from the Group Services subsystem. (When an Event Manager daemon successfully joins the peer group, Group Services requests a response from the Event Manager daemon every two minutes. If the daemon does not respond to the request within two minutes, it is considered to be non-responsive. A daemon is also considered to be non-responsive if it does not reply to the join requests of other daemons within one minute.) The Event Manager daemon status, as displayed by the **lssrc** command, indicates if a daemon cannot join the peer group. If this is the case, the **em.default.syspar_name** file of any other daemon in the peer group should be examined for errors indicating that an Event Manager daemon is non-responsive. If so, and the non-responsive daemon does not terminate itself within a few minutes, perform the User Response specified for the error.

Reading the EMCDB

Once the daemon has joined the peer group and has determined the EMCDB version, it reads the run-time EMCDB file from the **/etc/ha/cfg** directory. If the file does not exist, it is copied from the staging directory on the control workstation.

Once the daemon has read the file, it compares the version string in the EMCDB to the one it fetched (from the SDR or from the group state). If the two version strings do not match, and the daemon has not just copied the EMCDB from the control workstation, then it copies the run-time EMCDB from the control workstation. If the version strings still do not match, the daemon terminates with an error.

Whenever it is necessary to copy the EMCDB from the control workstation, the EMCDB version string is used to determine how the copy is done. If the EMCDB version string was obtained from the group state then it is used to copy a back level EMCDB from the staging directory on the control workstation. Otherwise, the staging file **/spdata/sys1/ha/cfg/em.syspar_name.cdb** is copied. Note that back

level copies of the EMCDB should be removed from the staging directory only if their version string suffix indicates a time stamp older than the current version string found in the daemon peer group state (the current version string is found in the Event Manager daemon status, as displayed by the **lssrc** command. See "Displaying the Status of the Event Manager Daemon" on page 382.

After the daemon has read and validated the EMCDB, it enables daemon communications. This permits EM clients to send requests to the daemon, command-based resource monitors to connect to the daemon, and peers to send messages to the daemon. At this point, the initialization of the Event Manager daemon is complete.

To copy the EMCDB from the control workstation to the **/etc/ha/cfg** directory, the Event Manager daemon uses the **/usr/lpp/ssp/install/bin/haemrcpcdb** script. This script uses the **rcp** command to perform the actual copy.

The way in which Event Manager daemons determine the EMCDB version has important implications for the configuration of the subsystem. To place a new version of the EMCDB into production (that is, to make it the run-time version that is used by the Event Management subsystem), you must stop each Event Manager daemon in the operational domain after the **haemcfg** command is run. Stopping the daemons dissolves the existing peer group. Once the existing peer group is dissolved, the daemons can be restarted. As they restart, the daemons form a new peer group. A new EMCDB version string can be submitted as the group state only when a peer group is formed.

EM Daemon Operation

Normal operation of the Event Management subsystem requires no administrative intervention. The subsystem recovers from temporary failures automatically. However, there are some operational characteristics that may be of interest to administrators.

For performance reasons, the Event Manager daemon has an internal limit of 256 open file descriptors. In practice, this limits the number of EM client sessions, either local or remote, to about 225. This file descriptor limit is per daemon; it does not limit the number of EM clients in the operational domain.

Once the Event Manager daemon has successfully connected to a resource monitor of type **server**, the daemon attempts to reconnect to the resource monitor if it should terminate. The reconnection is attempted at the rate of once per minute. However, reconnection attempts are constrained under the following circumstances:

1. If it is necessary that the daemon start the resource monitor before each reconnection attempt, then after three attempts within two hours the resource monitor is "locked" and no further attempts are made.
2. If the resource monitor is not startable by the Event Manager daemon, then after about three successful reconnections within two hours the resource monitor is "locked" and no further attempts are made.

The rationale for locking the resource monitor is that, if it cannot be started and stay running or successful connections are frequently being lost, then a problem exists with the resource monitor. Once the problem has been determined and corrected, the **haemunlkrm** command can be used to unlock the resource monitor and start it. Note that locking does not apply to **client** type resource monitors.

The primary function of the Event Manager daemon is to generate events, by observing resource variable values and applying predicates to those values. However, this function is performed for a resource variable only if an EM client has registered to receive events for that resource variable. The Event Manager daemon also observes a resource variable once to satisfy a query request, if the resource variable is not already being observed. When observations are necessary, the Event Manager daemon commands the appropriate resource monitor (if it has a connection type of **server**) to supply resource variable values. When observations are no longer necessary, the Event Manager daemon commands the resource monitor to stop supplying values. In this way, the Event Manager daemon performs no action for resource variables that are not of interest to clients.

Even if a resource monitor that has a connection type of **server** is running, it does not supply data to the Event Manager daemon except by command of the daemon.

The Event Manager daemon either observes a resource variable located in shared memory every X seconds, where X is the observation interval that is specified in the resource variable's resource class definition, or when the resource variable's value is sent to the Event Manager daemon by the resource monitor (transparently, via the RMAPI). The values of resource variables of value type Counter and Quantity are located in shared memory. The values of resource variables of value type State are not.

All resource variables that are located in shared memory with the same observation interval are observed on the same time boundary. This minimizes the observation overhead, no matter when the request for a resource variable is made.

Event Management Procedures

For the most part the Event Management subsystem runs itself without requiring administrator intervention. However, on occasion, you may need to check the status of the subsystem, or add or change some of the configuration data.

This section contains the procedures that you need to do these tasks, which include:

- Displaying the status of the Event Manager daemon
- Loading non-PSSP configuration data into the SDR
- Changing configuration data in the SDR
- Activating the configuration data in the SDR

Displaying the Status of the Event Manager Daemon

You can display the operational status of the Event Manager daemon by issuing the **lssrc** command.

On the control workstation, enter:

```
lssrc -l -s haem.syspar_name
```

where *syspar_name* is the name of the system partition of interest.

On a node, enter:

```
lssrc -l -s haem
```

In response, the **lssrc** command writes the status information to standard output. The information includes:

- The information provided by the **lssrc -s haem** command (short form)
- The names of any trace flags that are set

  For information on these flags, see the **haemtrcon** command in *PSSP Command and Technical Reference*.
- The EMCDB version string and an indication as to whether the version string is taken from the SDR or from the peer group state.
- The day and time the Event Manager daemon was started.
- A report on the daemon's progress through initialization:

  ```
  Daemon connected to group services: TRUE/FALSE
  Daemon has joined peer group:       TRUE/FALSE
  Daemon communications enabled :     TRUE/FALSE
  ```
- A count of the peer daemons that are currently in the peer group. The count does not include this daemon.
- The number and type of EM client connections. Note that when a daemon relays a request to another daemon, the sending daemon is treated as a client by the receiving daemon.
- A list of each resource monitor that is defined in the EMCDB, its connection type (**server**, **client**, or **internal**), and its connection status.

  The connection status is indicated by a file descriptor (FD). If the file descriptor is greater than or equal to 0, a connection is open. The process ID of the resource monitor is also listed; it is interpreted as follows:

  ID greater than 0 — Resource monitor has been successfully started by the Event Manager daemon

  ID equal to 0 — A resource monitor started by the daemon has terminated (or the resource monitor forked, the parent process exited and the child process is the actual monitor)

  ID equal to -1 — The resource monitor has never been started by the Event Manager daemon

  ID equal to -2 — The resource monitor is not startable by the Event Manager daemon.

  In addition, the Resource Monitor Information indicates whether or not a resource monitor is locked.
- The highest file descriptor in use.
- The peer daemon status.

  This lists the status of peer daemons by node number, starting with node 0, and up to and including the highest node number seen for a peer. Note that this may include node numbers that are not in the operational domain of this Event Manager daemon.

Following the node number are two characters. If both characters are **S**, the specified node is the number of the node on which this daemon is running. Otherwise, the characters can take on values as follows.

The first character is **I**, **O**, or **-**, where:

- **I** indicates that the peer on the specified node is a peer group member.
- **O** indicates that the peer is no longer a peer group member (but was at one time).
- **-** indicates that no peer on the specified node has ever joined the group.

The second character is either **A** or **R**, where:

- **A** indicates that this daemon is accepting join requests from the peer on the specified node.
- **R** means this daemon is rejecting join requests.

- A list of internal daemon counters, for use by IBM service personnel.

Loading Non-PSSP Configuration Data into the SDR

The default configuration data supplied with PSSP is normally loaded into the SDR and compiled into its binary format automatically by the **haemctrl** script. However, if resource monitors supplied by other IBM products or by third parties are installed on the SP system, you must load the configuration data supplied with the resource monitors into the SDR and activate it. To do this:

1. Login to the control workstation. Use an ID that has **root** authority.
2. Create a file in load list format with the data to be loaded, or identify the path name of the file that has been supplied.

   If you are creating a new file, use the format specified in the man page for the **haemloadlist** file.
3. Set the **SP_NAME** environment variable to the appropriate system partition name.
4. Load the data into the SDR using the **haemloadcfg** command. Enter:

   haemloadcfg *new_loadlist*

   where *new_loadlist* is the path name of the load list you previously created or identified.
5. Activate all of the Event Management data in the SDR, including your new data, using the procedure in “Activating the Configuration Data in the SDR” on page 385.

Changing Configuration Data in the SDR

With an optional flag, the **haemloadcfg** command can replace existing objects (identified by their key attributes) in the SDR. If you want to change an object that already exists, do the following:

1. Login to the control workstation. Use an ID that has **root** authority.
2. Change the object's attribute in the load list file.

   For PSSP configuration data, this is the default configuration data in the **/usr/lpp/ssp/install/config/haemloadlist** file. If you have non-PSSP or

third-party resource monitors, the object will be loaded from another load list file.

Do this step for as many objects as you are changing.

3. Set the **SP_NAME** environment variable to the appropriate system partition name.
4. Run the **haemloadcfg** command, specifying the **-r** flag and the changed load list file.

   In response, the **haemloadcfg** command replaces each object in the SDR matched by an object in the load list file.
5. Activate all of the Event Management data in the SDR, including your changed data, using the procedure in "Activating the Configuration Data in the SDR."

Activating the Configuration Data in the SDR

When you have added or changed the Event Management data in the SDR, you must activate it by recompiling the EMCDB and stopping and restarting the Event Manager daemons. To do this:

1. Login to the control workstation. Use an ID that has **root** authority.
2. Set the **SP_NAME** environment variable to the appropriate system partition name.
3. Recompile the EMCDB. Enter:

   ```
   haemcfg
   ```

   The output is placed in the staging directory in a file whose name indicates the system partition.
4. Stop the Event Manager daemons in this system partition.

   Issue the **haemctrl -k** command on the control workstation and on each of the nodes in the system partition.

   You can use the **dsh** or **Sysctl** commands to run the command on multiple nodes from the control workstation. For more information on using these commands, see Chapter 4, "Parallel Management Commands" on page 31.
5. Verify that all of the Event Manager daemons in this system partition have stopped.

   On the control workstation, issue the **lssrc -s haem.***syspar_name* command. On the nodes, issue the **lssrc -s haem** command.

   You can use the **dsh** or **Sysctl** commands to run the command on multiple nodes from the control workstation. For more information on using these commands, see Chapter 4, "Parallel Management Commands" on page 31.

   The status of each daemon should indicate that it is inactive.
6. Restart the Event Manager daemons in this system partition.

   Issue the **haemctrl -s** command on the control workstation and on each of the nodes in the system partition.

   You can use the **dsh** or **Sysctl** commands to run the command on multiple nodes from the control workstation. For more information on using these commands, see Chapter 4, "Parallel Management Commands" on page 31.

Chapter 25. Using the Problem Management Subsystem

The Problem Management subsystem (**pman**) provides an infrastructure for recognizing and acting on problem events in your SP system. This infrastructure is based on an Event Management application that provides configurable access to Event Management client and resource monitor function without the necessity of writing C programs that use the Event Management APIs.

To understand the information presented in this chapter, you should be familiar with Event Management concepts and terminology. Some important Event Management terms are:

event In Event Management, the notification that a predicate evaluated to true. This evaluation occurs each time an instance of a resource variable is observed.

instance vector In Event Management, a set of elements, where each element is a name/value pair of the form `name=value`, whose values uniquely identify the copy of the resource (and by extension, the copy of the resource variable) in the system.

predicate In Event Management, the relational expression between a resource variable and other elements (such as constants or the previous value of an instance of the variable) that, when true, generates an event. An example of a predicate is `X < 10` where X represents the resource variable `IBM.PSSP.aixos.PagSp.%totalfree` (the percentage of total free paging space). When the predicate is true, that is, when the total free paging space is observed to be less than 10%, the Event Management subsystem generates an event to notify the appropriate application.

rearm predicate In Event Management, a predicate used to generate an event that alternates with an original event predicate in the following way: the event predicate is used until it is true, then the rearm predicate is used until it is true, then the event predicate is used, and so on. The rearm predicate is commonly the inverse of the event predicate (for example, a resource variable is on or off). It can also be used with the event predicate to define an upper and lower boundary for a condition of interest.

resource In Event Management, an entity in the system that provides a set of services. Examples of resources include hardware entities such as processors, disk drives, memory, and adapters, and software entities such as database applications, processes, and file systems. Each resource in the system has one or more attributes that define the state of the resource.

resource monitor A program that supplies information about the resources in the system. It can be a command, a daemon, or part of an application or subsystem that manages any type of system resource.

resource variable In Event Management, the representation of an attribute of a resource. An example of a resource variable is `IBM.AIX.PagSp.%totalfree`, which represents the percentage of

© Copyright IBM Corp. 1995, 1997

total free paging space. `IBM.AIX.PagSp` specifies the resource name and `%totalfree` specifies the resource attribute.

For more information on Event Management services, refer to *IBM Parallel System Support Programs for AIX: Event Management Programming Guide and Reference*.

The major components of the Problem Management subsystem are:

- A Problem Management daemon (**pmand**) that provides access to events generated by Event Management on all nodes. In addition, the following function is provided to work in conjunction with **pmand**:
  - A command, **pmandef**, that provides for subscribing to Event Management events and associating actions for those events
  - A command, **pmanquery**, that queries the SDR for a description of a Problem Management subscription
  - A script, **pmandefaults**, that causes **pmand** to register for a set of default events
  - A script, **notify_event**, that mails an event notification when an event occurs
  - A script, **log_event**, that logs a record of an event to a regular wraparound file
- A resource monitor daemon, **pmanrmd**, that provides resource variables to Event Management. The following function is provided to work in conjunction with **pmanrmd**:
  - Sixteen resource variables. The resource variables are named **IBM.PSSP.pm.User_state1** through **IBM.PSSP.pm.User_state16**.
  - A command, **pmanrminput**, that provides for associating values with the sixteen supplied resource variables.
  - A sample file, **pmanrmd.conf**, for configuring the **pmanrmd** daemon.
  - A command, **pmanrmdloadSDR**, for loading the configuration information into the System Data Repository.

This chapter provides information on each of the components of Problem Management.

Understanding the Problem Management Daemon

The **pmand** daemon is a client of Event Management; it can be configured to register for Event Management events and perform actions when those events occur.

Event Management provides access to events throughout an SP system partition; therefore, **pmand** can monitor and react to events on the node on which it is running as well as on all other nodes in the system partition and the control workstation.

When you install your system, a **pmand** daemon is automatically configured on each node in a system partition. Additionally, there is a **pmand** daemon running on the control workstation for each system partition in the SP system that is running Event Management. When running on a node, the **pmand** daemon:

- Monitors events occurring on the node on which the daemon is running
- Monitors events on all other nodes in the system partition
- Monitors events not associated with a node, such as frame events, as supplied by Event Management

There are no restrictions on what a **pmand** daemon can monitor:

- Any number of **pmand** daemons can monitor and act on a single event
- A single **pmand** daemon can monitor any number of events locally or remotely
- A single **pmand** daemon can monitor the same event multiple times and all the actions associated with all the event registrations are taken by the daemon when the event occurs

In Figure 66 **pmand** daemons are running on a 4 node SP system partition.

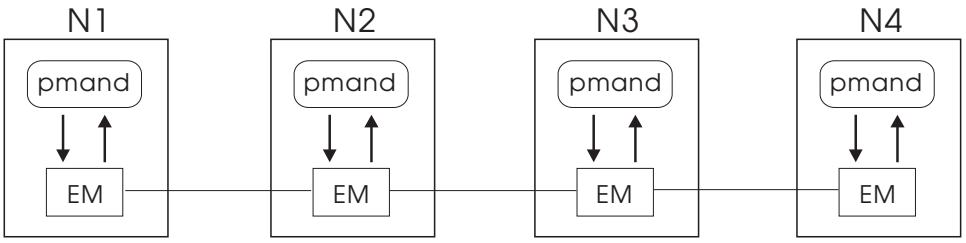


Figure 66. An Example of Problem Management Daemon Configuration

Each **pmand** daemon has access to events on the node on which it is running, as well as to the other nodes in the system partition. The access to the events is provided by Event Management, which is able, due to its distributed nature, to monitor resource variables throughout the system partition and to generate events based on the values of the resource variables. When an event that any of the daemons has subscribed to occurs (whether that event is local or remote), all the **pmand** daemons registered for it will perform the actions they are configured for (if any).

Because **pmand** is a daemon, its subscriptions to events are persistent. That is, the daemon continues to subscribe to events even after the process or user who created the subscription has gone away. A system administrator, for example, can set up automated operations with unattended monitoring and recovery actions.

Controlling pmand

The **pmand** daemon is under the System Resource Controller (SRC) and can be controlled by the following commands:

- **startsrc -s pman** starts **pmand**
- **stopsrc -s pman** stops **pmand**
- **refresh -s pman** causes **pmand** to update its internal configuration from the SDR and start pairing actions with events as specified by the **pmandef** command. If a **pmand** refresh occurs, all currently monitored events will be unregistered from before the configuration information is reread from the SDR. The SDR contains persistent information, so that a refresh results only in configuration changes that have been put into the SDR. If you haven't deleted

or modified a configuration record for a particular event, refreshing the daemon results in reregistering for the same event.

- **lssrc -ls** provides status on the **pmand** daemon. To check the status of **pmand** running on a node, enter:

  ```
  lssrc -ls pman
  ```

  To check the status of **pmand** running on the control workstation, enter:

  ```
  lssrc -ls pman.system_partition_name
  ```

  This command provides the following status information:

  - When **pmand** was started.
  - When **pmand** was last refreshed.
  - Whether tracing (debug mode) is on or off. When debug mode is on, all SRC requests and all events are logged to the **/var/adm/SPlogs/pman** directory.
  - Events for which registrations are as yet unacknowledged.
  - Events for which actions are currently being taken.
  - Events currently ready to be acted on by this daemon.

Creating Problem Management Subscriptions

The **pmandef** command is the mechanism provided for creating Problem Management subscriptions for the **pmand** daemon to Event Management services. The **pmandef** command provides for defining:

- Event Manager events to register for
- Actions to take when those events occur:
  - Run a command
  - Issue an SNMP trap
  - Write to the AIX Error Log or BSD syslog facilities

The **pmandef** command also provides for:

- Activating a Problem Management subscription
- Deactivating a Problem Management subscription
- Querying a Problem Management subscription
- Removing a Problem Management subscription

For more information on **pmandef**, refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Security Considerations

The **pmandef** command is based on Sysctl, which uses Kerberos for user authentication. All users of the **pmandef** command must have valid Kerberos credentials. In addition, the user's Kerberos principal must be listed in the **/etc/sysctl.pman.acl** file on the local node, in order to store the subscription in the SDR, as well as on all the nodes that are affected by the new subscription, in order for the affected Problem Management daemons to be notified of the new subscription. If the user's Kerberos principal is listed only in the **/etc/sysctl.pman.acl** file on the local node, the subscription will be stored in the SDR, but the Problem Management daemons will not act on the new subscription until the next time they are restarted.

The user's Kerberos principal is more than just a mechanism to validate access to the Problem Management subsystem. The Kerberos principal that was in effect when the subscription was created, via the **pmandef -s** command, is stored as part of the subscription definition, and it is used to establish the ownership of that subscription. Modifications to the subscription, via the **pmandef** command with the **-u**, **-d** and **-a** flags, are allowed only by a user with the same Kerberos principal that is stored in the subscription definition. The Kerberos principal that is stored in the subscription definition is also used by the Problem Management daemon to decide whether the action that is to result from the occurrence of an event should be allowed.

After **pmand** receives notification that an event has occurred, and before it performs the action for that event, it checks to see whether the Kerberos principal that is stored in the subscription definition is authorized to perform the requested action on the node where **pmand** is running. If the requested action is to write an entry in the AIX Error Log and BSD syslog or to generate an SNMP trap, then the Kerberos principal that owns the subscription must have been listed in the root user's **$HOME/.klogin** file. If the requested action is to execute a command, then the Kerberos principal must be listed in the **$HOME/.klogin** file of the user which will be used to run the command. The user which will be used to run the command is by default the same user who issued the **pmandef -s** command to create the subscription. A different user may be specified to **pmandef** by using the **-U** flag.

Running a Command

Use **pmandef** to specify a command to run when a specified event or rearm event occurs. For example:

```
pmandef -s Program_Monitor \
-e 'IBM.PSSP.Prog.pcount:NodeNum=12;ProgName=mycmd;UserName=bob:X@0==0' \
-r "X@0==1" -c "echo program has stopped >/tmp/myevent.out" \
-C "echo program has restarted >/tmp/myrearm.out"
```

Running this example on node 5 causes the command **echo program has stopped >/tmp/myevent.out** to run on node 5 whenever the number of processes named **mycmd** and owned by user **bob** on node 12 becomes 0 (the event). When this number increases back to 1 (the rearm event), the command **echo program has restarted >/tmp/myrearm.out** runs on node 5.

You can specify if you want the command to run on a node other than the one from which the **pmandef** command was issued. For example:

```
pmandef -s Program_Monitor \
-e 'IBM.PSSP.Prog.pcount:NodeNum=1-5,13;ProgName=mycmd;UserName=bob:X@0==0' \
-r "X@0==1" -c /usr/local/bin/start_recovery \
-C /usr/local/bin/stop_recovery -n 1-3,7
```

This example causes the commands to run on nodes 1, 2, 3 and 7, whenever **bob**'s program dies or gets restarted on any of nodes 1, 2, 3, 4, 5 or 13. If **bob**'s program dies on node 4, then the command **/usr/local/bin/start_recovery** runs on nodes 1, 2, 3 and 7.

Any number of commands can run simultaneously.

You can specify a timeout, in seconds, for each command. The minimum timeout that can be specified is 10 seconds. If the command has not exited before the specified timeout, the command is killed.

For information on command termination status, use the **lssrc -ls** command.

The Command Environment: The Problem Management subsystem makes all of the contents of an Event Management notification available in the command's environment when the command is run:

PMAN_HANDLE
The name that identifies this subscription to the Problem Management subsystem. This name was given as the argument of the **-s** flag to **pmandef**.

PMAN_PRINCIPAL
The name of the Kerberos principal that owns this subscription.

PMAN_RVNAME
The Event Management resource variable.

PMAN_IVECTOR
The Event Management instance vector.

PMAN_PRED Either the Event Management predicate or rearm predicate, depending on whether this is an event or rearm event.

PMAN_TIME The time that the event was reported to the Problem Management subsystem.

PMAN_LOCATION
The node number of the node on which the event was generated, usually (but not always) the node on which the event occurred.

PMAN_RVTYPE
One of **long**, **float** or **sbs**, depending on whether the type of the resource variable value is a long integer, a floating point value or a Structured Byte String.

Note: If the PMAN_RVTYPE is either **long** or **float**, then the resource variable value is stored in PMAN_RVVALUE, and PMAN_RVVALUE is to be interpreted as type PMAN_RVTYPE.

Note: If PMAN_RVTYPE is **sbs**, then the resource variable value is composed of one or more structure fields. There is no PMAN_RVVALUE environment variable. Instead there is a separate environment variable for each field, and the PMAN_RVCOUNT environment variable defines the number of fields. For example, if there are 3 structure fields within the Structured Byte String, the PMAN_RVCOUNT will be 3, and there will be 3 separate environment variables for the 3 structure fields: PMAN_RVFIELD0, PMAN_RVFIELD1 and PMAN_RVFIELD2. Each of these 3 environment variables contains a *name=value* pair, where *name* is the structure field name, and *value* is the structure field value.

For example, the following command:

```
pmandef -s example \
-e 'IBM.PSSP.Prog.pcount:NodeNum=9-11;ProgName=mycmd;UserName=root:X@0==0' \
-c "/usr/local/bin/recovery_cmd" -n 12
```

requests the **/usr/local/bin/recovery_cmd** command to run on node 12, when the number of processes named **mycmd** and owned by root on nodes 9, 10, or 11 becomes zero. If the **mycmd** program terminates on node 10, the command **/usr/local/bin/recovery_cmd** runs on node 12, and the following environment variables are included in its environment:

- PMAN_HANDLE (**example**)
- PMAN_PRINCIPAL (**root.admin@PPD.POK.IBM.COM**)
- PMAN_RVNAME (**IBM.PSSP.Prog.pcount**)
- PMAN_IVECTOR (**ProgName=mycmd;UserName=root;NodeNum=10**)
- PMAN_PRED (**X@0==0**)
- PMAN_TIME (**Thu Aug 22 00:42:08 1996**)
- PMAN_LOCATION (**10**)
- PMAN_RVTYPE (**sbs**)
- PMAN_RVCOUNT (**3**)
- PMAN_RVFIELD0 (**CurPIDCount=0**)
- PMAN_RVFIELD1 (**PrevPIDCount=1**)
- PMAN_RVFIELD2 (**CurPIDList=**)

This information could be used by any command. Two utilities that report this information are provided as part of the Problem Management subsystem: **notify_event** and **log_event**. (These commands are provided to get you started; you may want to write more sophisticated commands.)

notify_event captures event information and mails it to the user running the command on the local node.

log_event captures event information and logs it to a wraparound file. The syntax for **log_event** is:

```
/usr/lpp/ssp/bin/log_event log_filename
```

log_event uses the AIX 4.1 **alog** command to write to a wraparound file. The size of the wraparound file is limited to 64K. The **alog** command must be used to read the file. Refer to the AIX 4.1 **alog** man page for more information on this command.

Issuing an SNMP Trap

Use the **pmandef** command to subscribe to an Event Management event and specify that an SNMP trap be issued for that event. For example:

```
pmandef -s Filesystem_Monitor \
-e 'IBM.PSSP.aixos.FS%totused:NodeNum=10;VG=myvg;LV=mylv:X>95' \
-t 1234 -n 10
```

In this example, whenever the file system associated with the **mylv** logical volume and **myvg** volume group on node 10 becomes more that 95% full, an SNMP trap will be generated on node 10.

For complete information on how **pmand** can be configured to issue an SNMP trap when an event for which it is registered occurs, refer to Chapter 26, "Managing SP System Events in a Network Environment" on page 399.

Logging an Event

You can specify that **pmand** write event notification information, along with some optional specified text, to the AIX Error Log and BSD syslog facilities. For example:

```
pmandef -s Filesystem_Monitor \
-e 'IBM.PSSP.aixos.FS%totused:NodeNum=11;VG=myvg;mylv:X>95' \
-l "filesystem is almost full" -h local
```

In this example, whenever the file system associated with the **mylv** logical volume and **myvg** volume group on node 11 becomes more than 95 % full, the text

filesystem is almost full gets written to the AIX Error Log and BSD syslog facilities on node 11 (via the **-h local** option).

Requesting Problem Management Subscription Information

Use the **pmanquery** command to query the SDR for a description of a Problem Management subscription. The **pmanquery** command ouputs the details of the subscription information in raw format, which can then be used by other applications. The following example queries all subscriptions:

```
pmanquery -n all -k all
```

For more information on **pmanquery**, refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Monitoring Default Events

The Problem Management subsystem provides for a set of default events to be monitored in the **/usr/lpp/ssp/install/bin/pmandefaults** script. This script contains a series of **pmandef** commands that request an event notification to be mailed to root on the control workstation, when the specified event occurs. These are events that would interest many system administrators. Events are defined for all nodes in the current system partition, and all events are monitored from the control workstation. The list of events includes:

- The /**var** file system is more than 95 percent full.
- The /**tmp** file system is more than 95 percent full.
- An error log record of type **PERM** has been written to the AIX Error Log.
- The **inetd** daemon has terminated.
- The **sdrd** daemon has terminated (control workstation only).
- The **sysctld** daemon has terminated.
- The **hrd** daemon has terminated (control workstation only).
- The **fsd** daemon has terminated (nodes only).

This script is a suggested starting point for configuring the Problem Management subsystem on your SP system. You can choose to run the script as is, or you can make your own copy of the script and modify it to suit your needs, or you can choose not to run the script at all.

The script takes no arguments, and it should be executed as the **admin** Kerberos user (for example, **root.admin@PPD.POK.IBM.COM**). It should be executed once for each system partition; specify the system partition by setting the SP_NAME environment variable.

Understanding the Problem Management Resource Monitor Daemon

The **pmanrmd** daemon is a resource monitor daemon that provides resource variables to Event Management services. When you install your SP system, a **pmanrmd** daemon is automatically configured on each node in a system partition. Additionally, there is a **pmanrmd** daemon running on the control workstation for each system partition in the SP system.

Controlling pmanrmd

The **pmanrmd** daemon is under the System Resource Controller (SRC). Only the following SRC commands are available for controlling the **pmanrmd** daemon:

- **startsrc -s pmanrm** starts **pmanrmd**
- **stopsrc -s pmanrm** stops **pmanrmd**

No command is provided to refresh the **pmanrmd** daemon while it is running. To refresh the daemon, you must stop and start the daemon (using **stopsrc** and **startsrc**).

Creating Resource Monitors

The Problem Management subsystem provides sixteen resource variables, named **IBM.PSSP.pm.User_state1** through **IBM.PSSP.pm.User_state16**. You can associate values with the provided resource variables and configure **pmanrmd** to supply the values that you associate to Event Management services.

Associating Values with Problem Management Resource Variables

Two mechanisms are provided for associating values with the Problem Management resource variables:

1. The **pmanrminput** command
2. The **pmanrmd.conf** sample configuration file

The **/usr/lpp/ssp/bin/pmanrminput** command requires root authority to use. The command can be run from within scripts. It takes as arguments:

- The name of the SRC subsystem to which to communicate the request
- A string that consists of:
  - The name of one of the Problem Management supplied resource variables
  - A delimiter (+)
  - A string to provide as the value of the Problem Management supplied resource variable
  - Another delimiter (+)

For example, if you have a server application running on node 10 and you sometimes want to tell users on nodes 1-9 to stop using the server application, you could use the Problem Management subsystem to accomplish this. Create a Problem Management subscription on nodes 1-9 to run the **wall** command to inform the users about the state of the server application:

```
pmandef -s ready_or_not \
-e 'IBM.PSSP.pm.User_state3:NodeNum=10:X@0==“READY”' \
-r 'X@0!=“READY”' \
-c “wall Node 10 is ready” -C “wall Stop using node 10” -n 1-9
```

When the server application is ready for users, it uses the **pmanrminput** command to communicate its state to the Problem Management subsystem:

```
pmanrminput -s pman -a “IBM.PSSP.pm.User_state3+READY+”
```

The Problem Management subsystem passes the string **READY** to Event Management as the value of the resource variable **IBM.PSSP.pm.User_state3**, and that value satisfies the event definition, so the command **wall Node 10 is ready** runs on nodes 1-9. When the server application has a problem, it can use **pmanrminput** to communicate its state change:

```
pmanrminput -s pman -a "IBM.PSSP.pm.User_state3+NOT READY+"
```

The resource value **NOT READY** satisfies the rearm event definition, so the command **wall Stop using node 10** runs on nodes 1-9.

Another way to associate a value with one of the Problem Management resource variables is to provide the name of the resource variable and a command in a configuration file, based on the **pmanrmd.conf** sample file (for more information on the **pmanrmd.conf** sample file refer to "Configuring pmanrmd to Provide Resource Variables"). The standard output of the command will be provided to Event Management services as the value for the specified Problem Management resource variable.

Configuring pmanrmd to Provide Resource Variables

In the example in "Associating Values with Problem Management Resource Variables" on page 395, the resource, an application server, was able to communicate its state change directly to the Problem Management subsystem. However, this is not always possible. For example, if you want to run a command whenever the contents of the /**etc** directory on nodes 1-5 change, the /**etc** directory is the resource, and a directory is not capable of stating when it has changed. In this case, the Problem Management subsystem needs to be configured to periodically send the name and time stamp of the last changed file in the /**etc** directory as the resource variable value to Event Management. This can be accomplished by defining a resource monitor command to the Problem Management subsystem.

A sample **pmanrmd.conf** file is provided. Copy, rename, and edit this file to specify:

- The node(s) on which **pmanrmd** is to be configured
- The name of the Problem Management resource variable to provide via **pmanrmd**
- A sampling interval (in seconds) for each resource variable
- The command to run to produce the standard output that will be provided as the value for the Problem Management resource variable

For example:

```
TargetType=NODE_RANGE
Target=1-5
Rvar=IBM.PSSP.pm.User_state2
SampInt=600
Command="/bin/ls -tl /etc | /bin/head -2 | /bin/grep -v total"
```

In this example, the **pmanrmd** daemon will run the command /**bin**/**ls -tl** /**etc** | /**bin**/**head -2** | /**bin**/**grep -v total** every ten minutes, and it will send the output of that command as the value of the resource variable **IBM.PSSP.pm.User_state2** to Event Management. The output of this command is the name of the last file to be changed and its time stamp. You could then subscribe to changes in the /**etc** directory resource with the following command:

```
pmandef -s etc_file_changed \
-e 'IBM.PSSP.pm.User_state2:NodeNum=1-5:X@0!=X@P0' \
-c "echo somebody changed a file in /etc >/tmp/etc_change_log"
```

Whenever the most recent output of the /**bin**/**ls -tl** /**etc** | /**bin**/**head -2** | /**bin**/**grep -v total** command is different from the previous output, the event definition will be

satisfied, and the **“echo somebody changed a file in /etc >/tmp/etc_change_log”** command will get executed.

Loading pmanrmd Configuration Information into the SDR

pmanrmdloadSDR reads the configuration file and loads the information in the stanzas into the SDR. **pmanrmdloadSDR** runs on the control workstation.

The syntax for **pmanrmdloadSDR** is :

```
pmanrmdloadSDR configuration_file
```

After you load the configuration information into the SDR, refresh the affected **pmanrmd** daemons.

Chapter 26. Managing SP System Events in a Network Environment

IBM Parallel System Support Programs for AIX includes a Problem Management subsystem (refer to Chapter 25, “Using the Problem Management Subsystem” on page 387) that provides access to Event Management function without the necessity of writing C programs that use the Event Management APIs. The Problem Management subsystem gives you the ability to subscribe to Event Management events and to specify actions for those events. Issuing an SNMP trap in response to an event (which can come either from Event Management or the AIX Error Log) is one of the actions you can specify.

The ability to issue an SNMP trap in response to an event allows you to report problem events occurring in your SP system to a network manager existing on a remote node (a network manager application is not supplied with the SP). The Problem Management subsystem provides an SP SNMP proxy agent, **sp_configd**, (sometimes referred to as a ‘subagent daemon’) that runs on the control workstation and every SP processor node. The SP proxy agent provides the following functions:

1. A Management Information Base (MIB)
2. SNMP **GET** and **GET NEXT** command support that allows data in the MIB to be accessed by a network manager application
3. Creation and transmittal of SNMP traps to an installation-defined network manager application when the following events occur on the node on which the SP proxy agent is running:
   - A ‘cold start’ trap is issued when the agent is activated
   - An enterprise-specific trap is issued when an entry with an “alert=true” attribute is written in the AIX Error Log
   - An enterprise-specific trap is issued when an user-specified event is detected within Event Management services

This chapter provides overview and task information for making the SP part of a network management system. In such a system, for example, the SP presents error event information, in the form of SNMP traps, to a network management application, such as NetView for AIX, which displays and logs the trap, and may notify a problem management application, such as Trouble Ticket for AIX, of problem events occurring on the SP.

In order for the SP to be part of such a network management system, network managers must be notified when selected AIX Error Log entries are written and Event Management events occur. SP-specific configuration data is provided so that management applications can determine which nodes compose the SP system. You decide what Error Log and Event Management events will trigger manager notification.

We'll first discuss general concepts related to Network Management and then go on to present information specific to enabling the reporting of SP system events in a network environment. For further discussion of Simple Network Management Protocol and the Management Information Base, refer to *IBM AIX Version 4.1 Communications Programming Concepts*, SC23-2610.

© Copyright IBM Corp. 1995, 1997

Understanding Network Management

The Simple Network Management Protocol (SNMP) describes the means by which management data is packaged and transported through the network in order for SNMP agents and SNMP network managers to communicate. The Management Information Base (MIB) defines the structure of the data being transported.

Understanding SNMP

SNMP uses a client/server approach to management, defining two roles: Manager (client) and Agent (server).

The Manager (NetView for AIX is an example) oversees the overall network activity. Critical events in the form of SNMP traps are sent to this Manager to alert network operators of problems in the network. The Manager also acts as the central point for storing and displaying statistics maintained by the Agent (server). These statistics are maintained by the Agent in the form of MIB variables.

The Agent is responsible for reporting on and maintaining the data pertaining to a managed resource, such as a device or, in the case of the SP Agent, a set of configuration information identifying the nodes composing the SP system and tables of resource attributes within the SP which can be monitored. Agents can run on several different types of managed nodes.

Understanding the MIB

Every Agent supports a Management Information Base (MIB), a set of variables that represent the physical and logical resources of the managed systems or Agent. The MIB is not a database, in the sense of a monolithic collection of data, but rather it represents dynamic information. The values of the variables are maintained by different system functions such as the kernel, device drivers, or subsystems. The Agent obtains the variable values from these system functions.

The Manager can read variable values via these SNMP requests:

GET Requests the SNMP Agent to retrieve the value of the specified variable and return it to the requester (the Manager).

GET NEXT Requests the SNMP Agent to retrieve the value of the next variable, after the one specified in the request, and return it to the Manager. This is especially useful for retrieving tabular information, and multiple variables can be requested in it.

To summarize, the SNMP Manager determines the status of Agents by using the GET request to poll the Agents on a regular basis. However, if an Agent discovers an exceptional event, it will alert the Manager immediately, without waiting for a GET request, by sending a trap. The trap does not necessarily describe the problem in any detail; it simply informs the Manager that there is a change of state within the resources managed by the Agent, and the Manager may poll the Agent via an SNMP GET or GET NEXT request to determine more detail about the problem.

Enabling Network Management for the SP System

As we stated in the introduction to this chapter, in order for an SP to be part of a network management system, network managers must be notified of the AIX Error Log or Event Management events that you would like to centrally manage. A section of the SP proxy agent MIB (contained in the **/usr/lpp/ssp/config/snmp_proxy/ibmSPMIB.my** file) contains SP-specific SNMP configuration data, enabling Managers to view the SP as a system composed of processor nodes and a control workstation. In other words, there must be a way to create SNMP traps for SP error or problem events and there must be a MIB that contains configuration information specific to the SP system.

SNMP Traps for the SP from AIX Error Log Events

The **snmp_trap_gen** error notification method notifies the **sp_configd** daemon (or SP proxy agent) to create SNMP traps when selected types of errors are recorded in the AIX Error Log. (Note that the **trapgend** method from NetView for AIX also creates SNMP traps from AIX Error Log events. **trapgend** formats such events to be specifically compatible with SNA alerts. Only one of these methods should be used to avoid duplicate traps being generated for a single error.) The method and daemon run on each processor node and control workstation composing the SP system. In order for a trap to be created, using the **snmp_trap_gen** method, error log entries on each node and control workstation must contain an “alert=true” attribute. The following steps describe the flow of events for creating an SNMP trap for an AIX Error Log event:

1. An application or subsystem writes to the AIX Error Log facility.
2. If the Error Log entry contains an “alert=true” attribute, the **snmp_trap_gen** error notification method runs.
3. The **snmp_trap_gen** method uses the sequence number from the Error Log entry as input to the **errpt -a** command to obtain full information about the event.
4. **snmp_trap_gen** places the event information in a FIFO file, **/var/tmp/errlog_entry**
5. **sp_configd** reads the FIFO file, parses the information into objects within the **ibmSPErrlogVars** group of the **ibmSP** MIB and creates a trap from the objects whose instantiations contain non-null values.
6. **sp_configd** sends the trap to the AIX agent, **snmpd**, which in turn sends the trap to network managers specified in the **/etc/snmpd.conf** file existing on the node.

The following information is provided in SNMP traps from AIX Error Log events generated using the **snmp_trap_gen** method and the **sp_configd** daemon.

- The **enterprise** field contains the OID (object identifier) of the **sp_configd** subagent (this is the OID assigned to the **ibmSP** MIB)
- The **specific-trap** field contains the error ID from the Error Log entry (this is the convention expected by NetView for AIX when configuring events)
- When the corresponding information exists in the Error Log entry, the **variable bindings** field contains the following object values paired with their OIDs:
  - Error label
  - Error ID

- Error log entry time stamp
- Unique sequence number
- Machine ID parameter
- Node ID parameter
- Error class
- Error type
- Resource name
- Resource class
- Resource type
- Identifies the location code of a device
- Vital product data
- Error description
- Probable causes
- User causes
- User actions
- Install causes
- Install actions
- Failure causes
- Failure actions
- Detail data

Other fields in the trap are as specified by SNMP protocol.

Designating AIX Error Log Entries as Alertable

In order for **sp_configd** to pass any entries in the system error log as traps to the **snmpd** daemon for transmission to Network Managers, the Error Record Templates for the entries must have the specific "alert" field set to the value **true**. (Make sure you are authenticated as the root user.) This may be accomplished as follows:

1. On the system in which the **sp_configd** daemon is running, use the **errpt -t** command to list the templates for entries in the AIX Error Log.
2. Select the IDs of the entries for which you wish to have traps issued when they occur.
3. Create a file containing the following two lines for each entry to be designated alertable:

   ```
   = ID:
   alert = true
   ```

4. From the command line, enter:

   ```
   errupdate file
   ```

5. To verify that the alert status has been changed to "true" (or 1), enter:

   ```
   errpt -tF alert=1 | grep ID
   ```

   listing the error template. You should see the specified ID listed.

SNMP Traps for the SP from Event Management Events

The following steps describe the flow of events for creating an SNMP trap for an Event Management event:

Designating that an SNMP Trap is Issued for an Event Management Event

1. Use the **pmandef** command to subscribe to an Event Management event and specify that an SNMP trap be issued for that event. For example:

   ```
   pmandef -s Filesystem_Monitor /
   -e'IBM.PSSP.aixos.FS%totused:NodeNum=10;VG=myvg;LV=mylv:X>95' /
   -t 1234 -n 0
   ```

   (In this example, whenever the file system associated with the **mylv** logical volume and **myvg** volume group on node 10 becomes more that 95% full, an SNMP trap will be generated on the control workstation.)

2. **pmand** (the Problem Management Event Management client daemon) subscribes to the event as specified in the **pmandef** command. The event must occur on the local node on which **pmand** is running.
3. **pmand** writes the contents of the Event Manager subsystem-supplied event response and the user-specified event configuration information into a FIFO file
4. **sp_configd** reads the data and creates an SNMP trap from it
5. **sp_configd** sends the trap to the SNMP managers specified in the **/etc/snmpd.conf** file on the node.
6. You specify a particular trap ID in the configuration information for an Event Manager event.

The following information is provided in SNMP traps from Event Management events generated by the **sp_configd** daemon.

The Contents of SNMP Traps Containing Event Management Events

- The **enterprise** field contains the OID of the **ibmSP** MIB
- The **specific-trap** field contains the trap ID field from the **pmand** configuration for this event. This allows configuration of NetView for AIX actions based on the type of event (which is done based on the specific trap ID)
- The **variable bindings** field contains the following object values paired with their OIDs:
  - The event ID
  - Event flags
  - A time stamp indicating the time the Event Manager generated the event
  - The number of the node and the system partition address where the event occurred
  - The resource variable name
  - MIB table and instance information allowing further information about the variable involved with the event to be obtained from the **ibmSP** MIB.
  - The value of the resource variable that triggered the event
  - The predicate for which the event was triggered

Other fields in the trap are as specified by SNMP protocol.

The MIB for the SP

The **ibmSP** MIB is provided to define SP-specific information. The **ibmSP** MIB consists of an **ibmSPConfig** group, the **ibmSPErrlogVars** group, and the **ibmSPEMVariables** group. The **ibmSP** MIB is instantiated on each SP processor node and the control workstation.

The **ibmSP** MIB is defined in the source file **ibmSPMIB.my** in the **/usr/lpp/ssp/config/snmp_proxy** directory. For a sample of the **ibmSP** MIB file, see "ibmSPMIB.my" on page 516.

The **ibmSPConfig** group defines objects containing SP system configuration information. The group consists of objects containing the following information:

- The node number of the node on which the queried subagent is running
- The IP address of the system partition in which the queried node resides
- The host name of the primary control workstation.
- The operational state of the primary control workstation (up or down)
- The host name of the backup control workstation (will be **NULL** if none exists)
- The operational state of the backup control workstation (up or down)
- The version number of the IBM Parallel System Support Programs for AIX running on the control workstation
- A table of SPNodeEntrys. Each row in the table is indexed by the IP address of a system partition combined with a node number. Each row contains the following information about a node:
  - The IP address of the system partition in which it resides
  - The node number
  - The number of the frame containing the node
  - The lowest slot number in the frame containing the node
  - The number of slots in the frame occupied by the node
  - The reliable host name assigned to the node (this is the host name associated with the SP Ethernet)
  - The initial host name assigned to the node (this is the host name assigned to the node during the customization based on data obtained from the SP Perspectives GUI)
  - The name of the system partition in which the node resides
  - The version number of the IBM Parallel System Support Programs for AIX running on the node

The **ibmSPConfig** group, when instantiated on the control workstation, will identify the default system partition IP address as the IP address of the system partition containing the reporting node and **0** as the node number of the reporting node.

The **ibmSPErrlogVars** group is instantiated on each SP processor node and the control workstation. The group consists of a sequence of objects containing information about the latest Error Log write that caused an SNMP trap to be issued on the reporting node.

The **ibmSPEMVariables** group is instantiated on each SP processor node and the control workstation. The group consists of a sequence of objects containing information about the last Event Management event that caused an SNMP trap to

be issued on the reporting node and tables of objects representing Event Management variables. The group contains up to three tables:

- A table of objects (**ibmSPEMNodeDepVarsTable**) allowing access to the definitions of node-dependent resource variables (those defined with an Event Manager 'Locator' attribute). This table contains object instantiations for only those variables which have a Locator value equalling the node number of the node on which the SP proxy agent is running.
- A table of objects (**ibmSPEMNodeIndepVarsTable**) allowing access to the definitions of node-independent resource variables (those defined without an Event Manager 'Locator' attribute)
- A table of objects (**ibmSPEMVarValuesTable**) allowing access to the current values assigned to Event Management resource variables

The following information is provided through object instantiations for each of the Event Management resource variables in the **ibmSPEMNodeDepVarsTable** and **ibmSPEMNodeIndepVarsTable** tables:

- Strings containing the following Event Management variable information:
  - Name of the resource variable
  - Resource variable description
  - Variable type (counter, quantity, or state)
  - Data type (long, float, structured byte string)
  - Structure byte string format (only for a structured byte string data type)
  - Initial variable value
  - An index into the **ibmSPEMVarValuesTable**, which, when used in combination with an instantiation vector, provides access to an entry containing the current value of a variable instance
  - Variable's resource class
  - Instantiation vector definition
  - Description of each element in the instantiation vector
  - Name used to read and write the variable in the PTX shared (may be **NULL**) memory
  - Default predicate for event notification (may be **NULL**)
  - A description of the event
  - Locator value specified for the variable
  - Resource order group

The **ibmSPEMVarValuesTable** provides object instantiations for the current values of variables from both the **ibmSPEMNodeDepVarsTable** and the **ibmSPEMNodeIndepVarsTable** tables.

The portion of the **ibmSPEMVariables** group instantiated by the **sp_configd** subagent depends on the type of SP node the subagent is running on. The **ibmSPEMNodeDepVarsTable** and **ibmSPEMVarValuesTable** tables are instantiated on all the processor nodes and the control workstation. The **ibmSPEMNodeIndepVarsTable** table is instantiated on only the control workstation.

Viewing the ibmSP MIB

The AIX **snmpinfo** command may be used in place of a manager client, such as NetView for AIX, to view the contents of the **ibmSP** MIB. Use of this command requires **root** authentication. The following are examples of using **snmpinfo**:

1. To dump the contents of the **ibmSP** MIB located on the host *host_name* in verbose mode, enter the command:

   ```
   snmpinfo -m dump -v -h host_name ibmSP
   ```

2. To dump the contents of the **ibmSPConfig** group, enter the command:

   ```
   snmpinfo -m dump -v -h host_name ibmSPconfig
   ```

3. To get values for the MIB variable **SPhostnodenumber.0** located on the local host, enter:

   ```
   snmpinfo -m get ibmSPhostnodenumber.0
   ```

4. To get the value for the MIB variable following the **ibmSPhostnodenumber** variable, enter:

   ```
   snmpinfo -m next ibmSPhostnodenumber.0
   ```

Disabling Network Management for the SP System

If the SP is not part of a network management system, you can delete the **sp_configd** subsystem by issuing the following command on each SP processor node and on the SP control workstation:

```
/usr/lpp/ssp/bin/sp_configdctrl -c
```

This will delete the **sp_configd** subsystem from the SRC, and remove **sp_configd** entries from the following files:

- **/etc/inittab**
- **/etc/snmpd.conf**
- **/etc/snmpd.peers**

And it will remove the **ibmSP** MIB definitions appended to:

- **/etc/mib.defs**

Configuring the SP Proxy Agent

The SP proxy agents (**sp_configd**) residing on each managed SP node and control workstation are configured to be started automatically whenever AIX is started. The configuration changes necessary to accomplish this are performed during installation on the nodes and the control workstation.

If you have a management station that “listens” for traps, place the information for the trap destination in the **/etc/snmpd.conf** file on each SP node and control workstation. Instructions for doing this are in the **/etc/snmpd.conf** file. The following steps are performed automatically by the **syspar_ctrl** script on each node and the control workstation when the PSSP software is installed:

1. These steps presume that the standard AIX Agent is already configured to be managed by the desired Managers. If this is not the case, see the information contained in the **/etc/snmpd.conf** file for required updates to the file. This update must be made on every SP node and the control workstation.

2. The **/etc/snmpd.conf** file is modified to add the SP subagent (**sp_configd**) to the set of SMUX peers with a "smux" record
3. The SP MIB file **/usr/lpp/ssp/config/snmp_proxy/ibmSPMIB.defs** is appended to the **/etc/mibs.def** file. The **/usr/lpp/ssp/config/snmp_proxy/ibmSPMIB.defs** file is compiled from **ibmSPMIB.my**
4. Information for the new **sp_configd** subagent is added to the **/etc/snmpd.peers** file. The information is the OID of the Agent and the community name (password).
5. The **snmpd** daemon is refreshed.
6. The daemon is started under SRC control.
   a. An src_entry for **sp_configd** is added to the AIX Object Data Manager (ODM) so that the daemon can be controlled using the **startsrc** command. The **src_entry** invokes the **startsrc** command with the following parameters:

      ```
      -s sp_configd -a "-t 600" -p /usr/lpp/ssp/bin/spconfigd
      -u 0 -i /dev/null -o /dev/null -e /dev/null -S -n 15 -F 15
      ```

   b. An entry is added to the AIX **inittab** file to start the daemon by issuing the command:

      ```
      /usr/bin/startsrc -s sp_configd
      ```

   c. The **sp_configd** daemon supports a **-t** switch that determines the amount of time (in seconds) for which data is to be cached by the daemon and re-fetched when a new request for the data is received. The default value is 60 seconds. The default can be changed in the src_entry or the **startsrc** command contained in the **inittab** file, or by entering **startsrc** at the command line.
   d. The daemon can be stopped by issuing:

      ```
      /usr/bin/stopsrc -s sp_configd
      ```

Configuring NetView for AIX for the SP System

If you are using NetView for AIX as the network manager for your SP system, you must configure it to recognize SP trap and configuration information.

Configuring the SP MIB to be Used by NetView for AIX

1. Copy the file **/usr/lpp/ssp/config/snmp_proxy/ibmSPMIB.my** from an SP node into **/usr/OV/snmp_mibs** as **ibmSP.mib** on the node running NetView for AIX.
2. If you have dependent nodes attached to your SP system, copy the file **/usr/lpp/ssp/config/spmgrd/ibmSPDepNode.my** from the control workstation into **/usr/OV/snmp_mibs** as **ibmSPDepNode.mib** on the node running NetView for AIX.
3. Load the appropriate MIBs using the NetView for AIX menu option **Load/Unload MIBs**.

Configuring SP Events to be Used by NetView for AIX

Traps generated containing events from SP nodes and the control workstation are displayed automatically at the NetView for AIX **Control Desk** if the node on which the event originated is a managed node. If the node is not managed, the event is recorded in the event history file and is displayed only when you query the event history file. The traps contain object values indicating:

- The trap ID
- The variable name
- The event predicate definition that caused the event
- The value of the variable at the time the event was generated
- Information necessary to use the NetView for AIX browse facility to query the node for the current value of the variable or for more information about the characteristics of the variable and its value

You may tailor the information in the traps by using the **Trap Customization** selection under the **Event/Configuration** AIX menu option.

Chapter 27. Managing Error Logs

The IBM RS/6000 SP (SP) use both the BSD syslog and AIX Error Logging facilities, as well as a number of function-specific log files to record error events on each node. Refer to the *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide* for an overview of error logging and a description of SP specific log files.

This chapter describes error log management functions provided in the **ssp.sysman** install option of the IBM Parallel System Support Programs for AIX (PSSP). (Note that error log management function is provided as part of PSSP V2R1 and cannot be run on AIX 3.2.5 nodes.) Error log management consists of:

- SMIT panel interfaces
- A set of external commands
  - **splm** for general log viewing, archiving, and service collection
  - **psyslrpt** for generating reports of BSD syslog log files
  - **psyslclr** for trimming BSD syslog log files
  - **penotify** for creating, removing, or displaying Error Notification Objects
- Sysctl-based server functions

Methods are provided for performing the following log management tasks in parallel from a single point of control:

- Managing the AIX Error Log facility
- Managing the BSD syslog facility
- Archiving error logs
- Viewing error logs
- Collecting system data for IBM Service

To access the primary SMIT menu for error log management:

TYPE	**smit**
	▲ The System Management menu appears.
SELECT	RS/6000 SP System Management
	▲ The RS/6000 SP System Management menu appears.
SELECT	RS/6000 SP Log Management
	▲ The RS/6000 SP Log Management menu appears.

The fastpath invocation for the Log Management menu is:

```
smit splogmgt
```

Installing and Configuring Error Log Management

Log management functions are built upon the Sysctl facility, which uses the SP authentication services. Generating parallel AIX Error Log and BSD syslog reports and performing general log viewing require that the user issue the **kinit** command to be identified to the SP authentication services. All other log management commands additionally require that the user be defined as a principal in the **/etc/logmgt.acl** file. All users defined in the file **/etc/logmgt.acl** must also be placed in the authentication (PSSP or AFS) database as a principal. (Refer to “Adding Principals and Assigning Intitial Passwords” on page 182 for more

© Copyright IBM Corp. 1995, 1997

information.) Note that the majority of log management represents administrative tasks normally requiring root authority and that a user defined in the **logmgt.acl** file will execute commands as the root user.

The following is an example of an **/etc/logmgt.acl** file:

```
#acl#
# This sample acl file for log management commands contains
# a commented line for a principal
#_PRINCIPAL root.admin@HPSSL.KGN.IBM.COM
_PRINCIPAL root.admin@HPSSL.KGN.IBM.COM
_PRINCIPAL joe@HPSSL.KGN.IBM.COM
```

The file contains entries for the user root as principal **root.admin** and user joe as principal **joe**, giving them authority to execute log management commands.

The log management server functions executed by Sysctl are located in **/usr/lpp/ssp/sysctl/bin/logmgt.cmds**. During system installation, an include statement for this file is added to the default Sysctl configuration file **/etc/syctl.conf**. If an alternate Sysctl configuration file is used, it must be updated with a statement to include the **logmgt.cmds** file, and the **sysctld** daemon must be restarted to pick up this change. Refer to Chapter 15, “Sysctl” on page 195 for a description of the Sysctl function, ACL and AUTH callback authorizations, and using an alternate Sysctl configuration file.

Managing the AIX Error Log Facility

For detailed information on the AIX Error Log facility, refer to:

- *IBM RS/6000 Problem Solving Guide*
- *IBM General Concepts and Procedures for RS/6000*

SMIT menus as well as commands are provided to help you manage the AIX Error Log facility. You can perform the following management tasks for the AIX Error Log facility:

- Generate AIX Error Log reports
- Trim AIX Error Log reports
- Configure AIX Error Logs
- Manage error notification objects
- Manage error templates

When you execute commands related to AIX Error Log management you can specify whether the command should be executed on all nodes in the current system partition or you can specify node names or the name of a file containing a list of node names. The default is the local node.

From SMIT: To access the AIX Error Log SMIT menu, enter:

```
smit sperrlog
```

Generating AIX Error Log Reports

You can generate reports on entries in the AIX Error Log on a number of nodes. The report can be displayed or written to a file on the local node.

From SMIT: The fastpath invocation for the Generate an Error Report menu is:

```
smit perrpt
```

Trimming AIX Error Logs

You can trim records from error logs on a set of nodes.

From SMIT: The fastpath invocation for the Clean the Error Log menu is:

```
smit perrclear
```

Configuring the AIX Error Log

You can display the configuration parameters of the AIX Error Log to the local node.

From SMIT: The fastpath invocation for the Show Characteristics of the Error Log menu is:

```
smit perrdemon_shw
```

You can alter one or more of the configuration parameters for the AIX Error Log on a set of nodes. Because of the additional entries generated by SP system software, we suggest AIX Error Log file size be a minimum of 4M.

From SMIT: The fastpath invocation for the Change Characteristics of the Error Log menu is:

```
smit perrdemon_chg
```

Managing Error Notification Objects

Error Notification Objects are ODM objects held in the class errnotify that are used by the AIX Error Notification Facility to invoke methods upon occurrence of an error event. Fields in the errnotify class match to fields in an Error Template for selection. If an error is logged matching the selection criteria defined in a notification object, the method associated with that object is invoked. For more information on using the AIX Error Notification Facility, refer to *IBM General Concepts and Procedures for RS/6000*.

You can add, remove, and show notification objects in parallel on the SP system.

From SMIT:

- The fastpath invocation for the Add a Notification Object menu is:

  ```
  smit padd_en
  ```

- The fastpath invocation for the Remove a Notification Object menu is:

  ```
  smit prem_en
  ```

- The fastpath invocation for the Show a Notification Object menu is:

  ```
  smit pshw_en
  ```

From the Command Line: To add, remove, or show error notification objects in parallel on the SP system, enter:

```
penotify -f show
```

For complete details on the command, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Following is a description of the actions taken by the notification method EN_pend located under the **/spdata/sys1/err_methods** directory. This method can be installed and used to invoke pre- and post- action scripts and mail a report of the logged error. This script provides a suggested structure for notification methods and can be reused with different pre- or post- action scripts as described in the sections.

EN_pend Method Flow

1. EN_pend looks under the directory it resides in for a file with the same name and a **.envs** suffix. If found, it sources the file to pick up environment variables EN_RUNDEFAULT and EN_MAILLOC. An **EN_pend.envs** script is installed under the same directory.
2. **EN_pend.envs** sets the EN_RUNDEFAULT environment variable. It also sets the EN_MAILLOC to root at the control workstation, if possible, or to root at the local node.
3. **EN_pend** checks for a pre-action script under the same directory and name with a **.pre** suffix and executes it if found.
4. **EN_pend** mails an expanded report of the error using the sequence of the error passed by the notification facility to the EN_MAILLOC, if EN_RUNDEFAULT and EN_MAILLOC variables are set.
5. **EN_pend** checks for a post-action script under the same directory and name with a **.post** suffix and executes it if found.

Installing a Notifcation Object

To add the **EN_pend** method to all nodes in the current partition to send a report whenever an error of type PEND (loss of availability of a device is imminent) occurs, enter:

```
penotify -a -n "PEND_err" -P -t "PEND" -m '/spdata/sys1/err_methods/EN_pend $1'

penotify -a -n "pend_err" -P -t "pend" -m '/spdata/sys1/err_methods/EN_pend $1'

penotify -a -n "Pend_err" -P -t "Pend" -m '/spdata/sys1/err_methods/EN_pend $1'
```

The **-P** flag will cause the object to persist after the system is restarted. Three objects are added with variations on PEND because upper case is not always adhered to by all AIX LPPs and vendor functions. The **$1** argument causes the Error Notification Facility to pass the error sequence number to the notify method.

The **EN_pend** and **EN_pend.envs** scripts can be used to invoke different pre- and post- action scripts for different error events by creating links to them. **EN_pend** looks for **.envs**, **.pre** and **.post** scripts under the directory it is called from, and by the same basename. For example, to use **EN_pend** for reporting hdisk0 errors on nodes h0, h1, h2 and h3 and perform pre- and post- action:

1. Create .pre and .post action scripts on one of the nodes, for example, EN_hdisk0.pre, EN_hdisk0.post under the /**spdata/sys1/err_methods** directory on node h0.
2. Copy the pre- and post- scripts to nodes h1, h2 and h3 using **pcp**:

   ```
   pcp -w h1,h2,h3 EN_hdisk0.pre /spdata/sys1/err_methods/
   pcp -w h1,h2,h3 EN_hdisk0.post /spdata/sys1/err_methods/
   ```

3. Create links to the **EN_pend** and **EN_pend.envs** scripts:

   ```
   dsh -w h0,h1,h2,h3 ln -s /spdata/sys1/err_methods/EN_pend /spdata/sys1/err_methods/EN_hdisk0

   dsh -w h0,h1,h2,h3 ln -s /spdata/sys1/err_methods/EN_pend.envs /spdata/sys1/err_methods/
   EN_hdisk0.envs
   ```

4. Add the notification object:

   ```
   penotify -w h0,h1,h2,h3 -f add -P -n "hdisk0_err" -m '/spdata/sys1/err_methods/EN_hdisk0 $1' -N
   "hdisk0"
   ```

To display the notification object created, enter:

```
penotify -w h0,h1,h2,h3 -f show -n "hdisk0_err"
```

To remove this notification object, enter:

```
penotify -w h0,h1,h2,h3 -f remove -n "hdisk0_err"
```

Note: Notification methods need to be accessible to each node that the notify object is added to. We suggest that the notification scripts be kept local on each node in case of network failure. File collections should be used to maintain updates to notification methods.

Managing Error Templates

You can create a new error template in the Error Template repository for logging errors.

From SMIT: The fastpath invocation for the Add an Error Template menu is:

```
smit padd_et
```

You can remove a template from the Error Template Repository.

From SMIT: The fastpath invocation for the Remove an Error Template menu is:

```
smit prem_et
```

You can display entries from the Error Template Repository to the local host.

From SMIT: The fastpath invocation for the Show an Error Template menu is:

```
smit pshw_et
```

Managing the BSD syslog Facility

SP error logging utilizes the BSD syslog facility for recording error events.

From SMIT: To access the Syslog SMIT menu, enter:

```
smit spsyslog
```

Generating Reports on BSD syslog Log Files

The **syslogd** daemon, which logs the errors, is configured with a file designating filters for the incoming messages to determine their destination. The default configuration file is **/etc/syslog.conf**. BSD syslog errors are classified by the facility that is issuing the error, and by the error's priority value. Refer to **syslogd** in the *IBM AIX Version 4.1 Commands Reference*, SBOF-1851, for a list of facility and priority values. Entries in the configuration file determine the destination for each error message based on these values. Destinations can be a file, user ID, or the **syslogd** daemon on another machine. We suggest that error messages be logged locally rather than forwarded to a remote **syslogd** because of the increased network traffic. File collections can be used to maintain consistent configuration of the syslog facility.

The format of a syslog log file is:

```
MMM DD HH:MM:SS node_name resource[pid]: msg
```

Where:

MMM DD HH:MM:SS Timestamp (month day hour:minute:second)

node_name The name of the node that the error occurred on

resource The name of the failing system

pid The optionally logged process ID of the failing resource.

msg A free form error message.

Errors logged by SP components will contain in the message section the following additional information pertaining to the logging resource:

LPP LPP name

Fn filename

SID SID_level_of_the_file

L# Line number or function

From SMIT: The fastpath invocation for the Generate a Syslog Report menu is:

```
smit spsyslrpt
```

From the Command Line: To report to the local node all records logged by **ftp** for all syslog log files on nodes in the current system partition, enter:

```
psyslrpt -a -r ftp
```

To report all records that logged to files selected for the daemon and user facilities starting on March 3 and to report the records to the local node from nodes host1 and host2, enter:

```
psyslrpt -w host1,host2 -f user,daemon -s 03030000
```

For complete details on the **psyslrpt** command, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Trimming BSD syslog Log Files

The size of syslog log files is not configurable and will continue to grow until manually trimmed.

From SMIT: The fastpath invocation for the Trim Syslog Log Files menu is:

```
smit spsyslclr
```

From the Command Line: To trim all records older than 30 days from the log file **/var/adm/msgs** on the local node, enter:

```
psyslclr -y 30 -l /var/adm/msgs
```

To trim all records from all log files found in the alternate syslog configuration file **/var/adm/syslog.conf**, enter:

```
psyslclr -g /var/adm/syslog.conf -y 0
```

Note: The **syslogd** daemon is stopped during the trimming process, then restarted with either the default configuration file or an alternate file, if used (**-g** option). For complete details on the **psyslclr** command, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

psyslclr can be added as a **crontab** entry for performing scheduled syslog trimming. On the control workstation, **psyslclr** is used to trim daemon facility messages older than six days. This is done in **/usr/lpp/ssp/bin/cleanup.logs.ws**, which is run from the control workstation's **crontabs** file. (Refer to Chapter 13, “Maintaining the crontabs File” on page 153 for more information.)

Viewing Error Logs

You can view error log data.

From SMIT: To access the General Log Viewing SMIT menu, enter:

```
smit splogview
```

From the Command Line: To cause the contents of each file entry and output of each command entry on the target nodes in the log table **amd.tab** to be displayed to stdout of the node the **splm** command is executed on, enter:

```
splm -a view -t /spdata/sys1/logtables/amd.tab
```

For complete details on the **splm** command, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Archiving Error Logs

Error log archives can be created, removed, or gathered to a central location where they can be optionally written to a tape device or mailed to another location.

From SMIT: To access the Archive Logs SMIT menu, enter:

```
smit sparchive
```

Creating Archives

From SMIT: The fastpath invocation for the Create Archives menu is:

```
smit spcreate_archive
```

From the Command Line: To create an archive on each target node under directory **/var/archives/arch_weekly.tab**, enter:

```
splm -a archive -t /spdata/sys1/logtables/weekly.tab -c -d /var/archives
```

Removing Archives

From SMIT: The fastpath invocation for the Remove Archives menu is:

```
smit spremove_archive
```

From the Command Line: To remove all files and directories under and including **/var/archives/arch_weekly.tab**, enter:

```
splm -a archive -t /spdata/sys1/logtables/weekly.tab -r -d /var/archives
```

Gathering Archives

From SMIT: The fastpath invocation for the Gather Archives menu is:

```
smit spgather_archive
```

From the Command Line: To gather the compressed tar files created on each target node under directory **/var/archives/arch_weekly.tab** to the directory **/var/logrepos** on the local node.

```
splm -a gather -k archive -t /spdata/sys1/logtables/weekly.tab -d /var/archives -l /var/logrepos
```

For complete details on the **splm** command, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

Collecting System Data for IBM Service

Service collections can be created, removed and gathered in the same manner as the archive functions with the addition of interacting with the AIX command **snap** to gather additional system information. The default top level directory for service collections is **/tmp** and this path will end with **srvc**_*log_table_name*.

From SMIT: To access the Collect Logs for Service SMIT menu, enter:

```
smit spcollect
```

Creating Service Collections

From SMIT: The fastpath invocation for the Create Service Collections menu is:

```
smit spcreate_collect
```

From the Command Line: In the following example service collections are created using table **ssp.tab**. **snap -gfk** will be called to gather general system information, file system information and kernel information. Additional log and system data designated in the **ssp.tab** table will be collected under the other directory created by **snap**. The *nodename*.**tar.Z** compressed tar file will be created.

```
splm -a service -t /spdata/sys1/logtables/ssp.tab -p 'gfk' -c
```

Removing Service Collections

From SMIT: The fastpath invocation for the Remove Service Collections menu is:

```
smit spremove_collect
```

From the Command Line: In the following example, the service collections that were created using table **ssp.tab** are removed:

```
splm -a service -t /spdata/sys1/logtables/ssp.tab -r
```

Gathering Service Collections

From SMIT: The fastpath invocation for the Gather Service Collections menu is:

```
smit spgather_collect
```

From the Command Line: In the following example the compressed tar files from the nodes in table **ssp.tab** are gathered to the local node and written to tape device /**dev**/**rmt0**. The **-s** flag specifies to stagger the process by gathering the file from one node, writing it to tape, then removing it before moving onto the next node. This provides a method for gathering all the compressed tar files to a tape or disk device, or mailing them, without requiring disk space on the local node for all the service tar files.

```
splm -a gather -k service -t /spdata/sys1/logtables/ssp.tab -d /tmp -s -o /dev/rmt0
```

For complete details on the **splm** command, refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

For information on the **snap** command, refer to the *IBM AIX Version 4.1 Commands Reference*, SBOF-1851

System Performance

© Copyright IBM Corp. 1995, 1997

Chapter 28. Tuning the SP System

This chapter presents tuning information for the SP system. The chapter consists of six primary sections:

- The first section deals with how to approach tuning an SP system. In many cases tuning an SP system is different from, sometimes even opposite to, tuning a RS/6000 workstation.
- The second section deals with understanding the network topology of the SP system. Understanding the network topology aids in tuning for optimal performance.
- The third section deals specifically with tuning the networks. Network tuning is the key to achieving peak performance on the SP system. Network tuning must be done carefully due to the different approach to tuning required by the SP system.
- The fourth section addresses tuning the SP system switch adapters.
- The last two sections address specific problems and subsystems and how to adjust the tuning of the SP system for them.

We recommend that you read the first four sections before attempting to tune specific subsystems or to alleviate bottlenecks. Understanding all aspects of tuning an SP system before actually making any changes will help you to tune for optimal performance.

Once the SP system subsystems and networks are properly tuned, use standard AIX tuning within a node. Additional AIX tuning information can be found in *AIX Versions 3.2 and 4.1 Performance Tuning Guide*, SC23-2365.

Performance Tuning and Monitoring

When tuning or monitoring the performance of the SP system you must make sure that you know what you change, when you change it, when you collect any data, and what the data you collect is for. Not using change control, or having time stamps on performance data, makes tuning the SP system a very difficult task. We highly recommend that a change control system is used to track and monitor any changes to the tunables on any part of the SP system, including the control workstation and the nodes.

You must take into account the total system picture when you're tuning large configurations. For example, if you're tuning a partitioned SP system, subsystems that span several system partitions (such as the SP system Ethernet) can cause changes on more than one system partition.

Some administrative tasks, such as collecting and clearing log files, cause network traffic that could affect system performance. You should consider scheduling such tasks for off shift hours.

The approach to tuning the SP system is, in some situations, the opposite of how you would tune an AIX workstation. In tuning an AIX workstation or server, the most common approach is to tune the machine to handle the amount of traffic or services requested of it. In the case of a file server, the server is tuned to the

© Copyright IBM Corp. 1995, 1997

maximum amount of traffic it can receive. In general, the bottleneck in a high-end server is the capacity of the network through which services are requested.

In the SP system, a switch is faster than any other network available. With the non-blocking nature of a switch, the number of requests and volume of data that may be requested of a node can far exceed the node's capacity. To properly handle this situation, the application or tuning on the SP system must manage the volume of services requested of a server or destination node. Instead of trying to tune a node for the maximum amount of services requested, each of the nodes requesting services needs to manage the volume of requests made of another node. In other words, you should reduce the number or volume of requests at the client, rather than increase the capacity of the server. It is very easy on large SP system configurations to require more services than the most powerful node can currently deliver.

Changing Network Options on the Nodes

On a newly installed system where you have not selected an alternate network tuning file or created your own customized network tuning file, the nodes start off with default SP tunable settings. These network performance tuning settings have been selected by IBM for a typical SP environment and replace the default AIX settings (which are not appropriate for most SP environments). See the "Perform Additional Node Customization" step in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

There are three methods for changing the network performance tunables on the nodes after you have installed your SP system.

Changing Network Options on All Nodes: Using an IBM-Supplied File

IBM supplies three alternate tuning files which contain initial network performance tuning parameters for three different SP environments:

/usr/lpp/ssp/install/config/tuning.commercial Contains initial network performance tuning parameters for a typical commercial environment

/usr/lpp/ssp/install/config/tuning.development Contains initial network performance tuning parameters for a typical interactive or development environment

/usr/lpp/ssp/install/config/tuning.scientific Contains initial network performance tuning parameters for a typical engineering or scientific environment

To change the network option values for all nodes in your SP system, take the following steps:

1. On the control workstation, copy one of these three files to the **/tftpboot/tuning.cust** file. You can do this with SMIT or the **cptuning** command (see *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*).
2. Use SMIT or the **spbootins** command to set the nodes to **customize** mode.
3. To effect the changes on the nodes, you can either reboot the nodes (not a network boot) or simply run **/spdata/sys1/install/pssp/pssp_script** on each node. Use **dsh** to run **pssp_script** on several nodes at once. If you are using

boot/install servers, customize and boot the boot/install server first before customizing its client nodes. The revised **tuning.cust** file will be transferred from the control workstation to the boot/install server and then from the boot/install server to the client nodes.

After the node customization has completed, all subsequent boots will use the new network tunables on the nodes.

Note: Each of the supplied network tuning parameter files (including the default tuning parameter file) contains the line:

```
/usr/sbin/no -o ipforwarding=1
```

We recommend that on non-gateway nodes you change this to:

```
/usr/sbin/no -o ipforwarding=0
```

After a non-gateway node has been installed, migrated, or customized, you can make this change in the **/tftpboot/tuning.cust** file on that node.

Changing Network Options on the Node: User-Customized Settings

To change the network option values set during boot of the nodes from the control workstation, take the following steps:

1. On the control workstation, create the file **/tftpboot/tuning.cust**. You can use the file **/usr/lpp/ssp/samples/tuning.cust** as a starting point by first copying it to **/tftpboot/tuning.cust**. Or you can use one of the following files as a starting point by copying it to **tuning.cust**:
   - **/usr/lpp/ssp/install/config/tuning.commercial**
   - **/usr/lpp/ssp/install/config/tuning.development**
   - **/usr/lpp/ssp/install/config/tuning.scientific**
2. Select the tunable values that are best for you system. For tuning specific network adapters, refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*. For tuning for specific user environments, refer to Table 15 on page 433. This can be done only for AIX 4.1 and later nodes (not AIX 3.2.5).
3. Edit **tuning.cust**. Uncomment the **/usr/sbin/no** commands and change to the appropriate values.
4. Run **spbootins** to set the affected nodes to customize mode.
5. To make the changes to the nodes, you can either reboot the node or run **/spdata/sys1/install/pssp/pssp_script** on each affected node. Use **dsh** to run **pssp_script** on several nodes at once. If you are using boot/install servers, customize the boot/install server first before customizing its client nodes. The revised **tuning.cust** file will be transferred from the control workstation to the boot/install server and then from the boot/install server to the client nodes.

After the node customization has completed, all subsequent boots will use the new network tunables on the nodes.

Changing Network Options on Selected Nodes

You can change the network tunables for a node locally on the node. Take the following steps:

1. Login to the node as root and edit the file **/tftpboot/tuning.cust**. If there is no **/tftpboot/tuning.cust** file, make a new one by copying **/usr/lpp/ssp/samples/tuning.cust** to **/tftpboot/tuning.cust** on the node.
2. Select the tunable values that are best for the node. For tuning specific network adapters, refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*. For specific user environments, refer to Table 15 on page 433 This can be done only for AIX 4.1 and later nodes (not AIX 3.2.5).
3. Edit **tuning.cust**. Uncomment the **/usr/sbin/no** commands and change the default tunable values.
4. Once you have made the changes, run the **tuning.cust** script locally to make the changes on the node. There is no need to reboot for the new values to become effective. You may need to change the permission on the file to be able to execute it.

   Note: This is a temporary change. The next time the node is customized, the **tuning.cust** file will be transferred from the node's boot/install server and executed on the node, thus replacing any changes made locally. To make the network options permanent, change the **tuning.cust** file on the node's boot/install server, which in some cases may be the control workstation.

System Topology Considerations

When you set up a network for a large system you need to consider:

- The SP Ethernet
- Outside network connections
- Routers
- Gateways
- Switch traffic

The most important consideration in configuring the SP Ethernet is the number of subnets you configure. The routing through the SP Ethernet can be complicated because of the limitation on the number of simultaneous network installs per Ethernet subnet. More information regarding the topology of the SP Ethernet can be found in *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment*.

When configuring software that uses the SP Ethernet, consider such software's location in the SP Ethernet topology. For example, software such as LoadLeveler and POE use the SP Ethernet to communicate to other nodes. Installing such software on a far point in the network in a large SP configuration can cause bottlenecks on the network subnets and adapters connected to the control workstation. Installing such software on the control workstation causes the lowest possible traffic; however, the control workstation has to be powerful enough to act as the control workstation and, in addition, support the LoadLeveler and POE software.

If you connect the SP Ethernet to your external network, make sure that the user traffic does not overload the SP network. You can overload the SP Ethernet if you

route high speed network (FDDI or ATM, for example) traffic through it. Route gateway traffic over a switch for gateways to FDDI, ATM, and other high speed networks. Configure routers or gateways to distribute the network traffic so that one network or subnet is not a bottleneck.

Several gateways should be configured if a lot of traffic is expected. All traffic on these networks can be monitored using the standard network monitoring tools. Details about the monitoring tools and their usage can be found in *AIX Versions 3.2 and 4.1 Performance Tuning Guide*, SC23-2365.

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

A list of recommended network tunables for specific adapters can be found in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

Tuning the SP Ethernet

When you tune the SP network, you must consider the number of frames, the number of nodes, and the number of networks you are configuring.

The lower the amount of traffic in each network, the better the performance or response time for the system. If you run a small system of fewer than two frames, you probably connected the SP network to your own network through a gateway. For larger configurations, we recommend that you dedicate one node per frame as a gateway to the rest of the SP system. Routing from the gateway node can be either through the SP Ethernet or across a switch.

When routing across a switch, you should give careful consideration to other traffic moving across the switch. An environment which has many parallel jobs will suffer in performance if a lot of the user traffic is routed through a switch. Such traffic can interrupt the parallel jobs sharing a switch.

In larger SP system configurations, the network topology and the name server can cause problems. If there are many name server queries at the same time, and the topology from a node to the name server is complex, performance can suffer. Under these circumstances, there are three possible solutions:

1. Create a name server within the SP system complex so you are not at risk for traffic problems to an external name server.
2. Don't use the name server, but create an **/etc/hosts** file of all addresses that the SP system and the applications need, and change the nodes to resolve their addresses locally.
3. For an AIX 4.1.4 system or later, specify the name lookup hierarchy to first look locally in the **/etc/hosts** file, then the name server. (This is not possible in AIX systems earlier than 4.1.4.)

SP System-Specific Tunable Recommendations

The tunable values are set to either AIX defaults or the tunables you set during the installation process. See “Tuning for User Environments” on page 432 for the recommended defaults. Use the **netstat** and **no** commands to display tunable values. These commands and their parameters are described in the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*. The values you can tune are:

thewall The upper bound on the amount of real memory that can be used by the communications subsystem. The units are in 1K increments.

lowmbuf The minimum number of 256-byte mbuf buffers kept available at all times. (AIX 3.2.5 nodes only)

lowclust The minimum number of 4096-byte mbuf clusters kept available at all times from the mbuf pool. (AIX 3.2.5 nodes only)

mb_cl_hiwat The maximum number of 4096-byte mbuf clusters that can be used from the mbuf pool. (AIX 3.2.5 nodes only)

sb_max Upper limit on the size of the TCP and UDP buffers in mbufs of allocated space to a connection.

ipforwarding Specifies whether the kernel should forward packets. A value of **1** forwards packets when they are not for the local network; a value of **0** prevents forwarding.

tcp_sendspace The default size of the TCP send space value.

tcp_recvspace The default size of the TCP receive space value.

udp_sendspace The default size of the UDP send buffer. The safe maximum is 65536 (64K).

udp_recvspace The size of the UDP receive buffer.

rfc1323 Turns on several enhancements to the tcp_sendspace and tcp_recvspace tunables for high speed networks:

- The maximum window size is increased from 64K to 4 Gigabytes.
- ACKs are time-stamped for better round trip time estimates.
- Sequence numbers are protected against wrapping.

tcp_mssdflt The default maximum segment size used in communicating with remote networks.

MTU Maximum Transmission Unit on network adapters.

Refer to the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* for a list of initial tunable settings.

AIX 3.2.5 Differences

The way that TCP/IP works is that you allocate two different size allocations out of the mbuf pool. They are mbufs using 256 byte allocations, and mbuf clusters using 4096 byte allocations. Whenever the size of the buffer you are sending is greater than 935 bytes, you need to use an mbuf cluster for staging the data. In addition, if you need an mbuf cluster, you also need an mbuf to point to the mbuf cluster from the mbuf pool.

In AIX 3.2.5, the total mbuf pool can be available to either mbufs or mbuf clusters. The maximum amount of space for the mbuf pool is specified using the **thewall** value. This value is in 1K sizes. If you allocate 4096 as **thewall** value, you have 4 megabytes of space. The **thewall** space can be managed by adjusting allocation to mbufs or mbuf clusters. There are three parameters that allocate mbuf space to either mbufs or mbuf clusters:

- **lowmbuf**
- **lowclust**
- **mb_cl_hiwat**

The **lowmbuf** tunable determines the minimum number of mbufs that are kept available for mbuf allocation. They are not available for mbuf clusters. The **lowclust** is the minimum number of mbuf clusters that are kept available for mbuf clusters. Note that the **lowmbuf** is in 256-byte increments, and the **lowclust** is in 4096-byte increments. The **mb_cl_hiwat** tunable is the maximum number of mbuf clusters that can be used as mbuf clusters.

If you are running AIX 3.2.5, you need to make sure that **mb_cl_hiwat** is big enough. If you have the same number of **lowmbuf** and **lowclust** and you try to use up all the mbuf clusters for staging 4K data blocks, you run out of mbufs. The mbufs are used for other things besides pointing to mbuf clusters, so the **lowmbuf** becomes exhausted first.

When tuning for large file transfers, keep a large number of mbuf clusters available so that when data is transferred in buffers larger than 932 bytes, an mbuf cluster is available. For pure data streaming, a good starting point is to set aside 80% of the mbuf pool to mbuf clusters. Use the **mb_cl_hiwat** tunable to set the percentage.

In addition to opening up the TCP/IP window, adjust the size and allocation of the mbuf pool to give more space to 4KB data blocks on nodes running AIX 3.2.5.

Preventing Network Adapters from Dropping Packets

If traffic is routed through other external network routers, you may not be able to use the optimal maximum transmission unit (MTU) or **tcp_mssdflt** size for that network. If this happens, adjust the **tcp_sendspace** and **tcp_recvspace** values accordingly.

To find the optimal **tcp_sendspace** and **tcp_recvspace** sizes, get the largest MTU size that the router will handle to other networks of the same type. To get the optimal **tcp_sendspace** and **tcp_recvspace** sizes for single stream transfers, use the formula:

$t = m * q$

where:

t Optimal **tcp_sendspace** and **tcp_recvspace** sizes

m Largest MTU size that the router will handle to other networks

q Smaller of the transmit queue and receive queue size for the adapter

This does not apply to a switch because the switch adapter does not have transmit and receive queues. See "Tuning the Switch Device Driver Buffer Pools" on page 433 for more information on tuning **rpoolsize** and **spoolsize**.

The number produced by the formula is the largest that you can use before the adapter drops packets and TCP/IP has to do a retry on the dropped packets. You want to avoid dropping packets on the adapter to get optimal throughput performance.

When configuring the nodes, you must plan the amount of mbuf space for all interfaces to a node. If you have SP Ethernet, FDDI and a switch all on one node, the mbuf pool size must be sized for the aggregate traffic expected. An indication that the mbuf allocation (**thewall** in the **no** tunables) is too small is getting a `request for memory denied` count in the **netstat -m** output. Switch traffic can require a large number of mbufs. Monitoring and tuning the mbufs is detailed in the *AIX Versions 3.2 and 4.1 Performance Tuning Guide*, SC23-2365.

SP Specific Tuning

The following are rules of thumb for tuning an SP :

thewall	At least 25 % of real memory up to a limit of 64 meg
sb_max	At least twice the size of the largest of the **tcp_sendspace**, **tcp_recvspace**, **udp_sendspace** and **udp_recvspace** values described below.
subnetsarelocal	Always set to **1**
ipforwarding	Set to **1** if a gateway node
tcp_sendspace	Minimum of **65536** unless there are problems with external connections not handling greater than 32768 byte windows. Never higher than major network adapter transmit queue limit. To calculate this limit, use *Adapter queue size * MTU*
tcp_recvspace	Minimum of **65536** unless there are problems with external connections not handling greater than 32768 byte windows. Never higher than major network adapter transmit queue limit. To calculate this limit, use *Adapter queue size * MTU*
udp_sendspace	Always less than **udp_recvspace**, but never greater than **65536**
udp_recvspace	Always greater than **udp_sendspace** and sized to handle as many simultaneous UDP packets as can be expected. For single parent/multiple child configurations, set **udp_recvspace** to **udp_sendspace** times the maximum number of child nodes if UDP is used.
rfc1323	Always set to **1**
tcp_mssdflt	Set for major network adapter for routed IP traffic. *Adapter MTU - IP headersize - TCP headersize - \ optional rfc1323 headersize* or *Adapter MTU* - 20 - 20 - optional 12 (52)

The following are additional details for setting the network tunables for the SP system.

thewall

For AIX 4.1 systems, set to at least 25% of real memory. For servers using IP, it might be necessary to increase this to 50% if you have a small memory node. The maximum value you can set this to is **65536** or 64 megabytes of memory. The maximum amount of memory you can allocate to **thewall** is 64 meg, even if this value indicates a higher amount.

For AIX 3.2.5 nodes, see the recommended **thewall** size in Table 16 on page 433.

sb_max

This value is aggregate across **tcp_sendspace**, **tcp_recvspace**, **udp_sendspace**, and **udp_recvspace** PER SOCKET. Because it is in mbuf and mbuf clusters allocated, **sb_max** needs to be larger than **tcp_sendspace** and **tcp_recvspace** combined. The guideline is to set it to twice the largest of the four values.

mb_cl_hiwat (AIX 3.2.5 only)

This tunable sets the upper limit of 4K mbuf clusters that can be allocated from the mbuf pool. This value should be set to correspond to between 20% and 80% of the **thewall** value. This allows the maximum number of mbuf clusters in the pool, plus some space for other small packets. The 80% is used because each mbuf cluster needs to be pointed to by a regular mbuf. That means that each 4K cluster is actually 4352 bytes of mbuf pool space and if set to 80% of the mbuf pool, will actually use up 85% of the pool. The last 15% of the mbuf pool is needed for mbufs for the socket connections, packet headers and protocol control blocks.

To set **mb_cl_hiwat** to allocate the maximum recommended 80% of the mbufs for mbuf clusters, use the formula:

$m = (t/4) * .8$

where:

m	**mb_cl_hiwat** setting
t	**thewall** value from **no -a**

For small data transfer environments, we recommend that you set aside at least 20% for mbuf clusters. Use the formula:

$m = (t/4) * .2$

where:

m	**mb_cl_hiwat** setting
t	**thewall** value from **no -a**

lowclust (AIX 3.2.5 only)

This tunable sets a “free list” size for mbuf clusters kept available for incoming packets. A good start is 20% of the **mb_cl_hiwat**. For large I/O systems, this is a good guide initially. This means that when the system checks the mbuf allocations to see if there is space for incoming data, it will make sure that 20% of the allocation mbuf clusters are free and available. If this number is too small, the traffic will slow down because if an mbuf cluster is not available, the system has to wait to free one up, usually resulting in dropped packets.

lowmbuf (AIX 3.2.5 only)

This tunable is the same as the one above, except for the regular mbufs, not the clusters. A good start is to have 100 more mbufs than clusters on the “free lists”. If the two settings are too close in value, mbufs will run out before the cluster “free list” is exhausted, causing dropped packets even though there are free mbuf clusters.

tcp_recvspace and tcp_sendspace

These tunables establish the size of the TCP window on a per datagram socket connection. The size depends on the network you are sending over as well as the number of sockets open at any one time. The recommendations are set to avoid overrunning networks other than a switch. For more information related to these tunables, see "The Nagle Algorithm" on page 438 and "Adapter Queue Lengths" on page 440.

To properly set these tunables, you need a good understanding of the type of traffic your application will be sending. For optimal switch performance, these need to be set to **256K** or greater.

udp_sendspace

Set to **65536** because anything beyond 65536 is essentially ineffective. Since UDP transmits a packet as soon as it gets any data, and since IP has an upper limit of 65536 per packet, anything beyond 65536 runs the small risk of getting thrown away by IP.

udp_recvspace

We recommend starting at 10 times the **udp_sendspace** because UDP may not be able to pass a packet to an application before another one arrives. Also, several nodes can send to one node at the same time. To provide some staging space, this size is set to allow 10 packets to arrive before any are thrown away. For large parallel applications using UDP, the value may have to be increased.

rfc1323

If you are setting **tcp_recvspace** and **tcp_sendspace** to greater than 65536, you need to set **rfc1323** on both clients and servers. Without having **rfc1323** set, the effective values for **tcp_recvspace** and **tcp_sendspace** will be **65536** even if they are set higher.

tcp_mssdflt

This tunable is used to set the maximum packet size for communication with remote networks. This tunable is the same as the MTU for communication across a local network except for one small difference. The **tcp_mssdflt** size if for the size of the data only, in a packet. You need to reduce the **tcp_mssdflt** for the size of any headers so that you send full packets instead of full a packet and a fragment. The way to calculate this is as follows:

```
MTU of interface - TCP headersize - IP headersize - /
     rfc1323 headersize
```

which is:

```
MTU - 20 - 20 - 12      (MTU - 52)
```

Limiting data to the MTU - 52 bytes ensures that, where possible, only full packets will be sent.

CSS MTU size

The MTU of a switch is **65520** and we recommend that you keep this setting. However, under some circumstances (for example, when you want to avoid the Nagle algorithm causing very slow traffic), it may be necessary to reduce this value. You can reduce the MTU of a switch to as low as 32678 with only a 2 - 10% loss in throughput. However, the CPU utilization will be slightly higher due to the per packet overhead.

To reduce the MTU of the switch, use:

```
ifconfig css0 mtu new_size
```

This takes effect immediately, and needs to be run as root user.

You should always use the same MTU across all nodes.

Tuning the SP Network for Specific Workloads

There are different network tunables that optimize the system for different types of network traffic. The following sections describe the way that the system utilizes these optimal tunables to benefit the types of network traffic generated.

A difficulty in tuning networks is that different mixes of traffic and the way applications utilize the network can work against each other. A set of tunables that benefits a single stream point-to-point large data transfer can cause severe problems for a Parallel Operating Environment (POE). In many cases you cannot optimize for two types of traffic. In addition, if your configuration has several types of network media (token-ring, Ethernet, FDDI and ATM), it is not always possible to optimize for all network types.

Tuning for User Environments

The following tables provide network tunable settings designed as initial values for various user environments. Table 15 on page 433 provides values for AIX 4.1 nodes and Table 16 on page 433 provides values for AIX 3.2.5 nodes. You need to change these values on each of the installed nodes. To temporarily change these values, use the **no** command. If you want these values to be preserved across booting the nodes, we recommend that you change the tuning script on each node. Details on how to change the script are found in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*. See "tuning.cust File" on page 545 for a sample of the tuning script.

The settings in Table 15 on page 433 and Table 16 on page 433 are only initial recommendations. Start with them and realize you may need to change them.

The settings in Table 15 on page 433 are contained in the three IBM-supplied tuning files:

- **/usr/lpp/ssp/install/config/tuning.commercial**
- **/usr/lpp/ssp/install/config/tuning.development**
- **/usr/lpp/ssp/install/config/tuning.scientific**

Table 15. Initial Values for AIX 4.1 and 4.2

Tunable	Software Development Mixed Interactive	Commercial	Scientific/Engineering High Performance Large Data Streaming
thewall	16384	16384	16384
sb_max	163840	442368	1310720
subnetsarelocal	1	1	1
tcp_sendspace	65536	221184	327680
tcp_recvspace	65536	221184	327680
udp_sendspace	32768	59392	65536
udp_recvspace	65536	221184	655360
rfc1323	1	1	1
tcp_mssdflt	1448	1448	Varies

Table 16. Initial Values for AIX 3.2.5

Tunable	Software Development Mixed Interactive	Commercial	Scientific/Engineering High Performance Large Data Streaming
thewall	4096	6144	6144
lowclust	60	100	100
lowmbuf	50	200	200
mb_cl_hiwat	120	1220	1220
sb_max	163840	237568	1310720
subnetsarelocal	1	1	1
tcp_sendspace	65536	118784	655360
tcp_recvspace	65536	118784	655360
udp_sendspace	32768	59392	65536
udp_recvspace	65536	118784	655360
rfc1323	1	1	1
tcp_mssdflt	1448	1448	Varies

Tuning the Switch Device Driver Buffer Pools

When you tune the switch buffer pools, you need to make several considerations in deciding what the optimal settings should be. Because several types of buffers are involved in staging data for the switch, at different times different buffer types can cause performance problems due to running out of buffer space. And along with the switch pools, the TCP/IP mbufs are also used. Along with performance problems, insufficient buffer space could also cause TCP connections to stop transferring data and then timeout after a set amount of time because nothing comes ready on the sockets.

The most important buffers to tune are usually the **spoolsize** and **rpoolsize** (for the spool and rpool buffers) for the switch adapter. These pools are used to stage the data portion of IP packets for the switch. However, the allocation and sizes utilized from the pools can cause buffer starvation problems.

The spool and rpool are separate buffer pools, one for outgoing data (spool) and one for incoming data (rpool). When an IP packet is passed to the switch interface, if the size of the data is large enough, a buffer is allocated from the pool. If the amount of data fits in the IP header mbuf used in the mbuf pool, no spool space is allocated for the packet. This amount of space is a little less than 200 bytes, depending on what type of IP packet is being sent. The header size varies between using UDP or TCP, or having **rfc1323** turned on.

To see the current spool and rpool buffer sizes, run the following command on your nodes:

```
lsattr -E -l css0
```

The output will look like:

```
bus_mem_addr   0x04000000 Bus memory address        False
int_level      0xb        Bus interrupt level       False
int_priority   3          Interrupt priority        False
dma_lvl        8          DMA arbitration level     False
spoolsize      2097152    Size of IP send buffer    True
rpoolsize      2097152    Size of IP receive buffer True
adapter_status css_ready  Configuration status      False
```

In this particular example, the pool sizes have been increased from the default size of 524288.

The pool settings can be changed using the **chgcss** command, and require rebooting the node. See the **chgcss** command in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

The sizes of the buffers allocated by the device driver start at 4096 bytes and increase to 65536 bytes. All sizes in between are a power of 2 value. This means that the buffer sizes allocated from the pools are:

```
4096     4K
8192     8K
16384   16K
32768   32K
65536   64K
```

If the size of the data being sent is just slightly larger than one of these sizes, the buffer allocated from the pool is the next size up. This can cause a 50% efficiency is usage of the buffer pools. More than half of the pool can go unused in bad circumstances (described later).

When assembling a TCP/IP packet, there is always one mbuf from the IP mbuf pool used to assemble the packet header information in addition to any data buffers from the spool. If the mbuf pool size is too small, and the system runs out of mbufs, the packet will be dropped. The mbuf pool is used globally for all IP traffic, and is set using the **thewall** tunable in the **no** command.

When sending 4K of data over the switch, an mbuf from the mbuf pool will be used as well as one 4K spool buffer for the data. If the amount of data being sent is less

than 200 or so bytes, no buffer from the spool is allocated because there is space in the mbuf used for assembling the headers to stage the data.

However, if sending 256 bytes of data, you will end up using one mbuf for the IP headers, and one 4K spool buffer for the data. This is the worst case: you are wasting 15/16 of the buffer in the spool. These same scenarios apply to the rpool when a packet is received on a node.

The key for peak efficiency of the spool and rpool buffers is to send messages that are at or just below the buffer sizes allocated from the pools or less than 200 or so bytes.

The appropriate values for the tunables will be unique for each installation. Therefore, a sizing exercise like the one described here is necessary. When trying to determine the appropriate rpool and spool sizes, you need to get a profile of the message sizes that are being sent by all applications on a node. This will help determine how the rpool and spool will be allocated in total number of buffers. At a given pool size, you will get 16 times as many buffers allocated out of the pool for 4K messages as for 64K messages.

Once you have a profile of the packet or buffer sizes used by all applications using the switch on a node, you can then determine roughly how many of each size spool or rpool buffers will be needed. This then determines your initial rpool and spool settings. A node that stages a mix of packet sizes, of which 25% are smaller than 200 bytes, 50% are 5K and 25% are 32K, if the number of packets staged at any one time is 1024, then the spool initial setting should be 12582912 or 12 megabytes. None of the small packets need any spool space, the 5K packets each use 8K out of the spool and need about 4 megabytes of space, and the 32K packets need about 8 megabytes of spool space. The total estimated pool size needed is 12 megabytes.

The allocations from the pools are very transient. The actual number of buffers used at any time is very dynamic due to the large volume of packets the switch can handle. The above calculation is a conservative estimate in that it assumes all packets will be staged at once. In reality, as packets are being staged into the pool, other packets are being drained out, so the effective number of active buffers should be less.

The sizes allocated from the pool are not fixed. At any point in time, the device driver will divide the pool up into the sizes needed for the switch traffic. If there is free space in the spool, and smaller buffers than the current allocation has available are needed, then the device driver will carve out the small buffer needed for the current packet. As soon as the buffer is released, it is joined back with the rest of the 64K buffer it came from. The buffer pool manager tries to return to 64K block allocations as often as possible to maintain high bandwidth at all times. If the pool were fragmented, and a large buffer needed 64K, then there may not be 64K of contiguous space available in the pool. Such circumstances would degrade performance for the large packets.

If all buffer space is used, then the current packet will be dropped, and TCP/IP or the application will need to resend it, expecting that some of the buffer space was freed up in the mean time. This is the same way that the transmit queues are managed for Ethernet, Token Ring and FDDI adapters. If these adapters are sent more packets than their queues can handle, the adapter drops the packet.

Currently, the upper limit for the spool and rpool is 16 megabytes for each. This means you can get a maximum of 4096 4K or 256 64K buffers each for sending and receiving data. There are APARs available that allow the receive pool to spill over buffer allocations into the mbuf pool instead of throwing them away when the receive pool is full. The send pool already does this in all releases that use the rpool and spool. However, when this condition occurs, there will be a penalty because mbuf management is more expensive and slower than rpool and spool management. The APAR numbers are:

```
Release          APAR
-------          --------
2.1              IX62830
2.2              Ix64979
```

To see what the usage of the send pool has been since the switch was last started, there are unsupported commands called **vdidl***x*. **vdidl3** is for the SP Switch and **vdidl2** is for the High Performance Switch. These commands are found in the **/usr/lpp/ssp/css** directory on each node. To list the node usage, use the **-i** flag to get the adapter pool statistics. The following is a sample output of the first table from this command:

```
send pool: anchor@=0x0546ea00 start@=0x0a2d0000 tags@=0x0a170200
bkt   allocd     free  success     fail    split     comb    freed
 12        0        0      409        0      316        0        0
 13        0        0      220        0      161        0        0
 14        0        0        0        0        0        0        0
 15        0        0        0        0        0        0        0
 16        0        8        0        0        0        0        0
```

Because this is an unsupported command, not all fields are currently used and the format and fields that are used may change at any time.

As of PSSP 2.2 the following field description and usage is valid:

bkt Lists the pool allocation in power of 2 for the line it is on. The line starting with 12 means 2 to the 12th or 4K allocations and the line starting with 16 means 2 to the 16th or 64K allocations.

allocd Lists the current allocations at the time the command was run, for each of the size allocations in the first column. This is an instantaneous value and, therefore, will increase and decrease between successive executions of the command.

free Lists the number of buffers of each allocation that are allocated and unused at the time the command was run. In the above example, there are 8 64K allocations free for use. This is an instantaneous value and, therefore, will increase and decrease between successive executions of the command.

success Increments every time an allocation of the given size succeeded. This counter accumulates and, therefore, shows the number of successes since the adapter was last initialized.

fail Increments every time an allocation is not available for the size requested. However, it is possible that the allocation is made by splitting up a larger allocation, or combining smaller ones, into the size needed. A fail does NOT necessarily mean a packet was dropped. It is also possible that an allocation was split from

a larger allocation without incurring a failed count. This counter accumulates and, therefore, shows the number of fails since the adapter was last initialized.

split Indicates how many times the allocation size was extracted from the pool by carving the size needed from a larger allocation in the pool. This counter accumulates and, therefore, shows the number of splits since the adapter was last initialized.

comb Currently not used.

freed Currently not used.

For systems using the High Performance Switch, and running at least PSSP 2.1 with PTF 11, the **recv** pool counters will also accumulate data.

Be careful interpreting the output because some of the statistics have several meanings.

When tuning the rpool and spools, it is important to understand the network traffic expected. As seen above, if the size of the buffers for the applications are not ideal, much of the spool and rpool will be wasted. Such circumstances can cause the need for a larger rpool and spool because of inefficient usage. When allocating the rpool and spool, realize that this space is pinned kernel space in physical memory. This takes space away from user applications and is particularly important in small memory nodes.

If there are a small number of active sockets, then there is usually enough rpool and spool space that can be allocated. In systems where a node has a large number of sockets opened across the switch, it is very easy to run out of spool space when all sockets transmit at once. For example, 300 sockets each sending 33K of data will far exceed the 16 megabytes limit for the spool. Or, 1100 sockets each sending four 1K packets will also exceed the maximum limit.

On the receive side of a parallel or client/server implementation, where one node acts as a collector for several other nodes, the rpool runs into the same problem. Four nodes each with 600 sockets, each sending to one node two 1K packets, will exceed the rpool limit, but those same sockets each sending twice as much data, 4K in one 4K packet, will work! The key here is sending a single larger packet rather than several smaller ones.

Another situation where a node combination can cause buffer exhaustion is when a faster node sends data to a slower node. A P2SC node can generate traffic several times faster than an older thin node. So, if the P2SC node sends packets as fast as it can, the slower thin node can run out of rpool space. The rpool space can fill because the slower processor cannot drain the rpool as fast as the P2SC node can fill it from across the switch. One way to prevent overrunning the receiving node is to make sure that the aggregate TCP window for all active sockets is less than the rpool size. See additional information regarding the TCP window mechanism in *IBM AIX Versions 3.2 and 4 Performance Tuning Guide*.

The other situation that aggravates exhaustion of the pools is the use of 604 High Nodes. Only one CPU is used to manage the send or receive data streams over the switch. However, each of the other CPUs in the 604 High Node is capable of generating switch traffic. As the number of CPUs increase, so does the aggregate volume of TCP/IP traffic that can be generated. We recommend that for 604 High

Nodes, the spool size be scaled to the number of processors when compared to a single CPU setting.

For applications, changing the size of messages or buffers sent is key to preventing exhausting the switch buffer pools on the SP. Most applications tuned for peak efficiency for slower networks, such as Ethernet, will try and send 1K messages, causing problems on the switch. Increasing the buffer size for the SP not only reduces the requirements on the rpool and spool, but it can also improve performance. TCP/IP transfers data faster over the switch when you use larger packets, with the peak throughput starting with 32K data buffers.

The key to tuning the buffer pools on the SP is to understand the number of packets using the switch pools at any one time and the size of the data being sent. The rpool and spool sizes need to be set so there are enough buffers allocated to handle the size and number of packets being sent or received. Applications that don't efficiently use full buffers from the pool aggravate the pool size needed or can cause exhaustion of the pools.

Tuning Problems and Issues

The section covers topics on specific problems that can be encountered on an SP system and ways to alleviate then.

The Nagle Algorithm

A problem that often occurs on the SP system is that an application runs very slowly when using a switch, while performance is significantly better on an Ehternet or FDDI network interface. This problem is sometimes caused by the Nagle Algorithm (used by TCP/IP) interacting with the delayed ACK (acknowledgment) timer.

The Nagle Algorithm states that under some circumstances, there will be a waiting period of 200 msec before data is sent. Such circumstances are:

- There are no outstanding packets on a socket to return an ACK
- The data is smaller than the segment size (MTU)
- If multiple buffers are sent equal to the TCP window and the window size is less than the MTU (when using a datagram socket, TCP will not allow sending more than the window size before receiving acknowledgment from the destination)

Under such circumstances the Nagle Algorithm causes a waiting period of up to 200 msec for other data that can also be included in a larger packet in order to reduce the number of packets on the network.

The Nagle Algorithm uses the following parameters for traffic over a switch:

- Segment size = MTU
- Window size = smaller of **tcp_sendspace** and **tcp_recvspace**
- Data size = application data buffer

The following are the specific rules used by the Nagle Algorithm in deciding when to send data:

- If a packet is equal to or larger than the segment size (or MTU), and the TCP window is not full, send an MTU-size buffer immediately.
- If the interface is idle, or the **TCP_NODELAY** flag is set, and the TCP window is not full, send the buffer immediately.
- If there is less than 1/2 of the TCP window in outstanding data, send the buffer immediately.
- If sending less than a segment size buffer, and if more than 1/2 the window is outstanding, and **TCP_NODELAY** is not set, wait up to 200 msec for more data before sending the buffer.

In addition to the Nagle Algorithm, a delayed ACK timer can cause slow data transfer over a switch. This timer is set when a single TCP/IP packet is received at the receive side of a connection. If a second packet is received, then the timer expires, and an acknowledgment is sent back to the sending side. You rarely see this on small segment size (MTU) networks because a large buffer of data results in more than one packet being transmitted. However, on large segment size networks like a switch, writing a 32K buffer results in only one packet. That same buffer on smaller segment size networks results in multiple packets, and there is no use of the delay ACK timer.

To summarize, a 200 msec delay in sending data over a switch can be caused by either the Nagle Algorithm being invoked or by the application writing buffers so that only one packet is sent over the interface and the delayed ACK timer causes a 200 msec delay on the acknowledgment.

The following recommendations will help to avoid the Nagle algorithm (and not affect other parts of any system):

1. Control the segment, window, and data sizes at the application socket.
2. From within an application, you can increase the TCP window size on the socket by setting a different size using the **SO_SNDBUF** and **SO_RCVBUF** settings on the socket. For good performance across a switch, we recommend that both **SO_SNDBUF** and **SO_RCVBUF** be set to at least **327680** on both the client and server nodes. You need to set both sides since TCP uses the lowest common size to determine the actual TCP window.

   If you set the **SO_SNDBUF** and **SO_RCVBUF** sizes above 65536, you need to set **TCP_RFC1323** also on the socket unless the **no** options already have it set. Setting **TCP_RFC1323** to **1** will take advantage of window sizes greater than 65536.
3. Set **TCP_NODELAY** on the socket of the sending side to turn off the Nagle algorithm.

   All data sent will go immediately, no matter what the data size. However, if the application sends very small buffers, you will significantly increase the total number of packets on the network.
4. If you are running an application and don't have access to the source code, use the **no** command to increase the TCP window.

   Increasing the TCP window may not always be effective because increasing the **tcp_sendspace** and **tcp_recvspace** size on the sending and receiving nodes may cause other negative affects on the SP system or to other applications running on the system.

5. Change the MTU size of the switch.

   Changing the MTU of the switch will move the window size and buffer size combination where the 200 msec delay is invoked. When writing 32K buffers to a TCP connection, if the TCP/IP window is 65536, only 5 packets per second are transmitted. If you change the MTU of the switch interface to 32768 there is no delay on transmitting a 32768 buffer because it is the same size at the segment size of the switch. However, reducing the MTU of the switch to 32768 degrades the peak TCP/IP throughput between 2 and 10 percent. Reducing the MTU even further degrades the peak throughput even more.

Adapter Queue Lengths

The high throughput of a switch can cause problems with other network adapters connected to gateway nodes on the SP system. There is a fixed amount of memory on each network adapter for staging network traffic. The transmit adapter queue length specifies the maximum number of packets for each adapter (except a switch).

The transmit adapter queue length is tunable through SMIT, and the typical default size is too small to handle SP system traffic. The typical value for this queue is 30 to 64, of which each entry can handle a maximum of the MTU of that adapter. For an Ethernet adapter, a queue length of 30 Ethernet packet works out to be a maximum total data size of 45000 bytes. This is less than a single packet from a switch.

If transmit adapter length queue size is exceeded, subsequent packets are thrown away by the adapter device driver and result in dropped packets. We recommend that the transmit adapter length queue size for all network adapters on all nodes of the SP system, as well as the control workstation, be set to their maximum.

A problem with adapter queue sizes is indicated by `TCP/IP retries on the node` in the **netstat -s** output or `Ierr` and `Oerr` in the **netstat -i** output for the adapter. This indicates that the adapter queue size is too small, or the amount of TCP/IP traffic too much for the adapter. Because **netstat -s** is for TCP/IP, and not for a specific adapter, you will have to look at the amount of errors in **netstat -i** for each adapter to determine which adapter is the one causing the problem. On AIX 4.1.4 and lower systems, view adapter errors using the **entstat**, **tokstat**, and **fddistat** commands.

Another way to tune IP for the adapter queue is to reduce the TCP/IP window and **udp_sendspace** so that it is less than the transmit queue size times the segment size on the network. This usually results in lower throughput and higher system overhead for network traffic.

For Ethernet, Token Ring and FDDI, the adapter statistics, including a count of dropped packets, can be seen by using the **entstat**, **tokstat**, and **fddistat** commands, respectively. These commands are AIX 4.1 commands and are not found on AIX 3.2.5 systems.

Tuning For Single Parent–Multiple Child Applications

Some applications, including the Parallel Operating Environment (POE), consist of a parent or master node with multiple child or remote nodes. Such applications have a potential for the child nodes all sending large numbers of messages to one parent node. Figure 67 on page 442 is a simplified view of a multiple-child application.

The following information on this configuration uses the Parallel Operating Environment (POE) as an example. The problem and solution are applicable to any parallel application that uses datagram sockets.

In a POE job that sends a large number of messages to standard out and standard error, the amount of traffic generated can use up a lot of mbufs on the system. The following tunables are recommended as initial values for systems that run large POE sessions or other applications that use similar resources:

- **lowclust** = 60 (AIX 3.2.5 nodes only)
- **lowmbuf** = 200 (AIX 3.2.5 nodes only)
- **thewall** = 8192 (AIX 3.2.5) or 16834 (AIX 4.1)
- **mb_cl_hiwat** = 524 (AIX 3.2.5 nodes only)
- **tcp_sendspace** = 65536
- **tcp_recvspace** = 65536
- **rfc1323** = 1

These are only initial tunable values. For large jobs, depending on the number of messages from the remote nodes, the value for **thewall** may need to be increased. If the master node gets a count of requests for memory denied in the **netstat -m** output, you will need to increase the **thewall** value. On AIX 3.2.5 nodes, the value for **mb_cl_hiwat** will have to be increased in proportion to the increase of **thewall**.

These initial values will have a negative affect on the performance of large data transfers across a switch.

On any task assigned to the nodes assigned in a POE session, anything sent to standard out or standard error is routed back to the master node and then on to the user's session. Each of these messages requires its own mbuf. If you have a large parallel job, one using 32 or more nodes, the master node mbuf pool can be flooded by the traffic from the remote nodes. Because messages from standard out and standard error are synchronous, or blocking transmits, large numbers of messages will end up slowing down the execution of the program.

If messages are placed in the code that record a high number of events, the SP Ethernet (used to pass messages from the remote nodes to the master node, as well as for other uses) can become saturated. And, if you are connected to the master node either explicitly or implicitly using the SP Ethernet, you double the amount of traffic you are sending through the SP Ethernet. This causes backup of the network traffic and a rapid increase of mbufs needed on the master node.

One way to reduce the load on the Ethernet, as well as the mbuf requirements on the master node, is to reduce the size of the **tcp_recvspace** tunable on the master node. As explained earlier, the **tcp_recvspace** size determines the amount of data that a node can have outstanding on a socket. Shortening the TCP window

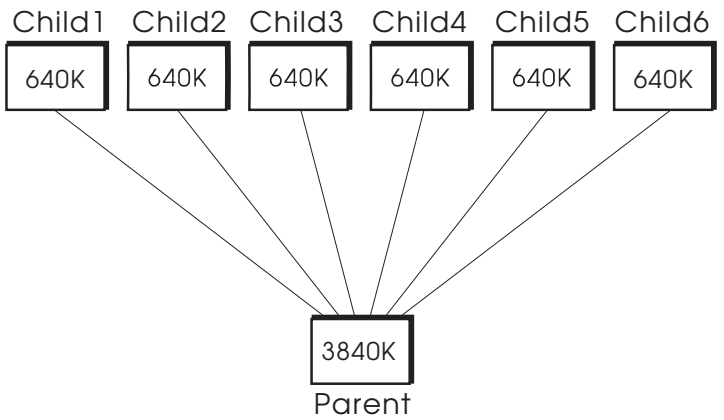


Figure 67. Simplified View of a Single Parent–Multiple Child Application

reduces the amount of traffic on the SP Ethernet. The traffic that would otherwise flood the SP Ethernet is held on the remote node until it can be transmitted. This prevents flooding the SP Ethernet and reduces the amount of mbufs needed on the master node. However, it causes traffic to back up on the remote nodes, and can cause the application to slow down.

Another side effect of using a smaller **tcp_recvspace** is that other users of TCP/IP, particularly those used for large data transfers, will see a decrease in data transfer throughput.

At this point, you need to decide whether to affect the performance of large data transfers from potential POE remote nodes or try to increase the mbufs on the master node. Increasing the amount of mbufs on the master node runs the risk that you will flood the SP Ethernet and cause other problems on the system.

A better solution is to set the socket **SO_RECV** size from within an application, so other traffic on the node is not affected.

Consolidated Systems

Consolidated systems present a larger tuning challenge than homogeneous systems. Consolidated systems are those in which nodes assume several different roles. In such a case, make a full examination of what is running on each node, the traffic through the node, and any interaction between applications. Defining an appropriate set of tunables for a consolidated system requires accounting for everything running on the SP system.

External Servers

External servers can restrict performance on the SP system. For example, some external servers may not support **rfc1323**, and, therefore, can use a maximum TCP/IP window of only 65536 to transfer data to or from an SP system.

In some cases external servers have to communicate with additional external servers. When such external servers establish a connection, the negotiated tunables may be set to small values. For example, the TCP/IP window size value is set to the least common denominator. In the case of an external server having a **tcp_recvspace** of 16384, that is the TCP/IP window size used. Such a small TCP/IP window size provides slow throughput if the sending node is on the far side of a switch from a gateway node.

Subsystem Tuning

This section provides information for specific subsystems in an SP environment.

Tuning for Large Data Transfers

Several typical SP applications and utilities transfer large blocks of data across a switch using TCP/IP. These applications and utilities include:

- ftp
- rcp

To achieve peak data transfer across a switch using TCP/IP, you need to increase the tunables that affect buffer sizes, queue sizes, and the TCP/IP window. The tunables to adjust are:

- **lowclust** (AIX 3.2.5 nodes only)
- **lowmbuf** (AIX 3.2.5 nodes only)
- **thewall**
- **mb_cl_hiwat** (AIX 3.2.5 nodes only)
- **tcp_sendspace**
- **tcp_recvspace**
- **rfc1323**

To get the best TCP/IP transfer rate, you need to size the TCP/IP window large enough to keep data streaming across a switch without stopping the IP stream. The switch has a Maximum Transmission Unit (MTU) of 65520 bytes. This is the largest buffer of data that it can send. When using TCP/IP, TCP will send as many buffers as it can until the total data sent without acknowledgment from the receiver reaches the **tcp_sendspace** value. Experimentation has found that having 4 to 10 buffers as the size of the window allows TCP/IP to achieve peak transfer rate. (The smaller number applies to AIX 4.1 and the larger to AIX 3.2.5.) Also, the faster nodes require a greater number of buffers. If you set **tcp_sendspace** and **tcp_recvspace** to 655360, it can hurt the performance of the other network adapters connected to the node.

Gateway or Router Nodes

Gateway or router nodes direct traffic between external networks and the SP system. Because different types of networks can be connected to router nodes, plan carefully. For slow speed networks such as Ethernet and Token Ring, several adapters can be configured. For higher performance networks, however, we recommend a single adapter.

If a lot of traffic is routed through a router node, any parallel job using that node will be affected. We recommend that router or gateway nodes not be assigned to parallel jobs unless there are no other nodes available, or the amount of expected network traffic is small.

Network File System (NFS)

Three aspects of NFS that are particularly challenging to tune for large SP system configurations are:

- The number of concurrent NFS requests to an NFS server
- The number of NFS daemons (**nfsd** and **biod**) running
- The NFS socket size

The first problem related to NFS that we generally see in large SP system configurations occurs when a single node is acting as the NFS server for a large number of nodes. In this scenario, the aggregate number of NFS requests can overwhelm the NFS socket or **nfsd** daemons on the server, or enough daemons cannot be configured.

When NFS is configured on a node, by default AIX configures 8 **nfsd**s and 6 **biod**s. The **nfsd**s handle NFS requests between a client node and the server. The **biod**s handle block I/O, primarily in NFS writes from the client node.

For the server configuration, **nfsd**s are the primary concern. If you have a 64 node configuration and configured one NFS server for the other 63 nodes, and if all 63 client nodes made an NFS request at the same time, you will need at least 63 **nfsd**s on the server. If you had only 8 **nfsd**s, as in the default configuration, then 55 NFS requests would have to wait and they could time out. Obviously, average NFS response time will be very slow.

However, there is a limit on how many **nfsd**s you can configure before the amount of processing on behalf of the NFS traffic overwhelms the server. Generally, when you configure more than 64 **nfsd**s you start to see NFS performance degradation, depending on the characteristics of the NFS traffic. The size of the requests, the mix of NFS operations, and the number of processes on each node that generates NFS requests influence the amount of NFS performance degradation. Examine **nfsstat** command output to check indications that you are generating too much NFS traffic to a server. If you see RPC timeouts or RPC retransmits when checking on the client nodes, then you probably are overwhelming the current number of **nfsd**s or **nfs_socket** size on the server.

Determining the number of **biod**s to configure on the client node is a bit complicated. The **biod**s are used when writing to an NFS server from a client node. Typically you'd configure enough **biod**s on a client node for the NFS write traffic that you expect. However, NFS has a limit of a maximum of 6 **biod**s on a client node per NFS mounted file system. If you have only one process on a client node writing to a single file system on the server, there is no reason to increase the number of **biod**s. And, if you have only one remote mounted NFS file system on each client, you can use only 6 **biod**s no matter how many processes are writing to the file system.

The limitation to 6 **biod**s on a client node per NFS mounted file system is imposed by NFS Version 2.0 (which is packaged with AIX 4.1). In NFS Version 2.0, all NFS writes use a fully synchronous 4K write model. When you write to an NFS mounted file system, you can send only 4K at a time from the client to the server. If you have only one I/O outstanding at a time, this is very slow. In a fully synchronous model the data must be written on the disk at the server before the client can send the next 4K block. To speed this up a bit, NFS allows staging 6 requests at the client per file system, but no more. Given the time it takes to write a block on disk in AIX, one block at a time is very slow.

If there is more than one NFS mounted file system on your client nodes, you can configure up to 6 **biod**s per file system. If the number of **biod**s gets too high, performance will suffer in the same way as when too many **nfsd**s are configured on the server. Again, a general rule is more than 64 **biod**s will start to degrade performance.

In general the same rules apply for **nfsd**s on the client node as on the server node. However, we don't recommend that you use the same number of **nfsd**s on the client nodes as the servers when a large number of nodes all use the same server (16 clients to 1 server for example). If all nodes have 16 **nfsd**s, each **nfsd** on the server can potentially be sent 16 NFS requests. In this situation, monitor **nfsstat** on both the client and server nodes to look for RPC timeouts and retransmits.

If you already have a lot of **nfsd**s configured on the server, the best solution is to split the NFS server duties up across several nodes, and keep the potential number of concurrent NFS requests below 64. As a general guideline, configuring 64 **nfsd**s on a server usually works well on the SP system.

For more detailed information regarding tuning NFS, refer to the *AIX 3.2.5 and 4.1 Performance Monitoring and Tuning Guide*, SC23-2365, and the *AIX System Management Guide: Communications and Networks*, GC23-2487.

ADSTAR Distributed Storage Manager (ADSM)

When running ADSM on the SP system, the performance of a backup or restore is affected by several tunable settings. However, depending on the backup destination, or restore source, the tunables and their settings are different. The following section lists the settings to initially try for running the ADSM server on MVS over an ESCON channel and running the ADSM server on an SP system node. Because ADSM is not always the only application running on a client node or server, some of these values may not be reasonable in your environment.

There are up to four places where various tunables can be set to optimize the performance of ADSM. The following sections list these places, and the initial values to use for ADSM over a switch and ADSM over ESCON to MVS.

In the following tables, use the switch settings for ADSM clients on another node. Use the ESCON to MVS settings for going to MVS on the client nodes.

SP Client Node

The following are the **no** tunable values that will achieve the best ADSM performance on the client nodes.

Parameter	Switch	ESCON to MVS
sb_max	1310720	131072
thewall	6000	6000
lowmbuf (3.2.5 only)	500	500
lowclust (3.2.5 only)	200	200
mb_cl_hiwat (3.2.5 only)	1200	1200
rfc1323	1	N/A
tcp_mssdflt	N/A	4056

If you already have a larger value for **thewall**, then you need to keep the larger **thewall** value, but keep the ratio between **mb_cl_hiwat** and **thewall** the same as in the above settings.

The following are the entries to use in the **/usr/lpp/adsm/bin/dsm.sys** file on the ADSM client nodes.

Parameter	Switch	ESCON to MVS
TCPWindowsize	256	64
TCPBuffsize	32	32
txnbyte	25600	25600
tcpnodelay	Y	Y

SP ADSM Server Node

The following are the **no** tunable values that will achieve the best ADSM performance on the SP ADSM server node.

Parameter	Switch	ESCON to MVS
sb_max	1310720	N/A
thewall	6000	N/A
lowmbuf (3.2.5 only)	500	N/A
lowclust (3.2.5 only)	200	N/A
mb_cl_hiwat (3.2.5 only)	1200	N/A
rfc1323	1	N/A
tcp_mssdflt	N/A	N/A

If you already have a larger value for **thewall**, you need to keep the larger **thewall** value, but keep the ratio between **mb_cl_hiwat** and **thewall** the same as in the above settings.

The following are the entries to use in the **/usr/lpp/adsm/bin/dsmserv.opt** file on the SP system ADSM server node.

Parameter	Switch	ESCON to MVS
TCPWindowsize	256	N/A
TCPBuffsize	32	N/A
txnbyte	25600	N/A
tcpnodelay	Y	N/A

ESCON Gateway Node

The following are the **no** tunable values that will achieve the best ADSM performance on the node acting as the gateway across ESCON.

Parameter	Switch	ESCON to MVS
sb_max	N/A	131072
thewall	N/A	6000
lowmbuf (3.2.5 only)	N/A	500
lowclust (3.2.5 only)	N/A	200
mb_cl_hiwat (3.2.5 only)	N/A	1200
rfc1323	N/A	N/A

If you already have a larger value for **thewall**, you need to keep the larger **thewall** value, but keep the ratio between **mb_cl_hiwat** and **thewall** the same as in the above settings.

You will also need to set the MTU size of the ESCON interface to 4096. To do that you can use:

```
ifconfig es0 -mtu 4096
```

MVS ADSM Server

The following tunable settings on MVS will get you peak throughput either across a switch, or across a single ESCON channel to MVS. Your actual throughput will be determined by the speed of the media being backed up to, or restored from, the SP system server or MVS server.

Table 17. TCP/IP(MVS) parameters/settings

DATABUFFERPOOLSIZE	240	32768
LARGEENVELOPEPOOLSIZE	240	8192

ESCON

If you are using CLIO to pass data from MVS to AIX, then TCP/IP over the switch to the thin node, you have to be careful on buffer sizes used. CLIO uses a 60K buffer to pass data in either direction, either to or from MVS. If the application running on the wide node is a catch and forward application, taking incoming traffic from one side, and passing it to the other side, then from the settings below, the TCP/IP packet size used across the switch is 64K. This can result in fragmentation, or re-assembly problems in the application on the wide node.

To achieve peak TCP/IP throughput over a switch and through ESCON to MVS, set the **tcp_mssdflt** size on the remote nodes to **4056** and the **MTU** of the ESCON interface to **4096**.

PCA Card and CLIO/S Tuning on the SP System

There are two sets of tunables to be concerned with when tuning the SP system for running with a PCA attached to MVS: the cat0 interface and the network mbufs.

Note: The PCA card is currently supported only on AIX 3.2.5. This section is included for your use if you are running AIX 3.2.5 in a system partition.

In setting up the cat0 interface, you need to define the number of buffers that are used for transmitting and receiving data from the channel. To display these parameters, enter the following command on the SP node:

```
lsattr -E -l cat0
```

This output looks like this:

```
dma_bus_mem  0x00F40000 Bus memory address             False
bus_intr_lvl 4          Bus interrupt level            False
dma_lvl      0x6        DMA arbitration level          False
xmitsz       4096       Transmit buffer SIZE
xmitno       26         NUMBER of transmit buffers
recvsz       4096       Receive buffer SIZE
recvno       26         NUMBER of receive buffers
rdto         0          Receive data transfer OFFSET
speed        4          Channel SPEED
first_sc     0x90       STARTING subchannel address
num_sc       8          NUMBER of subchannel addresses
clawset      0x90       CLAW mode subchannel set
switch       online     Online/Offline SWITCH
indicator    online     Online/Offline INDICATOR
ucode_type   1          N/A
host_name    KN1MVS     N/A
adapter_name P24N11A1   N/A
```

The items of interest are the **xmitno** and **recvno**. These parameters set up how much space is allocated in the PCA card buffer space for transmits and receives. In the example above we have set up the default of the same amount for both transmits and receives. If your system shows a large difference between transmits and receives rates, increase the number for the direction that is slow. You must not use more than 52 total buffers, assuming that **xmitsz** and **recvsz** are set to 4096. There is a total of 208K on the card; exceeding this total will render the interface unusable.

You also must not use fewer than 15 buffers for either transmits or receives if you are using CLIO/S across the interface. Because CLIO/S uses a 60K buffer size, you need a 60K minimum to store the data on the card. If you use fewer than 15 buffers, you will see transmission drop outs, stalls, or even communication shutdown on the node.

To change the number of buffers you can use the **chdev** command on the node. The command is:

```
chdev -P -l cat0 -a xmitno=37 -a recvno=15
```

You will need to reboot the node after issuing this command.

You can use whatever number of buffers you want as long as you do not exceed 52 4K buffers, or 208K total.

The **thewall** and **lowmbuf** tunables are for total mbufs in the system. Because CLIO/S over the PCA uses at least 4K buffers, it uses the mbuf clusters not individual mbufs. These tunables are the **lowclust** and **mb_cl_hiwat** tunables.

For total mbufs, you should make sure that ANY communication interface has enough mbufs available. The total mbufs are for all Ethernet, token-ring, PCA and switch interfaces. If all these interfaces need more mbufs than are set using **thewall**, it will need to be increased.

To change the **thewall** value, enter:

```
no -o thewall=12228
```

If you suspect that you are running out of mbufs, look at the output of the **netstat -m** statistics. This output looks like:

```
527 mbufs in use:
        252 mbufs allocated to data
        11 mbufs allocated to packet headers
        100 mbufs allocated to socket structures
        137 mbufs allocated to protocol control blocks
        15 mbufs allocated to routing table entries
        8 mbufs allocated to socket names and addresses
        4 mbufs allocated to interface addresses
252/401 mapped pages in use
1735 Kbytes allocated to network (65% in use)
126 requests for memory denied
0 requests for memory delayed
0 calls to protocol drain routines
```

This listing shows entries in `mbufs allocated to data` and `requests for memory denied`. These two counters are the ones to watch. If you are getting an increasing count under `requests for memory denied`, increase the value for **thewall**. The `mbufs allocated for data` shows the current number of mbufs containing data. If this count gets high, examine the `mbufs denied` value. The `mbufs denied` counts the number of times an mbuf request was denied because one was not available since the system last booted. The `mbufs allocated to data` is the count at the time the data was collected.

You will normally see approximately 45 `mbufs allocated to data` when running only CLIO/S over the PCA. This number will be affected by traffic to other interfaces, as well as stalls on any interface. A stall can show up as an increase in `mbufs allocated for data`, and may stay at a high value.

If you run out of mbufs and generate a lot of `mbuf denied` counts, the interfaces can go down.

You might want to try the following:

1. Add buffers using the **thewall** setting default of 2048. If you are running other interfaces, such as a switch, you probably have to increase the value more than if you are not running a switch.
2. Increase the **lowclust** value because CLIO/S uses mainly clusters of mbufs.
3. Increase the **mb_cl_hiwat** to maintain a large enough number of clusters for CLIO/S traffic.

These are only initial recommendations. Depending on the type of PCA interface traffic, and other network impacts, these will need to be changed.

Additional information on general mbuf tuning can be found in *AIX Versions 3.2 and 4.1 Performance Tuning Guide*, SC23-2365.

The Communications Low-Level Application Programming Interface

© Copyright IBM Corp. 1995, 1997

Chapter 29. Understanding and Using the Communications Low-Level Application Programming Interface (LAPI)

This chapter introduces you to the concepts and use of the Communications Low-level Application Programming Interface (LAPI).

Introducing the LAPI

The LAPI is a non-standard application programming interface designed to provide optimal communication performance on the SP Switch. It is based on an "active message style" programming mechanism that provides a one-sided communications model (that is, one process initiates an operation and the completion of that operation does not require any other process to take a complementary action). The LAPI library provides **PUT** and **GET** function and a general "active message" function that allows programmers to supply extensions by means of additions to the notification handlers. The LAPI is designed for use by libraries and power programmers for whom performance is more important than code portability.

Why Use the LAPI?

The LAPI provides the following advantages:

- Performance - The LAPI provides basic function for optimal performance. It is designed to especially provide low latency on short messages.
- Flexibility - The LAPI's one-sided communications model provides flexibility because the completion of an operation by one process does not require any other process to take a complementary action. Also, the LAPI provides a more primitive interface (than either MPI or IP) to the SP Switch, giving the programmer the choice of how much additional communications protocol needs to be added.
- Extendibility - The LAPI supports programmer-defined handlers that are invoked when a message arrives. Programmers can customize the LAPI to their specific environments.

General characteristics provided by LAPI are:

- Reliability - The LAPI provides guaranteed delivery of messages. Errors not directly related to the application are not propagated back to the application.
- Flow control
- Support for large messages
- Non-blocking calls
- Interrupt and polling modes
- Efficient exploitation of switch function
- By default, ordering is not guaranteed

© Copyright IBM Corp. 1995, 1997

General LAPI Functions

LAPI functions are divided into three parts:

1. A basic “active message” infrastructure that allows programmers to install a set of handlers that are invoked and executed in the address space of a target process on behalf of the process originating the active message. This generic interface allows programmers to customize the LAPI function to their unique environment.
2. A set of defined functions that is complete enough to satisfy the requirements of most programmers. These defined functions make the LAPI more usable and at the same time lend themselves to efficient implementation because their syntax and semantics are known.
3. A set of control functions for the initialization and eventual orderly shutdown of the LAPI layer.

Understanding the LAPI

To help you achieve a fuller understanding of the LAPI, this section presents further details on the “active message” infrastructure and the defined set of functions. In addition, concepts important to understanding the LAPI are explained.

The “Active Message” Infrastructure

“Active message” was selected as the underlying infrastructure for the LAPI. The “active message” infrastructure has the following characteristics:

- The “active message” includes the address of a user-specified handler. When the active message arrives at the target process, the specified handler is invoked and executes in the address space of the target process.
- The “active message” optionally may also bring with it the user header and data from the originating process.
- Operations are unilateral in the sense that the target process does not have to take explicit action for the active message to complete.
- Storage buffers for arriving data need to be provided by the invoked handler.

Writing Handlers

The ability for programmers to write their own handlers provides a generalized, yet efficient, mechanism for customizing the interface to one's specific requirements. The user is responsible for protecting shared structures and buffers where necessary by using the locking structures available in the AIX p-threads library.

The LAPI supports messages that can be larger than the size supported by the underlying SP Switch subsystem. Therefore, the data sent with the “active message” may arrive at the target in multiple packets and, further, these packets can arrive out of order. This situation places some requirements on how the handler is written.

When the “active message” brings with it data from the originating process, the architecture requires that the handler be written as two separate routines:

1. A “header handler” function. This function is specified in the “active message” call. It is called when the message first arrives at the target process and provides the LAPI dispatcher (the part of the LAPI that deals with the arrival of

messages and invocation of handlers) with an address where the arriving data must be copied, the address of the optional "completion handler," and a pointer to the parameter that will be passed to the "completion handler".

2. A "completion handler" that is called after the whole message has been received.

An Example of LAPI "Active Message" Function

In this example, a programmer writes a handler for the LAPI "active message" interface. Refer to *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on the **LAPI_Amsend** subroutine.

1. The desired function (accumulate) is to add vector (S) to another (D) on the target node and put the results in the vector at the target:

   ```
   D[0..N-1] = D[0..N-1] + S[0..N-1]
   ```

   where,
   S[N] is a vector of length N in the address space of the origin process (*origin_process*)
   D[N] is a vector of length N in the address space of the target process (*target_process*)

2. The generic "active message" call is defined as **LAPI_Amsend** (*hndl, tgt, hdr_hdl, uhdr, uhdr_len, udata, udata_len, tgt_cntr, org_cntr, cmpl_cntr*)

3. Before making the active message call, you must obtain the address of the target counter (*target_cntr_addr*) and the address of the header handler to be executed on the target process (*accumulate_addr*). The address of the header handler is obtained by the **LAPI_Address** function.

4. Initialize the *udhr* based on the header expected by *accumulate*. For example, the structure of *udhr* could be:

   ```
   typedef struct {
      void *target_addr;
      uint length;
   } put_add_hdr_t;

   put_add_hdr_t uhdr;

   uhdr.target_addr = D;
   uhdr.length = N;
   ```

5. Make the specific call

   ```
   LAPI_Amsend (hndl, target_process, accumulate_addr,
   &uhdr, sizeof(put_add_hdr_t), &S[0],
   N*sizeof(S[0]), target_cntr_addr, &origin_cntr, &completion_cntr)
   ```

6. When this message is received at the target (assuming that the entire origin data is contained within a packet), the *accumulate* handler you specified is invoked by the dispatcher. The structure of the header handler is:

   ```
   void *header_handler (lapi_handle_t hndl, void *uhdr,
      uint uhdr_len, uint msg_len, completion_handler_t,
      *completion_handler, void *user_info)
   ```

   The structure of the completion handler is:

   ```
   void completion_handler (lapi_handle_t hndl,
      void *user_info)
   ```

7. If any state information about the message is required by the completion handler, the information required must be saved in a user buffer by the header handler. The header handler passes the address of this buffer to the dispatcher through the parameter *user_info*. The dispatcher uses this pointer as a parameter (*user_info*) for the completion handler.
8. For this example operations performed at the target process are:
   - Within the dispatcher:
     a. The LAPI header is read.
     b. *uhdr* and *uhdr_len* are extracted from the LAPI header.
     c. The header handler is invoked.

        ```
        buf = (*accumulate_addr)(hndl, uhdr, uhdr_len, msg_len,
               &completion_handler, &user_info);
        ```

     d. *udata* is copied into *buf*.
     e. The completion handler is invoked.

        ```
        (*completion_handler)(&hndl, user_info);
        ```

     Note: If the message was not contained within a packet, the LAPI layer will save the necessary information and will invoke the completion handler after all the *udata* has arrived and copied into *buf*
   - User defined functions:
     – Header handler:

       ```
       accumulate(hndl, uhdr, uhdr_len, msg_len, completion_handler, user_info)
       {
          buf = addr where incoming data should be buffered
          save (target_addr=D, length=N, buf) in user_info
          completion_handler = complete_accumulate
          return buf
       }
       ```

     – Completion handler:

       ```
       complete_accumulate (hndl, user_info)
       {
          retrieve required data (namely D,N and buf) from user_info;
          for (i=0; i<N; i++) D[i]=D[i]
            + buf [i];
          return
       }
       ```

The accumulate handler is the header handler and is called by the LAPI layer when the message first arrives at the target process. The header handler saves the information required by *complete_accumulate* (*target_addr*, *length*, and *buf*) in *user_info* and passes back a pointer to the *complete_accumulate* handler as *user_info*. Additionally, the header handler returns address of a buffer *buf*.

Large active messages are generally transferred as multiple packets. In this case the LAPI layer stores the incoming data in the packets as they arrive into *buf*. When all the data has been received, it calls the *complete_accumulate* function which uses *user_info* to access the two vectors, adds them and stores them at the desired location. After the return from the *complete_accumulate* routine, the LAPI layer increments *tgt_ctr*. The *origin_cntr* is incremented when it is safe to return the origin buffer back to the user.

The *cmpl_cntr* is incremented after the completion handler has completed execution. The *cmpl_cntr*, therefore, is a reflection, at the origin, of the *tgt_cntr*.

The Defined Set of Functions

Fundamentally, the defined set of functions for the LAPI provides a Remote Memory Copy (RMC) interface. The primary characteristics of the defined set of functions provided by LAPI are:

- The basic data transfer operations are memory to memory copy operations that transfer data from one virtual address space to another virtual address space.
- The operations are unilateral. That is, one process initiates an operation and the completion of the operation does not require any other process to take some complementary action. (This is unlike a send and receive operation, where a send requires a complementary receive with matching parameters to be posted for completion.)
- The operations support both "pull" and "push". The **LAPI_Get** operation copies data from the address space of the target process into the address space of the origin process. The **LAPI_Put** operation copies data into the address space of the target process from the address space of the origin process.
- The initiating process specifies the virtual address of both the source and destination of the data (unlike a send and receive process where each side specifies the address in its own address space). To avoid the limitation of requiring that the address maps on the different processes be identical, the LAPI provides the **LAPI_Address_init** mechanism by which the different communicating processes can exchange information regarding the address map of shared data objects.
- Because data transfer operations are unilateral and no synchronization between the two processes is implied, additional primitives are provided for explicit process synchronization when it is necessary for program correctness.
- Functions are provided to detect completion and to enforce ordering.

Important LAPI Concepts

To use the LAPI, it is important to understand the following concepts:

- Origin and target
- Blocking and non-blocking calls
- Completion of communication operation
- Message ordering and atomicity
- Error handling
- Progress

Origin and Target

"Origin" denotes the task (or process or processor) that initiates a LAPI operation (**PUT**, **GET**, or "active message".). "Target" denotes the other task whose address space is accessed. Although multiple tasks may run on a single node, it is convenient to think of each task as running on a different node. Therefore the origin task may also be referred to as the "origin node" and the target task as the "target node". The origin and target can be the same for any of the calls, but if the origin and target data areas overlap, the results are undefined.

Blocking and Non-Blocking Calls

A blocking procedure is one that returns only after the operation is complete. There are no restrictions on the reuse of user resources.

A non-blocking procedure is one that may return before the operation is complete and before the user is allowed to reuse all the resources specified in the call. A non-blocking operation is considered to be complete only after a completion testing function, such as **LAPI_Waitcntr** or **Lapi_Getcntr**, indicates that the operation is complete.

Completion of Communication Operation

A communication operation is considered to be complete, with respect to the buffer, when the buffer is reusable.

A **PUT** is complete with respect to the *origin* buffer when the data has been copied out of the buffer at the origin and may be overwritten. A **GET** is complete with respect to the *origin* buffer when that origin buffer holds the new data that was obtained by **GET**.

A **PUT** is complete with respect to the *target* buffer when the new data is available at the target buffer. A **GET** is complete with respect to the *target* buffer when the data has been copied out of the buffer at target and the target task may overwrite that buffer.

Communication Behaviors: Two communication behaviors support two different definitions of "completion":

- In **standard** behavior, a communication operation is defined as complete at the *origin* task when it is complete with respect to the *origin* buffer; it is complete at the *target* task when it is complete with respect to the *target* buffer.
- In **synchronous** behavior, a communication operation is defined as complete at the *origin* task when it is complete with respect to both the *origin* buffer and *target* buffer. It is complete at the *target* task when it is complete with respect to the *target* buffer.

The LAPI defines both standard and synchronous behaviors for **PUT** operations. The LAPI defines only synchronous behavior for **GET** operations.

Message Ordering and Atomicity

Two LAPI operations that have the same *origin* task are considered to be ordered with respect to the *origin* if one of the operations starts after the other has completed at the *origin* task. Similarly, two LAPI operations that have the same *target* task are considered to be ordered with respect to the *target* if one of the operations starts after the other has completed at the *target* task. If two operations are not ordered, they are considered concurrent. The LAPI provides no guarantees of ordering for concurrent communication operations. The LAPI does provide mechanisms which an application can use to guarantee order.

As an example, consider the case where a node issues two standard behavior **PUT** operations to the same target node, where the targets overlap. These two operations may complete in any order, including the possibility of the first **PUT** overlapping the second, in time. The contents of the overlapping region will be undefined, even after both **PUT**s complete. Using synchronous behavior for both **PUT** operations, (waiting for the first to complete before starting the second) will

ensure that the overlapping region contains the result of the second after both **PUT**s have completed.

Error Handling

If an error occurs during a communications operation, the error may be signaled at the origin of operation, or the target or both. Some errors may be caught before the communication operation begins, and these will be signaled at the origin. However, some errors will not occur until the communication is in progress (a segmentation violation at the target, for example); these may be signaled at either or both ends of the communication.

Progress

All LAPI operations are unilateral by default and can complete successfully or fail, independent of the actions of other tasks. Specifically, a LAPI operation to a particular target should complete even if the target is stuck in an infinite loop (that is, when the target process is in interrupt mode).

Using the LAPI

In general, LAPI functions:

- Are non-blocking calls
- Provide both polling and interrupt mode
- Signal completion by incrementing counters at each end
- Provide both C and Fortran bindings

Complementary functions provide for checking completion of operations and for enforcing relative ordering if required. Additionally, functions allow processes to exchange addresses that will be used in LAPI operations.

Specific LAPI Functions

The LAPI provides the following specific functions.

Active Message

Active message function (**LAPI_Amsend**) is a non-blocking call that causes the specified active message handler to be invoked and executed in the address space of the target process. Completion of the operation is signaled if counters are specified. Both standard and synchronous behaviors are supported. The **LAPI_Amsend** function provides two counters (*org_cntr* and *cmpl_cntr*) which can be used to provide the two behaviors. The *org_cntr* is incremented when the origin buffer can be reused (standard). The *cmpl_cntr* is incremented after the completion handler has completed execution (synchronous).

Data Transfer

Data transfer functions are non-blocking calls that cause data to be copied from a specified region in the origin address space to the specified region in the target address space (in the case of a **LAPI_Put** operation) or from a specified region in the target address space to a specified region in the origin address space (in the case of a **LAPI_Get** operation). Completion of the operation is signaled if counters are specified. Both standard and synchronous operations are supported for **PUT**. Only synchronous operation is possible in the case of **GET**. Standard **PUT** is provided by incrementing the *org_cntr* when the origin buffer can be reused.

Synchronous **PUT** is provided by incrementing the *cmpl_cntr* after the data has been written into the target buffer. The **LAPI_Getcntr** (or **LAPI_Waitcntr**) function should be used in conjunction with the *org_cntr* and *cmpl_cntr* counters to guarantee the respective standard and synchronous behavior of the LAPI **PUT** operations.

Synchronizing

The **LAPI_Rmw** function is used to synchronize two independent operations such as two processes sharing a common data structure. The operation is performed at the target process and is atomic. The operation takes a variable from the origin and performs one of four selected operations on a variable from the target and replaces the target variable with the results of the operation. The original value of the target variable is returned to the origin. **LAPI_Rmw** provides four different read/modify/write operations:

- **SWAP**
- **COMPARE_AND_SWAP**
- **FETCH_AND_ADD**
- **FETCH_AND_OR**

Completion is signaled at the origin if the counter is specified.

Completion Checking

The following counter functions provide the means for a process to manage the completion state of the LAPI operations.

- **LAPI_Waitcntr** - Wait on a counter to reach a specified value and return when the counter is equal to or greater than that value (blocking)
- **LAPI_Getcntr** - Get the current value of a specified counter (non-blocking)
- **LAPI_Setcntr** - Set the counter to a specified value

These functions also provide an efficient means to order the flow of LAPI operations or the use of certain user managed resources (for example, buffers). For example, a series of **PUT**s to a single target and buffer requires that the contents of the buffer at the target remains in step with the order of execution of the **PUT**s at the origin. Using the *cmpl_cntr* counter in the **LAPI_Put** function in conjunction with the **LAPI_Waitcntr** function provides the necessary ordering.

Ordering

LAPI-Fence and **LAPI_Gfence** operations provide a means to enforce the order of execution of LAPI functions. LAPI functions initiated prior to these fencing operations are guaranteed to complete before LAPI functions initiated after the fencing functions. **LAPI_Fence** is a local operation which is used to guarantee that all LAPI operations initiated by the local process and the same process thread are complete. **LAPI_Gfence** is a collective operation involving all processes in the parallel program. **LAPI_Gfence** provides a barrier operation for the parallel program. Both **LAPI_Fence** and **LAPI_Gfence** operations are a data fence that guarantee that the data movement is complete. These are not an operation fence which would need to include “active message” completion handlers completing at the target.

Progress

The **LAPI_Probe** function is used in polling mode to transfer control to the communication subsystem in order to make progress on arriving messages.

Address Manipulation

The **LAPI_Address_init** collective operation allows processes to exchange operand addresses of interest. Such function is required if the processes do not have identical address maps. The **LAPI_Address** is used by Fortran programs when an address needs to be stored in an array. In Fortran there is no concept of "address" (as there is in the C language) and this function gives that ability to Fortran.

LAPI Setup

LAPI_Init and **LAPI_Term** operations are used to initialize and terminate the communication structures required to effect LAPI communications for a specific instance of LAPI. **LAPI_Init** returns a unique handle which is used to associate a communication channel with the LAPI instance. This handle is subsequently passed as a parameter to each of the other LAPI functions. The **LAPI_Term** function is used to terminate a specific instance of LAPI.

Error Handling and Messages

The **LAPI_Init** function provides a means for the user of LAPI to register an error handler. This error handler is specified as part of the *lapi_info* structure parameter which is passed to the **LAPI_Init** function. The **LAPI_Msg_String** function provides the means to translate a LAPI call return code value (integer) into a message string.

LAPI Environment

The **LAPI_Qenv** function is used to query the state of the LAPI communications subsystem. The **LAPI_Senv** function allows the programmer to specify the value of some of the LAPI communications subsystem's environment variables. An important value that can be specified is the interrupt state. The interrupt state is set by specifying **INTERRUPT_SET** as **on** (for interrupt mode) or **off** (for polling mode). The default setting for **INTERRUPT_SET** is **on**.

The LAPI Execution Model

The goal of LAPI is to provide a threads-safe environment and support an execution model that allows for maximum execution concurrency within the LAPI library.

Using the setup function (**LAPI_Init**), a user process establishes a LAPI context. Within a LAPI context, the LAPI library is threads-safe, and multiple threads may make LAPI calls within the same context. The different calls may execute concurrently with each other and with the user threads. However, in reality execution concurrency among these calls is limited by the locking required with LAPI to maintain integrity of its internal data structures and the need to share a single underlying communication channel.

As with any multi-threaded application, coherence of user data is the responsibility of the user. Specifically, if two or more LAPI calls from different threads can execute concurrently and if they specify overlapping user buffer areas, then the result is undefined. It is the responsibility of the user to coordinate the required synchronization between threads that operate on overlapping buffers.

The user application thread, as well as the completion handlers, cannot hold mutual exclusion resources before making LAPI calls; if they do, it is possible to run into deadlock situations.

Because user-defined handlers may be called concurrently from multiple threads, it is the user's responsibility to make them threads-safe.

The LAPI does not currently support multiple user communication threads. User communication thread is defined as a user created thread which makes calls to the LAPI library. Multiple user computational threads are supported. From the user application thread a second thread is created by **LAPI_Init** to process notification events and invoke the LAPI handlers as illustrated in Figure 68.

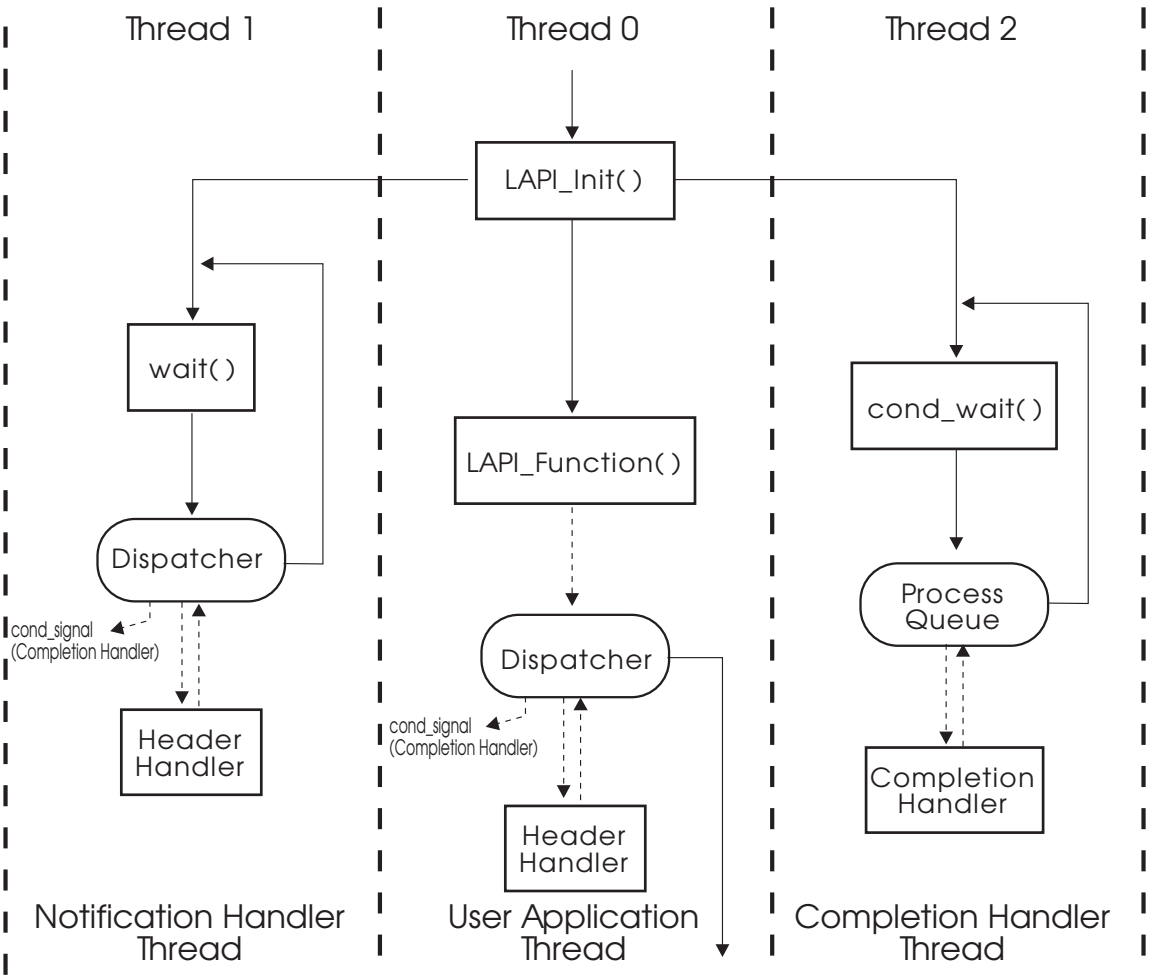


Figure 68. A LAPI Thread Model

The application thread, notification thread and the completion handler thread are shown in Figure 68.

Threads 0 and 1 (the application thread and the notification thread) attempt to invoke the LAPI dispatcher whenever possible; in this way, progress on incoming and outgoing messages can be made while minimizing additional overhead. Most LAPI calls (though not all) made by the application thread also result in the LAPI dispatcher being automatically run. The notification thread waits in the Kernel for the occurrence of a notification event. When an event occurs, the Kernel wakes up

the waiting thread. As shown in Figure 68, after the notification thread returns from waiting in the Kernel, it invokes the LAPI dispatcher.

The LAPI Dispatcher is the central control point that orchestrates the invocation of functions and threads necessary to process outstanding incoming and outgoing LAPI messages.

The LAPI Dispatcher can run from the application's user's thread, from the notification thread or from the completion handler thread. Locking is used to ensure that only one instance of the dispatcher runs at a time to maintain integrity. On incoming messages, the LAPI dispatcher manages the reassembly of data from different packets (which may arrive out-of-order) into the specified buffer, and then invokes the completion handler if necessary.

Thread 2 is created by **LAPI_Init** to execute completion handlers associated with "active messages". Completion handlers are written by users and may make LAPI function calls which in turn will invoke the LAPI Dispatcher. The completion handler thread processes work from the completion handler queue. When the queue is empty the thread waits using a **pthread_cond_wait()**. If an active message (**LAPI_Amsend**) includes a completion handler, the dispatcher queues a request on the completion queue after the whole message has arrived and has been reassembled in the specified buffer; the dispatcher then sends a **pthread_cond_signal** to the completion handler thread. If this thread was in a wait state it will begin processing the completion handler queue, otherwise, if it was not waiting, the thread signal is ignored.

LAPI handlers are not guaranteed to execute one at a time. It is important to note, however, that LAPI calls may execute concurrently with the origin or target or both. The same restrictions stated previously about not holding on to mutual exclusion resources when making LAPI calls still applies.

This discussion of a threads-safe environment and maximum execution concurrency within the LAPI library applies to both the polling and interrupt modes. In polling mode any calls to the communication library attempt to make progress on the context specified in the call. Further, the function **LAPI_Probe** is provided to allow applications to explicitly check for and handle incoming messages.

The execution model of the handlers consists of the following events:

Event	Action
Message Arrival	Copies the message from the network into the appropriate data access memory space.
Interrupt / Poll	Causes an interrupt if required, based on the mode.
Dispatcher Start	Invokes the dispatcher.
New Message Packet	Checks the LAPI header and determines (by checking the receive state message reassembly table) if the packet is part of a pending message or if it is a new message. Calls the header-handler function.

Return from Header-Handler
If the message is contained in more than one packet, the LAPI Dispatcher will log that there is a pending message, save the completion handler address, and save the user's buffer address to be used during the message reassembly of pending message packets.

Pending Message Packet
Copies the message to the appropriate portion of the user buffer specified through the header-handler. If the packet completes the message, the dispatcher queues the completion handler; otherwise the dispatcher returns to check for message arrivals.

Completion Handler
When the completion handler is executed, (after the return from the completion handler) updates the appropriate target counter before continuing.

Allocating Buffers

1. The user allocates as many buffers per origin as wanted.
   - Rule: At origin, if the target buffer allocation is exhausted, wait for the counter of previous requests.
   - Example: Use >= *P* buffers (one per origin). At origin with one request pending. Wait on counter before issuing next request.
2. If the header handler blocks, no further progress is made, including messages pending (that is, the communications adapter is stalled).

Executing LAPI Programs

The LAPI communications application programming interface must be used in conjunction with the IBM Parallel Environment for AIX (PE) product. Specifically the Parallel Operating Environment (POE) component of PE is used to compile and run LAPI parallel programs. POE also provides support for parallel programs to use the Message Passing Interface (MPI) both separately and in combination with the LAPI. The publication *IBM Parallel Environment for AIX: Operation and Use* provides detailed information about using the POE.

Compiling LAPI Programs

As with a serial application, you must compile a parallel C, C++, or Fortran program before you can run it. Instead of using the **cc_r**, **xlC_r**, or **xlf** commands, however, you use the commands **mpcc_r**, **mpCC_r**, or **mpxlf_r**, which support programs using the threaded LAPI library. The **mpcc_r**, **mpCC_r**, and **mpxlf_r** commands not only compile your program, but also link in the POE Partition Manager and PSSP Communication Subsystem (CSS) interfaces. When you later run the program, the CSS libraries will be dynamically linked with the executable program. Subroutines in these libraries enable POE's home node (the node from which the parallel program is invoked) to communicate with the parallel tasks, and tasks with each other.

The following table shows what to enter to compile a LAPI program.

Compile a C program	**mpcc_r program.c -o program**
Compile a C++ program	**mpCC_r program.C -o program**
Compile a Fortran program	**mpxlf_r program.f -o program**

Creating a Static Executable

Note: We discourage you from creating statically bound executables. If service is ever applied that affects any of the CSS libraries, you will need to recompile your application to create a new executable that will work with the new libraries. This could lead to a lot of work and may expose you to potential problems, which would be avoided if dynamic libraries are used.

The following is a sample makefile:

```
COMPILER        = mpcc_r
CFLAGS  = -g

FCOMPILER       = mpxlf_r
FFLAGS  = -g

STATIC = -bnso -bI:/usr/lib/syscalls.exp -bI:/usr/lib/threads.exp \
         -bI:/usr/lpp/ssp/css/libus/fs_ext.exp -lppe_r -lhal_r

all: Get Put Am Rmw Getf Putf Rmwf Amf Get_static

Get: Get.c
        $(COMPILER) $(CFLAGS) -v -o $@ $@.c

Getf: Getf.f
        $(FCOMPILER) $(FFLAGS) -v -o $@ $@.f

Put: Put.c
        $(COMPILER) $(CFLAGS) -v -o $@ $@.c

Putf: Putf.f
        $(FCOMPILER) $(FFLAGS) -v -o $@ $@.f

Am: Am.c
        $(COMPILER) $(CFLAGS) -v -o $@ $@.c

Amf: Amf.f
        $(FCOMPILER) $(FFLAGS) -v -o $@ $@.f

Rmw: Rmw.c
        $(COMPILER) $(CFLAGS) -v -o $@ $@.c

Rmwf: Rmwf.f
        $(FCOMPILER) $(FFLAGS) -v -o $@ $@.f

Get_static: Get.c
        $(COMPILER) $(CFLAGS) $(STATIC) -v -o $@
Get.c

clean:
        rm -rf Get Put Am Rmw Getf Putf Rmwf Amf Get_static
```

Running LAPI Programs

Before running your program, you need to set up your execution environment. There are a number of POE environment variables discussed throughout *IBM Parallel Environment for AIX: Operation and Use* and summarized in the "POE Environment Variables and Command-Line Flags" appendix of that book. LAPI programs must set the environment variable **MP_EUILIB** (or the command-line flag **-euilib**) to **us** (**us** is the User Space communication subsystem). The default setting for **MP_EUILIB** is **ip** (**ip** is the IP communication subsystem). The LAPI supports only User Space communication using the SP Switch. The LAPI does not support using the IP communication subsystem.

All other procedures for running a LAPI program using POE are common with the MPI programs and are described in *IBM Parallel Environment for AIX: Operation and Use*.

Specific LAPI Hints

The LAPI library uses the SIGBUS signal handler. If an application using LAPI also uses SIGBUS, the application's SIGBUS handler should be registered before the call is made to the **LAPI_Init** function.

The **LAPI_Senv** function should be used to enable error checking (**ERROR_CHK** should be set on) during application development and disabled (**ERROR_CHK** should be set off) for normal application use.

When the order of execution between any two LAPI functions within one task of a parallel program needs to be guaranteed, using **LAPI_Waitcntr** between the two LAPI functions will usually be more efficient than **LAPI_Fence**. **LAPI_Fence** requires that all LAPI operations initiated on the current thread before the **LAPI_Fence** be completed before any LAPI operation after the fence is allowed to start. **LAPI_Waitcntr** can be used to indicate the completion of a single LAPI function which had been initiated on the current thread before the **LAPI_Waitcntr**.

The scope of **LAPI_Fence** is per thread. For example, a **LAPI_Fence** which is issued from the completion handler thread will only guarantee that no LAPI operations initiated after the fence (on the completion handler thread) will start until all LAPI operations initiated before the fence have completed. In this case there are no guarantees about the order of LAPI operations initiated from the main application thread.

LAPI_Waitcntr can be used to indicate the completion of a single LAPI function which may have been initiated from an alternate thread (completion handler) within the same task. Therefore the possibility exists to use **LAPI_Waitcntr** to wait for the completion of another LAPI function which is initiated after the call to **LAPI_Waitcntr**.

LAPI_Waitcntr can be used to guarantee order of execution of **LAPI_Amsend** operations which are initiated from a single origin task. When **LAPI_Amsend** operations use the *cmpl_cntr* counter, this counter is incremented after the completion counter (or header handler if a completion handler is not specified) has executed at the target task. **LAPI_Fence** and **LAPI_Gfence** do not provide an indication that **LAPI_Amsend** operations have completed execution at the target.

LAPI_Waitcntr is a blocking call. If a user prefers to avoid this blocking operation a program loop comprised of the sequence

```
LAPI_Getcntr
a check of the value returned from Get
LAPI_Probe
```

will provide an equivalent logical operation and provide the user with added flexibility.

LAPI_Init must be called before any thread (the main thread or the completion handler thread) can make a LAPI call. In addition to this, **LAPI_Address_init** or **LAPI_Gfence** should be the second LAPI call. These two functions provide a barrier which guarantees that all other LAPI tasks have initialized their LAPI subsystems and are ready to receive requests from remote tasks. Failure to provide these barrier functions may result in dropped switch packets, low performance at start-up and unnecessary switch congestion. The instance of the LAPI subsystem should be quiesced before **LAPI_Term** is called to terminate the LAPI instance. This can be done by calling **LAPI_Gfence** before **LAPI_Term**.

When one task of a parallel job (using POE) determines a condition which requires that all tasks of the parallel job be terminated, the task needs to send a process terminating signal (for example, **SIGTERM** or **SIGQUIT**) to itself. The POE daemon (**PMD**) which spawned that task will:

1. Catch **SIGCHILD**
2. Send a message to the POE “home node”
3. The POE “home node” broadcasts a message to all the **PMD**s to terminate the job
4. Each of the **PMD**s sends **SIGTERM** to their child process and waits for **SIGCHILD**
5. Each task termination is reported back to the “home node”
6. The complete POE job ends

General LAPI Hints

Programs using the LAPI library are compiled with the POE commands **mpcc_r**, **mpCC_r** and **mpxlf_r**. The user's program may (but is not required to) create threads. The POE commands also link a POE version of the standard re-entrant C library, **libc_r.a**, which allows synchronization of the tasks at exit. All the PE and CSS libraries are dynamic shared libraries, enhancing the portability of the user's application between operating system levels and machine types. **mpcc_r** invokes the re-entrant C compiler with the compiler option to include the threaded (re-entrant) libraries and to generate thread-aware code. **mpcc_r** links in the threaded LAPI library, including the threaded versions of the POE utility libraries.

A threaded program has more than one independent instruction stream, but all threads share the same address space, file, and environment variables.

Core Dumps: If a task produces a core file, it is written to an appropriately named subdirectory of the user's current directory. The partial dump produced by default does not contain the stack and status information for all threads; thus it is of limited usefulness in trying to diagnose hang conditions. It is possible to request AIX to produce a full core file, but such files are generally larger than permitted by AIX user limits. The communication subsystem alone generates more than 64MB of core information. Thus, if possible, use the attach capability of **dbx**, **xldb** or **pdbx** to examine the task while it is still running.

Hangs: Coordinating the threads in a task requires careful locking and signaling. Program deadlocks waiting on locks that haven't been released are common, in addition to the deadlock possibilities offered by improper use of the LAPI calls.

Use with Non-Thread-Safe Libraries: A threaded LAPI program must meet the same criteria as any other threaded program. It must avoid using non-thread safe functions in more than one thread (for example, **strtok**). In addition, it must use only thread-safe libraries if library functions are called on more than one thread. In AIX Version 4.2.1, the **libm.a** and **libbsd.a** libraries are NOT thread-safe and their use in a program should be examined carefully.

Linking with Libraries Built with libc.a: Compiling a threaded LAPI program will cause the **libc_r.a** library to be used to resolve all the calls to the standard C library. If your program links with a library that has been built using the standard C library, it is still usable (assuming that it provides the necessary logical thread safety) under the following conditions:

- The library has been built using a reference to the C library shared object (this is the default, unless the **-bnso** flag was used to build it)
- The run-time library path resolves the file reference **libc.a** to the POE version of **libc_r.a**.

To explain further, the run-time library path for an executable is composed of the **AIX LIBPATH** environment variable concatenated with the library path string contained in the **a.out** file (and put there when the program was linked). All the shared libraries used by the executable must be found in the directories occurring in this string. Normally the **AIX LIBPATH** environment variable is empty, and the **cc** (**xlc**) compiler creates a string in which **/usr/lib** is the first entry. Thus, the executable looks for **/usr/lib/libc.a.** However, the **cc_r** (**xlc_r**) compiler creates a string starting with **/usr/lib/threads:/usr/lib**. Thus, the executable will look for **libc.a** in **/usr/lib/threads** first. And, in fact, there is exactly one member of **/usr/lib/threads**: **/usr/lib/threads/libc.a**, which is, however, a symbolic link to **/usr/lib/libc_r.a**! Thus, the executable looking for **libc.a** actually loads **libc_r.a**. Calls to symbols in **libc.a** are resolved by the second entry, **/usr/lib/libc_r.a**. Thus, all calls to the standard C library are resolved by the same library, providing a consistent internal state.

mpcc_r does the same sort of thing. **mpcc_r** puts **/usr/lpp/ppe.poe/lib/threads:/usr/lpp/ppe.poe/lib** in the executable's libpath string and expects **/usr/lpp/ppe.poe/lib/threads/libc.a** to be a symbolic link to **/usr/lpp/ppe.poe/lib/libc_r.a**

The **LIBPATH** environment variable can be set by the user directly. The current version of POE (Version 2.3) sets the default LIBPATH as follows: **LIBPATH=$MP_EUILIBPATH/$MP_EUILIB**, and expects that the **mpcc_r** command has caused **/usr/lpp/ppe.poe/lib** to be included in the libpath search string.

Use of Segment Registers (-bmaxdata Restrictions): The User Space LAPI library uses two segment registers of the 10 that are unassigned in the user's AIX process space. Thus, the user may use a maximum of 8 segments (**-bmaxdata=0x80000000**) for extended heap for large data arrays. The MPI library uses 3 segment registers. A program calling both the LAPI and MPI interfaces has a maximum of 5 segments available for shared memory or extended heap.

We recommend that you don't write threaded message passing programs until you are quite familiar with writing and debugging threaded single-task programs.

AIX requires that threaded programs that include <pthread.h> put this include first, before includes for <stdio.h> or other system includes. This is because <pthread.h> defines some conditional compile variables that modify the code generation of subsequent includes, particularly <stdio.h>. Note that <pthread.h> is not required unless the file uses thread-related calls or data.

POE catches asynchronous signals that normally terminate a program (**SIGINT**, **SIGQUIT**, **SIGTERM**, **SIGDANGER**) and terminates the entire parallel job. In addition, POE catches synchronous signals (**SIGSEGV**, **SIGBUS**) that are specific to a thread, terminates the entire parallel job, and causes a core file to be written at the point of interrupt on that thread. Since it is possible for several threads to generate synchronous signals simultaneously, the POE signal handler blocks subsequent occurrences.

The stacksize for the main thread is set as a characteristic of the user (in **/etc/security/limits** which is maintained by SMIT). The user can set the stacksize of any additional threads created; the default is 96K. Thread stacks are allocated in the process heap.

The main program should **exit**, not just **pthread_exit**. **pthread_exit** may not terminate all threads. **exit** will result in all threads being terminated.

The AIX XLF FORTRAN compiler does not produce thread-safe code. However, a program having one FORTRAN thread (and possibly additional C threads) can use the threaded LAPI library via the **mpxlf_r** command. This command links in the thread-tolerant FORTRAN library (**libxlf90_t.a**). This requires XLF FORTRAN Version 4.

In AIX Version 4.2.1, only the **libc_r.a** and **libsaa.a** are thread-safe. Applications using other libraries must ensure that they are used in a thread-safe way (usually by serializing entry to the function). This applies to all threaded programs under AIX, and is not specific to the threaded LAPI library.

If the user's application is also using the MPI protocol, the threaded MPI library (**libmpi_r.a**) must be used. Use of the MPI signal library (**libmpi.a**) along with the LAPI library is not supported.

The AIX compilers support a flag **-qarch** that allows the user to target code generation to a particular processor architecture. While this option can provide performance enhancements on specific platforms, it inhibits portability, particularly between the Power and PowerPC machines. The LAPI library is not targeted to a particular architecture, and is the same on PowerPC and Power nodes.

Once the LAPI library is initialized, if the process forks, only the forking thread exists in the child process. The child process doing LAPI communications is not supported.

IBM Parallel Environment for AIX: Hitchhiker's Guide provides additional information which has applicability to running LAPI parallel jobs using POE.

LAPI Programming Examples

Active Message (C)

This C program is an example of the use of the LAPI active message call:

```
/*Am.c
**Example Program illustrating use of the LAPI Active Message Call
*/

#include <lapi.h>

#define A_MAX     2
#define I_MAX     10

typedef struct {                    /* header for active message handler */
     compl_hndlr_t *cmpl_hndlr;     /* pointer to completion handler */
     int uinfo;                     /* uinfo passed to -- */
} usrhdr_t;                         /* the completion handler */

volatile int cont=0;

/*
/** Function:     The completion handler for the active call.
**                This is invoked at the target after all the data
**                of the active message send (LAPI_Amsend) call have
**                reached the target.
** Parameters:    hndl  -> pointer to the handle for the LAPI context
**                param -> pointer to the user param (specified
**                         by user in the header handler function)
*/

void
do_get(lapi_handle_t *hndl, void *param)                              do_get
{
     int loop, rc;
     int *buf;

     buf = (int *) param;
     printf("In Completion Handler: Result of AM call\n");

     /* Print Updated buffer */
     for (loop=0; loop < I_MAX; loop++) {
         printf("val[%d] = %d\n", loop, buf[loop]);
     }

     cont= 1;
}

/*
** Function:    User's active messsage header handler. This is invoked
**              at the target when the active message first arrives
**              at the target
** Parameters:  hndl    ->     pointer to the handle for the LAPI
**                             context
**              uhdr    ->     pointer to the user header
**              uhdrlen ->     pointer to the length of the user
```

```
**                              header
**              msglen  ->      pointer to the length of the message
**              compl_hndlr ->  pointer to the completion handler
**                              function pointer. This is to be set
**                              by the user in this function.  (CAN
**                              be NULL)
**              saved_info ->   pointer to the user_info.  This is set
**                              by the user in this function. This
**                              parameter is then passed to the
**                              completion handler when the completion
**                              handler is invoked.
*/

void *
hdr_hndlr(lapi_handle_t *hndl, void *uhdr, uint *uhdrlen, uint *msglen,
          compl_hndlr_t **cmpl_hndlr, void **saved_info)
{
    void          *buf;
    usrhdr_t      *vhdr;

    printf("In Header Handler\n");
    vhdr            = (usrhdr_t *) uhdr;
    *cmpl_hndlr     = (compl_hndlr_t *) vhdr->cmpl_hndlr;
    *saved_info     = (void *) vhdr->uinfo;
    buf             = (void *) vhdr->uinfo;

    return (buf);
}


int
main(int argc, char **argv)                                          main
{
    lapi_handle_t t_hndl;                /* LAPI context handle - */
                                         /* returned */
    lapi_info_t   t_info;                /* LAPI info structure */
    int           task_id,               /* My task id */
                  num_tasks;             /* Number of tasks in my */
                                         /* job */
    lapi_cntr_t   l_cntr;                /* Origin counter */
    lapi_cntr_t   t_cntr;                /* Target counter */
    lapi_cntr_t   c_cntr;                /* Completion counter */
    int           t_buf[I_MAX];          /* Buffer to manipulate */
    void          *global_addr[A_MAX];   /* Array to store */
                                         /* t_buf addr from all the */
                                         /* tasks.  The size of this */
                                         /* array needs to each */
                                         /* number of tasks */
    void          *tgt_addr[A_MAX];      /* Array to store target */
                                         /* counter addr from all */
                                         /* the tasks. */
    void          *hndlr_addr[A_MAX];    /* Array to store header */
                                         /* handlers */
    void          *cmpl_hndlr_addr[A_MAX]; /* Address of */
                                         /* completion handler */
    usrhdr_t      t_uhdr;                /* Store Header Handler */
                                         /* information */
    void          *uhdr, *udata;
```

```
int           uhdrlen, udatalen;
int           loop, rc, tgt, val, cur_val;
char          err_msg_buf[LAPI_MAX_ERR_STRING];

t_info.err_hndlr = NULL;   /* Not registering error handler */
                           /* function */

                                                                    ...main
if ((rc = LAPI_Init(&t_hndl, &t_info)) != LAPI_SUCCESS) {
    LAPI_Msg_string(rc, err_msg_buf);
    printf("Error Message: %s, rc = %d\n", err_msg_buf, rc);
    exit (rc);
}

rc = LAPI_Qenv(t_hndl, TASK_ID, &task_id);     /* Get task id */
                                               /* within job */
rc = LAPI_Qenv(t_hndl, NUM_TASKS, &num_tasks); /* Get no. of tasks */
                                               /* in job */

if (num_tasks != 2) {
    printf("Error Message: Program should run on 2 nodes\n");
    exit(1);
}

/* Turn off parameter checking - default is on */
rc = LAPI_Senv(t_hndl, ERROR_CHK, 0);

/* Initialize counters to be zero at the start */
rc = LAPI_Setcntr(t_hndl, &l_cntr, 0);
rc = LAPI_Setcntr(t_hndl, &t_cntr, 0);
rc = LAPI_Setcntr(t_hndl, &c_cntr, 0);

/*
** Exchange buffer address, tgt_cntr address and hdr_hndlr address
** and completion handler address of every task. Collective calls
*/
rc = LAPI_Address_init(t_hndl, t_buf, global_addr);
rc = LAPI_Address_init(t_hndl, &t_cntr, tgt_addr);
rc = LAPI_Address_init(t_hndl, (void *)&hdr_hndlr, hndlr_addr);
rc = LAPI_Address_init(t_hndl, (void *)&do_get, cmpl_hndlr_addr);

if (task_id == 0) { /* Task id is 0 , Origin */
    tgt = task_id + 1;
    for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
        t_buf[loop] = task_id - loop;
    }
    rc = LAPI_Gfence(t_hndl);  /* Global fence to sync before */
                               /* starting */

    /* Fill in uhdr and udata buffers for AM call */
    t_uhdr.cmpl_hndlr  = (compl_hndlr_t *) cmpl_hndlr_addr[1];
    t_uhdr.uinfo       = (int)global_addr[tgt];
    uhdr               = (void *)&t_uhdr;
    uhdrlen            = sizeof(usrhdr_t);
    udata              = (void *) t_buf;
    udatalen           = I_MAX*sizeof(int);

    rc = LAPI_Amsend(t_hndl, tgt, hndlr_addr[tgt], uhdr, uhdrlen,
```

```
                    (void *) udata, udatalen, tgt_addr[tgt], &l_cntr, &c_cntr);

                /* Wait for local AM completion */
                rc = LAPI_Waitcntr(t_hndl, &l_cntr, 1, &cur_val);

                /* Can now change local buffer */
                for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
                    t_buf[loop] = loop * task_id;
                                                                        ...main
                }

                /* Wait for target AM completion */
                rc = LAPI_Waitcntr(t_hndl, &c_cntr, 1, &cur_val);
                printf("Node %d, done issuing AM to node %d\n", task_id, tgt);

                rc = LAPI_Gfence(t_hndl);
                rc = LAPI_Get(t_hndl,tgt,I_MAX*sizeof(int), global_addr[tgt],
                                        (void *)t_buf,tgt_addr[tgt],&l_cntr);
                /* Wait for local Get completion */
                rc = LAPI_Waitcntr(t_hndl, &l_cntr, 1, NULL);

                printf("Node %d, done issuing Get from node %d\n", task_id, tgt);
                printf("Result of Get after the Am from node %d:\n", tgt);
                for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
                        printf("Val[%d] = %d\n", loop, t_buf[loop]);
                }
            } else { /* Task id is 1 , Target */
                tgt = task_id - 1;
                for (loop=0; loop < I_MAX; loop++) { /* Zero out buffer */
                    t_buf[loop] = 0;
                }
                rc = LAPI_Gfence(t_hndl);  /* Global fence to sync before */
                                           /* starting */
                /* Process AM */
                rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
                while (val < 1) {
                    sleep(1); /* Do some work */
                    rc = LAPI_Probe(t_hndl); /* Poll the adapter once */
                    rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
                }
                /* To reset the t_cntr value */
                rc = LAPI_Waitcntr(t_hndl, &t_cntr, 1, &cur_val);
                printf("Node %d, done doing work and processing AM\n", task_id);
                while (!cont) {
                    sleep(1); /* Do some work */
                }
                rc = LAPI_Gfence(t_hndl);
                /* Process Get */
                rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
                while (val < 1) {
                    sleep(1); /* Do some work */
                    rc = LAPI_Probe(t_hndl); /* Poll the adapter once */
                    rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
                }
                /* To reset the t_cntr value */
                rc = LAPI_Waitcntr(t_hndl, &t_cntr, 1, &cur_val);
                printf("Node %d, done doing work and processing Get\n", task_id);
            }
```

```
    rc = LAPI_Gfence(t_hndl); /* Global fence to sync before */
                              /* terminating job */
    rc = LAPI_Term(t_hndl);
 }
```

Active Message (Fortran)

This Fortran program is an example of the use of the LAPI active message call:

```
      /* Amf.f */
      INCLUDE 'lapif.h'

      INTEGER VOLATILE CONT
      COMMON /DATA/ CONT
      INTEGER T_HNDL, T_INFO(10)
      INTEGER TASKID, NUMTASKS
      INTEGER T_BUF(10)
      INTEGER L_CNTR, T_CNTR, C_CNTR
      INTEGER GLOBAL_ADDR(2)
      INTEGER TGT_ADDR(2), HNDLR_ADDR(2)
      INTEGER LOOP, IERROR, TGT, TGT2, VAL, CUR_VAL, LENGTH
      INTEGER T_ADDR
      INTEGER T_UHDR(2)
      INTEGER UHDR, UDATA
      INTEGER UHDRLEN, UDATALEN
      EXTERNAL DO_GET, HDR_HNDLR

      INTEGER ERR_MSG_BUF(40)

c     Not registering error handler function
      CALL LAPI_ADDRESS(MY_ERR_HNDLR, T_ADDR, IERROR)
      T_INFO(7) = T_ADDR                                          T_INFO

      CALL LAPI_INIT(T_HNDL, T_INFO, IERROR)
      IF (IERROR .NE. LAPI_SUCCESS) THEN                          IF
        VAL = IERROR
        CALL LAPI_MSG_STRING(VAL, ERR_MSG_BUF, IERROR)
        WRITE(6,*)'Error Message ',IERROR
        STOP 1
      ENDIF

c     GET task number and number of tasks in job
      CALL LAPI_QENV(T_HNDL, TASK_ID, TASKID, IERROR)
      CALL LAPI_QENV(T_HNDL, NUM_TASKS, NUMTASKS, IERROR)

      CALL LAPI_QENV(T_HNDL, TIMEOUT, TIME_OUT, IERROR)
      CALL LAPI_QENV(T_HNDL, INTERRUPT_SET, INTR_SET, IERROR)

      IF (TIME_OUT .gt. 30) THEN                                  IF
        VAL = 15
        CALL LAPI_SENV(T_HNDL, TIMEOUT, VAL, IERROR)
      ENDIF
      IF (INTR_SET .eq. 1) THEN                                   IF
c       Turn off interrupts
        VAL = 0
        CALL LAPI_SENV(T_HNDL, INTERRUPT_SET, VAL, IERROR)
      ENDIF
```

```
c     Turn off parameter checking - default is on
      VAL=0
      CALL LAPI_SENV(T_HNDL, ERROR_CHK, VAL, IERROR)

c     Initialize counters to be zero at the start
      CALL LAPI_SETCNTR(T_HNDL, L_CNTR, VAL, IERROR)
      CALL LAPI_SETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
      CALL LAPI_SETCNTR(T_HNDL, C_CNTR, VAL, IERROR)

      WRITE(6,*) "Node ",TASKID," Running AM fortran test."          WRITE

      IF (NUMTASKS .eq. 2) THEN                                      IF
c       Run only if number of tasks equal 2
c       Exchange buffer address to every task  -  Collective call
        CALL LAPI_ADDRESS(T_BUF, T_ADDR, IERROR)
        CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,GLOBAL_ADDR,IERROR)
        CALL LAPI_ADDRESS(T_CNTR, T_ADDR, IERROR)
        CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,TGT_ADDR,IERROR)
        CALL LAPI_ADDRESS(HDR_HNDLR, T_ADDR, IERROR)
        CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,HNDLR_ADDR,IERROR)

c       Task id is 0 , Origin
        IF (TASKID .eq. 0) THEN                                      IF
          TGT = TASKID + 1

c         Buffer in Fortran start at 1 and not 0
          TGT2 = TGT + 1
          LENGTH = 10*4

          DO LOOP = 1, 10
c           Update buffer
            T_BUF(LOOP) = TASKID - LOOP;                             T_BUF
          ENDDO

c         Global FENCE to sync before starting
          CALL LAPI_GFENCE(T_HNDL, IERROR)

c         Fill in uhdr and udata buffers for AM call
          CALL LAPI_ADDRESS(DO_GET, T_UHDR(1), IERROR)
          T_UHDR(2) = GLOBAL_ADDR(TGT2)                              T_UHDR
          UHDRLEN = 2 * 4

c         Issue AM from origin to target
          CALL LAPI_AMSEND(T_HNDL, TGT, HNDLR_ADDR(TGT2),
     1                     T_UHDR, UHDRLEN, T_BUF, LENGTH,
     2                     TGT_ADDR(TGT2), L_CNTR, C_CNTR, IERROR)

c         CALL LAPI_PUT(T_HNDL,TGT,LENGTH,GLOBAL_ADDR(TGT2),
c    1                  T_BUF, TGT_ADDR(TGT2),
c    2                  L_CNTR, C_CNTR, IERROR)

          VAL = 1
          CALL LAPI_WAITCNTR(T_HNDL, L_CNTR, VAL,
     1                       LAPI_ADDR_NULL, IERROR)

c         Local buffer can be reused now
          DO LOOP = 1, 10
            T_BUF(LOOP) = TGT2;                                      T_BUF
```

```
|                           ENDDO

|                           VAL = 1
|                           CALL LAPI_WAITCNTR(T_HNDL, C_CNTR, VAL,
|                     1                        LAPI_ADDR_NULL, IERROR)

|                           CALL LAPI_GFENCE(T_HNDL, IERROR)
|                     c     Issue GET from origin to target
|                           CALL LAPI_GET(T_HNDL,TGT,LENGTH,GLOBAL_ADDR(TGT2),
|                     1                T_BUF, TGT_ADDR(TGT2),
|                     2                L_CNTR, IERROR)                          ...T_BUF
|                           VAL = 1
|                           CALL LAPI_WAITCNTR(T_HNDL, L_CNTR, VAL,
|                     1                        LAPI_ADDR_NULL, IERROR)

|                           WRITE(6,*) "Node ",TASKID,                             WRITE
|                     1               "done issuing GET from node ", TGT
|                           WRITE(6,*) "Result of GET from node ", TGT             WRITE

|                           DO LOOP = 1, 10
|                             WRITE(6,*) "T_BUF(",LOOP,") = ", T_BUF(LOOP)         WRITE
|                           ENDDO
|                     c     Task id is 1 , Target
|                           ELSEIF (TASKID .eq. 1) THEN                            ELSEIF
|                             TGT = TASKID - 1
|                             DO LOOP = 1, 10
|                     c         Zero out buffer
|                             T_BUF(LOOP) = 0                                      T_BUF
|                           ENDDO

|                           VAL = 0
|                     c     Global FENCE to sync before starting
|                           CALL LAPI_GFENCE(T_HNDL, IERROR)
|                           CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
|                           DO WHILE (VAL .LT. 1)
|                     c       Can Do some work
|                             CALL LAPI_PROBE(T_HNDL, IERROR)
|                             CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
|                           ENDDO
|                           WRITE(6,*) "Result of AM from ",TGT, ":"             WRITE
|                           DO LOOP = 1, 10
|                             WRITE(6,*) "T_BUF(",LOOP,") = ", T_BUF(LOOP)       WRITE
|                           ENDDO

|                           DO WHILE (CONT .NE. 1)
|                     c       Can Do some work
|                             CALL LAPI_PROBE(T_HNDL, IERROR)
|                           ENDDO

|                           CALL LAPI_GFENCE(T_HNDL, IERROR)
|                     c     To clear the T_CNTR VALue
|                           VAL = 1
|                           CALL LAPI_WAITCNTR(T_HNDL, T_CNTR, VAL, CUR_VAL, IERROR)
|                           WRITE(6,*) "Node ", TASKID,                            WRITE
|                     1               "done doing work and processing AM"
|                           VAL = 0
|                           CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
```

```
|                            DO WHILE (VAL .LT. 1)
|                     c        Can Do some work
|                              CALL LAPI_PROBE(T_HNDL, IERROR)
|                              CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
|                            ENDDO

|                     c      To clear the T_CNTR VALue
|                            VAL = 1
|                            CALL LAPI_WAITCNTR(T_HNDL, T_CNTR, VAL, CUR_VAL, IERROR)
|                            WRITE(6,*) "Node ", TASKID,                              WRITE
|                          1            "done doing work and processing GET"
|                         ENDIF

|                          ENDIF                                                   ...WRITE
|                     c    Global FENCE to sync before terminating job
|                          CALL LAPI_GFENCE(T_HNDL, IERROR)

|                          CALL LAPI_TERM(T_HNDL, IERROR)
|                          END



|                          SUBROUTINE MY_ERR_HNDLR (HNDL, ERROR_CODE, ERR_TYPE,
|                         1                         TASKID, SRC)

|                          INCLUDE "lapif.h"
|                          INTEGER HNDL, ERROR_CODE, ERR_TYPE, TASKID, SRC
|                          INTEGER BUF(40)

|                          WRITE(6,*) "In my error handler, HNDL=",HNDL,            WRITE
|                         1           " ERROR_CODE=",ERROR_CODE," ERR_TYPE=",ERR_TYPE,
|                         2           " TASKID=",TASKID," SRC=",SRC
|                          CALL LAPI_MSG_STRING(ERROR_CODE, BUF, IERROR)
|                          WRITE(6,*) "In my error handler, error code = ", ERROR_CODE  WRITE
|                          IF (ERROR_CODE .ne. LAPI_ERR_TIMEOUT) THEN
|                     c      Cause program to exit
|                            STOP 2
|                          ENDIF

|                          RETURN
|                          END


|                          INTEGER FUNCTION HDR_HNDLR(HNDL, UHDR, UHDRLEN, MSGLEN,
|                         1                           COMPL_HNDLR, SAVED_INFO)
|                          INCLUDE "lapif.h"
|                          INTEGER HNDL
|                          INTEGER UHDR(*)
|                          INTEGER UHDRLEN
|                          INTEGER MSGLEN, COMPL_HNDLR, SAVED_INFO
|                          INTEGER T_ADDR

|                          WRITE(6,*) "In Header Handler"
|                          WRITE(6,*) "In Header Handler: UHDRLEN = ", UHDRLEN
|                          COMPL_HNDLR = UHDR(1)
|                          SAVED_INFO  = UHDR(2)
|                          HDR_HNDLR   = UHDR(2)
```

```
      RETURN
      END



      SUBROUTINE DO_GET (HNDL, PARAM)

      INCLUDE "lapif.h"
      INTEGER HNDL, PARAM(10)
      INTEGER LOOP
      INTEGER VOLATILE CONT
      COMMON /DATA/ CONT

                                                                    ...WRITE
      WRITE(6,*) "In Completion Handler: Result of AM call"
c     Print Updated buffer
      DO LOOP = 1, 10
          WRITE(6,*) "val[",LOOP,"] = ",PARAM(LOOP)
      ENDDO
      CONT = 1

      RETURN
      END
```

Get (C)

This C program is an example of the use of **LAPI_Get**:

```
/* Get.c
** Example Program illustrating the use of LAPI_Get
*/

#include <lapi.h>

#define A_MAX     2
#define I_MAX     10

int
main(int argc, char **argv)                                        main
{
   lapi_handle_t t_hndl;              /* LAPI context handle - */
                                      /* returned */
   lapi_info_t   t_info;              /* LAPI info structure */
   int           task_id,             /* My task id */
                 num_tasks;           /* Number of tasks in my job */
   int           t_buf[I_MAX];        /* Buffer to manipulate */
   lapi_cntr_t   l_cntr;              /* Origin counter */
   lapi_cntr_t   t_cntr;              /* Target counter */
   void          *global_addr[A_MAX]; /* Array to store t_buf addr */
                                      /* from all the tasks.  The */
                                      /* size of this array needs */
                                      /* to each number of tasks */
   void          *tgt_addr[A_MAX];    /* Array to store target */
                                      /* counter addr from all the */
                                      /* tasks. */
   int           loop, rc, tgt, val, cur_val;
   char          err_msg_buf[LAPI_MAX_ERR_STRING];
```

```
t_info.err_hndlr = NULL;   /* Not registering error handler /*
                           /* function */
if ((rc = LAPI_Init(&t_hndl, &t_info)) != LAPI_SUCCESS) }
    LAPI_Msg_string(rc, err_msg_buf);
    printf("Error Message: %s, rc = %d\n", err_msg_buf, rc);
    exit (rc);
}
rc = LAPI_Qenv(t_hndl, TASK_ID, &task_id);     /* Get task id */
                                               /* within job */
rc = LAPI_Qenv(t_hndl, NUM_TASKS, &num_tasks); /* Get no. of tasks */
                                               /* in job */

if (num_tasks != 2) }
    printf("Error Message: run with MP_PROCS set to 2\n");
    exit(1);
}

/* Turn off parameter checking - default is on */
rc = LAPI_Senv(t_hndl, ERROR_CHK, 0);

/* Initialize counters to be zero at the start */
rc = LAPI_Setcntr(t_hndl, &l_cntr, 0);
rc = LAPI_Setcntr(t_hndl, &t_cntr, 0);

/* Exchange buffer address and target_counter address of every task */
rc = LAPI_Address_init(t_hndl,t_buf,global_addr);  /* Collective */
                                                   /* call */
                                                                ...main
rc = LAPI_Address_init(t_hndl,&t_cntr,tgt_addr);   /* Collective */
                                                   /* call */

if (task_id == 0) { /* Task id is 0 , Origin */
    tgt = task_id + 1;
    for (loop=0; loop < I_MAX; loop++) { /* Zero out buffer */
        t_buf[loop] = 0;
    }
    rc = LAPI_Gfence(t_hndl);  /* Global fence to sync before */
                               /* starting */
    rc = LAPI_Get(t_hndl, tgt, I_MAX*sizeof(int), global_addr[tgt],
                          (void *)t_buf, tgt_addr[tgt], &l_cntr);
    rc = LAPI_Waitcntr(t_hndl, &l_cntr, 1, NULL);

    printf("Node %d, done issuing Get from node %d\n", task_id, tgt) ;
    printf("Result of Get from node %d:\n", tgt);
    for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
        printf("Val[%d] = %d\n", loop, t_buf[loop]);
    }
} else { /* Task id is 1 , Target */
    tgt = task_id - 1;
    for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
        t_buf[loop] = loop + task_id;
    }
    rc = LAPI_Gfence(t_hndl);  /* Global fence to sync before */
                               /* starting */
    rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
    while (val != 1) {
        sleep(1); /* Do some work */
        rc = LAPI_Probe(t_hndl); /* Poll the adapter once */
```

```
            rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
          }
          /* To clear the t_cntr value */
          rc = LAPI_Waitcntr(t_hndl, &t_cntr, val, &cur_val);
          printf("Node %d, done doing work and processing Get\n", task_id) ;
        }
        rc = LAPI_Gfence(t_hndl); /* Global fence to sync before */
                                  /* terminating job */
        rc = LAPI_Term(t_hndl);
}
```

Get (Fortran)

This Fortran program is an example of the use of **LAPI_GET**:

```
      /* Getf.f */
      INCLUDE 'lapif.h'

      INTEGER T_HNDL, T_INFO(10)
      INTEGER TASKID, NUMTASKS
      INTEGER T_BUF(10)
      INTEGER L_CNTR, T_CNTR
      INTEGER GLOBAL_ADDR(2)
      INTEGER TGT_ADDR(2)
      INTEGER LOOP, IERROR, TGT, TGT2, VAL, CUR_VAL, LENGTH
      INTEGER T_ADDR1, T_ADDR2, T_ADDR3, T_ADDR4
      INTEGER ERR_MSG_BUF(40)

c     Not registering error handler function
      CALL LAPI_ADDRESS(LAPI_ADDR_NULL, T_ADDR, IERROR)
      T_INFO(7) = T_ADDR                                          T_INFO

      CALL LAPI_INIT(T_HNDL, T_INFO, IERROR)
      IF (IERROR .NE. LAPI_SUCCESS) THEN                          IF
        VAL = IERROR
        CALL LAPI_MSG_STRING(VAL, ERR_MSG_BUF, IERROR)
        WRITE(6,*)'Error Message ',IERROR
        STOP 1
      ENDIF

c     GET task number and number of tasks in job
      CALL LAPI_QENV(T_HNDL, TASK_ID, TASKID, IERROR)
      CALL LAPI_QENV(T_HNDL, NUM_TASKS, NUMTASKS, IERROR)

c     Turn off parameter checking - default is on
      VAL=0
      CALL LAPI_SENV(T_HNDL, ERROR_CHK, VAL, IERROR)

c     Initialize counters to be zero at the start
      CALL LAPI_SETCNTR(T_HNDL, L_CNTR, VAL, IERROR)
      CALL LAPI_SETCNTR(T_HNDL, T_CNTR, VAL, IERROR)

      IF (NUMTASKS .eq. 2) THEN                                   IF
c       Run only if number of tasks equal 2
c       Exchange buffer address to every task  -  Collective call
        CALL LAPI_ADDRESS(T_BUF, T_ADDR, IERROR)
        CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,GLOBAL_ADDR,IERROR)
        CALL LAPI_ADDRESS(T_CNTR, T_ADDR, IERROR)
```

```
|                       CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,TGT_ADDR,IERROR)

|             c       Task id is 0 , Origin                                          IF
|                     IF (TASKID .eq. 0) THEN
|                       TGT = TASKID + 1

|             c         Buffer in Fortran start at 1 and not 0
|                       TGT2 = TGT + 1
|                       LENGTH = 10*4

|                       DO LOOP = 1, 10
|             c           Zero out buffer
|                         T_BUF(LOOP) = 0;                                          T_BUF
|                       ENDDO

|                                                                                   ...T_BUF
|             c         Global FENCE to sync before starting
|                       CALL LAPI_GFENCE(T_HNDL, IERROR)

|             c         Issue GET from origin to target
|                       CALL LAPI_GET(T_HNDL,TGT,LENGTH,GLOBAL_ADDR(TGT2),
|                  1                  T_BUF, TGT_ADDR(TGT2),
|                  2                  L_CNTR, IERROR)
|                       VAL = 1
|                       CALL LAPI_WAITCNTR(T_HNDL, L_CNTR, VAL,
|                  1                       LAPI_ADDR_NULL, IERROR)

|                       WRITE(6,*) "Node ",TASKID,                                  WRITE
|                  1              "done issuing GET from node ", TGT
|                       WRITE(6,*) "Result of GET from node ", TGT                  WRITE

|                       DO LOOP = 1, 10
|                         WRITE(6,*) "T_BUF(",LOOP,") = ", T_BUF(LOOP)              WRITE
|                       ENDDO
|             c       Task id is 1 , Target
|                     ELSEIF (TASKID .eq. 1) THEN                                   ELSEIF
|                       TGT = TASKID - 1
|                       DO LOOP = 1, 10
|             c           Update buffer
|                         T_BUF(LOOP) = LOOP + TASKID                               T_BUF
|                         WRITE(6,*) "T_BUF(",LOOP,") = ", T_BUF(LOOP)              WRITE
|                       ENDDO

|                       VAL = 0
|             c         Global FENCE to sync before starting
|                       CALL LAPI_GFENCE(T_HNDL, IERROR)
|                       CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
|                       DO WHILE (VAL .LT. 1)
|             c           Can Do some work
|                         CALL LAPI_PROBE(T_HNDL, IERROR)
|                         CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
|                       ENDDO

|             c         To clear the T_CNTR VALue
|                       VAL = 1
|                       CALL LAPI_WAITCNTR(T_HNDL, T_CNTR, VAL, CUR_VAL, IERROR)
|                       WRITE(6,*) "Node ", TASKID,                                 WRITE
|                  1              "done doing work and processing Put"
```

```
        ENDIF

      ENDIF
c     Global FENCE to sync before terminating job
      CALL LAPI_GFENCE(T_HNDL, IERROR)

      CALL LAPI_TERM(T_HNDL, IERROR)
      END
```

Put (C)

This C program is an example of the use of **LAPI_Put** and **LAPI_Get**:

```
/* Put.c
** Example program illustrating the use of LAPI_Put and LAPI_Get
*/

#include <lapi.h>

#define A_MAX     2
#define I_MAX     10

int
main(int argc, char **argv)                                          main
{
    lapi_handle_t t_hndl;              /* LAPI context handle - */
                                       /* returned */
    lapi_info_t   t_info;              /* LAPI info structure */
    int           task_id,             /* My task id */
                  num_tasks;           /* Number of tasks in my job */
    int           t_buf[I_MAX];        /* Buffer to manipulate */
    lapi_cntr_t   l_cntr;              /* Origin counter */
    lapi_cntr_t   t_cntr;              /* Target counter */
    lapi_cntr_t   c_cntr;              /* Completion counter */
    void          *global_addr[A_MAX]; /* Array to store t_buf */
                                       /* addr from all the tasks */
                                       /* The size of this array */
                                       /* needs to each number of tasks */
    void          *tgt_addr[A_MAX];    /* Array to store target */
                                       /* counter ad from all the */
                                       /* tasks. */
    int           loop, rc, tgt, val, cur_val;
    char          err_msg_buf[LAPI_MAX_ERR_STRING];

    t_info.err_hndlr = NULL;   /* Not registering error handler function */
    if ((rc = LAPI_Init(&t_hndl, &t_info)) != LAPI_SUCCESS) {
        LAPI_Msg_string(rc, err_msg_buf);
        printf("Error Message: %s, rc = %d\n", err_msg_buf, rc);
        exit (rc);
    }
    rc = LAPI_Qenv(t_hndl, TASK_ID, &task_id);     /* Get task number */
                                                   /* within job */
    rc = LAPI_Qenv(t_hndl, NUM_TASKS, &num_tasks); /* Get number of */
                                                   /* tasks in job */

    if (num_tasks != 2) }
        printf("Error Message: Run with MP_PROCS set to 2\n");
        exit(1);
```

```
}

/* Turn off parameter checking - default is on */
rc = LAPI_Senv(t_hndl, ERROR_CHK, 0);

/* Initialize counters to be zero at the start */
rc = LAPI_Setcntr(t_hndl, &l_cntr, 0);
rc = LAPI_Setcntr(t_hndl, &t_cntr, 0);
rc = LAPI_Setcntr(t_hndl, &c_cntr, 0);

/* Exchange buffer address and target counter address of every task */
rc = LAPI_Address_init(t_hndl,t_buf,global_addr);    /* Collective call */
                                                                ...main
rc = LAPI_Address_init(t_hndl,&t_cntr,tgt_addr);     /* Collective call */

if (task_id == 0) { /* Task id is 0 , Origin */
    tgt = task_id + 1;
    for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
        t_buf[loop] = task_id - loop;
    }
    rc = LAPI_Gfence(t_hndl);  /* Global fence to sync before starting */
    rc = LAPI_Put(t_hndl,tgt,I_MAX*sizeof(int),global_addr[tgt],
                    (void *)t_buf,tgt_addr[tgt],,&l_cntr,&c_cntr);
    /* Wait for local Put completion */
    rc = LAPI_Waitcntr(t_hndl, &l_cntr, 1, NULL);

    /* Can now change local buffer */
    for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
        t_buf[loop] = loop * task_id;
    }

    /* Wait for target Put completion at task 1 as well */
    rc = LAPI_Waitcntr(t_hndl, &c_cntr, 1, NULL);
    printf("Node %d, done issuing Put to node %d\n", task_id, tgt);

    rc = LAPI_Get(t_hndl,tgt,I_MAX*sizeof(int),global_addr[tgt],
                                    (void *)t_buf,tgt_addr[tgt],&l_cntr);
    /* Wait for local Get completion */
    rc = LAPI_Waitcntr(t_hndl, &l_cntr, 1, NULL);
    printf("Node %d, done issuing Get from node %d\n", task_id, tgt);
    printf("Result of Get after the Put from node %d:\n", tgt);
    for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
        printf("Val[%d] = %d\n", loop, t_buf[loop]);
    }
} else { /* Task id is 1 , Target */
    tgt = task_id - 1;
    for (loop=0; loop < I_MAX; loop++) { /* Zero out buffer */
        t_buf[loop] = 0;
    }
    rc = LAPI_Gfence(t_hndl);  /* Global fence to sync before starting */
    /* Process Put */
    rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
    while (val < 1) {
        sleep(1); /* Do some work */
        rc = LAPI_Probe(t_hndl); /* Poll the adapter once */
        rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
    }
    /* To clear the t_cntr value */
```

```
            rc = LAPI_Waitcntr(t_hndl, &t_cntr, 1, &cur_val);
            printf("Node %d, done doing work and processing Put\n", task_id);
            printf("Result of Put from %d:\n", tgt);
            for (loop=0; loop < I_MAX; loop++) { /* Update buffer */
                printf("Val[%d] = %d\n", loop, t_buf[loop]);
            }
            /* Process Get */
            rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
            while (val < 1) {
                sleep(1); /* Do some work */
                rc = LAPI_Probe(t_hndl); /* Poll the adapter once */
                rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
            }
                                                                    ...main
            /* To clear the t_cntr value */
            rc = LAPI_Waitcntr(t_hndl, &t_cntr, 1, &cur_val);
            printf("Node %d, done doing work and processing Get\n", task_id);
        }
        rc = LAPI_Gfence(t_hndl); /* Global fence to sync before terminating job */
        rc = LAPI_Term(t_hndl);
    }
```

Put (Fortran)

This Fortran program is an example of the use of **LAPI_PUT** and **LAPI_GET**:

```
      /* Putf.f */
      INCLUDE 'lapif.h'

      INTEGER T_HNDL, T_INFO(10)
      INTEGER TASKID, NUMTASKS
      INTEGER T_BUF(10)
      INTEGER L_CNTR, T_CNTR, C_CNTR
      INTEGER GLOBAL_ADDR(2)
      INTEGER TGT_ADDR(2)
      INTEGER LOOP, IERROR, TGT, TGT2, VAL, CUR_VAL, LENGTH
      INTEGER T_ADDR
      INTEGER ERR_MSG_BUF(40)

c     Not registering error handler function
      CALL LAPI_ADDRESS(LAPI_ADDR_NULL, T_ADDR, IERROR)
      T_INFO(7) = T_ADDR                                          T_INFO

      CALL LAPI_INIT(T_HNDL, T_INFO, IERROR)
      IF (IERROR .NE. LAPI_SUCCESS) THEN                          IF
        VAL = IERROR
        CALL LAPI_MSG_STRING(VAL, ERR_MSG_BUF, IERROR)
        WRITE(6,*)'Error Message ',IERROR
        STOP 1
      ENDIF

c     GET task number and number of tasks in job
      CALL LAPI_QENV(T_HNDL, TASK_ID, TASKID, IERROR)
      CALL LAPI_QENV(T_HNDL, NUM_TASKS, NUMTASKS, IERROR)

c     Turn off parameter checking - default is on
      VAL=0
      CALL LAPI_SENV(T_HNDL, ERROR_CHK, VAL, IERROR)
```

```
c     Initialize counters to be zero at the start
      CALL LAPI_SETCNTR(T_HNDL, L_CNTR, VAL, IERROR)
      CALL LAPI_SETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
      CALL LAPI_SETCNTR(T_HNDL, C_CNTR, VAL, IERROR)

      IF (NUMTASKS .eq. 2) THEN                                    IF
c       Run only if number of tasks equal 2
c       Exchange buffer address to every task  -  Collective call
        CALL LAPI_ADDRESS(T_BUF, T_ADDR, IERROR)
        CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,GLOBAL_ADDR,IERROR)
        CALL LAPI_ADDRESS(T_CNTR, T_ADDR, IERROR)
        CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,TGT_ADDR,IERROR)

c       Task id is 0 , Origin
        IF (TASKID .eq. 0) THEN                                    IF
          TGT = TASKID + 1

c         Buffer in Fortran start at 1 and not 0
          TGT2 = TGT + 1
          LENGTH = 10*4

          DO LOOP = 1, 10
c           Update buffer
            T_BUF(LOOP) = TASKID - LOOP;                           T_BUF
          ENDDO

c         Global FENCE to sync before starting                     ...T_BUF
          CALL LAPI_GFENCE(T_HNDL, IERROR)

c         Issue PUT from origin to target
          CALL LAPI_PUT(T_HNDL,TGT,LENGTH,GLOBAL_ADDR(TGT2),
     1                  T_BUF, TGT_ADDR(TGT2),
     2                  L_CNTR, C_CNTR, IERROR)
          VAL = 1
          CALL LAPI_WAITCNTR(T_HNDL, L_CNTR, VAL,
     1                       LAPI_ADDR_NULL, IERROR)

c         Local buffer can be reused now
          DO LOOP = 1, 10
            T_BUF(LOOP) = TGT2;                                    T_BUF
          ENDDO


          VAL = 1
          CALL LAPI_WAITCNTR(T_HNDL, C_CNTR, VAL,
     1                       LAPI_ADDR_NULL, IERROR)

c         Issue GET from origin to target
          CALL LAPI_GET(T_HNDL,TGT,LENGTH,GLOBAL_ADDR(TGT2),
     1                  T_BUF, TGT_ADDR(TGT2),
     2                  L_CNTR, IERROR)
          VAL = 1
          CALL LAPI_WAITCNTR(T_HNDL, L_CNTR, VAL,
     1                       LAPI_ADDR_NULL, IERROR)

          WRITE(6,*) "Node ",TASKID,                               WRITE
     1               "done issuing GET from node ", TGT
```

```
                    WRITE(6,*) "Result of GET from node ", TGT                WRITE

                    DO LOOP = 1, 10
                      WRITE(6,*) "T_BUF(",LOOP,") = ", T_BUF(LOOP)            WRITE
                    ENDDO
               c   Task id is 1 , Target
                    ELSEIF (TASKID .eq. 1) THEN                               ELSEIF
                      TGT = TASKID - 1
                      DO LOOP = 1, 10
               c        Zero out buffer
                        T_BUF(LOOP) = 0                                       T_BUF
                      ENDDO

                      VAL = 0
               c      Global FENCE to sync before starting
                      CALL LAPI_GFENCE(T_HNDL, IERROR)
                      CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
                      DO WHILE (VAL .LT. 1)
               c        Can Do some work
                        CALL LAPI_PROBE(T_HNDL, IERROR)
                        CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
                      ENDDO
                      WRITE(6,*) "Result of Put from ",TGT, ":"               WRITE
                      DO LOOP = 1, 10
                        WRITE(6,*) "T_BUF(",LOOP,") = ", T_BUF(LOOP)          WRITE
                      ENDDO

               c      To clear the T_CNTR VALue                               ...WRITE
                      VAL = 1
                      CALL LAPI_WAITCNTR(T_HNDL, T_CNTR, VAL, CUR_VAL, IERROR)
                      WRITE(6,*) "Node ", TASKID,                             WRITE
                     1           "done doing work and processing PUT"
                      VAL = 0
                      CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
                      DO WHILE (VAL .LT. 1)
               c        Can Do some work
                        CALL LAPI_PROBE(T_HNDL, IERROR)
                        CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
                      ENDDO

               c      To clear the T_CNTR VALue
                      VAL = 1
                      CALL LAPI_WAITCNTR(T_HNDL, T_CNTR, VAL, CUR_VAL, IERROR)
                      WRITE(6,*) "Node ", TASKID,                             WRITE
                     1           "done doing work and processing GET"
                    ENDIF

                   ENDIF
               c   Global FENCE to sync before terminating job
                   CALL LAPI_GFENCE(T_HNDL, IERROR)

                   CALL LAPI_TERM(T_HNDL, IERROR)
                   END
```

Rmw (C)

This C program is an example of **LAPI_Rmw** (read/modify/write function):

```
/* Rmw.c */
** Example Program showing use of read modify write function (LAPI_Rmw).
*/

#include <lapi.h>
#include <signal.h>
#include <unistd.h>

#define A_MAX     2

/*
** User error handler function.
*/

void my_err_hndlr (lapi_handle_t *hndl, int *error_code,
                   lapi_err_t *err_type, int *task_id, int *src)
{
    char buf[LAPI_MAX_ERR_STRING];

    printf("In my error handler, hndl=%d, error_code=%d, err_type=%d, "
           "task_id=%d, src=%d\n", *hndl,*error_code,*err_type,*task_id,*src);
    LAPI_Msg_string(*error_code,buf);
    printf("In my error handler, error code = %d, error reason = %s\n",
           *error_code, buf);
    if (*error_code != LAPI_ERR_TIMEOUT)
        kill(getpid(),SIGTERM);  /* Cause program to exit */
}

int
main(int argc, char **argv)                                        main
{
    lapi_handle_t t_hndl;              /* LAPI context handle - returned */
    lapi_info_t   t_info;              /* LAPI info structure */
    int           task_id,             /* My task id */
                  num_tasks;           /* Number of tasks in my job */
    int           t_buf;               /* Buffer to manipulate */
    int           t2_buf;              /* Temporary Buffer */
    lapi_cntr_t   l_cntr;              /* Origin counter */
    lapi_cntr_t   t_cntr;              /* Target counter */
    void          *global_addr[A_MAX]; /* Array to store t_buf addr from */
                                       /* all the tasks.  The size of */
                                       /* this array needs to each */
                                       /* number of tasks */
    void          *tgt_addr[A_MAX];    /* Array to store target counter */
                                       /* addr from all the tasks. */
    int           loop, rc, tgt, val, cur_val, prev_tgt_val;
    char          err_msg_buf[LAPI_MAX_ERR_STRING];
    int           time_out;            /* Get current timeout value */
    int           intr_set;            /* Get current interrupt setting */

    t_info.err_hndlr = my_err_hndlr;   /* register an error handler */
                                       /* function */

    if ((rc = LAPI_Init(&t_hndl, &t_info)) != LAPI_SUCCESS) {
```

```
    LAPI_Msg_string(rc, err_msg_buf);
    printf("Error Message: %s, rc = %d\n", err_msg_buf, rc);
    exit (rc);
                                                                    ...main
}

rc = LAPI_Qenv(t_hndl, TASK_ID, &task_id);
rc = LAPI_Qenv(t_hndl, NUM_TASKS, &num_tasks);

if (num_tasks != 2) {
    printf("Error Message: Run with MP_PROCS set to 2\n");
    LAPI_Term (t_hndl);
    exit(1);
}

rc = LAPI_Qenv(t_hndl, TIMEOUT, &time_out);  /* Value in seconds */
rc = LAPI_Qenv(t_hndl, INTERRUPT_SET, &intr_set);

if (time_out > 30) {
    rc = LAPI_Senv(t_hndl, TIMEOUT, 15);  /* Should be > */
                                          /* MIN_TIMEOUT */
}
if (intr_set == 1) {
    rc = LAPI_Senv(t_hndl, INTERRUPT_SET, 0);  /* Turn off */
                                               /* interrupts */
}
/* Turn off parameter checking - default is on */
rc = LAPI_Senv(t_hndl, ERROR_CHK, 0);

/* Initialize counters to be zero at the start */
rc = LAPI_Setcntr(t_hndl, &l_cntr, 0);
rc = LAPI_Setcntr(t_hndl, &t_cntr, 0);

/* Exchange buffer address to every task */
rc = LAPI_Address_init(t_hndl, &t_buf, global_addr);   /* Collective */
                                                       /* call */
rc = LAPI_Address_init(t_hndl, &t_cntr, tgt_addr);     /* Collective */
                                                       /* call */

if (task_id == 0) { /* Task id is 0 , Origin */
    tgt = task_id + 1;
    t_buf = 1;  /* Initial value to add at target */

    rc = LAPI_Gfence(t_hndl);  /* Global fence to sync before */
                               /* starting */
    rc = LAPI_Rmw(t_hndl, FETCH_AND_ADD, tgt, global_addr[tgt],
                                  &t_buf, &prev_tgt_val,&l_cntr);
    /* Wait for local Rmw completion */
    rc = LAPI_Waitcntr(t_hndl, &l_cntr, 1, NULL);

    printf("Node %d, done issuing Rmw to node %d\n", task_id, tgt);

    rc = LAPI_Get(t_hndl,tgt,sizeof(int),global_addr[tgt],
                          (void *)&t2_buf,tgt_addr[tgt],&l_cntr);
    /* Wait for local Get completion */
    rc = LAPI_Waitcntr(t_hndl, &l_cntr, 1, NULL);
    printf("Node %d, done issuing Get from node %d\n", task_id, tgt);
    printf("Result of Get after the Rmw from node %d:\n", tgt);
```

```
            printf("Correct value should be %d = %d\n",
                                              t_buf + prev_tgt_val, t2_buf);
        } else { /* Task id is 1 , Target */
            tgt = task_id - 1;                                          ...main
            t_buf = 5; /* Set initial buffer value */
            rc = LAPI_Gfence(t_hndl);  /* Global fence to sync before */
                                       /* starting */
            /* Process Get */
            rc=LAPI_Getcntr(t_hndl, &t_cntr, &val);
            while (val < 1) {
                sleep(1); /* Do some work */
                rc = LAPI_Probe(t_hndl); /* Poll the adapter once */
                rc = LAPI_Getcntr(t_hndl, &t_cntr, &val);
            }
            /* To clear the t_cntr value */
            rc = LAPI_Waitcntr(t_hndl, &t_cntr, 1, &cur_val);
            printf("Node %d, done doing work and processing Get\n", task_id);
        }
        rc = LAPI_Gfence(t_hndl); /* Global fence to sync before */
                                  /* terminating job */
        rc = LAPI_Term(t_hndl);
    }
```

Rmw (Fortran)

This Fortran program is an example of **LAPI_RMW** (read/write/modify function):

```
      /* Rmwf.f */
      INCLUDE 'lapif.h'

      INTEGER T_HNDL, T_INFO(10)
      INTEGER TASKID, NUMTASKS
      INTEGER T_BUF, T2_BUF, PREV_TGT_VAL
      INTEGER L_CNTR, T_CNTR
      INTEGER GLOBAL_ADDR(2)
      INTEGER TGT_ADDR(2)
      INTEGER LOOP, IERROR, TGT, TGT2, VAL, CUR_VAL, LENGTH
      INTEGER T_ADDR, TIME_OUT, INTR_SET
      INTEGER ERR_MSG_BUF(40)

c     Not registering error handler function
      CALL LAPI_ADDRESS(MY_ERR_HNDLR, T_ADDR, IERROR)
      T_INFO(7) = T_ADDR                                            T_INFO

      CALL LAPI_INIT(T_HNDL, T_INFO, IERROR)
      IF (IERROR .NE. LAPI_SUCCESS) THEN                            IF
        VAL = IERROR
        CALL LAPI_MSG_STRING(VAL, ERR_MSG_BUF, IERROR)
        WRITE(6,*)'Error Message ',IERROR
        STOP 1
      ENDIF

c     GET task number and number of tasks in job
      CALL LAPI_QENV(T_HNDL, TASK_ID, TASKID, IERROR)
      CALL LAPI_QENV(T_HNDL, NUM_TASKS, NUMTASKS, IERROR)

      CALL LAPI_QENV(T_HNDL, TIMEOUT, TIME_OUT, IERROR)
```

```
|                       CALL LAPI_QENV(T_HNDL, INTERRUPT_SET, INTR_SET, IERROR)

|                       IF (TIME_OUT .gt. 30) THEN                                    IF
|                         VAL = 15
|                         CALL LAPI_SENV(T_HNDL, TIMEOUT, VAL, IERROR)
|                       ENDIF
|                       IF (INTR_SET .eq. 1) THEN                                     IF
|                 c       Turn off interrupts
|                         VAL = 0
|                         CALL LAPI_SENV(T_HNDL, INTERRUPT_SET, VAL, IERROR)
|                       ENDIF

|                 c     Turn off parameter checking - default is on
|                       VAL=0
|                       CALL LAPI_SENV(T_HNDL, ERROR_CHK, VAL, IERROR)

|                 c     Initialize counters to be zero at the start
|                       CALL LAPI_SETCNTR(T_HNDL, L_CNTR, VAL, IERROR)
|                       CALL LAPI_SETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
|                       CALL LAPI_SETCNTR(T_HNDL, C_CNTR, VAL, IERROR)

|                       IF (NUMTASKS .eq. 2) THEN                                     IF
|                 c       Run only if number of tasks equal 2
|                 c       Exchange buffer address to every task  -  Collective call
|                         CALL LAPI_ADDRESS(T_BUF, T_ADDR, IERROR)
|                         CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,GLOBAL_ADDR,IERROR)
|                         CALL LAPI_ADDRESS(T_CNTR, T_ADDR, IERROR)
|                         CALL LAPI_ADDRESS_INIT(T_HNDL,T_ADDR,TGT_ADDR,IERROR)
|                                                                                     ...IF

|                 c       Task id is 0 , Origin
|                         IF (TASKID .eq. 0) THEN                                     IF
|                           TGT = TASKID + 1

|                 c         Buffer in Fortran start at 1 and not 0
|                           TGT2 = TGT + 1
|                 c         Get INTEGER size buffer
|                           LENGTH = 4

|                 c         Initial value to add at target
|                           T_BUF = 1

|                 c         Global FENCE to sync before starting
|                           CALL LAPI_GFENCE(T_HNDL, IERROR)

|                 c         Issue RMW from origin to target
|                           CALL LAPI_RMW(T_HNDL, FETCH_AND_ADD, TGT,
|                      1                GLOBAL_ADDR(TGT2), T_BUF,
|                      2                PREV_TGT_VAL, L_CNTR, IERROR)
|                           VAL = 1
|                           CALL LAPI_WAITCNTR(T_HNDL, L_CNTR, VAL,
|                      1                     CUR_VAL, IERROR)
|                           WRITE(6,*) "Node ",TASKID,                              WRITE
|                      1             "done issuing RMW from node ", TGT

|                 c         Issue GET from origin to target
|                           CALL LAPI_GET(T_HNDL,TGT,LENGTH,GLOBAL_ADDR(TGT2),
|                      1                T2_BUF, TGT_ADDR(TGT2),
```

```
     2                 L_CNTR, IERROR)
          VAL = 1
          CALL LAPI_WAITCNTR(T_HNDL, L_CNTR, VAL,
     1                       LAPI_ADDR_NULL, IERROR)

          WRITE(6,*) "Node ",TASKID,                                    WRITE
     1               "done issuing GET from node ", TGT
          WRITE(6,*) "Result of GET from node ", TGT                    WRITE

          VAL = T_BUF + PREV_TGT_VAL
          WRITE(6,*) "Correct value should be ",VAL,                    WRITE
     1               " = ", T2_BUF

c       Task id is 1 , Target
        ELSEIF (TASKID .eq. 1) THEN                                     ELSEIF
          TGT = TASKID - 1
c         Set initial buffer value
          T_BUF = 5

          VAL = 0
c         Global FENCE to sync before starting
          CALL LAPI_GFENCE(T_HNDL, IERROR)
          CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
          DO WHILE (VAL .LT. 1)
c           Can Do some work
            CALL LAPI_PROBE(T_HNDL, IERROR)
            CALL LAPI_GETCNTR(T_HNDL, T_CNTR, VAL, IERROR)
          ENDDO

                                                                        ...ELSEIF

c         To clear the T_CNTR VALue
          VAL = 1
          CALL LAPI_WAITCNTR(T_HNDL, T_CNTR, VAL, CUR_VAL, IERROR)
          WRITE(6,*) "Node ", TASKID,                                   WRITE
     1               "done doing work and processing GET"
        ENDIF

       ENDIF
c      Global FENCE to sync before terminating job
       CALL LAPI_GFENCE(T_HNDL, IERROR)

       CALL LAPI_TERM(T_HNDL, IERROR)
       END


       SUBROUTINE MY_ERR_HNDLR (HNDL, ERROR_CODE, ERR_TYPE,
      1                         TASKID, SRC)

       INCLUDE "lapif.h"
       INTEGER BUF(40)

       WRITE(6,*) "In my error handler, HNDL=",HNDL,                    WRITE
      1           " ERROR_CODE=",ERROR_CODE," ERR_TYPE=",ERR_TYPE,
      2           " TASKID=",TASKID," SRC=",SRC
       CALL LAPI_MSG_STRING(ERROR_CODE, BUF, IERROR)
       WRITE(6,*) "In my error handler, error code = ", ERROR_CODE      WRITE
```

```
|                         IF (ERROR_CODE .ne. LAPI_ERR_TIMEOUT) THEN                    IF
|                    c      Cause program to exit
|                           STOP 2
|                         ENDIF

|                         RETURN
|                         END
```

Appendixes

© Copyright IBM Corp. 1995, 1997

Appendix A. Using an Alternate Boot System Image

The Alternate Boot System Image function provides the ability to configure two different **rootvg** volume groups for an SP node or the control workstation. The **rootvg** volume group contains logical volumes for:

- Operating system file systems
- JFS logs (journaled file system logs)
- Dump space
- Paging space

You may select one of the two levels of software to execute by changing the boot list in NVRAM (nonvolatile random access memory) and booting the node or control workstation with the disk that contains the desired software level. The node or control workstation will, in effect, have two **rootvg** volume groups. Only one is active at any time and no changes from one **rootvg** volume group are propagated to the other **rootvg** volume group. For example, /**tmp** file changes on one **rootvg** volume group are not changed on the dormant **rootvg** volume group. External or non-**rootvg** volume groups on additional disks are available via export and import of the volume group on both **rootvg** volume groups.

This function is useful when a quick backout of a new software level is required. Such a backout may be desired for testing or to recover from problems found soon after a new software level is put into production.

Enabling Alternate Boot System Image Function

To use the Alternate Boot System Image function, a node or the control workstation must have the following hardware:

1. Two hard disks (one for each **rootvg** volume group)
2. A virtual battery on the SP nodes

SP Virtual Battery

The SP virtual battery is a power source that keeps NVRAM powered and the time-of-day clock running while the SP frame is plugged into a power outlet with power applied or until a node is removed from the SP frame. Use of the virtual battery provides for maintaining the boot list while the node is powered off. Power is supplied by the node supervisor card in each node. This card provides line power even when the node is powered off.

On wide SP nodes, a virtual battery is installed from the factory. On thin SP nodes, a no charge Engineering Change Action (ECA) can be ordered by your IBM representative. Order ECA number ECA007. On the control workstation, a battery is installed and no virtual battery ECA is required.

© Copyright IBM Corp. 1995, 1997

Using the Alternate Boot System Image Function

To use the Alternate Boot System Image function on an SP node, take the following steps:

1. Perform the following steps to install each system image on a single disk:
   a. Enter **spbootins -r install -h hdisk0 -l** *node_list*
   b. Run **setup_server** on the boot/install server
   c. Select **Select** and **Do Net Boot** from the Global Controls display on the SP System Monitor.
   d. Redo steps a-c and change the **spbootins -h** option to select the other install disk.
2. Set the boot list to the disk containing the image you want to boot.
   a. On the SP node, enter:

      ```
      /bin/bootlist -m prevboot -i
      ```

      This invalidates the previous boot list.
   b. On the SP node, enter:

      ```
      /bin/bootlist -m normal ent0 hdisk0
      ```

      This sets the boot list to boot from the SP Ethernet first (so an install can be done if required.) In this example, if an installation is not being performed, the boot list will boot from **hdisk0**.
   c. On the SP node, enter:

      ```
      spbootins -h hdisk0 -l 5
      ```

      This updates the SDR with the correct boot disk information. In this example, for **node_number 5** in the Node class of the SDR, the **bootdisk** attribute is set to **hdisk0**.

   All of the boot list manipulations can be done on the control workstation, using **dsh** for a group of nodes. The SDR command needs to be done for each node in the group.
3. Export any non-**rootvg** volume groups (**exportvg** *volume_group_name*).
4. Reboot the node to bring up the system image that was installed on **hdisk0**.
5. Import any non-**rootvg** volume groups (**importvg -y** *volume_group_name* **hdisk***x*).

To switch back to the other system image, redo step 2 and change the **-h hdisk0** value to the physical disk name of the alternate system image.

Setting the Boot List Manually

When a frame drops power or a node is removed, you must manually set the boot list on every node in that frame. In the case of nodes being removed from the frame, just the boot list on that node needs to be manually changed. To manually change the boot list for a specific node, perform the following steps from the SP System Monitor GUI on the control workstation:

1. Select **SP** → **Display Layouts**

2. Select the number of the frame you want to see
3. Select the number of the node within the frame
4. Select **Select Node Front Panel Layout**
5. Press **Display Node Layout**
6. Set the key switch position to **Secure**
7. Open the console for that node
8. **Power on** the node
9. Wait for the LED display to get to 200
10. Set the key switch position to **Service**
11. Press the **Reset** button
12. The Diagnostic Menu appears in the console window
13. Hit **Enter**
14. Select **Service Aids**
15. Select **Display/Alter Bootlist**
16. Select **Normal Mode Bootlist**
17. Select **Alter Current Bootlist**
18. Assign **1** to **en0** and **2** to **hdisk***x*
19. Press F7 to commit the change
20. Reboot the node

Using Alternate System Image on the Control Workstation

The alternate system image can be used on a control workstation with at least two hard disks available for the **rootvg**. Each hard disk must be large enough to hold the entire operating system and the /**spdata** directory.

Because the control workstation has a battery, the SP virtual battery is not needed.

To use the alternate system image function, you must first install both levels of software (AIX and PSSP) on each disk (one image per disk). Configure each level of software as a control workstation. When using alternate system images, complete and separate copies of the /**spdata** information are also made. The control workstation cannot have multiple system images while the SP nodes have multiple system images. You must first migrate the control workstation to a new software level and commit that software level (that is, you are not able to change back to the old level) before you create additional system images for the SP nodes. All alternate system images for the SP nodes must be created while the control workstation is at the new software level.

Because changes are not propagated to dormant **rootvg**s, you should not make any node configuration changes during the period when alternate system images are in use on the control workstation.

Steps to Switch Between Alternate System Images

1. Install the control workstation software (AIX and PSSP) on a single hard disk for the first system image.

   Use the AIX and PSSP documentation to install to a single disk. As an example, use **hdisk0**.

2. Install the control workstation software (AIX and PSSP) on a single hard disk for the second system image.

   Use the AIX and PSSP documentation to install to a single disk. As an example, use **hdisk1**.

3. Set the boot list to **hdisk0**.

   a. Enter the **diag** command at the shell prompt.

   b. The Diagnostic Menu appears in the console window

   c. Hit **Enter**

   d. Select **Service Aids**

   e. Select **Display/Alter Bootlist**

   f. Select **Normal Mode Bootlist**

   g. Select **Alter Current Bootlist**

   h. Assign **1** to **hdisk0***x*

   i. Press F7 to commit the change

   j. Reboot the node

To switch to the new level assign **1** to **hdisk1** in the **Alter Current Bootlist** menu, commit the change and reboot the node.

Appendix B. SP Daemons

The following daemons are specific to the SP servers and processor nodes:

fault_service_Worm_RTG	Fault-handling daemon for the High Performance Switch
fault_service_Worm_SP	Fault-handling daemon for the SP Switch
fs_monitor	Sticky interrupt handler for the High Performance Switch
fs_monitor_SP	Sticky interrupt handler for the SP Switch
haemd	Event Manager daemon; matches information about the state of system resources with information about resource conditions that are of interest to client programs to create events; runs on all nodes of a system partition; one instance per system partition runs on the control workstation.
hagsd	Group Services daemon; provides services for coordinating and monitoring changes to the state of a client program that is running on a set of nodes in an SP system; runs on all nodes of a system partition; one instance per system partition runs on the control workstation.
hagsglsmd	A Group Services subsystem daemon that provides global synchronization services for the High Performance Switch adapter membership group; runs on all nodes of a system partition; one instance per system partition runs on the control workstation.
hardmon	Interfaces to the serial port on each SP frame; does tasks such as polling hardware for changes in hardware status, sending state changes to interested parties, sending commands to change power state and key lock position, and providing an interface to the S1 serial port of each processing node
hatsd	Topology Services daemon; provides information to other PSSP subsystems about the state of the nodes and adapters on the SP system; runs on all nodes of a system partition; one instance per system partition runs on the control workstation.
hbd	Heartbeat daemon; runs on all nodes, including the control workstation
hrd	Host responds daemon; runs only on the control workstation (See Appendix F, “The System Data Repository” on page 575 for more information.)
jmd	SP Resource Manager

© Copyright IBM Corp. 1995, 1997

kerberos	Kerberos authentication server (doesn't run on nodes)
kpropd	Kerberos backup communication daemon (doesn't run on nodes)
kshd	Authenticated server for SP remote commands **rsh** and **rcp**
pmand	The Event Management client daemon that dispatches actions based on subscribed-to events
pmanrmd	Provides resource variables for problem management
sdrd	System Data Repository server
sp_configd	Creates and sends SNMP traps when selected types of errors are recorded
splogd	Logs SP hardware state changes, reports SP hardware errors, and provides a state change alert
spmgrd	An SNMP Manager daemon that runs on the control workstation and interacts with SNMP Agents for extension nodes
supfilesrv	Provides services for file collections
switchtbld	Switch Table API daemon
sysctld	Sysctl daemon; runs on the control workstation, all nodes, and other RS/6000s
xntpd	Network time protocol daemon, keeps clocks synchronized across the network

The following AIX daemons may also run on the SP servers and processor nodes:

automount	Automount daemon; automatically mounts a directory on request based on map entries
cron	Runs shell commands at specified intervals
errdemon	Standard AIX error daemon for recording error log information
inetd	Provides Internet server management for a network
named	Provides server functions for the Domain Name Protocol
nfsd,biod, statd, lockd	Standard NFS daemons
portmap	Provides **RPC** services
qdaemon	Schedules print jobs from the **enq** command
rlogind	Provides **rlogin** services
rpe.mountd	Responds to client NFS mount requests
rpestatd, rpe.lockd	Provide System V style of file and record locking with NFS

sendmail	Routes mail for local or network delivery
snmpd	Provides the Agent functions of the simple network management protocol
srcmstr	System resource controller master daemon
syslogd	Logs system messages
telnetd	Provides **telnet** services
uprintfd	Constructs and writes kernel messages
writesrv	Allows the **write** command to send and receive messages
xmservd	Performance Toolbox
ypserv	Runs on NIS servers and slaves to provide NIS services
ypbind	Runs on all NIS clients to communicate with respective server

Appendix C. Sample Files

afsclient.cust

This is a sample of the **afsclient.cust** file.

```
#############################################################################
#
# Module: <afsclient.cust>
#
#---------------------------------------------------------------
#
# Description: This script is a sample which is designed to be used
#       as a guide for users who may want to install AFS on SP nodes.
#       This script is based on a premise that the Control Workstation (CW)
#       has already been installed and activated as an AFS client.
#
#       This sample script is to be executed by "root" user on the SP node.
#       This may be used with the /tftpboot/script.cust with modification
#       based on the customer requirements and configuration.
#
#############################################################################
PATH=/bin:/usr/bin:/etc:/usr/sbin:/usr/bin/X11:/var/sysman:/var/sysman/etc:
/etc/amd:/usr/lpp/ssp/bin:/usr/lpp/ssp/install/bin:/usr/afsws/etc:/usr/vice/etc:

SERVER=‘cat /etc/ssp/server_hostname | cut -d" " -f3’

#############################################################################
# Create the required AFS directory on /usr  filesystem. If the /usr is to
# be used as client/server make sure the directories are installed on the
# SP node server first, and then made available to the SP node clients.
# Make sure you have adequate space (about 5000K) is required to load the
# AFS execucatbles and files on the /usr filesystem.
# We will create the directories on /usr, and then copy the required files
# from the CW to the /usr/vice/etc directory on the SP node.
# For our sample configuration we will reference the CW as "$SERVER"
/usr/sbin/chfs  -a size='+10000' /usr

/bin/mkdir -p /usr/vice/etc/C
/bin/mkdir /usr/vice/etc/dkload
# Activate the ticket granting ticket so the remote copy will work successfully
#to bring AFS files from the CW to the SP node.
/usr/lpp/ssp/rcmd/bin/rcmdtgt

# Copy the files from CW /usr/vice/etc directory to the SP node /usr/vice/etc
# directory on the SP node.

/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/usr/vice/etc/CellServDB /usr/vice/etc/CellServDB
/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/usr/vice/etc/ThisCell   /usr/vice/etc/ThisCell
/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/usr/vice/etc/afsd    /usr/vice/etc/afsd
/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/usr/vice/etc/cacheinfo /usr/vice/etc/cacheinfo
/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/usr/vice/etc/C/* /usr/vice/etc/C/
/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/usr/vice/etc/dkload/* /usr/vice/etc/dkload/
/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/usr/vice/etc/rc.afs  /usr/vice/etc/rc.afs
```

© Copyright IBM Corp. 1995, 1997

```
#############################################################################
#  This next step will setup the cache that will be used on the SP nodes
#  These values can be modified, based on the amount of users that will be
#  using afs services. We will use the same values as placed on the CW for our
#  configuration being used in the sample script. The values to be modified are
#  the filesystem size for /usr/vice/cache, and /usr/vice/etc/cacheinfo  size.
#  This configuration will setup for a medium afs configuration of 4-8 users.
#  We will use a cacheinfo of 80000K, and filesystem of 90000K.
#  We will create the /usr/vice/cache directory and then create a JFS that will
#  be mounted
#
/bin/mkdir /usr/vice/cache

/usr/sbin/crfs -v jfs -g rootvg -a size=180000 -m '/usr/vice/cache' -A 'yes' -p 'rw'
/bin/sleep 5
/usr/sbin/mount /usr/vice/cache

#############################################################################
#  This next step will setup the the afs as one of the supported filesystems
#  in the /etc/vfs  file. You can either "edit" and place the information
#  afs     4       none                    none
#  in the file, or append the information using the "echo" command.
#

/bin/echo "afs     4       none                    none" >> /etc/vfs

#############################################################################
#  This next step will setup the the afs directory, and then activiate the afs
#  daemon (afsd) that connect to the AFS server. This will be executed by
#  starting the /etc/rc.afs script.
#
#  The /etc/rc.afs will setup and execute the AFS libraries to be used as
#  kernel extension in the SP nodes. You may want to use different execution
#  variables based on you configuration and its use with nfs services
#  We expect the files to be available in the /usr/vice/etc/dkload directory
#
#  /usr/vice/etc/dkload/cfgexport -a export.ext
#  /usr/vice/etc/dkload/cfgafs  -a afs.ext
#
#############################################################################

/bin/mkdir  /afs
/usr/vice/etc/rc.afs

/bin/sleep 60
#############################################################################
#  This next step will setup a soft link of the /afs/<cellname>/@sys/usr/afsws
#  directory to /usr/afsws on the SP node. The exact pathnames will need
#  to reflect  the afs configuration cell name, and the system directory.
#  This will be based on where the afsws executables are located.
#
#  /bin/ln -s /afs/<cellname>/@sys/usr/afsws  /usr/afsws
#  My cell execution will be the following:
/bin/ln -s /afs/ppd.pok.ibm.com/rs_aix41/usr/afsws   /usr/afsws

#############################################################################
#  This next step will setup the itab entry for afs in the /etc/inittab file.
#
```

```
/usr/sbin/mkitab "rcafs:2:wait:/usr/vice/etc/rc.afs>/dev/console 2>&1 /
# Start AFS daemon "

#############################################################################
#  This next step will add the /usr/afsws/etc /usr/afsws/bin and the
#  /usr/vice/etc into your current path environment. You may want to
#  update the path in .profile, or ksh.rc or other login setup.
PATH=$PATH:/usr/afsws/etc:/usr/afsws/bin:/usr/vice/etc:
```

block_usr_sample

This script creates a file of user names in a column format called called /**tmp**/**usr.input**. The **usr.input** file can be used with the **spacs_cntrl** command. Once you are comfortable using Login Control, you can uncomment the **spacs_cntrl** command in this script. This file is in /**usr**/**lpp**/**ssp**/**samples**.

```
#!/usr/lpp/ssp/perl/bin/perl
##################################################
#  Description:
#
#   This sample script will build a file of users from the /etc/passwd
#   file to input to spacs_cntrl.
#
#   The following items should be checked and possibly changed to conform
#   to the policy at your site.
#
#   1)  uid threshhold to start adding users to the file.  Should be large
#       enough to prevent system users such as root, adm, bin, lpd, etc.
#       from being added to the file for spacs_cntrl.  The arbitary value
#       used is 125.
#
#   2)  flags for spacs_cntrl. The logging flag is issued to
#       allow you to follow the actions of spacs_cntrl.  You may wish to
#       remove this flag once you are familiar with running spacs_cntrl.
#       The -s flag suppresses error messages which are logged then -l is
#       issued.
#
##################################################


# required files.
$usrfile="/tmp/usr.input";


$uidstart = 125;

$allokay = 0;
$syserror = 2;



# open the usr.input file

unless ( open(USRFILE,">$usrfile") ) {
      print "block_usr:  Cannot open usr.input file.\n";
      exit ($syserror);
      }

# write any names with a uid of 125 or above into file

while ( ($uname,$passwd,$uid) = getpwent ) {
    if ( $uid >= $uidstart ) {
      print USRFILE $uname,"\n";
    }
}
```

```
close(USRFILE);

# make sure file has entries

if ( -z $usrfile ) {
   print "block_usr:  No entries in $usrfile.  Not executing spacs_cntrl.\n";
   exit($syserror);
}


# Uncomment the following for this script to automatically run the spacs_cntrl command
# and block users.
# issue spacs_cntrl to block users in the file.

#system "/usr/lpp/ssp/bin/spacs_cntrl -s -l -f $usrfile block";
#$rc=$?;
#if ( $rc != 0 ) {
#   $rc = ($rc >> 8);
#   print "block_usr:  Possible error from spacs_cntrl.  Return code = $rc\n";
#   exit($rc);
#}

exit ($allokay)
```

bootptab.info File

If it exists, the /**etc**/**bootptab.info** file is read by the **sphrdwrad** command for hardware Ethernet addresses to be placed in the SDR. If it does not exist, then the **sphrdwrad** command still runs but takes longer to collect the information from the hardware.

Each line should list a SP node, specified by node number or frame,slot, followed by a blank and the hardware Ethernet address. The file should look similar to this:

```
17 02608C2E48D9
19 02608C2D6712
21 02608C2E49A4
23 02608C2E48E2
```

ibmSPDepNode.my

A copy of this MIB can be found in file **/usr/lpp/ssp/config/spmgrd/ibmSPDepNode.my**. The file contains configuration information for a dependent node adapter on a dependent node. It is used by the **spmgrd** daemon to send configuration attribute values to the SNMP Agent running on the dependent node.

```
-- Licensed Materials - Property of IBM
--
-- 5765-529
--
-- (C) Copyright IBM Corp. 1997 All Rights Reserved.
--
-- US Government Users Restricted Rights - Use, duplication or disclosure
-- restricted by GSA ADP Schedule Contract with IBM Corp.
--
-- Script Name:  depnode.my
--
-- Description:
--      definition of the depnode mib.
--
-- "@(#)75 1.2
 src/ssp/spmgr/src/ibmSPDepNode.my,spmgr, ssp_rspr, rsprt2d9 2/7/9

DEPNODE-MIB  DEFINITIONS ::= BEGIN

IMPORTS
         Counter, Gauge, TimeTicks, IpAddress, DisplayString, enterprises
                 FROM RFC1155-SMI
         TRAP-TYPE
                 FROM RFC1215;

ibm              OBJECT IDENTIFIER ::= { enterprises 2 }

ibmProd          OBJECT IDENTIFIER ::= { ibm 6 }

ibmSP            OBJECT IDENTIFIER ::= { ibmProd 117 }

-----------------------------------------------------------------
--

--               IBM NODEDEP MIB

--
ibmSPDepNode       OBJECT IDENTIFIER ::= { ibmSP 4 }

ibmSPDepNodeTable OBJECT-TYPE
       SYNTAX  SEQUENCE OF IbmSPDepNodeEntry
       ACCESS  not-accessible
       STATUS  mandatory
       DESCRIPTION
               "This entity's table of dependent secondary nodes
                attached to IBM SP switch networks."
       ::= { ibmSPDepNode 1 }

ibmSPDepNodeEntry OBJECT-TYPE
       SYNTAX  IbmSPDepNodeEntry
```

```
|        ACCESS  not-accessible
|        STATUS  mandatory
|        DESCRIPTION
|                "Configuration information for a dependent
|                 secondary node on SP switch network."
|        INDEX { ibmSPDepNodeName }
|        ::= { ibmSPDepNodeTable 1 }

|  IbmSPDepNodeEntry ::=
|         SEQUENCE {
|         ibmSPDepNodeName
|            DisplayString,
|         ibmSPDepNodeNumber
|            INTEGER,
|         ibmSPDepSwToken
|            OCTET STRING,
|         ibmSPDepSwARP
|            INTEGER,
|         ibmSPDepSwNodeNumber
|            INTEGER,
|         ibmSPDepIPaddr
|            IpAddress,
|         ibmSPDepNetMask
|            IpAddress,
|         ibmSPDepIPMaxLinkPkt
|            INTEGER,
|         ibmSPDepIPHostOffset
|            INTEGER,
|         ibmSPDepConfigState
|            INTEGER,
|         ibmSPDepSysName
|            DisplayString,
|         ibmSPDepNodeState
|            INTEGER,
|         ibmSPDepSwChipLink
|            INTEGER,
|         ibmSPDepNodeDelay
|            INTEGER,
|         ibmSPDepAdminStatus
|            INTEGER
|         }

|  ibmSPDepNodeName OBJECT-TYPE
|        SYNTAX  DisplayString
|        ACCESS  read-only
|        STATUS  mandatory
|        DESCRIPTION
|                "Identifier assigned to this node in its
|                 system's administrative environment."
|        ::= { ibmSPDepNodeEntry 1 }

|  ibmSPDepNodeNumber OBJECT-TYPE
|        SYNTAX  INTEGER
|        ACCESS  read-write
|        STATUS  mandatory
|        DESCRIPTION
|                "Relative node number assigned to this node in the SP system.
|                 Note that this is distinct from switch node number."
```

```
        ::= { ibmSPDepNodeEntry 2 }

  ibmSPDepSwToken OBJECT-TYPE
        SYNTAX  OCTET STRING (SIZE(10))
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
                "Opaque correlator assigned in the SP system
                 for this node's switch connection."
        ::= { ibmSPDepNodeEntry 3 }

  ibmSPDepSwARP OBJECT-TYPE
        SYNTAX  INTEGER {
                    disabled(1),
                    enabled(2)
                }
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
                "Indicates whether ARP is enabled for the partition
                 of the switch network that contains the associated
                 dependent secondary node.
                 NOTE: An SNMP Agent may be responsible for switch
                 nodes in several SP system partitions, and ARP
                 may be enabled in none, some, or all of the partitions.

                 A value of disabled(1) indicates that ARP is
                 not to be used to resolve network protocol
                 addresses to switch node numbers for the switch
                 network.  Instead, the mapping of IP addresses
                 to switch node numbers is determined via the
                 algorithm in the description of ibmSPDepIPHostOffset.

                 A value of enabled(2) indicates that ARP is
                 to be used to resolve network protocol
                 addresses to switch node numbers for the switch
                 network."
        ::= { ibmSPDepNodeEntry 4 }

  ibmSPDepSwNodeNumber OBJECT-TYPE
        SYNTAX  INTEGER (0..65535)
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
                "The switch node number of the associated SP switch
                 interface.  In common networking terminology,
                 this is the physical address for the interface
                 on the SP switch network, and is also called
                 the Node Device ID in the SP dependent node
                 architecture."
        ::= { ibmSPDepNodeEntry 5 }

  ibmSPDepIPaddr OBJECT-TYPE
        SYNTAX  IpAddress
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
                "The IP address for the associated SP switch interface."
```

```
        ::= { ibmSPDepNodeEntry 6 }

  ibmSPDepNetMask OBJECT-TYPE
        SYNTAX  IpAddress
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
                 "The subnet mask associated with the IP address of
                  this entry.  The value of the mask is an IP
                  address with all the network bits set to 1 and all
                  the host bits set to 0."
        ::= { ibmSPDepNodeEntry 7 }

  ibmSPDepIPMaxLinkPkt OBJECT-TYPE
        SYNTAX  INTEGER
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
                "The maximum number of bytes in the data portion of
                 link-layer packets for IP datagrams on the
                 SP switch network."
        ::= { ibmSPDepNodeEntry 8 }

  ibmSPDepIPHostOffset OBJECT-TYPE
        SYNTAX  INTEGER
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
                "This value is the difference between switch node
                 numbers and the host portion of the corresponding
                 IP addresses when ARP is disabled on the SP switch
                 network.  It is the value to subtract from the
                 host portion of an IP address to calculate the
                 corresponding switch node number."
        ::= { ibmSPDepNodeEntry 9 }

  ibmSPDepConfigState OBJECT-TYPE
        SYNTAX  INTEGER {
                    notConfigured(1),
                    firmwareLoadFailed(2),
                    driverLoadFailed(3),
                    diagnosticFailed(4),
                    microcodeLoadFailed(5),
                    fullyConfigured(6)
                }
        ACCESS  read-only
        STATUS  mandatory
        DESCRIPTION
                "This value indicates the final configuration state
                 of the corresponding SP switch interface."
        ::= { ibmSPDepNodeEntry 10 }

  ibmSPDepSysName OBJECT-TYPE
        SYNTAX  DisplayString
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
                "The fully-qualified domain name of the SP system
```

```
|                partition that contains this dependent node. This
|                name is administratively assigned in the SP system."
|       ::= { ibmSPDepNodeEntry 11 }

|  ibmSPDepNodeState OBJECT-TYPE
|       SYNTAX  INTEGER {
|                   nodeUp(1),
|                   nodeDown(2)
|               }
|       ACCESS  read-only
|       STATUS  mandatory
|       DESCRIPTION
|               "This value indicates the state of the dependent
|                node with respect to its capability to handle
|                Switch Manager protocol via its SP switch interface.

|                A value of nodeUp(1) indicates that the node has
|                been reinitialized or reconfigured sufficiently
|                to handle the Switch Manager protocol to unfence
|                the node if it were previously fenced from the
|                SP switch network.

|                A value of nodeDown(2) indicates that the node is
|                no longer able to handle Switch Manager protocol
|                messages on the SP switch network."
|       ::= { ibmSPDepNodeEntry 12 }

|  ibmSPDepSwChipLink OBJECT-TYPE
|       SYNTAX  INTEGER
|       ACCESS  read-write
|       STATUS  mandatory
|       DESCRIPTION
|               "The send port number for packets from the
|                adjacent switch chip to the dependent node.
|                The dependent node needs this value to build
|                the packet that it sends to the adjacent
|                switch chip to trigger that chip to reply
|                with a clock synchronization packet."
|       ::= { ibmSPDepNodeEntry 13 }

|  ibmSPDepNodeDelay OBJECT-TYPE
|       SYNTAX  INTEGER
|       ACCESS  read-write
|       STATUS  mandatory
|       DESCRIPTION
|               "This value is an estimate of how many clock
|                cycles elapse before a clock synchronization packet
|                sent from the adjacent switch chip to the dependent node
|                results in loading of the switch time-of-day register
|                for the node.

|                The dependent node needs this value to build
|                the packet that it sends to the adjacent
|                switch chip to trigger that chip to reply with
|                a clock synchronization packet."
|       ::= { ibmSPDepNodeEntry 14 }

|  ibmSPDepAdminStatus OBJECT-TYPE
```

```
        SYNTAX  INTEGER {
                  up(1),
                  down(2),
                  reconfigure(3)
                }
        ACCESS  read-write
        STATUS  mandatory
        DESCRIPTION
               "This value indicates the desired state of the dependent
               node with respect to its capability to handle the
               Switch Manager protocol via its SP switch interface.
               If the dependent node is not in the desired state,
               then the SNMP Agent triggers actions to change the state
               to the described state.

               A value of up(1) indicates that the node should be
               initialized and configured sufficiently to handle
               the Switch Manager protocol.

               A value of down(2) indicates that the node should be
               reset such that it is not able to receive and handle
               Switch Manager protocol messages on the SP switch
               network.

               A value of reconfigure(3) indicates that the SNMP Agent
               must send the switchInfoNeeded trap to request new
               SP switch configuration information for the dependent
               node.  This causes the SNMP Manager to send the new
               configuration information for the dependent node MIB
               objects via set-request PDU(s).  Once the new values
               are set by the SNMP Manager, the SNMP Agent reinitializes
               and reconfigures the dependent node sufficiently to
               handle the Switch Manager protocol."
        ::= { ibmSPDepNodeEntry 15 }

  -- IBM enterprise-specific traps dependent secondary nodes on SP switch:

  switchInfoNeeded TRAP-TYPE
        ENTERPRISE  ibmSP
        VARIABLES   { ibmSPDepNodeName }
        DESCRIPTION
                "A switchInfoNeeded trap signifies that the sending
                 protocol entity needs SP switch configuration
                 information for a dependent secondary node within
                 the sender's configuration."
        ::= 1

  switchConfigState TRAP-TYPE
        ENTERPRISE  ibmSP
        VARIABLES   { ibmSPDepConfigState }
        DESCRIPTION
                "A switchConfigState trap indicates the final
                 configuration state of a dependent secondary node's
                 SP switch interface."
        ::= 2

  switchNodeUp TRAP-TYPE
        ENTERPRISE  ibmSP
```

```
        VARIABLES   { ibmSPDepNodeName }
        DESCRIPTION
                "A switchNodeUp trap indicates that a dependent
                 secondary node in the sender's configuration has
                 just been reinitialized or reconfigured and is able
                 to handle SP Switch Manager protocol messages."
        ::= 3

  switchNodeDown TRAP-TYPE
        ENTERPRISE  ibmSP
        VARIABLES   { ibmSPDepNodeName }
        DESCRIPTION
                "A switchNodeDown trap indicates that a dependent
                 secondary node in the sender's configuration
                 is no longer able to handle Switch Manager protocol
                 messages on the switch network."
        ::= 4

  END
```

ibmSPMIB.my

A copy of this MIB can be found in file **/usr/lpp/ssp/config/snmp_proxy/ibmSPMIB.my**. The file defines SP-specific information for the creation of SNMP traps for the SP from AIX Error Log and Event Management Events. It also contains information about the nodes and the control workstation that make up the SP. The MIB is supported by the **sp_configd** daemon.

```
-- Licensed Materials - Property of IBM
--
-- 5765-529
--
-- (C) Copyright IBM Corp. 1996 All Rights Reserved.
--
-- US Government Users Restricted Rights - Use, duplication or disclosure
-- restricted by GSA ADP Schedule Contract with IBM Corp.
--
-- Script Name:  ibmSPMIB.my
--
-- Description:
-- definition of the ibmSP mib.
--

IBMSP-MIB  DEFINITIONS ::= BEGIN

IMPORTS
        Counter, Gauge, TimeTicks, IpAddress, DisplayString, enterprises
  FROM RFC1155-SMI
 TRAP-TYPE
  FROM RFC1215;

ibm     OBJECT IDENTIFIER ::= { enterprises 2 }

ibmProd OBJECT IDENTIFIER ::= { ibm 6 }

ibmSP   OBJECT IDENTIFIER ::= { ibmProd 117 }

-------------------------------------------------------------------------
--

--              IBM SP MIB

--

ibmSPConfig OBJECT IDENTIFIER ::= { ibmSP 1 }

ibmSPhostnodenumber OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS read-only
 STATUS mandatory
 DESCRIPTION
  "A number from 0 to the total number of slots.  It identifies
  the relative node number assigned to this processor."
 ::= { ibmSPConfig 1 }

ibmSPhostpartaddr OBJECT-TYPE
 SYNTAX IpAddress
 ACCESS  read-only
```

```
 STATUS mandatory
 DESCRIPTION
  "ip address assigned to the system partition in which this host
  resides. If this host is acting as a control workstation (i.e.
  the value of ibmSPhostnodenumber.0 is 0), this will be the ip
  address of the default partition."
 ::= { ibmSPConfig 2 }

ibmSPCWScodeversion OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The ssp release that is installed on the control workstation.
   Values are of the form 2.0, 2.1, etc. A NULL value means the
   release is not known."
 ::= { ibmSPConfig 3 }

ibmSPprimaryCWSname OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "hostname of the primary control workstation."
 ::= { ibmSPConfig 4 }

ibmSPprimaryCWSoperstatus OBJECT-TYPE
 SYNTAX INTEGER {
  up(1),
  down(2)
  }
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "operational state of the primary control workstation."
 ::= { ibmSPConfig 5 }

ibmSPbackupCWSname OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "hostname of the backup control workstation."
 ::= { ibmSPConfig 6 }

ibmSPbackupCWSoperstatus OBJECT-TYPE
 SYNTAX INTEGER {
  up(1),
  down(2)
  }
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "operational state of the backup control workstation."
 ::= { ibmSPConfig 7 }

ibmSPSystemTable OBJECT-TYPE
 SYNTAX SEQUENCE OF IbmSPNodeEntry
```

```
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "A list of SPNodeEntrys."
 ::= { ibmSPConfig 8 }

ibmSPNodeEntry OBJECT-TYPE
 SYNTAX IbmSPNodeEntry
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "Identifies a processor node residing in an SP frame."
 INDEX { ibmSPpartitionaddr, ibmSPnodenumber }
 ::= { ibmSPSystemTable 1 }

IbmSPNodeEntry ::=
 SEQUENCE {
   ibmSPpartitionaddr
    IpAddress,
   ibmSPnodenumber
    INTEGER,
   ibmSPframenumber
    INTEGER,
   ibmSPslotnumber
    INTEGER,
   ibmSPslotsused
    INTEGER,
   ibmSPinitialhostname
    DisplayString,
   ibmSPreliablehostname
    DisplayString,
   ibmSPsysparname
    DisplayString,
   ibmSPcodeversion
    DisplayString
   }

ibmSPpartitionaddr OBJECT-TYPE
 SYNTAX IpAddress
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "ip address assigned to the partition in which this node resides."
 ::= { ibmSPNodeEntry 1 }

ibmSPnodenumber OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS read-only
 STATUS mandatory
 DESCRIPTION
  "A number from 1 to the total number of slots.  It identifies
  the relative node number assigned to the processor."
 ::= { ibmSPNodeEntry 2 }

ibmSPframenumber OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS read-only
 STATUS mandatory
```

```
 DESCRIPTION
  "A number from 1 to the number of frames.  It identifies
  the number of the frame in which the processor node resides."
 ::= { ibmSPNodeEntry 3 }

ibmSPslotnumber OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS read-only
 STATUS mandatory
 DESCRIPTION
  "A number from 1 to the number of slots in a frame.  It
  identifies the number of the first slot occupied by the
  processor node within the frame."
 ::= { ibmSPNodeEntry 4 }

ibmSPslotsused OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS read-only
 STATUS mandatory
 DESCRIPTION
  "A number from 1 to two.  It identifies the number of slots
  occupied by the processor."
 ::= { ibmSPNodeEntry 5 }

ibmSPinitialhostname OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS read-only
 STATUS mandatory
 DESCRIPTION
  "The initial hostname assigned to the node during the SP customization
  phase."
 ::= { ibmSPNodeEntry 6 }

ibmSPreliablehostname OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS read-only
 STATUS mandatory
 DESCRIPTION
  "The hostname associated with the SP ethernet."
 ::= { ibmSPNodeEntry 7 }

ibmSPsysparname OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS read-only
 STATUS mandatory
 DESCRIPTION
  "The name of the system partition containing this processor
  node."
 ::= { ibmSPNodeEntry 8}

ibmSPcodeversion OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The ssp release that is installed on the processor node.
  Values are of the form 2.0, 2.1, etc. A NULL value means the
  release is not known."
```

```
 ::= { ibmSPNodeEntry 9 }

    -- The format of traps containing errlog entries whose tmplates
    -- are defined with 'Alert=yes'.

-- ibmSPErrlogTrap TRAP-TYPE
-- ENTERPRISE ibmSP
-- VARIABLES  { ibmSPellabel, ibmSPelidentifier, ibmSPeldatetime,
--    ibmSPelsequencenum, ibmSPelmachineid, ibmSPelnodeid,
--    ibmSPelclass, ibmSPeltype, ibmSPelresource,
--    ibmSPelrscclass, ibmSPelrsctype, ibmSPellocation,
--    ibmSPelvpd, ibmSPetdescription, ibmSPetprobcauses,
--    ibmSPetusercauses, ibmSPetuseraction, ibmSPetinstcauses,
--    ibmSPetinstaction, ibmSPetfailcauses, ibmSPetfailaction,
--    ibmSPetdetaildata}
-- DESCRIPTION
--    "These traps contain the contents of errlog entries formatted
--    into objects defining the contents of the errlog entry.
--    Since any single errlog entry does not contain all of the
--    fields defined in the collection of errlog templates, when a
--    object contains a null value, it will not be included in the
--    trap."
--    ::= ibmSPelidentifier.0

ibmSPErrlogVars OBJECT-TYPE
 SYNTAX SEQUENCE OF IbmSPErrlogEntry
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "A single SPNodeErrlogEntry."
 ::= { ibmSP 2 }

IbmSPErrlogEntry ::=
 SEQUENCE {
  ibmSPellabel
   DisplayString,
  ibmSPelidentifier
   DisplayString,
  ibmSPeldatetime
   DisplayString,
  ibmSPelsequencenum
   DisplayString,
  ibmSPelmachineid
   DisplayString,
  ibmSPelnodeid
   DisplayString,
  ibmSPelclass
   DisplayString,
  ibmSPeltype
   DisplayString,
  ibmSPelresource
   DisplayString,
  ibmSPelrscclass
   DisplayString,
  ibmSPelrsctype
   DisplayString,
  ibmSPellocation
   DisplayString,
```

```
  ibmSPelvpd
   DisplayString,
  ibmSPetdescription
   DisplayString,
  ibmSPetprobcauses
   DisplayString,
  ibmSPetusercauses
   DisplayString,
  ibmSPetuseraction
   DisplayString,
  ibmSPetinstcauses
   DisplayString,
  ibmSPetinstaction
   DisplayString,
  ibmSPetfailcauses
   DisplayString,
  ibmSPetfailaction
   DisplayString,
  ibmSPetdetaildata
   DisplayString
  }

ibmSPellabel OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The label associated with the error identifier defined by
  the ibmSPelidentifier object."
 ::= { ibmSPErrlogVars 1 }

ibmSPelidentifier OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A unique identifier defining the type of error entry written
  to the system error log existing on the host from which the
  trap originated."
 ::= { ibmSPErrlogVars 2 }

ibmSPeldatetime OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A timestamp identifying the time that this error log entry
  was written to the system error log.  It is of the form:
  day   month   day_of_month   hour:min:sec"
 ::= { ibmSPErrlogVars 3 }

ibmSPelsequencenum OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A decimal number which is the sequence number assigned to this
  entry.  This is the value specified on the -l switch of the
```

```
  errpt command used to obtain the trap data on the host from
  which the trap originated."
 ::= { ibmSPErrlogVars 4 }

ibmSPelmachineid OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A decimal number which is the machine ID of the host on which
  the trap originated. This is the value returned by the AIX
  'uname -m' command when issued on the host from which the trap
  originated."
 ::= { ibmSPErrlogVars 5 }

ibmSPelnodeid OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "An alpha-numeric string which is the node ID of the host on
  which the trap originated.  This is the value returned by the AIX
  'uname -m' command when issued on the host from which the trap
  originated."
 ::= { ibmSPErrlogVars 6 }

ibmSPelclass OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "An alphabetic string which is the error class of this entry:
  H (hardware), S (software), O (errlogger command messages)."
 ::= { ibmSPErrlogVars 7 }

ibmSPeltype OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "An alphabetic string which is the severity of the error entry:
  'PEND' (impending loss of availability), 'PERF' (unacceptable
  performance degradation), 'PERM' (permanent), 'TEMP' (temporary),
  'UNKN' (unknown)."
 ::= { ibmSPErrlogVars 8 }

ibmSPelresource OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The resource name associated with the error. For hardware
  errors this is a device name, for software errors this is
  the name of the failing executable, for operator command
  messages this is 'OPERATOR'."
 ::= { ibmSPErrlogVars 9 }

ibmSPelrscclass OBJECT-TYPE
```

```
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The resource class associated with the error. For hardware
  errors this is a device class (or 'NONE')."
 ::= { ibmSPErrlogVars 10 }

ibmSPelrsctype OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The resource type associated with the error.  For hardware
  errors this is a device type (or 'NONE'), for software errors
  (when specified) this is an LPP."
 ::= { ibmSPErrlogVars 11 }

ibmSPellocation OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "For hardware errors, information about the location of the
  failing device (or 'NONE')."
 ::= { ibmSPErrlogVars 12 }

ibmSPelvpd OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "IBM or user supplied vital product data."
 ::= { ibmSPErrlogVars 13}

ibmSPetdescription OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "Error description."
 ::= { ibmSPErrlogVars 14}

ibmSPetprobcauses OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "Probable causes of the error."
 ::= { ibmSPErrlogVars 15 }

ibmSPetusercauses OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "User actions which may have caused the error."
 ::= { ibmSPErrlogVars 16 }
```

```
ibmSPetuseraction OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "Recommended actions the user may take to correct the error."
 ::= { ibmSPErrlogVars 17 }

ibmSPetinstcauses OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "Installation causes of the error."
 ::= { ibmSPErrlogVars 18 }

ibmSPetinstaction OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "User actions which may have been performed during installation
  to cause the error."
 ::= { ibmSPErrlogVars 19 }

ibmSPetfailcauses OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A list of candidates which may be the source of the error."
 ::= { ibmSPErrlogVars 20 }

ibmSPetfailaction OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A list of reccommended actions which may be taken to correct
  the possible failures."
 ::= { ibmSPErrlogVars 21 }

ibmSPeldetaildata OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "Detailed data about this particular error."
 ::= { ibmSPErrlogVars 22 }

ibmSPEMVariables OBJECT IDENTIFIER ::= { ibmSP 3 }

-- The format of traps containing errlog entries whose tmplates
-- are defined with 'Alert=yes'.

--  ibmSPEMEventTrap TRAP-TYPE
--  ENTERPRISE ibmSP
```

```
|  --  ibmSPEMEventVarValue
|  --  DisplayString,
|  --  ibmSPEMEventPredicate
|  --  DisplayString
   --  VARIABLES  { ibmSPEMEventID, ibmSPEMEventFlags, ibmSPEMEventTime,
   --   ibmSPEMEventLocation, ibmSPEMEventPartitionAddress,
   --   ibmSPEMEventVarsTableName, ibmSPEMEventVarsTableInstanceID,
   --   ibmSPEMEventVarName, ibmSPEMEventVarValueInstanceVector,
   --   ibmSPEMEventVarValuesTableInstanceID,
   --   ibmSPEMEventVariableValue, ibmSPEMEventPredicate,
   --  DESCRIPTION
   --   "These traps contain the contents of events generated from the
   --   PSSP Event Manager.  The events have been formatted
   --   into objects defining the contents of the event and where
   --   the variable pertaining to the event and its value are located
   --   in the ibmSP mib.
   --  ::= ibmSPEMEventID.0

   ibmSPEMEvent OBJECT-TYPE
    SYNTAX SEQUENCE OF IbmSPEMEventEntry
    ACCESS  not-accessible
    STATUS mandatory
    DESCRIPTION
     "A single IbmSPEMEventEntry."
    ::= { ibmSPEMVariables 1 }

   ibmSPEMEventEntry ::=
    SEQUENCE {
     ibmSPEMEventID
      INTEGER,
     ibmSPEMEventFlags
      INTEGER,
     ibmSPEMEventTime
      TimeTicks,
     ibmSPEMEventLocation
      INTEGER,
     ibmSPEMEventPartitionAddress
      IpAddress,
     ibmSPEMEventVarsTableName
      DisplayString,
     ibmSPEMEventVarsTableInstanceID
      DisplayString,
     ibmSPEMEventVarName
      DisplayString,
     ibmSPEMEventVarValueInstanceVector
      DisplayString,
     ibmSPEMEventVarValuesTableInstanceID
      DisplayString,
     ibmSPEMEventVarValue
      DisplayString,
     ibmSPEMEventPredicate
      DisplayString
     }

   ibmSPEMEventID OBJECT-TYPE
    SYNTAX INTEGER
    ACCESS  read-only
    STATUS mandatory
```

```
 DESCRIPTION
  "The trap identifier assigned to the occurence of the Event."
 ::= { ibmSPEMEvent 1}

ibmSPEMEventFlags OBJECT-TYPE
 SYNTAX INTEGER {
  re-arm(1),
  false-predicate(2),
  unregister-event(4)
  }
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "EM Flags."
 ::= { ibmSPEMEvent 2}

ibmSPEMEventTime OBJECT-TYPE
 SYNTAX TimeTicks
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "Elapsed time between the activation of the SP proxy sub-agent
  and the occurence of the event."
 ::= { ibmSPEMEvent 3}

ibmSPEMEventLocation OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The number of the SP node from which the event was generated."
 ::= { ibmSPEMEvent 4}

ibmSPEMEventPartitionAddress OBJECT-TYPE
 SYNTAX IpAddress
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The ip address of the SP partition from which the event was
  generated."
 ::= { ibmSPEMEvent 5}

ibmSPEMEventVarsTableName OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name of the table in the SP MIB in which contains the variable
  definition objects."
 ::= { ibmSPEMEvent 6}

ibmSPEMEventVarsTableInstanceID OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The instance identifier used to locate the row in the table named
  by the ibmSPEMEventVarsTableName object value which contains
```

```
  further information about the EM Event variable definition."
 ::= { ibmSPEMEvent 7}

ibmSPEMEventVarName OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name of the EM variable  about which the event is recorded."
 ::= { ibmSPEMEvent 8}

ibmSPEMEventVarValueInstanceVector OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The instantiation vector of the EM variable instance that resulted
  in the event."
 ::= { ibmSPEMEvent 9}

ibmSPEMEventVarValuesTableInstanceID OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The instance identifier used to locate the row in the
  ibmSPEMVarValuesTable identifying for which instantiation of the named
  EM variable the value is reported. This is used to obtain the current
  variable value in the table to see if it has changed since the time
  the trap was issued."
 ::= { ibmSPEMEvent 10}

ibmSPEMEventVarValue OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The value of the variable instance at the time of the event."
 ::= { ibmSPEMEvent 11}

ibmSPEMEventPredicate OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The predicate string which caused the event."
 ::= { ibmSPEMEvent 12}

ibmSPEMNodeDepVarsTable OBJECT-TYPE
 SYNTAX SEQUENCE OF IbmSPEMNodeDepVarEntry
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "A table of node-dependent EM Variable attributes. Variables
  in this table are only instantiated on the node containing the
  resource monitor. See the ibmSPEMNodeDepVarLocator object
  description."
 ::= { ibmSPEMVariables 2 }
```

```
ibmSPEMNodeDepVarEntry OBJECT-TYPE
 SYNTAX IbmSPEMNodeDepVarEntry
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "The attributes of an event manager variable.  Each octet of
  the variable-length index string is encoded in a separate
  sub-identifier. The number of the sub-identifiers in the index
  is IMPLIED."
 INDEX { ibmSPEMNodeDepVarName }
 ::= { ibmSPEMNodeDepVarsTable 1 }

IbmSPEMNodeDepVarEntry ::=
 SEQUENCE {
  ibmSPEMNodeDepVarName
   DisplayString,
  ibmSPEMNodeDepVarDescr
   DisplayString,
  ibmSPEMNodeDepVarType
   DisplayString,
  ibmSPEMNodeDepVarDataType
   DisplayString,
  ibmSPEMNodeDepVarSBSFormat
   DisplayString,
  ibmSPEMNodeDepVarInitValue
   DisplayString,
  ibmSPEMNodeDepVarCurValueIndex
   INTEGER,
  ibmSPEMNodeDepVarClass
   DisplayString,
  ibmSPEMNodeDepVarVecElDefn
   DisplayString,
  ibmSPEMNodeDepVarVecElDescr
   DisplayString,
  ibmSPEMNodeDepVarPTXName
   DisplayString,
  ibmSPEMNodeDepVarDefPred
   DisplayString,
  ibmSPEMNodeDepVarEventDescr
   DisplayString,
  ibmSPEMNodeDepVarLocator
   DisplayString,
  ibmSPEMNodeDepVarOrderGroup
   DisplayString
  }

ibmSPEMNodeDepVarName OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A resource variable name as defined to the Event Manager."
 ::= { ibmSPEMNodeDepVarEntry 1 }

ibmSPEMNodeDepVarDescr OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
```

```
 STATUS mandatory
 DESCRIPTION
  "A description of the variable, including its semantics."
 ::= { ibmSPEMNodeDepVarEntry 2 }

ibmSPEMNodeDepVarType OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "One of the strings Counter, Quantity or State."
 ::= { ibmSPEMNodeDepVarEntry 3 }

ibmSPEMNodeDepVarDataType OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "One of the strings long, float or SBS. SBS is only permitted
  for a variable of the type State. If the value of this
  object is SBS, then the definitions of the structured fields
  that comprise the structured byte string contained in the
  ibmSPEMNodeDepVarSBSFormat object."
 ::= { ibmSPEMNodeDepVarEntry 4 }

ibmSPEMNodeDepVarSBSFormat OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "If the ibmSPEMNodeDepVarDataType object with the corresponding
  instance id has a value of SBS, then this object dscribes the
  structured fields within the variable. Included for each structured
  field is the structured field name, followed by an equal sign,
  followed by the data type for the field. The structured fields
  are defined sequentially beginning with sequence number 0."
 ::= { ibmSPEMNodeDepVarEntry 5 }

ibmSPEMNodeDepVarInitValue OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The initial value of a resource variable before it is observed
  for the first time."
 ::= { ibmSPEMNodeDepVarEntry 6 }

ibmSPEMNodeDepVarCurValueIndex OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "An index into the ibmSPEMVarValuesTable to locate value
  instances for this variable."
 ::= { ibmSPEMNodeDepVarEntry 7 }

ibmSPEMNodeDepVarClass OBJECT-TYPE
 SYNTAX DisplayString
```

```
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name of the class to which this resource variable belongs."
 ::= { ibmSPEMNodeDepVarEntry 8 }

ibmSPEMNodeDepVarVecElDefn OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name(s) of instantiation vector elements associated with the
  resource variable."
 ::= { ibmSPEMNodeDepVarEntry 9 }

ibmSPEMNodeDepVarVecElDescr OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A description of the instantiation vector elements associated with
  the resource variable."
 ::= { ibmSPEMNodeDepVarEntry 10 }

ibmSPEMNodeDepVarPTXName OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name used to read and write the resource variable in the PTX
  shared memory."
 ::= { ibmSPEMNodeDepVarEntry 11 }

ibmSPEMNodeDepVarDefPred OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The default predicate to be applied to the resource variable."
 ::= { ibmSPEMNodeDepVarEntry 12 }

ibmSPEMNodeDepVarEventDescr OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A description of the event generated by the application of the
  default predicate."
 ::= { ibmSPEMNodeDepVarEntry 13 }

ibmSPEMNodeDepVarLocator OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name of the vector element whose value is the number of the
  node containing the variable instance."
 ::= { ibmSPEMNodeDepVarEntry 14 }
```

```
ibmSPEMNodeDepVarOrderGroup OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name of an Availability group. All events generated by the
  default predicates of resource variables within this order group
  name are guaranteed ordered delivery with respect to one
  another."
 ::= { ibmSPEMNodeDepVarEntry 15 }

ibmSPEMNodeIndepVarsTable OBJECT-TYPE
 SYNTAX SEQUENCE OF IbmSPEMNodeIndepVarEntry
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "A table of node-independent EM Variable attributes.  No EM Locator
  element is defined for these variables so the location of the
  resource monitor is unknown within the partition. Variables
  in this table are only instantiated on the Control Work Station.
  See the ibmSPEMNodeDepVarLocator object description."
 ::= { ibmSPEMVariables 3 }

ibmSPEMNodeIndepVarEntry OBJECT-TYPE
 SYNTAX IbmSPEMNodeIndepVarEntry
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "The attributes of an event manager variable.  Each octet of
  the variable-length index string is encoded in a separate
  sub-identifier. The number of the sub-identifiers in the index
  is IMPLIED."
 INDEX { ibmSPEMNodeIndepPartaddr, ibmSPEMNodeIndepVarName }
 ::= { ibmSPEMNodeIndepVarsTable 1 }

IbmSPEMNodeIndepVarEntry ::=
 SEQUENCE {
  ibmSPEMNodeIndepPartaddr
   IpAddress,
  ibmSPEMNodeIndepVarName
   DisplayString,
  ibmSPEMNodeIndepVarDescr
   DisplayString,
  ibmSPEMNodeIndepVarType
   DisplayString,
  ibmSPEMNodeIndepVarDataType
   DisplayString,
  ibmSPEMNodeIndepVarSBSFormat
   DisplayString,
  ibmSPEMNodeIndepVarInitValue
   DisplayString,
  ibmSPEMNodeIndepVarCurValueIndex
   INTEGER,
  ibmSPEMNodeIndepVarClass
   DisplayString,
  ibmSPEMNodeIndepVarVecElDefn
   DisplayString,
```

```
  ibmSPEMNodeIndepVarVecElDescr
   DisplayString,
  ibmSPEMNodeIndepVarPTXName
   DisplayString,
  ibmSPEMNodeIndepVarDefPred
   DisplayString,
  ibmSPEMNodeIndepVarEventDescr
   DisplayString,
  ibmSPEMNodeIndepVarOrderGroup
   DisplayString
  }

ibmSPEMNodeIndepPartaddr OBJECT-TYPE
 SYNTAX IpAddress
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "ip address assigned to the system partition in which the
  this resource variable resides."
 ::= { ibmSPEMNodeIndepVarEntry 1 }

ibmSPEMNodeIndepVarName OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A resource variable name as defined to the Event Manager."
 ::= { ibmSPEMNodeIndepVarEntry 2 }

ibmSPEMNodeIndepVarDescr OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A description of the variable, including its semantics."
 ::= { ibmSPEMNodeIndepVarEntry 3 }

ibmSPEMNodeIndepVarType OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "One of the strings Counter, Quantity or State."
 ::= { ibmSPEMNodeIndepVarEntry 4 }

ibmSPEMNodeIndepVarDataType OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "One of the strings long, float or SBS. SBS is only permitted
  for a variable of the type State. If the value of this
  object is SBS, then the definitions of the structured fields
  that comprise the structured byte string contained in the
  ibmSPEMNodeIndepVarSBSFormat object."
 ::= { ibmSPEMNodeIndepVarEntry 5 }

ibmSPEMNodeIndepVarSBSFormat OBJECT-TYPE
```

```
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "If the ibmSPEMNodeIndepVarDataType object with the corresponding
  instance id has a value of SBS, then this object describes the
  structured fields within the variable. Included for each structured
  field is the structured field name, followed by an equal sign,
  followed by the data type for the field. The structured fields
  are defined sequentially beginning with sequence number 0."
 ::= { ibmSPEMNodeIndepVarEntry 6 }

ibmSPEMNodeIndepVarInitValue OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The initial value of a resource variable before it is observed
  for the first time."
 ::= { ibmSPEMNodeIndepVarEntry 7 }

ibmSPEMNodeIndepVarCurValueIndex OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "An index into the ibmSPEMVarValuesTable table to locate value
  instances for this variable."
 ::= { ibmSPEMNodeIndepVarEntry 8 }

ibmSPEMNodeIndepVarClass OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name of the class to which this resource variable belongs."
 ::= { ibmSPEMNodeIndepVarEntry 9 }

ibmSPEMNodeIndepVarVecElDefn OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name(s) of instantiation vector elements associated with the
  resource variable."
 ::= { ibmSPEMNodeIndepVarEntry 10 }

ibmSPEMNodeIndepVarVecElDescr OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A description of the instantiation vector elements associated with
  the resource variable."
 ::= { ibmSPEMNodeIndepVarEntry 11 }

ibmSPEMNodeIndepVarPTXName OBJECT-TYPE
 SYNTAX DisplayString
```

```
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name used to read and write the resource variable in the PTX
  shared memory."
 ::= { ibmSPEMNodeIndepVarEntry 12 }

ibmSPEMNodeIndepVarDefPred OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The default predicate to be applied to the resource variable."
 ::= { ibmSPEMNodeIndepVarEntry 13 }

ibmSPEMNodeIndepVarEventDescr OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A description of the event generated by the application of the
  default predicate."
 ::= { ibmSPEMNodeIndepVarEntry 14 }

ibmSPEMNodeIndepVarOrderGroup OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The name of an Availability group. All events generated by the
  default predicates of resource variables within this order group
  name are guaranteed ordered delivery with respect to one
  another."
 ::= { ibmSPEMNodeIndepVarEntry 15 }

ibmSPEMVarValuesTable OBJECT-TYPE
 SYNTAX SEQUENCE OF IbmSPEMVarValuesEntry
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "A table of current EM Variable values. Variables."
 ::= { ibmSPEMVariables 4 }

ibmSPEMVarValuesEntry OBJECT-TYPE
 SYNTAX IbmSPEMVarValuesEntry
 ACCESS  not-accessible
 STATUS mandatory
 DESCRIPTION
  "The current value of an event manager variable instantiation.
  The value of the ibmSPEMVarValueIndex is assigned when the
  SP sub-agent (sp_configd) is initialized; it is contained in the
  ibmSPEMNodeDepVarCurValueIndex object within the ibmSPNodeDepVarsTable
  (if the variable is node-dependent) or in the
  ibmSPEMNodeIndepVarCurValueIndex object within the
  ibmSPEMNodeIndepVarTable (if the variable is node-independent).
  Each octet of the variable-length ibmSPEMVarValueInstanceVector value
  string is encoded in a separate sub-identifier, preceded by its
  length which may be 0 if its value is null. A 0 length indicates
```

```
  the function represented by the EM variable is not being monitored."
 INDEX { ibmSPEMVarValueIndex, ibmSPEMVarValueInstanceVector }
 ::= { ibmSPEMVarValuesTable 1 }

IbmSPEMVarValuesEntry ::=
 SEQUENCE {
  ibmSPEMVarValueIndex
   INTEGER,
  ibmSPEMVarValueInstanceVector
   DisplayString,
  ibmSPEMVarValuePartaddr
   IpAddress,
  ibmSPEMVarValueName
   DisplayString,
  ibmSPEMVarValue
   DisplayString
  }

ibmSPEMVarValueIndex OBJECT-TYPE
 SYNTAX INTEGER
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "An index used as part of the instance id to identify the
  object instance containing the current value of an EM
  resource variable instance "
 ::= { ibmSPEMVarValuesEntry 1}

ibmSPEMVarValueInstanceVector OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "The instantiation vector elements associated with the
  resource variable."
 ::= { ibmSPEMVarValuesEntry 2}

ibmSPEMVarValuePartaddr OBJECT-TYPE
 SYNTAX IpAddress
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "ip address assigned to the system partition in which the
  this resource variable resides."
 ::= { ibmSPEMVarValuesEntry 3}

ibmSPEMVarValueName OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
 DESCRIPTION
  "A resource variable name as defined to the Event Manager."
 ::= { ibmSPEMVarValuesEntry 4}

ibmSPEMVarValue OBJECT-TYPE
 SYNTAX DisplayString
 ACCESS  read-only
 STATUS mandatory
```

```
 DESCRIPTION
  "The current value of a resource variable."
 ::= { ibmSPEMVarValuesEntry 5}

END
```

jmd_config

This is a sample of the **jmd_config** file. You can find it in /**etc**/.

```
# General rules:
# 1) Comment lines start with a # in the first column.
# 2) All keywords must start in the first column of the file.
# 3) To continue a line put \ at the end of the line to be continued.
# 4) If a keyword has several values they must be separated by ";".
# 5) Any hostnames specified here must have a
#    corresponding IP address and node number in the SDR.
# 6) All blank lines must have "\n" (carriage return) in the first column.
#    Any lines with "\n" preceded by one or more blanks will cause a syntax
#    error.
#
# This file contains an example of an SP2 with sixteen wide nodes.
#
# JM_LIST contains the list of SP2 hostnames that are candidates to run
# the JM server daemon.  This list will be searched in the order specified.
#
# The JM_LIST keyword can be repeated on subsequent lines and the
# lists will be concatenated.
JM_LIST=r07n01;r07n03;r07n05

# ACCESS_CONTROL specifies if the system administrator wants the JM to use the
# access control system management tool (/usr/lpp/ssp/bin/spacs_cntrl)
# to allow user access to a parallel node while it is allocated.
# If not specified resource manager will default to no.
# The ACCESS_CONTROL keyword can be repeated on subsequent lines but the
# last value will be the only one used.
# Specify  <yes or no>
ACCESS_CONTROL = yes

# EN_ADAPTER selects the ethernet adapter network that will
# be used to allocate resources for job application requests.
# Valid values for EN_ADAPTER are either 0 or 1 for en0 or en1,
# respectively.  The default value is 0 for en0.  The EN_ADAPTER
# value will determine which adapter is used regardless of what
# the node names are for the Pool members.  Specify <0 or 1>.
EN_ADAPTER = 0

# General rules for pools:
# 1) To describe a pool, the POOL_ID and ATTRIBUTES keywords have to be
#    specified. If you specify none of the MEMBERS_INTERACTIVE,
#    MEMBERS_BATCH, MEMBERS_GENERAL keywords (i.e. the pool has no
#    member nodes) then the pool will be ignored.
# 2) POOL_ID must be an unique integer greater than or equal to -1.
# 3) ATTRIBUTES must be a string with no blanks or "="
#    within the string and less than 1024 characters.
# 4) Any of the MEMBERS_INTERACTIVE, MEMBERS_BATCH, or MEMBERS_GENERAL,
#    keywords can be repeated on subsequent lines and the
#    corresponding lists will be concatenated.
# 5) Each node on the SP2 can be in only one pool and only one subpool
#    within that pool.  Since different IP hostnames can be used
#    to refer to the same node, you will receive an error if you
#    use different names referring to the same node within
#    multiple pools.
```

```
#
# Serial pool:
#
#  There can only be one serial pool. This information is defined
#  here so that the system administrator can record her decision
#  on where the batch serial jobs run (serial batch nodes), and
#  where the users are allowed to log into (serial interactive
#  nodes). Serial nodes that can be batch and interactive are put
#  into the general subpools. For the sake of distinguishing
#  between serial and parallel pools, the serial pool is required
#  to have a pool_id of -1.  The resource manager does not allocate
#  nodes in the serial pool.

POOL_ID= -1
ATTRIBUTES = This_is_the_serial_test_pool
# In this example:
#   r07n01 will be where the users are allowed to log into,
#   r07n03 will be where the batch serial jobs run,
#   r07n05 can be used for either of the above reasons.
MEMBERS_INTERACTIVE = r07n01
MEMBERS_BATCH       = r07n03
MEMBERS_GENERAL     = r07n05

# Parallel pools:
#
#  There can be several parallel pools, each pool having its own
#  characteristics.  The Resource Manager allocates nodes from the
#  interactive subpools for use by interactive parallel jobs and
#  from the batch subpools for use by jobs submitted to batch
#  schedulers like
#  LoadLeveler.  Nodes will be allocated from the general subpools
#  if there are no nodes available from the more specific
#  subpools.

POOL_ID = 0
ATTRIBUTES = IP_pool

# In this example:
#   r07n07  will be allocated to interactive parallel jobs,
#   r07n09 and r07n11 will be allocated to jobs submitted to a batch
#          scheduler (like LoadLeveler).
#   r07n013, r07n15, r07n17 can be used for either of the above reasons.

MEMBERS_INTERACTIVE = r07n07;
MEMBERS_BATCH   = r07n09; \
                  r07n11;
MEMBERS_GENERAL = r07n13;
MEMBERS_GENERAL = r07n15;
MEMBERS_GENERAL = r07n17;

POOL_ID = 1

# In this example:
#   r07n19 will be allocated to interactive parallel jobs,
#   r07n21 will be allocated to jobs submitted to a batch
#          scheduler (like LoadLeveler).
#   r07n23-r07n31 can be used for either of the above reasons.
```

```
ATTRIBUTES = user_space_pool
MEMBERS_INTERACTIVE = r07n19
MEMBERS_BATCH = r07n021;
MEMBERS_GENERAL = r07n23;r07n25;r07n27;r07n29;r07n31;
```

script.cust File

The **script.cust** file is located in **/usr/lpp/ssp/samples**.

```
#!/bin/ksh
######################################################################
#
# Module: <script.cust>
#
#--------------------------------------------------------------------
#
# Description: This script is a sample which is designed to be used by the
# user as a guide for some functions that might be performed
# after a node has been netinstalled, but before it boots for the
# first time.
#
# If you want  to use some of the examples in this file, then you need
# to either copy the entire script to /tftpboot/script.cust and
# edit the lines you want to execute, or just copy over the portions
# of this sample that you want executed to /tftpboot/script.cust
#
#
######################################################################

PATH=/bin:/usr/bin:/etc:/usr/sbin:/usr/bin/X11:/var/sysman:/var/sysman/etc:/etc/amd:
/usr/lpp/ssp/bin:/usr/lpp/ssp/install/bin:
SSPDIR=/usr/lpp/ssp
SERVER= cat /etc/ssp/server_hostname | cut -d" " -f3
#
# Set the time zone
# use smit chtz to see the choices for time zones
# (EST5)           Eastern U.S.; Colombia  - daylight time unobserved
# (EST5EDT)        Eastern U.S.; Colombia  - daylight time observed
#chtz EST5EDT
#
#########################################################################
# Importing existing volume groups
#########################################################################
#
#importvg -y uservg hdisk1
#
# To activate a volume group use varyonvg.
#varyonvg uservg
#
#########################################################################
# if you are mirroring rootvg volume group, uncomment the following
# in order to preserve the disabling of volume group quorum across
# an install (using a mksysb image with rootvg mirrored).
#
# /usr/sbin/chvg -Qn rootvg
#
#
# defining name resolution
# name resolution is required on the nodes. If you have not created an
# image that contains a method for resolving hostnames and addresses, then
# one of these methods must be selected.
#########################################################################
```

```
#
#########################################################################
# To perform kerberos authenticated rcp use the ssp rcmdtgt to obtain
# a kerberos ticket first.
#########################################################################
#
# Define the file that will store the ticket
#export KRBTKFILE=/tmp/tkt_rcmd$$
#
# Get the kerberos ticket
#/usr/lpp/ssp/rcmd/bin/rcmdtgt
#
# If you are using /etc/resolv.conf to resolve hostnames and addresses
# then use this to copy the /etc/resolv.conf file from the boot/install server
#
# Perform kerberos authenticated rcp
#/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/etc/resolv.conf /etc/resolv.conf
#
# If you are using the /etc/hosts file to resolve hostnames and addresses
# then use the following commands to copy and append the /etc/hosts
# file from the boot/install server
#
# Perform kerberos authenticated rcp
#/usr/lpp/ssp/rcmd/bin/rcp $SERVER:/etc/hosts /tmp/hosts
#/bin/cat /tmp/hosts >>/etc/hosts
#/bin/rm /tmp/hosts
#
# The ticket file is removed at the end of the script in case there
# are other users of authenticated rcp in the script.
#
# If you are using NIS to resolve hostnames and addresses then use the
# next section to define you NIS environment
#
#########################################################################
# defining nis domains, and clients
#########################################################################
#
# To set the NIS domainname use the domainname command.
# To list the NIS domainname, issue domainname with no parameters and it
# will list the NIS domainname.
#
#/usr/etc/yp/chypdom '-B' <NIS domainname>
#
#
# Configuring NIS servers is not possible during install because
# the srcmstr daemon is not running and therefore cannot start
# the required yp daemons for NIS servers at the time script.cust runs.
#
#
# Make this node an NIS client at next boot (not now)
#
#mkclient -I
#
#########################################################################
# installing other lpp's
# If you installed other lpp's on your system, made a mksysb image of that
#  system, and are planning to install /usr clients from that mksysb image,
#  you do not need to issue any installp commands here related to those
```

```
#  lpp's.  The /usr parts will be available on your /usr server (assuming
#  you installed the server with the same mksysb image), and the root parts
#  will get updated automatically by the finishrclient script, which is
#  run automatically after every /usr client is installed.
###############################################################################
#
# To install Fortran on a node create a backup image of the installp in
# the /usr/sys/inst.images/ssp directory of the control workstation and
# name it xlf.installp (make sure the installp image has read and execute
# permissions by all - chmod +rx xlf.installp).
#
# Check to see if fortran is already installed on this system
#if [[  -z lslpp -h | grep xlf ]]
#then
#  # Fortran not installed - Mount the installp image and perform the installp
#  /usr/sbin/mount <control workstation address>:/usr/sys/inst.images/ssp /mnt
#  [[ -f /mnt/pssp.installp ]] && installp -ac -X -d/mnt/xlf.installp all
#
#  # Unmount the installp image
#  /usr/sbin/unmount /mnt
#fi
#
###############################################################################
# Modifying page space
###############################################################################
#
#
# To make a new logical volume mypage for page space with 5 logical
# partitions (4 MB per partition - 20 MB total) on the volume group vg04,
# issue the mklv command.
# -t specifies the logical volume type (jfs, jfslog, or paging)
# -y specifies the name of the logical volume to be used instead of the system
# generated logical volume name(pagelvXX - where XX starts at 00 and increase
# by one).
# Use lsps -a to list all available paging space and its status.
#
#mklv -t paging -y mypage vg04 5
#
# To cause the /dev/mypage device to become available, use the swapon command.
#
#swapon /dev/mypage
#
# To change page space to be active on the next and subsequent reboots
# us the chps command.
# -a specifies to use page spae at next reboot
# y specifies to use page space at subsequent reboots
#
#chps -a y mypage
#
# Change the paging space - to add 100 logical partitions (400 MB )
# to mypage page space
# note: The default logical partition size is 4 MB.
# note: To use smit to change page space issue: smit chps
#
# Sample check to see if we are greater than what we are asking for
# NOTE: read comments for increasing logical volume space below
#
#num_pps_tox=100         # number of PPs to extend page space by
```

```
#
## determine cur size of mypage in PPs
#pps=/usr/sbin/lsps -c mypage | tail +2 | cut -d":" -f4
#if (( pps < num_pps_tox ))
#then
#chps -s$num_pps_tox mypage
#fi
#
# Create a 40 logical partition (default 4 MB partions - 160 MB) page space
# on the root volume group
# -a option actviates page space on subsequent reboots
# -n option actviates page space immediately
# -s option is the size of the page space in partitions
#
#mkps -s'40' -n -a rootvg
#
################################################################################
# increasing logical volume space
#
# Code below is a sample test for increasing LV one time and for a large
# amount... If you need to extend for small amounts, uncomment just
# the extendlv command line or do the extend manually, but remember to
# comment it out for subsequent customizes; otherwise the LV will be
# extended by that amount on ALL customizes.
#
# Assumption: LV is a minimal size (4MB = 1PPs) and want it to be 160MB
# at initial install time ...
################################################################################
#
## increasing logical volume space 40 logical partition (default 4 MB partions - 160 MB)
## pps -> physical partitions equate to logical partitions when LV is COPY=1
#
#num_pps_tox=100    # number of PPs to extend LV by
#
## determine current number PPs in hd3
#cur_pps_str= lslv hd3 | grep "LPs" | cut -d":" -f2
#cur_pps=${cur_pps_str%%PP*}
#
#if (( num_pps_tox >= cur_pps )) # already that big
#then
# extendlv hd3 $num_pps_tox  # extend by 100
#fi
#
#
################################################################################
# increasing a filesystem size
################################################################################
# NOTE: read comments above on increasing logical volume space ...
#
## increasing the /var filesystem by 4 MB
#
##check to see if we did the increase once, by checking size > extend amount
#
#increase=8192  # specified in 512 byte blocks; represents 4M increase
#fssize= /usr/sbin/lsfs -c /var | tail +2 | cut -d":" -f6
#if (( fssize < increase ))
#then
# chfs -a size=+$increase /var
```

```
#fi
#
#
################################################################################
#
################################################################################
# changing the maximum number of of users that can be concurrently logged
# in according to the one-time AIX license fee for the system. The change is
# put into effect after the next restart of the system.
################################################################################
#
#chlicense -u30
#
################################################################################
#
# To set the allcast attribute of a Token Ring Network Interface to
# not confine broadcast to the local LAN:
#
# chdev -l tr0 -a allcast=off -P
################################################################################
################################################################################
#
# Copy user management files from server to node.
# These are only some of the files that are copy over.
# The server address is gotten from the /etc/ssp/server_hostname file.
# If the server can not get to the correct files (passwd, group, etc) then
# place the correct ip address of where those files are in place of $SERVER.
#   (e.g.) /usr/bin/rcp -p 129.40.62.11:/etc/passwd /etc/passwd
# Check the /var/adm/SPlogs/sysman/<hostname>.config.log log file for errors
#
################################################################################
#
# /usr/bin/rcp -p $SERVER:/etc/passwd /etc/passwd
# /usr/bin/rcp -p $SERVER:/etc/group /etc/group
# /usr/bin/rcp -p $SERVER:/etc/security/passwd /etc/security/passwd
# /usr/bin/rcp -p $SERVER:/etc/security/group /etc/security/group
# /usr/bin/rcp -p $SERVER:/etc/security/user /etc/security/user
```

tuning.cust File

The **tuning.cust** file is located in /**usr**/**lpp**/**ssp**/**samples**.

```
#!/bin/ksh
#"@(#)96   1.2   src/ssp/samples/tuning.cust, sysman, ssp_r2.4 6/12/95 19:55:43"
#####################################################################
#
# Module: <tuning.cust>
#
#--------------------------------------------------------------------
#
# Description: This script is a sample which is designed to be used by the
#   user as a guide to change the network options for TCP/IP and UDP/IP
#   communications on the SP system
#
#   If you want to use some of the examples in this file, then you need
#   to either copy the entire script to /tftpboot/tuning.cust and
#   edit the lines you want to execute, or just copy over the portions
#   of this sample that you want executed to /tftpboot/tuning.cust
#
#   These settings are only a sample and should be changes to the preferred
#   settings for your system.
#
#   For further recommendations on tuning, refer to the Administration Guide
#
#--------------------------------------------------------------------

#
#   Set default mbuf pool size for all TCP/IP and thread buffer
#
#  /usr/sbin/no -o thewall=16384

#
#   Set default sb_max size for all IP connections
#
#  /usr/sbin/no -o sb_max=163840

#
#   Set default TCP send window size for all TCP/IP connections
#
#  /usr/sbin/no -o tcp_sendspace=65536

#
#   Set default TCP receive window size for all TCP/IP connections
#
#  /usr/sbin/no -o tcp_recvspace=65536

#
#   Set default UDP send window size for all UDP/IP connections
#   Effective maximum is limited to 65536
#
#  /usr/sbin/no -o udp_sendspace=32768

#
#   Set default UDP receive window size for all UDP/IP connections
```

```
#
#  /usr/sbin/no -o udp_recvspace=65536

#
#   Set rfc1323 feature on for faster throughput for High Performance Networks.
#
#  /usr/sbin/no -o rfc1323=1

#
#   Set default Maximum Transmission Unit size for transfers across networks.
#
#  /usr/sbin/no -o tcp_mssdflt=1448
#   Enable the kernel to forward packets
#  /usr/sbin/no -o ipforwarding=1
```

/usr/lib/X11/app-defaults/Spmon

```
! "@(#)37   1.2   src/ssp/spmon/gui/Spmon, spmon, ssp_r2.1 9/16/94 10:29:33" :
! X resource file for the spmon -g or spmon_gui commands:
! tcl wants colon : in comment lines too
! There arent many fonts common across all servers:
Tk*font: 8x13bold
Dialog*font: 8x13bold
```

Appendix D. System Monitor Variables

These variables can be queried from the **spmon** command interface, the **hmmon** command interface, or from the System Monitor GUI **spmon -g**.

Notes:

1. Responses provided (in parentheses) with descriptions refer to **spmon** command values. Responses provided [in brackets] refer to **hmmon** command values. Colors refer to the System Monitor GUI.
2. On the System Monitor GUI in the All Node Summary, not all variables are valid for all node types. When a variable is not supported by the target node, that variable is blanked out.

Use the **hmmon** command to determine the type of frame, node, and switch.

For example, to determine the frame type for Frame 1:

```
hmmon -G -Q -s -v type 1:0
   1   0   type          18   0x883a   supervisor type
```

To determine the node type in Frame 1, slot 4:

```
hmmon -Q -s -v type 1:4
   1   4   type          97   0x943a   supervisor type
```

or

```
hmmon -G -Q -s -v type 1:4
   1   4   type          97   0x943a   supervisor type
```

To determine the switch type in Frame 1, slot 17:

```
hmmon -G -Q -s -v type 1:17
   1   17   type          49   0x8c3a   supervisor type
```

or to determine the switch type in Frame 11, slot 2:

```
hmmon -G -Q -s -v type 11:2
   11   2   type          49   0x8c3a   supervisor type
```

Refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information on the **hmmon** command.

Frame Hardware Variables Common to All Frame Supervisors

Variable Name	Display Type	Description
codeVersion	Constant	Supervisor microcode version number
controllerResponds	Boolean	Indicator that supervisor is responding to request and commands from host. (1 [True] (green) = OK, 0 [False] (red) = no response)
daemonPollRate	Integer	Daemon poll rate in seconds
diagByte	Integer	Diagnosis byte. Correlates with error indicators. This variable is not currently being used.

© Copyright IBM Corp. 1995, 1997

Variable Name	Display Type	Description
nodefail 1 ⋮ nodefail 16	Boolean	Loss of communication from frame supervisor to specified node supervisor (1 [True] (red) = communication loss, 0 [False] (green) = communication OK)
nodefail17	Boolean	Loss of communication from frame supervisor to switch supervisor (1 [True] (red) = communication loss, 0 [False] (green) = communication OK)
nodeLinkOpen 1-16	Multistate	Node serial link1 open (1 [True]=open, 0 [False]=closed)
rs232CTS	Boolean	Indicator that the control node has received a Clear To Send signal over the RS232 link (1[True]=Clear To Send signal received, 0[False]=Clear To Send signal not received)
rs232DCD	Boolean	Indicator that the control node has received a Carrier Detect signal over the RS232 link (1[True]=Carrier Detect signal received, 0[False]=Carrier Detect signal not received)
serialNum	Constant	Supervisor serial number. This variable is not currently being used.
temp	Celsius	Temperature
tempRange	Boolean	Temperature out of range (1 [True] = temperature out of range, 0 [False] = temperature within range)
timeTicks	Integer	Frame Supervisor “heartbeat” counter (in numbers of 100 ms tics)
type	Constant	Frame controller card type (in decimal)
voltP48	Volts	Voltage of +48 volt line
voltP48Range	Boolean	+48 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)

Frame Hardware Variables Unique to Card Type 17 and 20 Frame Supervisor

Variable Name	Display Type	Description
frACLED	Multistate	Represents the bottom green frame supervisor LED, showing status of AC power (actually monitors +12VDC present at frame supervisor) (2/3 [2] (gray) = LED blinking: AC Power failure, 1 [1] (green) = LED lit: AC power OK, 0 [0] (black) = LED not lit: no AC power exists)
frDCLED	Multistate	Represents of the top green frame supervisor LED, showing status of frame AC-to-DC power supplies. (2/3 [2] (gray) = LED blinking: DC Power failure, 1 [1] (green) = LED lit: DC power OK, 0 [0] (black) = LED not lit: no DC power exists)
frPowerDC_A	Boolean	Status of DC power supply A (1 [True] (green) = OK, 0 [False] (red) = not OK)
frPowerDC_B	Boolean	Status of DC power supply B (1 [True] (green) = OK, 0 [False] (red) = not OK)
frPowerDC_C	Boolean	Status of DC power supply C (1 [True] (green) = OK, 0 [False] (red) = not OK)
frPowerDC_D	Boolean	Status of DC power supply D (1 [True] (green) = OK, 0 [False] (red) = not OK)

Variable Name	Display Type	Description
frPowerOff_A	Boolean	Power switch for power supply A. Use the **-power** parameter on the **spmon** command to turn the power on or off. The frame supervisor returns values indicating whether the power is on or off. (1[True]=off, 0[False]=on)
frPowerOff_B	Boolean	Power switch for power supply B. Use the **-power** parameter on the **spmon** command to turn the power on or off. The frame supervisor returns values indicating whether the power is on or off. (1[True]=off, 0[False]=on)
frPowerOff_C	Boolean	Power switch for power supply C. Use the **-power** parameter on the **spmon** command to turn the power on or off. The frame supervisor returns values indicating whether power is on or off. (1[True]=off, 0[False]=on)
frPowerOff_D	Boolean	Power switch for power supply D. Use the **-power** parameter on the **spmon** command to turn the power on or off. The frame supervisor returns values indicating whether power is on or off. (1[True]=off, 0[False]=on)

Frame Hardware Variables Unique to SEPBU (Type 18) and HACWS SEPBU (Type 22) Frame Supervisors

Variable Name	Display Type	Description
amp_A	Amps	Current output of DC Module A
amp_ARange	Boolean	A current out of range (1 [True] = current out of range, 0 [False] = current within range)
amp_B	Amps	Current output of DC Module B
amp_BRange	Boolean	B current out of range (1 [True] = current out of range, 0 [False] = current within range)
amp_C	Amps	Current output of DC Module C
amp_CRange	Boolean	C current out of range (1 [True] = current out of range, 0 [False] = current within range)
amp_D	Amps	Current output of DC Module D
amp_DRange	Boolean	D current out of range (1 [True] = current out of range, 0 [False] = current within range)
frNoPowerModA	Boolean	Indicates if the DC power module A is installed (0 [False] = installed, 1 [True] = not installed)
frNoPowerModB	Boolean	Indicates if the DC power module B is installed (0 [False] = installed, 1 [True] = not installed)
frNoPowerModC	Boolean	Indicates if the DC power module C is installed (0 [False] = installed, 1 [True] = not installed)
frNoPowerModD	Boolean	Indicates if the DC power module D is installed (0 [False] = installed, 1 [True] = not installed)
frPowerModAbad	Boolean	Status of DC power module A (0 [False] (green) = OK, 1 [True] (red) = not OK)
frPowerModBbad	Boolean	Status of DC power module B (0 [False] (green) = OK, 1 [True] (red) = not OK)

Variable Name	Display Type	Description
frPowerModCbad	Boolean	Status of DC power module C (0 [False] (green) = OK, 1 [True] (red) = not OK)
frPowerModDbad	Boolean	Status of DC power module D (0 [False] (green) = OK, 1 [True] (red) = not OK)
frPowerLED	Boolean	Status of power supply (1 [1] = LED lit, power OK, 0 [0] = LED not lit, no power)
frPowerMods	Multistate	Represents the bottom green frame supervisor LED, showing whether power modules are present (2/3 [2] (gray)=LED blinking: no modules are present, 1 [1] (green)=LED lit: at least 1 module is present, 0 [0] (black)=LED not lit)
frPowerOff	Boolean	Frame power (1 [True] = off, 0 [False] = on)

Frame Hardware Variables Unique to HACWS (Type 20) and SEPBU HACWS (Type 22) Frame Supervisors

Variable Name	Display Type	Description
controllerIDMismatch	Boolean	(0 [True] = the frame ID stored in the HACWS supervisor card does not match the configured frame ID, 1 [False] = the frame ID stored in the HACWS supervisor card matches the configured frame ID)
controllerTail	Integer	(1 [1] = the monitor is connected via S1 to the frame supervisor, 2 [2] = the monitor is connected via S2 to the frame supervisor)
controllerTailActive	Boolean	(1 [True] (green) = the RS232 connection between the control workstation and the frame is active; the frame supervisor accepts hardware control commands from the monitor, 0 [False] (red) = the RS232 connection between the control workstation and the frame is not active; the frame supervisor does not accept hardware control commands from the monitor) In an HACWS environment, only one of the two connections should be active.
controllerTail1LED	Integer	Represents green LED 4 frame supervisor LED, showing the status of the S1 tail (1 [1] (green) LED on: S1 tail active, 0 [0] (red) LED off: S1 tail inactive)
controllerTail2LED	Integer	Represents green LED 3 frame supervisor LED, showing the status of the S2 tail (1 [1] (green) LED on: S2 tail active, 0 [0] (red) LED off: S2 tail inactive)
frRS232Active1	Integer	Represents the yellow LED 8 frame supervisor LED, showing the status of the RS-232Active1 communication with the control workstation (1 [1] (yellow) = LED on: RS232 active, 2/3 [2] (gray) = LED blinking: RS232 error, 0 [0] (black) = LED off: RS232 inactive)
frRS232Active2	Integer	Represents the yellow LED 7 frame supervisor LED, showing the status of the RS-232Active2 communication with the control workstation (1 [1] (yellow) = LED on: RS232 active, 2/3 [2] (gray) = LED blinking: RS232 error, 0 [0] (black) = LED off: RS232 inactive)
frTTNodeComm	Integer	Represents the yellow LED 6 frame supervisor LED, showing status of communication with nodes over the supervisor bus (I2C bus) (2/3 (gray)=LED blinking: bus dead-no nodes responding, 1 (yellow) = LED lit: 1 or more nodes not responding, 0 (black) = LED not lit: all nodes responding)

Frame Hardware Variables Unique to Card Type 17 and SEPBU (Type 18) Frame Superviors

Variable Name	Display Type	Description
frNodeComm	Integer	Represents the top yellow frame supervisor LED, showing status of communication with nodes over the supervisor bus (I2C link) (2/3 [2] (gray)=LED blinking: bus dead-no nodes responding, 1 [1] (yellow) = LED lit: 1 or more nodes not responding, 0 [0] (black) = LED not lit: all nodes responding)
frRS232Active	Integer	Represents the bottom yellow frame supervisor LED, showing status of serial (RS/232) communication with the control console (2/3 [2] (gray)=LED blinking: RS232 error, 1 [1] (yellow) = LED lit: RS232 activity, 0 [0] (black) = LED not lit: RS232 no activity)

Node Hardware Variables Common to All Node Supervisors

Variable Name	Display Type	Description
codeVersion	Constant	Supervisor microcode version number
commFail	Boolean	Indicates loss of communication to node supervisor from frame supervisor (1 [True] (red) = communication failure, 0 [False] (green) = communication OK) Display using the **spmon** command, use **-q -t frame*/nodefail*/value**
console	openConsole	Opens a window to the node's serial 1 port (no values are available for this object). The serialLinkOpen object indicates if a serial 1 port is open. To open a window to the serial 1 port using the **spmon** command, use the -open parameter or **s1term** command. Can't be queried with **hmmon** or **spmon**; can be selected from the All Node Summary of the System Monitor GUI.
envLED	Multistate	Reflects environment light on front of node. (2/3 [2] (gray) = LED blinking: power is shut off due to detection of a critical error (voltages, fans, temp., etc), 1 [1] (yellow) = LED lit: warning indicator-not yet critical, 0 [0] (black) = LED not lit: OK)
hostResponds	Boolean	Status of communication between the processor node and the control workstation (1 (green) = communication OK, 0 (red) = communication failure) Can't be queried from **hmmon**; can be selected from All Node Summary of System Monitor GUI.
keyModeSwitch	keySwitch	Three position key switch. To control the setting of the keyModeSwitch using the **spmon** command, use the **-key** parameter. To display the setting of the **keyModeSwitch** using the **spmon** command, use the -Key parmeter. The node supervisors return values indicating the position of the keyModeSwitch. (2 [2] (red) = service, 1 [1] (yellow) = secure, 0 [0] (green) = normal)
LED7SegA	7seg	** 1st digit of 3 digit display
LED7SegB	7seg	** 2nd digit of 3 digit display
LED7SegC	7seg	** 3rd digit of 3 digit display

Variable Name	Display Type	Description
nodePower	powerSwitch	To control the power using the **spmon** command, use the -power parameter and the -reset parameter. The node supervisors return values indicating whether nodePower is on or off. (1 [True] (green) = power switch on, 0 [False] (red) = power switch off)
powerLED	Multistate	Reflects power light on front of node. (2/3 [2] (gray) = LED blinking: power off but available, 1 [1] (green) = LED lit: power on, 0 [0] (black) = LED not lit: no power exists)
P48OK	Boolean	+48 volt input is OK (1 [True] (green) = OK, 0 [False] (red) = not OK)
serialLinkOpen	Multistate	Indicates whether the serial 1 port is open (1 [True] (yellow) = port opened, 0 [False] (black) = port closed)
switchResponds	Boolean	This variable is only defined when a switch is installed on the frame in which the node resides. On High Performance Switch systems, it indicates whether the switch is active on the node (1 (green) = active, 0 (red)= inactive, 2(yellow) = isolated) On SP Switch systems, it indicates whether the switch is active on the node (1 (green) = active, 0 (red) = not active on the switch for one of the following reasons • It was “fenced” without the **-autojoin** option • It faulted and was removed from the switch • The node's adapter failed to configure properly • The node's SP Switch adapter was recently reset by the **/usr/lpp/ssp/css/rc.switch** script. 2 (yellow) = “fenced” with the **-autojoin** option) Can't be queried from **hmmon**; can be selected from All Node Summary of System Monitor GUI.
temp	Celsius	Temperature (red = node shutdown due to temp failure, yellow = warning, temp is outside proper threshold, green = temp is OK)
tempRange	Boolean	Temperature out of range (1 [True] = temperature out of range, 0 [False] = temperature within range)
timeTicks	Integer	Node Supervisor “heartbeat” counter (in numbers of 100 ms tics)
type	Constant	Node Supervisor card type (in decimal)
3DigitDisplay	ledDisplay	Three-digit display. This variable cannot be used as a target of the **spmon** command. To display the contents of a three-digit display using the **spmon** command, use the -Led parameter. Information from **spmon -Led** can't be used with a 604 High Node. Can't be queried with **hmmon** or **spmon**; can be selected from the All Node Summary of the System Monitor GUI.

** Note: when querying the 3-digit LED values, the integers correspond to the following display characters:

Table 18. LED Translation Values	
192 = 0	248 = 7
249 = 1	128 = 8
164 = 2	144 = 9
176 = 3	167 = c
153 = 4	179 = d
146 = 5	157 = u
130 = 6	255 = blank

Node-Specific Hardware Variables for Card Types 33 and 65 Node Supervisor

Variable Name	Display Type	Description
currentShutdown	Boolean	Current shutdown (1 [True] = shutdown, 0 [False] = current OK)
diagByte	Integer	Diagnosis byte (correlates with error indicators) This variable is not currently being used.
fanfail1-3	Boolean	Environmental failure: Shutdown due to fan out of specifications condition (1 [True] (red) = shutdown, 0 [False] (green) = fan OK)
fanwarning1-3	Boolean	Environmental warning: Fan nearing out of specifications condition (1 [True] (red) = warning, 0 [False] (green) = fan OK)
memoryProtect	Boolean	Represents the MemoryProtect signal received from the I/O planar of the processor node. The MemoryProtect signal indicates whether the memory cards in the processor node are the correct type (voltage) for this processor node. (0 [False] (green) = OK, 1 [True] (red) = incorrect voltage type memory card)
nodePowerOn10Sec	Boolean	Indicates node has been operational more than 10 seconds and remains so (1 [True] (green) = powered on, 0 [False] (black) = powered off)
P48Off10Sec	Boolean	+48 volt input has been terminated due to error and remains off (1 [True] (red) = powered off, 0 [False] (black) = OK)
resetActive	Boolean	Indicates that the node is resetting (1 [True] = node is resetting, 0 [False] = node is not resetting)
serialNum	Constant	Supervisor serial number. This variable is not currently being used
shutdownN12High	Boolean	Environmental failure: Node shutdown due to -12 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownN12Low	Boolean	Environmental failure: Node shutdown due to -12 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5High	Boolean	Environmental failure: Node shutdown due to +5 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5Low	Boolean	Environmental failure: Node shutdown due to +5 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)

Variable Name	Display Type	Description
shutdownP12High	Boolean	Environmental failure: Node shutdown due to +12 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP12Low	Boolean	Environmental failure: Node shutdown due to +12 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP48Low	Boolean	Environmental failure: Node shutdown due to +48 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownTemp	Boolean	Environmental failure: Shutdown due to temperature out of specifications (1 [True] (red) = shutdown, 0 [False] (green) = temperature OK)
s1PortDTR	Boolean	Indicates serial 1 port DTR asserted. A console cannot be opened to the node's serial 1 port unless this flag is on. (1 [True] (green) = port asserted, 0 [False] (red) = port not asserted)
voltN12	Volts	Voltage on -12 volt line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltN12Range	Boolean	-12 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP5	Volts	Voltage on +5 volt line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP5Range	Boolean	+5 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP12	Volts	Voltage on +12 volt line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP12Range	Boolean	+12 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP48	Volts	Voltage on +48 volt line
voltP48Range	Boolean	+48 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
warningN12High	Boolean	Environmental warning: Node warning due to marginal specifications condition on -12 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningN12Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on -12 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP5High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP5Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP12High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +12 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)

Variable Name	Display Type	Description
warningP12Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +12 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
7segChanged	Multistate	Indicates the digits within the 3-digit display are changing faster than once per second. (1 [True] (black) digits changing faster than once per second, 0 [False] (yellow) digits changing less than once per second)

Node-Specific Hardware Variables for Card Type 97 Node Supervisors

Variable Name	Display Type	Description
diagByte	Integer	Diagnosis byte (correlates with error indicators) This variable is not currently being used.
fanfail1-3	Boolean	Environmental failure: Shutdown due to fan out of specifications condition (1 [True] (red) = shutdown, 0 [False] (green) = fan OK)
fanfail4	Boolean	Environmental failure: Shutdown due to fan out of specifications condition (1 [True] (red) = shutdown, 0 [False] (green) = fan OK)
fanwarning1-3	Boolean	Environmental warning: Fan nearing out of specifications condition (1 [True] (red) = warning, 0 [False] (green) = fan OK)
memoryProtect	Boolean	Represents the MemoryProtect signal received from the I/O planar of the processor node. The MemoryProtect signal indicates whether the memory cards in the processor node are the correct type (voltage) for this processor node. (0 [False] (green) = OK, 1 [True] (red) = incorrect voltage type memory card)
nodePowerOn10Sec	Boolean	Indicates node has been operational more than 10 seconds and remains so (1 [True] (green) = powered on, 0 [False] (black) = powered off)
P4dPresent	Multistate	Indicates daughter card 4v supply installed (1 [True] (green) = 4v card installed, 0 [False] (black) = 4v card not installed)
P48Off10Sec	Boolean	+48 volt input has been terminated due to error and remains off (1 [True] (red) = powered off, 0 [False] (black) = OK)
P5DCok	Boolean	Voltage on +5 volt line is good (1 [True] (green) = voltage OK, 0 [False] (red) = voltage not on)
P2_5dPresent	Multistate	Indicates daughter card 2.5v supply installed (1 [True] (green) = 2.5v card installed, 0 [False] (black) = 2.5v card not installed)
resetActive	Boolean	Indicates that the node is resetting (1 [True] = node is resetting, 0 [False] = node is not resetting)
serialNum	Constant	Supervisor serial number. This variable is not currently being used
shutdownN12High	Boolean	Environmental failure: Node shutdown due to -12 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownN12Low	Boolean	Environmental failure: Node shutdown due to -12 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP4dHigh	Boolean	Environmental failure: Node shutdown due to +4 volt daughter card line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK) / not installed

Variable Name	Display Type	Description
shutdownP4dLow	Boolean	Environmental failure: Node shutdown due to +4 volt daughter line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK) / not installed
shutdownP5High	Boolean	Environmental failure: Node shutdown due to +5 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5Low	Boolean	Environmental failure: Node shutdown due to +5 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP2_5dHigh	Boolean	Environmental failure: Node shutdown due to +2.5 volt daughter card line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK) / not installed
shutdownP2_5dLow	Boolean	Environmental failure: Node shutdown due to +2.5 volt daughter line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK) / not installed
shutdownP12High	Boolean	Environmental failure: Node shutdown due to +12 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP12Low	Boolean	Environmental failure: Node shutdown due to +12 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownTemp	Boolean	Environmental failure: Shutdown due to temperature out of specifications (1 [True] (red) = shutdown, 0 [False] (green) = temperature OK)
s1PortDTR	Boolean	Indicates serial 1 port DTR asserted. A console cannot be opened to the node's serial 1 port unless this flag is on. (1 [True] (green) = port asserted, 0 [False] (red) = port not asserted)
voltN12	Volts	Voltage on -12 volt line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltN12Range	Boolean	-12 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP4d	Volts	Voltage on +4 volt daughter card line; only valid when the P4dPresent value is 1 (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP4dRange	Boolean	+4 volt daughter out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP5	Volts	Voltage on +5 volt line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP5Range	Boolean	+5 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP2_5d	Volts	Voltage on +2.5 volt daughter card line; only valid when the P2_5dPresent value is 1 (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP2_5dRange	Boolean	+2.5 volt daughter out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)

Variable Name	Display Type	Description
voltP12	Volts	Voltage on +12 volt line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP12Range	Boolean	+12 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
warningN12High	Boolean	Environmental warning: Node warning due to marginal specifications condition on -12 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningN12Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on -12 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP4dHigh	Boolean	Environmental warning: Node warning due to marginal specifications condition on +4 volt daughter card line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK) / not installed
warningP4dLow	Boolean	Environmental warning: Node warning due to marginal specifications condition on +4 volt daughter card line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK) / not installed
warningP5High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP5Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP2_5dHigh	Boolean	Environmental warning: Node warning due to marginal specifications condition on +2.5 volt daughter card line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK) / not installed
warningP2_5dLow	Boolean	Environmental warning: Node warning due to marginal specifications condition on +2.5 volt daughter card line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK) / not installed
warningP12High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +12 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP12Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +12 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningTemp	Boolean	Environmental warning: Temperature nearing out of specifications condition (1 [True] (red) = warning, 0 [False] (green) = temperature OK)
7segChanged	Multistate	Indicates the digits within the 3-digit display are changing faster than once per second. (1 [True] (black) digits changing faster than once per second, 0 [False] (yellow) digits changing less than once per second)

Node-Specific Hardware Variables for Card Types 81 and 83 Node Supervisors

Variable Name	Display Type	Description
diagByte	Integer	Diagnosis byte (correlates with error indicators) This variable is not currently being used.

Variable Name	Display Type	Description
fanfail1d-5d	Boolean	Environmental failure: Shutdown due to fan out of specifications condition (1 [True] (red) = shutdown, 0 [False] (green) = fan OK)
fanwarning1-3	Boolean	Environmental warning: Fan nearing out of specifications condition (1 [True] (red) = warning, 0 [False] (green) = fan OK)
fanwarning4-5	Boolean	Environmental warning: Fan nearing out of specifications condition (1 [True] (red) = warning, 0 [False] (green) = fan OK)
nodePowerOn10Sec	Boolean	Indicates node has been operational more than 10 seconds and remains so (1 [True] (green) = powered on, 0 [False] (black) = powered off)
P48Off10Sec	Boolean	+48 volt input has been terminated due to error and remains off (1 [True] (red) = powered off, 0 [False] (black) = OK)
P5DCok	Boolean	Voltage on +5 volt line is good (1 [True] (green) = voltage OK, 0 [False] (red) = voltage not on)
resetActive	Boolean	Indicates that the node is resetting (1 [True] = node is resetting, 0 [False] = node is not resetting)
serialNum	Constant	Supervisor serial number. This variable is not currently being used
shutdownN12High	Boolean	Environmental failure: Node shutdown due to -12 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownN12Low	Boolean	Environmental failure: Node shutdown due to -12 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP4High	Boolean	Environmental failure: Node shutdown due to +4 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP4Low	Boolean	Environmental failure: Node shutdown due to +4 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5iHigh	Boolean	Environmental failure: Node shutdown due to +5 volt I/O supply line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5iLow	Boolean	Environmental failure: Node shutdown due to +5 volt I/O supply line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5mHigh	Boolean	Environmental failure: Node shutdown due to +5 volt memory supply line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5mLow	Boolean	Environmental failure: Node shutdown due to +5 volt memory supply line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP12High	Boolean	Environmental failure: Node shutdown due to +12 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP12Low	Boolean	Environmental failure: Node shutdown due to +12 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownTemp	Boolean	Environmental failure: Shutdown due to temperature out of specifications (1 [True] (red) = shutdown, 0 [False] (green) = temperature OK)

Variable Name	Display Type	Description
s1PortDTR	Boolean	Indicates serial 1 port DTR asserted. A console cannot be opened to the node's serial 1 port unless this flag is on. (1 [True] (green) = port asserted, 0 [False] (red) = port not asserted)
voltN12	Volts	Voltage on -12 volt line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltN12Range	Boolean	-12 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP4	Volts	Voltage on +4 volt supply line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP5i	Volts	Voltage on +5 volt I/O supply line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP5iRange	Boolean	+5 I/O voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP5m	Volts	Voltage on +5 volt memory supply line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP5mRange	Boolean	+5 memory voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP12	Volts	Voltage on +12 volt line (red = node shutdown due to voltage failure, yellow = warning, voltage is outside proper threshold, green = voltage OK)
voltP12Range	Boolean	+12 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
warningN12High	Boolean	Environmental warning: Node warning due to marginal specifications condition on -12 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningN12Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on -12 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP4High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +4 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP4Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +4 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP5iHigh	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt I/O supply line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP5iLow	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt I/O supply line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP5mHigh	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt memory supply line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)

Variable Name	Display Type	Description
warningP5mLow	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt memory supply line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP12High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +12 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP12Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +12 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningTemp	Boolean	Environmental warning: Temperature nearing out of specifications condition (1 [True] (red) = warning, 0 [False] (green) = temperature OK)
7segChanged	Multistate	Indicates the digits within the 3-digit display are changing faster than once per second. (1 [True] (black) digits changing faster than once per second, 0 [False] (yellow) digits changing less than once per second)

Node-Specific Hardware Variables For Card Type 161 Node Supervisor

Table 19 (Page 1 of 2). Node-Specific Hardware Variables For Card Type 161 Node Supervisor

Variable Name	Display Type	Description
boardLevel	Constant	Incremental number indicating level of hardware.
hardwareKeyID	Constant	A merging of “boardLevel” and “supLevel” constants to produce a Key ID that uniquely defines this level of hardware.
hardwareStatus	Hex	Hardware diagnostics byte as follows: bit 7: 1 = Busy condition on LCD bit 6: 1 = Bounces on power On/Off Switch bit 5: 1 = Bounces on reset button bit 4: 1 = Bounces on key switch bit 3: 1 = Wrong length command from BUMP bit 2: 1 = SIF timeout bit 1: 1 = BUMP timeout bit 0: 1 = Unknown command from BUMP
lcd1	String	16 bytes of ASCII data that is the first line of the LCD display.
lcd1flash	Boolean	Indicates line 1 of the LCD display is flashing. (1 [True] = flashing, [False] 0 = not flashing)
lcd2	String	16 bytes of ASCII data that is the second line of the LCD display.
lcd2flash	Boolean	Indicates line 2 of the LCD display is flashing. (1 [True] = flashing, [False] 0 = not flashing)
powerSwitchStatus	Boolean	Indicates power switch position and whether the node is capable of being powered on. (1 (on) = capable of being powered on, 0 (off) can not be powered on)
P5DCok	Boolean	Voltage on +5 volt line is good. (1 [True] (green) = voltage OK, 0 [False] (red) = voltage not on)
smpDiagLEDoff	Boolean	Environmental error: LED 5 (0 [False] = error condition, 1 [True] = no errors detected)
smpPowerLEDoff	Boolean	Indicates power status. LED 1 (0 [False] = DC power OK, 1 [True] = no DC power supplied)

Table 19 (Page 2 of 2). Node-Specific Hardware Variables For Card Type 161 Node Supervisor		
Variable Name	**Display Type**	**Description**
supLevel	Constant	Incremental number indicating level of application supervisor installed on board.
warningTemp	Boolean	Environmental warning: Temperature nearing out of specifications condition. (1 [True] (red) = warning, 0 [False] (green) = temperature OK)

Common Switch Variables

Variable Name	Display Type	Description
codeVersion	Constant	Supervisor microcode version number
diagByte	Integer	Diagnosis byte (correlates with error indicators). This variable is not currently being used.
envLED	Multistate	Reflects environment light on front of switch (2/3 [2] (gray) = LED blinking: power is shut off due to detection of a critical error (voltages, fans, temp., etc), 1 [1] (yellow) = LED lit: warning indicator-not yet critical, 0 [0] (black) = LED not lit: OK)
fanfail 1-5	Boolean	Environmental failure: Shutdown due to fan out of specifications condition (1 [True] (red) = shutdown, 0 [False] (green) = fan OK)
nodePower	powerSwitch	To control the power using the **spmon** command, use the **-power** parameter. The switch supervisor returns a value indicating whether switchPower is on or off. (1 [True] (green) = power switch on, 0 [False] (red) = power switch off)
nodePowerOn10Sec	Boolean	Indicates switch has been operational more than 10 seconds and remains so (1 [True] (green) = powered on, 0 [False] (black) = powered off)
powerLED	Multistate	Reflects power light on front of switch (2/3 [2] (gray) = LED blinking: power off but available, 1 [1] (green) = LED lit: power on, 0 [0] (black) = LED not lit: no power exists)
P48Off10Sec	Boolean	+48 volt input has been terminated due to error and remains off (1 [True] (red) = powered off, 0 [False] (black) = OK)
P48OK	Boolean	+48 volt input is OK (1 [True] (green) = OK, 0 [False] (red) = not OK)
serialNum	Constant	Supervisor serial number. This variable is not currently being used.
shutdownTemp	Boolean	Environmental failure: Shutdown due to temperature out of specifications (1 [True] (red) = shutdown, 0 [False] (green) = temperature OK)
temp	Celsius	Temperature (red = node shutdown due to temp failure, yellow = warning, temp is outside proper threshold, green = temp is OK)
tempRange	Boolean	Temperature out of range. (1 [True] = temp out of range, 0 [False] = temp within range)
timeTicks	Integer	Supervisor “heartbeat” counter (in numbers of 100 ms tics)
type	Constant	Switch Supervisor card type (in decimal)
warningTemp	Boolean	Environmental warning: Temperature nearing out of specifications condition (1 [True] (red) = warning, 0 [False] (green) = temperature OK)

High Performance Switch and High Performance Switch-LC8 Card Types 49 and 50

Variable Name	Display Type	Description
fanwarning 1-5	Boolean	Environmental warning: Fan nearing out of specifications condition (1 [True] (red) = warning, 0 [False] (green) = fan OK)
mux	Integer	Multiplexor setting to control the clock selection at each frame. Use the **-mux** parameter on the **spmon** command to select the proper input for the clock (3 [3] = input 3, 2[2] = input 2, 1[1] = input 1, 0[0] = internal clock)
shutdownN5High	Boolean	Environmental failure: Node shutdown due to -5 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownN5Low	Boolean	Environmental failure: Node shutdown due to -5 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5High	Boolean	Environmental failure: Node shutdown due to +5 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP5Low	Boolean	Environmental failure: Node shutdown due to +5 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP12High	Boolean	Environmental failure: Node shutdown due to +12 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP12Low	Boolean	Environmental failure: Node shutdown due to +12 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP48Low	Boolean	Environmental failure: Node shutdown due to +48 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
voltN5	Volts	Voltage on -5 volt line (red = node shutdown due to voltage failure, yellow = warning—voltage is outside proper threshold, green = voltage OK)
voltN5Range	Boolean	-5 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP5	Volts	Voltage on +5 volt line (red = node shutdown due to voltage failure, yellow = warning—voltage is outside proper threshold, green) = voltage OK)
voltP5Range	Boolean	+5 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP12	Volts	Voltage on +12 volt line (red = node shutdown due to voltage failure, yellow = warning—voltage is outside proper threshold, green) = voltage OK)
voltP12Range	Boolean	+12 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
voltP48	Volts	Voltage on +48 volt line (red = node shutdown due to voltage failure, yellow = warning—voltage is outside proper threshold, green) = voltage OK)

Variable Name	Display Type	Description
voltP48Range	Boolean	+48 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)
warningN5High	Boolean	Environmental warning: Node warning due to marginal specifications condition on -5 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningN5Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on -5 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP5High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP5Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +5 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP12High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +12 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP12Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +12 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)

SP Switch Variables for Card Type 129

Variable Name	Display Type	Description
boardLevel	Integer	Incremental number indicating level of hardware.
chip	Integer	Indicates which of the 8 chips the active clock is on when the value of the “clockSource” variable is set to 2 (STI). Value can be 0, 1, 2, 3, 4, 5, 6, or 7.
chipClkMissing	Boolean	Indicates whether the active chip, as defined by “chip,” is receiving its system clock from the current master clock source, as defined by “clockSource” ([True] = not receiving system clock, [False] = receiving system clock)
chipReceivedInit	Boolean	Indicates whether the active chip, as defined by “chip,” has received its initialization packet from the primary service node. The **Estart** command sets and distributes initialization data to chips. ([True] = received initialization data, [False] = did not receive initialization data)
chipSelftest	Boolean	Indicates whether the active chip, as defined by “chip,” is currently in selftest mode due to a power-on reset, as defined by “powerOnReset” ()True[= in selftest mode, [False] = not in selftest mode)
clockSource	Integer	Indicates whether the active clock source is Local (internal) or STI (external). (0 = Local, 2 = STI)
epromErase	Boolean	Indicates that FEEPROM was erased. This erasure is typically the result of an supervisor application microcode download ([True] = currently erased, [FALSE] = a supervisor application is loaded)
epromProgram	Boolean	Indicates whether FEEPROM was ever programmed with a supervisor application ([True] = programmed, [False] = not programmed)

Variable Name	Display Type	Description
mux	Integer	Indicates which of the 32 jacks the active clock is on as a function of chip and port when the value of the “clockSource” variable is set to 2 (STI). Value can be 2 through 34 per the following matrix: Port 0 1 2 3 Chip 0 34 33 32 31 1 10 9 8 7 2 26 25 24 23 3 18 17 16 15 4 22 21 20 19 5 14 13 12 11 6 30 29 28 27 7 6 5 4 3 Or, if the value is zero, indicates that the active clock is Local. In this case, the “clockSource” variable will be set to Local.
osc	Integer	Indicates which of the 2 Local (internal) oscillators are being used when the value of the “clockSource” variable is set to 0 (Local). Value can be 2 or 4.
pll	Integer	Indicates which of the four Phase Lock Loops are being used. Value can be 2, 3, 4, or 5.
port	Integer	Indicates which of the 4 ports the active clock is on when the value of the “clockSource” variable is set to 2 (STI). Value can be 0, 1, 2, or 3.
portClkMissing	Boolean	Indicates whether the active port, as defined by “port,” is receiving a valid STI clock, as defined by “clockSource” ([True] = not receiving STI clock, [False] = receiving STI clock)
powerOnReset	Boolean	Indicates power-on reset state of board ([True] = in state, [False] = not in state)
powerS1	Integer	Power supply #1 key ID.
powerS2	Integer	Power supply #2 key ID.
psParallel	Volts	Parallel power voltage.
psParallelFail	Boolean	Power detector failure: This variable indicates whether the ability to detect power supply failures is operational (0 [false] (green) = OK, 1 [true] (red) = not OK) If the value is 1 and either Power Supply #1 or Power Supply #2 failed, “ps1Fail” or “ps2Fail” could not be set to indicate the failure.
psParallelRange	Boolean	Parallel power voltage out of range ([True] = out of range, [False] = not out of range)
ps1Fail	Boolean	Power supply failure: Power Supply #1 failed and should be replaced (0 [false] (green) = OK, 1 [true] (red) = not OK)
ps1FuseGood	Volts	Fuse supply #1 voltage.
ps1FuseGoodRange	Boolean	Fuse supply #1 voltage out of range ([True] = out of range, [False] = not out of range)
ps1PowerGood	Volts	Power supply #1 voltage.
ps1PowerGoodRange	Boolean	Power supply #1 voltage out of range ([True] = out of range, [False] = not out of range)
ps2Fail	Boolean	Power supply failure: Power Supply #2 failed and should be replaced (0 [false] (green) = OK, 1 [true] (red) = not OK)

Variable Name	Display Type	Description
ps2FuseGood	Volts	Fuse supply #2 voltage.
ps2FuseGoodRange	Boolean	Fuse supply #2 voltage out of range ([True] = out of range, [False] = not out of range)
ps2PowerGood	Volts	Power supply #2 voltage.
ps2PowerGoodRange	Boolean	Power supply #2 voltage out of range ([True] = out of range, [False] = not out of range)
recPortNotTuned	Boolean	Indicates whether a port's receive port is communicating with it ([True] = receive port not communicating with active port, as defined by “port,”, [False] = receive port is communicating with active port, as defined by “port”)
sendPortNotTuned	Boolean	Indicates whether the active port, as defined by “port,” is communicating with its corresponding receive port ([True] = active port not communicating with receive port, [False] = active port is communicating with its receive port)
shutdownOC	Boolean	Threshold shutdown: Power Supply #1 or Power Supply #2 has exceeded maximum over current threshold. Power supply should be replaced (0 [false] (green) = OK, 1 [true] (red) = not OK)
shutdownP3_3High	Boolean	Environmental failure: Node shutdown due to +3.3 volt line out of specifications (high) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
shutdownP3_3Low	Boolean	Environmental failure: Node shutdown due to +3.3 volt line out of specifications (low) (1 [True] (red) = shutdown, 0 [False] (green) = voltage OK)
subtype	Boolean	Indicates the subtype of the SP Switch board. ([True] = SP Switch (16 port), [False] = SP Switch-8 (8 port).
supLevel	Integer	Incremental number indicating level of application Supervisor installed on board.
synchReset	Boolean	Indicates synchronous reset state of board ([True] = in state, [False] = not in state)
warningOC	Boolean	Threshold warning: Either Power Supply #1 or Power Supply #2 is approaching maximum over current threshold (0 [false] (green) = OK, 1 [true] (red) = not OK)
warningP3_3High	Boolean	Environmental warning: Node warning due to marginal specifications condition on +3.3 volt line (high) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
warningP3_3Low	Boolean	Environmental warning: Node warning due to marginal specifications condition on +3.3 volt line (low) (1 [True] (red) = warning, 0 [False] (green) = voltage OK)
voltP3_3	Volts	Voltage on +3.3 volt line (red = node shutdown due to voltage failure, yellow = warning—voltage is outside proper threshold, green = voltage OK)
voltP3_3Range	Boolean	+3.3 voltage out of range (1 [True] = voltage out of range, 0 [False] = voltage within range)

Basecode Mode Hardware Variables for the SP Switch Card Type 128, SEPBU Frame Card Type 16 and 604 High Node Type 16

Variable Name	Display Type	Description
applicationLevel	Integer	Is the level of application Supervisor loaded in FEEPROM.
checksumError	Boolean	Indicates whether the application Supervisor download process detected a data transmission checksum error ([True] = checksum error, [False] = no checksum error)
feepromEmpty	Boolean	Indicates whether the programmable FEEPROM is empty (not programmed) ([True] = not programmed, [False] = programmed)
feepromEraseCount	Integer	Incremental value that indicates the number of times the programmable FEEPROM was erased
hardwareKeyID	Hex	Is the value of the Hardware Key taken from the hardware (card) itself. The value is an incremental version number.
keyidMismatch	Boolean	Indicates whether a compatible supervisor application was downloaded to the card based on a match of the “softwareKeyID” and “hardwareKeyID” ([True] = mismatch, [False] = no mismatch)
softwareKeyID	Hex	Is the value of the Software Key taken from the supervisor application microcode file during download to the card. The value is an incremental version number.
supDataChecksum	Hex	Checksum for the application Supervisor loaded in FEEPROM.
supFailure	Boolean	Indicates that the application Supervisor loaded into FEEPROM cannot be executed due to an error generated during download or because FEEPROM was not programmed ([True] = not executable, [False] = executable) supFailure will be True on any of the following conditions occur: • keyidMismatch = True • checksumError = True • feepromEmpty = True • supDataChecksum != supRomChecksum
supRomChecksum	Hex	Saved supDataChecksum. If checksumError is False, this will match supDataChecksum.

Appendix E. Listing SP System Resource Variables

As described in Chapter 24, "The Event Management Subsystem" on page 365, PSSP supplies a number of resource monitors that collect resource data and send it to the Event Management subsystem. You may also install resource monitors that:

- Are supplied by other IBM products or third parties.
- Have been written by programmers at your installation.

Once all of these resource monitors are running, you may wish to determine what resource data is being collected in the system.

There are several ways to do this:

- You can look in the load list files that contain the Event Management configuration data for each of the resource monitors in your system.

  These files contain the data that is loaded into the SDR as Event Management objects by the **haemloadcfg** command.

- You can use commands.

  You can use the PSSP **SDRGetObjects** command to list Event Management objects that have been stored in the SDR. You can use the AIX **dspmsg** and **dspcat** commands to list individual messages and the contents of message catalogs that contain descriptions of resource variables.

- You can use the SP Perspectives graphical user interface.

  You can use the Event Management portion of SP Perspectives to list information about the resource variables and instance vectors that have been defined in the system.

Note that there are no Event Management resource variables for dependent nodes.

Looking at the Event Management Load List Files

The load list file for the resource variables supplied by PSSP is the **/usr/lpp/ssp/install/config/haemloadlist** file. This file is a plain text file that consists of stanzas, each of which describes an object in one of the Event Management SDR classes. For detailed information about the format of a load list file, see the man page for the **haemloadlist** file in *PSSP Command and Technical Reference*.

If you have installed other resource monitors, look at the load list files that contain their configuration data.

Getting Information about Event Management Objects in the SDR

To get information about all of the event management objects in the SDR, you can issue the **SDRGetObjects** command for each class of objects. You can also direct the output to a file, which you can then view or print. To do this, enter the following commands:

© Copyright IBM Corp. 1995, 1997

```
SDRGetObjects -G EM_Resource_Variable > rv_data
SDRGetObjects -G EM_Resource_Class > class_data
SDRGetObjects -G EM_Structured_Byte_String > sbs_data
SDRGetObjects -G EM_Instance_Vector > iv_data
SDRGetObjects -G EM_Resource_Monitor > rm_data
```

Specifying the **-G** flag ensures that you receive information about all of the resource variables in all of the system partitions in the SP system. If you are only interested in the resource variables in the current system partition, omit the **-G** flag.

Figure 69 on page 571 contains an example of the output of each **SDRGetObjects** command for the **IBM.PSSP.SP_HW.Node.lcd1** resource variable and its associated resource class, structured byte string, instance vector, and resource monitor definitions. For ease of presentation, the output has been reformatted vertically. The actual output of the command is horizontally formatted.

For a complete description of the attributes of Event Management SDR classes, see *IBM Parallel System Support Programs for AIX: Event Management Programming Guide and Reference*.

You can also restrict the information you receive. For example, to get only the **rvName** and **rvClass** information for the **IBM.PSSP.SP_HW.Node.lcd1** resource variable, enter:

```
SDRGetObjects -G EM_Resource_Variable \
    rvName==IBM.PSSP.SP_HW.Node.lcd1 rvName rvClass
```

In response, the system displays:

```
rvName       rvClass
IBM.PSSP.SP_HW.Node.lcd1          IBM.PSSP.SP_HW
```

For more information about the **SDRGetObjects** command, see *PSSP Command and Technical Reference*.

```
Information from EM_Resource_Variable class for Resource Variable IBM.PSSP.SP_HW.Node.lcd1

    rvName                  IBM.PSSP.SP_HW.Node.lcd1
    rvDescription           244
    rvValue_type            State
    rvData_type             SBS
    rvInitial_value         0
    rvClass                 IBM.PSSP.SP_HW
    rvPTX_name              ""
    rvPTX_description       ""
    rvPTX_min               ""
    rvPTX_max               ""
    rvPredicate             ""
    rvEvent_description     ""
    rvLocator               ""
    rvOrder_group           ""
    rvDynamic_instance      0
    rvIndex_vector          NodeNum

Information from EM_Resource_Class class for Resource Class IBM.PSSP.SP_HW

    rcClass                 IBM.PSSP.SP_HW
    rcResource_monitor      IBM.PSSP.hmrmd
    rcObservation_interval  0
    rcReporting_interval    0

Information from EM_Structured_Byte_String class for Resource Variable IBM.PSSP.SP_HW.Node.lcd1

    sbsVariable_name        IBM.PSSP.SP_HW.Node.lcd1
    sbsField_name           STRING
    sbsField_type           cstring
    sbsField_SN             0
    sbsField_init_val       ""

Information from EM_Instance_Vector class for Resource Name IBM.PSSP.SP_HW.Node

    ivResource_name         IBM.PSSP.SP_HW.Node
    ivElement_name          NodeNum
    ivElement_description   256

Information from EM_Resource_Monitor class for Resource Monitor IBM.PSSP.hmrmd

    rmName                  IBM.PSSP.hmrmd
    rmPath                  /usr/lpp/ssp/bin/haemRM/hmrmd
    rmArguments             IBM.PSSP.hmrmd
    rmMessage_file          hmrmd_des.cat
    rmMessage_set           1
    rmConnect_type          server
    rmPTX_prefix            0
    rmPTX_description       0
    rmPTX_asnno             0
```

Figure 69. Sample Output from PSSP SDRGetObjects commands

Displaying Event Management Descriptions

As shown in Figure 69 on page 571, the description of resource variable **IBM.PSSP.SP_HW.Node.lcd1** is message number 244 of message set 1 of the **hmrmd_des.cat** message catalog.

To display the description, enter:

```
dspmsg -s 1 hmrmd_des.cat 244
```

In response, the system displays:

```
"(string) lcd1: 'LCD line 1'."
```

To list all of the messages in the catalog and direct the output to a file, enter:

```
dspcat hmrmd_des.cat > hmrmdcat
```

For more information on the AIX **dspmsg** and **dspcat** commands, see *AIX Version 4.1 Commands Reference*.

Getting Resource Variable Information from SP Perspectives

You can use the Event Management portion of SP Perspectives to get information about the resource variables and instance vectors that have been defined in the system. To do this:

- Start the Event Management portion of SP Perspectives.

  Enter:

  ```
  spevent
  ```

  If you have more than one system partition defined, click on the icon that represents the system partition in which you are interested.

- Open the Create Condition dialog box.

  From the Menu bar of the Event Perspective, click on the Actions button and, from the cascading menus, select Event Definitions and then Create. In response, the system displays a notebook that is open to the Create Event Definition page.

  On the Create Event Definition page, click on the Create Condition button. In response, the system displays the Create Condition dialog box.

- Click the down arrow next to the Resource Variable Name field. In response, the system displays a list of all of the resource variables that are defined in the system partition you selected.

  To get information about any particular resource variable, select it. In response, the system displays the resource variable's name in the Resource Variable Name field, its description in the Resource Variable Description field, and the format of the instance vector in the Resource Identifier Format field.

  For an example of the Create Condition dialog box after the **IBM.PSSP.Prog.pcount** resource variable has been selected, see Figure 70 on page 573.

Note that SP Perspectives uses slightly different terminology than the Event Management APIs, as follows:

- Where the Event Management APIs use the term "predicate," SP Perspectives uses "expression."

- Where the Event Management APIs use the term "instance vector," SP Perspectives uses "resource identifier."

For complete information on using SP Perspectives, see the *PSSP Administration Guide* and the SP Perspectives help panels.

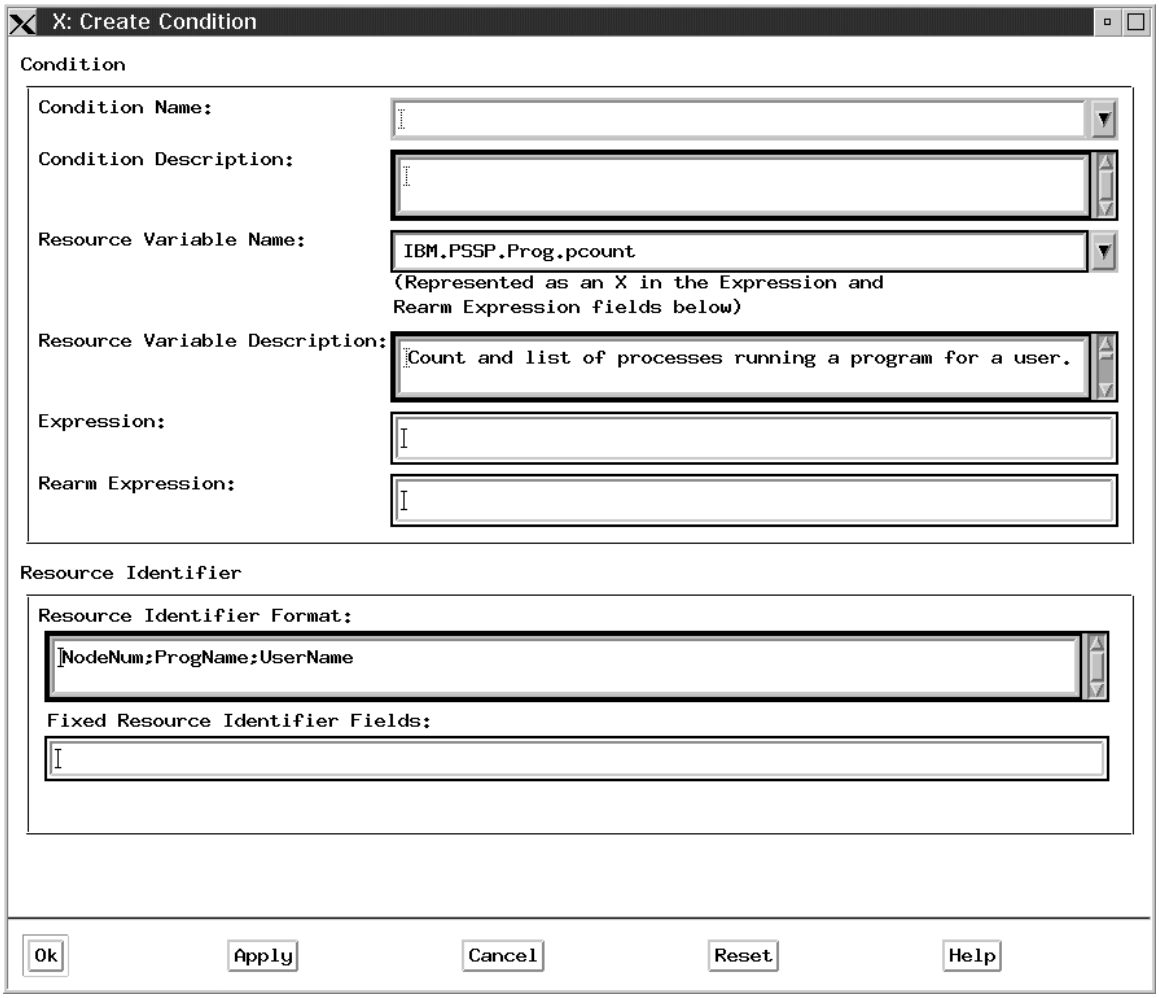


Figure 70. Using the Event Perspective to Get Resource Variable Information

Appendix F. The System Data Repository

The System Data Repository (SDR) is where your SP System configuration is stored. Information entered through the SMIT panels during installation goes into the SDR.

Understanding the System Data Repository

The SDR is composed of objects and their attributes. Objects differ according to class. An object belonging to, say, the node class, has a different set of attributes from an object in the adapter class.

Objects are *instances* of the object class. Some classes have many more objects than others. For example, there is only one object in the SP class for each SP system, but there are as many objects in the node class as there are nodes in the system.

An SDR object, however, is really only a specific set of attributes. Object attributes contain the meaningful data. To illustrate, Figure 71 compares two object classes in a sample configuration, the Frame class and the Adapter class.

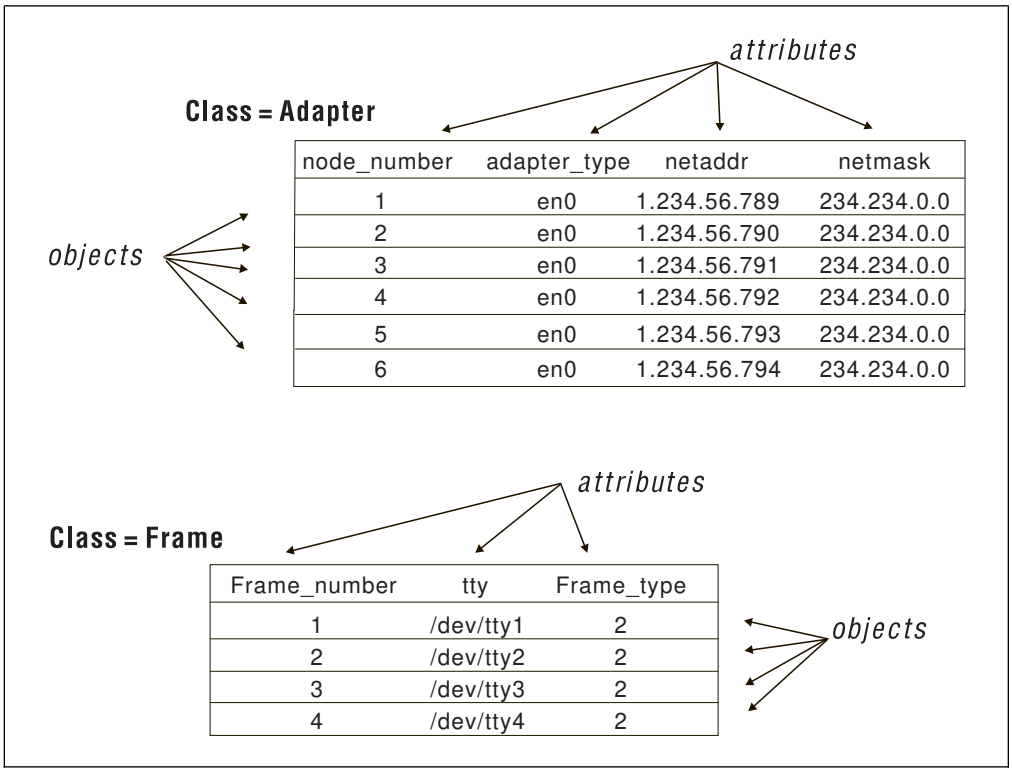


Figure 71. Comparing Two SDR Classes

The SP system provides *system partitioning* as a method for organizing the system into groups of nodes for various purposes such as testing new software and creating multiple production environments. Therefore there are two types of SDR classes: system and partitioned.

- System classes (contain objects that are common to all system partitions)

© Copyright IBM Corp. 1995, 1997

  - Class = Frame
  - Class = SP
  - Class = SP_ports
  - Class = Switch
  - Class = SysNodeGroup
  - Class = Syspar_map

- Partitioned classes (contain objects that are specific to a system partition. The *definition* of a Partitioned class is common to all system partitions, but each system partition contains different objects for the class.)

  - Class = Node
  - Class = DependentNode
  - Class = Pool
  - Class = Adapter
  - Class = DependentAdapter
  - Class = host_responds
  - Class = switch_responds
  - Class = Dont_care_pool_list
  - Class = JM_server_nodes
  - Class = JM_domain_info
  - Class = JM_job_info
  - Class = JM_node_usage
  - Class = Switch_partition
  - Class = VSD_Global_Volume_Group
  - Class = HSD_Table
  - Class = VSD_Table
  - Class = VSD_Minor_Number
  - Class = HSD_Minor_Number
  - Class = VSD_Fence
  - Class = GMT_Global_smt_nds
  - Class = Syspar
  - Class = pmandConfig
  - Class = pmanrmdConfig
  - Class = NodeGroup
  - Class = TS_Config
  - Class = Syspar_ports
  - Class = EM_Condition

For a complete list of attributes for each object, see "System Data Repository Classes and Attributes" on page 585.

In addition, there is a set of partitioned classes that are used by Event Management application programs:

- Class = EM_Resource_Variable
- Class = EM_Structured_Byte_String
- Class = EM_Instantiation_Vector
- Class = EM_Resource_Class
- Class = EM_Resource_Monitor

For more information about these classes and their attributes, see *IBM Parallel System Support Programs for AIX: Event Management Programming Guide and Reference*.

Accessing the System Data Repository

There are two methods you can use to access the data in the SDR: SMIT panels and the SDR command line interface. The SDR command line interface is used by various SP commands. There should be no need to use the SDR command line directly and you should not do so unless you are instructed to by an IBM representative. In addition, many system management commands, such as **spbootins**, manipulate SDR data. You must be authorized in order to access the SDR. The SDR also performs authentication before allowing read-write access.

Authorization

The SDR has two levels of authorization, read-write and read-only. Table 20 shows the conditions under which each level of authorization is given.

Table 20. SDR Authorizations

Machine	Root User	Non-Root User
Control Workstation	read-write	read-only
SP Node	read-write	read-only
Other	read-only	read-only

Command entries made without necessary authorization will fail.

Authentication

Read-write authorization is given after authentication. This consists of checking the user ID. If a root user is running on the same machine as the SDR (the control workstation), then read-write authority is given. For a root user who is not on the control workstation, the SDR server checks that the adapter connecting the client machine has an adapter class object in the SDR. If so, the user is given read-write authority. If not, the user gets read-only authority.

All nonroot users get read-only authority.

Locating the SDR Server

There are three ways in which a process can locate the SDR server. They are selected in the following order of preference but when one method is selected, no other methods are attempted. This means that if method 1 is available, only method 1 is attempted. If it fails, methods 2 and 3 are *not* attempted.

1. The destination is passed as the third parameter in the **SDROpenSession** command. This only works for library routines and therefore may be used by some SP subsystems, such as the Resource Manager. The destination is the hostname or TCP/IP address of the control workstation where the SDR runs.
2. The **SP_NAME** environment variable is set to the hostname or TCP/IP address of the control workstation where the SDR runs.
3. The **/etc/SDR_dest_info** file is present and has the **primary** record set to the hostname or TCP/IP address of the control workstation where the SDR runs. This file is installed on the control workstation and all SP System nodes along with the IBM Parallel System Support Programs for AIX.

The SDR_dest_info File

The **/etc/SDR_dest_info** file is created on the control workstation at system installation, and propagated to all nodes in the SP system. The **/etc/SDR_dest_info** file has the following format:

```
* comments have an asterisk in column 1
default: <default system partition>
primary: <TCP/IP address of control workstation>
```

The **default** record identifies the default system partition. The default record is used at boot time so that the node can determine if it has changed system partitions.

Only the **primary** record is used by the SDR to locate the control workstation where the SDR server runs.

The SDR Daemon Log

The SDR daemon writes information to a log named **/var/adm/SPlogs/sdr/sdrdlog.***syspar_ip_addr.pid*, where *syspar_ip_addr* is the IP address of the system partition and *pid* is the process ID of the SDR daemon (**sdrd** process). This log will contain the date and time the process started, as well as problems encountered by the daemon in the course of operation.

Method 1 — Using SMIT

The RS/6000 SP System Management SMIT panel provides options for accessing configuration data in the SDR. To invoke this panel:

TYPE **smit**

▲ The System Management Menu appears.

SELECT **RS/6000 SP System Management**

▲ The RS/6000 SP System Management menu appears.

The RS/6000 SP System Management menu offers the following options:

1. RS/6000 SP Configuration Database Management

   The dialogs available through this path allow you to enter, list, and change information during the installation process about your nodes, primary and secondary external LANs, and switch connections, as well as site environment information. These tasks are explained in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

   The choices from this menu are:

   - *Enter Database Information*
     - *Site Environment Information*
     - *Frame Information*
     - *Node Database Information*
     - *Node Group Information*
     - *System Partition Configuration*
     - *Extension Node Database Information*
     - *Run setup_server Command*
   - *List Database Information*

- *List Site Environment Database Information*
- *List Frame Database Information*
- *List Node Database Information*
- *List Node Group Database Information*
- *List System Partition Database Information*
- *List LAN Database Information*
- *List Extension Node Database Information*
- *List Extension Node Adapters Database Information*

- *Delete Database Information*
  - *Delete Frame Information*
  - *Delete Node Information*
  - *Delete Adapter Information*
  - *Delete Node Group Information*
  - *Delete Extension Node Information*
  - *Delete Extension Node Adapter Information*

2. RS/6000 SP Cluster Management

   The choices from this menu are:

   - *Run setup_server Command*
   - *Select System Tuning Parameters*
   - *Perform Switch Operations*
   - *Run enadmin Command*

3. RS/6000 SP Configuration Information

   Using this path, you can display configuration information about your nodes, networks, file systems, and paging spaces.

   The choices from this menu are:

   - *List Node Hardware Information*
   - *List Node Network Information*
   - *List Node File System Information*

4. RS/6000 SP Users

   Using this path, you can add and delete users, as well as change user management attributes such as passwords.

   The choices from this menu are:

   - *Add a User*
   - *Change/Show Characteristics of a User*
   - *Remove a User*

5. RS/6000 SP Installation/Configuration Verification

   Use this path to check that your IBM Parallel System Support Programs for AIX options are installed correctly.

   The choices are:

   - *System Monitor Installation*
   - *System Monitor Configuration*
   - *System Data Repository*
   - *System Management*
   - *Communication Subsystem*
   - *System Partition Configuration*
   - *Switch Table API Installation*

- *Resource Manager Installation*
- *Resource Manager Configuration*

6. RS/6000 SP Supervisor Manager

   The choices are:

   - *Check for Supervisors That Require Action (Single Message Issued)*
   - *List Status of Supervisors (Report Form)*
   - *List Status of Supervisors (Matrix Form)*
   - *List Supervisors That Require Action (Report Form)*
   - *List Supervisors That Require Action (Matrix Form)*
   - *Update *ALL* Supervisors That Require Action (Use Most Current Level)*
   - *Update Selectable Supervisors That Require Action (Use Most Current Level)*

7. RS/6000 SP Resource Manager

   The choices are:

   - *Change/Show Configuration Data*
   - *Start the Resource Manager*
   - *Reconfigure the Resource Manager*
   - *Stop the Resource Manager*

8. RS/6000 SP Log Management

   The choices are:

   - *AIX Error Log*
   - *Syslog*
   - *General Log Viewing*
   - *Archive Logs*
   - *Collect Logs for Service*

The menu selections run standard AIX commands and, in some cases, add information from the Hardware Monitor and reformat the output for useability. Let's examine the options on the list menu in more detail.Figure 72 on page 581 contains options that invoke the AIX commands listed in Table 21 on page 581.

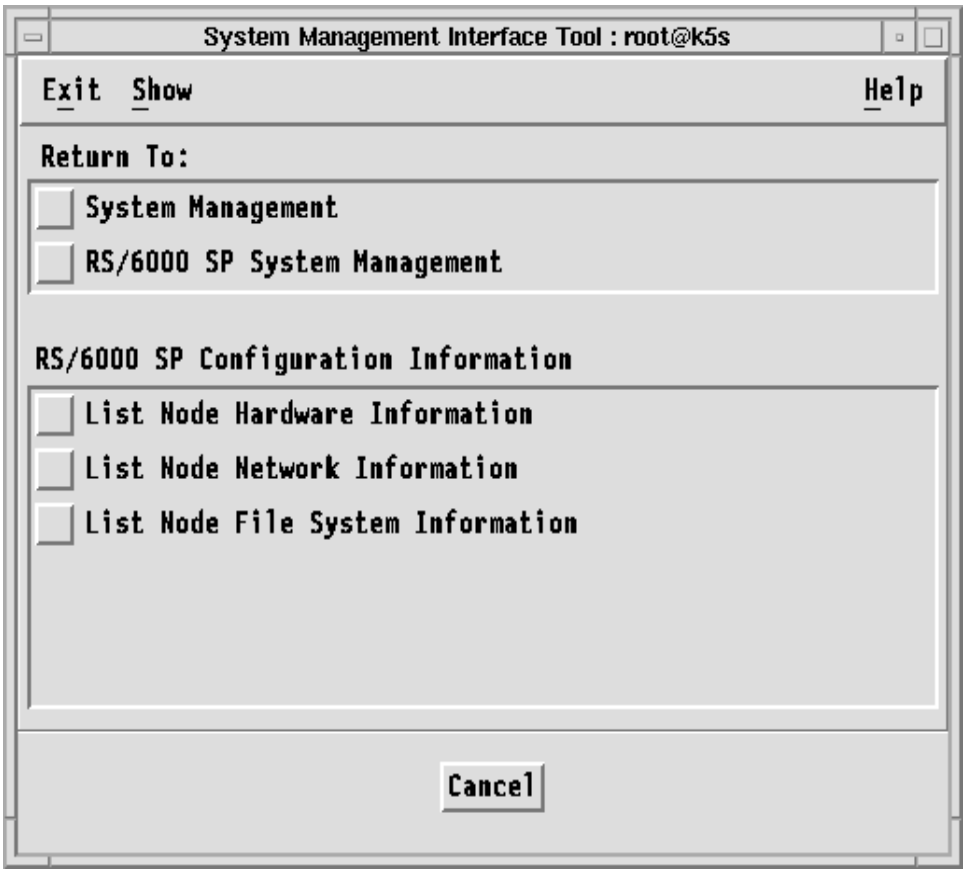


Figure 72. List RS/6000 SP Configuration Information

Table 21. Commands Invoked by SMIT Panels

Select:	To:
List Node Hardware Information	Run the AIX **lscfg** command to display the name, location, and description of the devices related to the nodes
List Node Network Information	Run the AIX **netstat -in** command to show the state of the nodes' configured interfaces
List Node File System Information	Run the AIX **df** command to display the total spaces and available space on the node file systems

SELECT Any of the three choices

▲ SMIT runs the corresponding AIX command for all the nodes and displays the information in a scrollable window.

You can also list the configuration data using the **splstdata** command. Refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for complete syntax and examples.

Method 2 — - Using the SDR Command Line Interface

A command line interface allows you to display, change, or delete the contents of an SDR object without invoking SMIT panels.

These commands are used by the IBM Parallel System Support Programs for AIX to operate on SDR data. You should not need to use these commands directly. Should you choose to use them, do so with caution. SDR contents can be corrupted or made inaccessible.

The following list briefly describes these commands. See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for exact syntax.

SDRAddSyspar
The IBM Parallel System Support Programs for AIX use this command to create a new daemon using the System Resource Controller (SRC).

SDRArchive
The IBM Parallel System Support Programs for AIX use this command to create an archive file containing all current SDR classes attributes.

SDRChangeAttrValues
The IBM Parallel System Support Programs for AIX use this command to change the attribute values of an existing object.

SDRClearLock
The IBM Parallel System Support Programs for AIX use this command to unlock a class that is locked, regardless of who has the lock. This is for system administration use only and should be used with caution.

SDRCreateAttrs
The IBM Parallel System Support Programs for AIX use this command to create new attributes for an SDR class.

SDRCreateClass
The IBM Parallel System Support Programs for AIX use this command to create a new class of objects and its attributes.

SDRCreateFile
The IBM Parallel System Support Programs for AIX use this command to create an SDR file from an AIX file.

SDRCreateObjects
The IBM Parallel System Support Programs for AIX use this command to create one or more new objects and define their attribute values.

SDRCreateSystemClass
The IBM Parallel System Support Programs for AIX use this command to create a system class.

SDRCreateSystemFile
The IBM Parallel System Support Programs for AIX use this command to create a file that can be retrieved from any system partition.

SDRDeleteFile
The IBM Parallel System Support Programs for AIX use this command to delete an SDR file.

SDRDeleteObjects
The IBM Parallel System Support Programs for AIX use this command to delete target objects.

SDRGetObjects
The IBM Parallel System Support Programs for AIX use this command to query the values of target objects and attributes and prints them to stdout.

SDRListClasses
The IBM Parallel System Support Programs for AIX use this command to list the class names in the SDR.

SDRListFiles
The IBM Parallel System Support Programs for AIX use this command to first list all the files in the system area, then list all the files in the system partition area.

SDRMoveObjects
The IBM Parallel System Support Programs for AIX use this command to move objects from one system partition to another.

SDRRemoveSyspar
The IBM Parallel System Support Programs for AIX use this command to remove the entire contents of the subdirectory under system partitions. It uses the SRC to remove the daemon that serves the system partition.

SDRReplaceFile
The IBM Parallel System Support Programs for AIX use this command to replace the specified SDR file with the specified AIX file.

SDRRestore
The IBM Parallel System Support Programs for AIX use this command to overwrite the current SDR with the contents of an archived SDR file.

SDRRetrieveFile
The IBM Parallel System Support Programs for AIX use this command to create an AIX file from an SDR file.

SDR_test
The IBM Parallel System Support Programs for AIX use this command to verify that the installation and configuration of the SDR completed successfully.

SDRWhoHasLock
The IBM Parallel System Support Programs for AIX use this command to query the lock transaction ID for a specified object class.

Updating the host_responds Class (the hrd Daemon)

The **host_responds** class in the SDR is updated automatically by the **hrd** daemon. The **hrd** daemon is called from the **hr** script, which is under SRC control. **hr** will respawn if it, or the **hrd** daemon, is killed. **hr** and **hrd** run on the control workstation. Both the **hr** script and the **hrd** daemon are in directory **/usr/lpp/ssp/bin**. (The **hrctrl** script provides the same function as the **hr** script, but it follows the syntax of the **syspar_ctrl** command.)

The **hrd** daemon monitors the SP nodes in one of two ways. The default method is to use the system heartbeat (**hbd** daemon). The **hrd** daemon communicates with the **hbd** daemon on the control workstation, getting updates about the status of the nodes. The **hbd** runs on the control workstation and all of the SP nodes.

The second method by which **hrd** can get node status is to use **fping**. **fping** is a program that runs **ping** to many nodes asyncronously. This second method also issues **snmpinfo** calls to each node, to detect a situation where **ping** succeeds but the node is not responding to user requests.

The **hr** script determines which method will be used. An environment variable named `HR_FPING` is set to **0** for heartbeat or **1** for **fping**. The default is **0**.

The line that chooses between the methods looks like this:

```
typeset -x HR_FPING=0
```

After making a change to this file, you should reset the **hrd** daemon by issuing **hr reset**. This will cause it to respawn and the change will be put in effect. Other options for configuring **hrd** are documented in comments in the **hr** script (in **/usr/lpp/ssp/bin**).

The reason for the second method (**fping**) is that it has different characteristics than the system heartbeat. It does not require a daemon to be on each node, so it may reduce cpu usage on the nodes by a small amount.

Backing Up the SDR

Backing up the SDR regularly is a good way to insure against the loss of your data. Whenever you change your SP system configuration, use the **SDRArchive** command to take a snapshot of the SDR contents and save it in a **tar** file. The *IBM Parallel System Support Programs for AIX: Command and Technical Reference* contains syntax and usage information.

You can restore a saved image of the SDR using the **SDRRestore** command. Before restoring the SDR, be sure to stop the Resource Manager. (Refer to Chapter 7, “Managing SP Resources” on page 83.)

SDR Shadow Files

The SDR classes are stored in subdirectories under directory **/spdata/sys1/sdr**. The subdirectories of interest are **/system/classes** and **/partitions/***partition_ip_address*/**classes**. The names of the files in those directories correspond to the names of the classes.

Before SDR commits changes to a class, the class contents are moved to a backup file in the same directory as the class. The backup file is named *class.***shadow**, where *class* is the name of the class being written. If a power loss occurs or the SDR daemon is killed while the SDR class is being written, the class file that was being written at the time may not exist or may be corrupted. You should check for the existence of a *class*.**shadow** file. If one exists, the following steps should be taken to restore the class.

1. The corrupted *class* file (if one exists) should be removed.
2. The *class*.**shadow** file should be renamed to *class*.
3. The SDR daemon (**sdrd**) should be restarted. This can be accomplished by issuing **sdr -g sdr reset**.

The SDR daemon will restart automatically and recognize the renamed shadow file as the class.

System Data Repository Classes and Attributes

The tables below show all the SDR classes. You can specify these classes as targets in **SDRxxxxx** commands. The value in the **Type** column is the datatype of the attribute: S=string, I=integer, F=floating point.

Class = Node

Attribute Name	*Type*	*Description*	*Comments*
node_number	I	Relative node number ((frame#-1)*16)+slot	This number is computed from frame info and put here by system management.
hdw_enet_addr	I	Universal hardware ethernet identifier for this node.	Obtained via system monitor during system initialization.
frame_number	I	Frame number	Number from 1 to number of frames. Set during object creation.
slot_number	I	Node slot in frame	Number from 1 to number of nodes. Set during object creation.
slots_used	I	Nodes use 1, 2 or 4 slots. Distinguishes thin, wide or 604 High Node nodes.	Set by data obtained from system monitor.
switch_node_number	I	Switch relative node number.	Set during object creation from data obtained from system monitor.
switch_chip_port	I	Switch port this node is connected to.	Set during object creation from data obtained from system monitor.
switch_chip	I	Switch chip number this node is connected to.	Set during object creation from data obtained from system monitor.
switch_number	I	Switch number this node is connected to.	Set during object creation from data obtained from system monitor.
initial_hostname	S	Hostname assigned to this node during customization.	Customer supplied from CMI at installation.

Attribute Name	*Type*	*Description*	*Comments*
reliable_hostname	S	Hostname associated with SP ethernet.	Set during CMI processing from name entered for en0 interface name.
default_route	S	IP address of default route.	Customer supplied from CMI at installation.
boot_server	I	Relative node number of node that will answer bootp request when necessary. (0 represent control workstation.)	Set during creation to be node 1 of frame. For node 1 in frame set to 0 to represent control workstation.
bootdisk	S		
install_image	S	Path of net install image for this node. This field is initially set to **DEFAULT**, meaning that the default image is used.	Optionally setable by customer from CMI interface.
install_disk	S	The hard disks on which the install image is installed	
last_install_time	S	Time and date that this system was last successfully installed	Set during network installation.
last_install_image	S	server:/filepath of last image installed	Set during network installation.
switch_protocol	S	This is the string **LSP** or **CSSCI**	
switch_partition_number	I	Switch partition this node belongs to	Default is **1**
bootp_response	S	**install**=full net install and customize; **disk**=use local disk to boot (ignore bootp request); **maintenance**=on reboot, go into Installation/Maintenance menus; **customize**=just use information from SDR to customize node's ODM; **restore**=restore of backup image via network install	Initially set to **customize** during object creation. Changeable through CMI interface.
boot_device			
usr_maint	S	If set to **true** a node will perform usr client maintenance on boot	Initialized to **false**, set through CMI

Attribute Name	*Type*	*Description*	*Comments*
JM_domain_id	S	Identifier for a JM domain	
VSD_adapter	S	The adapter on this node used by VSD	This is an adapter type to be joined with the adapter_type attribute of the Adapter class, where the node and Adapter node_number fields match.
VSD_max_buffer_count	I	Buffer cache size for vsd	
VSD_request_blocks	I	The number of request blocks the VSD should allocate	
VSD_pbufs	I	The number of pbufs the VSD should allocate	
VSD_init_buffer_count	I	Buffer cache size for VSD	
VSD_min_buddy_buffer_size	I	The minimum VSD buddy buffer size allocated to a single request	
VSD_max_buddy_buffer_size	I	The maximum VSD buddy buffer size allocated to a single request	
VSD_max_buddy_buffers	I	The size of the VSD buddy buffer, given in the number of buffers of max_buddy size.	
cfs_adapter	S	The adapter on this node used by Calypso and Global Mount	
acct_class_id	S	An ID selected by administrator for grouping and merging acct data.	Default = **default**
acct_enable	S	Whether acct is enabled for a node	Default = **default**
acct_job_charge	S	Used to determine the number of charge fee units to charge a user for exclusive use of the node. Its value is in units of "seconds per charge fee unit"	Default = **1.0**

Attribute Name	*Type*	*Description*	*Comments*
acct_excluse_enable	S	Indicates whether stop and end job accounting records will be generated for jobs having exclusive use of a node.	Default = **false**
usr_server_ip	S	This is either an IP address or **local**	Default is **local**
usr_client_adapter	S	This will be **en0**, **en1**, **tr0** or **tr1**	Default is **en0**
has_usr_clients	S	This will be either true or false	Default is false
code_version	S	Contains the ssp release that is installed on this node.	Default is nothing (unknown)
usr_gateway_ip	S	The IP address of the gateway to get to /**usr** server	Default is **0**, meaning /**usr** server is not on a gateway.
VSD_maxIPmsgsz	I	Largest sized block of data that the IBM Virtual Shared Disk sends over the network for an I/O request	Default is **24576**
processor_type	S	Specifies if the node is a uniprocessor or multiprocessor.	**UP** or **MP**
processors_installed	I	The total number of processors in the node.	
VSD_do_ip_checksum	S	Test to determine when to do IP checksum calculation.	
lppsource_name	S	The name of the LPP source resource from which to obtain AIX file sets for the installation of this node.	

Class = DependentNode

Attribute Name	*Type*	*Description*	*Comments*
node_number	I	Represents the relative node number for the dependent node.	This is set by the **-n** flag of **endefnode**.
entension_node_identifier	S	A unique identifier in the administrative environment of an extension node. For example, for the Ascend GRF switched IP router, the slot number of the SP Switch Router Adapter on the GRF.	This is specified by the **-i** flag of the **endefnode** command.
reliable_hostname	S	Represents the host name associated with the dependent nodes network interface on the administrative network.	This is specified by the **-a** flag on the **endefnode** command.
management_agent_hostname	S	Represents the host name of the dependent node's SNMP Agent that interacts with the SNMP Manager on the SP. This host name must resolve to an IP address.	This is specified by the **-s** flag of the **endefnode** command. May be the same as or different from the **reliable_hostname**, depending on the characteristics of the extension node.
snmp_community_name	S	Represents the SNMP community name for the authentication field in the SNMP messages exchanged between the SNMP Agent on the dependent node and the SNMP Manager on the SP	This is specified by the **-c** flag of the **endefnode** command.
switch_node_number	I	Represents the switch port to which the dependent node is connected.	
switch_number	I	Represents the switch board to which the dependent node is connected.	

Attribute Name	*Type*	*Description*	*Comments*
switch_chip	I	Represents the number of the switch chip to which the dependent node is connected.	
switch_chip_port	I	Represents the number of the switch chip port to which the dependent node is connected.	
switch_partition_number	I	Represents the number of the switch partition for this dependent node.	

Class = Frame

Attribute Name	*Type*	*Description*	*Comments*
frame_number	I	Frame number	
tty	S	The tty port to be used for this frame (like /dev/tty1)	From customer via CMI during configuration
frame_type	S	Switch or no switch	From information returned by the monitor.
MACN	S	Stands for Monitor and Control Node	Hostname of the machine that monitors this frame
backup_MACN	S	The backup monitor and control node for hardware monitoring with High Availability Control Workstation Connectivity function.	
slots	I	Frame height	A full-size frame is 79" (2.00 meters). An SP Switch-8 frame is 49" (1.25 meters).
frame_in_config	I	Relative position of frame within expansion frame configuration (0,1,2,3).	
snn_index	I	Internal attribute used to index to determine correct switch node numbers.	
switch_config	I	Internal attribute used to determine switch configuration.	

Class = Pool

Attribute Name	*Type*	*Description*	*Comments*
pool_id	I	The ID of the pool this node belongs to.	Identifies the pool a node is in.
node_name	S	A name that can be used to contact the node.	Comes from node_type_file.
pool_type	I	InteractiveOnly, BatchOnly, or General (1,2, or 3)	
attribute	S	User description of pool	Customer description of pool

Class = Switch

Attribute Name	*Type*	*Description*	*Comments*
switch_number	S	Switch Number	If less than 1000, switches connected to nodes. If greater than 1000, intermediate switches.
frame_number	I	Frame number	
slot_number	I	Slot number within frame	Node switches equal 17. Intermediate switches equal 2,4,6,8,10,12,14,16.
switch_partition_number	I	Switch partition that the switch is on.	Default is **1**
switch_type	I	Switch supervisor card type (49, 50, or 129)	
clock_input	S	Clock input source	High Performance Switch possible values: 0, 1, 2, 3 SP Switch possible values: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 27, 28, 29, 30, 31, 32, 33, 34
switch_level	I	Microcode level from the switch supervisor	

Attribute Name	*Type*	*Description*	*Comments*
switch_name	S	Short model name of the switch	
clock_source	I	Switch number of the switch providing the clocking signal for the current switch	
clock_change	S	Clock input has changed since last **Eclock**	

Class = Switch_partition

Attribute Name	*Type*	*Description*	*Comments*
switch_partition_number	I	The partition number for High Performance Switch use	
topology_filename	S	The filename of the topology file that is stored in the SDR	
primary_name	S	The reliable hostname of the switch primary node	
arp_enabled	S	Address resolution protocol	Default is **no**
switch_node_number_used	S	Switch adapter's IP addresses were assigned based on the switch node numbers	Default is **yes**
run_phase_duration	I	Length of time in minutes that the switch stays in run phase. Range is 1–117281 minutes.	Default is **2**
primary_backup_name	S	Reliable host name of this system partition's primary backup node	
oncoming_primary_name	S	Reliable host name of the node that will become the primary node on the next **Estart**	
oncoming_primary_backup_name	S	Reliable host name of the node that will become the primary backup node on the next **Estart**	
num_nodes_success	I	Number of nodes in the system partition that successfully received the topology file	

Attribute Name	*Type*	*Description*	*Comments*
switch_max_ltu	I	Represents the maximum packet length of the data portion of an IP packet sent over the switch. This value is determined by internal switch support, and should not be changed.	
switch_link_delay	I	Represents the switch link delay value used by switch support when initializing the switch network. This value is determined by internal switch support, and should not be changed.	

Class = Adapter

Attribute Name	*Type*	*Description*	*Comments*
node_number	I	Relative node number	
adapter_type	S	Adapter type as recorded in ODM (lsp0, enet0, enet1, fi0, fi1 etc.)	Set via customer through CMI. Only en0 is required.
netaddr	S	ip address for this interface.	Set via customer through CMI.
netmask	S	netmask for this interface.	Set via customer through CMI.
tok_rate	I	Speed of token rate, (4 or 16 mb)	
enet_type	S	**bnc** or **dix** or **N/A**	
ucode_version	S	The file name of the microcode that is loaded	
prog_logic_version	S	Programmable logic	Currently not used

Class = DependentAdapter

Attribute Name	*Type*	*Description*	*Comments*
node_number	I	Relative node number for the dependent node.	This is specified by the **-n** flag of the **endefadapter** command.

Attribute Name	*Type*	*Description*	*Comments*
netaddr	S	Represents the IP address assigned to the dependent node.	This is specified by the **-a** flag of the **endefadapter** command.
netmask	S	Represents the netmask associated with this dependent node.	This is specified by the **-m** flag of the **endefadapter** command.

Class = host_responds

Attribute Name	*Type*	*Description*	*Comments*
node_number	I	Relative node number	This attribute gets mapped to frame/slot to retrieve DE value
host_responds	I	0 or 1 to indicate host is down or up.	This is really a query into the frame info written by the System Monitor.

Class = switch_responds

Attribute Name	*Type*	*Description*	*Comments*
node_number	I	Relative node number.	
switch_responds	I	0 or 1 if switch is inactive/active on the node.	This is written by switch code.
autojoin	I	Should this node join switch on startup.	
isolated	I	**1** or **0** if this node is isolated from the switch or not	
adapter_config_status	S	Status of the switch adapter in the node	

Class = SP

Attribute Name	*Type*	*Description*	*Comments*
control_workstation	S	Hostname of control workstation	
cw_ipaddrs	S	IP addresses of control workstation	
install_image	S	Full path of net install image.	Customer supplied from CMI at installation
remove_image	S	If set to **true** the install images are erased after all nodes have been installed from this server.	Customer supplied from CMI at installation
primary_node	S	Ethernet hostname of primary node.	Default is node 1 frame 1.
ntp_config	S	The kind of NTP installation done: **none**=Do not configure NTP for this site. **consensus**=setup NTP to configure Control Workstation as the NTP server and boot servers as NTP peers. **timemaster**=site has an existing NTP server. Configure NTP to user these. NTP_SERVERS will contain the NTP server hostnames. **internet**=The Control Workstation has access to the internet. Configure the Control Workstation to be an NTP server using the internet time server(s) defined in NTP_SERVERS.	From customer input from CMI.
ntp_server	S	Hostname of NTP server.	From customer input from CMI
ntp_version	S	Default is 3.	Customer can override using CMI
amd_config	S	**True** or **False** for whether the SP will provide automounter support	Customer can input in CMI, default is false.
print_config	S	**false**=do not print **secure**=install print in secure mode **open**=install print in open mode	Default is false, changed from customer via CMI

Attribute Name	*Type*	*Description*	*Comments*
print_id	S	User name to rsh to for secure mode printing	From customer via CMI
usermgmt_config	S	**false**=do not install SP user management **true**=install SP user management code and SMIT interface	Default is false, changed from customer via CMI
passwd_file	S	File containing password information (/etc/passwd, or other)	Default is /etc/passwd
passwd_file_loc	S	Location of password file	
homedir_server	S	Hostname of default user directory server for user management code	From customer via CMI, default is control workstation
homedir_path	S	Default path to user home directories.	From customer via CMI, default is /home/$homedir_server
filecoll_config	S	**True** or **false** if file collection management code should be installed	From customer via CMI, default is false
supman_uid	S	UID for supman	
supfilesrv_port	I	File collections port number	
spacct_enable	S	Flag indicating whether, by default, accounting is enabled on all nodes that have an accounting enabled attribute of **default**	default is **false**
spacct_actnode_thresh	I	The SP accounting active node threshold indicates the percentage of nodes for which accounting data must be available, for merging and reporting of the data for a cycle to take place.	default is 80
spacct_excluse_enable	S	The cluster exclusive use accounting enabled attribute-indicates whether accounting start and end job records will be generated for jobs having exclusive use of the node.	True or false. Default is **false**.
acct_master	I	Specifies which node is to act as the accounting master.	default is 0

Attribute Name	*Type*	*Description*	*Comments*
cw_has_usr_clients	S	Either **true** or **false**	Default is **false**
code_version	S	Contains the ssp release that is installed on the CWS	Default is nothing (unknown)
layout_dir	S	Pathname of layout directory for current system partition configuration	
authent_server	S	The type of authentication server used. The possible values are **ssp** (SP authentication services) (**ssp.authent** server, either primary or slave, on the control workstation), **afs** (afs server on either the control workstation or elsewhere), **krb** (SP authentication services or other MIT Kerberos 4 server(s) on other workstations only).	The default is **krb**. The value is set by **install_cw**.
backup_cw	S	Kernel hostname of backup control workstation	If value is null, configuration is not HACWS
ipaddrs_bucw	S	IP addresses of backup control workstation that are not service adapters and are not involved with fail over	If value is null, the backup control workstation is not reachable as the backup during fail over periods
active_cw	S	Active control workstation, either control workstation or backup control workstation	If value is null, configuration is not HACWS
cw_lppsource_name	S	The name of the LPP source resource from which to obtain NIM file sets for the control workstation.	

Class = Dont_care_pool_list

Attribute Name	*Type*	*Description*	*Comments*
pool_id	I	A pool in the don't care list.	
order_number	I	The position of the pool in the dont_care_pool_list	User must enforce no duplicates or holes.

Class = JM_server_nodes

Attribute Name	*Type*	*Description*	*Comments*
node_name	S	A name that can be used to contact the node.	Comes from node_type_file.
order_number	I	The position of this node in the JM_server_nodes list	User must enforce no duplicates or holes.

Class = JM_domain_info

Attribute Name	*Type*	*Description*	*Comments*
JM_domain_id	S	Identifier for a JM domain	
server_name	S	The node on which the JM server in this domain is running.	This may change as JM moves to a backup node.
pid	I	The process ID for the JM server in this domain.	
backup_name	S	The node that should server as backup to the JM server.	
backup_pid	I	The process ID of the JM server on the backup node.	Backup servers are always running. They just wake up when needed.
access_control	I	Whether authentication should be checked for user on node	0=**No**=Don't check, 1=**Yes**=Check
num_pools	I	Number of pools	
num_server_nodes	I	Number of JM server candidate nodes	
num_dont_care_pools	I	Number of don't care pools	
en_adapter	I	Represents what Ethernet adapter to use for Ethernet job applications. Corresponds to the `EN_ADAPTER` field in the **jm_config** file.	Valid values are **0** and **1**. Default for **en0** is **0**.

Class = JM_job_info

Attribute Name	*Type*	*Description*	*Comments*
job_id	I	Identifier for a JM job	Job ID for whole job. Subjobs identified by this ID + node.
group_id	I	The job group this job belongs to.	
client_node	S	The node name of the client that asked for the job.	This a hostname.
client_pid	I	The process ID of the client this job belongs to.	
client_uid	I	The user ID of the client this job belongs to.	
job_description	S	The job description associated with the job	Only set when **node_number=1**
adapter_type	I	The type of adapter the job is using: Ethernet, FDDI, and so on.	
pool_type	I	The type of request pool. Interactive = **0**, batch = **1**	Only set when **node_number=1**
req_auth	I	What action must be taken during deallocation	Only set when **node_number=1**
time_alloc	S	The time when the request was received by the server	Only set when **node_number=1**
node_number	I	Data about entire job for node number **1**. Data about specific node in the job for all other node numbers.	
cpu_usage	I	What type of usage was requested of node	Only set when node number is > 0.
adapter_usage	I	What type of usage was requested of node	Only set when node number is > 0.
num_virtual_tasks	I	The number of virtual tasks running on the node	Only set when node number is > 0.
virtual_task_ids	S	The virtual task IDs running on the node	Only set when node number is > 0.
rc	I	Indicates the return code associated with the node	Only set when node number is > 0.

Class = JM_node_usage

Attribute Name	*Type*	*Description*	*Comments*
node_number	I	The number of the node whose usage data is described in this object	
adapter_type	I	The adapter type whose usage data is described in this object	
current	I	How many jobs are currently using the specified CPU or adapter	
remaining	I	Indicates whether another job can use this CPU or adapter	**0** = no other job can use; **1** = other jobs can use if willing to share

Class = VSD_Global_Volume_Group

Attribute Name	*Type*	*Description*	*Comments*
global_group_name	S	Global Volume Group name	
local_group_name	S	Local Volume Group name	
primary_node	I	Primary node for this volume group	This is a relative node number.
secondary_node	I	Secondary node for this volume group	This is a relative node number and is OPTIONAL.
eio_recovery	I	Denotes if recovery should be made from a hardware error	
recovery	I	Denotes if recovery was made from a hardware error	**0** = recovery was not made; **1** = recovery was made
primary_ts	S	Primary node for a tail	
secondary_ts	S	Secondary node for a tail	

Class = HSD_Table

Attribute Name	*Type*	*Description*	*Comments*
minor_number	I	Unique minor numbers	
stripe_size	I	Stripe width	
num_vsds	I	Number of VSDs in HSD	
option	S	Whether the HSD will skip the first stripe on each underlying IBM Virtual Shared Disk in an HSD	Value is **protect_lvcb** or **not_protect_lvcb**
HSD_name	S	Hashed Shared Device name	
VSD_name	S	Virtual Shared Device name	
size_in_MB	I	Size of table	

Class = VSD_Table

Attribute Name	*Type*	*Description*	*Comments*
VSD_name	S	Volume group name	
global_group_name	S	Global Volume Group name	Same names as global_group_name in VSD_Global_Volume_group
logical_volume_name	S	The name of the logical volume comprising this vsd.	
minor_number	I	Minor number from minor table	
option	S	Options associated with this vsd	**cache** or **nocache** are the permissible values
size_in_MB	I	The size of the logical volume pointed to by an IBM VSD	
lv_blk0_pdev	I	The physical device pointed to by the logical volume	
lv_blk0_pbn	I	The offset a logical volume resides on on the physical device	

Class = VSD_Minor_Number

Attribute Name	*Type*	*Description*	*Comments*
minor_number	I	Unique minor numbers	

Class = HSD_Minor_Number

Attribute Name	*Type*	*Description*	*Comments*
minor_number	I	Unique minor numbers	

Class = VSD_Fence

Attribute Name	*Type*	*Description*	*Comments*
minor_number	I	The unique minor number of an IBM VSD that is fenced out of more than one node	
commit	I		**0** if in process of being committed; **1** if committed
Map	S	A hex dump of the map designating the nodes that an IBM VSD is fenced out of	

Class = SP_ports

Attribute Name	*Type*	*Description*	*Comments*
daemon	S	Name of a daemon	
hostname	S	Host that the daemon runs on	
port	I	Port that the daemon uses on node specified in **hostname**	

Class = GMT_Global_smt_nds

Attribute Name	*Type*	*Description*	*Comments*
node_1	I	Global Mount uses this node as primary smart node.	
node_2	I	Global Mount uses this node as backup smart node.	
node_3	I	Global Mount uses this node as backup smart node.	
node_4	I	Global Mount uses this node as backup smart node.	
node_5	I	Global Mount uses this node as backup smart node.	

Class = Syspar_map

Attribute Name	*Type*	*Description*	*Comments*
syspar_name	S	Short hostname of a system partition	
syspar_addr	S	IP address of a system partition	
node_number	I	A node number in this system partition	
switch_node_number	I	The switch node number of this node	
used	I	Is this node on IC bus of frame controller?	
node_type	S	Represents the type of node that this entry in the Syspar_map is. The values are **standard** and **dependent**.	The default is **standard**

Class = Syspar

Attribute Name	*Type*	*Description*	*Comments*
syspar_name	S	Hostname of system partition	
ip_address	S	IP address of system partition	
install_image	S	The image installed on this system partition	
syspar_dir	S	The directory from which this system partition was created	
code_version	S	IBM Parallel System Support Programs for AIX code version in this partition	
haem_cdb_version	S	Event Management Configuration Database (EMCDB) version string	

Class = pmandConfig

Attribute Name	*Type*	*Description*	*Comments*
pmNodenumber	I	The node number of the host on which **pmand** is to subscribe to the event.	If null, the **pmand** daemon on all nodes in the partition will subscribe to the event.
pmEventid	I	The ID for this event on this node as returned by the Event Manager	
pmTargetType	S	Identifies the type of the parameter in the **pmTarget** field.	**NODE_LIST**, **NODE_RANGE**, or **NODE_GROUP**
pmTarget	S	Lists the targets on which daemons are to be configured. Can be a list of host names (**NODE_LIST**) separated by commas, a range of nodes (**NODE_RANGE**) as specified by **-n** argument to **hostlist**, or the name of a node group (**NODE_GROUP**).	A pre-defined node group, **ALLPMAN**, is provided. This node group consists of all the nodes in the current system partition.

Attribute Name	*Type*	*Description*	*Comments*
pmRvar	S	Specifies the resource variable name for the event.	Required.
pmIvec	S	Specifies a fully-qualified instantiation vector for the event.	Required.
pmPred	S	Specifies a predicate for the event.	Required.
pmRearmPred	S	Specifies a predicate for the rearm event.	Optional.
pmCommand	S	The command to issue when one of the specified events occurs.	Optional.
pmCommandTimeout	S	The amount of time, in seconds, to wait before assuming that the **pmCommand** has hung, and killing it.	Optional.
pmHandle	S	An identifier (handle) for the event-action pair represented by this entry.	Required.
pmTrapid	I	Specifies an ID to be put into an SNMP trap to be sent when the event occurs.	Optional.
pmPPSlog	I	If non-zero, specifies logging via ppslog().	Optional.
pmRearmCommand	S	The command to issue when the rearm event occurs.	Optional.
pmRearmCommandTimeout	S	The amount of time to wait, in seconds, before assuming the command has hung, and killing it.	Optional.
pmRearmTrapid	I	Specifies a specific id to be put into an SNMP trap to be sent when the rearm event occurs.	Optional.
pmRearmPPSlog	I	If non-zero, specifies logging via ppslog().	Optional.
pmUsername	S	The user name of the subscriber event.	Defaults to root.

Attribute Name	*Type*	*Description*	*Comments*
pmPrincipal	S	The Kerberos principal of the requesting client (the client authorized to modify the subscription).	Required.
pmHost	S	The host from which the subscription was commanded.	Defaults to the local node.
pmActivated	S	Indicates whether a specific event-action pair is activated (subscribed to) or not.	Default is activated.
pmText	S	Optional text to be placed in the log record.	
pmRearmtext	S	Optional text to be placed into the log record.	
pmUserLabel	S	Stores the name of the **Named_Condition** from the Event Management subsystem, if one was used to subscribe to a problem management record.	Used by the SP Perspectives GUI.

Class = pmanrmdConfig

Attribute Name	*Type*	*Description*	*Comments*
pmrmNodenumber	I	The node number of the host on which **pmanrmd** is to instantiate the resource variable being defined.	
pmrmTargetType	S	Identifies the type of the parameter in the **pmTarget** field.	**NODE_LIST**, **NODE_RANGE**, or **NODE_GROUP**

Attribute Name	*Type*	*Description*	*Comments*
pmrmTarget	S	Lists the targets on which daemons are to be configured. Can be a list of host names (**NODE_LIST**) separated by commas, a range of nodes (**NODE_RANGE**) as specified by the **-n** argument to **hostlist**, or the name of a node group (**NODE_GROUP**).	
pmrmRvar	S	Specifies the resource variable being defined.	Must be **IBM.PSSP.User_state***nn*, where *nn* is 1-16.
pmrmSampInt	I	Specifies a sampling interval, in seconds, for the resource variable.	
pmrmCommand	S	Specifies a command to run at the specified interval.	

Class = SysNodeGroup

Attribute Name	*Type*	*Description*	*Comments*
name	S	Name associated with a node group.	The first character of a node group name must be a letter. The remaining characters may be letters, numbers, . (period), or _ (underbar).
defList	S	Nodes and node groups which are members of this node group.	For a node group named “A”, a list of nodes and node groups that belong to “A” directly (not through member node groups).

Class = NodeGroup

Attribute Name	*Type*	*Description*	*Comments*
name	S	Name associated with a node group.	The first character of a node group name must be a letter. The remaining characters may be letters, numbers, . (period), or _ (underbar).
defList	S	Nodes and node groups which are members of this node group.	For a node group named “A”, a list of nodes and node groups that belong to “A” directly (not through member node groups).

Class = TS_Config

For the default settings of the attributes in this class, see the man pages for the **hbctrl** and **hatsctrl** commands in *PSSP Command and Technical Reference*.

Attribute Name	*Type*	*Description*	*Comments*
Frequency	I	The frequency, in heartbeats per second, with which heartbeats are sent out	
Sensitivity	I	The number of heartbeats from the neighboring node that can be missed before the neighbor is declared inoperative	
Run_FixPri	I	Whether the execution priority is fixed or floating (1=fixed, 0=floating)	
FixPri_Value	I	The value of the execution priority that is used on the **set_priority** system call	

Class = Syspar_ports

Attribute Name	*Type*	*Description*	*Comments*
subsystem	S	Name of a system partition-sensitive subsystem	
port	I	Port number used for intrasubsystem communication	

Class = EM_Condition

Attribute Name	*Type*	*Description*	*Comments*
name	S	The name of this condition	
variable	S	The resource variable name this condition applies to	
predicate	S	The condition	
rearm	S	The rearm condition	
specified	S	The specified instance elements	
unspecified	S	The unspecified instance elements	
description	S	A description of this condition	

Appendix G. IP Address and Host Name Changes for SP Systems

The information in this appendix describes the activities that the administrator needs to perform when changing either the primary Ethernet IP address and host name for the SP nodes, or the primary IP address and host name for the control workstation. We strongly recommend that you avoid making any host name or IP address changes if possible. The tasks are tedious and in some cases require rerunning the SP installation steps. For more information on SP networks, reference the section "Planning your Network" located in *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment*.

This chapter provides procedures that are necessary for changing the IP address and host name for PSSP systems with control workstations running the PSSP 2.3 release. If your control workstation is at an earlier PSSP level, refer to *IP Address and Host Name Changes for PSSP Systems* from your IBM marketing representative, or Service Bulletin # 10 on the web (the current address is **http://www.rs6000.ibm.com/support/sp/sp_secure/bulletins**) under the IBM RS/6000 home page.

These IP address and host name procedures support SP nodes at PSSP levels PSSP 1.2 (AIX 3.2.5), PSSP 2.1 (AIX 4.1), PSSP 2.2 (AIX 4.1-4.2), and PSSP 2.3 (AIX 4.2) systems. The PSSP 2.3 release supports both SP node coexistence and system partitioning. There are special considerations relating to PSSP 1.2 level nodes when working with system partitions.

Changing the IP address or host names of SP nodes affect the entire SP system. IP address and host names are located in the System Data Repository (SDR) using objects and attributes. IP address and host names are also kept in system-related files that are located on SP nodes and the control workstation. It is imperative that all new IP addresses and host names being used are resolvable on your SP system.

The following PSSP components must be considered when changing the IP address and host names:

- Network Installation Manager (NIM)
- System partitioning
- IBM Virtual Shared Disk
- High Availability Control Workstation (HACWS)
- High availability services
- Problem management subsystem
- Performance monitor services
- Extension nodes

SDR Objects with IP Address or Host Names

The following SDR objects reference the host name and IP address in the SP system for PSSP systems:

Adapter Specifies the IP addresses used with the **ent**, **css0**, **tr0**, and **fi0** adapters

© Copyright IBM Corp. 1995, 1997

Frame	Specifies the Monitor and Control Node **MACN**, and HACWS **backup_MACN** attributes on the control workstation that work with host names
JM_domain_info	Works with the host names for Resource Manager domains
JM_Server_Nodes	Works with the host names for Resource Manager server nodes
Node	Works with the initial or reliable host names and uses the IP address for SP nodes and boot servers. The nodes are organized by system partitions.
Pool	Works with host names for Resource Manager pools
SP	Works with control workstation IP address and host names. Uses the host name when working with Network Time Protocol (NTP) printing, user management, and accounting services
SP_ports	Works with the host name used with **hardmon** and the control workstation
Switch_partition	Works with host name for primary and backup nodes used to support the **css** SP switch
Syspar	Works with the IP address and SP_NAME with system partitions
Syspar_map	Provides the host name and IP address on the control workstation for system partitions
pmandConfig	Captures the SP node host name data working problem management
SPDM	Works with host name for Performance Monitor status data
SPDM_NODES	Works with host name for SP nodes and organized by system partition

The following SDR objects are added for PSSP 2.3 systems:

DependentNode Works with host name for dependent extension node

DependentAdapter Works with IP address for dependent extension node adapter

SP System Files with IP Address and Host Names

The following files contain the IP address or host name that exist on SP nodes and the control workstation. We recommend that you look through these files when completing the procedures for changing host names and IP addresses for your SP system. The following files are available for PSSP systems:

/.rhosts	Contains host names used exclusively with **rcmd** services.
/.klogin	Contains host names used with authentication **rcmd** services.
/etc/hosts	Contains IP addresses and host names used with the SP system.
/etc/resolv.conf	Contains the IP address for Domain Name Service (DNS) (Optional).

/var/yp/*nis* References the host name and IP address with Network Information Service (NIS).

/etc/krb.conf Works with the host name for the authentication server.

/etc/krb.realms
Works with the host name of the SP nodes and authentication realm.

/etc/krb-srvtab Provides the authentication service key using host name.

/etc/SDR_dest_info
Specifies the IP address of the control workstation and the SDR

/etc/ssp/cw_name
Specifies IP address of control workstation host name on SP nodes that work with node installation and customization.

/etc/ssp/server_name
Specifies IP address and host name of the SP boot/install servers on SP nodes working with node customization.

/etc/ssp/server_hostname
Specifies IP address and host name of the SP install servers on SP nodes working with node installation.

/etc/ssp/reliable_hostname
Specifies IP address and host name of the SP node working with node installation and customization.

/etc/ntp.conf Works with the IP address of the NTP server (Optional).

/etc/filesystems
Can contain the IP address or host name of NFS systems (mainly used on **/usr** client systems).

/tftpboot/*host***.config_info**
Contains the IP address and host name for each SP node. It is found on the control workstation and boot servers.

/tftpboot/*host***.intstall_info**
Contains the IP address and host name for each SP node. It is found on the control workstation and boot servers.

/tftpboot/*host***-new-srvtab**
Provides authentication service keys using host name. It is found on the control workstation and boot servers.

/etc/rc.net Contains the alias IP addresses used with system partitions

/etc/niminfo Works with the NIM configuration for NIM master information

/etc/sysctl.acl Uses host name that works with Sysctl ACL support

/etc/logmgt.acl
Uses host name that works with Error Log Mgt ACL support

/spdata/sys1/spmon/hmacls
Uses short host name that works with hardmon authentication services

/etc/jmd_config.*SP_NAME*
Works with host names for Resource Management on the control workstation for all defined SP_NAME syspars.

/usr/lpp/csd/vsdfiles/VSD_ipaddr
Contains the SP node IBM Virtual Shared Disk adapter IP address

/spdata/sys1/ha/cfg/em.*SP_NAME***cdb.***data*
Uses Syspar host name that works with configuration files for Event Management services.

/var/ha/run/*Availability Services*
Uses Syspar host name that contains the run files for the Availability Services, which include **haem** (Event Manager), **hats** (Topology Services) **hags** (Group Services-Ethernet), and **hagsglam** (Group Services- switch).

/var/ha/log/*Availability Services*
Uses Syspar host name that contains the log files for the Availability Services.

/var/adm/SPlogs/pman/*data*
Uses Syspar host name that contains the log files for the Problem Management subsystem.

/etc/services Specifies short host name based on *SP_NAME* partition that work with Availability Services port numbers

/etc/auto/maps/auto.u
Contains host names of the file servers providing NFS mounts to Automount

/etc/amd/amd-maps/amd.u
Contains host names of the file servers providing NFS mounts to Amd

Changing the IP Address and Host Names for PSSP 2.3 Systems

Changing the IP address and host names for your SP system will cause changes to the control workstation and SP nodes. You may want to create a **mksysb** backup for the control workstation and backup critical file systems on the control workstation and SP nodes.

Changing a host name or IP address for SP nodes or the control workstation requires the administrator to run the procedures based on the configuration of the SP system. There are network and SP configuration updates required on both the control workstation and SP nodes. We recommend that you document and understand all the required IP address and host name changes prior to performing this activity. It is necessary to understand which SP nodes will be the boot server nodes in your SP configuration.

You will need to update the **en0** network interface and update system files on each SP node that have IP address or host name changes. You will need to know the new selected IP addresses and host names being used by the SP nodes, control workstation, and gateway servers. This is required so that the SP node can properly communicate with the control workstation during reboot activity.

You must work on the control workstation and make the necessary changes to the SDR. This includes updates to reset PSSP daemons, PSSP system resource objects, and system partition files. You need to update files on the control workstation to reflect the new host names and IP addresses. During this activity,

the SP system should not be available to users. The SP system is not ready until the SP nodes reference the new IP address and host name changes.

You must set the SDR boot response to **customize** and then execute a reboot for the SP nodes. After the SP nodes complete the reboot and are customized, you should verify that the files on the SP nodes reflect the new IP address and host name changes.

The PSSP installation support requires the **admin** user to replace the NIM master objects. It may be necessary to update the NIM network and machine resources to match the new IP address or host name. The NIM updates are needed for the control workstation and boot/install server nodes. There are installation commands in PSSP that support the implemetation of NIM interfaces.

The PSSP system provides High Availability Services which are subsystems and interfaces that are installed on the control workstation and SP nodes. These subsystems (**hags, hats, haem, pman, sp_configd, hr, and hb**) have replaced the host responds (**hr**) and heartbeat (**hb**) interfaces from PSSP 2.1.

If you use system partitioning on your SP system, you may want to rerun the system partitioning steps on your SP system. This allows the necessary Syspar config files to reflect all updates. If you support PSSP 1.2 (AIX 3.2.5) nodes, you must manually change the SP-based files on the PSSP 1.2 boot server node.

If your SP configuration supports HACWS, you must reconfigure the current HACWS configuration with the new IP address and host name changes. You will need to execute specific HACWS procedures to re-create the HACWS configuration. Refer to “Working with HACWS Configurations” on page 625 for more information.

If your SP system supports IBM Virtual Shared Disk configurations, you must execute some IBM Virtual Shared Disk commands on the SP nodes. You may need to stop and unconfigure the IBM Virtual Shared Disk, make IP address and host name changes, and then configure and start the IBM Virtual Shared Disk on the SP nodes.

If your SP system supports extension nodes (such as a switched IP router) you will need to reconfigure the extension node's host name and IP address.

Updating the PSSP 2.2/2.3 Node Interfaces

You need to make the following updates on all PSSP 2.3 and PSSP 2.2 nodes prior to making updates to the control workstation.

Note to the SP Administrator

If your SP node is running at PSSP 1.2 level, you can refer to “Updating the PSSP 1.2 Boot Server Node on PSSP 2.x Systems” on page 619. You will need to manually update files on your PSSP 1.2 boot server node prior to completing all control workstation steps.

If your SP node is running at PSSP 2.1 level, you can refer to “Updating the PSSP 2.1 Node Interfaces” on page 617.

The following steps need to be verified on each PSSP 2.3 and PSSP 2.2 level node. Some of the steps can be performed by implementing user shell scripts and by using the SP distributed shell (**dsh**) services from the control workstation to the target nodes.

1. Stop/Remove the Availability Services

   You need to shutdown the daemons and source master objects that work with PSSP Availability Services. The Availability services include Topology Services, Group Services, Event Manager, Problem Management, and Performance Monitor. Execute the **syspar_ctrl** command to stop and remove these resources.

   ```
   /usr/lpp/ssp/bin/syspar_ctrl -c
   ```

2. Update the NIM files and configuration

   If available, update the /**etc**/**niminfo** file to provide the new host name and IP address changes for the NIM client, NIM master, and NIM file references. If the SP node is currently a NIM master (boot/install server), you need to remove the NIM ODM objects and NIM master configuration files from the SP node by issuing the **delnimmast** command:

   ```
   /usr/lpp/ssp/bin/delnimmast -l node_number
   ```

3. Unconfigure IBM Virtual Shared Disks (Optional)

   You need to determine if there are any IBM Virtual Shared Disk configurations available on the SP node by issuing the /**usr**/**lpp**/**csd**/**bin**/**vsdatalst -n** command. If you receive an error message indicating “not found” or your SP node is not listed in the IBM Virtual Shared Disk configuration, you do not have to be concerned about the IBM Virtual Shared Disk. If your SP node is listed as having an IBM Virtual Shared Disk configuration, you will need to suspend, stop, and unconfigure the IBM Virtual Shared Disk for this SP node.

   ```
   /usr/lpp/csd/bin/suspendvsd -a
   /usr/lpp/csd/bin/stopvsd -a
   /usr/lpp/csd/bin/ucfgvsd -a
   ```

4. Update PSSP files for updates on each PSSP node

   You need to update the following files to reflect the new host name and IP address for the SP node and servers:

   a. Update the /**etc**/**SDR_dest_info** file using the new control workstation and SDR IP addresses and host names.

   b. Update the /**etc**/**ssp cw_name**, **server_hostname** and **server_name** files to reference the new boot server host name and IP address.

   c. Update the /**etc**/**ssp**/**reliable_hostname** file to reference the new SP node client host name and IP address.

   d. Make sure that the SP_NAME environment variable is updated or set to blank (export SP_NAME=).

   e. Issue /**usr**/**lpp**/**ssp**/**kerberos**/**bin**/**kdestroy** to remove any active Kerberos ticket-granting-tickets. You may need to update the /**etc**/**krb.conf** file if the authentication server is changed. We recommend renaming the /**etc**/**krb-srvtab** file on the SP node.

5. Updating the PSSP node **en0** interface

You now need to update the **en0** interface on each SP node to allow for proper communication paths between the SP node and the control workstation and you need to change the **en0** interface to the new IP address and host name. You must also supply the proper gateway and route information for the **en0** adapter. Make sure that all new IP addresses and host names are resolvable on the SP node. You do not need to modify any other adapter interfaces on the SP node. To accomplish this task, use the **mktcpip** command. You can do this using SMIT **mktcpip** or by issuing the **mktcpip** command on each SP node.

The following examples use the **mktcpip** command:

```
mktcpip -h new node hostname -a new node IP addr -m subnet mask \
-i en0 -n name server -d domain \
-g gateway IP addr -t bnc
```

The following example changes the host name of node k22n06 to a new IP address 129.40.88.70 and host name k88n06 in domain ppd.pok.ibm.com:

```
/usr/sbin/mktcpip -h'k88n06.ppd.pok.ibm.com' -a'129.40.88.70' \
-m'255.255.255.192' -i'en0' -n'129.40.70.1' -d'ppd.pok.ibm.com' \
-g'129.40.88.126' -t'bnc'
```

After issuing **mktcpip**, the node loses the current network connection to the control workstation and the other SP nodes. Any further updates to the SP node need to be performed using the **tty0 console** using the **s1term** command. You may be able to kill the "telnet" or "rlogin" process from the control workstation or the home machine to regain the window.

You have now completed the required updates for the PSSP 2.2 or PSSP 2.3 node. You now need to make the appropriate changes at the PSSP 2.3 control workstation. Do *not* reboot the PSSP nodes until instructed to do so. For information on making the appropriate changes at the control workstation, refer to “Execution on the PSSP 2.3 Control Workstation” on page 621.

Updating the PSSP 2.1 Node Interfaces

You need to make the following updates on all PSSP 2.1 nodes prior to making updates to the control workstation.

Note to the SP Administrator

If your SP node is running at PSSP 1.2 and uses system partitioning, you need to manually update files on your PSSP 1.2 boot server node prior to completing the control workstation steps. Please refer to “Updating the PSSP 1.2 Boot Server Node on PSSP 2.x Systems” on page 619.

The following steps need to be verified on each PSSP 2.1 node. Some steps can be performed by implementing user shell scripts and by using the SP distributed shell (**dsh**) services from the control workstation to the target nodes.

1. Update the NIM files and configuration

   If available, update the /**etc**/**niminfo** file to provide the new host name and IP address changes for the NIM client, master, and file references. If the SP node is a NIM master or a boot server, you need to remove the NIM ODM objects and configuration files from the SP node by issuing the following:

```
/usr/sbin/nim -o unconfig master
/usr/sbin/installp -u bos.sysmgt.nim.master
```

2. Unconfigure IBM Virtual Shared Disks (Optional)

   You need to determine if there are any IBM Virtual Shared Disk configurations available on the SP node by issuing the **/usr/lpp/csd/bin/vsdatalst -n** command. If you receive an error message indicating "not found" or your SP node is not listed in the IBM Virtual Shared Disk configuration, you do not have to be concerned about the IBM Virtual Shared Disk. If your SP node is listed as having an IBM Virtual Shared Disk configuration, you will need to suspend, stop, and unconfigure the IBM Virtual Shared Disk for this SP node.

   ```
   /usr/lpp/csd/bin/suspendvsd -a
   /usr/lpp/csd/bin/stopvsd -a
   /usr/lpp/csd/bin/ucfgvsd -a
   ```

3. Update PSSP files for updates on each PSSP 2.1 node

   You need to update the following files to reflect the new host name and IP address for the SP node and servers:

   a. Update the **/etc/SDR_dest_info** file using the new control workstation and SDR IP addresses

   b. Update the **/etc/ssp cw_name**, **server_hostname**, and **server_name** files to reference the new boot server host name and IP address.

   c. Update the **/etc/ssp/reliable_hostname** file to reference the new SP node client host name and IP address.

   d. Make sure that the SP_NAME environment variable is updated or set to blank (export SP_NAME=).

   e. Issue **/usr/lpp/ssp/kerberos/bin/kdestroy** to remove any active Kerberos ticket-granting-tickets. You may need to update the **/etc/krb.conf** file if the authentication server is changed. We recommend renaming the **/etc/krb-srvtab** file on the SP node.

4. Updating the PSSP 2.1 node **en0** interface

   You now need to update the **en0** interface on each SP node to allow for proper communication paths between the SP node and the control workstation and you need to change the **en0** interface to the new IP address and host name. You must also supply the proper gateway and route information for the **en0** adapter. It is important that all new IP addresses and host names be resolvable on the SP node. You do not need to modify any other adapter interfaces on the SP node. To accomplish this task, use the **mktcpip** command. You can do this using SMIT **mktcpip** or by issuing the **mktcpip** command on each SP node.

   The following examples use the **mktcpip** command:

   ```
   mktcpip -h new node hostname -a new node IP addr -m subnet mask \
   -i en0 -n name server -d domain \
   -g gateway IP addr -t N/A
   ```

   The following example changes the host name of node k22n06 to a new IP address 129.40.88.70 and host name k88n06 in domain ppd.pok.ibm.com:

   ```
   /usr/sbin/mktcpip -h'k88n06.ppd.pok.ibm.com' -a'129.40.88.70'  \
   -m'255.255.255.192' -i'en0' -n'129.40.70.1' -d'ppd.pok.ibm.com' \
   -g'129.40.88.126' -t'N/A'
   ```

After issuing **mktcpip**, the node loses the current network connection to the control workstation and the other SP nodes. Any further updates to the SP node will need to be performed using the **tty0 console** using the **s1term** command.

You have now completed the required updates for the PSSP 2.1 node. You now need to make the appropriate changes at the control workstation. Do **not** reboot the PSSP 2.1 nodes until instructed to do so. For information on making the appropriate changes at the control workstation, refer to "Execution on the PSSP 2.3 Control Workstation" on page 621.

Updating the PSSP 1.2 Boot Server Node on PSSP 2.x Systems

Updating of the PSSP 1.2 boot server node for a PSSP 1.2 (AIX 3.2.5) syspar can be tedious. All PSSP 1.2 system files that contain IP address and host names will need to be manually updated. Refer to "SP System Files with IP Address and Host Names" on page 612 to review the possible PSSP 1.2 files that need to be updated.

1. Update the files for the SDR, control workstation, and Kerberos servers.

   a. Update the **/etc/SDR_dest_info** file to point to the control workstation and the PSSP 1.2 (SP_NAME) SDR IP address. Set the SP_NAME variable to the correct system partition name or set it to blank.

   b. Update the **/etc/ssp** files **cw_name**, **server_hostname**, and **server_name** to reference the new control workstation and boot/install server host name and IP address.

   c. Issue **/usr/lpp/ssp/kerberos/bin/kdestroy** to remove any active Kerberos ticket-granting-tickets. You can update the **/etc/krb.conf** file to point to the correct Kerberos server. You may have to update the **/etc/krb.realms** file to match the **/etc/krb.realms** file found on the control workstation.

   You may need to extract a new **/etc/krb-srvtab** from the Kerberos database if changes are made to the Kerberos server. Issue the following command:

   ```
   /usr/lpp/ssp/kerberos/etc/ext_srvtab node
   ```

   on the control workstation. You can **ftp** the node's **srvtab** file *node***-new-srvtab** from the control workstation, and replace the **/etc/krb-srvtab** file on the boot server node.

   You need to update the **/.rhosts** and **/.klogin** files to have the proper **rcmd** support. You can reference the files from the control workstation.

2. Update the files to resolve IP address and host names.

   Based on your network configuration, you need to update the following files for all resolvable host names. You can **ftp** or **rcp** the files from the control workstation to the boot server node. The following are the network-related files:

   /etc/hosts Contains IP addresses and host names used with the SP system

   /etc/resolv.conf Contains IP address for domain name service (Optional)

   /var/yp/*nis* Contains NIS host name and IP address

Update files on the boot server to reflect its new host name and IP address. The following are some of the files:

/etc/ntp.conf Works with the IP address of the NTP server (Optional)

/etc/filesystems
Can contain the IP address or host name of NFS systems

3. Updating the PSSP 1.2 node adapters

   It may be necessary to make the appropriate updates in regard to the proper route tables and gateway servers for the network changes. You should verify that the current route tables are correct by issuing **netstat -nr**. If they are not correct, you can *flush* the route table and remove the route stanzas from the ODM. You can do this using the SMIT TCPIP panels or by issuing the **route -n -f** and **odmdelete** commands.

   Make sure that you re-create the proper default route tables using the SMIT TCPIP or the **mktcpip** command.

   Update the **en0** interface for the boot server node by issuing SMIT **mktcpip** or the **mktcpip** command. If you have additional adapters (**en1**, **tr0**, **fi0**) in the boot server node, issue the SMIT **mktcpip** or the **mktcpip** command for each adapter.

   ```
   mktcpip -h new node hostname -a new node IP addr -m subnet mask \
   -i en0 -n name server -d domain -g gateway IP addr
   ```

4. If the switch IP address was updated, you will need to update the **css0** adapter data. You need to update the ODM files to include new **css** IP address and netmask. You can do this by issuing the **odmdelete** and **odmadd** commands. A suggested path will be to customize this boot server node from another PSSP 1.2 node when the node has been properly updated.

   Make sure that the node number on this PSSP 1.2 boot server node matches the boot server node number listed in the SDR. If the node number in the SDR is not listed as a boot server node, issuing the **setup_server** command will remove required boot server files from **/usr/sys/inst.images/ssp** directory. You can validate this by issuing the following command on the PSSP 1.2 node:

   ```
   odmget -q "attribute=node_number" CuAt
   ```

   Compare the value to the node number listed in the SDR using the **splstdata -b** command.

   If the node numbers are different, you must update the SDR or the ODM to reflect the correct node number. You may need to use the **odme** editor or work with the **odmdelete** and **odmadd** commands.

5. Update the files on the **/tftpboot** directory.

   You must make sure that the files created on the **/tftpboot** directory are updated to reflect the new IP address and host names for each PSSP 1.2 node being installed by the server node. You will also need to update the **/tftpboot/firstboot.cust** file to make any IP address or host name changes.

   You should issue the **savebase** command to save the appropriate updates made to the ODM. We suggest issuing the **shutdown -Fr** command to reinitialize the PSSP 1.2 boot server node with the IP address and host name. When the node reboots, it executes the **setup_server** to create new PSSP 1.2 node **install_info**, **config_info**, and **srvtab** files in the **/tftpboot** directory. The following are some file examples:

node_hostname.**install_info**
node_hostname.**config_info**
node_short_hostname-**new-srvtab**

6. Review all files for updates

   At this time, we suggest reviewing all the files that were listed in the "SP System Files with IP Address and Host Names" on page 612 section. After you have verified the files on the PSSP 1.2 boot server node, you should be ready to perform a network boot for the remaining PSSP 1.2 nodes in that system partition.

   You may need to open the **tty0** console using **s1term** for the PSSP 1.2 node during the node customization if the LED indicates a hang state of 231.

7. Customize the main PSSP 2.1 server

   After the PSSP 1.2 nodes have been customized successfully, you should now set up the main PSSP 1.2 boot server node to be customized by the backup PSSP 1.2 boot server. You should reference the **install**, **config**, and **srvtab** files in the /**tftpboot** directory, then perform a network boot on the main boot server node as was done in the previous step.

Execution on the PSSP 2.3 Control Workstation

The following tasks need to be performed by the **admin** (root) user on the control workstation when changing the Ethernet IP address or host name.

You may want to backup the current /**spdata** file system, and create a **mksysb** backup before making any IP address and host name changes to the control workstation.

If you want to save the current SDR attributes, issue the command /**usr**/**lpp**/**ssp**/**bin**/**SDRArchive**. The support in PSSP allows the **admin** user to work with multiple system partition configurations. PSSP 2.3 supports coexistence of SP nodes at the PSSP 1.2, PSSP 2.1, PSSP 2.2, and PSSP 2.3 levels within a system partition.

If you are currently in an HACWS configuration, you should properly disable the backup control workstation and unconfigure HACWS prior to the changing any IP address and host names for the control workstation. Refer to "Working with HACWS Configurations" on page 625 for more information.

The following steps 1 through 10 are required steps when there are IP address and host name changes made to the control workstation. If there are only changes being made to the SP nodes, you may go to step 11 on page 623 in this section.

1. Stop PSSP daemons and remove current source master objects (**sdrd**, **hardmon** and Availability services) on the control workstation. The **syspar_ctrl -c** command will stop and remove all active resources that work with the PSSP Availability Services. The **stopsrc** command will stop the other PSSP resources. Remove the SDR source master object for each defined system partition (*SP_NAME*).

   ```
   /usr/lpp/ssp/bin syspar_ctrl -G -c
   /bin/stopsrc -g sdr
   /bin/stopsrc -s hardmon
   /bin/stopsrc -s supfilesrv
   /usr/lpp/ssp/bin/sdr -spname SP_NAME rmsrc
   ```

2. Using **mktcpip** or the **chdev** command, specify IP address or host name changes required for the control workstation. This includes changes being made for any affected adapter interfaces. It is also important to make appropriate updates to the route tables, netmasks, and gateway servers.
3. If multiple system partitions exist, update the **/etc/rc.net** file to reference the new alias address used for each system partition. Issue **/etc/rc.net** or execute **ifconfig** command to configure the new alias IP addresses. You can issue the **netstat -ni** command to validate the new alias addresses. The old alias addresses used with system partitions will be removed during the next reboot of the control workstation.
4. Manually update the **/etc/SDR_dest_info** and **/spdata/sys1/spmon/hmacls** files to the new IP address and host name for the SDR and hardmon interfaces.

   Manually update the SDR system partition map file **/spdata/sys1/sdr/system/classes/Syspar_map** to reflect the new IP address and their SP_NAME values with each system partition.

   Move the system partition directories found at location **/spdata/sys1/sdr/partitions** from the old IP address to the new IP address for each system partition being modified.

   ```
   /bin/mv old_IP_addr new_IP_addr
   ```

   Manually update the SDR system partition file **/spdata/sys1/sdr/partitions**/*new_IP_addr*/**classes/Syspar** to reflect the new IP address and their SP_NAME host name values for each defined system partition.
5. This step should only be performed for host name and domain changes for the control workstation. It is not required for changes made to control workstation IP addresses or to updates for SP nodes.

   Issue the **setup_authent** script to create authentication services for the new host names being used. See the step “Initialize RS/6000 SP Authentication Services” in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

   Manually check that the authentication files **/etc/krb.conf /etc/krb.realms** reference the proper host names and domain. You can issue **/usr/lpp/ssp/kerberos/bin/ksrvutil list** to make sure that the **rcmd** and **hardmon** services reference the new host names in the **/etc/krb-srvtab** file. You may also need to recreate Kerberos principals for any users that were previously defined in the Kerberos database.
6. Create the new source master resources for each SDR object, and then start the SDR and hardmon daemons.

   ```
   sdr -spname SP_NAME mksrc new_IP_addr (for each Syspar)
   startsrc -g sdr
   startsrc -s hardmon
   ```
7. After the SDR daemon is properly activated on the SP system, manually issue **SDRChangeAttrValues** for the control workstation.
   - To change the **hostname** attribute using the new host name for **SP_ports**, enter:

     ```
     SDRChangeAttrValues SP_ports hostname=new_CWS_hostname
     ```

- To change the **MACN** attribute using the new host name for **Frame**, enter:

  ```
  SDRChangeAttrValues Frame MACN=new_CWS_hostname
  ```

- To change the **control_workstation** attribute using the short host name for **SP**, enter:

  ```
  SDRChangeAttrValues SP control_workstation=new_CWS_short_hostname
  ```

8. Remove the current Network Installation Manager (NIM) ODM database and configuration files on the control workstation. You need to issue:

   /usr/lpp/ssp/bin/delnimmast -l 0

   It may be necessary to update the **/etc/niminfo** file to reference any changes to the new IP address and host names.

   You need to remove entries from the NFS export list that reference the old IP address and host names. You can do this by updating the **/etc/exports** file, and then stopping and starting NFS subsystem.

9. Using CMI or the **spsitenv** command, specify any host name changes that may be referenced for NTP, printing, and user management. See the step "Enter Site Environment Information" in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* (Optional).

10. Reboot the control workstation now. This establishes a clean system to reflect the IP address and host name changes. You should verify that all PSSP daemons are activated from **/etc/inittab** and the system resource master.

    Usage Note

    The remaining steps for the control workstation involve updating the SDR and SP system files for the SP node objects. Based on your SP configuration changes, some of the steps may be optional.

11. Update the SDR node objects by using commands or the SMIT-based SP Configuration Management Interface (CMI). You can perform the following tasks for each system partition by exporting the SP_NAME variable (for example, export SP_NAME=*SP_NAME*).

    a. Using CMI or the **spethernt** command, specify the new Ethernet IP address or host name changes required for the SP nodes. See the step "Enter Required Node Information" in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

    b. Using CMI or the **spadaptrs** command, reset the **css0** and other additional adapters that need to reference the new IP address or host names being changed. See the step "Configure Additional Adapters for Nodes" in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

    c. Using CMI or the **sphostnam** command, reset the initial host name that you want to use in your system. See the step "Configure Initial Host Names for Nodes" in *IBM Parallel System Support Programs for AIX: Installation and Migration Guide* (Optional).

    d. Using CMI or the **spbootins** command, reset all boot/install servers to reference the new Ethernet IP address and host names. It is important that you set your SP node boot/install servers into the proper configuration for your SP system. If you have system partitions, you should have already designated the proper PSSP boot server nodes.

```
spbootins -s no -n boot_node -l node_list
```

You need to set the bootp response to **customize** for all the SP nodes.

```
spbootins -r customize -l node_list
```

You can issue the **splstdata -G -b** command to verify the correct boot information for the SP nodes.

The **spbootins** without the **-s no** flag runs the **setup_server** command which creates all the NIM-based files and resources required for installation. It also creates the authentication **rcmd** principals for the new SP node host names.

You may want to validate that the following files have the proper IP adresses and host names defined.

/etc/ntp.conf
/tftpboot/*host***.config_info**
/tftpboot/*host***.install_info**
/tftpboot/*host***-new-srvtab**

Note to the SP Administrator

If this SP system supports any PSSP 1.2 nodes, you should manually update the designated PSSP 1.2 boot server node. Refer to the "Updating the PSSP 1.2 Boot Server Node on PSSP 2.x Systems" on page 619 section.

12. If your SP system supports system partitions, you may reissue the system partitioning steps found Chapter 17, "Managing System Partitions" on page 239. The system partitioning **spapply_config** command re-creates the proper PSSP daemons and resynchronizes the SDR objects to reflect any IP address and host name changes.

    If you do not execute the system partitioning step, you will then need to create the source master objects and start the daemons for the Availability Services using the **syspar_ctrl -G -A** command.

    ```
    /usr/lpp/ssp/bin/syspar_ctrl -G -A
    ```

13. Update configuration files on the control workstation

    If your SP system supports extension nodes, you may need to update an IP address or host name for an extension node. You may use CMI or commands **endefnode** and **endefadapter** to update extension node configuration information. See Chapter 19, "Managing Extension Nodes" on page 269.

    Various AIX and SP files may need to be updated to reflect IP address or host name changes. You should look through the following files for required updates.

    a. Update any files that are involved with host name resolution. The files are **/etc/hosts**, **/etc/resolv.conf** (DNS) and **/var/yp/*** (NIS).

    b. Update the **.rhosts** or **.klogin** files.

    c. Update the **/etc/filesystems** and **/etc/jmd_config.***SP_NAME* files for your SP configuration.

    d. Make sure your **/tftpboot/script.cust** file is updated to reflect IP address and host name changes. Instead of hardcoding host names, you can reference the $SERVER and $CWS variables.

14. For each PSSP boot server node (except PSSP 1.2), you should have already unconfigured NIM during PSSP node update activities. During the node customization, **setup_server** creates the proper install files for the nodes they are to customize. Since the SP nodes are in **customize** mode, most configuration files are updated to reflect the IP address and host name changes.

    Some files will not be updated on the nodes during the customization. You can **rcp** these files from the control workstation to the SP nodes by including them in the **/tftpboot/script.cust** file. These can include the **/etc/resolv.conf**, **/.rhosts**, and other SP customer-owned files.

    Perform a **REBOOT** for each PSSP node (except PSSP 1.2). You can do this by using the System Monitor GUI, or by issuing the **hmcmds** command.

    You need to follow the proper install sequence by customizing each of the boot server nodes first. After the SP boot/install server node completes the installation setup, you can customize the remaining SP nodes.

    Note to the SP Administrator

    For PSSP 1.2 nodes, you must first update all required files on the PSSP 1.2 boot server node. You must then perform a network boot on each PSSP 1.2 node, except for the current PSSP 1.2 boot server node. When the PSSP 1.2 nodes are properly updated, you can then perform a network boot on the primary PSSP 1.2 boot server node.

    You will now need to verify that the customization was successful. For more information, refer to the appropriate section (for PSSP 1.2 nodes, see "Execution on PSSP 1.2 Nodes" on page 628; for PSSP 2.1 nodes, see "Execution on PSSP 2.1 Nodes" on page 627; and for PSSP 2.2 and PSSP 2.3 nodes, see "Execution on PSSP 2.2 and PSSP 2.3 Nodes" on page 626).

15. Consider not performing the HACWS reconfiguration until your SP system has been properly updated and is stable on the new IP addresses and host names. When you are ready to reconfigure HACWS, refer to "Working with HACWS Configurations."

Working with HACWS Configurations

The following additional information for working with an HACWS configuration in relation to working with 1 through 10 in "Execution on the PSSP 2.3 Control Workstation" on page 621.

You may also need to reference the HACWS information "Installing and Configuring the High Availability Control Workstation" in the *IBM Parallel System Support Programs for AIX: Installation and Migration Guide*.

1. Stop daemons. Stop HACMP on both the primary and backup control workstations (this automatically stops all HACWS related daemons).
2. Use **mktcpip**. Manually issue the appropriate **ifconfig** commands to configure the control workstation service addresses (backup CWS) on the primary control workstation. Make sure the host name and IP address are configured on the backup control workstation. If you are using an alias IP address, make the required changes to the **/etc/rc.backup_cw_alias** script.

   Manually vary on the external volume group which contains the **/spdata** file system, and then mount the **/spdata** filesystem.

3. Update **rc.net**. Make updates to the /**spdata**/**sys1**/**hacws**/**rc.syspar_aliases** file to reference the new alias address used for system partition.
4. Update **SDR_dest_info**. No updates required.
5. Issue **setup_authent**. No updates required.
6. Create the new source master. Update the new SDR source master objects on the backup control workstation as well as the primary control workstation.

   ```
   sdr -spname SP_NAME mksrc new_IP_addr (for each Syspar)
   ```

   You can then start the SDR and hardmon daemons on the primary control workstation.
7. Update the SDR. Update the Frame object to include the host name for the **backup_MACN** using **SDRChangeAttrValues**.

   ```
   SDRChangeAttrValues Frame backup_MACN=backup_CWS_hostname
   ```

   You should also now execute the HACWS installation script **install_hacws** on the primary control workstation.

   ```
   /usr/sbin/hacws/install_hacws -p primary_name  -b backup_name
   -s
   ```
8. Remove the current NIM. No updates required.
9. Update the SDR node objects. No updates required.
10. Reboot. To reconfigure the cluster topology, follow the standard HACMP procedures documented in *HACMP/6000: Administration Guide*.

    Verify that the HACWS configuration is properly setup by executing the /**usr**/**sbin**/**hacws**/**hacws_verify** command on the control workstation.

    Also verify that you can properly **fail over** from the control workstation to the backup control workstation.

Execution on PSSP 2.2 and PSSP 2.3 Nodes

The following verification steps need to be performed on each PSSP 2.3 or PSSP 2.2 node for changes made for the IP address and host name:

1. After reboot is complete and the SP nodes are customized, verify that the following have the correct IP address and host name specified:
   - The updated SP Ethernet address specified for the SP node
   - The updated default route and gateway are correct
   - The additional adapters are correct (**en1**, **tr0**, **fi0**, **css0**)
   - The host name resolution files were updated correctly (/**etc**/**hosts**)
   - The /**etc**/**SDR_dest_info** file points to the control workstation and SDR Syspar IP address.
   - The acct_master host name is updated for the NFS export list directory /**var**/**adm**/**acct** (SP Accounting)
   - The files in the /**etc**/**ssp** directory *cw_name*, *server_hostname*, *server_name*, and *reliable_hostname* reference the correct IP address and host names.
   - The /**etc**/**krb.conf**, /**etc**/**krb.realms**, and /**etc**/**krb-srvtab** files have the correct host names defined

2. You may need to modify the following files on the SP nodes to reflect the updated IP addresses and host name definitions:

   /etc/filesystems
   /etc/ntp.conf
   /etc/resolv.conf
   /.rhosts
   /.klogin

   If any of the files are incorrect, make the proper updates for the correct IP address or host name.

3. Verify that the NIM resources were built on the PSSP boot server nodes by executing the **lsnim** command. If there were NIM problems, remove the current NIM database and configuration files.

   ```
   /usr/lpp/ssp/bin/delnimmast -l node_number
   ```

   Then issue **setup_server** on the PSSP boot server node.

4. Verify that the Availability Services have been activated on the PSSP nodes. Execute **lssrc -a** command to list all the active subsystems.

   ```
   /bin/lssrc -a   | grep Availability_susbsystem
   ```

   If you suspect that any of the expected Availability Services are inoperative or not available on the SP node, it is best to remove and then add the Availability services using the **syspar_ctrl** command.

   ```
   /usr/lpp/ssp/bin/syspar_ctrl -c
   /usr/lpp/ssp/bin/syspar_ctrl -A
   ```

5. Reconfigure and start the IBM Virtual Shared Disk for the SP nodes that support the IBM Virtual Shared Disk configuration. You can do this by issuing the following IBM Virtual Shared Disk commands:

   ```
   /usr/lpp/csd/bin/cfgvsd -a
   /usr/lpp/csd/bin/startvsd -a
   ```

   You can validate that the IP address and host names are correctly specified by issuing the following command:

   ```
   /usr/lpp/csd/bin/vsdatalst -n
   ```

6. If there were modifications made to any SP node system files, it would be best to reboot the PSSP nodes to reflect the IP address and host name changes. When the SP nodes are initialized, your SP system should be activated using the new IP addresses and host names.

Execution on PSSP 2.1 Nodes

The following verification steps need to be performed on each PSSP 2.1 nodes for changes made for the IP address and host name:

1. After reboot is complete and the SP nodes are customized, verify that the following have the correct IP address and host name specified:
   - The updated SP Ethernet address specified for the SP node
   - The updated default route and gateway are correct
   - The additional adapters are correct (**en1**, **tr0**, **fi0**, **css0**)
   - The host name resolution files were updated correctly (/**etc**/**hosts**)

- The **/etc/SDR_dest_info** file points to the control workstation and SDR Syspar IP address.
- The acct_master host name is updated for the NFS export list directory **/var/adm/acct** (SP Accounting)
- The files in the **/etc/ssp** directory *cw_name*, *server_hostname*, *server_name*, and *reliable_hostname* reference the correct IP address and host names.
- The **/etc/krb.conf**, **/etc/krb.realms**, and **/etc/krb-srvtab** files have the correct host names defined

2. You may need to modify the following files on the SP nodes to reflect the updated IP addresses and host name definitions:

   /etc/filesystems
   /etc/ntp.conf
   /etc/resolv.conf
   /.rhosts
   /.klogin

   If any of the files are incorrect, make the proper updates for the correct IP address or host name.

3. Verify that the NIM resources were built on the PSSP 2.1 boot server nodes by executing the **lsnim** command. If there were NIM problems, remove the current NIM database and configuration files.

   ```
   /usr/sbin/nim -o unconfig master
   /usr/sbin/installp -u bos.sysmgt.nim.master
   ```

   Then issue **setup_server** on the PSSP 2.1 boot server node.

4. Reconfigure and start the IBM Virtual Shared Disk for the SP nodes that support the IBM Virtual Shared Disk configuration. You can do this by issuing the following IBM Virtual Shared Disk commands:

   ```
   /usr/lpp/csd/bin/cfgvsd -a
   /usr/lpp/csd/bin/startvsd -a
   ```

   You can validate that the IP address and host names are correctly specified by issuing the following command:

   ```
   /usr/lpp/csd/bin/vsdatalst -n
   ```

5. If there were modifications made to any SP node system files, it would be best to reboot the PSSP 2.1 nodes to reflect the IP address and host name changes. When the SP nodes are initialized, your SP system should be activated using the new IP addresses and host names.

Execution on PSSP 1.2 Nodes

The following steps are performed on each PSSP 1.2 node to validate the IP address and host name updates for the control workstation and SP nodes. Perform the following verification checks on each of the SP nodes:

1. After the network boot is complete and the SP nodes are initialized, verify that they are using the correct IP address and host name.
   - Updated Ethernet and additional adapters are using the correct address for each SP node
   - Updated default routes are correct

- Host name resolution files were updated correctly (**/etc/hosts**, **dns**)
- The **/etc/SDR_dest_info** file points to the control workstation and the SDR
- The acct_master host name is updated for the NFS export list for the **/var/adm/acct** directory (SP Accounting)
- The **/etc/ssp/cw_name**, **/etc/ssp/server_hostname**, and **/etc/ssp/server_name** files point to the correct IP address and host names
- The **/etc/krb.conf**, **/etc/krb.realms**, and **/etc/krb-srvtab** files have the correct host names defined.

2. You may need to modify the following files on the SP nodes to reflect the updated IP addresses and host name definitions:

   /etc/filesystems
   /etc/ntp.conf
   /etc/resolv.conf
   /.rhosts
   /.klogin

   If any of the files are not correct, make the proper updates for the correct IP address and host name manually.

3. You may have to reset the routing tables after PSSP 1.2 node customization. You should verify that the current route tables are correct by issuing **netstat -nr**. If they are not correct, you can *flush* the route table and remove the route stanzas from the ODM. You can do this using the SMIT TCPIP panels or by issuing the **route -n -f** and **odmdelete** commands.

   Make sure that you re-create the proper default route tables using SMIT TCPIP or the **mktcpip** command.

4. If the node positions were changed, make sure that the correct node number is listed for the PSSP 1.2 boot server node. If the node number in the SDR is not listed as a boot server node, the **setup_server** command will remove required boot server files from the **/usr/sys/inst.images/ssp** directory.

   Make sure that the node number on the PSSP 1.2 boot server node matches the boot server node number listed in the SDR. To do this, on the PSSP 1.2 node, enter:

   ```
   odmget -q attribute=node_number CuAt
   ```

   Compare the value to the node number listed in the SDR using the **splstdata -b** command.

   If the node numbers are different, you will need to update the SDR or the ODM to reflect the proper node number. You may need to use the **odme** editor or work with the **odmdelete** and **odmadd** commands.

5. We recommend issuing the **savebase** command to save your current ODM configuration to the boot device.

   If any manual updates were required to reflect the new IP address and host name changes, you should reboot the SP node. When the SP nodes are initialized, the SP system should now be using the new IP address and host names.

Glossary of 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. 1995, 1997

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.

AIX Version 3. The IBM AIX Version 3 for RISC System/6000 (also known as AIX/6000). It is based on UNIX System V, conforms with POSIX IEEE Standard 1003.1, is compatible with Berkeley Software Distribution 4.3 (4.3 BSD), and is designed to meet Trusted Computing Base level C2 security.

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 using an application program.

ARP. Address Resolution Protocol.

ATM. Asynchronous Transfer Mode. (See *TURBOWAYS 100 ATM Adapter*.)

Authentication. The process of validating the identity of 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.

CDE. Common Desktop Environment. A graphical user interface for UNIX.

charge feature. An optional feature for either software or hardware for which there is a charge.

CLI. Command Line Interface.

client. * (1) A function that requests services from a server and makes them available to the user. * (2) A term used in an environment to identify a machine that uses the resources of the network.

Client Input/Output Sockets (CLIO/S). A 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.

css. Communication subsystem.

D

daemon. A process, not associated with a particular user, that performs system-wide functions such as administration and control of networks, execution of time-dependent activities, line printer spooling and so forth.

DASD. Direct Access Storage Device. Storage for input/output data.

DCE. Distributed Computing Environment.

dependent node. A type of extension node that depends on standard SP nodes for certain functions, but implements much of the switch-related protocol that standard nodes use on the SP Switch.

DFS. distributed file system. A subset of the IBM Distributed Computing Environment.

DNS. Domain Name Service. A hierarchical name service which maps high level machine names to IP addresses.

E

Error Notification Object. An object in the SDR that is matched with an error log entry. When an error log entry occurs that matches the Notification Object, a user-specified action is taken.

ESCON. Enterprise Systems Connection. The ESCON channel connection allows the RS/6000 to communicate directly with a host System/390; the host operating system views the system unit as a control unit.

Ethernet. (1) Ethernet is the standard hardware for TCP/IP local area networks in the UNIX marketplace. It is a 10-megabit per second baseband type LAN that allows multiple stations to access the transmission medium at will without prior coordination, avoids contention by using carrier sense and deference, and resolves contention by collision detection (CSMA/CD). (2) A passive coaxial cable whose interconnections contain devices or components, or both, that are all active. It uses CSMA/CD technology to provide a best-effort delivery system.

Ethernet network. A baseband LAN with a bus topology in which messages are broadcast on a coaxial cabling using the carrier sense multiple access/collision detection (CSMA/CD) transmission method.

event. In Event Management, the notification that a predicate evaluated to true. This evaluation occurs each time an instance of a resource variable is observed.

expect. Programmed dialogue with interactive programs.

extension node. A non-standard node that extends the SP system's capabilities or scope, but can't be used in all the same ways as standard SP nodes.

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 IBM Virtual Shared Disk. The device driver lets application programs stripe data across physical disks in multiple IBM Virtual Shared Disks, thus reducing I/O bottlenecks and hot spots.

I

instance vector. In Event Management, a set of elements, where each element is a name/value pair of the form `name=value`, whose values uniquely identify the copy of the resource (and by extension, the copy of the resource variable) in the system.

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 on frames that contain nodes.

NSB. Node Switch Board.

NTP. Network Time Protocol.

O

ODM. Object Data Manager. In AIX, a hierarchical object-oriented database for configuration data.

P

parallel environment. 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.

perspective. The primary window for each SP Perspectives application, so called because it provides a unique view of an SP system.

POE. Formerly Parallel Operating Environment, now Parallel Environment for AIX.

port. (1) An end point for communication between devices, generally referring to physical connection. (2)

A 16-bit number identifying a particular TCP or UDP resource within a given TCP/IP node.

predicate. In Event Management, the relational expression between a resource variable and other elements (such as constants or the previous value of an instance of the variable) that, when true, generates an event. An example of a predicate is `X < 10` where X represents the resource variable `IBM.PSSP.aixos.PagSp.%totalfree` (the percentage of total free paging space). When the predicate is true, that is, when the total free paging space is observed to be less than 10%, the Event Management subsystem generates an event to notify the appropriate application.

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.

rearm predicate. In Event Management, a predicate used to generate an event that alternates with an original event predicate in the following way: the event predicate is used until it is true, then the rearm predicate is used until it is true, then the event predicate is used, and so on. The rearm predicate is commonly the inverse of the event predicate (for example, a resource variable is on or off). It can also be used with the event predicate to define an upper and lower boundary for a condition of interest.

remote host. *See foreign host*.

resource. In Event Management, an entity in the system that provides a set of services. Examples of resources include hardware entities such as processors, disk drives, memory, and adapters, and software entities such as database applications, processes, and file systems. Each resource in the system has one or more attributes that define the state of the resource.

resource monitor. A program that supplies information about resources in the system. It can be a command, a daemon, or part of an application or subsystem that manages any type of system resource.

resource variable. In Event Management, the representation of an attribute of a resource. An example of a resource variable is `IBM.AIX.PagSp.%totalfree`, which represents the percentage of total free paging space. `IBM.AIX.PagSp` specifies the resource name and `%totalfree` specifies the resource attribute.

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.

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 nonoverlapping nodes on a switch chip boundary that act as a logical SP system.

T

tar. Tape ARchive, is a standard UNIX data archive utility for storing data on tape media.

Tcl. Tool Command Language.

TclX. Tool Command Language Extended.

TCP. Acronym for Transmission Control Protocol, a stream communication protocol that includes error recovery and flow control.

TCP/IP. Acronym for Transmission Control Protocol/Internet Protocol, a suite of protocols designed to allow communication between networks regardless of the technologies implemented in each network. TCP provides a reliable host-to-host protocol between hosts in packet-switched communications networks and in interconnected systems of such networks. It assumes that the underlying protocol is the Internet Protocol.

Telnet. Terminal Emulation Protocol, a TCP/IP application protocol that allows interactive access to foreign hosts.

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.

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, IBM. The function that allows application programs executing at different nodes of a system partition to access a raw logical volume as if it were local at each of the nodes. In actuality, the logical volume is local at only one of the nodes (the server node).

W

workstation. * (1) A configuration of input/output equipment at which an operator works. * (2) A terminal or microcomputer, usually one that is connected to a mainframe or to a network, at which a user can perform applications.

X

X Window System. A graphical user interface product.

Index

Special Characters

/etc/.profile file 53
/etc/aliases file 132
/etc/cshutSeq
  definition of 20
  example 28
  if empty 20
  if none exists 20
  syntax of 26
/etc/cstartSeq
  definition of 20
  example 28
  if empty 20
  if none exists 20
  syntax of 26
/etc/environment file 54, 55
/etc/ha/cfg directory
  EMCDB version string files 375
  run-time EMCDB files 375
/etc/inittab file
  use by Event Management 378
  use by Group Services 359
/etc/krb.conf file 175
/etc/krb.realms file 175
/etc/ntp.conf file 99
/etc/rc.tcpip file 132
/etc/SDR_dest_info file 578
/etc/sendmail.cf file 130
/etc/services file 174
  use by Event Management 369, 370
  use by Group Services 356
/etc/subsysSeq
  definition of 20
  example 29
  if none exists or if empty 20
  syntax of 26
/etc/syslog.conf file 301
/spdata/sys1/ha/cfg directory
  staging area for EMCDB files 373
/usr/lib/X11/app-defaults/SPmon file 547
/usr/local/fonts file 58
/usr/lpp/ssp/bin/newpass.log file 70
/usr/lpp/ssp/config/cron_template 148
/usr/lpp/ssp/install/config/haemloadlist file
  listing resource variable information 569
/var file system
  and Event Management tracing 378
  and Group Services tracing 359
/var/adm/acct/nite directory 150
/var/adm/cacct/ directory 151
/var/adm/cacct/fiscal directory 152
/var/adm/cacct/sum directory 151
/var/adm/SPlogs/cs/cshut.*MMDDhhmmss.pid* 25
/var/adm/SPlogs/cs/cstart.*MMDDhhmmss.pid* 25
/var/ha/lck directory
  Event Management lock files 374
  Group Services lock files 357
/var/ha/log directory
  Event Management log files 374
  Group Services log files 357
/var/ha/run directory
  Event Manager daemon working files 374
  Group Services daemon working files 357
  registration cache 374
  resource monitor working directories 374
/var/ha/soc directory
  Event Management socket files 373
  Group Services socket files 356
/var/sysman/file.collections file 42
/var/sysman/sup/.resident file 42
/var/sysman/sup/lists file 42
/var/sysman/sup/node.root/prefix file 55
/var/sysman/sup/user.admin/list file 55
/var/sysman/sup/user.admin/prefix file 55
/var/sysman/sup.resident file 59
.forward file 132, 136
.klogin file 171
.profile 63
.resident file 59
.rhosts file 171

Numerics

604 High Node
  time to power on 21
604 High Nodes 6

A

access to the configuration data 577
accounting 10
  accounting
  additional setup tasks 147
  by node or by class 141
  differences between SP and standard AIX 141
  entering node accounting information 146
  entering system information 144
  files 148
  for exclusive use of node jobs 142
  HACWS and 344
  installation 147
  report and summary files 149

© Copyright IBM Corp. 1995, 1997

accounting *(continued)*
setting up SP 143
system partitioning 144
template for crontabs 148
Active Message
C example 470
definition of function 459
Fortran example 474
infrastructure 454
writing handlers 454
writing handlers example 455
adapter failures 339
Adapter Membership
resource monitor 369
adapters, extension node
adding 271
changing 273
listing 277
reconfiguring 275
removing 274
adapters, network connectivity 8
adding and deleting files in a file collection 54
adding extension nodes 271
adding subsystems
Event Management (haemctrl) 377
Group Services (hagsctrl) 359
adding user accounts 70
administrative tasks, scheduling 421
ADSM
tuning for 445
AFS
adding principals to 184
administering AFS authentication services 193
changing passwords in 170
AFS authentication services 193
afsclient.cust 503
AIX 9
AIX error log
Group Services entries 361
use by Event Manager daemon 374
AIX error log facility
configuring 411
SMIT, from 411
creating SNMP traps from 402
generating reports 411
managing 410
SMIT, from 410
trimming logs 411
SMIT, from 411
AIX operating system resources
resource monitor 369
AIX operating system statistics
obtained from SPMI 376
aixos resource monitor 369
alternate boot system image
enabling 495
alternate boot system image *(continued)*
using 495
Amd map files
converting 123
amd.u map file
converting 123
APIs
Communications Low-Level Application Programming Interface (LAPI) 2
LAPI 2
applications
Ascend GRF 269
authentication
adding IBM RS/6000 workstations to your realm 173
adding principals 182
administering SP authentication services
adding principals 182
changing admin instance password 186
changing passwords 185
setting passwords 182
administration roles 172
AFS
adding principals 184
administering 193
changing passwords in 170
AFS services 193
authentication client system 178
authorization as an authentication database administrator 185
backing up database 179
changing admin instance password 186
changing passwords 185
changing your password 170
configuration files 174
credentials 166
destroying tickets 169
directory paths for services 178
HACWS and 345
identifying yourself 166
implementations 173
listing tickets 169
managing information for authenticated services
changing service keys 190
managing the database 179
primary server 176
protecting your database 172
realms 164
remote commands 171
secondary servers 173
service keys 189
services on the SP 157, 170
services, using 166
setting passwords 182
setting up a secondary server 177
setting up an authentication client system 178

authentication *(continued)*
terminology 158
understanding 157
using the SP authentication services
changing your password 170
destroying tickets 169
identifying yourself 166
listing tickets 169
authentication, SDR 577
authority levels for SP System Monitor
control authority 299
monitor authority 299
authorization, SDR 577
automount daemon
restarting 112
stopping 112
automounter
coexistence considerations 121
control of other file systems 109
customizing function 112
files 112
diagnostics 108
file collections 108
HACWS and 343
managing 103
map file 110
creating 110
map files 105, 111
adding to file collections 111
distributing to boot/install servers and nodes 111
map for users' directories 108
master map file 110
updating 110
master map files 107
migration considerations 122
requesting support 104
understanding 103
automounter support 2
availability 10

B

backing up authentication database 179
basecode mode hardware variables for the SP Switch 568
battery, SP virtual 495
boot/install server
distributing automounter map files to 111
sequencing of startup and shutdown 19
bootlist, setting 496, 497
bootptab 508
BSD syslog facility
generating reports 414
managing 413
building a file collection 56

C

cache, registration
Event Management 374
cardtype 17 and 20 unique frame hardware variables 550
cfgauto.cust script 113
change management 10
changing extension nodes 273
changing user accounts 71
changing your password 72
characteristics of SP system
checkauto.cust script 116
chkp 187
cleaning subsystems
Event Management (haemctrl) 378
Group Services (hagsctrl) 359
client
type of resource monitor 368
client communication
with Event Management subsystem 367
with Group Services subsystem 355
client, Group Services
definition 353
CLIO/S, tuning 447
closing windows to end the display manager session 303
cluster
and Event Management 367
cluster workstation
and Event Management 367
cmonacct 142
coexistence support 3
collecting system data for IBM service 416
command
type of resource monitor 368
command files, nrunacct 149
command line, from the 412
command line, using the 246, 248, 250, 253, 254, 256, 257, 258
commands
cshutdown 23
cstartup command 20
df 581
dsh 31
dspcat
displaying event management descriptions 572
dspmsg
displaying event management descriptions 572
Eprimary
Estart
hostlist 33
jm_config 89
jm_start 92
jm_status 96
jm_stop 95

commands *(continued)*
jm_verify 88
locate_jm 96
lscfg 581
netstat 581
omit 64
omitany 64
p_cat 35
passwd 73
pcp 35
pdf 37
pexec 34
pfind 35
pfps 37
pls 35
pmv 35
ppred 36
pps 35
prm 35
SDRGetObjects
listing resource variable information 569
output example 571
spchuser 67
spevent
listing resource variable information 572
output example 573
splsuser 67
spmkuser 67
sprmuser 67
supper activate 46
supper install 48
supper offline 61
supper reset 47
supper scan 48
supper status 43
supper update 48
symlinkall 65
yppasswd 73
communication subsystem support 10
communication, client
with Event Management subsystem 367
with Group Services subsystem 355
Communications Low-level Application Programming Interface (LAPI) 453
communications, Event Management
between Event Manager daemons 369
local EM clients 370
remote EM clients 370
resource monitors 370
communications, Group Services
between Group Services daemons 355
local GS clients 356
configuration data
activating 385
changing in the SDR 384
for third-party resource monitors 371, 373, 384
configuration data (SDR)
component of Event Management 370
configuration database
component of Event Management 372
configuration load utility
component of Event Management 373
configuration management 9
configuration utility
component of Event Management 371
configuring resource manager 86
configuring the system monitor 300
configuring the user home environment 301
connectivity adapters, network 8
connectivity, switch 231
control script
component of Event Management 372
component of Group Services 356
control workstation 8
booting for HACWS 338
failovers 340
high availability 333
preparing
controlling login 75
core file
Event Manager daemon 374
Group Services daemon 357
cprdaily 142
create class command 203
create proc command 203
create var command 203
creating custom layouts
credentials
deleting 169
displaying 169
managing 166
obtaining 166
crontabs file 45, 153
crontabs, accounting template for 148
cshutdown command
log file 25
using 23
CSS MTU size, tuning 431
CSS subsystem
resource monitor 368
cssMembership group 369
use by Event Management 376
cstartup command
log file 25
safety warning 20
customizing
automounter function 112
system partitions 251

D

daemon
  type of resource monitor 368
daemons 499
  sendmail 128, 131, 132
  srcmstr 139
data files 149
default configuration data
  component of Event Management 373
  loading into the SDR with haemloadcfg 373
  location for Event Management 373
default predicate
  changing 371
default system partition 241
deleting subsystems
  Event Management (haemctrl) 378
  Group Services (hagsctrl) 359
deleting user accounts 72
dependent nodes
  switch considerations 224
descriptions, displaying event management
  using the dspcat command 572
  using the dspmsg command 572
determining if nodes are operating 322
determining switch connectivity 231
df command 581
diagnostics
  automounter 108
dialog boxes of SP Perspectives 289
disks for HACWS 335
displaying event management descriptions
  using the dspcat command 572
  using the dspmsg command 572
displaying frame and node layouts 307
displaying the configuration data 580
distributed shell 31
distribution of mail-related files 139
domain, operational
  for Event Management 366
  for Group Services 354
dsh 31
dsh based 34
dsh command 31
dspcat command
  displaying event management descriptions 572
dspmsg command
  displaying event management descriptions 572

E

EM client (Event Management client)
  definition 365
  registration request cache 374
  restrictions on number 381
EM_Instance_Vector SDR class
  key attributes for loading 373
  use by Event Management 370
EM_Resource_Class SDR class
  key attributes for loading 373
  use by Event Management 371
EM_Resource_Monitor SDR class
  key attributes for loading 373
  use by Event Management 371
EM_Resource_Variable SDR class
  key attributes for loading 373
  use by Event Management 370
EM_Structured_Byte_String SDR class
  key attributes for loading 373
  use by Event Management 370
EMAPI (Event Management Application Programming Interface)
  component of Event Management 367
EMAPI libraries
  location 367
embedded resource monitor
  type of resource monitor 368
EMCDB (Event Management Configuration Database)
  and Event Manager daemon initialization 380
  compiling with haemcfg 371
  placing a new version in production 381
  run-time directory 375
  run-time file 372
  staging directory 373
  staging file 371
  version string 371
  version string and joining peer group 380
  version string and peer group state 376
  version string directory 375
Emonitor 235
Emonitor subsystem
  and system partitioning 260
en_pend method flow 412
enMembership group 369
  use by Event Management 376
Eprimary command
Equiesce 236
error log management
  archiving error logs 415
  collecting system data for IBM service 416
  creating archives 416
  creating service collections 416
  gathering archives 416
  gathering service collections 417
  installing and configuring 409
  removing archives 416
  removing service collections 417
  viewing error logs 415
error logs
  archiving 415
  creating archives 416

error logs *(continued)*
gathering archives 416
managing 409
removing archives 416
viewing 415
error message output log
Event Management 374
error notification objects, managing 411
error templates
managing 413
ESCON
tuning for 447
Estart command
Event Management
abnormal termination of haemctrl add 378
disk space and tracing 378
performance and tracing 378
shared variables with performance monitoring 376
Event Management API (EMAPI)
component of Event Management 367
Event Management client (EM client)
definition 365
Event Management communications
between Event Manager daemons 369
local EM clients 370
remote EM clients 370
resource monitors 370
Event Management Conditions subsystem
and system partitioning 260
Event Management configuration
changing 371
SDR classes 370
Event Management Configuration Database (EMCDB)
and Event Manager daemon initialization 380
compiling with haemcfg 371
creating 372
placing a new version in production 381
run-time directory 375
staging directory 373
version string and joining peer group 380
version string and peer group state 376
version string directory 375
Event Management directories
/etc/ha/cfg 375
/spdata/sys1/ha/cfg 373
/var/ha/lck 374
/var/ha/log 374
/var/ha/run 374
/var/ha/soc 373
Event Management SDR classes
key attributes for loading 373
Event Management subsystem
activating configuration data 385
adding with haemctrl command 377
and system partitioning 260
changing configuration data in the SDR 384
Event Management subsystem *(continued)*
cleaning with haemctrl command 378
client communication 367
component summary 365
components 365—375
configuration 377—378
configuring and operating 365—385
deleting with haemctrl command 378
dependencies 375—376
Event Manager daemon initialization 379—381
Event Manager daemon operation 381—382
getting subsystem status 382—384
installation 377
introducing 365
loading non-PSSP data into the SDR 384
operational domain 366
recovery from failure (automatic) 381
starting with haemctrl command 378
stopping with haemctrl command 378
tracing with haemctrl command 378
Event Manager daemon
abnormal termination core file 374
communications 369
component of Event Management 366
current working directory 374
definition 365
error message log file 374
getting status 382—384
group membership 376
group subscriptions 376
haemctrl control script 372
initialization 379—381
message trace output log file 374
operation 381—382
recovery from failure (automatic) 381
trace output log file 374
use of AIX Error Log 374
event processing 348
examples
resource data
output from the SDRGetObjects command 571
using SP Perspectives 573
expression
synonymous with predicate 572
extension nodes
adding 271
changing 273
implementations of 269
listing 277
managing 269
reconfiguring 275
removing 274
resetting 276
SNMP 271
switch considerations 224
transferring configuration information for 271

extension nodes *(continued)*
understanding 269

F

failovers of control workstation 340
fencing nodes 233
file collection
adding and deleting files 54
adding automounter map to 111
automounter 108
building 56
HACWS and 342
installing 59
from a specific server
managing 41
modifying the hierarchy 61
organization 45
refusing files 61
removing 63
structure 45
tasks
updating 52
using scan to verify 51
file management 10
files 108
/etc/.profile 53
/etc/aliases 132
/etc/environment 54, 55
/etc/ntp.conf 99
/etc/rc.tcpip 132
/etc/sendmail.cf 130
/etc/syslog.conf 301
/usr/lib/X11/app-defaults/SPmon 547
/usr/local/fonts 58
/usr/lpp/ssp/bin/newpass.log 70
/usr/lpp/ssp/config/cron_template 148
/var/sysman/file.collections 42
/var/sysman/sup/.resident 42
/var/sysman/sup/lists 42
/var/sysman/sup/node.root/prefix 55
/var/sysman/sup/user.admin/list 55
/var/sysman/sup/user.admin/prefix 55
/var/sysman/sup.resident 59
.forward 132, 136
.profile 63
.resident 59
accounting report and summary 149
afsclient.cust 503
crontabs 45
ibmSPDepNode.my 509
ibmSPMIB.my 516
jmd_config 537
last 48
list 54
logs 53
files *(continued)*
prefix 54
refuse 61
sendmail.cf 127
when 48
files and directories
component of Event Management 373
component of Group Services 356
files, accounting 148
files, data 149
files, nrunacct command 149
filesets, Event Management
ssp.clients 377
ssp.ha 377
filesets, Group Services
ssp.ha 358
first boot customization file (script.cust)
sample file 540
first boot customization file (tuning.cust)
sample file 545
frame hardware variables 549, 562
frame supervisor cables 350
frame supervisor cards for HACWS 336
frames 7
from SMIT 411, 413
from the command line 412

G

group membership list
definition 353
Group Services
abnormal termination of hagsctrl add 359
disk space and tracing 359
performance and tracing 359
Group Services API (GSAPI)
component of Group Services 355
Group Services client
definition 353
Group Services communications
between Group Services daemons 355
local GS clients 356
Group Services daemon
abnormal termination core file 357
communications 355
component of Group Services 354
current working directory 357
getting status 362—363
hagsctrl control script 356
initialization 360—361
operation 361—362
recovery from failure (automatic) 361
trace output log file 357
Group Services directories
/var/ha/lck 357
/var/ha/log 357

Group Services directories *(continued)*
/var/ha/run 357
/var/ha/soc 356
Group Services subsystem
adding with hagsctrl command 359
and Event Manager daemon initialization 379
and system partitioning 259
cleaning with hagsctrl command 359
client communication 355
component summary 354
components 354—357
configuration 358—359
configuring and operating 353—363
deleting with hagsctrl command 359
dependencies 357—358
dependency by Event Management 376
getting subsystem status 362—363
Group Services daemon initialization 360—361
Group Services daemon operation 361—362
initialization errors 361
installation 358
introducing 353
operational domain 354
recovery from failure (automatic) 361
starting with hagsctrl command 359
stopping with hagsctrl command 359
system groups 369
tracing with hagsctrl command 359
group state
and joining the ha_em_peers group 380
EMCDB version string 376
group state value
definition 353
group, Group Services
definition 353
groups
Group Services
restrictions on number per client 361
restrictions on number per domain 361
GS client (Group Services client)
restrictions on number 361
GS nameserver
establishing 360
GSAPI (Group Services Application Programming Interface)
component of Group Services 355
GSAPI libraries
location 355

H

ha_em_peers group
joining 379
use by Event Management 376
HACMP/6000
configuring resources for 348
HACMP/6000 *(continued)*
customizing for HACWS 348
event processing 348
HACWS foundation 336
operating for the control workstation 338
HACWS
See also High Availability Control Workstation (HACWS)
managing 333
HACWS (High Availability Control Workstation)
and Event Management 378
and Group Services 359
haemcfg command
compiling the EMCDB 371
haemctrl command
adding the Event Management subsystem 377
cleaning the Event Management subsystem 378
control script for Event Management 372
deleting the Event Management subsystem 378
starting the Event Management subsystem 378
stopping the Event Management subsystem 378
summary of functions 377
tracing the Event Management subsystem 378
haemd daemon
location 366
haemloadcfg command
differences from SDRCreateObjects 373
Event Management configuration load utility 373
use of loadsdr program 373
haemloadlist file
default Event Management configuration data 373
haemrcpcdb script
use by Event Management 381
hagsctrl command
adding the Group Services subsystem 359
cleaning the Group Services subsystem 359
control script for Group Services 356
deleting the Group Services subsystem 359
starting the Group Services subsystem 359
stopping the Group Services subsystem 359
summary of functions 358
tracing the Group Services subsystem 359
hagsd daemon
location 354
halting the SP 23
hardware monitor resource monitor 368
hardware overview 5
Heartbeat subsystem
and system partitioning 259
help for SP Perspectives 293
High Availability Control Workstation (HACWS)
and Event Management 378
and Group Services 359
configuration changes 341
customizing HACMP/6000 for 348
failovers of control workstations 340

High Availability Control Workstation (HACWS) *(continued)*
HACWS and 347
hardware 335
managing subsystems for 342, 343, 344, 345, 346, 347
accounting 344
automounter 343
dependent nodes 347
extension nodes 347
file collections 342
High Performance Switch 345
mail service 344
parallel management commands 342
Resource Manager 343
security features 345
SP Switch 346
starting up and shutting down system 342
Sysctl 345
system partitions 346
time synchronization 343
user accounts 343
software 336
terminology 334
understanding 334
high availability services
Event Management subsystem 365
Group Services subsystem 353, 365
Topology Services subsystem 353, 365
high performance switch 7
HACWS and 345
High Performance Switch variables 564
high-performance switch
configuring
highlights of PSSP software 9
Host Membership
resource monitor 369
host name changes for SP systems 611
host_responds
updating 583
Host_Responds daemon
dependency by Event Management 376
host_responds SDR class
and Event Management 376
resource monitor 369
Host_Responds subsystem
and system partitioning 260
hostlist command 33
HostMembership group 369
use by Event Management 376
hr script
Host_Responds daemon 376

I

IBM parallel system support programs for AIX (PSSP) 9
IBM products, other 15
IBM.PSSP.harmld resource monitor 368
IBM.PSSP.harmpd resource monitor 368
IBM.PSSP.hmrmd resource monitor 368
IBM.PSSP.Membership.LANAdapter.state resource variable 376
IBM.PSSP.Membership.Node.state resource variable 376
IBM.PSSP.pm.User_state resource variables 395
IBM.PSSP.pmanrmd resource monitor 368
IBM.PSSP.Response.Host.state resource variable 376
IBM.PSSP.Response.Switch.state resource variable 376
ibmSPDepNode.my 509
ibmSPMIB.my 516
include command 202
independent software vendor products 15
inetd.conf file 174
installation 9
Event Management subsystem 377
Group Services subsystem 358
installing a file collection 59
installing a file collection from a specific server
installing resource manager 84
instance vector
configuration in SDR 370
integrating mail services 127
using /etc/hosts 136
using DNS 128
using NIS 132
integrating print services
internal resource monitor
definition 368
supplied by PSSP 368
introduction 1
IP address changes for SP systems 611
IP router
isolating nodes 322

J

jm_config command 89
jm_start command 92
jm_status 96
jm_stop 95
jm_verify command 88
jmd_config file 537
job management 10
joining a Group Services group
Event Management subsystem 379

K

Kerberos 160
  client and server interaction 160
  DES 162
  management improvements 3
  password, master 179
  principals 162
  restarting 192
  stopping 192
  tickets 164
key attributes
  Event Management SDR objects 373
keys
  changing 190
  creating 189
  displaying information for 189
  distributing to client systems 191
  managing 189

L

LAPI
  Active Message 459
    C example 470
    Fortran example 474
    infrastructure 454
    writing handlers example 455
  address manipulation 461
  advantages 453
  allocating buffers 464
  blocking and non-blocking calls 458
  characteristics of functions 459
  communication behaviors 458
  compiling programs 464
  completion checking 460
  completion of communication operation 458
  concepts 457
  core dumps 467
  data transfer 459
  defined functions 457
  environment 461
  error handling 459, 461
  executing programs 464
  execution model 461
  general functions 454
  Get
    C example 478
    Fortran example 480
  hangs 468
  hints 466, 467
  linking with libraries 468
  message ordering and atomicity 458
  messages 461
  new feature in PSSP 2.3 2
  non-thread-safe-libraries 468
LAPI *(continued)*
  origin and target 457
  overview 453
  programming examples 470
  progress 459, 461
  PUT
    C example 482
    Fortran example 484
  Rmw
    C example 487
    Fortran example 489
  running programs 466
  segment registers 468
  setup 461
  specific functions 459
  synchronizing 460
  understanding 454
  using 459
  why use 453
  writing handlers 454
LAPI_Address 461
LAPI_Address_init: 461
LAPI_Amsend 459
LAPI_Getcntr 460
LAPI_Init 461
LAPI_Msg_String 461
LAPI_Probe 461
LAPI_Qenv 461
LAPI_Rmw 460
LAPI_Senv 461
LAPI_Setcntr 460
LAPI_Term 461
LAPI_Waitcntr 460
large data transfers, tuning for 443
last file 48
Launch Pad
  bringing up 282
  customizing 284
  description of 283
  elements of the 284
line, from the command 412
line, using the command 246, 248, 250, 253, 254, 256, 257, 258
list file 54
list file keywords and operands 64
list node file system information 581
list node hardware information 581
list node network information 581
listing a user account 72
listing extension nodes 277
listing resource variable information
  from SP Perspectives 572
  using load list files 569
  using the SDRGetObjects command 569
load command 204

LoadLeveler
  resource monitor 368
loadsdr program
  use by haemloadcfg command 373
local EM clients
  Event Management communications 370
local GS clients
  Group Services communications 356
locate_jm 96
lock file
  Event Management 374
  Group Services 357
log file
  Event Management 374
  Group Services 357
log_event 393
login control 75
logs file 53
lowclust, tuning 430
lowmbuf, tuning 430
lsautoent.cust script 120
lscfg command 581
lskp 180
lssrc command
  getting Event Management status 382
  getting Group Services status 362

M

mail 127
mail services 127
  HACWS and 344
mail.aliases map 132
maintaining the crontabs file 153
Management Information Base
  SP 404
management, configuration 9
management, user 10
managing
  file collection 41
  user accounts 67
managing error notification objects 411
managing mail 127
managing user information 67
manual pages for public code xxviii
map files for automounter
  adding to file collections 111
  creating 110
  distributing to boot/install servers and nodes 111
master map files for automounter
  updating 110
mb_cl_hiwat, tuning 430
Membership resource monitor 369
message boxes of SP Perspectives 292
message trace output log
  Event Management 374
MIB (Management Information Base)
  understanding 400
migration support 3
mkautoent.cust script 118
mkkp 184
modifying the file collection hierarchy 61
monitoring and operating the SP system 297

N

Nagle algorithm 438
nameserver, Group Services
  establishing 360
netstat command 581
Netview for AIX
  configuring for the SP system 407
network boot 320
network connections for HACWS 336
network connectivity adapters 8
Network File Sustem (NFS)
  tuning 443
Network File System (NFS)
network tuning options
  changing on control workstation 423
  changing on selected nodes 424
  changing on the node 423
new features of PSSP
NFS
  See also Network File System (NFS)
  mounts over a switch 237
node groups
  commands for creating, modifying, and deleting 265
  commands for examining 265
  commands that take node groups as
    arguments 266
  creating 264
  creating, modifying, and deleting 265
  definition list 263
  deleting 266
  examining 265, 266
  example of resolving 265
  flattening 264
  global 263
  managing 263
  modes 263
  names 263
  partition-bound 263
  resolving 264
  SDR information 263
  SMIT panels for creating, modifying, and
    deleting 265
  SMIT panels for deleting 266
  SMIT panels for examining 266
  understanding 263
node hardware variables 553, 555, 557, 559

node isolation 233, 322
node jobs, accounting for exclusive use of 142
node.root file collection 45
nodes
  dependent 2
  extension 2, 7
  Group Services
    restrictions on number per domain 361
  monitoring on a switch 235
  overview of sequencing 19
nodes, 604 High 6
nodes, primary 226
nodes, primary backup 226
nodes, processor 6
nodes, thin 6
nodes, wide 6
notebooks of SP Perspectives 290
notifcation object, installing a 412
notification objects, managing error 411
notify_event 393
notifying users of impending shutdown 24
nrunacct command files 149
NTP 99
NTP options 99

O

objects, managing error notification 411
observation interval
  changing 371
omit command 64
omitany command 64
operating frame environment controls 313
operating the front panel controls 313
operating the global controls 320
operation 10
overview of RS/6000 SP
overview, hardware 5
overview, software 9

P

p_cat 35
packaging 11
parallel management commands 31
  HACWS and 342
  sample 34
Parallel System Support Programs (PSSP)
  new features of 1
Parallel System Support Programs for AIX (PSSP),
 IBM 9
partitioning, system
  applying a configuration 254
  configuration 246, 248
    displaying information for 248
    selecting 246
partitioning, system *(continued)*
  HACWS and 346
  layouts 250
    selecting 250
  overview 242
  steps for 245
  understanding 239
partitions, system
  changed 256
    rebooting 256
  changing 253
  customizing 251
  default system partition 241
  displaying configuration information 256
  managing system partition-sensitive
   subsystems 259
  verifying configuration information 258
partititons, system
  managing 239
passwd command 73
passwork, Kerberos master 179
PCA card, tuning 447
pcp 35
pdf 37
peer group, Event Management
  joining 379
Performance Aide for AIX
  dependency by Event Management 376
performance monitoring
  shared variables with Event Management 376
Performance Toolbox for AIX
  dependency by Event Management 376
  use of shared resource variables 376
Performance Toolbox Parallel Extensions (PTPE)
  use of shared resource variables 376
perspective
  customizing a 288
  description of 285
  elements of 285
perspectives 282
pexec 34
pexscr 35
pfck 38
pfind 35
pfps 37
pls 35
pman 387
pmand
  controlling 389
  understanding 388
pmandef
  issuing an SNMP trap for an event 393
  logging an event 393
  purpose 390
  running a command for an event 391
  security considerations 390

pmandefaults 394
pmanquery 394
pmanrmd
  configuring 396
  controlling 395
  SDR information 397
  understanding 394
pmanrmd.conf 395, 396
pmanrmdloadSDR 397
pmanrminput 395
pmv 35
port numbers
  component of Event Management 369
  component of Group Services 355
power_system file collection 44
powering on the SP
  safety warning 20
pps 35
predicate
  definition 365
prefix file 54
primary authentication server 176
primary node takeover 222
primary windows of SP Perspectives 285
principals
  adding 171
  adding and assigning passwords 182
  deleting 188
  displaying information 181
  listing 180
  managing 180
  service 163
  types 162
  user 163
print management 125
  removal of SP print management system 4
prm 35
problem determination
  Event Management subsystem
    abnormal termination core file 374
    abnormal termination of haemctrl add 378
    getting subsystem status 382
    tracing 378
  Group Services subsystem
    abnormal termination core file 357
    abnormal termination of hagsctrl add 359
    getting subsystem status 362
    tracing 359
problem management 10
  resource monitor 368
Problem Management subsystem
  and system partitioning 260
  components of 388
  creating resource monitors 395
  creating subscriptions 390
  default events 394
Problem Management subsystem *(continued)*
  environment variables 392
  IBM.PSSP.pm.User_state resource variables 395
  log_event 393
  logging an event 393
  notify_event 393
  pmand 388
  pmandef 390
  pmandefaults 394
  pmanquery 394
  pmanrmd 394
  pmanrmd.conf 395
  pmanrmdloadSDR 397
  pmanrminput 395
  requesting information about a subscription 394
  resource monitor daemon 394
  resource variables and values 395
  running a command for an event 391
  SNMP trap, issuing 393
  understanding 387
process resource monitor 368
processor nodes 6
protocol, Group Services
  definition 353
protocol, switch communication 221
provider
  definition 353
PSSP 9, 11

Q

querying resource manager 96
quiescing the SP Switch 236

R

rcp 171
realms 164
reboot
  and Event Management 378
  and Group Services 359
rebooting the SP 23
reconfiguring
  Event Management subsystem 384
reconfiguring extension nodes 275
reconfiguring resource manager 94
recovery from failure
  Event Management 381
  Group Services 361
refauto.cust script 115
refuse file 61
refusing files in a file collection 61
registration cache
  and Event Manager daemon 379
  Event Management 374

remote EM clients
  Event Management communications 370
removing a file collection 63
removing extension nodes 274
repartitioning, SP Switch 235
report and summary files, accounting 149
reporting file collection information 49
required file collections 43
resetting extension nodes 276
resource class
  configuration in SDR 371
resource data, listing
  using the SDRGetObjects command 569
resource identifier
  synonymous with instance vector 572
resource manager 83
  configuring 86
  HACWS and 343
  installing 84
  querying 96
  reconfiguring 94
  recovery 97
  starting 88
  stopping 95
  understanding 83
resource monitor
  component of Event Management 368
  configuration in SDR 371
  connection types 368
  current working directory 374
  definition 365
  Event Management communications 370
Resource Monitor Application Programming Interface (RMAPI)
  component of Event Management 367
  use of EMCDB version string directory 375
  use of shared memory key 374
resource monitor, non-PSSP
  loading configuration data for 371, 373, 384
resource monitor, third-party
  loading configuration data for 371, 373, 384
resource monitors
  creating 395
resource monitors supplied by PSSP
  aixos (internal) 369
  IBM.PSSP.harmld 368
  IBM.PSSP.harmpd 368
  IBM.PSSP.hmrmd 368
  IBM.PSSP.pmanrmd 368
  Membership (internal) 369
  Response (internal) 369
resource variable
  configuration in SDR 370
  definition 365
  IBM.PSSP.Membership.LANAdapter.state 376
  IBM.PSSP.Membership.Node.state 376
resource variable *(continued)*
  IBM.PSSP.Response.Host.state 376
  IBM.PSSP.Response.Switch.state 376
resource variable information, displaying
  using the spevent command 572
resource variable, PSSP-supplied
  listing 569
Response resource monitor 369
restarting the SP 20, 23
restrictions
  Event Management 381
    number of EM clients 381
    observation of resource variable values 382
  Group Services 361
    groups per client 361
    groups per domain 361
    nodes per domain 361
    number of GS clients 361
rfc1323, tuning 431
RMAPI (Resource Monitor Application Programming Interface)
  component of Event Management 367
  use of EMCDB version string directory 375
  use of shared memory key 374
RMAPI libraries
  location 367
rmautoent.cust script 119
router nodes, tuning 443
RS/6000 SP
  characteristics of 1
  overview 5
rsh 171
run-time EMCDB
  creating 372, 380

S

sample files 503
  script.cust 540
  tuning.cust 545
sb_max, tuning 430
SBS (structured byte string)
  configuration in SDR 370
scheduling administrative tasks 421
script.cust
  sample file 540
SDR 575, 585
  archiving 245
  pmanrmd information 397
  shadow files 584
SDR (System Data Repository)
  and haemctrl clean 378
  and haemctrl delete 378
  and hagsctrl clean 359
  and hagsctrl delete 359
  dependency by Event Management 376

SDR (System Data Repository) *(continued)*
dependency by Group Services 357
key attributes for Event Management objects 373
use by Event Management 365
SDR authentication 577
SDR authorization 577
SDR daemon log 578
SDR server 577
SDRCreateObjects command
differences from haemloadcfg command 373
SDRGetObjects command
example of output 571
listing resource variable information 569
secondary authentication server 177
secondary servers 173
security 10, 157
selecting nodes 226
primary 226
primary backup 226
selecting primary and primary backup nodes 226
sendmail daemon 128, 131, 132
sendmail.cf file 127
SEPBU unique frame hardware variables 551
sequencing nodes
about 19
default 20, 27
defining 26
sequencing of nodes
overview 19
server
type of resource monitor 368
service collections
creating 416
gathering 417
removing 417
service, collecting system data for IBM 416
set command 204
setting time 101
setting up a primary authentication server 176
setting up a secondary authentication server 177
setting up SP accounting 143
shadow files 584
shared memory key
use by RMAPI 374
shutting down and rebooting the SP 23
shutting down the SP 23
checking the status of 25
notifying users before 24
overview 19
phases of shutdown processing 24
single point of control. 31
SMIT
See System Management Interface Tool (SMIT)
SMIT interface 69
SMIT interface access 578
SMIT, from 411, 413
SMIT, using 243, 245, 247, 248, 250, 251, 253, 255, 256, 258
SNMP
contents of traps 403
extension nodes 271
traps from AIX Error Log events 401
traps from Event Management events 402
understanding 400
SNMP Support subsystem
and extension nodes 271
and system partitioning 260
SNMP traps
pmandef to specify issuing 393
snmp_trap_gen 401
sockets
component of Event Management 369
component of Group Services 355
names used by Event Management 370
software overview 9
SP
overview of starting up and shutting down 19
powering on 20
restarting 20
starting 20
SP Ethernet
tuning 425
SP hardware state
resource monitor 368
SP Perspectives
administrative tasks 294
applications 283
authorizing users for 282
customizing primary windows 288
dialog boxes 289
example of displaying resource data 573
getting started 281
graphical elements 283
help for 293
Launch Pad 283
listing resource variable information 572
message boxes 292
notebooks 290
performance enhancements 4
perspectives 285
primary windows 285
starting 282
tasks that can be performed with 281
terminology differences from APIs 572
understanding the interface 283
usability enhancements 4
Using 281
SP switch
HACWS and 346
SP Switch repartitioning 235

SP Switch Router Adapter 269
SP Switch variables 565
SP system
  tuning 421, 426
SP System Monitor GUI 302
SP user management options 68
sp_configd 399, 401
SP_NAME environment variable
  and haemcfg command 371
  and haemctrl command 372
  and haemloadcfg command 373
  and hagsctrl command 356
SP_ports SDR class
  use by Event Management 370
spchuser command 67, 68, 71
  home attribute 69
  SMIT interface 69
special subsystems
  creating 29
spevent command
  example of output 573
  listing resource variable information 572
spimg 13
splsuser command 67, 72
  SMIT interface 69
SPMI (System Performance Monitor Interface)
  dependency by Event Management 376
SPMI library
  shared memory key 374
spmkuser command 67, 68, 70
  home attribute 69
  SMIT interface 69
sprmuser command 67, 72
  SMIT interface 69
SRC (System Resource Controller)
  and Event Manager daemon 379
  and Group Services daemon 360
  dependency by Group Services 357
srcmstr daemon 139
ssp.authent 13
ssp.basic 11
ssp.clients 13
ssp.clients fileset, PSSP
  and Event Management 367, 377
  and Group Services 358
ssp.csd 14
ssp.css 12
ssp.docs 13
ssp.gui 11
ssp.ha 12
ssp.ha fileset, PSSP
  and Event Management 377
  and Group Services 358
ssp.hacws 14
ssp.jm 12
ssp.perlpkg 13
ssp.pman 13
ssp.public 13
ssp.spmgr 12
ssp.st 12
ssp.sysctl 13
ssp.sysman 11
ssp.top 12
staging file for EMCDB
startauto.cust script 114
starting resource manager 88
starting subsystems
  Event Management (haemctrl) 378
  Group Services (hagsctrl) 359
starting the display manager interface 303
starting the SP 20
  overview 19
starting the switch 231
starting up the SP
  checking the status of 25
status, Event Management
  output of lssrc command 382—384
status, Group Services
  output of lssrc command 362—363
stopping resource manager 95
stopping subsystems
  Event Management (haemctrl) 378
  Group Services (hagsctrl) 359
structured byte string (SBS)
  configuration in SDR 370
structured field
  configuration in SDR 370
subscriber
  definition 353
subsystem
  Event Management 365—385
  Group Services 353—363
  system partition-sensitive 356, 372
subsystem status
  for Event Management 382—384
  for Group Services 362—363
subsystems
  special 29
  tuning 443
subsystems, system partition-sensitive
  Emonitor subsystem 260
  Event Management Conditions subsystem 260
  Event Management subsystem 260
  general structure 260
  Group Services subsystem 259
  Heartbeat subsystem 259
  Host_Responds subsystem 260
  listing 259
  managing 259
  Problem Management subsystem 260
  SNMP Support subsystem 260

subsystems, system partition-sensitive *(continued)*
summary 260
summary and report files, accounting 149
sup.admin file collection 44
supper activate command 46
supper install command 48
supper offline command 61
supper program 41
supper reset command 47
supper scan command 48
supper status command 43
supper subcommands table 49
supper update command 48
support, automounter 2
switch 221
clock API 4
clock reset 225
clocks 224
dependent nodes 224
extension nodes 224
setting switch clock source 225
SMIT panels 225
switch topology files 229
switch clocks
reset of 225
selecting 224
switch connectivity 231
switch protocol 221
switch queue, tuning 440
switch topology files
annotating 229
managing 229
switch variables 563
switch_responds SDR class
and Event Management 376
resource monitor 369
switch, high performance 7
switch, starting 231
switch, system partitioning 223
switch, using 221
switches
monitoring nodes on 235
quiescing 236
tuning 433
symlinkall command 65
synchronization, time 101
syntax rules 64
sysctl
application development
passing parameters to a script 214
sending a script 212
specifying server options 213
architecture 197
Authentication Services, relationship to 197
authorization
adding new callbacks 205
components 198
sysctl *(continued)*
authorization *(continued)*
levels 198
variables 211
built-in commands 216
commands
built-in sysctl commands 216
Extended Tcl command summary 215
Tcl command summary 215
components, illustration of 197
configuration
debugging problems 201
procedure 200
configuration commands
create class 203
create proc 203
create var 203
include 202
load 204
set 204
description of 195
executables
sysctl 199
sysctld 199
Extended Tcl command summary 215
files
sysctl.conf 199
sysctld.log 200
HACWS and 345
installation 200
Tcl command summary 215
using
directing commands to remote hosts 207
interactive mode 206
non-interactive mode 207
prerequisites 206
sending commands to a collection of nodes 208
sending commands to a server 207
Sysctl-based 36
Syspar Controller
managing system partition-sensitive subsystems 259
Syspar SDR class 371
syspar_ctrl command
and Event Management subsystem 372
and Group Services subsystem 356
listing system partition-sensitive subsystems 259
managing system partition-sensitive subsystems 259
Syspar_ports SDR class
use by Event Management 369
use by Group Services 355
System Data Repository 575, 585
default sequencing based on 20
System Data Repository (SDR)
and haemctrl clean 378

System Data Repository (SDR) *(continued)*
and haemctrl delete 378
and hagsctrl clean 359
and hagsctrl delete 359
archiving 245
configuration data for Event Management 370
dependency by Event Management 376
dependency by Group Services 357
key attributes for Event Management objects 373
use by Event Management 365
System Management Interface Tool (SMIT)
system partition
and Event Management 366
and Event Management SDR classes 371
and Group Services 354
specifying for haemcfg command 371
specifying for haemctrl command 372
specifying for haemloadcfg command 373
specifying for hagsctrl command 356
system partitioning and the switch 223
System Performance Monitor Interface (SPMI)
dependency by Event Management 376
System Resource Controller (SRC)
and Event Manager daemon 379
and Group Services daemon 360
dependency by Group Services 357
system resources
listing PSSP-supplied resource variables 569

T

takeover, primary node 222
TCP port
use by Event Management 369
TCP socket
Event Management client communication 367
tcp_mssdflt, tuning 431
tcp_recvspace, tuning 431
tcp_sendspace, tuning 431
terminology differences
between APIs and SP Perspectives 572
thewall, tuning 429
thin nodes 6
third-party resource monitor
loading configuration data for 371, 373, 384
tickets
cache files 165
lifetimes 165
understanding 164
time limits
Event Management
connection to Group Services 379
observation intervals 382
peer group joining 380
reconnection to resource monitors 381
Group Services
connection to Topology Services 360
time setting 101
time synchronization 101
HACWS and 343
Topology Services subsystem
and Group Services daemon initialization 360
and system partitioning 259
dependency by Group Services 357
trace output log
Event Management 374
Group Services 357
tracing subsystems
Event Management (haemctrl) 378
Group Services (hagsctrl) 359
trademarks xv
trimming log files 415
troubleshooting
Event Management subsystem
abnormal termination core file 374
abnormal termination of haemctrl add 378
getting subsystem status 382
tracing 378
Group Services subsystem
abnormal termination core file 357
abnormal termination of hagsctrl add 359
getting subsystem status 362
iniitalization errors 361
tracing 359
tuning
ADSM 445
AIX 3.2.5 considerations 427
CLIO/S 447
consolidated systems 442
ESCON 447
external server considerations 442
general recommendations 421
large data transfers 443
Network File System (NFS) 443
network options 2, 422
PCA card 447
problems 438
router nodes 443
single parent–multiple child applications 441
SP Ethernet 425
SP system-specific 426
specific workloads 432
subsystems 443
switch queue lengths 440
switches 433
system topology considerations 424
the SP system 421
user environments 432
tuning.cust
sample file 545

U

UDP port
  use by Event Management 369
  use by Group Services 355
udp_recvspace, tuning 431
udp_sendspace, tuning 431
understanding file collections 41
understanding resource manager 83
understanding the file collection structure 45
  directory and master files 46
understanding the system data repository 575
unfencing nodes 233
Unix domain socket
  Event Management client communication 367
  Group Services client communication 355
  resource monitor communication 367
  use by Event Management 369
  use by Group Services 355
user accounts
  adding 70
  changing 71
  deleting 72
  HACWS and 343
  list 72
  manage 67
user directories
  automounter map for 108
  changing 109
user environments, tuning 432
user management 10
user.admin file collection 44
using a switch 221
using SMIT 245, 247, 248, 250, 251, 253, 255, 256, 258
using the command line 246, 248, 250, 253, 254, 256, 257, 258
using the display manager interface 302

V

verifying file collections using scan 51
version string for EMCDB 371
VSD subsystem
  enhancements 3
  resource monitor 368

W

when file 48
wide nodes 6
window elements 303
workstation outside SP
  and Event Management 367
workstation, control 8

Y

yppasswd command 73

Communicating Your Comments to IBM

IBM Parallel System Support
Programs for AIX
Administration Guide

Publication No. GC23-3897-04

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 Parallel System Support
Programs for AIX
Administration Guide

Publication No. GC23-3897-04

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
GC23-3897-04

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

Program Number: 5765-529

Printed in the United States of America
on recycled paper containing 10%
recovered post-consumer fiber.

GC23-3897-04