

IBM

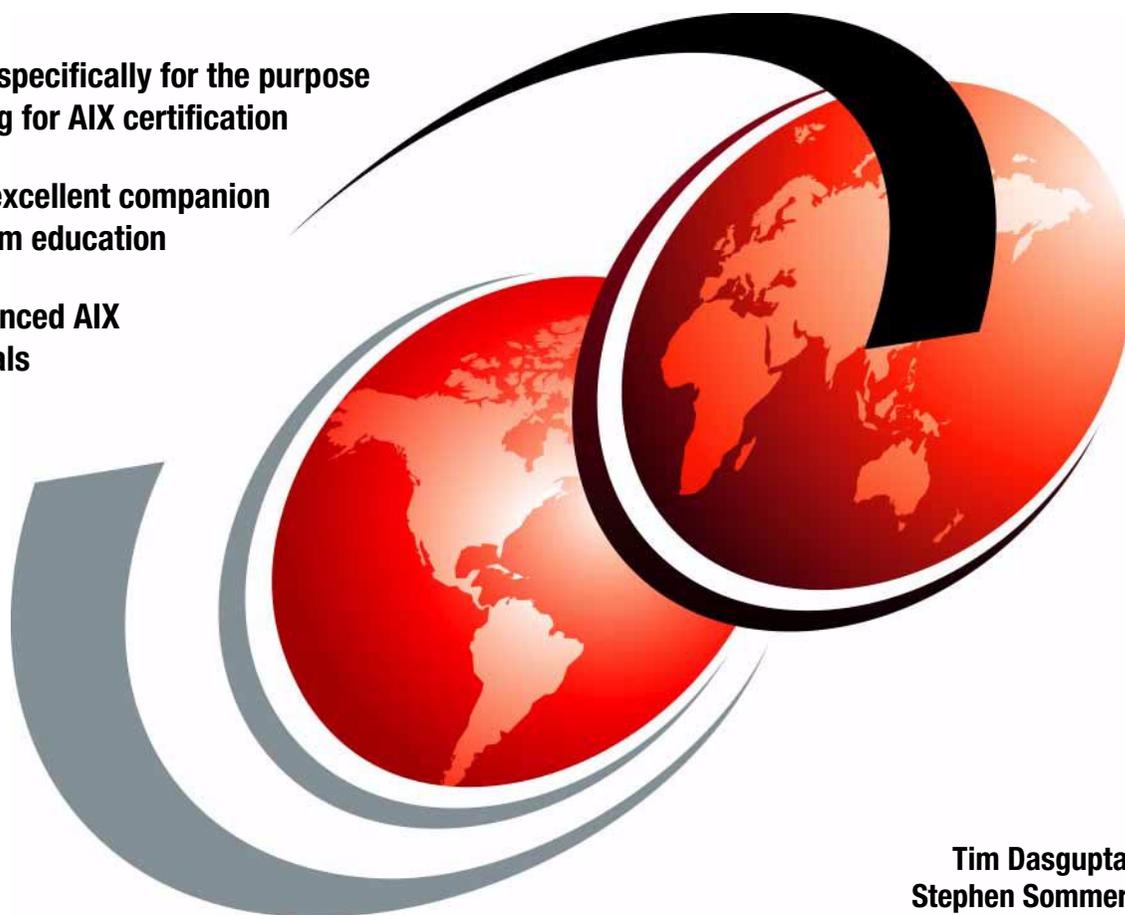


IBM **e**server™ Certification Study Guide - AIX 5L Problem Determination Tools and Techniques

Developed specifically for the purpose
of preparing for AIX certification

Makes an excellent companion
to classroom education

For experienced AIX
professionals



Tim Dasgupta
Stephen Sommer

ibm.com/redbooks

Redbooks



International Technical Support Organization

**IBM @server Certification Study Guide -
AIX 5L Problem Determination Tools and
Techniques**

January 2003

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

Second Edition (January 2003)

This edition applies to AIX 5L Version 5.1 (5765-E61) and subsequent releases running on an IBM @server pSeries or RS/6000 server.

© Copyright International Business Machines Corporation 2000, 2003. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	xi
Tables	xiii
Notices	xv
Trademarks	xvi
Preface	xvii
The team that wrote this redbook	xviii
Become a published author	xix
Comments welcome	xx
Chapter 1. Certification overview	1
1.1 Certification requirements	2
1.1.1 Required prerequisite	2
1.1.2 Recommended prerequisite	2
1.1.3 Information and registration for the certification exam	2
1.1.4 Core requirements	2
1.2 Certification education courses	6
Chapter 2. Customer relations	7
2.1 Defining the problem	8
2.2 Collecting information from the user	8
2.3 Collecting information about the system	10
2.4 Quiz	10
2.4.1 Answers	12
Chapter 3. Booting problem determination	13
3.1 A general overview of the boot process	14
3.2 BIST - POST	15
3.2.1 MCA systems	15
3.2.2 PCI systems	20
3.3 Boot phase 1	22
3.4 Boot phase 2	24
3.4.1 LED 551, 555, or 557	26
3.4.2 LED 552, 554, or 556	26
3.4.3 LED 518	28
3.4.4 The alog command	28
3.5 Boot phase 3	29

3.5.1	The /etc/inittab file	31
3.5.2	LED 553	32
3.5.3	LED c31	32
3.5.4	LED 581	32
3.5.5	pSeries servers	34
3.6	Boot-related information in the error log	35
3.7	Boot summary	36
3.8	Command summary	38
3.8.1	The errpt command	38
3.8.2	The w command	39
3.9	Quiz	39
3.9.1	Answers	42
3.10	Exercises	42
Chapter 4. Hardware problem determination		43
4.1	Hardware basics	44
4.1.1	Hardware inventory	44
4.2	Running diagnostics	46
4.2.1	Concurrent mode	47
4.2.2	Stand-alone diagnostics from disk - service mode	50
4.2.3	Stand-alone diagnostics from CD-ROM	51
4.2.4	Task selection or service aids	52
4.3	Serial Storage Architecture disks	54
4.3.1	General SSA setup rules	54
4.3.2	SSA devices	55
4.3.3	SSA disk considerations	55
4.3.4	Three-digit display values	57
4.3.5	Common boot time LEDs	57
4.3.6	888 in the three-digit display	58
4.4	Command summary	60
4.4.1	The chdev command	61
4.4.2	The lsattr command	61
4.5	Quiz	62
4.5.1	Answers	64
4.6	Exercises	64
Chapter 5. System dumps		65
5.1	Configuring the dump device	66
5.2	Starting a system dump	68
5.2.1	Using the command line	68
5.2.2	Using the SMIT interface	69
5.2.3	Using the Reset button	70
5.2.4	Using special key sequences	71

5.2.5 System Hang Detection	73
5.3 System dump status check	74
5.3.1 Status codes	74
5.4 Increasing the size of the dump device	76
5.5 Configuring remote dump devices	76
5.6 Copying a system dump	77
5.7 Reading dumps	79
5.8 Core dumps	81
5.8.1 Checking for core dump	81
5.8.2 Locating a core dump	82
5.8.3 Determining the program that caused the core dump	82
5.9 Command summary	83
5.9.1 The crash command	83
5.9.2 Types of crashes	95
5.9.3 The snap command	95
5.9.4 The strings command	97
5.9.5 The sysdumpdev command	98
5.9.6 The sysdumpstart command	100
5.10 Quiz	101
5.10.1 Answers	106
5.11 Exercises	106
Chapter 6. Error reports	107
6.1 The error daemon	108
6.2 The errdemon command	108
6.3 The errpt command	110
6.3.1 Error classes	121
6.4 The errclear command	125
6.5 Accounting	127
6.5.1 Setting up an accounting system	128
6.5.2 Setting up disk-usage accounting	129
6.6 The syslogd daemon	130
6.7 Quiz	133
6.7.1 Answers	136
6.8 Exercises	136
Chapter 7. LVM, file system, and disk problem determination	137
7.1 LVM data	138
7.1.1 Physical volumes	138
7.1.2 Volume groups	138
7.1.3 Logical volumes	139
7.1.4 Object Data Manager (ODM)	139
7.2 LVM problem determination	139

7.2.1	Data relocation	140
7.2.2	Backup data	140
7.2.3	ODM resynchronization	140
7.2.4	Understanding importvg problems	141
7.2.5	Extending the number of max physical partitions	143
7.3	Disk replacement	144
7.3.1	Replacing a disk	144
7.3.2	Recovering an incorrectly removed disk	148
7.4	The AIX JFS	150
7.4.1	Creating a JFS	150
7.4.2	Increasing the file system size	152
7.4.3	File system verification and recovery	152
7.4.4	Sparse file allocation	154
7.4.5	Unmount problems	154
7.4.6	Removing file systems	155
7.4.7	Different output from du and df commands	156
7.4.8	Enhanced journaled file system	156
7.4.9	The /proc file system	162
7.4.10	Disk quota	164
7.5	Paging space	166
7.5.1	Recommendations for creating or enlarging paging space	166
7.5.2	Determining if more paging space is needed	167
7.5.3	Reducing and removing paging space	168
7.6	Command summary	169
7.6.1	The lsvg command	169
7.6.2	The chvg command	170
7.6.3	The importvg command	171
7.6.4	The rmlvcopy command	172
7.6.5	The reducevg command	172
7.6.6	The rmdev command	172
7.6.7	The syncvg command	173
7.7	Quiz	173
7.7.1	Answers	176
7.8	Exercises	176
Chapter 8. Network problem determination		177
8.1	Network interface problems	178
8.2	Routing problems	181
8.2.1	Dynamic or static routing	184
8.3	Name resolution problems	185
8.3.1	The tcpdump and iptrace commands	186
8.4	NFS troubleshooting	189
8.4.1	General steps for NFS problem solving	189

8.4.2	NFS mount problems	190
8.4.3	Increasing NFS Socket Buffer Size	191
8.4.4	The biod and nfsd daemons	191
8.5	Command summary	192
8.5.1	The chdev command	192
8.5.2	The exportfs command	192
8.5.3	The ifconfig command	193
8.5.4	The iptrace command	193
8.5.5	The lsattr command	194
8.5.6	The netstat command	194
8.5.7	The route command	195
8.5.8	The tcpdump command	195
8.6	Quiz	196
8.6.1	Answers	200
8.7	Exercises	200
Chapter 9. System access and printer problem determination		201
9.1	User license problems	202
9.2	Telnet troubleshooting	203
9.2.1	Network problems	203
9.2.2	The telnet subserver	204
9.2.3	Slow telnet login	205
9.2.4	Telnet error	206
9.3	FTP troubleshooting	206
9.3.1	File limits	206
9.4	System settings	207
9.4.1	Adjusting AIX kernel parameters	207
9.4.2	The su command	208
9.4.3	A full file system	210
9.5	Tracing	211
9.5.1	Trace hook IDs	212
9.5.2	Starting a trace	213
9.5.3	Trace reports	213
9.5.4	Tracing example	214
9.6	Managing mail logging	217
9.7	TTY troubleshooting	218
9.7.1	Respawning too rapidly errors	218
9.8	Printing	220
9.8.1	Local printing problem	220
9.8.2	Remote printing problem	221
9.8.3	The /var file system	221
9.8.4	Default printing subsystem	221
9.9	Command summary	225

9.9.1	The lslicense command	225
9.9.2	The lssrc command	225
9.9.3	The startsrc command	226
9.9.4	The trace command	226
9.9.5	The trcrpt command	227
9.10	Quiz	228
9.10.1	Answers	231
9.11	Exercises	231
Chapter 10. Performance problem determination		233
10.1	CPU-bound system	235
10.1.1	The sar command	235
10.1.2	The vmstat command	238
10.1.3	The ps command	241
10.1.4	The tprof command	243
10.2	Memory-bound system	246
10.2.1	The vmstat command	247
10.2.2	The ps command	251
10.2.3	The svmon command	253
10.2.4	The schedtune command	254
10.3	Disk I/O bound system	258
10.3.1	The iostat command	260
10.3.2	The filemon command	264
10.3.3	The fileplace command	264
10.3.4	The sar command	264
10.4	Network I/O bound system	265
10.4.1	The netstat command	266
10.4.2	The nfsstat command	273
10.4.3	The netpmon command	274
10.5	Workload Manager (WLM)	274
10.5.1	WLM concepts and architecture	276
10.5.2	Automatic assignment	284
10.5.3	Manual assignment	285
10.5.4	Backward compatibility	289
10.5.5	Resource sets	289
10.5.6	rset registry	291
10.6	System debuggers	293
10.6.1	The dbx command	294
10.6.2	The kdb command	296
10.7	Summary	299
10.8	Command summary	300
10.8.1	The sar command	300
10.8.2	The ps command	300

10.8.3	The netstat command	301
10.8.4	The nfsstat command	302
10.9	Quiz	302
10.9.1	Answers	305
Chapter 11. Software updates		307
11.1	Overview	308
11.1.1	Terminology	308
11.1.2	Software layout	308
11.1.3	Software states	309
11.2	Installing a software patch	311
11.2.1	Software patch installation procedure	312
11.3	Software inventory	315
11.4	Command summary	315
11.4.1	The lspp command	315
11.4.2	The installp command	316
11.4.3	The instfix command	317
11.4.4	The lppchk command	317
11.5	Quiz	318
11.5.1	Answers	319
11.6	Exercises	319
Chapter 12. Online documentation		321
12.1	Installing the Web browser	323
12.2	Installing the Web server	323
12.3	Installing the Documentation Library Service	324
12.4	Configuring the Documentation Library Service	325
12.5	Installing online manuals	326
12.6	Invoking the Documentation Library Service	326
12.7	Exercises	329
Abbreviations and acronyms		331
Related publications		341
IBM Redbooks		341
Other resources		342
Referenced Web sites		342
How to get IBM Redbooks		343
IBM Redbooks collections		343
Index		345

Figures

3-1	General boot order	14
3-2	Function selection menu in diag	17
3-3	Task selection menu in diag	18
3-4	Display/alter bootlist menu in diag	18
3-5	SMS main menu	21
3-6	Boot phase 1	23
3-7	Boot phase 2, part one.	24
3-8	Boot phase 2, part two.	25
3-9	Boot phase 3	30
3-10	Example of rc.boot 3 in /etc/inittab.	31
4-1	Main diagnostics menu	48
4-2	Format of the 103 code message	60
4-3	Case study	63
5-1	SMIT dump screen.	69
5-2	SMIT Add a TTY screen - Remote reboot options.	72
5-3	Error log case study.	102
5-4	Case study	104
6-1	The errpt command error log report process	111
6-2	The errpt command error record template repository process.	112
7-1	Disk problem mail from Automatic Error Log Analysis (diagela)	145
7-2	JFS organization	151
8-1	Case study	197
9-1	SMIT menu to change the number of licensed users	203
9-2	The smitty crjfsbf fast path	207
9-3	SMIT screen for changing AIX operating system characteristics.	208
9-4	Output display of the topas command	214
9-5	Change/Show Current Print Subsystem option	224
10-1	General performance tuning flowchart.	234
10-2	CPU penalty example	257
10-3	Web-based System Manager Overview and Tasks dialog	276
10-4	Hierarchy of classes.	277
10-5	Resources cascading through tiers	281
10-6	SMIT with the class creation attributes screen	281
10-7	SMIT panel shows the additional localshm attribute	283
10-8	Resource set definition to a specific class	290
10-9	SMIT main panel for resource set management	291
10-10	SMIT panel for rset registry management	292
10-11	SMIT panel to add a new resource set	293

10-12 Performance tuning flowchart	299
11-1 SMIT Software Maintenance and Utilities panel	311
11-2 Install and Update Software panel	313
12-1 Netscape filesets	323
12-2 HTTP filesets	324
12-3 Documentation Library Service filesets	325
12-4 Documentation Library Service	328

Tables

3-1	Common MCA LED codes	19
3-2	MCA POST LEDs	37
3-3	Boot phase 2 LED codes	37
3-4	Boot phase 3 LED codes	37
3-5	Commonly used flags of the errpt command	38
3-6	Commonly used flags of the w command	39
4-1	SSA adapter information	55
4-2	Common MCA LED codes	58
4-3	Location code mapping table	60
4-4	Commonly used flags of the chdev command	61
4-5	Commonly used flags of the lsattr command	61
5-1	REMOTE reboot ENABLE settings	73
5-2	SHD default values	73
5-3	Differences in system dump	80
5-4	The vmerrlog structure	93
5-5	Commonly used flags of the snap command	96
5-6	Commonly used flags of the strings command	97
5-7	Commonly used flags of the sysdumpdev command	99
5-8	Commonly used flags of the sysdumpstart command	101
6-1	Commonly used flags of the errdemon command	108
6-2	Commonly used flags of the errpt command	112
6-3	List of error classes	121
6-4	Commonly used flags of the errclear command	125
6-5	System Resource Controller commands	130
7-1	Journalized file system specifications	161
7-2	Functions of pseudo files in /proc/<pid> directory	163
7-3	Commonly used flag of the lsvg command	170
7-4	Commonly used flag of the chvg command	171
7-5	Commonly used flags of the importvg command	171
7-6	Commonly used flags of the reducevg command	172
7-7	Commonly used flags of the rmdev command	172
7-8	Commonly used flag of the syncvg command	173
8-1	Commonly used flags of the chdev command	192
8-2	Commonly used flags of the exportfs command	192
8-3	Commonly used flags of the ifconfig command	193
8-4	Commonly used flags of the iptrace command	193
8-5	Commonly used flags of the lsattr command	194
8-6	Commonly used flags of the netstat command	194

8-7	Commonly used flags of the route command	195
8-8	Commonly used flags of the tcpdump command	196
9-1	Managing TTY devices	219
9-2	Comparison of print subsystem functions	221
9-3	Commonly used flags of the lssrc command	225
9-4	Commonly used flags of the startsrc command	226
9-5	Commonly used flags of the trace command	227
9-6	Commonly used flags of the trcrpt command	227
10-1	CPU-related ps output	241
10-2	Memory-related ps output	251
10-3	Some commonly used flags for the schedtune command	258
10-4	List of process types	287
10-5	Examples of class assignment rules	288
10-6	Sample dbx subcommands	294
10-7	Sample kdb subcommands	296
10-8	Commonly used flags of the sar command	300
10-9	Commonly used flags of the ps command	300
10-10	Commonly used flags of the netstat command	301
10-11	Commonly used flags of the nfsstat command	302
11-1	Commonly used flags of the lppchk command	314
11-2	Commonly used flags of the lspp command	316
11-3	Commonly used flags of the installp command	316
11-4	Commonly used flags of the instfix command	317
11-5	Commonly used flags of the lppchk command	317

Notices

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AFP™
AFS®
AIX®
AIX 5L™
DFS™
@server
IBM®
Infoprint®
IPDS™

LoadLeveler®
Micro Channel®
Perform™
PowerPC Reference Platform®
PowerPC®
PowerPC Reference Platform®
pSeries™
PTX®
QMF™

Redbooks™
Redbooks(logo)™ 
RS/6000®
Sequent®
SPT™
TotalStorage™
Versatile Storage Server™

The following terms are trademarks of International Business Machines Corporation and Lotus Development Corporation in the United States, other countries, or both:

Domino™

Lotus®

Notes®

The following terms are trademarks of other companies:

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

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

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

C-bus is a trademark of Corollary, Inc. in the United States, other countries, or both.

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

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

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

Preface

The AIX and IBM @server pSeries Certifications offered through the Professional Certification Program from IBM are designed to validate the skills required of technical professionals who work in the powerful and often complex environments of AIX and IBM @server pSeries. A complete set of professional certifications is available. It includes:

- ▶ IBM Certified AIX User
- ▶ IBM Certified Specialist - Business Intelligence for RS/6000
- ▶ IBM Certified Specialist - Domino for RS/6000
- ▶ IBM @server Certified Specialist - p690 Solutions Sales
- ▶ IBM @server Certified Specialist - p690 Technical Support
- ▶ IBM @server Certified Specialist - pSeries Sales
- ▶ IBM @server Certified Specialist - pSeries AIX System Administration
- ▶ IBM @server Certified Specialist - pSeries AIX System Support
- ▶ IBM @server Certified Specialist - pSeries Solution Sales
- ▶ IBM Certified Specialist - RS/6000 SP and PSSP V3
- ▶ IBM Certified Specialist - Web Server for RS/6000
- ▶ IBM @server Certified Specialist - pSeries HACMP for AIX
- ▶ IBM @server Certified Advanced Technical Expert - pSeries and AIX 5L

Each certification is developed by following a thorough and rigorous process to ensure the exam is applicable to the job role and is a meaningful and appropriate assessment of skill. Subject matter experts who successfully perform the job participate throughout the entire development process. They bring a wealth of experience into the development process, making the exams much more meaningful than the typical test that only captures classroom knowledge, and ensuring that the exams are relevant to the *real world*. Thanks to their effort, the test content is both useful and valid. The result of this certification is the value of appropriate measurements of the skills required to perform the job role.

This IBM Redbook is designed as a study guide for professionals wishing to prepare for the AIX 5L Problem Determination Tools and Techniques certification exam as a selected course of study in order to achieve the IBM @server Certified Advanced Technical Expert - pSeries and AIX 5L certification.

This IBM Redbook is designed to provide a combination of theory and practical experience needed for a general understanding of the subject matter. It also provides sample questions that will help in the evaluation of personal progress and provide familiarity with the types of questions that will be encountered in the exam.

This publication does not replace practical experience, nor is it designed to be a stand-alone guide for any subject. Instead, it is an effective tool that, when combined with education activities and experience, can be a very useful preparation guide for the exam.

For additional information about certification and instructions on how to register for an exam, visit our Web site at:

<http://www.ibm.com/certify>

The team that wrote this redbook

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

Tim Dasgupta is an IBM Certified AIX Advanced Technical Expert (CATE). He works as a Senior Systems Architect at IBM Global Services in Canada. He has over eight years of experience in the areas of AIX, RS/6000, and pSeries. He is currently the Team Leader of Midrange Architecture Group in Montreal, Canada.

Stephen Sommer is an IBM Certified AIX Advanced Technical Expert (CATE), AIX Version 4.3.3 and 5.1. He works as a Senior IT Specialist at Faritec Services, an IBM Business Partner in Johannesburg, South Africa. He has eight years of experience in Midrange Support for AIX, RS/6000, and pSeries, both in South Africa and the United Kingdom.

The authors of the first edition are:

Thomas C. Cederlöf	IBM Sweden
André de Klerk	IBM South Africa
Thomas Herlin	IBM Denmark
Tomasz Ostaszewski	Prokom Software SA in Poland

The project that produced this publication was managed by:

Scott Vetter	IBM Austin
---------------------	------------

Special thanks to:

Darin Hartman	Program Manager, AIX Certification
Shannan L DeBrule	IBM Atlanta

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

Jesse Alcantar	IBM Austin
-----------------------	------------

John Agombar	IBM U.K
Greg Althaus	IBM Austin
Atsushi Baba	IBM Japan
Karl Borman	ILS Austin
Larry Brenner	IBM Austin
Malin Cederberg	ILS Sweden
Greg Flaig	IBM Austin
Edward Geraghty	IBM Boston
John Hance	IBM Australia
Adnan Ikram	IBM Pakistan
Peter Mayes	IBM U.K.
Shawn Mullen	IBM Austin
Brian Nicholls	IBM Austin
Robert Olsson	ILS Sweden
Michelle Page-Rivera	IBM Atlanta
Christopher Snell	IBM Raleigh
Darrin Woodard	IBM Toronto

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You'll team with IBM technical professionals, Business Partners and/or customers.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you'll develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

- ▶ Use the online **Contact us** review redbook form found at:

ibm.com/redbooks

- ▶ Send your comments in an Internet note to:

redbook@us.ibm.com

- ▶ Mail your comments to:

IBM Corporation, International Technical Support Organization
Dept. JN9B Building 003 Internal Zip 2834
11400 Burnet Road
Austin, Texas 78758-3493



Certification overview

This chapter provides an overview of the skill requirements needed to obtain an IBM Advanced Technical Expert certification. The following chapters are designed to provide a comprehensive review of specific topics that are essential for obtaining the certification IBM @server Certified Advanced Technical Expert - pSeries and AIX 5L.

This level certifies an advanced level of pSeries and AIX knowledge and understanding, both in breadth and depth. It verifies the ability to perform in-depth analysis, apply complex AIX concepts, and provide resolution to critical problems, all in a variety of areas within AIX, including the hardware that supports it.

1.1 Certification requirements

To attain the IBM @server Certified Advanced Technical Expert - pSeries and AIX 5L certification, you must pass four tests.

One test is the prerequisite in either pSeries AIX System Administration or pSeries AIX System Support. The other three tests are selected from a variety of pSeries and AIX topics. These requirements are explained in greater detail in the sections that follow.

1.1.1 Required prerequisite

Prior to attaining the IBM @server Certified Advanced Technical Expert - pSeries and AIX 5L certification, you must be certified as either an:

- ▶ IBM @server Certified Specialist - pSeries AIX System Administration
- or
- ▶ IBM @server Certified Specialist - pSeries AIX System Support

1.1.2 Recommended prerequisite

A minimum of six to 12 months of experience in performing in-depth analysis and applying complex AIX concepts in a variety of areas within AIX is a recommended prerequisite.

1.1.3 Information and registration for the certification exam

For the latest certification information, see the following Web site:

<http://www.ibm.com/certify>

1.1.4 Core requirements

Select three of the following exams. You will receive a Certificate of Proficiency for tests when passed.

AIX 5L Installation and System Recovery

Test 233 was developed for this certification.

Preparation for this exam is the topic of *IBM @server Certification Study Guide - AIX 5L Installation and System Recovery*, SG24-6183.

AIX 5L Performance and System Tuning

Test 234 was developed for this certification.

Preparation for this exam is the topic of *IBM @server Certification Study Guide - AIX 5L Performance and System Tuning, SG24-6184*.

AIX 5L Problem Determination Tools and Techniques

The following objectives were used as a basis when the certification test 235 was developed. Some of these topics have been regrouped to provide better organization when discussed in this publication.

Preparation for this exam is the topic of this publication.

Section I - Identify problem

To identify the problem:

1. Clarify the problem description.
 - a. Get an exact description of the symptom from the person reporting it.
 - b. Determine whether the problem occurs at particular times.
 - c. Determine what steps are required to recreate the problem or symptom.
2. Determine the source of the problem.
 - a. Determine if the system is hung or slow.
 - b. Analyze error logs for messages relating to the problem/symptom to isolate the source.
 - c. Check the system LED panel for pertinent information.
 - d. Determine basic network functionality by using network tools.
 - e. Examine a core file or attached running process.
 - f. Check the system console for error messages.
 - g. Force system dump on the hung system.
 - h. Isolate the cause of the printer problem.

Section II - Perform problem analysis

To perform problem analysis:

1. Identify the cause of the problem description.
 - a. Examine system dumps.
 - b. Examine log files for messages related to problem.
 - c. Check the AIX error log for hardware failures and paging space problems.
 - d. Determine if network interfaces are active.

- e. Verify that file systems are mounted and accessible using commands such as **mount**, **df**, and **ls**.
 - f. Check for software installation problems in install logs and using the **ls1pp** command.
 - g. Check **syslog.conf** to see if additional error logging is enabled.
 - h. Manage processes and daemons.
 - i. Diagnose Logical Volume Manager (LVM) problems.
 - j. Determine if file systems and kernel are compatible (**jfs/jfs2**).
2. Identify recent system changes preceding the problem.
 - a. Use the AIX error log report to determine if reboot introduced the problem.
 - b. Check with customer/vendor for recent software/hardware changes.
 - c. Identify recent system and application configuration changes.
 - d. Check for **vm tune**, **sched tune**, or **wlm** tuning changes.
 - e. Check for changes to printing subsystems.
 3. Analyze error messages.
 - a. Look up the system dump code.
 - b. Use the **alog -t boot -o** command to check for a problem during boot.

Section III - Use tools/create solution

To use tools and create a solution:

1. Use system tools to evaluate the problem.
 - a. Use the **vmstat** command to look for basic CPU, memory, and I/O bottlenecks.
 - b. Use the **topas** command to look for basic system performance problems.
 - c. Use the **iostat** command to identify unusual disk activity.
 - d. Use the **sar** command to check for excessive system call and fork rates.
 - e. Use **iptrace**, **netstat**, **traceroute**, and **ping** commands to check for network connectivity and adapter problems.
 - f. Use **tprof** to identify abnormal CPU usage.
 - g. Examine the hung process stack using the **crash/kdb** and/or **dbx** command.
 - h. Enable debug/verbose modes to gather more information.
 - i. Use dump/trace tools to examine system behavior.
 - j. Enable accounting/auditing to monitor application behavior.

- k. Use **ipc** commands for managing IPC (shared memory and semaphores, for example).
 - l. Use the **df** command to check file system status.
 - m. Use **lpstat** to determine printer status.
2. Use tools to create solutions.
- Use **wlm** to control resources.

Section IV - Perform problem avoidance

To implement a plan:

1. Plan implementation.
 - a. Use supported hardware/software configurations.
 - b. Develop a test plan for new hardware/software changes.
2. Monitor system health.
 - a. Monitor AIX error logs for hardware failure messages.
 - b. Monitor AIX error logs for software failure messages.
 - c. Monitor syslog output.
 - d. Monitor application logs.
 - e. Monitor for low memory, paging space, and file system space.
 - f. Log system performance.
 - g. Review error alerts.
 - h. Monitor the system operating environment (for example, power and temperature).
3. Implement system modification control.
 - a. Schedule system changes.
 - b. Restrict root access.
 - c. Evaluate the impact of any modifications as they relate to the enterprise.

AIX 5L Communications

Test 236 was developed for this certification.

Preparation for this exam is the topic of *IBM @server Certification Study Guide - AIX 5L Communications*, SG24-6186.

pSeries HACMP for AIX

Test 187 was developed was developed for this certification.

Preparation for this exam is the topic of *IBM @server Certification Study Guide - pSeries HACMP for AIX*, SG24-6187.

RS/6000 SP and PSSP V3.1

Test 188 was developed for this certification.

Preparation for this exam is the topic of *IBM @server Certification Study Guide - RS/6000 SP*, SG24-5348.

p690 Technical Support

Test 195 was developed for this certification.

An IBM Redbook is planned for first quarter 2003 on this subject.

1.2 Certification education courses

Courses are offered to help you prepare for the certification tests. For a current list, visit the following Web site, locate your test number, and select the education resources available:

<http://www.ibm.com/certify/tests/info.shtml>



Customer relations

The following topics are discussed in this chapter:

- ▶ Problem definition
- ▶ Collecting information from the user
- ▶ Collecting information from the system

This chapter is intended for system support people who have to assist customers with a certain problem. The intention is to provide methods for describing a problem and collecting the necessary information about the problem in order to take the best corrective course of action.

2.1 Defining the problem

The first step in problem resolution is to define the problem. It is important that the person trying to solve the problem understands exactly what the users of the system perceive the problem to be. A clear definition of the problem is useful in two ways:

- ▶ It can give you a hint as to the cause of the problem.
- ▶ It is much easier to demonstrate to the users that the problem has been solved if you know how the problem is seen from their point of view.

For example, consider the situation where a user is unable to print a document. The problem may be due to the /var file system running out of space. The person solving the problem may fix this and demonstrate that the problem has been fixed by using the `df` command to show that the /var file system is no longer full.

This example can also be used to illustrate another difficulty with problem determination. Problems can be hidden by other problems. When you fix the most visible problem, another one may come to light. The problems that are unearthed during the problem determination process may be related to the one that was initially reported. In other words, there may be multiple problems with the same symptoms. In some cases, you may discover problems that are completely unrelated to the one that was initially reported.

In the previous printing example, simply increasing the amount of free space in the /var file system may not solve the problem being experienced by the user. The printing problem may turn out to be a cable problem, a problem with the printer, or perhaps a failure of the lpd daemon. This is why understanding the problem from the user's perspective is so important. In this example, a better way of proving that the problem has been resolved is to get the user to successfully print her document.

2.2 Collecting information from the user

The best way of understanding the problem from the users' perspective is to ask questions. From their perception of the situation, you can deduce if they have a problem, and the time scale in which they expect it to be resolved. Their expectations may extend beyond the scope of the machine or the application it is running.

The following questions should be asked when collecting information from the user during problem determination:

▶ What is the problem?

Try to get the users to explain what the problem is and how it affects them. Depending on the situation and the nature of the problem, this question can be supplemented by either of the following two questions:

- What is the system doing?
- What is the system *not* doing?

Once you have determined what the symptoms of the problem are, you should try to establish the history of the problem.

- ▶ How did you first notice the problem? Did you do anything differently that made you notice the problem?
- ▶ When did it happen? Does it always happen at the same time (for example, when the same job or application is run)?
- ▶ Does the same problem occur elsewhere? Is only one machine experiencing the problem or are multiple machines experiencing the same problem?
- ▶ Have any changes been made recently?
This refers to any type of change made to the system, ranging from adding new hardware or software to configuration changes of existing software.
- ▶ If a change has been made recently, were all of the prerequisites met before the change was made?

Software problems most often occur when changes have been made to the system, and either the prerequisites have not been met (for example, system firmware is not at the minimum required level), or instructions have not been followed exactly in order (for example, the person following the instructions second guesses what the instructions are attempting to do and decides they know a quicker route). The second guess then means that, because the person has taken a perceived better route, prerequisites for subsequent steps may not have been met, and the problem develops into the situation you are confronted with.

Other changes, such as the addition of hardware, bring their own problems, such as cables incorrectly assembled, contacts bent, or addressing misconfigured.

The How did you first notice the problem? question may not help you directly, but it is very useful in getting the person to talk about the problem. Once he starts talking, he invariably tells you things that will enable you to determine the starting point for problem resolution.

If the problem occurs on more than one machine, look for similarities and differences between the situations.

2.3 Collecting information about the system

The second step in problem determination is collecting information about the system. Some information will have already been obtained from the user during the process of defining the problem.

The user is not the only source that can provide information regarding a problem. By using various commands, it is possible to determine how the machine is configured, the errors that are being produced, and the state of the operating system.

The use of commands, such as **lsdev**, **lspv**, **lsvg**, **lslpp**, **lsattr**, **df**, **mount**, and others, enable you to gather information on how the system is configured. Other commands, such as **errpt**, can give you an indication of any errors being logged by the system.

If the system administrator uses SMIT or Web-based System Manager to perform administrative tasks, examine the log files for these applications to look for recent configuration changes. The log files are, by default, contained in the home directory of the root user and, by default, are named `/smit.log` for SMIT and `/websm.log` for the Web-based System Manager.

If you are looking for something specific based on the problem described by the user, then other files are often viewed or extracted so that they can be sent to your IBM support function for analysis, such as system dumps or checkstop files.

2.4 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

1. A user complains that she is no longer able to get into the system. Which of the following procedures should be performed to determine the cause?
 - A. Check the `/etc/security/passwd` file.
 - B. Inform the user that she is not doing something right.
 - C. Identify the steps the user is performing to access the system.
 - D. Ignore the user's definition and attempt to determine the problem from scratch.

2. A user explains that a problem was first noticed after a software update occurred. Which of the following procedures should the system administrator perform next to fix the updated software?
- A. Reboot the system.
 - B. Add more memory to the system.
 - C. Load a backup made prior to the update.
 - D. Check for the prerequisites and updates of software applied.

2.4.1 Answers

The following are the preferred answers to the questions provided in this section.

1. C
2. D



Booting problem determination

The following topics are discussed in this chapter:

- ▶ A general overview of the boot process
- ▶ Differences between MCA and PCI systems
- ▶ AIX boot phase 1 - configuring the base devices
- ▶ AIX boot phase 2 - activating the root volume group
- ▶ AIX boot phase 3 - configuring the remaining devices
- ▶ Common boot problem scenarios and how to fix them

Because boot problems are among the most common problems, an overall discussion on the subject is useful. This chapter begins with a general overview of the boot process, then expands on the details and discusses the process along with the LED codes for each stage of the boot process in further detail. A summary of the LED codes can be found in “LED codes” on page 37.

3.1 A general overview of the boot process

Both hardware and software problems can cause the system to halt during the boot process. The boot process is also dependent on which hardware platform is used. In the initial startup phase, there are some important differences between MCA and PCI systems, and these differences will determine the way to handle a hardware-related boot problem. These differences are covered in 3.2, “BIST - POST” on page 15.

The general workflow of the boot process is shown in Figure 3-1.

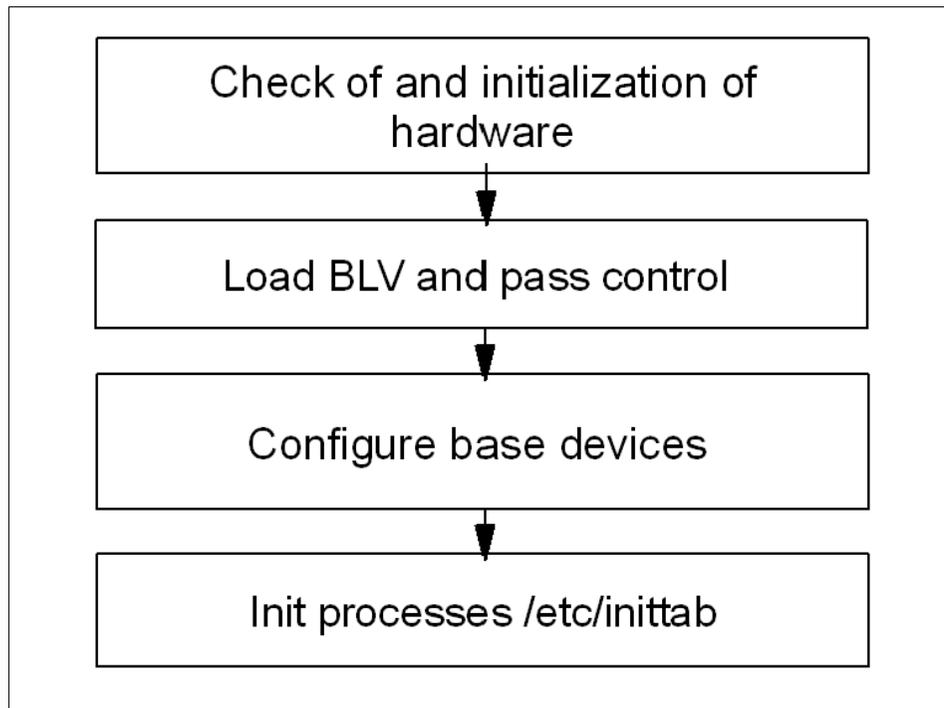


Figure 3-1 General boot order

The initial hardware check is to verify that the primary hardware is okay. This phase is divided into two separate phases on an MCA system. The first is the built-in self test (BIST), and the second a power-on self test (POST). On PCI systems, it is handled by a single POST. After this, the system loads the boot logical volume (BLV) into a RAM file system (RAMFS) and passes control to the BLV.

Note: The contents of the BLV are as follows:

▶ AIX kernel

The kernel is always loaded from the BLV. There is a copy of the kernel in /unix (soft link to /usr/lib/boot/unix_mp or unix_up). This version is used to build the hd4 file system where the kernel image is read during system boot.

▶ rc.boot

This is the configuration script that will be called three times by the init process during boot.

▶ Reduced ODM

Device support is provided only to devices marked as base devices in the ODM.

▶ Boot commands

For example, **cfgmgr** or **bootinfo**.

Because the rootvg is not available at this point, all the information needed for boot is included in the BLV used for creation of the RAMFS in memory. After this, the init process is loaded and starts to configure the base devices. This is named boot phase 1 (the init process executes the rc.boot script with an argument of 1).

The next step, named boot phase 2, attempts to activate rootvg, and this is probably the phase where the most common boot problems occur (for example, a file system or the jfslog is corrupt). Next, the control is passed to the rootvg init process and the RAMFS is released.

Finally, the init process, now loaded from disk (not the BLV), executes the rc.boot script with an argument of 3 to configure the remaining devices. This final stage is done from the /etc/inittab file. This is named boot phase 3.

3.2 BIST - POST

As mentioned before, there are differences between the classic RS/6000 system with MCA architecture and the PCI systems that are delivered today. The MCA system is discussed first.

3.2.1 MCA systems

At a system startup of an MCA system, the first thing that happens is a BIST. These tests are stored on EPROM chips, and the tests performed by BIST are

mainly to components on the motherboard. LED codes shown during this phase of the startup will be in the range of 100–195, defining the hardware status. After this, the POST will be initialized.

The task of the POST is to find a successful hardware path to a BLV. All hardware that is required to load a boot image is tested. The LED codes at this stage are in the range of 200–2E7. Both hardware and software problems can cause a halt in the startup process during this stage.

On an MCA system, the load of the BLV starts with checking the bootlist. The bootlist is defined by the key position (a physical key switch is located on the outside of many of the MCA models). When the key is in the normal position, applications will be started as well as network services. This is done when the init process reads the `/etc/inittab` file and executes the configuration scripts referenced in that file. A normal boot is represented by run level 2. The `/etc/inittab` file is discussed in further detail in 3.5.1, “The `/etc/inittab` file” on page 31. To manipulate the boot list for normal mode, use the following command:

```
# bootlist -m normal hdisk0 hdisk1 rmt0 cd0
```

This command will set the system to search `hdisk0` first for a usable BLV. If there is no BLV on `hdisk0`, then `hdisk1` will be searched, and so on.

The service bootlist is used when booting the system for maintenance tasks. The key is switched to the service position. No applications or network services will be started. To check the service bootlist, use the `-o` flag, which was introduced with AIX Version 4.2, as follows:

```
# bootlist -m service -o
fd0
cd0
rmt0
hdisk2
ent0
```

Another feature introduced with AIX Version 4.2 is the use of generic device names. Instead of pointing out the specified disk, such as `hdisk0` or `hdisk1`, you can use the generic definition of SCSI disks. For example, the following command uses the generic SCSI definition:

```
# bootlist -m service cd rmt scdisk
```

This command will request the system to probe any CD-ROM or DVD-ROM, then probe any tape drive, and finally probe any SCSI disk, for a BLV. The actual probing of the disk is a check of sector 0 for a boot record that contains data that points out the location of the boot image.

Changes to the bootlist can also be made through the **diag** command menus. At the Function Selection menu, choose **Task Selection**, as shown in Figure 3-2.

```
FUNCTION SELECTION                                     801002

Move cursor to selection, then press Enter.

Diagnostic Routines
  This selection will test the machine hardware. Wrap plugs and
  other advanced functions will not be used.
Advanced Diagnostics Routines
  This selection will test the machine hardware. Wrap plugs and
  other advanced functions will be used.
Task Selection(Diagnostics, Advanced Diagnostics, Service Aids, etc.)
  This selection will list the tasks supported by these procedures.
  Once a task is selected, a resource menu may be presented showing
  all resources supported by the task.
Resource Selection
  This selection will list the resources in the system that are supported
  by these procedures. Once a resource is selected, a task menu will
  be presented showing all tasks that can be run on the resource(s).

F1=Help           F10=Exit           F3=Previous Menu
```

Figure 3-2 Function selection menu in diag

In the list of tasks, choose **Display or Change Bootlist**, as shown in Figure 3-3 on page 18.

```
TASKS SELECTION LIST 801004

From the list below, select a task by moving the cursor to
the task and pressing 'Enter'.
To list the resources for the task highlighted, press 'List'.

[MORE...18]
  Display Firmware Device Node Information
  Display Hardware Error Report
  Display Hardware Vital Product Data
  Display Microcode Level
  Display Previous Diagnostic Results
  Display Resource Attributes
  Display Service Hints
  Display Software Product Data
  Display System Environmental Sensors
  Display Test Patterns
  Display or Change Bootlist
  Download Microcode
[MORE...12]

F1=Help           F4=List           F10=Exit         Enter
F3=Previous Menu
```

Figure 3-3 Task selection menu in diag

Finally, you have to choose whether to change the Normal mode bootlist or the Service mode bootlist, as shown in Figure 3-4.

```
DISPLAY/ALTER BOOTLIST 802590

Select an option, then press Enter.

  Normal mode bootlist
  This selection allows displaying, altering, or erasing
  the normal mode bootlist.
  Service mode bootlist
  This selection allows displaying, altering, or erasing
  the service mode bootlist.

F3=Cancel           F10=Exit
```

Figure 3-4 Display/alter bootlist menu in diag

At this point, a lot of things can cause a boot problem. The bootlist could point to a device that does not have a BLV, or the devices pointed to are not accessible because of hardware errors.

The following sections cover several problems that can cause a halt. All problems at this stage of the startup process have an error code defined, which is shown on the LED display on the operator panel of the system. These LED values are covered again in 4.3.5, “Common boot time LEDs” on page 57, with the additional codes not related to initial system startup.

LED 200

A LED code 200 is related to the secure key position. When the key is in the secure position, the boot will stop until the key is turned, either to the normal position or the service position. The boot will then continue.

LED 299

A LED code of 299 indicates that the BLV will be loaded. If this LED code is passed, then the load is successful. If, after passing 299, you get a stable 201, then you have to recreate the BLV, as discussed in “How to recreate the BLV” on page 19.

MCA LED codes

Table 3-1 provides a list of the most common LED codes on MCA systems. More of these can be found in the AIX base documentation.

Table 3-1 Common MCA LED codes

LED	Description
100–195	Hardware problem during BIST.
200	Key mode switch in secure position.
201	If LED 299 passed, recreate BLV. If LED 299 has not passed, POST encountered a hardware error.
221 721 221–229 223–229 225–229 233–235	The bootlist in NVRAM is incorrect (boot from media and change the bootlist), or the bootlist device has no bootimage (boot from media and recreate the BLV), or the bootlist device is unavailable (check for hardware errors).

How to recreate the BLV

When the LED code indicates that the BLV cannot be loaded, you should start diagnosis by checking for hardware problems, such as cable connections. The

next step is to start the system in maintenance mode from an external media, such as an AIX installation CD-ROM. Use the Access this Volume Group startup menu after booting from the installation media, and start a shell menu for recreation of the BLV (this menu is also used if the boot problem was due to an incorrect bootlist). Execute the following command if you want to recreate the BLV on hdisk0:

```
# bosboot -ad /dev/hdisk0
```

Another scenario where you may want to create a BLV with the **bosboot** command is with a mirrored rootvg. Mirroring this volume group does not make the disks containing the mirrored data bootable. You still have to define the disks in the bootlist and execute the **bosboot** command on the mirrored devices.

The following is a short summary on how to access the maintenance menus. For more detailed information see Chapter 10, “Accessing a system that will not boot”, in the *AIX Version 5.0 Installation Guide*, SC23-4112.

1. Boot the system from the installation media.
2. At the installation menu, choose **Start Maintenance for System Recovery**.
3. On the next menu, choose **Access a Root Volume Group**.
4. A list of accessible disks is shown. Choose the rootvg disk.
5. Finally, choose **Access this Volume Group** and start a shell when you want to recreate the BLV. Change the bootlist or forgotten root password.

Choose **Access this Volume Group** and start a shell before mounting file systems if the file systems or the jfslog in rootvg are corrupt.

3.2.2 PCI systems

When booting PCI systems, there are important differences from the MCA systems. It has already been mentioned that there is an absence of BIST. Another difference is the absence of the key switch. Modern PCI systems use a logical keymode switch, which is handled by the use of function keys. Also, the diag function is missing on some older PCI systems. The following section discusses how to change the bootlist and the support of the normal and service boot options on PCI systems.

Changing the bootlist on PCI systems

All PCI systems have System Management Services (SMS) menus. On most systems, these menus can be accessed by pressing function key 1 (F1) or 1 when the console is initiated (the use of 1 or F1 depends on the use of graphical display or ASCII terminal). At this time, a double beep is heard. Depending on the PCI model, there are three or four choices in the SMS main menu. One of these is named boot. Under this menu, you can define the bootlist. The SMS

main menu from an RS/6000 Model 43P-140 is shown in Figure 3-5. Newer PCI systems also have an additional selection called multiboot.



Figure 3-5 SMS main menu

Changing the boot order can also be done with the `bootlist` command.

Normal boot and service boot on PCI systems

Some PCI systems do not support service mode (for example, the 7248-43P). The only way to boot in another mode, such as maintenance mode, is to change the normal bootlist. This can be done with the `bootlist -m normal` command, if the system is accessible. If the system is not accessible, this can be done by booting from installation media and changing the bootlist through the SMS menus.

All PCI systems have a default bootlist. On modern PCI systems, this default bootlist can be accessed (and from `diag`) by using the F5 function key. This is a good option to use when booting the system in single user mode for accessing stand-alone `diag` functions. This cannot be done on older PCI systems. Instead, a single bootlist is provided and can be reset to the default values by removing the battery for about 30 seconds. This is because the bootlist is stored in NVRAM, and the NVRAM is only non-volatile as long as the battery is maintaining the memory.

Newer PCI architecture machines (for example, the 43P-150) support a service bootlist. The simplest way to find out if a particular system supports the service boot option is to execute:

```
# bootlist -m service -o
0514-220 bootlist: Invalid mode (service) for this model
```

If you receive the previous error message, the system does not support the service boot option.

All new PCI systems support the following key allocations as standard:

- ▶ F1 or 1 on ASCII terminal: Starts System Management Services
- ▶ F5 or 5 on ASCII terminal: Boot diag (use default boot list of fd, cd, scdisk, or network adapter)
- ▶ F6 or 6 on ASCII terminal: Boot diag (use of custom service boot list)

POST LED codes on PCI systems

On old PCI systems, such as the 7020-40P or the 7248-43P, the LED display is missing, so there will be no LED codes to help solve boot problems. Fortunately, this has been changed on modern PCI systems, but the error codes generated during this phase of the system startup differs from model to model. The only way to figure out the exact meaning of an error code is to refer to the service guide delivered with the system. IBM provides a Web page where service guides for most PCI systems are available in HTML and PDF format. The URL is:

http://www.ibm.com/servers/eserver/pseries/library/hardware_docs/index.html

3.3 Boot phase 1

So far, the system has tested the hardware, found a BLV, created the RAMFS, and started the init process from the BLV. The rootvg has not yet been activated. From this step on, the boot sequence is the same on both MCA systems and PCI systems.

The workflow for boot phase 1 is shown in Figure 3-6 on page 23.

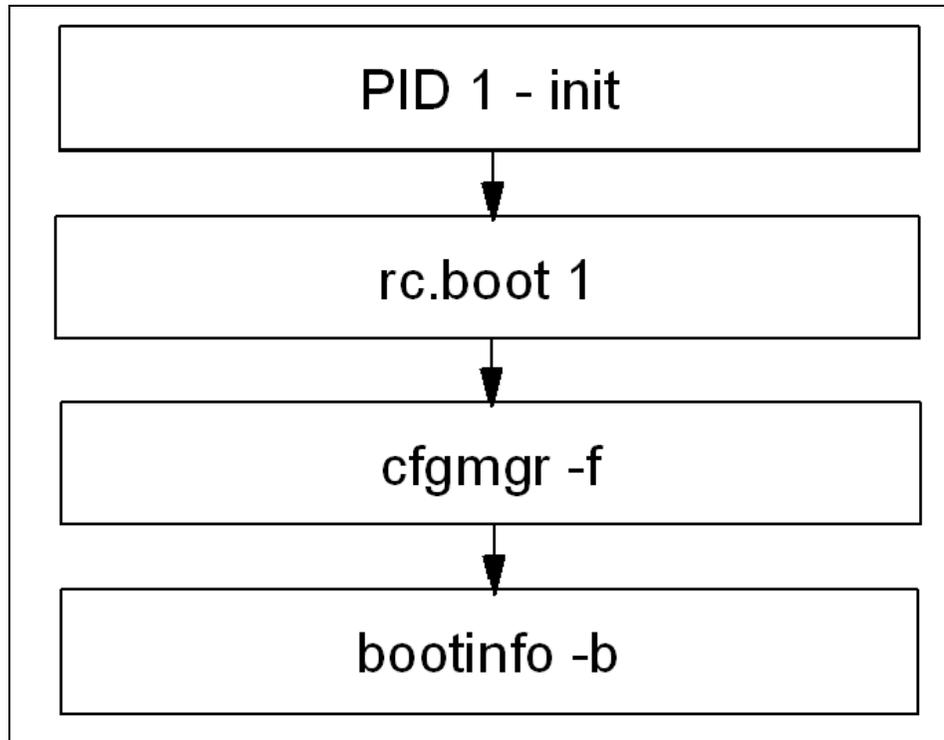


Figure 3-6 Boot phase 1

During this phase, the following steps are taken:

1. The init process started from RAMFS executes the boot script rc.boot 1. At this stage, the **restbase** command is called to copy the reduced ODM from the BLV into the RAMFS. If this operation fails, a LED code of 548 is presented.
2. After this, the **cfgmgr -f** command reads the Config_Rules class from the reduced ODM. In this class, devices with the attribute phase=1 will be considered base devices. Base devices are all devices that are necessary to access rootvg. The process invoked with rc.boot 1 attempts to configure devices so that rootvg can be activated in the next rc.boot phase.
3. At the end of boot phase 1, the **bootinfo -b** command is called to determine the last boot device. At this stage, the LED shows 511.

3.4 Boot phase 2

In boot phase 2, the rc.boot script is passed to the parameter 2. The first part of this phase is shown in Figure 3-7.

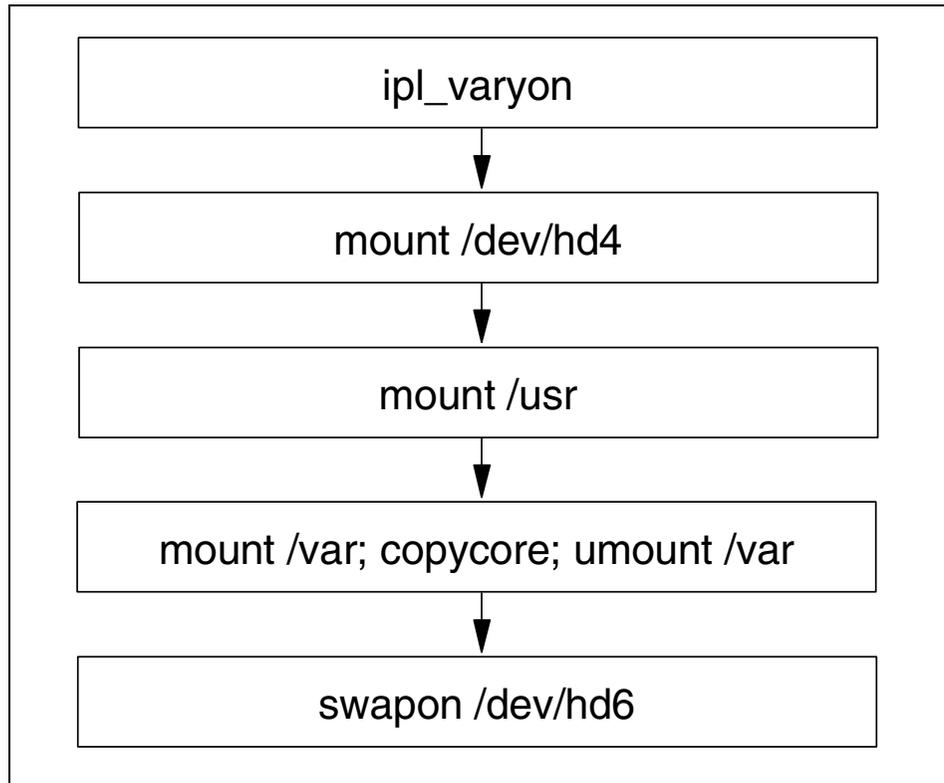


Figure 3-7 Boot phase 2, part one

During this phase, the following steps are taken.

1. The rootvg volume group will be varied on with the special **ipl_varyon** command. If this command is not successful, one of the following LED codes will appear: 552, 554, or 556.
2. After the successful execution of **ipl_varyon**, the root file system (/dev/hd4) is mounted on a temporary mount point (/mnt) in RAMFS. If this fails, 555 or 557 will appear in the LED display.
3. Next, the /usr and /var file systems are mounted. If this fails, the LED 518 appears. The mounting of /var, at this point, enables the system to copy an eventual dump from the default dump devices, /dev/hd6, to the default copy directory, /var/adm/ras.

4. After this, rootvg's primary paging space, /dev/hd6, will be activated.

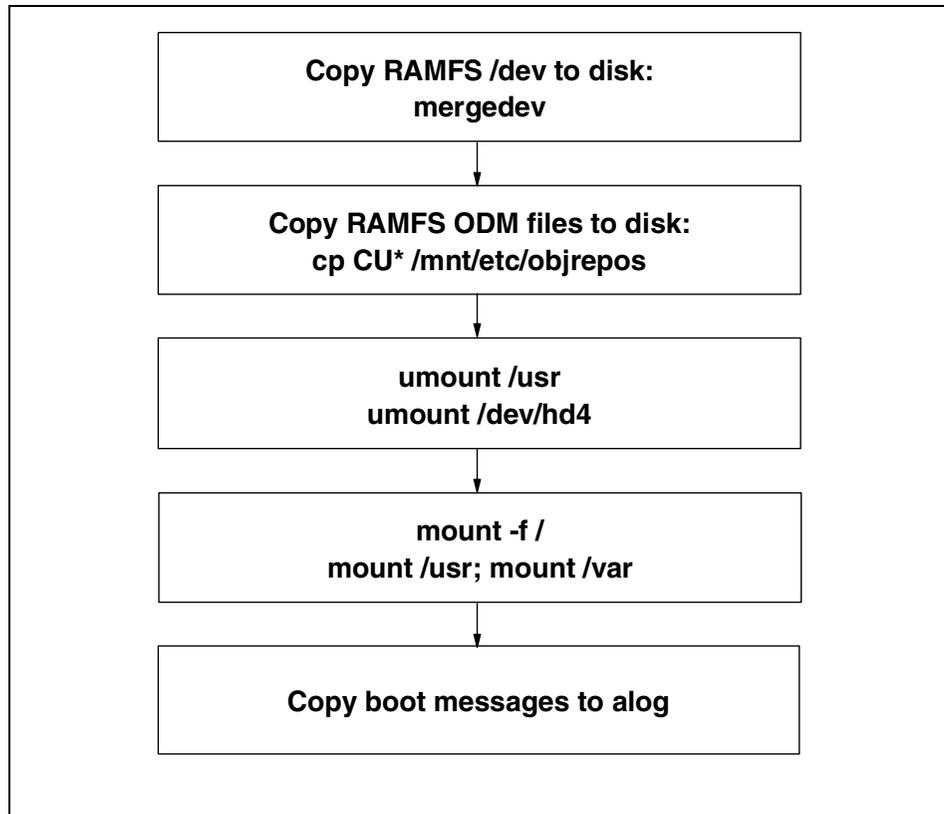


Figure 3-8 Boot phase 2, part two

The second part of this phase is shown in Figure 3-8, and the following steps are taken:

1. The copy of rootvg's RAMFS' ODM and /dev directories will occur (**mergedev**). This is possible because the temporary mount point, /mnt, is used for the mounted root file system.
2. Next, the /usr and /var from the RAMFS are unmounted.
3. Finally, the root file system from rootvg (disk) is mounted over the root file system from the RAMFS. The mount points for the rootvg file systems become available. Now the /var and /usr file systems from the rootvg can be mounted again on their ordinary mount points.

There is no console available at this stage, so all boot messages will be copied to **alog**. The **alog** command can maintain and manage logs.

As mentioned, there are a lot of different possible problems in this phase of the boot. The following sections discuss how to correct some of them.

3.4.1 LED 551, 555, or 557

There can be several reasons for a system to halt with LED codes 551, 555, or 557. For example:

- ▶ A damaged file system
- ▶ A damaged journaled file system (JFS) log device
- ▶ A failing disk in the machine that is a member of the rootvg

To diagnose and fix these problems, you need to boot from a bootable media, access the maintenance menus, choose **Access a Volume Group**, and start a shell before mounting file systems, and then do one or all of the following actions:

- ▶ To ensure file system integrity, run the **fsck** command to fix any file systems that may be corrupted:

```
# fsck -y /dev/hd1
# fsck -y /dev/hd2
# fsck -y /dev/hd3
# fsck -y /dev/hd4
# fsck -y /dev/hd9var
```

- ▶ To ensure the correct function of the log device, run the **logform** command on /dev/hd8 to recreate the log device:

```
# /usr/sbin/logform /dev/hd8
```

- ▶ If the BLV is corrupted, recreate the BLV and update the bootlist:

```
# bosboot -a -d /dev/hdisk0
# bootlist -m normal hdisk0
```

3.4.2 LED 552, 554, or 556

A LED code of 552, 554, or 556 during a standard disk-based boot indicates that a failure occurred during the varyon of the rootvg volume group. This can be the cause of:

- ▶ A damaged file system
- ▶ A damaged journaled file system (JFS) log device
- ▶ A bad IPL-device record or bad IPL-device magic number (the magic number indicates the device type)
- ▶ A damaged copy of the Object Data Manager (ODM) database on the boot logical volume

- ▶ A hard disk in the inactive state in the root volume group
- ▶ A damaged superblock

To diagnose and fix the problem, you need to boot from the installation media, navigate the menus to access the volume group, and start a shell before mounting the file systems.

If the **fsck** command indicates that block 8 could not be read when used, as shown in 3.4.1, “LED 551, 555, or 557” on page 26, the file system is probably unrecoverable. The easiest way to fix an unrecoverable file system is to recreate it. This involves deleting it from the system and restoring it from a backup. Note that `/dev/hd4` cannot be recreated. If `/dev/hd4` is unrecoverable, you must reinstall AIX.

A corrupted ODM in the BLV is also a possible cause for these LED codes. To create a usable one, run the following commands that remove the system's configuration and save it to a backup directory:

```
# /usr/sbin/mount /dev/hd4 /mnt
# /usr/sbin/mount /dev/hd2 /usr
# /usr/bin/mkdir /mnt/etc/objrepos/bak
# /usr/bin/cp /mnt/etc/objrepos/Cu* /mnt/etc/objrepos/bak
# /usr/bin/cp /etc/objrepos/Cu* /mnt/etc/objrepos
# /usr/sbin/umount all
# exit
```

After this, you must copy this new version of the ODM in the RAMFS to the BLV. This is done with the **savebase** command. Before that, make sure you place it on the disk used for normal boot by executing:

```
# lsiv -m hd5
```

Save the clean ODM database to the boot logical volume. For example:

```
# savebase -d /dev/hdisk0
```

Finally, recreate the BLV and reboot the system. For example:

```
# bosboot -ad /dev/hdisk0
# shutdown -Fr
```

Another possible reason for these error codes is a corrupted superblock. If you boot in maintenance mode and receive error messages such as Not an AIX file system or Not a recognized file system type, it is probably due to a corrupted superblock in the file system.

Each file system has two super blocks: One in logical block 1 and a copy in logical block 31. To copy the superblock from block 31 to block 1 for the root file system, issue the following command (before you use this command, check the

product documentation for the AIX release you are using to make sure all of the parameters shown are correct):

```
# dd count=1 bs=4k skip=31 seek=1 if=/dev/hd4 of=/dev/hd4
```

3.4.3 LED 518

The 518 LED code has an unclear definition in the product documentation, which reads:

Display Value 518

Remote mount of the / (root) and /usr file systems during network boot did not complete successfully.

This is not the entire problem. If the system runs into problems while mounting the /usr from disk (locally, not a network mount), you will get the same error. Fix this problem using the same procedure as you would for any other rootvg file system corruption.

3.4.4 The alog command

Up until this stage, the system has not yet configured the console, so there is no stdout defined for the boot processes. At this stage, the **alog** command is useful.

The **alog** command can maintain and manage logs. All boot information is sent through the **alog** command. To look at the boot messages, use the following command options:

```
# alog -ot boot
***** no stderr *****
-----
Time: 12          LEDS: 0x538
invoking top level program -- "/usr/lib/methods/definnet > /dev/null
2>&1;opt=~ /u
sr/sbin/lstatr -E -l inet0 -a bootup_option -F value`
      if [ $opt = "no" ];then nf=/etc/rc.net
      else nf=/etc/rc.bsdnet
      fi;$nf -2;x=$?;test $x -ne 0&&echo $nf failed. Check for invalid
command
s >&2;exit $x"
Time: 21          LEDS: 0x539
return code = 0
***** no stdout *****
```

The **alog** command with the **-C** option hangs the attributes for a specified log type.

Assuming that you have an existing log type *sample* in the `alog` configuration database, to change the name of the log file for the log type *sample* to `/var/sample.log`, enter:

```
# alog -C -t sample -f /var/sample.log
```

The next step of the boot process checks the `bootup_option` to determine if a BSD-style configuration of TCP/IP services are to be used, or if the default of ODM-supported configuration should be used. During this stage, the LED codes 538 and 539 are shown, as provided in the preceding `alog` example.

3.5 Boot phase 3

In the boot process, the following boot tasks have been accomplished:

- ▶ Hardware configuration performed during BIST and POST
- ▶ The load of the BLV
- ▶ Phase 1, where base devices are configured to prepare the system for activating the rootvg
- ▶ Phase 2, where rootvg is activated

Finally, phase 3 is initiated by the `init` process loaded from rootvg. An outline of this phase is shown in Figure 3-9 on page 30.

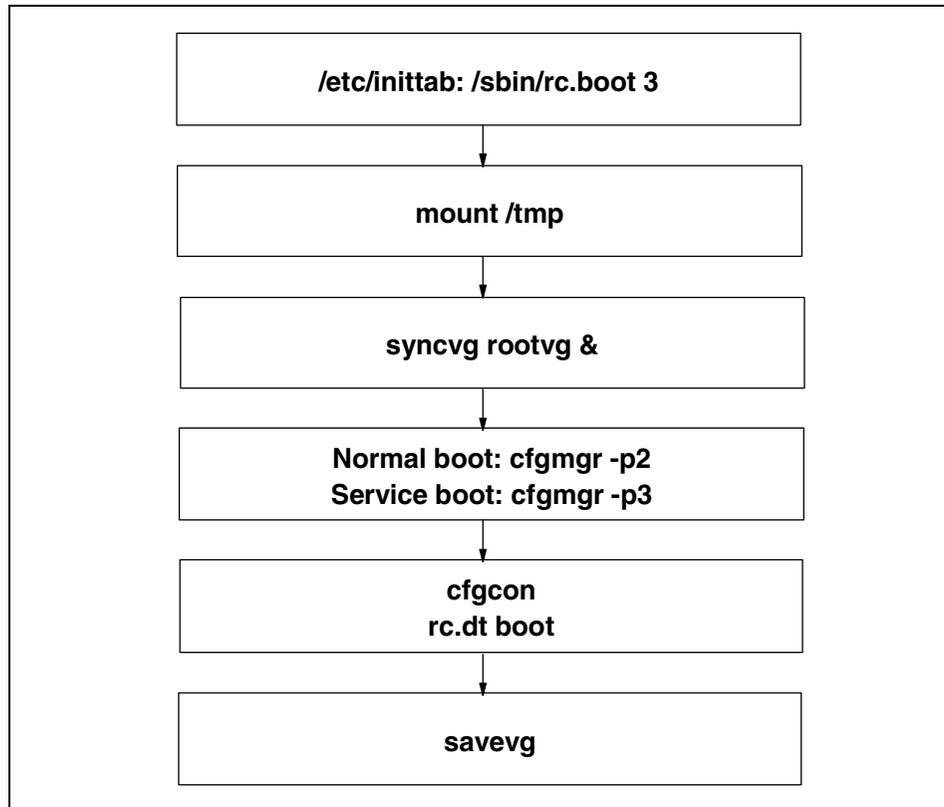


Figure 3-9 Boot phase 3

The order of boot phase 3 is as follows:

1. Phase 3 is started in /etc/inittab.
2. The /tmp file system is mounted.
3. The rootvg is synchronized. This can take some time. This is why the **syncvg rootvg** command is executed as a background process. At this stage, the LED code 553 is shown.
4. At this stage, the **cfgmgr -p2** process for normal boot and the **cfgmgr -p3** process for service mode are also run. **cfgmgr** reads the Config_rules file from ODM and checks for devices with phase=2 or phase=3.
5. Next, the console will be configured. LED codes shown when configuring the console are shown on the following page. After the configuration of the console, boot messages are sent to the console if no STDOUT redirection is made. Many of these boot messages scroll past at a fast pace, so there is not

always time to read all of the messages. However, all missed messages can be found in `/var/adm/ras/conslog`.

6. Finally, the synchronization of the ODM in the BLV with the ODM from the `/` (root) file system is done by the `savebase` command.

When the `cfgcon` process is called, different LED codes are shown depending on which device is configured.

The `cfgcon` LED codes include:

- c31** Console not yet configured. Provides instructions to select console.
- c32** Console is an LFT terminal.
- c33** Console is a TTY.
- c34** Console is a file on the disk.

3.5.1 The `/etc/inittab` file

The `/etc/inittab` file supplies configuration scripts to the `init` process. In Figure 3-10, the highlighted line is the file record that runs `rc.boot` with parameter 3.

```
(C) COPYRIGHT International Business Machines Corp. 1989, 1993
: All Rights Reserved
: Licensed Materials - Property of IBM
:
: US Government Users Restricted Rights - Use, duplication or
: disclosure restricted by GSA ADP Schedule Contract with IBM Corp.
:
: Note - initdefault and sysinit should be the first and second entry.
:
init:2:initdefault:
brc::sysinit:/sbin/rc.boot 3 >/dev/console 2>&1 # Phase 3 of system boot
powerfail::powerfail:/etc/rc.powerfail 2>&1 | alog -tboot > /dev/console # Power
Failure Detection
rc:2:wait:/etc/rc 2>&1 | alog -tboot > /dev/console # Multi-User checks
fbcheck:2:wait:/usr/sbin/fbcheck 2>&1 | alog -tboot > /dev/console # run /etc/fi
rstboot
srcmstr:2:respawn:/usr/sbin/srcmstr # System Resource Controller
rctcpip:2:wait:/etc/rc.tcpip > /dev/console 2>&1 # Start TCP/IP daemons
rcnfs:2:wait:/etc/rc.nfs > /dev/console 2>&1 # Start NFS Daemons
cron:2:respawn:/usr/sbin/cron
piobe:2:wait:/usr/lib/lpd/pio/etc/pioint >/dev/null 2>&1 # pb cleanup
qdaemon:2:wait:/usr/bin/startsrc -sqdaemon
writesrv:2:wait:/usr/bin/startsrc -swritesrv
```

Figure 3-10 Example of `rc.boot 3` in `/etc/inittab`

The `/etc/inittab` file is composed of entries that are position dependent and have the following format:

Identifier:RunLevel:Action:Command

The first line in `/etc/inittab` (`initdefault`) defines what run level is to be considered as a default run level. In the example provided, the run level is 2, which means a normal multi-user boot. In the case of a multi-user boot, all records with run level 2 will be executed from the `/etc/inittab` file. If this value is missing, you are prompted at boot to define the run level.

The `rc.boot` line is to be executed on all run levels (this equals run level 0123456789). The action defined, `sysinit`, has to finish before continuing with the next line in `/etc/inittab`. From `rc.boot 3`, among other things, the `rootvg` is synchronized, the mirroring is started, and the `/tmp` directory mounted. A detailed description of `/etc/inittab` is provided in *IBM @server Certification Study Guide - pSeries AIX System Support*, SG24-6199.

3.5.2 LED 553

A LED code of 553 is caused when the `/etc/inittab` file cannot be read. To recover from a LED 553, check `/dev/hd3` and `/dev/hd4` for space problems and erase unneeded files to free up disk space. Check the `/etc/inittab` file for corruption and correct the errors if necessary. Typical syntax errors found in `/etc/inittab`, as seen at the support centers, are entries that are incorrectly defined in the file. When editing `/etc/inittab`, the `inittab` commands should be issued. For example:

- ▶ `mkitab`
- ▶ `chitab`

It is helpful to remember that `/etc/inittab` is very sensitive to even the most trivial syntax error. A misplaced dot can halt the system boot.

3.5.3 LED c31

LED code `c31` is not really an error code, but the system is waiting for input from the keyboard. This is usually encountered when booting from CD-ROM or a `mksysb` tape. This is normally the dialog to select the system console.

3.5.4 LED 581

LED code 581 is not really an error code. LED 581 is shown during the time that the configuration manager configures TCP/IP and runs `/etc/rc.net` to do specific adapter, interface, and host name configuration.

A problem is when the system hangs while executing `/etc/rc.net`. The problem can be caused by either a system or a network problem that happens because TCP/IP waits for replies over an interface. If there are no replies, the wait eventually times out and the system marks the interface as down. This time-out period varies and can range from around three minutes to an indefinite period.

The following problem determination procedure is used to verify that the methods and procedures run by `/etc/rc.net` are causing the LED 581 hang:

1. Boot the machine in service mode.
2. Move the `/etc/rc.net` file to a safe location:

```
mv /etc/rc.net /etc/rc.net.save
```
3. Reboot in normal mode to see if the system continues past the LED 581 and allows you to log in.

Note: The previous steps assume that DNS or NIS is not configured.

If you determine that the procedures in `/etc/rc.net` are causing the hang, that is, the system continued past LED 581 when you performed the steps above, the problem may be one of the following:

- ▶ Ethernet or token-ring hardware problems
Run diagnostics and check the error log.
- ▶ Missing or incorrect default route
- ▶ Networks not accessible
Check that the gateways, name servers, and NIS masters are up and available.
- ▶ Bad IP addresses or masks
Use the `iptrace` and `ipreport` commands for problem determination.
- ▶ Corrupt ODM
Remove and recreate network devices.
- ▶ Premature name or IP address resolution
Either `named`, `ybind/ypserv`, or `/etc/hosts` may need correction.
- ▶ Extra spaces at the ends of lines in configuration files
Use the `vi` editor with the `set list` subcommand to check files, such as the `/etc/filesystems` file, for this problem.
- ▶ LPP installations or configurations with errors
Reinstall the LPP.

A specific LED 581 hang case occurs when ATMLE is being used with DNS. If you are experiencing this problem, you can either work around the problem by adding a host=local,bind entry to the /etc/netsvc.conf file or by adding the following lines to the /etc/rc.net file:

```
#####
# Part III - Miscellaneous Commands.
#####
# Set the hostid and uname to `hostname`, where hostname has been
# set via ODM in Part I, or directly in Part II.
# (Note it is not required that hostname, hostid and uname all be
# the same).
export NSORDER="local"          <<=====NEW LINE ADDED HERE
/usr/sbin/hostid `hostname`    >>$LOGFILE 2>&1
/bin/uname -S`hostname|sed 's/\..*$//'\` >>$LOGFILE 2>&1
unset NSORDER                  <<=====NEW LINE ADDED HERE
#####
```

3.5.5 pSeries servers

Although new pSeries servers have the same basic boot steps, the LEDs are displayed differently. Please consult your specific pSeries service guide for more information. Following is an example of IBM @server pSeries 690 boot phases.

The IPL process starts when AC power is connected to the system. The IPL process has the following phases.

Phase 1: Service processor initialization

Phase 1 starts when AC power is connected to the system and ends when 0K is displayed in the media subsystem operator panel. 8xxx checkpoints are displayed during this phase. Several 9xxx codes may also be displayed. Service processor menus are available at the end of this phase by striking any key on the console keyboard.

Phase 2: Hardware initialization by the service processor

Phase 2 starts when system power-on is initiated by pressing the power on button on the media subsystem operator panel. 9xxx checkpoints are displayed during this time. 91FF, the last checkpoint in this phase, indicates that the transition to phase 3 is taking place.

Phase 3: System firmware initialization

On a full system partition, at phase 3, a system processor takes over control and continues initializing partition resources. During this phase, checkpoints in the

form Exxx are displayed. E105, the last checkpoint in this phase, indicates that control is being passed to the AIX boot program. On a partitioned system, there is a global system-wide initialization phase 3, during which a system processor continues the initialization process. Checkpoints in this phase are of the form Exxx. This global phase 3 ends with a LPAR... on the operator panel. As a logical partition begins a partition-initialization phase 3, one of the system processors assigned to that partition continues initialization of resources assigned to that partition. Checkpoints in this phase are also of the form Exxx. This partition phase 3 ends with an E105 displayed on the partition's virtual operator panel on the HMC, indicating control has been passed to that logical partition's AIX boot program. For both the global and partition phase 3, location codes may also be displayed on the physical operator panel and the partition's virtual terminal, respectively.

Phase 4: AIX boot

When AIX starts to boot, checkpoints in the form 0xxx and 2xxx are displayed. This phase ends when the AIX login prompt displays on the AIX console.

3.6 Boot-related information in the error log

Because the function of the error log should be familiar to you from your previous certification training, this section will only cover boot-related messages.

The error log facility provides historical information on system boots and what may have caused them. One way to find the reboot time stamp is to check for when error logging has been turned on, as shown in the following example:

```
# errpt
IDENTIFIER  TIMESTAMP    T C RESOURCE_NAME  DESCRIPTION
499B30CC    0711125600  T H ent1          ETHERNET DOWN
1104AA28    0711125200  T S SYSPROC      SYSTEM RESET INTERRUPT RECEIVED
9DBCDFDEE   0711125500  T O errdemon     ERROR LOGGING TURNED ON
499B30CC    0707114100  T H ent1          ETHERNET DOWN
499B30CC    0707113700  T H ent1          ETHERNET DOWN
C60BB505    0705101400  P S SYSPROC      SW PROGRAM ABNORMALLY TERMINATED
35BFC499    0705101100  P H cd0          DISK OPERATION ERROR
0BA49C99    0705101100  T H scsi0        SCSI BUS ERROR
9DBCDFDEE   0704153700  T O errdemon     ERROR LOGGING TURNED ON
192AC071    0704153700  T O errdemon     ERROR LOGGING TURNED OFF
9DBCDFDEE   0704152600  T O errdemon     ERROR LOGGING TURNED ON
```

Every time the system is booted, the error log facility is started. In the previous example, the system has been gracefully shut down two times on the 4th of July.

When the system is gracefully shut down, the error logging facility is also shut down, as the error log entry 192AC071 shows. In the case of the reboot on the 11th of July, there is no stop of the error log facility reported; in other words, that shutdown cannot be considered graceful. Three minutes before the reboot (12:55), a system reset is reported (the line above with the 12:52 time stamp). The reason for the non-graceful reboot is often reported sequentially later than the reboot. The reason for the reboot (the use of the Reset button) is shown highlighted in the following example:

```
# errpt -aj 1104AA28
```

```
-----  
LABEL:          SYS_RESET  
IDENTIFIER:     1104AA28  
  
Date/Time:      Tue Jul 11 12:52:54  
Sequence Number: 12  
Machine Id:     000BC6DD4C00  
Node Id:        server3  
Class:          S  
Type:           TEMP  
Resource Name:  SYSPROC
```

Description
SYSTEM RESET INTERRUPT RECEIVED

Probable Causes
SYSTEM RESET INTERRUPT

Detail Data
KEY MODE SWITCH POSITION AT BOOT TIME
normal
KEY MODE SWITCH POSITION CURRENTLY
normal

3.7 Boot summary

The following section provides short summaries of the boot phases and some common LED codes.

Boot phases

BIST and POST are used to test hardware and to find a successful hardware path to a BLV.

Boot phase 1 (init rc.boot 1) is used to configure base devices.

Boot phase 2 (init rc.boot 2) is used to activate the rootvg.

Boot phase 3 (init /sbin/rc.boot 3) is used to configure the rest of the devices.

LED codes

The LED codes during POST on an MCA system are listed in Table 3-2.

Table 3-2 MCA POST LEDs

LED	Reason/action
100–195	Hardware problem during BIST.
200	Key mode switch in secure position.
201	If LED 299 passed, recreate BLV. If LED 299 has not passed, POST encountered a hardware error.
221 721 221–229 223–229 225–229 233–235	The bootlist in NVRAM is incorrect (boot from media and change the bootlist), or the bootlist device has no bootimage (boot from media and recreate the BLV), or the bootlist device is unavailable (check for hardware errors).

The LED codes shown during boot phase 2 are listed in Table 3-3.

Table 3-3 Boot phase 2 LED codes

LED	Reason/action
551 555 557	Corrupted file system, use the fsck -y device command. Corrupted jfslog, use the /usr/sbin/logform /dev/hd8 command. Corrupted BLV, use the bosboot -ad device command.
552 554 556	The ipl_varyon failed. Except for the reason mentioned above (551, 555, or 557): <ul style="list-style-type: none"> ▶ Corrupted ODM, backup ODM; recreate with savebase. ▶ Superblock dirty; copy in superblock from block 31.
518	/usr cannot be mounted: If /usr should be mounted over the network, check for network problem. If /usr is to be mounted locally, fix the file system.

The LED codes shown during boot phase 3 are listed in Table 3-4.

Table 3-4 Boot phase 3 LED codes

LED	Reason/action
553	Syntax error in /etc/inittab.

LED	Reason/action
c31	Define the console.

3.8 Command summary

The following section provides a list of the key commands discussed in this chapter.

3.8.1 The `errpt` command

The `errpt` command is used to check for errors reported by the error log facility.

The syntax of the `errpt` command is provided in the following examples.

To process a report from the error log, the syntax is:

```
errpt [ -a ] [ -A ] [ -c ] [ -d ErrorClassList ] [ -D ] [ -e EndDate ] [ -g ]
[ -i File ] [ -I File ] [ -j ErrorID [ ,ErrorID ] ] | [ -k ErrorID [ ,ErrorID ] ]
[ -J ErrorLabel [ ,ErrorLabel ] ] | [ -K ErrorLabel [ ,ErrorLabel ] ]
[ -l SequenceNumber ] [ -m Machine ] [ -n Node ] [ -s StartDate ]
[ -F FlagList ] [ -N ResourceNameList ] [ -P ] [ -R ResourceTypeList ]
[ -S ResourceClassList ] [ -T ErrorTypeList ] [ -y File ] [ -z File ]
```

To process a report from the error record template repository, the syntax is:

```
errpt [ -a ] [ -A ] [ -I File ] [ -t ] [ -d ErrorClassList ]
[ -j ErrorID [ ,ErrorID ] ] | [ -k ErrorID [ ,ErrorID ] ]
[ -J ErrorLabel [ ,ErrorLabel ] ] | [ -K ErrorLabel [ ,ErrorLabel ] ]
[ -F FlagList ] [ -P ] [ -T ErrorTypeList ] [ -y File ] [ -z File ]
```

Some useful `errpt` command flags are provided in Table 3-5.

Table 3-5 Commonly used flags of the `errpt` command

Flags	Description
-a	Detailed output
-j <i>error identifier</i>	Includes only the error-log entries specified by the ErrorID (error identifier) variable
-s <i>StartDate</i>	Specifies all records posted on and after the StartDate variable
-T <i>ErrorTypeList</i>	Limits the error report to error types specified by the valid ErrorTypeList variables: INFO, PEND, PERF, PERM, TEMP, and UNKN

3.8.2 The w command

The `w` command prints a summary of current system activity.

The syntax of the `w` command is:

```
w [ -h ] [ -u ] [ -w ] [ -l | -s ] [ User ]
```

Some useful `w` command flags are provided in Table 3-6.

Table 3-6 Commonly used flags of the `w` command

Flags	Description
-u	Prints the time of day, amount of time since last system startup, number of users logged on, and number of processes running. Same output as the <code>uptime</code> command.

3.9 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

1. While a machine is booting up, several error messages are appearing on the screen. The user is not able to write down all of the errors. However, the user can refer to the console log file stored by default. Which of the following indicates where the console log file is located by default?
 - A. `/tmp/conslog`
 - B. `/tmp/console.log`
 - C. `/var/adm/ras/conslog`
 - D. `/var/adm/ras/console.log`
2. A system is hanging with a LED code of 581. This means that the system is hanging while running `/etc/rc.net`. Which of the following procedures should be performed next?
 - A. Run `rm /etc/rc.net` and then reboot.
 - B. Replace the network interface adapter.
 - C. Reboot the system into service mode and run `rmdev -d ent0`.
 - D. Reboot the system into service mode and run `mv /etc/rc.net /etc/rc.net.save`.

3. A file system is being mounted but failed. After running the **fsck** command the problem is still not resolved. Which of the following commands should run next?
- A. Run **savebase**.
 - B. Run **logform**.
 - C. Run **sync1vadm**.
 - D. Restore file system from **mksysb**.
4. After applying patches, no preventative steps were taken. As a result, the system hangs during the reboot with the following message: `starting tcp/ip daemons`. All of the following procedures are applicable to fixing the problem *except*:
- A. Checking `/etc/inittab`
 - B. Checking `/etc/rc.tcpip`
 - C. Checking name resolution
 - D. Running the **bosboot** command to fix bootable image
5. A system was rebooted in normal mode but hung after the multi-user initialization completed. The **alog -ot boot** command indicated the following:
- ```
Saving Base Customize Data to boot disk
Starting the sync daemon
Starting the error daemon
Starting Multi-user Initialization
Performing auto-varyon of Volume Groups
Activating all paging spaces
Swapon: Paging device /dev/hd6 activated.
/dev/rhd1 (/home): #;#; Unmounted cleanly - Check suppressed
Performing all automatic mounts
Multi-user initialization completed
```
- Once the system is booted into maintenance mode, which of the following procedures should be performed to first attempt to fix the problem?
- A. Run **fsck**.
  - B. Check the `/etc/rc` file for any incorrect entries.
  - C. Check `/etc/filesystems` for any incorrect entries.
  - D. Edit `/etc/inittab` and comment out the `rctcpip` and `rcnfs` lines.

6. A system was running fine until a power outage occurred during the reboot. As a result, the system LED displayed cycles between 223 and 229. Which of the following procedures should be performed next?
- A. Power off/up the system.
  - B. Boot into maintenance mode and run **fsck**.
  - C. Boot into maintenance mode and run **bosboot**.
  - D. Boot into maintenance mode and check the bootlist.

### 3.9.1 Answers

The following are the preferred answers to the questions provided in this section.

1. C
2. D
3. B
4. D
5. D
6. D

### 3.10 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

Do not perform these exercises on a production system or data:

1. Create a file system for this exercise and copy in some files to the file system, then destroy the first super block. This can be done by copying 4 KB from `/dev/zero` to block 1 on your logical volume. For example:

```
dd count=1 bs=4k seek=1 if=/dev/zero of=/dev/thomasclv
```

Try to mount the file system and run **fsck** on the file system to determine the problem. Finally, fix the problem as described in this chapter.

2. Still on your test system, with verified `mksysb` at hand, make a backup of `/etc/inittab`. Remove the first uncommented line and try to reboot. You are, at reboot, prompted for what?

After the boot has finished, edit the `/etc/inittab` and change a dot to a comma or a colon to semicolon on a line with `action=wait`. What happens? Which LED code is displayed? What do you have to do to fix this?



# Hardware problem determination

The following topics are discussed in this chapter:

- ▶ Hardware basics
- ▶ Running diagnostics
- ▶ SSA problem determination
- ▶ Three-digit display codes

This chapter discusses common hardware-related problem determination. It provides problem-resolving procedures based on the system architecture.

## 4.1 Hardware basics

RS/6000 and pSeries servers are available in a variety of models. A server can be in single processor or multiprocessor configurations. Currently, models comply to a number of architecture specifications, such as Micro Channel, PowerPC Reference Platform (PREP), Common Hardware Reference Platform (CHRP), and RS/6000 Platform Architecture (RPA).

The hardware platform type is an abstraction that allows machines to be grouped according to fundamental configuration characteristics, such as the number of processors or I/O bus structure. Machines with different hardware platform types have basic differences in the way their devices are dynamically configured at boot time. Currently available hardware platforms, which are able to be differentiated by software, in the RS/6000 and pSeries family are:

|                |                                                     |
|----------------|-----------------------------------------------------|
| <b>rs6k</b>    | Micro Channel-based uni-processor models            |
| <b>rs6ksmp</b> | Micro Channel-based symmetric multiprocessor models |
| <b>rspc</b>    | ISA-bus models                                      |
| <b>chrp</b>    | PCI-bus models                                      |

In order to determine the hardware platform type on your machine, enter the following command:

```
bootinfo -p
chrp
```

### 4.1.1 Hardware inventory

To determine a system's hardware inventory, use either the **lsdev** command or the **lscfg** command. These commands show different aspects of installed devices. The **lsdev** command displays information about devices in the device configuration database.

```
lsdev -C
sys0 Available 00-00 System Object
sysplanar0 Available 00-00 System Planar
pci0 Available 00-fef00000 PCI Bus
pci1 Available 00-fee00000 PCI Bus
pci2 Available 00-fed00000 PCI Bus
isa0 Available 10-58 ISA Bus
sa0 Available 01-S1 Standard I/O Serial Port
sa1 Available 01-S2 Standard I/O Serial Port
scsi1 Available 30-58 Wide SCSI I/O Controller
cd0 Available 10-60-00-4,0 SCSI Multimedia CD-ROM Drive
mem0 Available 00-00 Memory
proc0 Available 00-00 Processor
```

|          |                       |                                      |
|----------|-----------------------|--------------------------------------|
| proc1    | Available 00-01       | Processor                            |
| proc2    | Available 00-02       | Processor                            |
| proc3    | Available 00-03       | Processor                            |
| L2cache0 | Available 00-00       | L2 Cache                             |
| sioka0   | Available 01-K1-00    | Keyboard Adapter                     |
| fd0      | Available 01-D1-00-00 | Diskette Drive                       |
| rootvg   | <b>Defined</b>        | Volume group                         |
| hd5      | Defined               | Logical volume                       |
| tok0     | Available 10-68       | IBM PCI Tokenring Adapter (14103e00) |
| ent0     | Available 10-80       | IBM PCI Ethernet Adapter (22100020)  |
| ent1     | Available             |                                      |

The output shows whether the device is in the Available or Defined state.

Use the **lscfg** command to display vital product data (VPD) such as part numbers, serial numbers, microcode level, and engineering change levels from either the Customized VPD object class or platform-specific areas. To display all of these features for `hdisk1`, enter:

```
lscfg -vp -l hdisk1
 DEVICE LOCATION DESCRIPTION

 hdisk1 10-60-00-9,0 16 Bit SCSI Disk Drive (9100 MB)

 Manufacturer.....IBM
 Machine Type and Model.....DNES-309170W
 FRU Number.....25L3101
 ROS Level and ID.....53414730
 Serial Number.....AJ286572
 EC Level.....F42017
 Part Number.....25L1861
 Device Specific.(Z0).....000003029F00013A
 Device Specific.(Z1).....25L2871
 Device Specific.(Z2).....0933
 Device Specific.(Z3).....00038
 Device Specific.(Z4).....0001
 Device Specific.(Z5).....22
 Device Specific.(Z6).....F42036
```

PLATFORM SPECIFIC

```
Name: sd
Node: sd
Device Type: block
```

The most important fields in the previous example are:

- |                         |                                                                                               |
|-------------------------|-----------------------------------------------------------------------------------------------|
| <b>FRU Number</b>       | Use this number to order the same device in case of damage to the original one.               |
| <b>ROS Level and ID</b> | This is the microcode level, and it is used to determine the firmware version in your device. |

To display attribute characteristics and possible values of attributes for devices in the system, use the **lsattr** command:

```
lsattr -El hdisk1
pvid 000bc6ddc63c40380000000000000000 Physical volume identifier False
queue_depth 3 Queue DEPTH False
size_in_mb 9100 Size in Megabytes False
```

**Note:** It is a good practice to have a printout from the **lscfg**, **lsdev**, and **lsattr** commands to maintain and track your system inventory.

## 4.2 Running diagnostics

Hardware diagnostics can be run in three different ways:

- ▶ The first way is concurrent mode, where the system is up and running with users online, all processes running, and all volume groups being used.
- ▶ The second way is service mode. This is when you have the machine with AIX running, but with the minimum amount of processes started and only rootvg varied on.
- ▶ The third way is stand-alone diagnostics from CD-ROM. The CD-ROM-based diagnostics are a completely isolated version of AIX, so any diagnostics run are totally independent of the AIX setup on the machine being tested.

What method you select depends upon the circumstances, such as:

- ▶ Are you able to test the device? Is the device in use?
- ▶ Do you need to decide if the problem is related to hardware or AIX? Stand-alone diagnostics from CD-ROM or diskette are independent of the machine operating system. Advanced diagnostics run using the diagnostic CD-ROM or diskettes and completing successfully should be taken as proof of no hardware problem.

**Note:** If you are going to boot from a CD-ROM or a mksysb tape on a machine that has a configuration with two or more SCSI adapters sharing the same SCSI bus, check that no SCSI adapters on the shared bus are set at address 7. If you boot from bootable media, the bootable media will automatically assign address 7 to all SCSI adapters on the machine being booted. This will cause severe problems on any other machines sharing the same SCSI bus that have address 7 IDs set on their adapters.

The method you use to run diagnostics varies with the machine type. The next sections describe how to run all the diagnostic modes on the most common machine types.

There are a few RS/6000 models that do not have the capability to run AIX-based diagnostics. The most common of these are the 7020-40P and 7248-43P. To run diagnostics on these models, you must have the SMS diskette for the machine.

Maintenance mode is a function of the **shutdown -m** command, which is sometimes referred to as single-user mode. It provides a limited working environment where networking services and user access is limited.

## 4.2.1 Concurrent mode

Concurrent mode diagnostics are run while AIX is running on the machine and potentially sharing the environment with users. To run diagnostics concurrently, you must have root authority and use one of the following methods:

1. To run diagnostics on a specific device, use the following command:

```
diag -d [resource name]
```

This command enables you to test a specific device directly without the need to pass through a number of menus. The diagnostic process run is the Advanced Diagnostic process.

2. To go directly to the main diagnostics menu, use the **diag** command.
3. Using SMIT, select the following menus in the order provided.
  - a. Problem Determination
  - b. Hardware Diagnostics
  - c. Current shell

Methods 2 and 3 will present the entry screen of the diagnostics menu. If you press Enter, you will be provided a menu, as shown in Figure 4-1 on page 48.

Move cursor to selection, then press Enter.

**Diagnostic Routines**

This selection will test the machine hardware. Wrap plugs and other advanced functions will not be used.

**Advanced Diagnostics Routines**

This selection will test the machine hardware. Wrap plugs and other advanced functions will be used.

**Task Selection(Diagnostics, Advanced Diagnostics, Service Aids, etc.)**

This selection will list the tasks supported by these procedures. Once a task is selected, a resource menu may be presented showing all resources supported by the task.

**Resource Selection**

This selection will list the resources in the system that are supported by these procedures. Once a resource is selected, a task menu will be presented showing all tasks that can be run on the resource(s).

F1=Help

F10=Exit

F3=Previous Menu

Figure 4-1 Main diagnostics menu

The first three menu options shown in Figure 4-1 are explained in the following sections.

## Diagnostic Routines

This set of routines is primarily aimed at the operator of the machine. When the diagnostics are run using this option, there will be no prompts to unplug devices or cables, and no wrap plugs are used. Therefore, the testing done by this method is not as comprehensive as the testing performed under Advanced Diagnostics. In some cases, it can produce a No Trouble Found result when there is an actual problem.

## Advanced Diagnostics Routines

This set of routines will run diagnostic tests that will ask you to remove cables, plug and unplug wrap plugs, and use various other items. As a result, the tests run are as detailed as possible. Generally, if you get a No Trouble Found result using Advanced Diagnostics, you can be reasonably certain the devices tested have no hardware defects.

## Task Selection

This section is sometimes referred to as Service Aids. There are many useful tools within this section. The use of this option is discussed in 4.2.4, "Task selection or service aids" on page 52.

After you have selected the level of diagnostics you wish to run, you are presented with a menu to select the Problem Determination method or the System Verification method.

## **Problem Determination**

This selection will run the diagnostic routine and search the AIX error log for any errors posted in the previous 24 hours against the device you are testing. It will then use the sense data from any error log entry for the device being tested in conjunction with the results of the diagnostic testing of the device to produce a Service Request Number (SRN). This method must be used to determine the cause of any machine checks and checkstops on 7025 and 7026 machine types. If you are performing diagnostics more than seven days after the machine check occurred, then you will need to set the system date and time to within seven days of the machine check time stamp. The seven-day period is required when using AIX Version 4.3.1 and later. If you are using AIX Version 4.3.0 or earlier, the system date and time must be within 24 hours of the checkstop entry.

## **System Verification**

Use this selection if you have just replaced a part or performed a repair action. System verification runs a diagnostic routine on the device but does not refer to the AIX error log, so it reflects the machine's condition at the time of running the test. You can also use system verification when you just want to run a direct test to a device or whole machine.

Concurrent mode provides a way to run diagnostics online to system resources while AIX is up and running and users are logged on.

Since the system is running in normal operation, some resources cannot be tested in concurrent mode. The following list shows which resources cannot be tested:

- ▶ SCSI adapters used by disks connected to paging devices
- ▶ Disk drives used for paging
- ▶ Memory
- ▶ Processors

Depending on the status of the device being tested, there are four possible test scenarios in concurrent mode:

- ▶ Minimal testing is used when the device is under the control of another process.
- ▶ Partial testing occurs when testing is performed on an adapter or device that has some processes controlling part of it. For example, testing unconfigured ports on an 8-port RS232 adapter.

- ▶ Full testing requires that the device be unassigned and unused by any other process. Achieving this condition may require commands to be run prior to the commencement of the diagnostic testing.
- ▶ When tests are run for CPU or memory, the diagnostics refer to an entry in the NVRAM that records any CPU or memory errors generated during initial testing done at system power on. By analyzing these entries, the diagnostics produce any relevant SRNs.

## 4.2.2 Stand-alone diagnostics from disk - service mode

Service mode enables you to run tests to the devices that would ordinarily be busy if you ran diagnostics with the machine up in normal mode boot (for example, the network adapter ent0). However, you still will not be able to test any SCSI device that is attached to the same SCSI adapter as disks containing paging space or rootvg. Stand-alone diagnostics from disk are started when you boot up the machine in service mode boot. The method that you employ to get a service mode boot depends upon the type of machine.

### MCA machines

To start a service mode boot, power off the machine, then perform the following steps:

1. Set the key mode switch of the machine to the Service position.
2. Power on the machine without a CD-ROM, tape, or diskette in the machine.

After a period of time, you will see the Diagnostics Entry screen appear on the console. Press Enter and proceed to the screen that gives you the choice of diagnostics to run.

### PCI machines

This section applies to machines of model type 7017, 7024, 7025, 7026, 7043, 7046, and newer. It does not apply to PCI machine types 7020 or 7248.

To start a service mode boot, power off the machine, then perform the following steps:

1. Turn on the machine power.
2. After a short period of time, you will see the Icons screen. At this point, press F6 if using a graphics console, or 6 if using an ASCII terminal. If you are using the graphics console, the display device may have power saving enabled, and it will take time to warm up and display the icon images. This can cause you to miss the Icon screen being displayed. In this situation, observe the power LED on the display device, and when it changes from orange to green, press the F6 key.

Once the keyboard input has been processed, the machine will display a Software Starting screen. This is followed by more information that indicates that the SCSI ID of the boot device is being used. Once diagnostics have been loaded, you will have the Diagnostic Entry screen displayed.

### **4.2.3 Stand-alone diagnostics from CD-ROM**

Stand-alone diagnostics run from CD-ROM or diskettes is a good way of proving if the problem is a hardware or an AIX problem. The CD-ROM or diskettes load a totally independent version of AIX onto the machine as a RAM image. If you get a No Trouble Found result using advanced diagnostics using all of the test equipment asked for during the diagnostic, the probability of there being a hardware problem is extremely small. In such cases, the underlying cause of the problem is most often software related.

#### **MCA machines**

This section describes how to boot from CD-ROM on MCA machines and from diskette for the early level of MCA machines.

##### ***Boot from CD-ROM***

To boot from CD-ROM, complete the following steps:

1. Power off the machine.
2. Turn the key mode switch to the Service position.
3. Power on the machine and place the Diagnostic CD-ROM in the drive.

For the machine to boot from the Diagnostic CD-ROM, there must be an entry in the bootlist that includes the CD-ROM. Using the code on the CD-ROM, the machine will boot, eventually pausing when displaying c31 in the LED panel. The code c31 is an indication that you need to select a system console. After selecting a console at the prompt, the Diagnostic Entry screen is displayed, followed by subsequent screens. One of these subsequent screens will prompt you to enter the terminal type. Make sure you know the type before you proceed, since a wrong entry could result in you having to restart the process from the beginning.

#### **PCI bus machines**

This section applies to machines of model type 7017, 7024, 7025, 7026, 7043, 7046, and newer. It does not apply to PCI machine types 7020 or 7248.

To start a CD-ROM boot, use the following procedures:

1. Power off the machine.
2. Turn on machine power.

3. Place the CD-ROM into the drive.
4. After a short period of time, you will see the Icons screen. At this point, press F5 if you are using a graphics console, or 5 if you are using an ASCII terminal. If you are using the graphics console, sometimes the display screen will have power saving enabled, and will take time to warm up before anything can be seen on the screen. This can cause you to miss the Icon screen display. In this situation, observe the power LED on the display device, and when it changes from orange to green, then press the F5 key (E1F1 LEDs are shown).

After performing the previous steps, you will get various screens displayed, one of which will indicate to you the SCSI address of the device that the machine is booting from. Following this screen, the Diagnostic Entry screen is displayed.

## 4.2.4 Task selection or service aids

The diagnostics described in this section are known by two names: *task selection* or *service aids*, depending upon the level of diagnostics you are using. Task selection is the name used by AIX Version 4.3.2; however, in AIX Version 4.1.4, the same menu is known as service aids. This portion of the diagnostic package is equally as useful in the diagnosis of faults as the diagnostic routines themselves. The next few sections will cover a selection of the service aids available.

### Local area network service aid

This service aid is useful in the diagnosis of network problems. It enables you to type in IP addresses of both a source machine and a target machine. When activated, it will tell you if it managed to connect to the target machine. If it failed, it will try and give you a reason why it could not reach the destination host. The result of this can help in fault diagnosis.

### Microcode download

Using this service aid makes manipulation of microcode much easier than doing it from the command line. As a result, you are less liable to make a mistake.

The microcode download facility is also available when using the Diagnostic CD-ROM. This enables you to download microcode to devices that are not capable of being updated when AIX is running.

### SCSI bus analyzer

This is one of the most useful service aids. It enables you to issue a SCSI inquiry command to any device on any SCSI bus connected to the machine. The results that are returned give you a good idea of the problem.

The results returned are:

- ▶ The exerciser transmitted a SCSI inquiry command and did not receive any response back. Ensure that the address is valid, then try this option again.
- ▶ The exerciser transmitted a SCSI inquiry command and received a valid response back without any errors being detected.
- ▶ A check condition was returned from the device.

To run this service aid, perform the following steps:

1. From the Task Selection menu, select **SCSI Bus Analyzer**.
2. Next, select the adapter that has the device that you wish to test attached to it.
3. Use the Tab key to increment the SCSI ID field to the number you want to test.
4. Press F7 to confirm your selection.
5. Press Enter to commence the test.

If the device is working correctly, an affirmative system message should be returned almost instantly. If there is a problem, it should return an answer after a few seconds. Sometimes a device that has a severe check condition will hang the service aid. If this is the case, press Ctrl+C to exit from the service aid.

## Disk maintenance

The disk-to-disk copy will only work with SCSI disks that pass diagnostics and ideally have minimal errors when the certify process is run. If the error rate is too high when a disk-to-disk copy is being run, the program will fail. You will find it useful if the customer situation is such that they have no backup and the disk is unstable but running. Disk-to-disk copy differs from an AIX-based migrate operation because it does not alter the source disk when finished, as the **migratepv** command does. Disk-to-disk copy is best run from CD-ROM diagnostics, which requires you to have the exclusive use of the machine while the disk copying takes place. Also, the disk to be copied to *must not* be smaller or more than 10 percent larger in size than the source disk. The copied disk will have the same PVID as the original, so the defective disk must be removed from the machine before starting AIX.

## SSA service aids

This service aid can be used to help diagnose SSA subsystem problems. It is also used to physically identify and control SSA disks in the tower or drawer. This function greatly speeds the locating of specific disks, especially in very large installations.

**Note:** This service aid is only present when SSA devices are configured on the machine.

## 4.3 Serial Storage Architecture disks

The Serial Storage Architecture (SSA) disk subsystem is capable of being externally connected to one or more RS/6000 or pSeries systems. Certain models of RS/6000 and pSeries can also be configured with internal SSA disks. SSA devices are connected through two or more SSA links to an SSA adapter that is located in the system used. The devices, SSA links, and SSA adapters are configured in loops. Each loop provides a data path that starts at one connector of the SSA adapter and passes through a link (SSA cable) to the devices. The loop continues through the devices, then returns through another link to a second connector on the SSA adapter. Each adapter is capable of supporting two loops. Each loop can have between one and 48 devices. A loop can have as many as eight SSA adapters connected in up to eight systems, but this is dependent on the type of SSA adapter being used and how they are configured. Again, dependent on adapters, disk subsystem, and cables in use, the aggregate loop speed per adapter can either be 80 MB/s or 160 MB/s. As you can see, the number of possible combinations is almost endless and changes at each product announcement. The SSA configuration rules provided in the following section cover basic considerations.

### 4.3.1 General SSA setup rules

The following rules must be followed when connecting a 7133 or similar SSA subsystem:

- ▶ Each SSA loop must be connected to a valid pair of connectors on the SSA adapter card. A1 and A2 form one loop, and B1 and B2 form another loop.
- ▶ Only one pair of connectors of an SSA adapter can be connected in a particular SSA loop. A1 or A2, with B1 or B2, cannot be in the same SSA loop.
- ▶ A maximum of 48 disks can be connected in an SSA loop.
- ▶ A maximum of three dummy disk drive modules can be connected next to each other.
- ▶ A maximum of two adapters can be in the same host per SSA loop.
- ▶ Cables joining SSA nodes should not exceed 25 meters.
- ▶ There is no addressing setup for any SSA device.
- ▶ There is no termination since all connections should form a loop.

The maximum number of adapters per SSA loop at the time of this writing is provided in Table 4-1.

Table 4-1 SSA adapter information

| Feature code | Description                      | Identifier | Maximum number per loop                                                                                     |
|--------------|----------------------------------|------------|-------------------------------------------------------------------------------------------------------------|
| 6214         | MCA adapter                      | 4-D        | 2                                                                                                           |
| 6216         | MCA Enhanced SSA 4 port adapter  | 4-G        | 8                                                                                                           |
| 6217         | MCA SSA RAID adapter             | 4-I        | 1                                                                                                           |
| 6218         | PCI SSA RAID adapter             | 4-J        | 1                                                                                                           |
| 6219         | MCA Enhanced RAID adapter        | 4-M        | Between one and eight per loop, depending on microcode level and whether RAID and Fast Write Cache are used |
| 6215         | PCI Enhanced RAID adapter        | 4-N        |                                                                                                             |
| 6225         | PCI Advanced Serial RAID adapter | 4-P        |                                                                                                             |
| 6230         | Advanced SerialRAID Plus adapter | 4-P        |                                                                                                             |

For the most comprehensive and up-to-date information on SSA adapters, refer to the following URL:

<http://www.storage.ibm.com/hardsoft/products/ssa>

The user guides for each SSA adapter are also available on this Web site. They contain information about the valid adapter combinations allowed on the same loop.

### 4.3.2 SSA devices

SSA subsystem components use microcode to control their function. When working on SSA problems, you should ensure that the microcode level and any drivers on all devices in the loop are at the latest published level.

### 4.3.3 SSA disk considerations

If you configure an SSA disk into a system and it only shows as a pdisk with no corresponding hdisk, the most probable cause is that the disk was originally part of a RAID array set up on another machine. If disks are removed from a RAID array for any reason to be incorporated into any other system as a normal disk, the following procedure must be used:

1. Enter **smitty ssaraid** (the fast path to SSA RAID SMIT panels).

2. Select **Change Show use of an SSA Physical disk**. The disk must be returned to general use as an AIX system disk.
3. If the disk is to be removed from the system, use the relevant AIX commands. Do not remove the pdisk until you have removed the disk from the system using the SSA service aids.

If you are presented with this situation, and the disk with the problem was not a member of a RAID set on this machine, your only option to return this disk to normal use is to do a low-level format using the SSA service aid. This can take time if the disk is 9 GB or larger.

## SSA RAID

The SSA subsystem is capable of being operated by some adapters as either single system disks or as RAID LUNs. Provided that all has been set up correctly, then the RAID implementation works well. If you have any doubts as to how the RAID is set up, refer to *SSA Adapters: User's Guide and Maintenance Information*, SA33-3272.

If you need to do anything involving an SSA RAID array, then use the relevant procedure listed. This will ensure that the integrity of the RAID set is maintained at all times.

## Changing SSA disks

SSA disks are hot swappable. When preparing AIX for the removal of an SSA disk, do not use the **rmdev** command to remove the pdisk prior to physically removing the disk from the enclosure. You will need the pdisk to do the following steps. Use the **rmdev** command for the pdisk only when all steps are completed.

1. Use the SSA Service Aid to power the disk off prior to removal. This is done by using the set service mode and identify facility. This will put the disks on either side of the one you want to remove into string mode and power off the disk to be removed.
2. When the replacement disk or blanking module is inserted, use the same Service Aid to reset service mode. This will initialize the new disk and take the other disks out of string mode.
3. At this point, you can now run the **rmdev** command to remove the pdisk allocated to the disk you removed.
4. The disk change procedures will then tell you to run the **cfgmgr** command to create a new pdisk for the replaced physical disk.

**Note:** The `cfgmgr` command should *not* be executed on any system that is in an HACMP cluster. To do so may seriously damage the configuration of the machine, possibly resulting in the cluster going down.

If the disk to be changed is a defective RAID disk and was in use by the system, then you need to follow the procedures in *SSA Adapters: User's Guide and Maintenance Information*, SA33-3272. Read these procedures carefully because some of the earlier editions of this publication indicate you have finished the procedure when, in fact, you need to perform other steps to return the array to a protected state. Below is a list of the important steps that need to be completed before you can be sure that the array will function correctly.

Steps involved in the replacement of a RAID SSA disk are:

1. Addition of the replacement disk to the system using the `cfgmgr` command or the `mkdev` command on HACMP systems.
2. Make the disk an array candidate or hot spare using SMIT.

If the disk was removed from a RAID array leaving it in an exposed or degraded state, you now need to add the disk to the array using SMIT. While the array is being rebuilt, error messages will be seen each hour in the error log. These will cease when the array is completely rebuilt. It is best to schedule disk swaps during scheduled downtime to minimize the effects on the system.

### 4.3.4 Three-digit display values

Three-digit display messages are system-error indicators that display on the system operator panel. Most of the three-digit display values are progress indicators that only display briefly. This section enables you to interpret the codes displayed on the system operator panel.

### 4.3.5 Common boot time LEDs

The following sections cover some hardware-related problems that can cause a halt. All problems at this stage of the startup process have an error code defined, which is shown in the LED display on the front panel.

#### **LED 200**

The LED code 200 is connected to the secure key position. When the key is in the secure position, the boot will stop until the key is turned, either to the normal position or the service position; then the boot will continue.

## LED 299

A LED code of 299 shows that the BLV will be loaded. If this LED code is passed, then the load has been successful. If, after passing 299, you get a stable 201, then you have to recreate the BLV.

## MCA LED codes

Table 4-2 provides a list of the most common LED codes on MCA systems. More of these can be found in the AIX Version 4 base documentation.

Table 4-2 Common MCA LED codes

| LED                                                    | Description                                                                                                                                                                                                                            |
|--------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 100–195                                                | Hardware problem during BIST.                                                                                                                                                                                                          |
| 200                                                    | Key mode switch in secure position.                                                                                                                                                                                                    |
| 201                                                    | If LED 299 passed, recreate BLV.<br>If LED 299 has not passed, POST encountered a hardware error.                                                                                                                                      |
| 221<br>721<br>221–229<br>223–229<br>225–229<br>233–235 | The bootlist in NVRAM is incorrect (boot from media and change the bootlist), or<br>the bootlist device has no bootimage (boot from media and recreate the BLV), or<br>the bootlist device is unavailable (check for hardware errors). |

### 4.3.6 888 in the three-digit display

A flashing 888 indicates that a problem was detected, but could not be displayed on the console. A message is encoded as a string of three-digit display values. The 888 will be followed by either a 102, 103, or 105. The Reset button is used to scroll the message.

#### The 102 code

A 102 indicates that a dump has occurred and your AIX kernel crashed due to component failure. A LED code description is provided in the following list:

- ▶ 888 - This value flashes to indicate a system crash.
- ▶ 102 - This value indicates an unexpected system halt.
- ▶ nnn - This value is the cause of the system halt (reason code).
- ▶ 0cx - The value 0cx indicates dump status.

The reason code is the second value displayed after 888 appears. Also, this code can be found using the `stat` subcommand in `crash`.

- ▶ 000 - Unexpected system interrupt (hardware related).
- ▶ 2xx - Machine check. A machine check can occur due to hardware problems (for example, bad memory) or because of a software reference to a non-existent address.
- ▶ 3xx - Data storage interrupt (DSI). A page fault always begins as a DSI, which is handled in the exception processing of the VMM. However, if a page fault cannot be resolved, or if a page fault occurs when interrupts are disabled, the DSI will cause a system crash. The page fault may not be resolved if, for example, an attempt is made to read or write a pointer that has been freed; in other words, the segment register value is no longer valid, and the address is no longer mapped.
- ▶ 400 - Instruction access exception. This is similar to a DSI, but occurs when fetching instructions, not data.
- ▶ 5xx - External interrupt. Interrupt arriving from an external device.
- ▶ 700 - Program interrupt. Usually caused by a trap instruction that can be a result of failing an *assert*, or hitting a *panic* within a kernel or kernel extension code.
- ▶ 800 - Floating point unavailable. An attempt is made to execute a floating point instruction but the floating point available bit in the Machine Status Register (MSR) is disabled.

For more information about system dumps, see Chapter 5, “System dumps” on page 65.

### The 103 and 105 codes

A 103 message indicates that a Service Request Number (SRN) follows the 103. The SRN consists of the two sets of digits following the 103 message. This number together with other system-related data is used to analyze the problem. Record and report the SRN to your service representative.

A 105 message indicates that an encoded SRN follows the 105. Record and report SRN 111-108 to your service representative. The format is shown in Figure 4-2 on page 60.

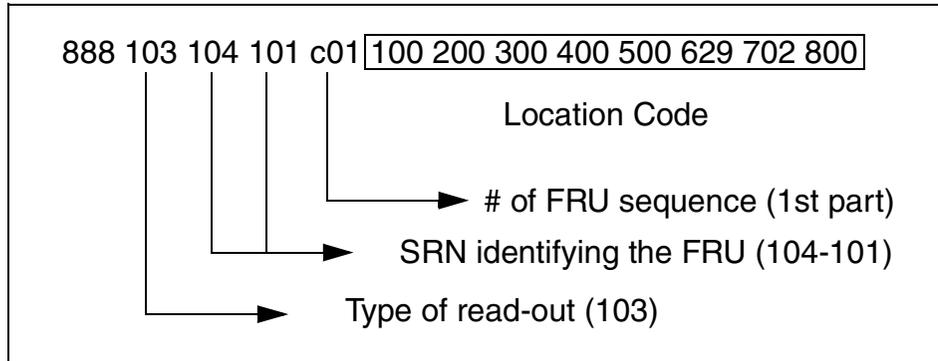


Figure 4-2 Format of the 103 code message

The fifth value identifies the FRU number (number of the defective part). Because more than one part could be described in the 888 message, the next eight identifiers describe the location code of the defective part. These should be mapped with the values provided in Table 4-3 to identify the location code.

Table 4-3 Location code mapping table

|        |        |        |        |
|--------|--------|--------|--------|
| 00 = 0 | 09 = 9 | 19 = I | 28 = S |
| 01 = 1 | 11 = A | 20 = J | 30 = T |
| 02 = 2 | 12 = B | 21 = K | 31 = U |
| 03 = 3 | 13 = C | 22 = L | 32 = V |
| 04 = 4 | 14 = D | 23 = M | 33 = W |
| 05 = 5 | 15 = E | 24 = N | 34 = X |
| 06 = 6 | 16 = F | 25 = O | 35 = Y |
| 07 = 7 | 17 = G | 26 = P | 36 = Z |
| 08 = 8 | 18 = H | 27 = R |        |

## 4.4 Command summary

The following section provides a list of the key commands discussed in this chapter.

## 4.4.1 The chdev command

The **chdev** command changes the characteristics of a device. The command has the following syntax:

```
chdev -l Name [-a Attribute=Value ...]
```

The commonly used flags are provided in Table 4-4.

Table 4-4 Commonly used flags of the chdev command

| Flag               | Description                                                                                                                                        |
|--------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| -l Name            | Specifies the device logical name, specified by the name parameter, in the Customized Devices object class whose characteristics are to be changed |
| -a Attribute=Value | Specifies the device attribute value pairs used for changing specific attribute values                                                             |

## 4.4.2 The lsattr command

The **lsattr** command displays attribute characteristics and possible values of attributes for devices in the system. The command has the following syntax:

```
lsattr -E -l Name [-a Attribute] ...
```

The commonly used flags are provided in Table 4-5.

Table 4-5 Commonly used flags of the lsattr command

| Flag         | Description                                                                                                                  |
|--------------|------------------------------------------------------------------------------------------------------------------------------|
| -E           | Displays the attribute names, current values, descriptions, and user-settable flag values for a specific device              |
| -l Name      | Specifies the device logical name in the Customized Devices object class whose attribute names or values are to be displayed |
| -a Attribute | Displays information for the specified attributes of a specific device or kind of device                                     |

## 4.5 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

1. Which of the following commands should be used to determine the microcode level of a system?
  - A. `lsattr -E1`
  - B. `lscfg -vl`
  - C. `lsCs ssa`
  - D. `lsdev -Cc disk`
2. After a legacy, the microchannel system has gone down with flashing 888s. Which of the following procedures is the best way to diagnose the problem?
  - A. Turn the power off and back on.
  - B. Reboot the system in maintenance mode.
  - C. Turn the key to service and press the Reset button to take a dump.
  - D. Verify that the key is in normal mode and press the Reset button to obtain error codes.
3. Which of the following AIX commands should be used to determine if there is a Service Request Number (SRN) on a device?
  - A. `diag`
  - B. `lssrn`
  - C. `lsdev`
  - D. `errpt`
4. Using information provided in Figure 4-3 on page 63, all of the following are true except:
  - A. The error was unrecoverable.
  - B. The termination of the SCSI bus failed.
  - C. A defective cable or terminator probably caused the error.
  - D. The presence of all zeros in the sense data indicates invalid data.

```
LABEL: SCSI_ERR1
ID: 0502F666

Date/Time: Jun 19 22:29:51
Sequence Number: 95
Machine ID: 123456789012
Node ID: host1
Class: H
Type: PERM
Resource Name: scsi0
Resource Class: adapter
Resource Type: hscsi
Location: 00-08
VPD:
Device Driver Level.....00
Diagnostic Level.....00
Displayable Message.....SCSI
EC Level.....C25928
FRU Number.....30F8834
Manufacturer.....IBM97F
Part Number.....59F4566
Serial Number.....00002849
ROS Level and ID.....24
Read/Write Register Ptr....0120
Description
ADAPTER ERROR

Probable Causes
ADAPTER HARDWARE CABLE
CABLE TERMINATOR DEVICE

Failure Causes
ADAPTER
CABLE LOOSE OR DEFECTIVE

Recommended Actions
PERFORM PROBLEM DETERMINATION PROCEDURES
CHECK CABLE AND ITS CONNECTIONS

Detail Data
SENSE DATA
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
```

Figure 4-3 Case study

- 5. A mirrored SSA data disk volume group must have a disk replaced. Which of the following concerns should be considered?
  - A. Schedule down time for rebooting.
  - B. Schedule down time for replacement of disk.
  - C. Schedule down time for replacement of disk and reboot.
  - D. Schedule disk replacement for non-peak usage time.

### 4.5.1 Answers

The following are the preferred answers to the questions provided in this section.

1. B
2. D
3. A
4. D
5. B

### 4.6 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

1. Take a hardware inventory of your system.
2. Check all possible menus in the concurrent mode diagnostics.



## System dumps

In this chapter, the system dump is discussed with respect to how that dump is managed and read. The way to set up the dump device will also be discussed.

A system dump is created when the system has an unexpected system halt or a system failure. The dump will be a snapshot of the system at the time of the dump; it does not collect data about what happened before the system dump. This dump is written to the primary dump device; if this is not available, it will write the dump to the secondary device. A system dump can also be initiated by a user using a different device (if required).

## 5.1 Configuring the dump device

Prior to AIX Version 4.1, the default dump device is /dev/hd7. In AIX versions after 4.1, the default dump device is /dev/hd6, which is the default paging space logical volume (/dev/paging $nn$  for dumps). The secondary dump device is /dev/sysdumpnull. Once the system is booted, this image is copied from /dev/hd6 to the directory /var/adm/ras. AIX 5L Version 5.1 servers with a real memory size larger than 4 GB will, at installation time, have a dedicated dump device created. This dump device is automatically created and no user intervention is required. The default name of the dump device is lg\_dumplv.

The current dump configuration can be determined by running the **sysdumpdev** command as follows:

```
sysdumpdev
primary /dev/hd6
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump FALSE
dump compression OFF
```

The primary dump devices must always be in the root volume group for permanent dump devices. The secondary device may be outside the root volume group unless it is a paging space.

**Note:** Do not use a mirrored or copied logical volume as the active dump device. Carefully check your AIX release to see if this function is available. System dump error messages will not be displayed, and any subsequent dumps to a mirrored logical volume will fail.

Do not use a diskette drive as your dump device.

AIX Version 4.2.1 and later supports using any paging device in the root volume group (rootvg) as the secondary dump device.

The **sysdumpdev** command can be used to configure remote dump devices. The following conditions must be met before a remote dump device can be configured:

- ▶ The local and the remote host must have Transmission Control Protocol/Internet Protocol (TCP/IP) installed and configured.
- ▶ The local host must have the Network File System (NFS) installed.
- ▶ The remote host must support NFS.

- ▶ The remote host must be operational and on the network. This condition can be tested by issuing the **ping** command.
- ▶ The remote host must have an NFS exported directory defined such that the local host has read and write permissions as well as root access to the dump file on the remote host.
- ▶ The remote host cannot be the same as the local host.

To change a primary dump device permanently, use the **sysdumpdev** command as follows:

```
sysdumpdev -P -p /dev/hd3
primary /dev/hd3
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump FALSE
dump compression OFF
```

This will remain the permanent dump device until it is changed again with the **sysdumpdev** command.

To change the secondary device permanently, use the **sysdumpdev** command as follows:

```
sysdumpdev -P -s /dev/rmt0
primary /dev/hd3
secondary /dev/rmt0
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump FALSE
dump compression OFF
```

To temporarily change the primary device to another device, use the **sysdumpdev** command as follows:

```
sysdumpdev -p /dev/rmt0
primary /dev/rmt0
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump FALSE
dump compression OFF
```

This will temporarily change the primary dump device to `/dev/rmt0` until the next system reboot.

## 5.2 Starting a system dump

A user-initiated dump is different from a dump initiated by an unexpected system halt because the user can designate which dump device to use. When the system halts unexpectedly, a system dump is automatically initiated to the primary dump device. Do not start a system dump if the flashing 888 number shows in your operator panel display. This number indicates that your system has already created a system dump and written the information to your primary dump device. If you start your own dump before copying the information in your dump device, your new dump will overwrite the existing information.

You can start a system dump by using one of the methods listed below.

If you have the Software Service Aids Package installed, you have access to the **sysdumpstart** command and can start a dump using one of these methods:

- ▶ Use the command line.
- ▶ Use SMIT.

If you do not have the Software Services Aids Package installed, you must use one of these methods to start a dump:

- ▶ Use the Reset button.
- ▶ Use special key sequences.

### 5.2.1 Using the command line

To create a system dump, use the following steps to choose a dump device, initiate the system dump, and determine the status of the system dump.

Check which dump device is appropriate for your system (the primary or secondary device) by using the following **sysdumpdev** command:

```
sysdumpdev -l
primary /dev/hd6
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump FALSE
dump compression OFF
```

This command lists the current dump devices. You can use the **sysdumpdev** command to change device assignments.

Start the system dump by entering the following **sysdumpstart** command:

```
sysdumpstart -p
```

This command starts a system dump on the default primary dump device. You can use the `-s` flag to specify the secondary dump device. If a code shows in the operator panel display, refer to 5.3, “System dump status check” on page 74, for more information.

If the dump was successful, reboot the system. During the boot process, if the forced copy flag is set to `TRUE`, a menu will be displayed on the primary console requesting the removable media to copy the dump to `/dev/rmtx` or `/dev/fd0`. (You are prompted to choose which location.) The size of the dump in `/dev/hd6` is also displayed. It is advisable to not use `/dev/fd0` for the copy of the dump. Once the copy has been completed, exit the copy screen and the system will continue the boot process.

The `sysdumpdev -K` command will set the force copy flag to true.

## 5.2.2 Using the SMIT interface

Use the following SMIT command to choose a dump device and start the system dump:

```
smit dump
```

The Choose the Show Current Dump Devices option can be used to note the available dump devices.

Select either the primary or secondary dump device to hold your dump information, as shown in Figure 5-1.

```
System Dump
Move cursor to desired item and press Enter.

Show Current Dump Devices
Show Information About the Previous System Dump
Show Estimated Dump Size
Change the Primary Dump Device
Change the Secondary Dump Device
Change the Directory to which Dump is Copied on Boot
Start a Dump to the Primary Dump Device
Start a Dump to the Secondary Dump Device
Copy a System Dump from a Dump Device to a File
Copy a System Dump from a Dump Device to Diskette
Always ALLOW System Dump
System Dump Compression

F1=Help F2=Refresh F3=Cancel F8=Image
F9=Shell F10=Exit Enter=Do
```

Figure 5-1 SMIT dump screen

A command status screen will be displayed, and once the dump has completed, the system will need to be reset.

If the dump was successful, reboot the system. During the boot process, if the forced copy flag is set to TRUE, a menu will be displayed on the primary console requesting the removable media to copy the dump to /dev/rmtx or /dev/fd0. (You are prompted to choose which location.) The size of the dump in /dev/hd6 is also displayed. It is advisable to not use /dev/fd0 for the copy of the dump. Once the copy has been completed, exit the copy screen and the system will continue the boot process.

### 5.2.3 Using the Reset button

To start a dump with the Reset button, the key switch must be in the service position. If the system does not have a key switch, set the always allow system dump value to true. To set this, use the **sysdumpdev** command as follows:

```
sysdumpdev -K
```

The value can be checked using the **sysdumpdev** command without flags as follows:

```
sysdumpdev
primary /dev/hd6
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump TRUE
dump compression OFF
```

To obtain the system dump, press the Reset button. This will initiate the system dump and may take some time.

If the dump was successful, reboot the system. During the boot process, if the forced copy flag is set to TRUE (**sysdumpdev -D dumpdev**), a menu will be displayed on the primary console requesting the removable media to copy the dump to /dev/rmtx or /dev/fd0. (You are prompted to choose which location.) The size of the dump in /dev/hd6 is also displayed. It is advisable to not use /dev/fd0 for the copy of the dump. Once the copy has been completed, exit the copy screen and the system will continue the boot process.

If the system does not have a key switch, set the always allow dump option back to false. Use the **sysdumpdev** command as follows:

```
sysdumpdev -k
```

Ensure that the always allow dump option has been set back to FALSE using the **sysdumpdev** command, as follows:

```
sysdumpdev
primary /dev/hd6
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump FALSE
dump compression OFF
```

## 5.2.4 Using special key sequences

To start a dump with a key sequence, you must have the key switch in the service position, or have set the always allow dump value to true. To set this, use the **sysdumpdev** command as follows:

```
sysdumpdev -K
```

The value can be checked using the **sysdumpdev** command without flags, as follows:

```
sysdumpdev
primary /dev/hd6
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump TRUE
dump compression OFF
```

Press the Ctrl+Alt 1 key sequence to write the dump information to the primary dump device.

Press the Ctrl+Alt 2 key sequence to write the dump information to the secondary dump device.

Both of these key sequences will initiate the system dump, and this process may take some time.

If the dump was successful, reboot the system. During the boot process, if the forced copy flag is set to TRUE, a menu will be displayed on the primary console requesting the removable media to copy the dump to /dev/rmtx or /dev/fd0. (You are prompted to choose which location.) The size of the dump in /dev/hd6 is also displayed. It is advisable to not use /dev/fd0 for the copy of the dump. Once the copy has been completed, exit the copy screen and the system will continue the boot process.

If the system does not have a key switch to set the always allow dump value back to false, use the **sysdumpdev** command as follows:

```
sysdumpdev -k
```

Ensure that the always allow dump option has been set back to FALSE by using the **sysdumpdev** command, as follows:

```
sysdumpdev
primary /dev/hd6
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump FALSE
dump compression OFF
```

## The TTY remote reboot

AIX Version 4.3.2 has added the ability to do a remote reboot of a system across native serial ports by using a user-defined string. This feature is configured by setting up two ODM attributes that have been added to the native serial ports. Figure 5-2 shows the options as they are set up in the SMIT screen.

```

 Add a TTY

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

[MORE...14] [Entry Fields]
STTY attributes for RUN time [hupcl,cread,brkint,icr> +
STTY attributes for LOGIN [hupcl,cread,echoe,cs8]
LOGGER name []
STATUS of device at BOOT time [available] +
REMOTE reboot ENABLE no +
REMOTE reboot STRING [#@reb@#]
TRANSMIT buffer count [16] +#
RECEIVE trigger level [3] +#
STREAMS modules to be pushed at OPEN time [ldterm] +
INPUT map file [none] +
OUTPUT map file [none] +
CODESET map file [sbcs] +

[MORE...17]

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

```

Figure 5-2 SMIT Add a TTY screen - Remote reboot options

The settings for the REMOTE reboot ENABLE attribute are described in Table 5-1 on page 73.

Table 5-1 REMOTE reboot ENABLE settings

| REMOTE reboot ENABLE settings | Description                                                                            |
|-------------------------------|----------------------------------------------------------------------------------------|
| no                            | Remote reboot is disabled and no action will be taken if the reboot string is entered. |
| reboot                        | If the reboot string is entered, the system will reboot.                               |
| dump                          | When the reboot string is entered, the system will execute a system dump.              |

The REMOTE reboot STRING option is a user-defined string that can be used to perform the function as set up in the REMOTE reboot ENABLE option.

## 5.2.5 System Hang Detection

AIX 5L introduces System Hang Detection (SHD) feature which provides a mechanism to detect a system hang situation and allows you a method to recover. This feature is implemented as the shdaemon daemon that runs at the highest process priority. This daemon queries the kernel for the lowest priority thread run over a specified interval. The shdaemon daemon is launched by init and it runs at priority 0 (zero).

If the priority is above a configured threshold, the daemon can take one of several actions. Each of these actions can be independently enabled, and each can be configured to trigger at any priority and over any time interval. Table 5-2 lists the default values when the SHD is enabled.

Table 5-2 SHD default values

| Option                      | Enablement | Priority | Timeout seconds |
|-----------------------------|------------|----------|-----------------|
| Log an error in errlog file | Disable    | 60       | 120             |
| Display a warning message   | Disable    | 60       | 120             |
| Give a recovering getty     | Enabled    | 60       | 120             |
| Launch a command            | Disabled   | 60       | 120             |
| Reboot the system           | Disabled   | 39       | 300             |

You can manage the SHD configuration using the `smitty shd` fast path to get to the SHD configuration menu.

The **shconf** command is invoked when SHD is enabled. It manages the system hang detection parameters.

The **shconf** command syntax is as follows:

```
shconf [-D | -E] [-O] -l Name
shconf -l Name -a Attribute=Value ...
```

## 5.3 System dump status check

When a system dump is taking place, status and completion codes are displayed in the operator panel display. When the dump is complete, a 0cx status code displays if the dump was user initiated, and a flashing 888 displays if the dump was system initiated.

You can check whether the dump was successful, and if not, what caused the dump to fail, if a 0cx code is displayed.

**Note:** If the dump fails, upon reboot, look for an error log entry with the label DSI\_PROC or ISI\_PROC. If the Detailed Data area shows an EXVAL of 000 0005, this is probably a paging space I/O error. If the paging space is the dump device or on the same hard drive as the dump device, the dump may have failed due to a problem with the hard drive. Diagnostics should be run against that disk.

### 5.3.1 Status codes

The following are the list of status codes for the system dump:

- 000** The kernel debugger is started. If there is an ASCII terminal attached to one of the native serial ports, enter **q dump** at the debugger prompt (>) on that terminal and then wait for the flashing 888s to appear in the operator panel display. After the flashing 888 appears, go to 5.6, “Copying a system dump” on page 77, which describes how to check the dump status.
- 0c0** The dump completed successfully. Go to 5.6, “Copying a system dump” on page 77.
- 0c1** An I/O error occurred during the dump.
- 0c2** A user-requested dump is not finished. Wait at least one minute for the dump to complete and for the operator panel display value to change. If the operator panel display value changes, find the new value on this list. If the value does not change, then the dump did not complete due to an

unexpected error. Complete the Problem Summary Form, and report the problem to your software service department.

- 0c4** The dump ran out of space. A partial dump was written to the dump device, but there is not enough space on the dump device to contain the entire dump. To prevent this problem from occurring again, you must increase the size of your dump media. Go to 5.4, “Increasing the size of the dump device” on page 76.
- 0c5** The dump failed due to an internal error. Wait at least one minute for the dump to complete and for the operator panel display value to change. If the operator panel display value changes, find the new value on the list. If the value does not change, then the dump did not complete due to an unexpected error. Complete the Problem Summary Form and report the problem to your software service department.
- 0c7** A network dump is in progress, and the host is waiting for the server to respond. The value in the operator panel display should alternate between 0c7 and 0c2 or 0c9. If the value does not change, then the dump did not complete due to an unexpected error. Complete the Problem Summary Form, and report the problem to your software service department.
- 0c8** The dump device has been disabled. The current system configuration does not designate a device for the requested dump. Enter the `sysdumpdev` command to configure the dump device.
- 0c9** A dump started by the system did not complete. Wait at least one minute for the dump to complete and for the operator panel display value to change. If the operator panel display value changes, find the new value on the list. If the value does not change, then the dump did not complete due to an unexpected error. Complete the Problem Summary Form and report the problem to your software service department.
- 0cc** (For AIX Version 4.2.1 and later only) An error occurred dumping to the primary device; the dump has switched over to the secondary device. Wait at least one minute for the dump to complete and for the three-digit display value to change. If the three-digit display value changes, find the new value on this list. If the value does not change, then the dump did not complete due to an unexpected error. Complete the Problem Summary Form and report the problem to your software service department.
- c20** The kernel debugger exited without a request for a system dump. Enter the quit dump subcommand. Read the new three-digit value from the LED display.

## 5.4 Increasing the size of the dump device

The size required for a dump is not a constant value, because the system does not dump paging space; only data that resides in real memory can be dumped. Paging space logical volumes will generally hold the system dump. However, because an incomplete dump may not be usable, use the procedure below to make sure that you have enough dump space.

When a system dump occurs, all of the kernel segment that resides in real memory is dumped (the kernel segment is segment 0). Memory resident user data (such as u-blocks) is also dumped.

The minimum size for the dump space can best be determined using the **sysdumpdev -e** command. This provides an estimated dump size, taking into account the memory currently in use by the system, as shown in the following example:

```
sysdumpdev -e
0453-041 Estimated dump size in bytes: 38797312
```

If the dump device is the default dump device of /dev/hd6, use the **lspv -a** command to check paging space available, as follows:

```
lspv -a
Page Space Physical Volume Volume Group Size %Used Active Auto Type
hd6 hdisk0 rootvg 512MB 1 yes yes lv
```

If the size of the dump device needs to be increased, use the **smit chps** command and change the paging space size. If the dump device is a file, ensure that the file system has enough space; if not, use the **smit chfs** command to increase the size of the file system.

Starting with AIX 5.x, the **dumpcheck** facility will notify you if your dump device needs to be larger, or the file system containing the copy directory is too small. It will also automatically turn compression on if this will alleviate these conditions. This notification appears in the system error log.

## 5.5 Configuring remote dump devices

The **sysdumpdev** command can also be used to configure remote dump devices. The following conditions must be met before a remote dump device can be configured:

- ▶ The local and remote host must have Transmission Control Protocol/Internet Protocol (TCP/IP) installed and configured.
- ▶ The local host must have Network File System (NFS) installed.

- ▶ The remote host must support NFS.
- ▶ The remote host must be configured on the network (ping address).
- ▶ The remote host must have an NFS exported directory defined such that the local host has read and write permissions as well as root access to the dump file on the remote host.
- ▶ The remote host cannot be the same as the local host.
- ▶ The network device driver must support remote dump.

To designate remote dump file `/var/adm/ras/systemdump` on host `server4` for a primary dump device, enter:

```
sysdumpdev -p server4:/var/adm/ras/systemdump
```

A colon (:) must be inserted between the host name and the file name.

## 5.6 Copying a system dump

If the dump is not copied to an external device during boot, it can be copied to the external device using the **snap** command. The **snap** command will check for an existing dump on the system and copy it to tape or, if no dump is available on the system, it will prompt for the dump to be copied from the external device.

The last system dump can be checked using the **sysdumpdev** command as follows:

```
sysdumpdev -L
0453-039

Device name: /dev/hd6
Major device number: 10
Minor device number: 2
Size: 42568192 bytes
Date/Time: Wed Jul 12 14:53:55 CDT 2000
Dump status: 0
dump completed successfully
Dump copy filename: /usr/dumpdir/vmcore.0
```

In this case, the dump was successfully completed and it can be copied to an external media device, such as tape.

Use the **snap** command (as follows) to copy the dump to tape. The flags indicate that general operating system, file system, and kernel information, along with the kernel dump, are copied to a tape device:

```
snap -gfkD -o /dev/rmt0
```

```

Setting output device to /dev/rmt0... done.
Checking space requirement for general
information.....
..... done.
Checking space requirement for kernel information..... done.
Checking space requirement for dump information..... done.
Checking space requirement for fileysys information.....
done.
Checking for enough free space in filesystem... done.

*****Checking and initializing directory structure
Directory /tmp/ibmsupt/filesys already exists... skipping
Directory /tmp/ibmsupt/dump already exists... skipping
Directory /tmp/ibmsupt/kernel already exists... skipping
Directory /tmp/ibmsupt/general already exists... skipping
Directory /tmp/ibmsupt/testcase already exists... skipping
Directory /tmp/ibmsupt/other already exists... skipping
*****Finished setting up directory /tmp/ibmsupt

Gathering general system
information.....
..... done.
Gathering kernel system information..... done.
Gathering dump system information.... done.
Gathering fileysys system information..... done.

Copying information to /dev/rmt0... Please wait... done.

***** Please Write-Protect the output device now...

***** Please label your tape(s) as follows:
***** snap blocksize=512
***** problem: xxxxx Wed Jul 12 15:41:42 CDT 2000
***** 'your name or company's name here'

The dump file can be copied from the external device using the tar -x
command. To view the contents of the tape device, use the following command:

tar -tvf /dev/rmt0
drwx----- 0 0 0 Jul 12 13:48:44 2000 ./dump/

```

```

-rw----- 0 0 2555 Jul 12 15:40:21 2000 ./dump/dump.snap
-rw----- 0 0 1770955 Jul 12 13:48:29 2000 ./dump/unix.Z
-rwx----- 0 0 41761792 Jul 12 11:03:29 2000 ./dump/dump_file
...
drwx----- 0 0 0 Jul 12 11:23:06 2000 ./kernel/
-rw----- 0 0 75122 Jul 12 15:40:21 2000 ./kernel/kernel.snap
drwx----- 0 0 0 Jul 12 11:22:58 2000 ./testcase/
drwx----- 0 0 0 Jul 12 11:22:58 2000 ./other/

```

The files `dump.snap`, `unix.Z`, and `dump_file` should exist on the tape device and should be greater than 0 bytes in size.

## 5.7 Reading dumps

To check that the dump is readable, start the **crash** command (or use the **kdb** command on AIX 5L systems). Refer to 10.6.2, “The kdb command” on page 296, for details on the dump files, using the command syntax: **crash dump unix**. The **crash** command needs a kernel file (`unix`) to match the dump file. If you do not specify a kernel file, **crash** uses the file `/unix` by default:

```

crash dump unix
>

```

If you do not see a message from **crash** about dump routines failing, you probably have a valid dump file. Run the `stat` subcommand at the `>` prompt, as in the following example:

```

crash dump unix
> stat
 sysname: AIX
 nodename: sp5i
 release: 3
 version: 4
 machine: 000126774C00
 time of crash: Tue May 4 04:56:10 CDT 1999
 age of system: 4 min.
 xmalloc debug: disabled
 abend code: 300
 csa: 0x2ff3b400
 exception struct:
 dar: 0x00000003
 dsisr: 0x00000000:
 srv: 0x04000000
 dar2: 0x3c160040
 dsirr: 0x06001000: "(unknown reason code)"

```

Look at the time of the dump and the abend code. If these are related to the problem causing the dump, then perform some initial analysis. Refer to 5.9.1, “The crash command” on page 83, for more information.

A message stating dumpfile does not appear to match namelist means the dump is not valid. For example:

```
crash dump unix
Cannot locate offset 0x02052b8 in segment 0x000000.
endcomm 0x00000000/0x011c5e70
WARNING: dumpfile does not appear to match namelist
Cannot locate offset 0x00ccf10 in segment 0x000000.
0452-179: Cannot read v structure from address 0x ccf10.
Symbol proc has null value.
Symbol thread has null value.
Cannot locate offset 0x00ccf10 in segment 0x000000.
0452-179: Cannot read v structure from address 0x ccf10.
Cannot locate offset 0x00034c4 in segment 0x000000.
0452-1002: Cannot read extension segment value from address 0x 34c4
```

Any other messages displayed when starting **crash** may indicate that certain components of the dump are invalid, but these are generally handled by **crash**. If a required component of the dump image is missing, additional messages will indicate this, and the dump should be considered invalid. To prevent problems, it is a good idea to use **crash** from the same level of AIX as that from the machine that created the dump.

Table 5-3 highlights the differences in dump between AIX Version 5.x and 4.x.

*Table 5-3 Differences in system dump*

| AIX Version 5.x                                                                                       | AIX Version 4.x                                                              |
|-------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| Use the <b>kdb</b> command to read dumps.                                                             | Use the <b>crash</b> command to read dumps.                                  |
| Use the <b>dumpcheck</b> command to check if the dump device is large enough.                         | Use the <b>sysdumpdev</b> command to check if the dump devices large enough. |
| An application can create a core file by using the <b>coredump()</b> system call.                     | Not available in AIX Version 4.x.                                            |
| Servers with a real memory size larger than 4 GB will have a dedicated dump device created.           | The paging space is used as the default dump device created.                 |
| A feature called System Hang Detection provides a SMIT-configurable mechanism to detect system hangs. | Not available in AIX Version 4.x.                                            |

## 5.8 Core dumps

When a system encounters a core dump, a core file is created in the current directory when various errors occur. Errors such as memory-address violations, illegal instructions, bus errors, and user-generated quit signals commonly cause a core dump. The core file that is created contains a memory image of the terminated process. A process with a saved user ID that differs from the real user ID does not produce a memory image.

### 5.8.1 Checking for core dump

When a core dump is created, an error will be reported, and the following entry can be seen in the error report:

```
errpt
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
...
C60BB505 0705101400 P S SYSPROC SOFTWARE PROGRAM ABNORMALLY TERMINATED
...
```

From the previous report, it can be seen that the error has an identifier of C60BB505. A detailed report of the error can be displayed as follows:

```
errpt -a -j C60BB505

LABEL: CORE_DUMP
IDENTIFIER: C60BB505

Date/Time: Wed Jul 5 10:14:59
Sequence Number: 8
Machine Id: 000BC6DD4C00
Node Id: client1
Class: S
Type: PERM
Resource Name: SYSPROC

Description
SOFTWARE PROGRAM ABNORMALLY TERMINATED

Probable Causes
SOFTWARE PROGRAM

User Causes
USER GENERATED SIGNAL

Recommended Actions
CORRECT THEN RETRY
```

Failure Causes  
SOFTWARE PROGRAM

Recommended Actions  
RERUN THE APPLICATION PROGRAM  
IF PROBLEM PERSISTS THEN DO THE FOLLOWING  
CONTACT APPROPRIATE SERVICE REPRESENTATIVE

Detail Data  
SIGNAL NUMBER  
4  
USER'S PROCESS ID:  
15394  
FILE SYSTEM SERIAL NUMBER  
5  
INODE NUMBER  
2  
PROGRAM NAME  
netscape\_aix4  
ADDITIONAL INFORMATION  
Unable to generate symptom string.  
Too many stack elements.

In the previous output, it can be seen that the program that created the core dump was netscape\_aix4. See the following section to help you determine where the core file is located.

## 5.8.2 Locating a core dump

When a system does a core dump, it writes a file named core. This file may be written anywhere in the system, including on networked file systems, and it will need to be found using the **find** command, as follows:

```
find / -name core -ls
737 10188 -rw-r--r-- 1 root system 10430807 Jul 5 10:14 /core
```

From the command example, the file is located in the root directory.

## 5.8.3 Determining the program that caused the core dump

There are two ways to determine which program caused the core dump: One is using the **strings** command, and the other is using the **lquerypv** command. Although this information should be in the error report, there may be occasion when the error report is not available or has been cleared out.

The **strings** command will give the full path name of the program, and is used as follows:

```
strings core | grep _=
_=/usr/netscape/communicator/us/netscape_aix4
```

The **lquerypv** command is run as follows:

```
lquerypv -h core 6b0 64
000006B0 7FFFFFFF FFFFFFFF 7FFFFFFF FFFFFFFF |.....|
000006C0 00000000 000007D0 7FFFFFFF FFFFFFFF |.....|
000006D0 00120000 137084E0 00000000 00000016 |....p.....|
000006E0 6E657473 63617065 5F616978 34000000 |netscape_aix4...|
000006F0 00000000 00000000 00000000 00000000 |.....|
00000700 00000000 00000000 00000000 0000085E |.....^|
00000710 00000000 00000F5A 00000000 00000776 |.....Z.....v|
```

Using the information provided, the file was dumped by the `netscape_aix4` program as displayed in the error report.

## 5.9 Command summary

The following section provides a list of the key commands and provides examples of their use.

### 5.9.1 The **crash** command

This section provides you with information on common problems using the **crash** command, and assists you in making a basic determination as to what caused the problem.

#### Uses of **crash**

The **crash** command can be used on a running system. Invoking **crash** with no parameters essentially allows you to view the memory and state of the currently running system by examining `/dev/mem`. The `alter` subcommand in **crash** allows you to modify the running kernel. This should only be used under the direction of IBM support, since incorrect use can cause the system to fail. The user must be in the system group to run the **crash** command on the live system.

The **crash** command can also be used on a system dump. It is the primary tool used to analyze a dump resulting from a system failure. Invoking the **crash** command with a parameter specifying a dump file allows you to examine a dump file for problem analysis.

Using **crash**, you can examine:

- ▶ Addresses and symbols
- ▶ Kernel stack traceback
- ▶ Kernel extensions
- ▶ The process table
- ▶ The thread table
- ▶ The file table
- ▶ The inode table
- ▶ System registers

In addition to the items listed, you can use **crash** to look at anything else contained in the kernel memory.

### **What the kernel is**

The kernel is the program that controls and protects system resources. It runs in privileged mode. It operates directly with the hardware. The major functions of the kernel are:

- ▶ Creation and deletion of processes/threads
- ▶ CPU scheduling
- ▶ Memory management
- ▶ Device management
- ▶ Synchronization and communication tools for processes

In contrast to a user program, which creates a core dump and halts. If the kernel has an error the machine will fail.

The **crash** command is used to debug these kernel problems.

### **Examining a system dump**

The **crash** command needs a kernel `/unix` file to match the dump file under analysis. For example:

```
itsostrv1:/dumptest> crash dumpfile unix
>
```

If no kernel file is specified, the default is `/unix`.

```
itsostrv1:/dumptest> crash dumpfile
Using /unix as the default namelist file.
>
```

The **crash** command uses the kernel file to interpret symbols and allows for symbolic translation and presentation. If the kernel file does not match the dump, you will get an error message when you start **crash**.

## Basic crash subcommands

Once you initiate the **crash** command, the prompt character is the greater than sign (>). For a list of the available subcommands, type the question mark (?) character. To exit, type q. You can run any shell command from within the **crash** command by preceding it with an exclamation mark (!).

The following is a list of common **crash** subcommands:

- ▶ stat  
Shows dump statistics.
- ▶ proc [-] [-r] [processTableEntry]  
Displays the process table (proc.h). Alias p and ps.
- ▶ user [ProcessTableEntry]  
Displays user structure of named process (user.h). Alias u.
- ▶ thread [-] [-r] [-p] [threadTableEntry]  
Displays the thread table (thread.h).
- ▶ mst [addr]  
Displays the mstsave portion of the utthread structure (utthread.h, mstsave.h).
- ▶ ds [addr]  
Finds the data symbol closest to the given address.
- ▶ knlist [symbol]  
Displays address of symbol name given. It is the opposite of ds.
- ▶ trace [-k | s][[-m][[-r][ThreadTableEntry]  
Displays kernel stack trace. Alias t.
- ▶ le  
Displays loader entries.
- ▶ nm [symbol]  
Displays symbol value and type as found in the /unix file.
- ▶ od [symbol name or addr] [count] [format]  
Dumps count number of data words starting at symbol name or address in the format specified by format.

- ▶ ? or help[]  
Lists all subcommands.  
Provides information about crash subcommands.
- ▶ cm [thread slot][seg\_no]  
Changes the map of the **crash** command internal pointers for any process thread segment not paged out. Resets the map of internal pointers if no parameters are used.
- ▶ fs [thread slotNumber]  
Dumps the kernel stack frames for the specified thread.
- ▶ dlock [tid] | -p [processor\_num]  
Displays deadlock information about all types of locks: Simple, complex, and lockl.
- ▶ errpt [count]  
Displays error log messages. The errpt subcommand always prints all messages that have not yet been read by the errdemon. Count specifies the number of messages to print.
- ▶ du  
Dump user area of process.
- ▶ ppd  
Display per processor data area; useful for multiprocessor systems. Shows all data that varies for each processor, such as Current Save Area (CSA).
- ▶ symptom  
If your system supports symptom, this is a useful subcommand to obtain a quick snapshot of dump information.

### ***The stat subcommand***

The stat subcommand provides plenty of useful information about a dump, such as the dump code, the panic string, time of the crash, version and release of the operating system, name of the machine that crashed, and how long the machine had been running since the last crash or power off of the system. For example:

```
> stat
 sysname: AIX
 nodename: kmdvs
 release: 3
 version: 4
 machine: 000939434C00
 time of crash: Mon May 3 17:49:46 KORST 1999
 age of system: 2 day, 4 hr., 28 min.
```

```

xmalloc debug: disabled
dump code: 700
csa: 0x384eb0
exception struct:
 0x00000000 0x00000000 0x00000000 0x00000000 0x00000000
panic: HACMP for AIX dms timeout - ha

```

The stat subcommand should always be the first command run when examining a system crash.

### ***The trace -m subcommand***

The trace -m subcommand gives you a kernel stack traceback.

This is typically the second command you will run when examining a system dump.

This subcommand provides information about what was happening in the kernel when the failure occurred. The trace -m subcommand provides a history of function calls and interrupt processing at the time of failure. If the failure occurred while interrupt processing was going on, this subcommand will be very useful in determining the cause. This subcommand traces the linked list of mstsave areas. The mstsave areas basically contain a history of what interrupt processing was going on in the system.

The machine state save area, or MST, contains a saved image of the machine's process context. The process context includes the general purpose and floating point registers, the special purpose registers, and other information necessary to restart a thread when it is dispatched. For example:

```

> trace -m
Skipping first MST

```

#### MST STACK TRACE:

```

0x002baeb0 (excpt=00000000:00000000:00000000:00000000:00000000) (intpri=3)
 IAR: .[atmle_dd:atmle_ready_ind]+d8 (01b05cb0): tweqi r5,0x0
 LR: .[atmle_dd:atmle_ready_ind]+34 (01b05c0c)
002ba940: .[atmle_dd:atmle_receive_ether_data]+1ec (01b0c35c)
002ba9a0: .[atm_demux:atm_dmx_receive]+204 (01adc0e8)
002baa00: .[atmdd:atm_deqhandler]+1254 (01ac7e6c)
002babc0: .[atmdd:atm_HandleCardRsp]+1a4 (01aba084)
002baca0: .[atmdd:atm_handler]+48 (01aba350)
002bad40: .[atmdd:atm_intr]+ac (01ac4a04)
002bad90: .i_poll_soft+9c (0001ef84)
002badf0: .i_softmod+c8 (0001e964)
002bae70: flih_603_patch+c0 (0000bb9c)

0x2ff3b400 (excpt=00000000:00000000:00000000:00000000:00000000) (intpri=11)
 IAR: .waitproc+c0 (0000edb0): lwz r3,0x6c(r28)

```

```
LR: .waitproc+d4 (0000edc4)
2ff3b388: .procentry+14 (00045414)
2ff3b3c8: .low+0 (00000000)
```

In this example, there are two levels of stack traceback. The first level shows the Instruction Address Register (IAR) pointing to a trap instruction, `tweqi r5, 0x0`, as shown.

The following registers are worth considering:

**IAR** Instruction Address Register. The address of the instruction that caused the crash.

**LR** Link Register that called the fatal function or where last call returns to.

This trap instruction is what you will see when you get a crash of type Program Interrupt, or Dump Status = 700. This was probably the result of assert or panic. It can be seen that the interrupt priority is 3 (`intpri=3`). In this case, it can be seen that interrupt processing was occurring when the crash happened, because the interrupt priority was less than 11 or 0xB, which is the base interrupt priority. This is the level at which a normal process runs.

The first entry on the stack traceback was the most recently running function, which was called by the function below it, which was called by the function below it, and so on. Therefore, in the case of the middle stack traceback in our example, it can be seen that `i_softmod` called `i_poll_soft`, which called some functions in `atmdd` and `atm_demux` modules, which in turn called `atmle_receive_ether_data`, which called `atmle_ready_ind`, and an assert was hit in `atmle_ready_ind`. Look at the code for this string to try to find out the cause of the assert action. You can deduce that the `atmle_dd` module did something wrong or the parameters passed in to the function were incorrect.

Make sure the failing module is at the latest version. Problems are frequently resolved in later versions of software. You can use the `le` subcommand in `crash` and the `ls1pp -w` command to find the fileset that contains the specific module.

### ***The le subcommand***

Use the `le` subcommand with the address listed in the IAR of the topmost MST area as the argument. The address is displayed in brackets after the name of the module. For example:

```
> le 01b05cb0
LoadList entry at 0x04db7780
 Module start:0x00000000_01b016e0 Module filesize:0x00000000_00030fbc
 Module *end:0x00000000_01b3269c
 *data:0x00000000_0125ef40 data length:0x00000000_0000375c
 Use-count:0x000c load_count:0x0001 *file:0x00000000
 flags:0x00000272 TEXT KERNELEX DATAINTEXT DATA DATAEXISTS
```

```

*exp:0x04e0e000 *lex:0x00000000 *deferred:0x00000000 *expsize:0x69626f64
Name: /usr/lib/drivers/atmle_dd
ndepend:0x0001 maxdepend:0x0001
*depend[00]:0x04db7580
le_next: 04db7380

```

One of the fields listed by the `le` subcommand is the name of the module. You can then use the `ls1pp -w` command to determine the fileset that contains the module. For example:

```

itsostrv1: /> ls1pp -w /usr/lib/drivers/atmle_dd
File Fileset Type

/usr/lib/drivers/atmle_dd bos.atm.atmle File

```

This command is available in AIX Version 4.2 or later.

Consider the following line:

```
002ba940: .[atmle_dd:atmle_receive_ether_data]+1ec (01b0c35c)
```

The address of the entry on the stack is in the first column. The last column contains the return address of the code (01b0c35c). This address corresponds to the function shown, `atmle_receive_ether_data`, which is contained in the module `atmle_dd`. The square brackets around the `[module:function]` pair indicate that this is a kernel extension. In addition, the instruction at this return address is at offset 0x1ec from the beginning of the function `atmle_receive_ether_data()`.

The last of the stack trace backs indicates that the user-level process (`intpri=b`) and the running process is `wait`. If the `crash` user subcommand is run, it will be seen that the running process is `wait`. However, `wait` did not cause the problem here; the problem was caused by a program running at interrupt level, and looking at the MST stack traceback is the only way to see the real problem.

When a Data Storage Interrupt (DSI) with dump code 300 occurs, the exception structure is filled in as follows:

```
0x2ff3b400 (excpt=DAR:DSISR:SRV:DAR2:DSIRR) (intpri=?)
```

The exception structure shows various machine registers and the interrupt level. The registers shown in the exception structure are defined as follows:

|              |                                        |
|--------------|----------------------------------------|
| <b>DAR</b>   | Data Address Register                  |
| <b>DSISR</b> | Data Storage Interrupt Status Register |
| <b>SRV</b>   | Segment Register Value                 |
| <b>DAR2</b>  | Secondary Data Address Register        |

## DSIRR

## Data Storage Interrupt Reason Register

The interrupt priority of the running context is shown in the (intpri=?) field at the end of the line. The intpri value ranges from 0xb (INTBASE) to 0x0 (INTMAX).

The exception structure is not used for code 700 dumps.

The le subcommand can indicate the kernel extension that an address belongs to. Take, for example, the address 0x0123cc5c. This is a kernel address, since it starts 0x01, which indicates it is in segment 0, the kernel segment. To find the kernel module that contains the code at this address, use the le subcommand. For example:

```
> le 0123cc5c
LoadList entry at 0x04db7780
 Module start:0x00000000_012316e0 Module filesize:0x00000000_00030fbc
 Module *end:0x00000000_0126269c
 *data:0x00000000_0125ef40 data length:0x00000000_0000375c
 Use-count:0x000c load_count:0x0001 *file:0x00000000
 flags:0x00000272 TEXT KERNELEX DATAINTEXT DATA DATAEXISTS
 *exp:0x04e0e000 *lex:0x00000000 *deferred:0x00000000 *expsize:0x69626f64
 Name: /usr/lib/drivers/pse/pse
 ndepend:0x0001 maxdepend:0x0001
 *depend[00]:0x04db7580
 le_next: 04db7380
```

In this case, it can be seen that the code at address 0x0123cc5c is in module /usr/lib/drivers/pse/pse. The le subcommand is only helpful for modules that are already loaded into the kernel.

### ***The proc subcommand***

The proc subcommand displays entries in the process table. The process table is made up of entries of type struct proc, one per active process. Entries in the process table are pinned so that they are always resident in physical memory. The process table contains information needed when the process has been swapped out in order to get it running again at some point in the future. For example:

```
> proc - 0
SLT ST PID PPID PGRP UID EUID TCNT NAME
 0 a 0 0 0 0 0 1 swapper
 FLAGS: swapped_in no_swap fixed_pri kproc

Links: *child:0xe3000170 *siblings:0x00000000 *uid1:0xe3001fa0
 *ganchor:0x00000000 *pgrp1:0x00000000 *tty1:0x00000000
Dispatch Fields: pevent:0x00000000 *synch:0xffffffff
 lock:0x00000000 lock_d:0x01390000
Thread Fields: *threadlist:0xe6000000 threadcount:1
```

```

 active:1 suspended:0 local:0 terminating:0
Scheduler Fields: fixed pri: 16 repage:0x00000000 scount:0 sched_pri:0
 *sched_next:0x00000000 *sched_back:0x00000000 cpticks:0
 msgcnt:0 majfltsec:0
Misc: adspace:0x0001e00f kstackseg:0x00000000 xstat:0x0000
 *p_ipc:0x00000000 *p_dblist:0x00000000 *p_dbnext:0x00000000
Signal Information:
 pending:hi 0x00000000,lo 0x00000000
 sigcatch:hi 0x00000000,lo 0x00000000 sigignore:hi 0xffffffff,lo 0xffff7fff
Statistics: size:0x00000000(pages) audit:0x00000000
 accounting page frames:0 page space blocks:0

 pctcpu:0 minflt:1802 majflt:7

```

The fields in the first few lines of the output are as follows:

- SLT** This is the process slot number, and simply indicates the process position in the process table. Use this number to tell the **crash** command which specific process block or u-block to display. Note that the slot numbers are in decimals.
- ST** This is a one-character field indicating the status of the process, and may be a=active, i=idle, t=stopped, or z=zombie.
- PID** This is the actual process ID by which the process is known to the system. The process slot number is used to generate the process ID.
- PPID** Parent process ID.
- PGRP** Process group ID.
- UID** User ID.
- EUID** Effective user ID.
- TCNT** Thread count.
- NAME** Program name.
- FLAGS** Status flags.

### ***The thread subcommand***

The thread table contains per-thread information that can be used by other threads in a process. There is one structure allocated per active thread. Entries that are in use are pinned to avoid page faults in kernel critical sections. For example:

```

> thread - 0
SLT ST TID PID CUID POLICY PRI CPU EVENT PROCNAME
 0 s 3 0 unbound FIFO 10 78 swapper
 t_flags: wakeon sig kthread

```

```

Links: *procp:0xe3000000 *uthreadp:0x2ff3b400 *userp:0x2ff3b6e0

```

```

*prevthread:0xe6000000 *nextthread:0xe6000000, *stackp:0x00000000
*wchan1(real):0x00000000 *wchan2(VMM):0x00000000 *swchan:0x00000000
wchan1sid:0x00000000 wchan1offset:0x00000000
pevent:0x00000000 wevent:0x00000001 *slist:0x00000000
Dispatch Fields: *prior:0xe6000000 *next:0xe6000000
polevel:0x0000000a ticks:0x0139 *synch:0xffffffff result:0x00000000
*eventlst:0x00000000 *wchan(hash):0x00000000 suspend:0x0001
thread waiting for: event(s)
Scheduler Fields: cpuid:0xffffffff scpuid:0xffffffff pri: 16 policy:FIFO
affinity:0x0003 cpu:0x0078 lpri: 0 wpri:127 time:0x00
sav_pri:0x10
Misc: lockcount:0x00000000 ulock:0x00000000 *graphics:0x00000000
dispct:0x000000e4 fpuct:0x00000001 boosted:0x0000
userdata:0x00000000
Signal Information: cursig:0x00 *scp:0x00000000
pending:hi 0x00000000,lo 0x00000000 sigmask:hi 0x00000000,lo 0x00000000

```

The fields in the output of the thread subcommand are as follows:

|                 |                                                                                                                   |
|-----------------|-------------------------------------------------------------------------------------------------------------------|
| <b>SLT</b>      | Slot number.                                                                                                      |
| <b>ST</b>       | Status. This may be i=idle, r=running, s=sleeping, w=swapped out, t=stopped, or z=zombie.                         |
| <b>TID</b>      | Thread ID.                                                                                                        |
| <b>PID</b>      | Process ID of the associated process. There may be multiple threads per process, but only one process per thread. |
| <b>CPUID</b>    | CPU ID of the CPU running the thread. On a uniprocessor system, this will always be 0.                            |
| <b>POLICY</b>   | This is the scheduling policy used for the thread and may have the values FIFO, RR, or other.                     |
| <b>PRI</b>      | Dispatch priority. This is not the <b>nice</b> value.                                                             |
| <b>CPU</b>      | CPU utilization. This value is used for scheduling.                                                               |
| <b>PROCNAME</b> | The name of the process for this thread.                                                                          |
| <b>EVENTS</b>   | This is the wait channel if not zero.                                                                             |
| <b>FLAGS</b>    | Status flags.                                                                                                     |

### ***The od subcommand***

To display and examine memory areas from the dump, use the **od** subcommand. The syntax of the subcommand is as follows:

```
od [symbol name] [count] [format]
```

Formats are ASCII, octal, decimal, hex, byte, character, instruction, long octal, and long decimal.

For example:

```
> od vmker 15
000bde48: 00002001 00006003 00000000 00008004
000bde58: 00200000 00000012 0000000d 00000200
000bde68: 00080000 00000017 00078c93 00066320
000bde78: 00000ab2 00020000 00002870

> od 0xbde48 15 a
000bde48: 00002001 00006003 00000000 00008004 |.. ^.....|
000bde58: 00200000 00000012 0000000d 00000200 |.|
000bde68: 00080000 00000017 00078c93 00066320 |.....c |
000bde78: 00000ab2 00020000 00002870 |.....(p|
```

### ***The errpt subcommand***

To examine the last few error log entries from the dump, use the errpt subcommand. For example:

```
> errpt
ERRORS NOT READ BY ERRDEMON (MOST RECENT LAST):
Sun Apr 6 01:01:11 1997 : DSI_PROC data storage interrupt : processor
Resource Name: SYSVMM
42000000 007fffff 80000000 ffffffff
>
```

### ***The symptom subcommand***

The symptom[-e] subcommand displays the symptom string for a dump. It is not valid on a running system. The -e option will create an error log entry containing the symptom string and is normally only used by the system and not manually. The symptom string can be used to identify duplicate problems.

## **VMM error log**

When the Dump Status code indicates a DSI or an ISI, look at the VMM error log. This is done using the od subcommand and looking at the vmerrlog structure. See Table 5-4 for valid offset codes. For example:

```
> od vmerrlog 9 a
000c95b0: 9d035e4d 53595356 4d4d2000 00000000 |.. ^MSYSVMM|
000c95c0: 00000000 0a000000 00000000 0000000b |.....|
000c95d0: 00000086 |....|
```

Table 5-4 The vmerrlog structure

| Offset | Meaning                                            |
|--------|----------------------------------------------------|
| 0x14   | The Data Storage Interrupt Status Register (DSISR) |
| 0x1C   | Faulting address                                   |
| 0x20   | VMM return code                                    |

In this example, the VMM return code 0x86 means protection exception. The various VMM return codes, symbolic names, and meanings are provided in the following:

- 0000000E** This return code indicates an EFAULT. It comes from errno.h (14) and is returned if you attempt to access an invalid address.
- FFFFFFFFFA** This return code indicates that you tried to access an invalid page that is not in memory. This is usually the result of a page fault. This will be returned if you try to access something that is paged out while interrupts are disabled.
- 00000005** This is a hardware problem. An I/O error occurred when you tried to page in or page out, or you tried to access a memory mapped file and could not do it. Check the error log for disk or SCSI errors.
- 00000086** This return code indicates a protection exception. This means that you tried to store to a location that is protected. This is usually caused by low kernel memory.
- 0000001C** This return code indicates no paging space. This means that the system has exhausted its paging space.

## Handling crash output

Some **crash** subcommands generate many more lines than can fit on one screen. Also, **crash** does not pause the output after each screen is full. You will want to have some way of seeing the scrolled data.

In the past, the **script** or **tee** commands were used for this. For example:

```
tee -a outf | crash /tmp/dump /unix | tee -a outf
```

There is now a new way to obtain a log file by using the set logfile subcommand. For example:

```
>set logfile crash.log
```

Once this has been entered, **crash** starts logging all input and output to the specified file. The set variable subcommand is available in AIX Version 4.1.5, 4.2.1, 4.3, and later.

In addition to the logfile support, command pipeline support was added to **crash**, allowing you to pipe long output to other commands, such as **more**, **pg**, and **grep**. For example:

```
> 1e 0123cc5c | grep Name
Name: /usr/lib/drivers/pse/pse
```

## 5.9.2 Types of crashes

Common problems requiring crash dump analysis include those discussed in the following sections.

### Kernel panic or trap

A kernel panic or trap is usually the cause of a system crash with the LED sequence 888-102-700-0cx.

In AIX, kernel panics manifest themselves as traps. The panic() routine in the kernel puts its message into a buffer, writes it to the debug TTY using the kernel debug program, and calls brkpoint(). If the kernel debugger is loaded, and an ASCII terminal is connected on a serial port, this will start the debugger; otherwise, it will cause a dump. If a panic or assert occurs, you must examine the source code to understand the condition that caused the panic or assert.

### Addressing exception of data storage interrupt

An addressing exception of data storage interrupt is accompanied by the LED sequence 888-102-300-0cx.

The 300 in the LED sequence indicates an addressing exception (a data storage interrupt or DSI). This is usually caused by a bad address being accessed, or page fault occurring when interrupts are disabled. When you get this type of crash, check the VMM return code.

### System hang

A dump can be forced when the system locks up (to determine the cause of the hang).

A system hang is a total system lockup. A dump forced by turning the key to the Service position and pressing the Reset button can be examined to see what locks are being held by whom. Refer to 5.2, “Starting a system dump” on page 68, for more information.

## 5.9.3 The snap command

The **snap** command gathers system configuration information and compresses the information into a TAR file. The file can then be downloaded to disk or tape, or transmitted to a remote system. The information gathered with the **snap** command may be required to identify and resolve system problems.

The **snap** command syntax is as follows:

```
snap [-a] [-A] [-b] [-c] [-D] [-f] [-g] [-G] [-i] [-k]
[-l] [-L] [-n] [-N] [-p] [-r] [-s] [-S] [-t]
[-o OutputDevice] [-d Dir] [-v Component]
```

Commonly used **snap** command flags are listed in Table 5-5.

Table 5-5 Commonly used flags of the snap command

| Flag          | Description                                                                                                                                                                                                                                                                                     |
|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -a            | Gathers all system configuration information. This option requires approximately 8 MB of temporary disk space.                                                                                                                                                                                  |
| -A            | Gathers asynchronous (TTY) information.                                                                                                                                                                                                                                                         |
| -b            | Gathers SSA information.                                                                                                                                                                                                                                                                        |
| -c            | Creates a compressed TAR image (snap.tar.Z file) of all files in the /tmp/ibmsupt directory tree or other named output directory.                                                                                                                                                               |
| -D            | Gathers dump and /unix information. The primary dump device is used. If <b>bosboot -k</b> was used to specify the running kernel to be other than /unix, the incorrect kernel will be gathered. Make sure that /unix is, or is linked to, the kernel in use when the dump was taken.            |
| -d <i>Dir</i> | Identifies the optional <b>snap</b> command output directory (/tmp/ibmsupt is the default).                                                                                                                                                                                                     |
| -f            | Gathers file system information.                                                                                                                                                                                                                                                                |
| -g            | Gathers the output of the <b>ls1pp -hBc</b> command, which is required to recreate exact operating system environments. Writes output to the /tmp/ibmsupt/general/ls1pp.hBc file. Also collects general system information and writes the output to the /tmp/ibmsupt/general/general.snap file. |
| -G            | Includes predefined Object Data Manager (ODM) files in general information collected with the -g flag.                                                                                                                                                                                          |
| -i            | Gathers installation debug vital product data (VPD) information.                                                                                                                                                                                                                                |
| -k            | Gathers kernel information.                                                                                                                                                                                                                                                                     |
| -l            | Gathers programming language information.                                                                                                                                                                                                                                                       |
| -L            | Gathers LVM information.                                                                                                                                                                                                                                                                        |
| -n            | Gathers Network File System (NFS) information.                                                                                                                                                                                                                                                  |
| -N            | Suppresses the check for free space.                                                                                                                                                                                                                                                            |

| Flag                   | Description                                                                                                                          |
|------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
| -o <i>OutputDevice</i> | Copies the compressed image onto diskette or tape.                                                                                   |
| -p                     | Gathers printer information.                                                                                                         |
| -r                     | Removes <b>snap</b> command output from the /tmp/ibmsupt directory.                                                                  |
| -s                     | Gathers Systems Network Architecture (SNA) information.                                                                              |
| -S                     | Includes security files in general information collected with the -g flag.                                                           |
| -t                     | Gathers Transmission Control Protocol/Internet Protocol (TCP/IP) information.                                                        |
| -v <i>Component</i>    | Displays the output of the commands executed by the <b>snap</b> command. Use this flag to view the specified name or group of files. |

## 5.9.4 The strings command

The **strings** command looks for printable strings in an object or binary file. A string is any sequence of four or more printable characters that end with a new-line or a null character. The **strings** command is useful for identifying random object files.

The **strings** command syntax is as follows:

```
strings [-a] [-] [-o] [-t Format] [-n Number] [-Number]
[File]
```

Commonly used **strings** command flags are listed in Table 5-6.

Table 5-6 Commonly used flags of the strings command

| Flag             | Description                                                                                                                                                            |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -a or -          | Searches the entire file, not just the data section, for printable strings.                                                                                            |
| -n <i>Number</i> | Specifies a minimum string length other than the default of four characters. The maximum value of a string length is 4096. This flag is identical to the -Number flag. |
| -o               | Lists each string preceded by its octal offset in the file. This flag is identical to the -t o flag.                                                                   |

| Flag             | Description                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>-t Format</i> | Lists each string preceded by its offset from the start of the file. The format is dependent on the character used as the format variable:<br><br>d writes the offset in decimal.<br>o writes the offset in octal.<br>x writes the offset in hexadecimal.<br><br>When the <i>-o</i> and the <i>-t Format</i> flags are defined more than once on a command line, the last flag specified controls the behavior of the <b>strings</b> command. |
| <i>-Number</i>   | Specifies a minimum string length other than the default of four characters. The maximum value of a string length is 4096. This flag is identical to the <i>-n Number</i> flag.                                                                                                                                                                                                                                                               |
| <i>File</i>      | Binary or object file to be searched.                                                                                                                                                                                                                                                                                                                                                                                                         |

## 5.9.5 The sysdumpdev command

The **sysdumpdev** command changes the primary or secondary dump device designation in a system that is running. The primary and secondary dump devices are designated in a system configuration object. The new device designations are in effect until the **sysdumpdev** command is run again, or the system is restarted.

If no flags are used with the **sysdumpdev** command, the dump devices defined in the SWservAt ODM object class are used. The default primary dump device is /dev/hd6. The default secondary dump device is /dev/sysdumpnull.

**Note:** Do not use a mirrored or copied logical volume as the active primary dump device. System dump error messages will not be displayed, and any subsequent dumps to a mirrored logical volume will fail.

Do not use a diskette drive as your dump device.

If you use a paging device, only use hd6, the primary paging device. AIX Version 4.2.1 or later supports using any paging device in the root volume group (rootvg) as the secondary dump device.

The **sysdumpdev** command syntax is as follows:

```
sysdumpdev [-c | -C] -P { -p Device | -s Device } [-q]
sysdumpdev [-c | -C] [-p Device | -s Device] [-q]
sysdumpdev [-c | -C] [-d Directory | -D Directory | -e | [-k | -K] | -l | -L
| -p Device | -q | -r Host: Path | -s Device | -z]
```

Commonly used **sysdumpdev** command flags are listed in Table 5-7.

Table 5-7 Commonly used flags of the **sysdumpdev** command

| Flag                | Description                                                                                                                                                                                                                                                                                                                                                                                                              |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -c                  | Specifies that dumps will not be compressed. The -c flag applies to only AIX Version 4.3.2 and later versions.                                                                                                                                                                                                                                                                                                           |
| -C                  | Specifies that all future dumps will be compressed before they are written to the dump device. The -C flag applies to only AIX Version 4.3.2 and later versions.                                                                                                                                                                                                                                                         |
| -d <i>Directory</i> | Specifies the directory the dump is copied to at system boot. If the copy fails at boot time, the -d flag ignores the system dump.                                                                                                                                                                                                                                                                                       |
| -D <i>Directory</i> | Specifies the directory the dump is copied to at system boot. If the copy fails at boot time, using the -D flag allows you to copy the dump to an external media. When using the -d <i>Directory</i> or -D <i>Directory</i> flags, the following error conditions are detected:<br><br>Directory does not exist.<br>Directory is not in the local journaled file system.<br>Directory is not in the rootvg volume group. |
| -e                  | Estimates the size of the dump (in bytes) for the current running system.                                                                                                                                                                                                                                                                                                                                                |
| -k                  | Requires the key mode switch to be in the service position before a dump can be forced with the Reset button or the dump key sequences. This is the default setting.                                                                                                                                                                                                                                                     |
| -K                  | The Reset button or the dump key sequences will force a dump with the key in the normal position. On a machine without a key mode switch, a dump cannot be forced with the Reset button or the key switch without this value set.                                                                                                                                                                                        |
| -l                  | Lists the current value of the primary and secondary dump devices, copy directory, and forcecopy attribute.                                                                                                                                                                                                                                                                                                              |
| -L                  | Displays statistical information about the most recent system dump. This includes date and time of last dump, number of bytes written, and completion status.<br><br>The status value will correspond to the dump LED codes as follows:<br>0 = 0c0 - Dump completed successfully<br>-1 = 0c8 - No primary dump device<br>-2 = 0c4 - Partial dump<br>-3 = 0c5 - Dump failed to start                                      |
| -P                  | Makes permanent the dump device specified by the -p or -s flags. The -P flag can only be used with the -p or -s flags.                                                                                                                                                                                                                                                                                                   |

| Flag                | Description                                                                                                                                                                                                                                                                                                                                                            |
|---------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -p <i>Device</i>    | Temporarily changes the primary dump device to the specified device. The device can be a logical volume or a tape device. For a network dump, the device can be a host name and a path name.                                                                                                                                                                           |
| -q                  | Suppresses all messages to standard output. If this flag is used with the -l, -r, -z, or -L flag, the -q flag will be ignored.                                                                                                                                                                                                                                         |
| -r <i>Host:Path</i> | Frees space used by the remote dump file on server host. The location of the dump file is specified by the path.                                                                                                                                                                                                                                                       |
| -s <i>Device</i>    | Temporarily changes the secondary dump device to the specified device. The device can be a logical volume or a tape device. For a network dump, the device can be a host name and a path name.                                                                                                                                                                         |
| -z                  | Determines if a new system dump is present. If one is present, a string containing the size of the dump in bytes and the name of the dump device will be written to standard output. If a new system dump does not exist, nothing is returned. After the <b>sysdumpdev -z</b> command is run on an existing system dump, the dump will no longer be considered recent. |

## 5.9.6 The sysdumpstart command

The **sysdumpstart** command provides a command line interface to start a kernel dump to the primary or secondary dump device. When the dump completes, the system halts. Use the **crash** command to examine a kernel dump. Use the **sysdumpdev** command to reassign the dump device.

The **sysdumpstart** command syntax is as follows:

```
sysdumpstart { -p | -s [-f] }
```

During a kernel dump, the following values can be displayed on the three-digit terminal display as follows:

- 0c0** Indicates that the dump completed successfully.
- 0c1** Indicates that an I/O error occurred during the dump. This value only applies to AIX Version 4.2.1 or later.
- 0c2** Indicates that the dump is in progress.
- 0c4** Indicates that the dump device is too small.
- 0c5** Indicates a dump internal error.
- 0c6** Prompts you to make the secondary dump device ready. This value does not apply for AIX Version 4.2.1 or later.

- 0c7** Indicates that the dump process is waiting for a response from the remote host.
- 0c8** Indicates that the dump was disabled. In this case, no dump device was designated in the system configuration object for dump devices. The **sysdumpstart** command halts, and the system continues running.
- 0c9** Indicates that a dump is in progress.
- 0cc** Indicates that the system switched to the secondary dump device after attempting a dump to the primary device. This value only applies to AIX Version 4.2.1 and later.

The **sysdumpstart** command flags are listed in Table 5-8.

*Table 5-8 Commonly used flags of the sysdumpstart command*

| Flag | Description                                                                                                            |
|------|------------------------------------------------------------------------------------------------------------------------|
| -f   | Suppresses the prompt to make the secondary dump device ready. This flag does not apply to AIX Version 4.2.1 or later. |
| -p   | Initiates a system dump and writes the results to the primary dump device.                                             |
| -s   | Initiates a system dump and writes the results to the secondary dump device.                                           |

## 5.10 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

1. Using the following error log entries as shown, which of the following conclusions best explains why the system crashed?
  - A. The system crash was caused by a signal 11.
  - B. The **calc** command caused the system to crash.
  - C. The system crashed due to an invalid inode number.
  - D. The system took a dump to paging space after the **calc** command core dumped.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>LABEL: DUMP_STATS<br/>IDENTIFIER: 5D66BBC4</p> <p>Date/Time: Mon Jun 19 22:17:48<br/>Sequence Number: 180<br/>Machine Id: 000005937100<br/>Node Id: satori<br/>Class: S<br/>Type: UNKN<br/>Resource Name: SYSDUMP</p> <p>Description<br/>SYSTEM DUMP</p> <p>Probable Causes<br/>UNEXPECTED SYSTEM HALT</p> <p>User Causes<br/>SYSTEM DUMP REQUESTED BY USER</p> <p>Recommended Actions<br/>PERFORM PROBLEM DETERMINATION PROCEDURES</p> <p>Failure Causes<br/>UNEXPECTED SYSTEM HALT</p> <p>Recommended Actions<br/>PERFORM PROBLEM DETERMINATION PROCEDURES</p> <p>Detail Data<br/>DUMP DEVICE<br/>/dev/hd6<br/>MAJOR DEVICE NUMBER<br/>10<br/>MINOR DEVICE NUMBER<br/>1<br/>DUMP SIZE<br/>11149824<br/>TIME<br/>Fri Feb 24 09:09:45 1995<br/>DUMP TYPE (1 = PRIMARY, 2 = SECONDARY)<br/>1<br/>DUMP STATUS<br/>0</p> | <p>LABEL: CORE_DUMP<br/>IDENTIFIER: DE0A8DC4</p> <p>Date/Time: Mon Jun 19 21:42:05<br/>Sequence Number: 179<br/>Machine Id: 000005937100<br/>Node Id: satori<br/>Class: S<br/>Type: PERM<br/>Resource Name: SYSPROC</p> <p>Description<br/>SOFTWARE PROGRAM ABNORMALLY TERMINATED</p> <p>Probable Causes<br/>SOFTWARE PROGRAM</p> <p>User Causes<br/>USER GENERATED SIGNAL</p> <p>Recommended Actions<br/>CORRECT THEN RETRY</p> <p>Failure Causes<br/>SOFTWARE PROGRAM</p> <p>Recommended Actions<br/>RERUN THE APPLICATION PROGRAM<br/>IF PROBLEM PERSISTS THEN DO THE FOLLOWING<br/>CONTACT APPROPRIATE SERVICE REPRESENTATIVE</p> <p>Detail Data<br/>SIGNAL NUMBER<br/>11<br/>USER'S PROCESS ID:<br/>9238<br/>FILE SYSTEM SERIAL NUMBER<br/>9<br/>INODE NUMBER<br/>17<br/>PROGRAM NAME<br/>calc</p> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Figure 5-3 Error log case study

2. Which of the following crash subcommands should be used to verify the date and time of a system dump?
  - A. le
  - B. t -s
  - C. stat
  - D. errpt

3. After a legacy microchannel system has gone down with flashing 888s, which of the following procedures is the best way to diagnose the problem?
  - A. Turn the power off and back on.
  - B. Reboot the system in maintenance mode.
  - C. Turn the key to service and press the Reset button to make a dump.
  - D. Verify that the key is in normal mode and press the Reset button to obtain error codes.
4. All of the following are contained in a system dump except:
  - A. The entire contents of memory
  - B. The history of kernel function calls
  - C. The process/thread that was active
  - D. The final history of error log events leading up to the system crash
5. Which of the following devices is a valid primary dump device?
  - A. /dev/hd5
  - B. /dev/null
  - C. /var/adm/ras
  - D. /dev/paging00
6. A system dump must be performed in order to have a permanent snapshot of the current state of the system. Which of the following commands should be run?
  - A. **snap**
  - B. **crash**
  - C. **sysdumpdev**
  - D. **sysdumpstart**
7. After a user initiates a dump request on an AIX system, the entry shown in Figure 5-4 on page 104 is recorded in the system's error log.

| F      | S | UID | PID  | PPID  | C | PRI | NI | ADDR | SZ  | WCHAN   | TTY    | TIME   | CMD      |
|--------|---|-----|------|-------|---|-----|----|------|-----|---------|--------|--------|----------|
| 202803 | S | 0   | 1    | 0     | 0 | 60  | 20 | 1004 | 528 |         | -      | 131:33 | init     |
| 260801 | S | 0   | 1406 | 1     | 0 | 60  | 20 | 4550 | 208 |         | -      | 0:00   | srcmstr  |
| 240801 | S | 0   | 1694 | 1     | 0 | 60  | 20 | 37cd | 176 | 5a6a024 | -      | 0:01   | cron     |
| 260801 | S | 0   | 2448 | 1     | 0 | 60  | 20 | 5d57 | 144 |         | -      | 3:18   | portmap  |
| 240801 | S | 0   | 2836 | 1     | 0 | 60  | 20 | 34cd | 72  | 5e34398 | -      | 42:42  | syncd    |
| 42801  | S | 0   | 3606 | 1     | 0 | 60  | 20 | 50d4 | 284 | cc98    | -      | 0:01   | errdemon |
| 260801 | S | 0   | 5255 | 1     | 0 | 60  | 20 | 5535 | 148 |         | -      | 1:00   | syslogd  |
| 240801 | S | 0   | 5541 | 1     | 0 | 59  | 20 | 7d9f | 60  | 3f2f8   | -      | 0:00   | llbd     |
| 240801 | S | 200 | 5750 | 79989 | 0 | 60  | 20 | a83  | 180 |         | pts/10 | 0:00   | -ksh     |
| 260801 | S | 0   | 6040 | 1     | 0 | 60  | 20 | 1565 | 468 |         | -      | 7:53   | snmpd    |
| 60801  | S | 0   | 6299 | 1     | 0 | 60  | 20 | 2d6b | 224 |         | -      | 0:00   | x_st_mgr |
| 240801 | S | 0   | 6502 | 1406  | 0 | 60  | 20 | 14e7 | 184 | 12d0440 | -      | 0:05   | qdaemon  |
| 40001  | S | 0   | 6659 | 1     | 0 | 23  | -- | 3aae | 312 | 1fca54  | hft/0  | 0:32   | userprog |

Figure 5-4 Case study

Which of the following conclusions can be drawn?

- A. The dump request failed.
  - B. The dump request was initiated due to an invalid sequence number.
  - C. The dump request was partially successful.
  - D. The dump request was initiated due to a problem with the device with the major/minor numbers 10,1.
8. What does the LED c20 indicate?
- A. A system dump failure.
  - B. A partial system dump.
  - C. A complete system dump.
  - D. The low level debugger is invoked.
9. If a system crashes and the LED c20 is displayed, what does this indicate?
- A. A dump has been forced on the system.
  - B. The system has automatically crashed.
  - C. The kernel debugger has been invoked.
  - D. The system took a partial dump and failed.

10. Which of the following codes best indicates that a crash was caused by a software problem?
- A. 888-102-000-0c8
  - B. 888-102-207-0c0
  - C. 888-102-300-0c0
  - D. 888-102-500-0c4
11. Which of the following codes best indicates that the crash was caused by a hardware problem?
- A. 888-102-207-0c0
  - B. 888-102-300-0c4
  - C. 888-102-300-0c8
  - D. 888-102-700-0c0
12. Which of the following dump codes best describes the LED 0c2 condition?
- A. Dump completed successfully
  - B. Dump is incomplete
  - C. No valid dump device
  - D. User requested dump started

### 5.10.1 Answers

The following are the preferred answers to the questions provided in this section.

1. D
2. C
3. D
4. A
5. D
6. D
7. A
8. D
9. C
10. C
11. A
12. D

### 5.11 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

1. Describe the different ways to start a system dump.
2. On a core dump, name the two ways that can be used to find the program that caused the core dump.
3. Briefly describe how the **crash** command can be used to analyze system dumps.



## Error reports

AIX records system errors and other information, such as system shutdowns and other system functions, in the error log. The contents of the error log can be viewed using the error report. This chapter will cover the use of the error report and how it can be used to obtain information about problems, and how the report can be maintained.

## 6.1 The error daemon

The error logging daemon is started with the **errdemon** command and writes entries to the error log.

The error logging daemon reads error records from the `/dev/error` file and creates error log entries in the system error log. Besides writing an entry to the system error log, the error logging daemon performs error notification as specified in the error notification database `/etc/objrepos/errnotify`. The default system error log is maintained in the `/var/adm/ras/errlog` file. The last error entry is placed in nonvolatile random access memory (NVRAM). During system startup, this last error entry is read from NVRAM and added to the error log or dump when the error logging daemon is started.

The error logging daemon does not create an error log entry for the logged error if the error record template specifies `Log=FALSE`.

If you use the error logging daemon without flags, the system restarts the error logging daemon using the values stored in the error log configuration database for the error log file name, file size, and internal buffer size.

Use the **errclear** command to remove entries from the system error log.

**Note:** The error logging daemon is normally started during system initialization. Stopping the error logging daemon can cause error data temporarily stored in internal buffers to be overwritten before it can be recorded in the error log file.

## 6.2 The errdemon command

The **errdemon** command syntax is as follows:

```
errdemon [[-B BufferSize] [-i File] [-s LogSize] | -l]
```

Commonly used **errdemon** flags are shown in Table 6-1.

Table 6-1 Commonly used flags of the errdemon command

| Flag                 | Description                                                                                                                                                       |
|----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>-i File</code> | Uses the error log file specified by the file variable. The specified file name is saved in the error log configuration database and is immediately put into use. |
| <code>-l</code>      | Displays the values for the error log file name, file size, and buffer size from the error log configuration database.                                            |

| Flag                 | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -s <i>LogSize</i>    | <p>Uses the size specified by the <code>LogSize</code> variable for the maximum size of the error log file. The specified log file size limit is saved in the error log configuration database, and it is immediately put into use. If the log file size limit is smaller than the size of the log file currently in use, the error logging daemon renames the current log file by appending <code>.old</code> to the file name. The error logging daemon creates a new log file with the specified size limit. Generate a report from the old log file using the <code>-i</code> flag of the <code>errpt</code> command.</p> <p>If this parameter is not specified, the error logging daemon uses the log file size from the error log configuration database.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| -B <i>BufferSize</i> | <p>Uses the number of bytes specified by the <code>BufferSize</code> parameter for the error log device driver's in-memory buffer. The specified buffer size is saved in the error log configuration database. If the <code>BufferSize</code> parameter is larger than the buffer size currently in use, the in-memory buffer is immediately increased. If the <code>BufferSize</code> parameter is smaller than the buffer size currently in use, the new size is put into effect the next time the error logging daemon is started after the system is rebooted. The buffer cannot be made smaller than the hard-coded default of 8 KB.</p> <p>If this parameter is not specified, the error logging daemon uses the buffer size from the error log configuration database.</p> <p>The size you specify is rounded up to the next integral multiple of the memory page size (4 KB). The memory used for the error log device driver's in-memory buffer is not available for use by other processes (the buffer is pinned). Be careful not to impact your system's performance by making the buffer excessively large. On the other hand, if you make the buffer too small, the buffer can become full if error entries arrive faster than they can be read from the buffer and put into the log file. When the buffer is full, new entries are discarded until space becomes available in the buffer. When this situation occurs, the error logging daemon creates an error log entry to inform you of the problem. You can correct the problem by enlarging the buffer.</p> |

An example of the `errdemon` command is provided in the following example.

To check the attributes of the error log file, use the `errdemon` command as follows:

```
/usr/lib/errdemon -l
Error Log Attributes

Log File /var/adm/ras/errlog
Log Size 23899 bytes
Memory Buffer Size 8192 bytes
```

To change the current log file, the **errdemon** command is used as follows:

```
/usr/lib/errdemon -i /var/adm/ras/myerrlog
```

To change the error log file size, the **errdemon** command is used as follows:

```
/usr/lib/errdemon -s 47798
```

To change the error log buffer size, the **errdemon** command is used as follows:

```
/usr/lib/errdemon -B 16384
0315-175 The error log memory buffer size you supplied will be rounded up
to a multiple of 4096 bytes.
```

The new status can be checked using the **errdemon** command as follows:

```
/usr/lib/errdemon -l
Error Log Attributes

Log File /var/adm/ras/myerrlog
Log Size 47798 bytes
Memory Buffer Size 16384 bytes
```

The **errdemon** command without flags will start the error daemon if it is not running as follows:

```
/usr/lib/errdemon
```

If the error daemon is running, an error will be reported as follows:

```
/usr/lib/errdemon
0315-100 The error log device driver, /dev/error, is already open.
The error demon may already be active.
```

## 6.3 The **errpt** command

The **errpt** command generates an error report from entries in an error log. It includes flags for selecting errors that match specific criteria. By using the default condition, you can display error log entries in the reverse order in which they occurred and were recorded. By using the **-c** (concurrent) flag, you can display errors as they occur. If the **-i** flag is not used with the **errpt** command, the error log file processed by **errpt** is the one specified in the error log configuration database (by default `/var/adm/ras/errlog`).

The default summary report contains one line of data for each error. You can use flags to generate reports with different formats.

**Note:** The **errpt** command does not perform error log analysis; for analysis, use the **diag** command.

To process a report from the error log, use the following syntax:

```
errpt [-a] [-c] [-d ErrorClassList] [-e EndDate] [-g] [-i File]
[-j ErrorID [,ErrorID]] | [-k ErrorID [,ErrorID]] [-J ErrorLabel
[,ErrorLabel]] | [-K ErrorLabel [,ErrorLabel]] [-l SequenceNumber]
[-m Machine] [-n Node] [-s StartDate] [-F FlagList]
[-N ResourceNameList] [-R ResourceTypeList] [-S ResourceClassList]
[-T ErrorTypeList] [-y File] [-z File]
```

Figure 6-1 shows how the **errpt** command processes a report from the error log.

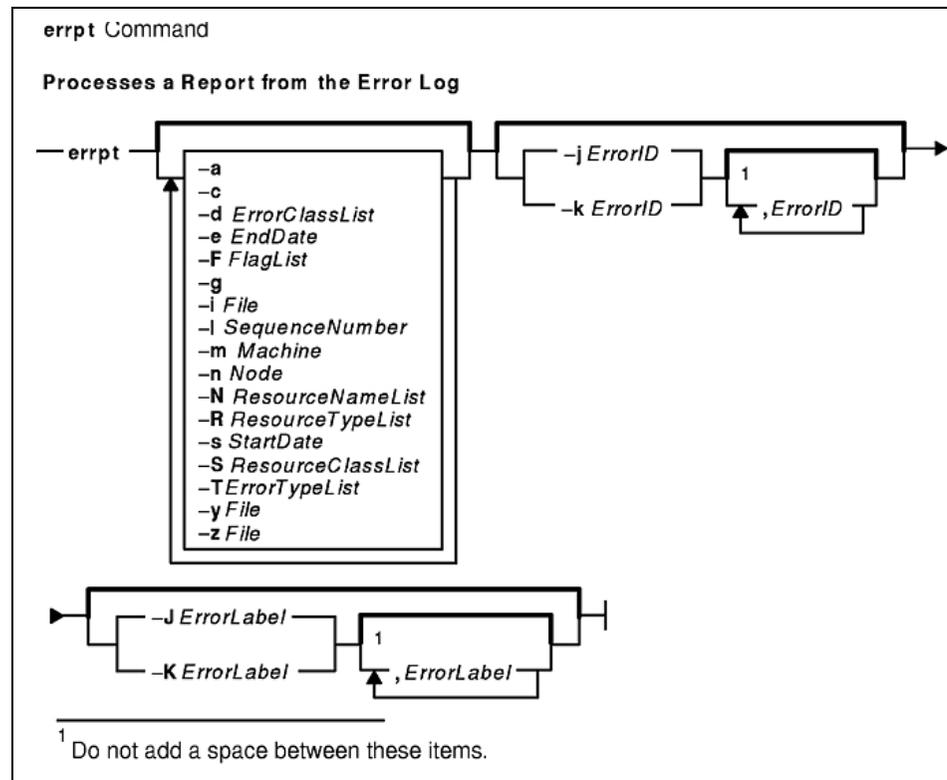


Figure 6-1 The **errpt** command error log report process

To process a report from the error record template repository, use the following syntax:

```
errpt [-a] [-t] [-d ErrorClassList] [-j ErrorID [,ErrorID]] |
[-k ErrorID [,ErrorID]] [-J ErrorLabel [,ErrorLabel]] |
[-K ErrorLabel [,ErrorLabel]] [-F FlagList] [-T ErrorTypeList]
[-y File] [-z File]
```

Figure 6-2 shows how the **errpt** command processes a report from the error record template.

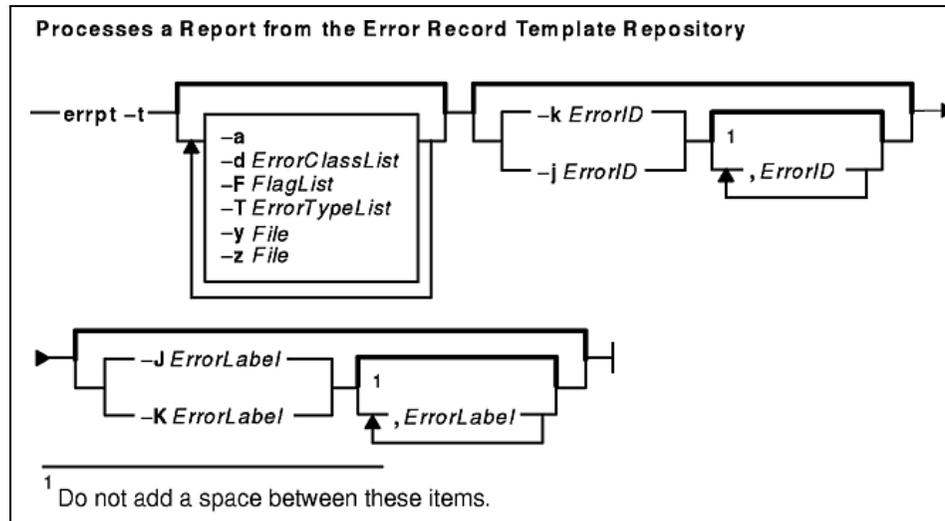


Figure 6-2 The **errpt** command error record template repository process

Table 6-2 is a listing of the commonly used **errpt** command flags.

Table 6-2 Commonly used flags of the **errpt** command

| Flag | Description                                                                                                                                                                                   |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -a   | Displays information about errors in the error log file in detailed format. If used in conjunction with the - t flag, all the information from the template file is displayed.                |
| -c   | Formats and displays each of the error entries concurrently, that is, at the time they are logged. The existing entries in the log file are displayed in the order in which they were logged. |

| <b>Flag</b>              | <b>Description</b>                                                                                                                                                                                                                                                                                                                                      |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>-d ErrorClassList</i> | Limits the error report to certain types of error records specified by the valid ErrorClassList variables: H (hardware), S (software), O ( <b>err1ogger</b> command messages), and U (undetermined). The ErrorClassList variable can be separated by commas (,), or enclosed in double quotation marks (") and separated by commas or space characters. |
| <i>-e EndDate</i>        | Specifies all records posted before the EndDate variable, where the EndDate variable has the form mmddhhmmyy (month, day, hour, minute, and year).                                                                                                                                                                                                      |

| Flag           | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -g             | <p>Displays the ASCII representation of unformatted error log entries. The output of this flag is in the following format:</p> <p><i>el_sequence</i> Error log stamp number<br/> <i>el_label</i> Error label<br/> <i>el_timestamp</i> Error log entry time stamp<br/> <i>el_crcid</i> Unique cyclic-redundancy-check (CRC) error identifier<br/> <i>el_machineid</i> Machine ID variable<br/> <i>el_nodeid</i> Node ID variable<br/> <i>el_class</i> Error class<br/> <i>el_type</i> Error type<br/> <i>el_resource</i> Resource name<br/> <i>el_rclass</i> Resource class<br/> <i>el_rtype</i> Resource type<br/> <i>el_vpd_ibm</i> IBM vital product data (VPD)<br/> <i>el_vpd_user</i> User VPD<br/> <i>el_in</i> Location code of a device<br/> <i>el_connwhere</i> Hardware-connection ID (location on a specific device, such as slot number)<br/> <i>et_label</i> Error label<br/> <i>et_class</i> Error class<br/> <i>et_type</i> Error type<br/> <i>et_desc</i> Error description<br/> <i>et_probcauses</i> Probable causes<br/> <i>et_usercauses</i> User causes<br/> <i>et_useraction</i> User actions<br/> <i>et_instcauses</i> Installation causes<br/> <i>et_instaction</i> Installation actions<br/> <i>et_failcauses</i> Failure causes<br/> <i>et_failaction</i> Failure actions<br/> <i>et_detail_length</i> Detail-data field length<br/> <i>et_detail_descid</i> Detail-data identifiers<br/> <i>et_detail_encode</i> Description of detail-data input format<br/> <i>et_logflg</i> Log flag<br/> <i>et_alertflg</i> Alertable error flag<br/> <i>et_reportflg</i> Error report flag<br/> <i>el_detail_length</i> Detail-data input length<br/> <i>el_detail_data</i> Detail-data input</p> |
| -i <i>File</i> | <p>Uses the error log file specified by the file variable. If this flag is not specified, the value from the error log configuration database is used.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |

| Flag                           | Description                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -j<br><i>ErrorID[,ErrorID]</i> | Includes only the error log entries specified by the ErrorID (error identifier) variable. The ErrorID variables can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters. When combined with the -t flag, entries are processed from the error-template repository. (Otherwise, entries are processed from the error-log repository.) |
| -k<br><i>ErrorID[,ErrorID]</i> | Excludes the error log entries specified by the ErrorID variable. The ErrorID variables can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters. When combined with the -t flag, entries are processed from the error-template repository. (Otherwise, entries are processed from the error-log repository.)                         |
| -l<br><i>SequenceNumber</i>    | Selects a unique error log entry specified by the SequenceNumber variable. This flag is used by methods in the error-notification object class. The SequenceNumber variable can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                                                                                 |
| -m <i>Machine</i>              | Includes error log entries for the specified machine variable. The <b>uname -m</b> command returns the machine variable value.                                                                                                                                                                                                                                                             |
| -n <i>Node</i>                 | Includes error log entries for the specified node variable. The <b>uname -n</b> command returns the node variable value.                                                                                                                                                                                                                                                                   |
| -s <i>StartDate</i>            | Specifies all records posted after the StartDate variable, where the StartDate variable has the form mmddhhmmyy (month, day, hour, minute, and year).                                                                                                                                                                                                                                      |
| -t                             | Processes the error record template repository instead of the error log. The -t flag can be used to view error record templates in report form.                                                                                                                                                                                                                                            |
| -y <i>File</i>                 | Uses the error record template file specified by the file variable. When combined with the -t flag, entries are processed from the specified error template repository. (Otherwise, entries are processed from the error log repository, using the specified error template repository.)                                                                                                   |
| -z <i>File</i>                 | Uses the error logging message catalog specified by the file variable. When combined with the -t flag, entries are processed from the error template repository. (Otherwise, entries are processed from the error log repository.)                                                                                                                                                         |

| Flag                               | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -F <i>FlagList</i>                 | <p>Selects error record templates according to the value of the Alert, Log, or Report field of the template. The FlagList variable can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters. The -F flag is used with the -t flag only.</p> <p>Valid values of the FlagList variable include:</p> <p><i>alert=0</i>, which selects error record templates with the Alert field set to False.</p> <p><i>alert=1</i>, which selects error record templates with the Alert field set to True.</p> <p><i>log=0</i>, which selects error-record templates with the Log field set to False.</p> <p><i>log=1</i>, which selects error record templates with the Log field set to True.</p> <p><i>report=0</i>, which selects error record templates with the Report field set to False.</p> <p><i>report=1</i>, which selects error record templates with the Report field set to True.</p> |
| -J <i>ErrorLabel</i>               | <p>Includes the error log entries specified by the ErrorLabel variable. The ErrorLabel variable values can be separated by commas, or enclosed in double-quotation marks and separated by commas or blanks. When combined with the -t flag, entries are processed from the error template repository. (Otherwise, entries are processed from the error log repository.)</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| -K <i>ErrorLabel</i>               | <p>Excludes the error log entries specified by the ErrorLabel variable. The ErrorLabel variable values can be separated by commas, or enclosed in double-quotation marks and separated by commas or blanks. When combined with the -t flag, entries are processed from the error template repository. (Otherwise, entries are processed from the error log repository.)</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| -N<br><i>ResourceName<br/>List</i> | <p>Generates a report of resource names specified by the ResourceNameList variable. For hardware errors, the ResourceNameList variable is a device name. For software errors, it is the name of the failing executable. The ResourceNameList variable can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| -R<br><i>ResourceTypeLi<br/>st</i> | <p>Generates a report of resource types specified by the ResourceTypeList variable. For hardware errors, the ResourceTypeList variable is a device type. For software errors, it is the LPP value. The ResourceTypeList variable can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |

| Flag                           | Description                                                                                                                                                                                                                                                                                              |
|--------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -S<br><i>ResourceClassList</i> | Generates a report of resource classes specified by the ResourceClassList variable. For hardware errors, the ResourceClassList variable is a device class. The ResourceClassList variable can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters. |
| -T <i>ErrorTypeList</i>        | Limits the error report to error types specified by the valid ErrorTypeList variables: INFO, PEND, PERF, PERM, TEMP, and UNKN. The ErrorTypeList variable can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                 |

To display a complete summary report, enter:

```
errpt
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
9DBCfDEE 0713172600 T O errdemon ERROR LOGGING TURNED ON
9DBCfDEE 0713172400 T O errdemon ERROR LOGGING TURNED ON
192AC071 0713172400 T O errdemon ERROR LOGGING TURNED OFF
9DBCfDEE 0713172300 T O errdemon ERROR LOGGING TURNED ON
192AC071 0713171700 T O errdemon ERROR LOGGING TURNED OFF
...
35BfC499 0707112300 P H cd0 DISK OPERATION ERROR
0BA49C99 0707112300 T H scsi0 SCSI BUS ERROR
35BfC499 0707104000 P H cd0 DISK OPERATION ERROR
0BA49C99 0707104000 T H scsi0 SCSI BUS ERROR
369D049B 0706151600 I O SYSPFS UNABLE TO ALLOCATE SPACE IN FILE
SYSTEM
```

To display a complete detailed report, enter:

```
errpt -a

LABEL: ERRLOG_ON
IDENTIFIER: 9DBCfDEE

Date/Time: Thu Jul 13 17:26:11
Sequence Number: 143
Machine Id: 000FA17D4C00
Node Id: server2
Class: 0
Type: TEMP
Resource Name: errdemon

Description
ERROR LOGGING TURNED ON
```

Probable Causes  
ERRDEMON STARTED AUTOMATICALLY

User Causes  
/USR/LIB/ERRDEMON COMMAND

Recommended Actions  
NONE

...  
Date/Time: Thu Jul 6 15:16:09  
Sequence Number: 8  
Machine Id: 000FA17D4C00  
Node Id: server2  
Class: 0  
Type: INFO  
Resource Name: SYSPFS

Description  
UNABLE TO ALLOCATE SPACE IN FILE SYSTEM

Probable Causes  
FILE SYSTEM FULL

Recommended Actions  
USE FUSER UTILITY TO LOCATE UNLINKED FILES STILL REFERENCED  
INCREASE THE SIZE OF THE ASSOCIATED FILE SYSTEM  
REMOVE UNNECESSARY DATA FROM FILE SYSTEM

Detail Data  
MAJOR/MINOR DEVICE NUMBER  
002B 0003  
FILE SYSTEM DEVICE AND MOUNT POINT  
/dev/lv00, /u

To display a detailed report of all errors logged for the error identifier 369D049B,  
enter:

```
errpt -a -j 369D049B
```

```

LABEL: JFS_FS_FULL
IDENTIFIER: 369D049B
```

```
Date/Time: Thu Jul 6 15:16:09
Sequence Number: 8
Machine Id: 000FA17D4C00
Node Id: server2
Class: 0
Type: INFO
Resource Name: SYSPFS
```

Description  
UNABLE TO ALLOCATE SPACE IN FILE SYSTEM

Probable Causes  
FILE SYSTEM FULL

Recommended Actions  
USE FUSER UTILITY TO LOCATE UNLINKED FILES STILL REFERENCED  
INCREASE THE SIZE OF THE ASSOCIATED FILE SYSTEM  
REMOVE UNNECESSARY DATA FROM FILE SYSTEM

Detail Data  
MAJOR/MINOR DEVICE NUMBER  
002B 0003  
FILE SYSTEM DEVICE AND MOUNT POINT  
/dev/lv00, /u

To display a detailed report of all errors logged in the past 24 hours, enter:

```
date
Fri Jul 14 14:08:35 CDT 2000
errpt -a -s 0714140800
```

To list error record templates for which logging is turned off for any error log entries, enter:

```
errpt -t -F log=0
Id Label Type CL Description
AF6582A7 LVM_MISSPVRET UNKN S PHYSICAL VOLUME IS NOW ACTIVE
```

To view all entries from the alternate error log file */var/adm/ras/myerrorlog*, where *myerrorlog* is an alternative error log as specified with the **errdemon -i** command, enter:

```
errpt -i /var/adm/ras/myerrlog
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
192AC071 0713172300 T 0 errdemon ERROR LOGGING TURNED OFF
9DBCfDEE 0713172100 T 0 errdemon ERROR LOGGING TURNED ON
192AC071 0713172100 T 0 errdemon ERROR LOGGING TURNED OFF
9DBCfDEE 0713171900 T 0 errdemon ERROR LOGGING TURNED ON
192AC071 0713171900 T 0 errdemon ERROR LOGGING TURNED OFF
9DBCfDEE 0713171700 T 0 errdemon ERROR LOGGING TURNED ON
```

To view all hardware entries from the alternate error log file */var/adm/ras/errlog.alternate*, enter:

```
errpt -i /var/adm/ras/errlog.alternate -d H
```

To display a detailed report of all errors logged for the error label ERRLOG\_ON, enter:

```
errpt -a -J ERRLOG_ON
```

```

LABEL: ERRLOG_ON
IDENTIFIER: 9DBCFDÉE
```

```
Date/Time: Thu Jul 13 17:26:11
Sequence Number: 143
Machine Id: 000FA17D4C00
Node Id: server2
Class: 0
Type: TEMP
Resource Name: errdemon
```

```
Description
ERROR LOGGING TURNED ON
```

```
Probable Causes
ERRDEMON STARTED AUTOMATICALLY
```

```
User Causes
/USR/LIB/ERRDEMON COMMAND
```

```
Recommended Actions
NONE
```

```
...
```

```
LABEL: ERRLOG_ON
IDENTIFIER: 9DBCFDÉE
```

```
Date/Time: Fri Jul 7 17:00:46
Sequence Number: 14
Machine Id: 000FA17D4C00
Node Id: server2
Class: 0
Type: TEMP
Resource Name: errdemon
```

```
Description
ERROR LOGGING TURNED ON
```

```
Probable Causes
ERRDEMON STARTED AUTOMATICALLY
```

```
User Causes
/USR/LIB/ERRDEMON COMMAND
```

```
Recommended Actions
NONE
```

## 6.3.1 Error classes

The Error Notification object class specifies the conditions and actions to be taken when errors are recorded in the system error log. The user specifies these conditions and actions in an Error Notification object.

Each time an error is logged, the error notification daemon determines whether the error log entry matches the selection criteria of any of the Error Notification objects. If matches exist, the daemon runs the programmed action, also called a notify method, for each matched object.

The Error Notification object class is located in the `/etc/objrepos/errnotify` file. Error Notification objects are added to the object class by using Object Data Manager (ODM) commands. Error Notification objects contain the class descriptors provided in Table 6-3.

Table 6-3 List of error classes

| Class of error log entry | Description                                  |
|--------------------------|----------------------------------------------|
| H                        | Hardware error class                         |
| S                        | Software error class                         |
| O                        | Messages from the <b>errorlogger</b> command |
| U                        | Undetermined error class                     |

The following is an error entry example of when the system is rebooted. It is a Class S temporary error:

```
LABEL: REBOOT_ID
IDENTIFIER: 2BFA76F6

Date/Time: Mon Sep 16 11:24:12 CDT
Sequence Number: 202
Machine Id: 000BC6FD4C00
Node Id: server2
Class: S
Type: TEMP
Resource Name: SYSPROC
```

```
Description
SYSTEM SHUTDOWN BY USER
```

```
Probable Causes
SYSTEM SHUTDOWN
```

```
Detail Data
```

```
USER ID
 0
0=SOFT IPL 1=HALT 2=TIME REBOOT
 0
TIME TO REBOOT (FOR TIMED REBOOT ONLY)
 0
```

The following is an error entry example of when the system lost power for sysplanar0. It is a Class H permanent error:

```
LABEL: EPOW_SUS_CHRP
IDENTIFIER: BE0A03E5
```

```
Date/Time: Thu Sep 12 12:56:20 CDT
Sequence Number: 184
Machine Id: 00015F8F4C00
Node Id: server4
Class: H
Type: PERM
Resource Name: sysplanar0
Resource Class: planar
Resource Type: sysplanar_rspc
Location: 00-00
```

```
Description
ENVIRONMENTAL PROBLEM
```

```
Probable Causes
Power Turned Off Without a Shutdown
POWER OR FAN COMPONENT
```

```
Recommended Actions
RUN SYSTEM DIAGNOSTICS.
PERFORM PROBLEM DETERMINATION PROCEDURES
```

```
Detail Data
POWER STATUS REGISTER
0000 0005
PROBLEM DATA
03B4 0040 0000 004E C600 9500 1515 2300 2002 0917 0000 0005 4020 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 4942 4D00 0000 0000 000C 4646 0406 00B1
0000 0000 0010 4646 0406 00B0 0001 0085 0000 0000 0002 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
```



Resource Class: planar  
Resource Type: sysplanar\_rspc  
Location: 00-00

Description  
UNDETERMINED ERROR

Failure Causes  
UNDETERMINED

Recommended Actions  
RUN SYSTEM DIAGNOSTICS.

Detail Data  
PROBLEM DATA  
0344 0003 0000 0082 C600 8D08 1150 1500 2002 0913 4942 4D00 4000 0000 0000 0000  
0000 0000 0000 0000 0000 0000 0000 0000 4942 4D00 0000 0000 0050 3032 B119 4690  
03A0 005D A2B0 0000 0000 0000 0000 0701 0000 0000 0000 0000 0000 0000 0000 0000  
.....  
.....  
.....

The following is an error entry example of when an operator notification has been added to the error log. It is a Class O temporary error:

LABEL: OPMSG  
IDENTIFIER: AA8AB241

Date/Time: Thu Sep 12 12:56:30 CDT  
Sequence Number: 219  
Machine Id: 00015F8F4C00  
Node Id: server4  
Class: 0  
Type: TEMP  
Resource Name: OPERATOR

Description  
OPERATOR NOTIFICATION

User Causes  
ERRLOGGER COMMAND

Recommended Actions  
REVIEW DETAILED DATA

Detail Data  
MESSAGE FROM ERRLOGGER COMMAND  
This a test message!!!

## 6.4 The `errclear` command

The `errclear` command deletes error log entries older than the number of days specified by the `days` parameter. To delete all error log entries, specify a value of 0 for the `days` parameter.

If the `-i` flag is not used with the `errclear` command, the error log file cleared by `errclear` is the one specified in the error log configuration database. (To view the information in the error log configuration database, use the `errdemon` command.)

The `errclear` command syntax is as follows:

```
errclear [-d ErrorClassList] [-i File]
[-J ErrorLabel [,ErrorLabel]] | [-K ErrorLabel [,ErrorLabel]]
[-l SequenceNumber] [-m Machine] [-n Node] [-N ResourceNameList] [-R
ResourceTypeList] [-S ResourceClassList] [-T ErrorTypeList]
[-y FileName] [-j ErrorID [,ErrorID]] |
[-k ErrorID [,ErrorID]] Days
```

Table 6-4 provides the commonly used flags of the `errclear` command.

Table 6-4 Commonly used flags of the `errclear` command

| Flag                                            | Description                                                                                                                                                                                                                                                                                                                         |
|-------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>-d <i>List</i></code>                     | Deletes error log entries in the error classes specified by the list variable. The list variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters. The valid list variable values are H (hardware), S (software), O (errlogger messages), and U (undetermined). |
| <code>-i <i>File</i></code>                     | Uses the error log file specified by the file variable. If this flag is not specified, the <code>errclear</code> command uses the value from the error log configuration database.                                                                                                                                                  |
| <code>-j <i>ErrorID</i>[,<i>ErrorID</i>]</code> | Deletes the error log entries specified by the <code>ErrorID</code> (error identifier) variable. The <code>ErrorID</code> variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                                                            |
| <code>-J <i>ErrorLabel</i></code>               | Deletes the error log entries specified by the <code>ErrorLabel</code> variable. The <code>ErrorLabel</code> variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                                                                         |

| Flag                           | Description                                                                                                                                                                                                                                                                                                                                                          |
|--------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -k<br><i>ErrorID[,ErrorID]</i> | Deletes all error log entries except those specified by the ErrorID (error identifier) variable. The ErrorID variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                                                                                                          |
| -K <i>ErrorLabel</i>           | Deletes all error log entries except those specified by the ErrorLabel variable. The ErrorLabel variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                                                                                                                       |
| -l<br><i>SequenceNumber</i>    | Deletes error log entries with the specified sequence numbers. The SequenceNumber variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                                                                                                                                     |
| -m <i>Machine</i>              | Deletes error log entries for the machine specified by the machine variable. The <code>uname -m</code> command returns the value of the machine variable.                                                                                                                                                                                                            |
| -n <i>Node</i>                 | Deletes error log entries for the node specified by the node variable. The <code>uname -n</code> command returns the value of the node variable.                                                                                                                                                                                                                     |
| -N <i>List</i>                 | Deletes error log entries for the resource names specified by the list variable. For hardware errors, the list variable is a device name. For software errors, the list variable is the name of the unsuccessful executable. The list variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters. |
| -R <i>List</i>                 | Deletes error log entries for the resource types specified by the list variable. For hardware errors, the list variable is a device type. For software errors, the value of the list variable is LPP. The list variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                        |
| -S <i>List</i>                 | Deletes error log entries for the resource classes specified by the list variable. For hardware errors, the list variable is a device class. The list variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                                                                 |
| -T <i>List</i>                 | Deletes error log entries for error types specified by the list variable. Valid list variable values are: PERM, TEMP, PERF, PEND, INFO, and UNKN. The list variable values can be separated by commas, or enclosed in double quotation marks and separated by commas or space characters.                                                                            |

| Flag               | Description                                                                    |
|--------------------|--------------------------------------------------------------------------------|
| -y <i>FileName</i> | Uses the error record template file specified by the <i>FileName</i> variable. |

To delete all entries from the error log, enter:

```
errclear 0
```

To delete all entries in the error log classified as software errors, enter:

```
errclear -d S 0
```

To clear all entries from the alternate error log file `/var/adm/ras/errlog.alternate`, enter:

```
errclear -i /var/adm/ras/myerrlog 0
```

To clear all hardware entries from the alternate error log file `/var/adm/ras/errlog.alternate`, enter:

```
errclear -i /var/adm/ras/myerrlog -d H 0
```

**Note:** Once the **errclear** command has been run, it clears the error log, and this data is no longer available. To get this error information, the error log would have to be restored from a backup prior to running the **errclear** command.

## 6.5 Accounting

The accounting system allows you to collect and report on individual, group, and Workload Manager (WLM) class use of various system resources. This accounting information can be used to bill users for the system resources they utilize, and to monitor selected aspects of the system operation. To assist with billing, the accounting system provides the resource-usage totals defined by members of the adm group, and, if the **chargefee** command is included, factors in the billing fee. The fileset `bos.acct` contains the accounting commands and configuration files.

The accounting system also provides data to assess the adequacy of current resource assignments, set resource limits and quotas, forecast future needs, and order supplies for printers and other devices.

## 6.5.1 Setting up an accounting system

The following is an overview of the steps you must take to set up an accounting system. Refer to the commands and files noted in these steps for more specific information.

1. Use the **nulladm** command to ensure that each file has the correct access permission: read (r) and write (w) permission for the file owner and group, and read(r) permission for others by typing:

```
/usr/sbin/acct/nulladm wtmp pacct
```

This provides access to the **pacct** and **wtmp** files.

2. Update the **/etc/acct/holidays** file to include the hours you designate as prime time and to reflect your holiday schedule for the year.
3. Turn on process accounting by adding the following line to the **/etc/rc** file or by deleting the comment symbol (#) in front of the line if it exists:

```
/usr/bin/su - adm -c /usr/sbin/acct/startup
```

The **startup** procedure records the time that accounting was turned on and cleans up the previous day's accounting files.

4. Identify each file system you want included in disk accounting by adding the following line to the stanza for the file system in the **/etc/filesystems** file:

```
account = true
```

5. Specify the data file to use for printer data by adding the following line to the queue stanza in the **/etc/qconfig** file:

```
acctfile = /var/adm/qacct
```

6. As the **adm** user, create a **/var/adm/acct/nite**, a **/var/adm/acct/fiscal**, and a **/var/adm/acct/sum** directory to collect daily and fiscal period records:

```
su - adm
cd /var/adm/acct
mkdir nite fiscal sum
exit
```

7. Set daily accounting procedures to run automatically by editing the **/var/spool/cron/crontabs/root** file to include the **dodisk**, **ckpacct**, and **runacct** commands.
8. Set the monthly accounting summary to run automatically by including the **monacct** command in the **/var/spool/cron/crontabs/root** file. For example, type:

```
15 5 1 * * /usr/sbin/acct/monacct
```

Be sure to schedule this procedure early enough to finish the report. This example starts the procedure at 5:15 a.m. on the first day of every month.

9. To submit the edited cron file, type:

```
/crontab /var/spool/cron/crontabs/root
```

## 6.5.2 Setting up disk-usage accounting

The **dodisk** command initiates disk-usage accounting by calling the **diskusg** command and the **acctdisk** command. If you specify the **-o** flag with the **dodisk** command, a more thorough but slower version of disk accounting by login directory is initiated using the **acctdusg** command. Normally, the cron daemon runs the **dodisk** command. Syntax for this command is:

```
/usr/sbin/acct/dodisk [-o] [File ...]
```

The **-o** flag calls the **acctdusg** command, instead of the **diskusg** command, to initiate disk accounting by login directory.

By default, the **dodisk** command does disk accounting only on designated files with stanzas in the **/etc/filesystems** file and that contain the attribute **account=true**. If you specify file names with the **File** parameter, disk accounting is done on only those files.

If you do not specify the **-o** flag, the file parameter should contain the special file names of mountable file systems. If you specify both the **-o** flag and the file parameter, the files should be mount points of mounted file systems.

**Note:** You should not share accounting files among nodes in a distributed environment. Each node should have its own copy of the various accounting files.

This command should grant execute (x) access to members of the **adm** group only.

To start automatic disk-usage accounting, add the following to the **/var/spool/cron/crontabs/root** file:

```
0 2 * * 4 /usr/sbin/acct/dodisk
```

This example shows the instructions that the cron daemon will read and act upon. The **dodisk** command will run at 2 a.m. (0 2) each Thursday (4). This command is only one of the accounting instructions normally given to the cron daemon.

## 6.6 The syslogd daemon

The syslogd daemon receives information from other daemons and sends it to files, terminals, users, or other machines. Depending on its configuration, it only logs critical errors or debugging output. By default the file has no active entries. The syslogd daemon is started by /etc/rc.tcpip. The daemon is configured by /etc/syslog.conf, in which the priority of the information, its source, and where it should be sent to are specified.

The syslogd daemon is a subsystem of the System Resource Controller (SRC) and can be manipulated by the SRC commands described in Table 6-5.

Table 6-5 System Resource Controller commands

| SRC commands | Description                                                         |
|--------------|---------------------------------------------------------------------|
| startsrc     | Starts a subsystem, group of subsystems, or a subserver             |
| stopsrc      | Stops a subsystem, group of subsystems, or a subserver              |
| lssrc        | Gets the status of a subsystem, group of subsystems, or a subserver |

The syslogd daemon creates the /etc/syslog.pid file, which contains a single line with the command process ID used to end or reconfigure the syslogd daemon.

The /usr/include/sys/syslog.h include file defines the facility and priority codes used by the configuration file. Locally written applications use the definitions contained in the syslog.h file to log messages using the syslogd daemon.

The syslog subroutine writes messages onto the system log maintained by the **syslogd** command. The message is obtained from the errno global variable. Messages are read by the **syslogd** command and written to the system console or log file, or forwarded to the **syslogd** command on the appropriate host. Messages are tagged with codes indicating the type of priority for each. A priority is encoded as a facility, which describes the part of the system generating the message, and as a level, which indicates the severity of the message.

The syslogd daemon reads a datagram socket and sends each message line to a destination described by the /etc/syslog.conf configuration file. The syslogd daemon reads the configuration file when it is activated and when it receives a hangup signal. The syslogd daemon creates the /etc/syslog.pid file, which contains a single line with the command process ID used to end or reconfigure the syslogd daemon. A terminate signal sent to the syslogd daemon ends the daemon. The syslogd daemon logs the end-signal information and terminates immediately.

Each entry in the `/etc/syslog.conf` file must consist of two parts:

- ▶ A selector field to determine the message priorities to which the line applies
- ▶ An action

Each line can contain an optional part:

- ▶ Rotation

The fields must be separated by one or more tabs or spaces.

The selector field names a facility and a priority level. Separate facility names with a comma. Separate the facility and priority-level portions of the selector field with a period. Separate multiple entries in the same selector field with a semicolon (;). To select all facilities, use an asterisk (\*).

The Action field identifies a destination (file, host, or user) to receive the messages. If routed to a remote host, the remote system will handle the message as indicated in its own configuration file. To display messages on a user's terminal, the Destination field must contain the name of a valid, logged-in system user.

The Rotation field identifies how rotation is used. If the Action field is a file, then rotation can be based on size or time, or both. You can also compress and archive the rotated files.

Each message is one line. A message can contain a priority code, marked by a digit enclosed in angle braces (< >) at the beginning of the line. Messages longer than 900 bytes may be truncated.

Use the following system facility names in the Selector field:

|               |                           |
|---------------|---------------------------|
| <b>kern</b>   | Kernel                    |
| <b>user</b>   | User level                |
| <b>mail</b>   | Mail subsystem            |
| <b>daemon</b> | System daemons            |
| <b>auth</b>   | Security or authorization |
| <b>syslog</b> | syslogd daemon            |
| <b>lpr</b>    | Line-printer subsystem    |
| <b>news</b>   | News subsystem            |
| <b>uucp</b>   | uucp subsystem            |
| <b>*</b>      | All facilities            |

Use the following message priority levels in the Selector field. Messages of the specified priority level and all levels above it are sent as directed.

|                |                                                                                                                                                                          |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>emerg</b>   | Specifies emergency messages (LOG_EMERG). These messages are not distributed to all users. LOG_EMERG priority messages can be logged into a separate file for reviewing. |
| <b>alert</b>   | Specifies important messages (LOG_ALERT), such as a serious hardware error. These messages are distributed to all users.                                                 |
| <b>crit</b>    | Specifies critical messages not classified as errors (LOG_CRIT), such as improper login attempts. LOG_CRIT and higher-priority messages are sent to the system console.  |
| <b>err</b>     | Specifies messages that represent error conditions (LOG_ERR), such as an unsuccessful disk write.                                                                        |
| <b>warning</b> | Specifies messages for abnormal, but recoverable, conditions (LOG_WARNING).                                                                                              |
| <b>notice</b>  | Specifies important informational messages (LOG_NOTICE). Messages without a priority designation are mapped into this priority message.                                  |
| <b>info</b>    | Specifies informational messages (LOG_INFO). These messages can be discarded, but are useful in analyzing the system.                                                    |
| <b>debug</b>   | Specifies debugging messages (LOG_DEBUG). These messages may be discarded.                                                                                               |
| <b>none</b>    | Excludes the selected facility. This priority level is useful only if preceded by an entry with an asterisk (*) in the same selector field.                              |

Use the following message destinations in the Action field.

|                          |                                                |
|--------------------------|------------------------------------------------|
| <b>File name</b>         | Full path name of a file opened in append mode |
| <b>@host name</b>        | Host name, preceded by an at sign (@)          |
| <b>User[, User][...]</b> | User names                                     |
| <b>*</b>                 | All users                                      |

Use the following rotation keywords in the Rotation field.

|               |                                                        |
|---------------|--------------------------------------------------------|
| <b>rotate</b> | This keyword must be specified after the Action field. |
|---------------|--------------------------------------------------------|

|                 |                                                                                                                                                             |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>size</b>     | This keyword specifies that rotation is based on size. It is followed by a number and either a k (kilobytes) or m (megabytes).                              |
| <b>time</b>     | This keyword specifies that rotation is based on time. It is followed by a number and either an h (hour), d (day), w (week), m (month), or y (year).        |
| <b>files</b>    | This keyword specifies the total number of rotated files. It is followed by a number. If not specified, then there is an unlimited number of rotated files. |
| <b>compress</b> | This keyword specifies that the saved rotated files will be compressed.                                                                                     |
| <b>archive</b>  | This keyword specifies that the saved rotated files will be copied to a directory. It is followed by the directory name.                                    |

Following is an example of a `/etc/syslog.conf` configuration file modification to send all system messages, except those from the mail facility, to a host named `server4`. Type:

```
*.debug;mail.none @server4
```

## 6.7 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

1. A system has been experiencing intermittent problems. Which of the following procedures should be performed to determine if the problem has been occurring at a specific time or frequency rate?
  - A. Check the error log.
  - B. Look at the SMIT log.
  - C. Talk to all of the users.
  - D. Look at the crontab file.
2. Which of the following should be used as the device for reporting and buffering errors with time stamps?
  - A. `/etc/error`
  - B. `/dev/error`
  - C. `/usr/bin/error`
  - D. `/etc/dev/error`

3. A system operator accidentally deleted the error log with the **errclear 0** command before backing it up to a file. Which of the following procedures should be performed to retrieve the data?
  - A. **errpt -v**
  - B. **errpt -i /var/adm/ras/errlog**
  - C. **cat /var/adm/ras/errlog**
  - D. The data is lost and cannot be retrieved.
4. The syslogd daemon is typically started by:
  - A. /etc/inetd
  - B. /etc/rc.tcpip
  - C. /etc/rc.boot
  - D. /etc/inittab
5. The default /etc/syslog.conf configuration specifies that messages be logged for which of the following facilities?
  - A. auth
  - B. kern
  - C. none
  - D. daemon
6. The **dodisk** command is part of which fileset?
  - A. bos.adt.samples
  - B. devices.scsi.disk.rte
  - C. bos.acct
  - D. bos.rte.filesystems
7. To be able to monitor disk usage on a per-user basis, which of the following commands needs to be run?
  - A. **df**
  - B. **fileplace**
  - C. **dodisk**
  - D. **delta**

8. Which of the following options is the valid combination of selector, action, and rotation to use in the `/etc/syslog.conf` configuration file?
- A. \*.debug, @hostname, archive
  - B. system.alert, \*, backup
  - C. system.none, process ID, directories
  - D. tcp.none, process ID, files

## 6.7.1 Answers

The following are the preferred answers to the questions provided in this section.

1. A
2. B
3. D
4. B
5. C
6. C
7. C
8. A

## 6.8 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

1. Describe the two methods the **errpt** command uses to process a report.
2. Once the **errclear** command has been run to clear all entries in the error report, what is the only way to restore the error log?
3. What is the file name for the error log and the directory where it is kept?



## LVM, file system, and disk problem determination

The following topics are discussed in this chapter:

- ▶ Logical Volume Manager (LVM) problems
- ▶ Replacement of physical volumes
- ▶ JFS problems and their solutions
- ▶ Paging space creation and removal, as well as recommendations about paging space

To understand the problems that can happen on an AIX system with volume groups, logical volumes, and file systems, it is important to have a detailed knowledge about how the storage is managed by the Logical Volume Manager. This chapter does not cover the fundamentals of the LVM; they are considered prerequisite knowledge required to understand the issues addressed in this chapter.

## 7.1 LVM data

The Logical Volume Manager (LVM) data structures that are required for the LVM to operate are stored in a number of structures. This logical layout is described in the following sections.

### 7.1.1 Physical volumes

Each disk is assigned a Physical Volume Identifier (PVID) when it is first assigned to a volume group. The PVID is a combination of the serial number of the machine creating the volume group and the time and date of the operation. The PVID is stored on the physical disk itself and is also stored in the Object Data Manager (ODM) of a machine when a volume group is created or imported.

You should not use the **dd** command to copy the contents of one physical volume to another, since the PVID will also be copied; this will result in two disks having the same PVID, which can confuse the system.

### 7.1.2 Volume groups

Each volume group has a Volume Group Descriptor Area (VGDA). There are (commonly) multiple copies of the VGDA in a volume group. A copy of the VGDA is stored on each disk in the volume group. The VGDA stores information about the volume group, such as the logical volumes and the disks in the volume group.

The VGDA is parsed by the **importvg** command when importing a volume group into a system. It is also used by the **varyonvg** command in the quorum voting process to decide whether a volume group should be varied on.

For a single disk volume group, there are two VGDA's on the disk. When a second disk is added to make a two disk volume group, the original disk retains two VGDA's and the new disk gets one VGDA.

Adding a third disk results in the extra VGDA from the first disk moving to the third disk for a quorum of three with each disk having one vote. Adding this additional disk adds a new VGDA per disk.

A volume group with quorum checking enabled (the default) must have at least 51 percent of the VGDA's in the volume group available before it can be varied on. Once varied on, if the number of VGDA's falls below 51 percent, the volume group will automatically be varied off.

In contrast, a volume group with quorum checking disabled must have 100 percent of the VGDA's available before it can be varied on. Once varied on, only one VGDA needs to remain available to keep the volume group online.

A volume group also has a Volume Group Identifier (VGID), a soft serial number for the volume group similar to the PVID for disks.

Each disk in a volume group also has a Volume Group Status Area (VGSA), a 127 byte structure used to track mirroring information for up to the maximum 1016 physical partitions on the disk.

### 7.1.3 Logical volumes

Each logical volume has a Logical Volume Control Block (LVCB) that is stored in the first 512 bytes of the logical volume. The LVCB holds important details about the logical volume, including its creation time, mirroring information, and mount point (if it contains a journaled file system [JFS]).

Each logical volume has a Logical Volume Identifier (LVID) that is used to represent the logical volume to the LVM libraries and low-level commands. The LVID is made up of `VGID.num`, where *num* is the order in which it was created in the volume group.

### 7.1.4 Object Data Manager (ODM)

The Object Data Manager is used by the LVM to store information about the volume groups, physical volumes, and logical volumes on the system. The information held in the ODM is placed there when the volume group is imported or when each object in the volume group is created.

There exists an ODM object known as the `vg-lock`. Whenever an LVM modification command is started, the LVM command will lock the `vg-lock` for the volume group being modified. If for some reason the lock is inadvertently left behind, the volume group can be unlocked by running the `varyonvg -b` command, which can be run on a volume group that is already varied on.

## 7.2 LVM problem determination

The most common LVM problems are related to disk failures. Depending on the extent of the failure, you may be able to recover the situation with little or no data loss. However, a failed recovery attempt may leave the system in a worse condition. This leaves restoring from backup as the only way to recover. Therefore, always take frequent backups of your system.

## 7.2.1 Data relocation

When a problem occurs with a disk drive, data relocation may take place. There are three types of data relocation, namely:

- ▶ Internal to the disk
- ▶ Hardware relocation ordered by LVM
- ▶ Software relocation

Relocation typically occurs when the system fails to perform a read or write due to physical problems with the disk platter. In some cases, the data I/O request completes but with warnings. Depending on the type of recovered error, the LVM may be wary of the success of the next request to that physical location, and it orders a relocation to be on the safe side.

The lowest logical layer of relocation is the one that is internal to the disk. These types of relocations are typically private to the disk and there is no notification to the user that a relocation occurred.

The next level up in terms of relocation complexity is a hardware relocation called for by the LVM device driver. This type of relocation will instruct the disk to relocate the data on one physical partition to another portion (reserved) of the disk. The disk takes the data in physical location A and copies it to a reserved portion of the disk (location B). However, after this is complete, the LVM device driver will continue to reference physical location A, with the understanding that the disk itself will handle the true I/O to the real location B.

The top layer of data relocation is the *soft* relocation handled by the LVM device driver. In this case, the LVM device driver maintains a bad block directory, and whenever it receives a request to access logical location A, the LVM device driver will look up the bad block table and translate it to actually send the request to the disk drive at physical location B.

## 7.2.2 Backup data

The first step you should perform if you suspect a problem with LVM is to make a backup of the affected volume group and save as much data as possible. This may be required for data recovery. The integrity of the backup should be compared with the last regular backup taken before the problem was detected.

## 7.2.3 ODM resynchronization

Problems with the LVM tend to occur when a physical disk problem causes the ODM data to become out of sync with the VGDA, VGSA, and LVCB information stored on disk.

ODM corruption can also occur if an LVM operation terminates abnormally and leaves the ODM in an inconsistent state. This may happen, for example, if the file system on which the ODM resides (normally root, /) becomes full during the process of importing a volume group.

If you suspect that the ODM entries for a particular volume group have been corrupted, a simple way to resynchronize the entries is to varyoff and export the volume group from the system, then import and varyon to refresh the ODM. This process can only be performed for non-rootvg volume groups.

For the rootvg volume group, you can use the **redefinevg** command, which examines every disk in the system to determine which volume group it belongs to and then updates the ODM. For example:

```
redefinevg rootvg
```

If you suspect that the LVM information stored on disk has become corrupted, use the **sync1vodm** command to synchronize and rebuild the LVCB, the device configuration database, and the VGDA's on the physical volumes. For example:

```
sync1vodm -v myvg
```

If you have a volume group where one or more logical volumes is mirrored, use the **syncvg** command if you suspect that one or more mirrored copies has become stale. The command can be used to resynchronize an individual logical volume, a physical disk, or an entire volume group. For example:

```
syncvg -l lv02
```

The above command synchronizes the mirror copies of the logical volume lv02.

```
syncvg -v myvg
```

The above command synchronizes all of the logical volumes in the volume group myvg.

## 7.2.4 Understanding importvg problems

If importing a volume group into a system is not possible using the **importvg** command, the following areas are the typical problem areas:

- ▶ AIX version level
- ▶ Invalid PVID
- ▶ Disk change while volume group was exported
- ▶ Shared disk environment

In general, if the **importvg** command is unsuccessful, check the error log for information that can point to the problem.

## AIX version level

Verify that the volume group you are importing is supported by the level of AIX running on the system. Various new features have been added to the LVM system at different levels of AIX, such as support for large volume groups. A number of these features require a change to the format of the VGDA stored on the disk, and thus will not be understood by previous levels of AIX.

## Invalid PVID

Check that all of the disks in the volume group you are trying to import are marked as available to AIX and have valid PVIDs stored in the ODM. This can be checked using the **lspv** command. If any disks do not have a PVID displayed, use the **chdev** command to resolve the problem. For example:

```
lspv
hdisk0 000bc6fdc3dc07a7 rootvg
hdisk1 000bc6fdbff75ee2 testvg
hdisk2 000bc6fdbff92812 testvg
hdisk3 000bc6fdbff972f4 None
hdisk4 None None
chdev -l hdisk4 -a pv=yes
hdisk4 changed
lspv
hdisk0 000bc6fdc3dc07a7 rootvg
hdisk1 000bc6fdbff75ee2 testvg
hdisk2 000bc6fdbff92812 testvg
hdisk3 000bc6fdbff972f4 None
hdisk4 000bc6fd672864b9 None
```

In this example, the PVID for hdisk4 is not shown by the **lspv** command. This is resolved by running the **chdev** command. The PVID is read from the disk and placed in the ODM, if the disk is accessible. It will only write a new PVID if there truly is no PVID on the disk. Alternately, the disk can be removed using the **rmdev** command and, by running the configuration manager **cfgmgr** command, the device is recreated with the correct PVID. After this, an import of the volume group with the **importvg** command should be possible.

## Disk change while volume group was exported

If the **importvg** command fails with an error message similar the following, the physical volume is marked missing and it is possible that some disk change to the disks defined in the volume group was made while the volume group was exported:

```
0516-056 varyon testvg: The volume group is not varied on because a
physical volume is marked missing. Run diagnostics.
```

Check the error log with the **errpt** command in order to see what happened to the respective disk.

In order to force the volume group to be varied online, use the **-f** flag of the **importvg** command. This makes it possible to operate on the volume group and, depending on the situation, reconfigure the volume group by excluding the disk that is marked missing with the **reducevg** command.

### Shared disk environment

In a shared disk environment, such as an SSA disk system, used by two or more systems, it is possible that the physical volumes defined are not accessible because they are already imported and *varied on* by another machine. Check the volume groups on both machines and compare the PVIDs by using the **lspv** command.

## 7.2.5 Extending the number of max physical partitions

When adding a new disk to a volume group, you may encounter an error due to there being too few PP descriptors for the required number of PVs. This may occur when the new disk has a much higher capacity than existing disks in the volume group.

This situation is typical on older installations, due to the rapid growth of storage technology. To overcome this, a change of the volume group LVM metadata is required.

The **chvg** command is used for this operation using the **-t** flag and applying a factor value, as shown in the following example:

```
lsvg testvg
VOLUME GROUP: testvg VG IDENTIFIER: 000bc6fd5a177ed0
VG STATE: active PP SIZE: 16 megabyte(s)
VG PERMISSION: read/write TOTAL PPs: 542 (8672 megabytes)
MAX LVs: 256 FREE PPs: 42 (672 megabytes)
LVs: 1 USED PPs: 500 (8000 megabytes)
OPEN LVs: 0 QUORUM: 2
TOTAL PVs: 1 VG DESCRIPTORS: 2
STALE PVs: 0 STALE PPs: 0
ACTIVE PVs: 1 AUTO ON: yes
MAX PPs per PV: 1016 MAX PVs: 32
chvg -t 2 testvg
0516-1193 chvg: WARNING, once this operation is completed, volume group testvg
cannot be imported into AIX 430 or lower versions. Continue (y/n) ?
y
0516-1164 chvg: Volume group testvg changed. With given characteristics testvg
can include upto 16 physical volumes with 2032 physical partitions
each.
```

This example shows that the volume group testvg with a current 9.1 GB disk has a maximum number of 1016 PPs per physical volume. Adding a larger 18.2 GB disk would not be possible; the maximum size of the disk is limited to 17 GB unless the maximum number of PPs is increased. Using the **chvg** command to increase the maximum number of PPs by a factor of 2 to 2032 PPs allows the volume group to be extended with physical volumes of up to approximately 34 GB.

## 7.3 Disk replacement

AIX, like all operating systems, can be problematic when you have to change a disk. AIX provides the ability to prepare the system for the change using the LVM. You can then perform the disk replacement and then use the LVM to restore the system back to how it was before the disk was changed. This process manipulates not only the data on the disk itself, but is also a way of keeping the Object Data Manager (ODM) intact.

The ODM within AIX is a database that holds device configuration details and AIX configuration details. The function of the ODM is to store the information between reboots, and also provide rapid access to system data, eliminating the need for AIX commands to interrogate components for configuration information. Since this database holds so much vital information regarding the configuration of a machine, any changes made to the machine, such as the changing of a defective disk, need to be done in such a way as to preserve the integrity of the database.

### 7.3.1 Replacing a disk

The following scenario shows a system that has a hardware error on a physical volume. However, since the system uses a mirrored environment, which has multiple copies of the logical volume, it is possible to replace the disk while the system is active. The disk hardware in this scenario are hot-swappable SCSI disks, which permit the replacement of a disk in a production environment.

One important factor is detecting the disk error. Normally, mail is sent to the system administrator (root account) from the Automatic Error Log Analysis (diagela). Figure 7-1 on page 145 shows the information in such a diagnostics mail.

```

Message 13:
From root Fri Jul 14 03:00:33 2000
Date: Fri, 14 Jul 2000 03:00:33 -0500
From: root
To: root
Subject: diagela

A PROBLEM WAS DETECTED ON Fri Jul 14 03:00:26 CDT 2000 801014

The Service Request Number(s)/Probable Cause(s)
(causes are listed in descending order of probability):

440-129: Error log analysis indicates a SCSI bus problem.
n/a FRU: n/a 10-60-00-12,0
 SCSI bus problem: cables, terminators or other SCSI
 devices
hdisk4 FRU: 25L3101 10-60-00-12,0
 16 Bit SCSI Disk Drive (9100 MB)
pci0 FRU: 03N2826 P2
 PCI Bus
n/a FRU: n/a 10-60-00-12,0
 Software

? █

```

Figure 7-1 Disk problem mail from Automatic Error Log Analysis (diagela)

Automatic Error Log Analysis (diagela) provides the capability to do error log analysis whenever a permanent hardware error is logged. Whenever a permanent hardware resource error is logged, the diagela program is invoked. Automatic Error Log Analysis is enabled by default on all platforms.

The diagela message shows that the hdisk4 has a problem. Another way of locating a problem is to check the state of the logical volume using the `lsvg` command, as in the following example:

```

lsvg -l mirrorvg
mirrorvg:
LV NAME TYPE LPs PPs PVs LV STATE MOUNT POINT
lvdb01 jfs 500 1000 2 open/syncd /u/db01
lvdb02 jfs 500 1000 2 open/stale /u/db02
loglv00 jfslog 1 1 1 open/syncd N/A

```

The logical volume lvdb02 in the volume group mirrorvg is marked with the status `stale`, indicating that the copies in this LV are not synchronized. Look at the error log using the error-reporting `errpt` command, as in the following example:

```

errpt
EAA3D429 0713121400 U S LVDD PHYSICAL PARTITION MARKED STALE
F7DDA124 0713121400 U H LVDD PHYSICAL VOLUME DECLARED MISSING
41BF2110 0713121400 U H LVDD MIRROR WRITE CACHE WRITE FAILED
35BFC499 0713121400 P H hdisk4 DISK OPERATION ERROR

```

This error information displays the reason why the LV lvdb02 is marked *stale*. The hdisk4 had an DISK OPERATION ERROR and the LVDD could not write the mirror cache.

Based on the information in the example, hdisk4 needs to be replaced. Before taking any action on the physical disk of the mirrored LV are recommended that you do a file system backup in case anything should go wrong. Since the other disk of the mirrored LV is still functional, all the data should be present. If the LV contains a database, then the respective database tools for backup of the data should be used.

## Removing a bad disk

If the system is a high-availability (24x7) system, you might decide to keep the system running while performing the disk replacement, provided that the hardware supports an online disk exchange with hot-swappable disks. However, the procedure should be agreed upon by the system administrator or customer before continuing. Use the following steps to remove a disk:

1. To remove the physical partition copy of the mirrored logical volume from the erroneous disk, use the **rm1vcopy** command as follows:

```
rm1vcopy lvdb02 1 hdisk4
```

The logical volume lvdb02 is now left with only one copy, as shown in the following:

```
ls1v -l lvdb02
lvdb02:/u/db02
PV COPIES IN BAND DISTRIBUTION
hdisk3 500:000:000 21% 109:108:108:108:067
```

2. Reduce the volume group by removing the disk you want to replace from its volume group:

```
reducevg -f mirrorvg hdisk4
```

```
lsvg -l mirrorvg
mirrorvg:
LV NAME TYPE LPs PPs PVs LV STATE MOUNT POINT
lvdb01 jfs 500 1000 2 open/syncd /u/db01
lvdb02 jfs 500 500 1 open/syncd /u/db02
loglv00 jfs1log 1 1 1 open/syncd N/A
```

3. Remove the disk as a device from the system and from the ODM database with the **rmdev** command:

```
rmdev -d -l hdisk4
hdisk4 deleted
```

This command is valid for any SCSI disk. If your system is using SSA, then an additional step is required. Since SSA disks also define the device pdisk, the

corresponding `pdisk` device must be deleted as well. Use the SSA menus in `SMIT` to display the mapping between `hdisk` and `pdisk`. These menus can also be used to delete the `pdisk` device.

4. The disk can now be safely removed from your system.

## Adding a new disk

Continuing the scenario from the previous section, this section describes how to add a new disk into a running environment. After `hdisk4` has been removed, the system is now left with the following disks:

```
lsdev -Cc disk
hdisk0 Available 30-58-00-8,0 16 Bit SCSI Disk Drive
hdisk1 Available 30-58-00-9,0 16 Bit SCSI Disk Drive
hdisk2 Available 10-60-00-8,0 16 Bit SCSI Disk Drive
hdisk3 Available 10-60-00-9,0 16 Bit SCSI Disk Drive
```

Use the following steps to add a new disk:

1. Plug in the new disk and run the configuration manager `cfgmgr` command. The `cfgmgr` command configures devices controlled by the Configuration Rules object class, which is part of the device configuration database. The `cfgmgr` command will see the newly inserted SCSI disk and create the corresponding device. Although the command requires no option, the `-v` flag specifies verbose output, which helps in troubleshooting, as shown in the following:

```
cfgmgr -v
cfgmgr is running in phase 2

Time: 0 LEDS: 0x538
Invoking top level program -- "/etc/methods/cfgprobe -c
/etc/drivers/coreprobe.ext"
Time: 0 LEDS: 0x539
Return code = 0
*** no stdout ****
*** no stderr ****

Time: 0 LEDS: 0x538
Invoking top level program -- "/etc/methods/defsys"
Time: 0 LEDS: 0x539
Return code = 0
***** stdout *****
sys0
.....
.....
```

The result is a new `hdisk4` added to the system:

```
lsdev -Cc disk
```

```

hdisk0 Available 30-58-00-8,0 16 Bit SCSI Disk Drive
hdisk1 Available 30-58-00-9,0 16 Bit SCSI Disk Drive
hdisk2 Available 10-60-00-8,0 16 Bit SCSI Disk Drive
hdisk3 Available 10-60-00-9,0 16 Bit SCSI Disk Drive
hdisk4 Available 10-60-00-12,0 16 Bit SCSI Disk Drive

```

- The new hdisk must now be assigned to the volume group mirrorvg by using the LVM **extendvg** command:

```
extendvg mirrorvg hdisk4
```

- To re-establish the mirror copy of the LV, use the **mk1vcopy** command.

```
mk1vcopy lvdb02 2 hdisk4
```

The number of copies of LV is now two, but the LV stat is still marked as *stale* because the LV copies are not synchronized with each other:

```
lsvg -l mirrorvg
mirrorvg:
LV NAME TYPE LPs PPs PVs LV STATE MOUNT POINT
lvdb01 jfs 500 1000 2 open/syncd /u/db01
lvdb02 jfs 500 1000 2 open/stale /u/db02
loglv00 jfslog 1 1 1 open/syncd N/A
```

- To get a fully synchronized set of copies of the LV lvdb02, use the **syncvg** command:

```
syncvg -p hdisk4
```

The **syncvg** command can be used with logical volumes, physical volumes, or volume groups. The synchronization process can be quite time consuming, depending on the hardware characteristics and the amount of data.

After the synchronization is finished, verify the logical volume state using either the **lsvg** or **lslv** command:

```
lsvg -l mirrorvg
mirrorvg:
LV NAME TYPE LPs PPs PVs LV STATE MOUNT POINT
lvdb01 jfs 500 1000 2 open/syncd /u/db01
lvdb02 jfs 500 1000 2 open/syncd /u/db02
loglv00 jfslog 1 1 1 open/syncd N/A
```

The system is now back to normal.

### 7.3.2 Recovering an incorrectly removed disk

If a disk was incorrectly removed from the system, and the system has been rebooted, the **sync1vadm** command will need to be run to rebuild the logical volume control block, as shown in the following examples.

In the examples, a disk has been incorrectly removed from the system and the logical volume control block needs to be rebuilt.

The disks in the system before the physical volume was removed is shown in the following command output:

```
lsdev -Cc disk
hdisk0 Available 30-58-00-8,0 16 Bit SCSI Disk Drive
hdisk1 Available 30-58-00-9,0 16 Bit SCSI Disk Drive
hdisk2 Available 10-60-00-8,0 16 Bit SCSI Disk Drive
hdisk3 Available 10-60-00-9,0 16 Bit SCSI Disk Drive
```

The allocation of the physical volumes before the disk was removed are shown as follows:

```
lspv
hdisk0 000bc6fdc3dc07a7 rootvg
hdisk1 000bc6fdbff75ee2 volg01
hdisk2 000bc6fdbff92812 volg01
hdisk3 000bc6fdbff972f4 volg01
```

The logical volumes on the volume group:

```
lsvg -l volg01
volg01:
LV NAME TYPE LPs PPs PVs LV STATE MOUNT POINT
logvol01 jfs 1000 1000 2 open/syncd /userfs01
loglv00 jfslog 1 1 1 open/syncd N/A
```

The logical volume distribution on the physical volumes is shown using the **lslv** command:

```
lslv -l logvol01
logvol01:/userfs01
PV COPIES IN BAND DISTRIBUTION
hdisk1 542:000:000 19% 109:108:108:108:109
hdisk3 458:000:000 23% 109:108:108:108:025
```

The system after a reboot has the following physical volumes:

```
lspv
hdisk0 000bc6fdc3dc07a7 rootvg
hdisk1 000bc6fdbff75ee2 volg01
hdisk3 000bc6fdbff972f4 volg01
```

When trying to mount the file system on the logical volume, the error may look similar to the following example:

```
mount /userfs01
mount: 0506-324 Cannot mount /dev/logvol01 on /userfs01: There is an input or output error.
```

To synchronize the logical volume, the following command should be run:

```
syncldvdm -v volg01
syncldvdm: Physical volume data updated.
syncldvdm: Logical volume logvol01 updated.
syncldvdm: Warning, lv control block of loglv00 has been over written.
0516-622 syncldvdm: Warning, cannot write lv control block data.
syncldvdm: Logical volume loglv00 updated.
```

The system can now be repaired. If the file system data was spread across all the disks, including the failed disk, it may need to be restored from the last backup.

## 7.4 The AIX JFS

Similar to the LVM, most JFS problems can be traced to problems with the underlying physical disk.

As with volume groups, various JFS features have been added at different levels of AIX, which preclude those file systems being mounted if the volume group was imported on an earlier version of AIX. Such features include large file enabled file systems, file systems with non-default allocation group size, and JFS2.

### 7.4.1 Creating a JFS

In a journaled file system (JFS), files are stored in blocks of contiguous bytes. The default block size, also referred to as fragmentation size in AIX, is 4096 bytes (4 KB). The JFS i-node contains an information structure of the file with an array of eight pointers to data blocks. A file that is less than 32 KB is referenced directly from the i-node.

A larger file uses a 4-KB block, referred to as an indirect block, for the addressing of up to 1024 data blocks. Using an indirect block, a file size of  $1024 \times 4 \text{ KB} = 4 \text{ MB}$  is possible.

For files larger than 4 MB, a second block, the double indirect block, is used. The double indirect block points to 512 indirect blocks, providing the possible addressing of  $512 \times 1024 \times 4 \text{ KB} = 2 \text{ GB}$  files. Figure 7-2 on page 151 illustrates the addressing using double indirection.

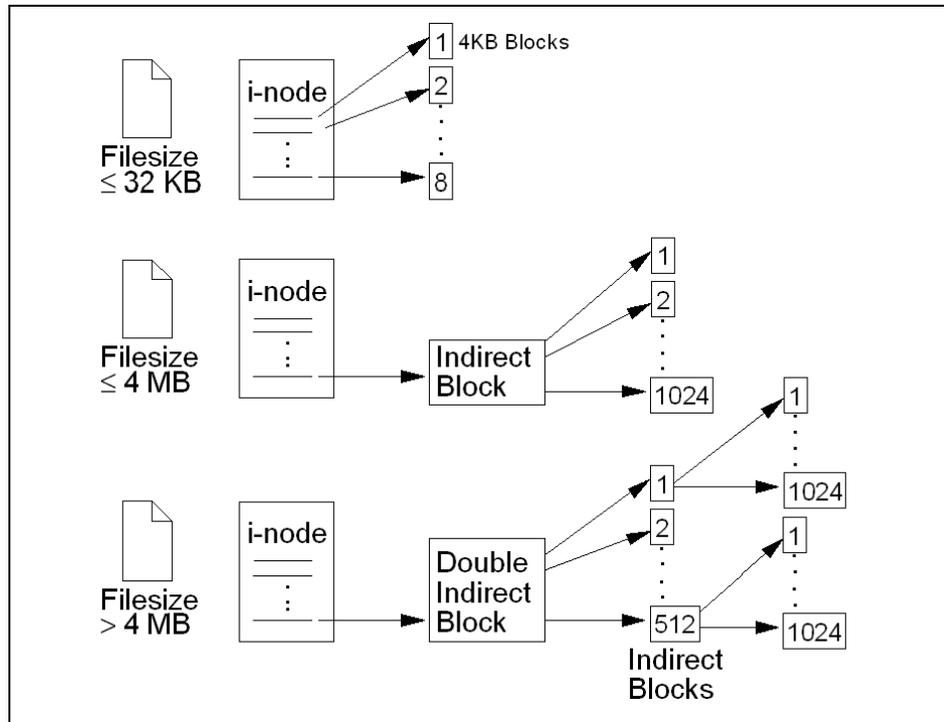


Figure 7-2 JFS organization

AIX Version 4.2 and later supports even larger files by defining a new type of JFS named the *bigfile* file system. In the bigfile file system, the double indirect use references 128-KB blocks rather than 4-KB blocks. However, the first indirect block still points to a 4-KB block, so the large blocks are only used when the file size is above 4 MB. This provides a new maximum file size of just under 64 GB.

When creating a JFS, the structure is defined on either a new logical volume or an already defined logical volume. The parameters of a defined JFS can be displayed either using SMIT menus (**smitt jfs**) or by using the **lsjfs** command:

```
lsjfs /u/testfs
#MountPoint:Device:Vfs:Nodename:Type:Size:Options:AutoMount:Acct:OtherOptions:L
vSize:FfsSize:FragSize:Nbpi:Compress:Bf:AgSize:
/u/testfs:/dev/1v03:jfs:::425984:rw:yes:no::425984:425984:4096:4096:no:false:8:
```

The **lsjfs** command shows the JFS attributes directly using colon (:) and delimiter.

## 7.4.2 Increasing the file system size

In many instances, the size of a file system needs to be increased because the demand for storage has increased. In AIX, this is a common procedure, and it is possible to do by using the **chfs** command, as in the following example:

```
chfs -a size=+300000 /u/testfs
Filesystem size changed to 458752
```

This example shows how the file system `testfs` is extended with 300000 512-byte blocks. When the file system is extended, the logical volume holding the JFS is also extended, with the number of logical partitions that is needed to fulfill the space request. If the system does not have enough free space, the volume group can either be extended with an additional physical volume, or the size specified for the **chfs** command must be lowered so that it matches the number of free LPs.

## 7.4.3 File system verification and recovery

The **fsck** command checks and interactively repairs inconsistent file systems. You should run this command before mounting any file system. You must be able to read the device file on which the file system resides (for example, the `/dev/hd0` device).

Normally, the file system is consistent, and the **fsck** command merely reports on the number of files, used blocks, and free blocks in the file system. If the file system is inconsistent, the **fsck** command displays information about the inconsistencies found and prompts you for permission to repair them. If the file system cannot be repaired, restore it from backup.

Mounting an inconsistent file system may result in a system crash. If you do not specify a file system with the `FileSystem` parameter, the **fsck** command will check all the file systems with the attribute `check=TRUE` in `/etc/filesystems`.

**Note:** By default, the `/`, `/usr`, `/var`, and `/tmp` file systems have the `check` attribute set to `false` (`check=false`) in their `/etc/filesystems` stanzas. The attribute is set to `false` for the following reasons:

- ▶ The boot process explicitly runs the **fsck** command on the `/`, `/usr`, `/var`, and `/tmp` file systems.
- ▶ The `/`, `/usr`, `/var`, and `/tmp` file systems are mounted when the `/etc/rc` file is run. The **fsck** command will not modify a mounted file system, and **fsck** results on mounted file systems are unpredictable.

## Fixing a bad superblock

If you receive one of the following errors from the **fsck** or **mount** commands, the problem may be a corrupted superblock, as shown in the following example:

```
fsck: Not an AIX3 file system
fsck: Not an AIXV3 file system
fsck: Not an AIX4 file system
fsck: Not an AIXV4 file system
fsck: Not a recognized file system type
mount: invalid argument
```

The problem can be resolved by restoring the backup of the superblock over the primary superblock using the following command (care should be taken to check with the latest product documentation before running this command):

```
dd count=1 bs=4k skip=31 seek=1 if=/dev/lv00 of=/dev/lv00
```

The following is an example of when the superblock is corrupted and copying the backup helps solve the problem:

```
mount /u/testfs
mount: 0506-324 Cannot mount /dev/lv02 on /u/testfs: A system call received a
parameter that is not valid.
fsck /dev/lv02
```

```
Not a recognized filesystem type. (TERMINATED)
```

```
dd count=1 bs=4k skip=31 seek=1 if=/dev/lv02 of=/dev/lv02
1+0 records in.
1+0 records out.
```

```
fsck /dev/lv02
```

```
** Checking /dev/lv02 (/u/tes)
** Phase 0 - Check Log
log redo processing for /dev/lv02
** Phase 1 - Check Blocks and Sizes
** Phase 2 - Check Pathnames
** Phase 3 - Check Connectivity
** Phase 4 - Check Reference Counts
** Phase 5 - Check Inode Map
** Phase 6 - Check Block Map
8 files 2136 blocks 63400 free
```

Once the restoration process is complete, check the integrity of the file system by issuing the **fsck** command:

```
fsck /dev/lv00
```

In many cases, restoration of the backup of the superblock to the primary superblock will recover the file system. If this does not resolve the problem, recreate the file system and restore the data from a backup.

#### 7.4.4 Sparse file allocation

Some applications, particularly databases, maintain data in sparse files. Files that do not have disk blocks allocated for each logical block are called sparse files. If the file offsets are greater than 4 MB, then a large disk block of 128 KB is allocated. Applications using sparse files larger than 4 MB may require more disk blocks in a file system enabled for large files than in a regular file system.

In the case of sparse files, the output of the `ls` command is not showing the actual file size, but is reporting the number of bytes between the first and last blocks allocated to the file, as shown in the following example:

```
ls -l /tmp/sparsefile
-rw-r--r-- 1 root system 100000000 Jul 16 20:57 /tmp/sparsefile
```

The `du` command can be used to see the actual allocation, since it reports the blocks actually allocated and in use by the file. Use `du -rs` to report the number of allocated blocks on disk.

```
du -rs /tmp/sparsefile
256 /tmp/sparsefile
```

**Note:** The `tar` command does not preserve the sparse nature of any file that is sparsely allocated. Any file that was originally sparse before the restoration will have all space allocated within the file system for the size of the file. New AIX 5L options for the `backup` and `restore` command are useful for sparse files.

Using the `dd` command in combination with your own backup script will solve this problem.

#### 7.4.5 Unmount problems

A file system cannot be unmounted if any references are still active within that file system. The following error message will be displayed:

Device busy

or

A device is already mounted or cannot be unmounted

The following situations can leave open references to a mounted file system.

- ▶ Files are open within a file system. These files must be closed before the file system can be unmounted. The **fuser** command is often the best way to determine what is still active in the file system. The **fuser** command will return the process IDs for all processes that have open references within a specified file system, as shown in the following example:

```
umount /home
umount: 0506-349 Cannot unmount /dev/hd1: The requested resource is busy.
fuser -x -c /home
/home: 11630
ps -fp 11630
 UID PID PPID C STIME TTY TIME CMD
 guest 11630 14992 0 16:44:51 pts/1 0:00 -sh
kill -1 11630
umount /home
```

The process having an open reference can be killed by using the **kill** command (sending a **SIGHUP**), and the unmount can be accomplished. A stronger signal may be required, such as **SIGKILL**.

- ▶ If the file system is still busy and still cannot be unmounted, this could be due to a kernel extension that is loaded but exists within the source file system. The **fuser** command will not show these kinds of references, since a user process is not involved. However, the **genkex** command will report on all loaded kernel extensions.
- ▶ File systems are still mounted within the file system. Unmount these file systems before the file system can be unmounted. If any file system is mounted within a file system, this leaves open references in the source file system at the mount point of the other file system. Use the **mount** command to get a list of mounted file systems. Unmount all the file systems that are mounted within the file system to be unmounted.

## 7.4.6 Removing file systems

When removing a JFS, the file system must be unmounted before it can be removed. The command for removing file systems is **rmfs**.

In the case of a JFS, the **rmfs** command removes both the logical volume on which the file system resides and the associated stanza in the `/etc/filesystems` file. If the file system is not a JFS, the command removes only the associated stanza in the `/etc/filesystems` file, as shown in the following example:

```
lsvg -l testvg
testvg:
LV NAME TYPE LPs PPs PVs LV STATE MOUNT POINT
loglv00 jfslog 1 1 1 open/syncd N/A
```

```
lv02 jfs 2 2 1 open/syncd /u/testfs
rmfs /u/testfs
rmfs: 0506-921 /u/testfs is currently mounted.
umount /u/testfs
rmfs /u/testfs
rmlv: Logical volume lv02 is removed.
lsvg -l testvg
testvg:
LV NAME TYPE LPs PPs PVs LV STATE MOUNT POINT
loglv00 jfslog 1 1 1 closed/syncd N/A
```

This example shows how the file system testfs is removed. The first attempt fails because the file system is still mounted. The associated logical volume lv02 is also removed. The jfslog remains defined on the volume group.

### 7.4.7 Different output from du and df commands

Sometimes **du** and **df** commands are used to get a free block value. **df** is used to report the total block count, and then the value returned by **du -s /filesystem\_name** is subtracted from that total to calculate the free block value. However, this method of calculation yields a value that is greater than the free block value reported by **df**. At AIX Version 4.1 and later, both **df** and **du** default to 512-byte units. Sample output from the **du** and **df** commands is below:

```
du -s /tmp
152 /tmp
df /tmp
Filesystem 512-blocks Free %Used Iused %Iused Mounted on
/dev/hd3 24576 23320 6% 33 1% /tmp
```

Here (total from **df**) - (used from **du**) + (false free block count): 24576 - 152 = 24424.

24424 is greater than 23320. The reason for this discrepancy involves the implementation of **du** and **df**. **du -s** traverses the file tree, adding up the number of blocks allocated to each directory, symlink, and file as reported by the `stat()` system call. This is how **du** arrives at its total value. **df** looks at the file system disk block allocation maps to arrive at its total and free values.

### 7.4.8 Enhanced journaled file system

The enhanced journaled file system (JFS2) contains several architectural differences over the standard JFS, including:

- ▶ Variable number of i-nodes for enhanced journaled file system

JFS2 allocates i-nodes as needed. Therefore, the number of i-nodes available is limited by the size of the file system itself.

- ▶ Specifying file system block size

File system block size is specified during the file system's creation with the **crfs** and **mkfs** commands or by using the SMIT. The decision of file system block size should be based on the projected size of files contained by the file system.
- ▶ Identifying file system block size

The file system block size value can be identified with the **lsfs** command or the System Management Interface Tool (SMIT). For application programs, the **statfs** subroutine can be used to identify the file system block size.
- ▶ Compatibility and migration

The enhanced journaled file system (JFS2) is a new file system and is not compatible with AIX Version 4.
- ▶ Device driver limitations

A device driver must provide disk block addressability that is the same or smaller than the file system block size.
- ▶ Performance costs

Although file systems that use block sizes smaller than 4096 bytes as their allocation unit might require substantially less disk space than those using the default allocation unit of 4096 bytes, the use of smaller block sizes can incur performance degradation.
- ▶ Increased allocation activity

Because disk space is allocated in smaller units for a file system with a block size other than 4096 bytes, allocation activity can occur more often when files or directories are repeatedly extended in size. For example, a write operation that extends the size of a zero-length file by 512 bytes results in the allocation of one block to the file, assuming a block size of 512 bytes. If the file size is extended further by another write of 512 bytes, an additional block must be allocated to the file. Applying this example to a file system with 4096-byte blocks, disk space allocation occurs only once, as part of the first write operation. No additional allocation activity is performed as part of the second write operation since the initial 4096-byte block allocation is large enough to hold the data added by the second write operation.
- ▶ Increased block allocation map size

More virtual memory and file system disk space might be required to hold block allocation maps for file systems with a block size smaller than 4096 bytes. Blocks serve as the basic unit of disk space allocation, and the allocation state of each block within a file system is recorded in the file system block allocation map.

- ▶ Understanding enhanced journaled file system size limitations

The maximum size for an enhanced journaled file system is architecturally limited to 4 Petabytes. I-nodes are dynamically allocated by JFS2, so you do not need to consider how many i-nodes you may need when creating a JFS2 file system. You need to consider the size of the file system log.

- ▶ Enhanced journaled file system log size issues

In most instances, multiple journaled file systems use a common log configured to be 4 MB in size. When file systems exceed 2 GB or when the total amount of file system space using a single log exceeds 2 GB, the default log size might not be sufficient. In either case, scale log sizes upward as the file system size increases. The JFS log is limited to a maximum size of 256 MB.

- ▶ JFS2 file space allocation

File space allocation is the method by which data is apportioned physical storage space in the operating system. The kernel allocates disk space to a file or directory in the form of logical blocks. A logical block refers to the division of a file or directory contents into 512, 1024, 2048, or 4096 byte units. When a JFS2 file system is created the logical block size is specified to be one of 512, 1024, 2048, or 4096 bytes. Logical blocks are not tangible entities; however, the data in a logical block consumes physical storage space on the disk. Each file or directory consists of zero or more logical blocks.

- ▶ Full and partial logical blocks

A file or directory may contain full or partial logical blocks. A full logical block contains 512, 1024, 2048, or 4096 bytes of data, depending on the file system block size specified when the JFS2 file system was created. Partial logical blocks occur when the last logical block of a file or directory contains less than the file system block size of data.

For example, a JFS2 file system with a logical block size of 4096 with a file of 8192 bytes is two logical blocks. The first 4096 bytes reside in the first logical block and the following 4096 bytes reside in the second logical block. Likewise, a file of 4608 bytes consists of two logical blocks. However, the last logical block is a partial logical block containing the last 512 bytes of the file's data. Only the last logical block of a file can be a partial logical block.

- ▶ JFS2 file space allocation

The default block size is 4096 bytes. You can specify smaller block sizes with the `mkfs` command during a file system's creation. Allowable fragment sizes are 512, 1024, 2048, and 4096 bytes. You can use only one block's size in a file system.

The kernel allocates disk space so that only the last file system block of data receives a partial block allocation. As the partial block grows beyond the limits of its current allocation, additional blocks are allocated.

Block reallocation also occurs if data is added to logical blocks that represent file holes. A file hole is an "empty" logical block located prior to the last logical block that stores data. (File holes do not occur within directories.) These empty logical blocks are not allocated blocks. However, as data is added to file holes, allocation occurs. Each logical block that was not previously allocated disk space is allocated a file system block of space.

Additional block allocation is not required if existing data in the middle of a file or directory is overwritten. The logical block containing the existing data has already been allocated file system blocks.

JFS tries to maintain contiguous allocation of a file or directory's logical blocks on the disk. Maintaining contiguous allocation lessens seek time because the data for a file or directory can be accessed sequentially and found on the same area of the disk. The disk space required for contiguous allocation may not be available if it has already been written to by another file or directory.

The file system uses a bitmap called the block allocation map to record the status of every block in the file system. When the file system needs to allocate a new block, it refers to the block allocation map to identify which blocks are available. A block can only be allocated to a single file or directory at a time.

► Extents

An extent is a sequence of contiguous file system blocks allocated to a JFS2 object as a unit. Large extents may span multiple allocation groups.

Every JFS2 object is represented by an i-node. I-nodes contain the expected object-specific information such as time stamps, file type (regular verses directory, etc.). They also contain a B+ tree to record the allocation of extents.

A file is allocated in sequences of extents. An extent is a contiguous variable-length sequence of file system blocks allocated as a unit. An extent may span multiple allocation groups. These extents are indexed in a B+ tree.

There are two values needed to define an extent, the length and the address. The length is measured in units of the file system block size. 24-bit value represents the length of an extent, so an extent can range in size from 1 to 224 -1 file system blocks. Therefore the size of the maximum extent depends on the file system block size. The address is the address of the first block of the extent. The address is also in units of file system blocks. It is the block offset from the beginning of the file system.

An extent-based file system combined with user-specified file system block size allows JFS2 to not have separate support for internal fragmentation. The user can configure the file system with a small file system block size (such as

512 bytes) to minimize internal fragmentation for file systems with large numbers of small size files.

In general, the allocation policy for JFS2 tries to maximize contiguous allocation by allowing a minimum number of extents, with each extent as large and contiguous as possible. This allows for larger I/O transfer resulting in improved performance.

► Data fragmentation

JFS2 supports fragmented file systems. Fragmentation saves disk space by allowing a logical block to be stored on the disk in units or fragments smaller than the full block size of 4096 bytes. In a fragmented file system, only the last logical block of files no larger than 32 KB is stored in this manner, so that fragmented support is only beneficial for filesystems containing numerous small files.

The use of fragments increases the potential for fragmentation of disk free space. Fragments allocated to a logical block must be contiguous on the disk. A file system experiencing free space fragmentation might have difficulty locating enough contiguous fragments for a logical block allocation, even though the total number of free fragments may exceed the logical block requirements. The JFS2 alleviates free space fragmentation by providing the **defragfs** program, which defragments a file system by increasing the amount of contiguous free space. The disk space savings gained from fragments can be substantial, while the problem of free space fragmentation remains manageable.

AIX 5L introduced a feature to set all file systems in the rootvg as JFS2-type file systems, but the server must be capable of running a 64-bit kernel, since having all JFS2 file systems in rootvg requires 64-bit kernel enablement. In AIX 5L, the 64-bit kernel is installed only on 64-bit capable pSeries servers by default, but it is inactive.

**Note:** You cannot import a JFS2 file system into an AIX Version 4.3 or earlier system.

While installing a system with the complete overwrite option, you can enable the 64-bit kernel and JFS2. If this option is enabled, the installation task will create JFS2 file systems in the rootvg.

Use the following **ls1pp** command to determine if the 64-bit kernel is installed:

```
ls1pp -h bos.mp64
Fileset Level Action Status Date Time

Path: /usr/lib/objrepos
bos.mp64
```

```
5.2.0.0 COMMIT COMPLETE 09/12/02 17:09:02
```

```
Path: /etc/objrepos
 bos.mp64
```

```
5.2.0.0 COMMIT COMPLETE 09/12/02 17:09:04
```

To query if the 64-bit kernel is enabled, enter:

```
bootinfo -K
64
```

The -K flag is not a widely documented flag. The **getconf** command in AIX 5L Version 5.2 is intended to replace the currently unsupported **bootinfo** command. To get kernel information with the **getconf** command, enter:

```
getconf KERNEL_BITMODE
64
```

The following example is one of the many ways to determine the file system type by viewing the `/etc/filesystems` file:

```
/test:
 dev = /dev/testlv
 vfs = jfs2
 log = /dev/lv04
 mount = false
 account = false
```

The `vfs` shows that it is a `jfs2` file system.

Table 7-1 provides a comparison chart between the JFS2 and the standard JFS.

*Table 7-1 Journalled file system specifications*

| Function                               | JFS2                           | JFS                                |
|----------------------------------------|--------------------------------|------------------------------------|
| Fragments/block size                   | 512–4096 block sizes           | 512-4096 fragments                 |
| Architectural maximum file             | 1 PB                           | 64 GB                              |
| Architectural maximum file system size | 4 PB                           | 1 TB                               |
| Maximum file size tested               | 1 TB                           | 64 GB                              |
| Maximum file system size tested        | 16 TB 64-bit kernel            | 1 TB                               |
| Number of nodes                        | Dynamic, limited by disk space | Fixed, set at file system creation |
| Directory organization                 | B-tree                         | Linear                             |
| Online defragmentation                 | Yes                            | Yes                                |

| Function                      | JFS2        | JFS     |
|-------------------------------|-------------|---------|
| Compression                   | No          | Yes     |
| Default ownership at creation | root.system | sys.sys |
| SGID of default file mode     | SGID=off    | SGID=on |
| Quotas                        | No          | Yes     |
| Extended ACL                  | Yes         | Yes     |

**Note:** The JFS2 file systems must have a JFS2 log device. By default, it is named jfs2log.

## 7.4.9 The /proc file system

AIX 5L provides support of the /proc file system. This pseudo file system maps processes and kernel data structures to corresponding files. The output of the **mount** and **df** commands showing /proc is provided in the following examples:

```
mount
node mounted mounted over vfs date options

/dev/hd4 /
/dev/hd2 /usr
/dev/hd9var /var
/dev/hd3 /tmp
/dev/hd1 /home
/proc /proc
procfs Sep 11 16:53 rw
```

```
df
Filesystem 512-blocks Free %Used Iused %Iused Mounted on
/dev/hd4 65536 27760 58% 2239 14% /
/dev/hd2 1507328 242872 84% 22437 12% /usr
/dev/hd9var 32768 16432 50% 448 11% /var
/dev/hd3 557056 538008 4% 103 1% /tmp
/dev/hd1 32768 31608 4% 47 2% /home
/proc - - - - - /proc
```

The entry in the /etc/vfs file appears as follows:

```
lsvfs procfs
procfs 6 none none
```

Each process is assigned a directory entry in the /proc file system with a name identical to its process ID. In this directory, several files and subdirectories are created corresponding to internal process control data structures. Most of these

files are read-only, but some of them can also be written to and be used for process control purposes.

The ownership of the files in the /proc file system is the same as for the processes they represent. Therefore, regular users can only access /proc files that belong to their own processes. For example, a process is waiting for standard input. When you look at an active process, a lot of the information would constantly change:

```
errpt -a|more
.....
.....
ps -ef|grep errpt
 root 12586 17594 1 16:53:00 pts/3 0:00 grep errpt
 root 22214 8882 0 16:49:39 pts/1 0:00 errpt -a
ls -al /proc/22214
total 0
dr-xr-xr-x 1 root system 0 Sep 17 16:56 .
dr-xr-xr-x 1 root system 0 Sep 17 16:56 ..
-rw----- 1 root system 0 Sep 17 16:56 as
-r----- 1 root system 128 Sep 17 16:56 cred
--w----- 1 root system 0 Sep 17 16:56 ctl
dr-xr-xr-x 1 root system 0 Sep 17 16:56 lwp
-r----- 1 root system 0 Sep 17 16:56 map
dr-x----- 1 root system 0 Sep 17 16:56 object
-r--r--r-- 1 root system 448 Sep 17 16:56 psinfo
-r----- 1 root system 1024 Sep 17 16:56 sigact
-r----- 1 root system 1520 Sep 17 16:56 status
-r--r--r-- 1 root system 0 Sep 17 16:56 sysent
```

Table 7-2 provides the functions of the pseudo files listed in the previous output.

Table 7-2 Functions of pseudo files in /proc/<pid> directory

| Pseudo file name | Function                                                     |
|------------------|--------------------------------------------------------------|
| as               | Read/write access to address space                           |
| cred             | Credentials                                                  |
| ctl              | Write access to control process, for example, stop or resume |
| lwp directory    | Kernel thread information                                    |
| map              | Virtual address map                                          |
| object directory | Map file names                                               |
| psinfo           | Information for the ps command; readable by everyone         |
| sigact           | Signal status                                                |

| Pseudo file name | Function                                                          |
|------------------|-------------------------------------------------------------------|
| status           | Process state information, such as address, size of heap or stack |
| sysent           | Information about system calls                                    |

### 7.4.10 Disk quota

The disk quota system allows you to control the number of files and data blocks that can be allocated to users or groups. The quota system can be defined for individual users or groups, and is maintained for each journaled file system.

The disk quota system establishes limits based on three parameters that can be changed with the **edquota** command:

- ▶ User's or group's soft limits
- ▶ User's or group's hard limits
- ▶ Quota grace period

The soft limit defines the number of 1-KB disk blocks or files below which the user must remain. The hard limit defines the maximum amount of disk blocks or files the user can accumulate under the established disk quotas. The quota grace period allows the user to exceed the soft limit for a short period of time (the default value is one week).

The disk quota system tracks user and group quotas in the `quota.user` and `quota.group` files that reside in the root directories of file systems enabled with quotas.

You should consider implementing the disk quota system under the following conditions:

- ▶ Your system has limited disk space.
- ▶ You require more file system security.
- ▶ Your disk-usage levels are large, such as at many universities.

Typically, only those file systems that contain user home directories and files require disk quotas. The disk quota system works only with the journaled file system.

**Note:** Because many editors and system utilities create temporary files in the `/tmp` file system, it must be free of quotas.

## Setting up the disk quota

In the following example, a quota system is being set up for the /home file system.

1. To verify that you have the bos.sysmgt.quota fileset installed, enter:

```
ls1pp -l |grep bos.sysmgt.quota
bos.sysmgt.quota 5.1.0.25 COMMITTED Filesystem Quota Commands
```

1. Determine which file system requires quotas. In this example, it is for /home file system.
2. Use the **chfs** command to include the userquota and groupquota quota configuration attributes in the /etc/filesystems file. To enable both user and group quotas on the /home file system, enter:

```
chfs -a "quota = userquota,groupquota" /home
```

The corresponding entry in the /etc/filesystems is displayed as follows:

```
/home:
dev = /dev/hd1
vfs = jfs
log = /dev/hd8
mount = true
options = rw
account = false
quota = userquota,groupquota
```

3. Mount the specified file systems, if not previously mounted.
4. Set the desired quota limits for each user or group. Use the **edquota** command to create each user's or group's soft and hard limits for allowable disk space and maximum number of files. To set the quota for user quotausr, enter:

```
edquota -u quotausr
```

The **edquota** command also invokes the vi editor (or the editor specified by the EDITOR environment variable) on the temporary file so that quotas can be added and modified.

```
Quotas for user quotausr:
/home: blocks in use: 0, limits (soft = 100, hard = 150)
 inodes in use: 0, limits (soft = 200, hard = 250)
~
~
```

To set the quota for the entire group named quotagrp, enter:

```
edquota -g quotagrp
```

Modify the temporary file to set the limits:

```
Quotas for group quotagrp:
```

```
/home: blocks in use: 0, limits (soft = 100, hard = 250)
 inodes in use: 0, limits (soft = 200, hard = 250)
~
~
```

5. To enable the quota system with the **quotaon** command, enter:  

```
quotaon -u /home
```
6. Use the **quotacheck** command to check the consistency of the quota files against actual disk usage:  

```
quotacheck /home
```

## 7.5 Paging space

On AIX systems, the following list indicates possible problems associated with paging space:

- ▶ All paging spaces defined on one physical volume
- ▶ Page space nearly full
- ▶ Imbalance in allocation of paging space on physical volumes
- ▶ Fragmentation of a paging space in a volume group

### 7.5.1 Recommendations for creating or enlarging paging space

Do not put more than one paging space logical volume on a physical volume.

All processes started during the boot process are allocated paging space on the default paging space logical volume (hd6). After the additional paging space logical volumes are activated, paging space is allocated in a round robin manner in 4-KB units. If you have paging space on multiple physical volumes and put more than one paging space on one physical volume, you are no longer spreading paging activity over multiple physical volumes.

Avoid putting a paging space logical volume on the same physical volume as a heavily active logical volume, such as that used by a database.

It is not necessary to put a paging space logical volume on each physical volume.

Make each paging space logical volume roughly equal in size. If you have paging spaces of different sizes and the smaller ones become full, you will no longer be spreading your paging activity across all of the physical volumes.

Do not extend a paging space logical volume into multiple physical volumes. If a paging space logical volume is spread over multiple physical volumes, you will not be spreading paging activity across all the physical volumes. If you want to allocate space for paging on a physical volume that does not already have a paging space logical volume, create a new paging space logical volume on that physical volume.

For best system performance, put paging space logical volumes on physical volumes that are each attached to a different disk controller.

## 7.5.2 Determining if more paging space is needed

Allocating more paging space than necessary results in unused paging space that is simply wasted disk space. If you allocate too little paging space, a variety of negative symptoms may occur on your system. To determine how much paging space is needed, use the following guidelines:

- ▶ Enlarge paging space if any of the following messages appear on the console or in response to a command on any terminal:

```
INIT: Paging space is low
ksh: cannot fork no swap space
Not enough memory
Fork function failed
fork () system call failed
Unable to fork, too many processes
Fork failure - not enough memory available
Fork function not allowed. Not enough memory available.
Cannot fork: Not enough space
```

- ▶ Enlarge paging space if the %Used column of the `lspvs -s` output is greater than 80.

Use the `iostat`, `vmstat`, and `lspvs` commands to determine if you need to make changes regarding paging space logical volumes:

```
iostat
vmstat
lspvs
```

If you wish to remove a paging space from the system or reduce the size of a paging space, this should be performed in two steps. The first step in either case is to change the paging space so that it is no longer automatically used when the system starts. This is done with the `chpvs` command. For example:

```
chpvs -a n paging00
```

### 7.5.3 Reducing and removing paging space

In AIX Version 4.x, to remove a paging space you need to reboot the system, since there is no way to dynamically bring a paging space offline. Once the system reboots, the paging space will not be active. At this point, you can remove the paging space logical volume.

In AIX Version 4.x, to reduce the size of the paging space, you should remove the logical volume, and then create the new paging space with the desired size. The new paging space can be activated without having to reboot the machine using the **mkps** command.

Removing paging space can be done by using the following procedure involving the **chps** and the **rmps** commands.

**Note:** Removing default paging spaces incorrectly can prevent the system from restarting.

The paging space must be deactivated before it can be removed. A special procedure is required for removing the default paging spaces (hd6, hd61, and so on). These paging spaces are activated during boot time by shell scripts that configure the system. To remove one of the default paging spaces, these scripts must be altered and a new boot image must be created.

This scenario describes how to remove an existing paging space, paging00, from the system. This disk layout is as follows:

```
lsps -a
Page Space Physical Volume Volume Group Size %Used Active Auto Typ
paging00 hdisk2 testvg 3200MB 1 yes yes lv
hd6 hdisk0 rootvg 1040MB 1 yes yes lv
```

1. The paging00 paging space is automatically activated. Use the **chps** command to change its state:

```
chps -a n paging00
```

2. The paging space is in use; a reboot of the system is required. Make sure that the system dump device is still pointing to a valid paging space, as follows:

```
sysdumpdev -l
primary /dev/hd6
secondary /dev/sysdumpnull
copy directory /var/adm/ras
forced copy flag TRUE
always allow dump FALSE
dump compression OFF
```

3. Remove the paging00 paging space using the **rmpps** command:

```
rmpps paging00
rmlv: Logical volume paging00 is removed.
```

If the paging space you are removing is the default dump device, you must change the default dump device to another paging space or logical volume before removing the paging space. To change the default dump device, use the **sysdumpdev -P -p /dev/new\_dump\_device** command.

In AIX 5L, you can use the **swapoff** command to move all pages out of a paging space that is to be removed. After pages are moved out, use the **rmpps** command to remove the paging space. In the following example, paging00 is removed:

```
swapoff /dev/paging00
rmpps paging00
rmlv: Logical volume paging00 is removed.
```

With AIX 5L, managing paging space also became much simpler. You can reduce the size of paging space dynamically. For example, remove one logical partition from paging00 by running:

```
chps -d 1 paging00
shrinkps: Temporary paging space paging01 created.
shrinkps: Paging space paging00 removed.
shrinkps: Paging space paging00 recreated with new size.
```

Rebooting of the server is not needed.

The **chps** command automatically creates the temporary paging spaces as needed in AIX 5L.

## 7.6 Command summary

The following sections review commands discussed in this chapter and the flags most often used.

### 7.6.1 The **lsvg** command

The **lsvg** command sets the characteristics of a volume group. The command has the following syntax:

```
lsvg [-L] [-o] | [-n DescriptorPhysicalVolume] | [-i] [-l | -M | -p]
VolumeGroup ...
```

The most commonly used flag is provided in Table 7-3 on page 170.

Table 7-3 Commonly used flag of the lsvg command

| Flag | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -l   | <p>Lists the following information for each logical volume within the group specified by the VolumeGroup parameter:</p> <p>LV - A logical volume within the volume group.<br/>           Type - Logical volume type.<br/>           LPs - Number of logical partitions in the logical volume.<br/>           PPs - Number of physical partitions used by the logical volume.<br/>           PVs - Number of physical volumes used by the logical volume.<br/>           Logical volume state - State of the logical volume. Opened/stale indicates the logical volume is open but contains partitions that are not current. Opened/syncd indicates the logical volume is open and synchronized. Closed indicates the logical volume has not been opened.</p> |

## 7.6.2 The chvg command

The **chvg** command sets the characteristics of a volume group. The command has the following syntax:

```
chvg [-a AutoOn { n | y }] [-c | -l] [-Q { n | y }] [-u]
[-x { n | y }] [-t [factor]] [-B] VolumeGroup
```

The most commonly used flag is provided in Table 7-4 on page 171.

Table 7-4 Commonly used flag of the chvg command

| Flag               | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -t <i>[factor]</i> | <p>Changes the limit of the number of physical partitions per physical volume, specified by factor. The factor should be between one and 16 for 32-disk volume groups, and one and 64 for 128-disk volume groups.</p> <p>If factor is not supplied, it is set to the lowest value such that the number of physical partitions of the largest disk in volume group is less than factor x 1016.</p> <p>If factor is specified, the maximum number of physical partitions per physical volume for this volume group changes to factor x 1016.</p> <p>Notes:</p> <ul style="list-style-type: none"> <li>▶ If the volume group is created in AIX Version 3.2 and 4.1.2 in violation of 1016 physical partitions per physical volume limit, this flag can be used to convert the volume group to a supported state. This will ensure proper stale/fresh marking of partitions.</li> <li>▶ The factor cannot be changed if there are any stale physical partitions in the volume group.</li> <li>▶ Once a volume group is converted, it cannot be imported into AIX Version 4.3 or earlier versions.</li> <li>▶ This flag cannot be used if the volume group is varied on in concurrent mode.</li> </ul> |

### 7.6.3 The importvg command

The **importvg** command imports a new volume group definition from a set of physical volumes. The command has the following syntax:

```
importvg [-V MajorNumber] [-y VolumeGroup] [-f] [-c] [-x] |
[-L VolumeGroup] [-n] [-F] [-R]PhysicalVolume
```

The most commonly used flags are provided in Table 7-5.

Table 7-5 Commonly used flags of the importvg command

| Flag                     | Description                                                                                                                                                                                                                                                                                                                                                                     |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -y<br><i>VolumeGroup</i> | <p>Specifies the name to use for the new volume group. If this flag is not used, the system automatically generates a new name.</p> <p>The volume group name can only contain the following characters: "A" through "Z," "a" through "z," "0" through "9," or "_" (the underscore), "-" (the minus sign), or "." (the period). All other characters are considered invalid.</p> |
| -f                       | Forces the volume group to be varied online.                                                                                                                                                                                                                                                                                                                                    |

## 7.6.4 The rmlvcopy command

The **rmlvcopy** command removes copies from each logical partition in the logical volume. The command has the following syntax:

```
rmlvcopy LogicalVolume Copies [PhysicalVolume ...]
```

## 7.6.5 The reducevg command

The **reducevg** command removes physical volumes from a volume group. When all physical volumes are removed from the volume group, the volume group is deleted. The command has the following syntax:

```
reducevg [-d] [-f] VolumeGroup PhysicalVolume ...
```

The most commonly used flags are provided in Table 7-6.

Table 7-6 Commonly used flags of the reducevg command

| Flag | Description                                                                                                                                                                                         |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -d   | Deallocates the existing logical volume partitions and then deletes resultant empty logical volumes from the specified physical volumes. User confirmation is required unless the -f flag is added. |
| -f   | Removes the requirement for user confirmation when the -d flag is used.                                                                                                                             |

## 7.6.6 The rmdev command

The **rmdev** command removes a device from the system. The command has the following syntax:

```
rmdev -l Name [-d | -S] [-f File] [-h] [-q] [-R]
```

The most commonly used flags are provided in Table 7-7.

Table 7-7 Commonly used flags of the rmdev command

| Flag    | Description                                                                                                        |
|---------|--------------------------------------------------------------------------------------------------------------------|
| -l Name | Specifies the logical device, indicated by the name variable, in the Customized Devices object class.              |
| -d      | Removes the device definition from the Customized Devices object class. This flag cannot be used with the -S flag. |

## 7.6.7 The syncvg command

The **syncvg** command synchronizes logical volume copies that are not current. The command has the following syntax:

```
syncvg [-f] [-i] [-H] [-P NumParallelLps] { -l | -p | -v } Name ...
```

The most commonly used flag is provided in Table 7-8.

Table 7-8 Commonly used flag of the syncvg command

| Flag | Description                                                                |
|------|----------------------------------------------------------------------------|
| -p   | Specifies that the name parameter represents a physical volume device name |

## 7.7 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

1. A file system shows the following output from the **df** command:

```
Filesystem 1024-blocks Free %Used Iused %Iused Mounted on
/dev/hd3 57344 47888 17% 184 2% /tmp
```

However, a long listing of the file `/tmp/myfile` shows the following:

```
2066 -rwxrwxrwx 1 lucinda staff 100000000 May 07 14:41 /tmp/myfile
```

Which of the following file types is `/tmp/myfile`?

- A. Dense
  - B. Sparse
  - C. Fragmented
  - D. Compressed
2. A volume group `vg00` has two 4.5-GB disk drives (`hdisk4` and `hdisk5`). When a 9.1-GB disk (`hdisk6`) was added, the following error occurred:

```
0516-1162 extendvg: Warning, The Physical Partition Size of 16 requires the
creation of 3258 partitions for hdisk6. The limitation for volume group
rootvg is 1016 physical parititons per physical volume.
```

```
0516-792 extendvg: Unable to extend volume group to the rootvg
```

Which of the following commands should be used to successfully add the 9.1-GB drive to vg00?

- A. **chvg -t -2 vg00**
  - B. **sync1vodm -v vg00**
  - C. **lqueryvg -Atp hdisk6**
  - D. **redefinevg -d hdisk6 vg00**
3. Which of the following should be avoided with regards to paging space?
- A. Placing paging space on non-rootvg disk
  - B. Placing paging spaces on non-SCSI drives
  - C. Using paging space less than the RAM size
  - D. Using multiple paging spaces on the same drive
4. During a routine check of an error log, the system is receiving SSA errors. The error log reports an SSA link open failure error. The system is working fine, but the SSA errors keep repeating every 30 minutes. All of the following are steps to fix the error *except*:
- A. Run **cfgmgr -v**.
  - B. Check the cabling.
  - C. Run link verification.
  - D. Check the microcode for the SSA drive.
5. Which of the following statements best describes the relationship between a logical volume and a journaled file system?
- A. Increasing the size of a logical volume also increases the size of the file system.
  - B. Increasing the size of a file system requires file systems to be unmounted first.
  - C. Increasing the size of a file system also increases the size of the logical volume when necessary.
  - D. Reducing the size of a logical volume requires reducing the size of the file system first.

6. Which of the following statements is *false* regarding the process for administering file systems?
- A. A logical volume must be created prior to defining a file system.
  - B. A file system must be empty before it can be removed.
  - C. Defining a file system imposes a structure on a logical volume.
  - D. Removing a file system using `rmfs` automatically removes the underlying logical volume.

## 7.7.1 Answers

The following are the preferred answers to the questions provided in this section:

1. B
2. A
3. D
4. A
5. C
6. B

## 7.8 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

1. Verify the maximum number of PPs on your system, using rootvg, as an example. What is the maximum disk size that can be added to your system?
2. If you have access to a test system that is equipped with a hot-swappable SCSI disk, try the disk replacement example in 7.3.1, “Replacing a disk” on page 144.



# Network problem determination

The following topics are discussed in this chapter:

- ▶ Network interface problems
- ▶ Routing problems
- ▶ Name resolution problems
- ▶ NFS troubleshooting

This chapter discusses network problem source identification and resolution.

## 8.1 Network interface problems

If host name resolution does not work and you cannot ping any address in the routing table, the interface itself may be the culprit. The first step to determine if this is true should be to check the installed adapter types and states using the **lsdev -Cc adapter** and **lsdev -Cc if** commands, as shown in the following:

```
lsdev -Cc adapter
pmc0 Available 01-A0 Power Management Controller
fda0 Available 01-C0 Standard I/O Diskette Adapter
ide0 Available 01-E0 ATA/IDE Controller Device
ide1 Available 01-F0 ATA/IDE Controller Device
....
ppa0 Available 01-D0 Standard I/O Parallel Port Adapter
ent0 Available 04-D0 IBM PCI Ethernet Adapter (22100020)
tok0 Available 04-01 IBM PCI Tokenring Adapter (14101800)
lsdev -Cc if
en0 Available Standard Ethernet Network Interface
et0 Defined IEEE 802.3 Ethernet Network Interface
lo0 Available Loopback Network Interface
tr0 Available Token Ring Network Interface
```

As shown, there are two network adapters and four network interfaces. All interfaces can be administrated by either the **chdev** or the **ifconfig** command.

To determine the state of the interface, use the **ifconfig** command. The following examples show the en0 interface in the up, down, and detach state.

The en0 interface in the up state is shown in the following:

```
ifconfig en0
en0:
flags=e080863<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT>
inet 10.47.1.1 netmask 0xffff0000 broadcast 10.47.255.255
```

The down state of the interface keeps the system from trying to transmit messages through that interface. Routes that use the interface are not automatically disabled.

```
ifconfig en0 down
ifconfig en0
en0:
flags=e080862<BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT>
inet 10.47.1.1 netmask 0xffff0000 broadcast 10.47.255.255
```

The interface in the detach state is removed from the network interface list. If the last interface is detached, the network interface driver code is unloaded.

```
ifconfig en0 detach
```

```
ifconfig en0
en0: flags=e080822<BROADCAST,NOTRAILERS,SIMPLEX,MULTICAST,GROUPRT,64BIT>
```

All changes made to the network interface as shown can also be made by the **chdev** command. Changes made by this command are permanent because they are made directly to the ODM database. To list the parameters of the network interface tr0 that you can change by the **chdev** command, enter:

```
lsattr -El tr0
mtu 1492 Maximum IP Packet Size for This Device True
mtu_4 1492 Maximum IP Packet Size for This Device True
mtu_16 1492 Maximum IP Packet Size for This Device True
mtu_100 1492 Maximum IP Packet Size for This Device True
remmtu 576 Maximum IP Packet Size for REMOTE Networks True
netaddr 9.3.240.59 Internet Address True
state up Current Interface Status True
arp on Address Resolution Protocol (ARP) True
allcast on Confine Broadcast to Local Token-Ring True
hwloop off Enable Hardware Loopback Mode True
netmask 255.255.255.0 Subnet Mask True
security none Security Level True
authority Authorized Users True
broadcast Broadcast Address True
netaddr6 N/A True
alias6 N/A True
prefixlen N/A True
alias4 N/A True
rfc1323 N/A True
tcp_nodelay N/A True
tcp_sendspace N/A True
tcp_recvspace N/A True
tcp_msdfilt N/A True
```

For example, to set up the broadcast address for the tr0 interface, enter:

```
chdev -l tr0 -a broadcast='9.3.240.255'
tr0 changed
```

To check the new value of the broadcast parameter, enter:

```
lsattr -El tr0 -a broadcast
broadcast 9.3.240.255 Broadcast Address True
```

When you have a network performance problem and you suspect that the network interface could be the cause of it, you should check the interface statistics. To display the statistics for the en0 interface, enter:

```
netstat -I en0
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en0 1500 link#2 8.0.5a.fc.d2.e1 28982 0 579545 0 0
en0 1500 10.47 server4_ 28982 0 579545 0 0
```

As you can see, the output shows the number of input and output errors and the number of input and output packets.

**Note:** The collision count for Ethernet interfaces is not displayed by the **netstat** command. It always shows 0.

To see more detailed statistics, use the **entstat** command:

```
entstat -d en0

ETHERNET STATISTICS (en0) :
Device Type: IBM PCI Ethernet Adapter (22100020)
Hardware Address: 08:00:5a:fc:d2:e1
Elapsed Time: 1 days 1 hours 21 minutes 29 seconds

Transmit Statistics: Receive Statistics:

Packets: 579687 Packets: 55872
Bytes: 49852606 Bytes: 4779893
Interrupts: 0 Interrupts: 55028
Transmit Errors: 0 Receive Errors: 0
Packets Dropped: 0 Packets Dropped: 0
Bad Packets: 0

Max Packets on S/W Transmit Queue: 2
S/W Transmit Queue Overflow: 0
Current S/W+H/W Transmit Queue Length: 0

Broadcast Packets: 2327 Broadcast Packets: 0
Multicast Packets: 0 Multicast Packets: 0
No Carrier Sense: 0 CRC Errors: 0
DMA Underrun: 0 DMA Overrun: 0
Lost CTS Errors: 0 Alignment Errors: 0
Max Collision Errors: 0 No Resource Errors: 0
Late Collision Errors: 0 Receive Collision Errors: 0
Deferred: 34 Packet Too Short Errors: 0
SQE Test: 0 Packet Too Long Errors: 0
Timeout Errors: 0 Packets Discarded by Adapter: 0
Single Collision Count: 4 Receiver Start Count: 0
Multiple Collision Count: 12
Current HW Transmit Queue Length: 0

General Statistics:

No mbuf Errors: 0
Adapter Reset Count: 4
Driver Flags: Up Broadcast Running
 Simplex AlternateAddress 64BitSupport
```

## IBM PCI Ethernet Adapter Specific Statistics:

```

Chip Version: 16
Packets with Transmit collisions:
 1 collisions: 4 6 collisions: 2 11 collisions: 0
 2 collisions: 1 7 collisions: 1 12 collisions: 0
 3 collisions: 4 8 collisions: 1 13 collisions: 0
 4 collisions: 1 9 collisions: 0 14 collisions: 0
 5 collisions: 1 10 collisions: 1 15 collisions: 0
```

To test for dropped packets, use the **ping** command with the **-f** flag. The **-f** flag *floods*, or outputs, packets as fast as they come back or one hundred times per second, whichever is more. For every ECHO\_REQUEST sent, a period is printed, while for every ECHO\_REPLY received, a backspace is printed. This provides a rapid display of how many packets are being dropped. Only the root user can use this option.

## 8.2 Routing problems

If you are not able to ping by host name or IP address, you may have a routing problem.

First, check the routing tables as follows:

1. Use the **netstat -nr** command to show the content of your local routing table using IP addresses.

```
netstat -nr
Routing tables
Destination Gateway Flags Refs Use If PMTU Exp
Groups
```

Route Tree for Protocol Family 2 (Internet):

```
default 9.3.240.1 UGc 0 0 tr0 - -
9.3.240/24 9.3.240.58 U 31 142091 tr0 - -
10.47.1.2 9.3.240.59 UGH 0 2 tr0 - -
127/8 127.0.0.1 UR 0 3 lo0 - -
127.0.0.1 127.0.0.1 UH 3 761 lo0 - -
195.116.119/24 195.116.119.2 U 2 406 en0 - -
```

Route Tree for Protocol Family 24 (Internet v6):

```
::1 ::1 UH 0 0 lo0 16896 -
```

2. Check the netmask displayed and ensure that it is correct (ask the network administrator what it should be if you are unsure).

```
lsattr -El tr0 -a netmask -F value
255.255.255.0
```

3. If there is a default route, attempt to ping it.

```
ping 9.3.240.1
PING 9.3.240.1: (9.3.240.1): 56 data bytes
64 bytes from 9.3.240.1: icmp_seq=0 ttl=64 time=1 ms
64 bytes from 9.3.240.1: icmp_seq=1 ttl=64 time=0 ms
^C
----9.3.240.1 PING Statistics----
2 packets transmitted, 2 packets received, 0% packet loss
round-trip min/avg/max = 0/0/1 ms
```

The **ping -c** command will limit the number of pings to the value specified after the **-c** flag.

```
ping -c 1 server2
PING server2.itsc.austin.ibm.com: (9.3.240.57): 56 data bytes
64 bytes from 9.3.240.57: icmp_seq=0 ttl=255 time=0 ms
```

```
----server2.itsc.austin.ibm.com PING Statistics----
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0/0/0 ms
```

```
export NSORDER=local,bind,nis
```

```
ping -c 1 server2
```

```
PING server2: (9.3.240.57): 56 data bytes
64 bytes from 9.3.240.57: icmp_seq=0 ttl=255 time=0 ms
```

```
----server2 PING Statistics----
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0/0/0 ms
```

4. If you have more than one network interface, attempt to determine if any interfaces are working.

If you cannot ping your default route, either the default gateway is down, or your local network connection may be down. Attempt to ping all of the gateways listed in the routing table to see if any portion of your network is functioning.

If you cannot ping any host or router interface from among those listed in the routing table, try to ping your loopback interface lo0 with the following command:

```
ping localhost
PING localhost: (127.0.0.1): 56 data bytes
64 bytes from 127.0.0.1: icmp_seq=0 ttl=255 time=1 ms
^C
----localhost PING Statistics----
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 1/1/1 ms
```

If the ping is successful, you have an adapter or network hardware problem or a routing problem. The **ping -f** (flood ping) command outputs packets as fast as they come back or one hundred times per second, whichever is more. The

command **ping -f** (flood ping) requires root access to the system. For every ECHO\_REQUEST sent, a period '.' is printed, while for every ECHO\_REPLY received, a backspace is printed. This provides a rapid display of how many packets are being dropped.

If the ping is not successful, you need to:

1. Ensure that the inetd process is actively using the **lssrc -g tcpip** command. If inetd is not active, issue the **startsrc -s inetd** or **startsrc -g tcpip** command.

```
lssrc -g tcpip
Subsystem Group PID Status
routed tcpip 5424 active
inetd tcpip 6192 active
snmpd tcpip 6450 active
gated tcpip inoperative
named tcpip inoperative
----- the output was edited for brevity -----
```

2. Check the state of the loopback interface (lo0) with the **netstat -i** command. If you see lo0\* in the output, check the /etc/hosts file for an uncommented local loopback entry, as follows:

```
netstat -I lo0 -n
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
lo0* 16896 link#1 412934 0 414344 0 0
lo0* 16896 127 127.0.0.1 412934 0 414344 0 0
lo0* 16896 ::1 412934 0 414344 0 0
grep localhost /etc/hosts
127.0.0.1 localhost localhost # loopback (lo0) name/address
```

A splat (\*) after the interface name in the output from the **netstat** command indicates that the interface is down. Use the following command to start the lo0 interface:

```
ifconfig lo0 inet 127.0.0.1 up
```

If you cannot reach a host that is in a different network, you can check the connection using the **traceroute** command. The **traceroute** command output shows each gateway that the packet traverses on its way to finding the target host. If possible, examine the routing tables of the last machine shown in the **traceroute** output to check if a route exists to the destination from that host. The last machine shown is where the routing is incorrectly set.

```
traceroute 9.3.240.56
traceroute to 9.3.240.56 (9.3.240.56), 30 hops max, 40 byte packets
 1 server4e (10.47.1.1) 1 ms 1 ms 0 ms
 2 server1 (9.3.240.56) 1 ms 1 ms 1 ms
```

If you are using the **route** command to change the routing table on your machine and you want this change to be permanent, insert the appropriate line in the `/etc/rc.net` file.

## 8.2.1 Dynamic or static routing

If you have a problem with the dynamic routing protocol, follow the procedure provided in this section.

If your system is set up to use the routed daemon:

1. Check if the routed daemon is running; if not, start it with the **startsrc -s routed** command.
2. If routed cannot identify the route through queries, check the `/etc/gateways` file to verify that a route to the target host is defined and that the target host is running the RIP.
3. Make sure that gateways responsible for forwarding packets to the host are up and that they are running the RIP (routed or gated active). Otherwise you will need to define a static route.
4. Run the routed daemon with the debug option to log information such as bad packets received. Invoke the daemon from the command line using the following command:

```
startsrc -s routed -a "-d"
```

5. Run the routed daemon using the `-t` flag, which causes all packets sent or received to be written to standard output. When routed is run in this mode, it remains under the control of the terminal that started it. Therefore, an interrupt from the controlling terminal kills the daemon.

If your system is set up to use the gated daemon:

1. Check if gated is running; if not, start it with the **startsrc -s gated** command.
2. Verify that the `/etc/gated.conf` file is configured correctly and that you are running the correct routing protocols.
3. Make sure that the gateway on the source network is using the same protocol as the gateway on the destination network.
4. Make sure that the machine that you are trying to communicate with has a return route back to your host machine.

You should set static routes under the following conditions:

- ▶ The destination host is not running the same protocol as the source host, so it cannot exchange routing information.

- ▶ The host must be reached by a distant gateway (a gateway that is on a different autonomous system than the source host). The RIP can be used only among hosts on the same autonomous system.

If you are using dynamic routing, you should not attempt to add static routes to the routing table using the **route** command.

As a very last resort, you may flush the routing table using the **route -f** command, which will cause all the routes to be removed and eventually replaced by the routing daemons. Since any networking that was functioning before will be temporarily cut off once the routes are removed, be sure no other users will be affected by this.

If your system is going to be configured as a router (it has two or more network interfaces), then it needs to be enabled as a router by the **no** command. The network option that controls routing from one network to another is ipforwarding and by default is disabled. To enable it, enter:

```
no -o ipforwarding=1
```

This is not a permanent setting and after the next system reboot it will be lost. To make this permanent, add this command to the end of the `/etc/rc.net` file.

**Note:** When you add the second network interface to your system, a new entry will appear in the routing table. This is a route associated with the new interface.

## 8.3 Name resolution problems

If network connections seem inexplicably slow sometimes but fast at other times, it is a good idea to check the name resolution configuration for your system. Do a basic diagnostic for name resolving. You can use either the **host** command or the **nslookup** command.

```
host dhcp240.itsc.austin.ibm.com
dhcp240.itsc.austin.ibm.com is 9.3.240.2
```

The name resolution can be served through either a remote DNS server or a remote NIS server. If one of them is down, you may have to wait until TCP timeout occurs. The name can be resolved by an alternate source, which can be a secondary name server or the local `/etc/hosts` file.

First check the `/etc/netsvc.conf` file and the NSORDER environment variable for your particular name resolution ordering. The NSORDER variable overrides the hosts settings in the `/etc/netsvc.conf` file. Check the `/etc/resolv.conf` file for the IP

address of the named server and try to ping it. If you can, then it is reachable. If not, try different name resolution ordering.

**Note:** When you can ping the name server, it does not mean that the named daemon is active on this system.

By default, resolver routines attempt to resolve names using BIND and DNS. If the `/etc/resolv.conf` file does not exist, or if BIND or DNS could not find the name, NIS is queried (if it is running). NIS is authoritative over the local `/etc/hosts`, so the search will end here if it is running. If NIS is not running, then the local `/etc/hosts` file is searched. If none of these services could find the name, then the resolver routines return with `HOST_NOT_FOUND`. If all of the services are unavailable, then the resolver routines return with `SERVICE_UNAVAILABLE`.

If you want to change the name resolution ordering so that NIS takes precedence over the BIND and DNS, your `/etc/netsvc.conf` file should look like the following example:

```
cat /etc/netsvc.conf
hosts = nis,bind
```

You can override this setting by using the `NSORDER` environment variable:

```
export NSORDER=local,bind
```

In this situation the `/etc/hosts` file will be examined for name resolution first.

### 8.3.1 The `tcpdump` and `iptrace` commands

You may need to see the real data *crossing the wire* to solve a problem. There are two commands that let you see every incoming and outgoing packet from your interface: `tcpdump` and `iptrace`.

The `tcpdump` command prints out the headers of packets captured on a specified network interface. The following example shows a telnet session between hosts 9.3.240.59 and 9.3.240.58:

```
tcpdump -i tr0 -n -I -t dst host 9.3.240.58
9.3.240.59.44183 > 9.3.240.58.23: S 1589597023:1589597023(0) win 16384 <mss
1452> [tos 0x10]
9.3.240.58.23 > 9.3.240.59.44183: S 1272672076:1272672076(0) ack 1589597024 win
15972 <mss 1452>
9.3.240.59.44183 > 9.3.240.58.23: . ack 1 win 15972 [tos 0x10]
9.3.240.59.44183 > 9.3.240.58.23: . ack 1 win 15972 [tos 0x10]
9.3.240.59.44183 > 9.3.240.58.23: P 1:16(15) ack 1 win 15972 [tos 0x10]
9.3.240.59.44183 > 9.3.240.58.23: P 1:16(15) ack 1 win 15972 [tos 0x10]
9.3.240.59.44183 > 9.3.240.58.23: . ack 6 win 15972 [tos 0x10]
```

```

9.3.240.59.44183 > 9.3.240.58.23: . ack 6 win 15972 [tos 0x10]
9.3.240.58.23 > 9.3.240.59.44183: P 6:27(21) ack 1 win 15972 (DF)
9.3.240.59.44183 > 9.3.240.58.23: P 1:27(26) ack 27 win 15972 [tos 0x10]
9.3.240.59.44183 > 9.3.240.58.23: P 1:27(26) ack 27 win 15972 [tos 0x10]
9.3.240.58.23 > 9.3.240.59.44183: P 27:81(54) ack 27 win 15972 (DF)
9.3.240.59.44183 > 9.3.240.58.23: P 27:30(3) ack 81 win 15972 [tos 0x10]
9.3.240.59.44183 > 9.3.240.58.23: P 27:30(3) ack 81 win 15972 [tos 0x10]

```

The first line indicates that TCP port 44183 on host 9.3.240.59 sent a packet to the telnet port (23) on host 9.3.240.58. The S indicates that the SYN flag was set. The packet sequence number was 1589597023 and it contained no data. There was no piggy-backed ack field, the available receive field win was 16384 bytes, and there was a max-segment-size (mss) option requesting a mss of 1452 bytes. Host 9.3.240.58 replies with a similar packet, except it includes a piggy-backed ack field for host 9.3.240.59 SYN. Host 9.3.240.59 then acknowledges the host 9.3.240.58 SYN. The period (.) means no flags were set. The packet contains no data, so there is no data sequence number. On the eleventh line, host 9.3.240.59 sends host 9.3.240.58 26 bytes of data. The PUSH flag is set in the packet. On the twelfth line, host 9.3.240.58 says it received data sent by host 9.3.240.59 and sends 54 bytes of data; it also includes a piggy-backed ack for sequence number 27.

The `iptrace` daemon records IP packets received from configured interfaces. Command flags provide a filter so that the daemon traces only packets meeting specific criteria. Packets are traced only between the local host on which the `iptrace` daemon is invoked and the remote host. To format `iptrace` output, run the `ipreport` command. The following example shows the query from host 9.3.240.59 to DNS server 9.3.240.2. The output from the `nslookup` command is shown in the following:

```

nslookup www.prokom.pl
Server: dhcp240.itsc.austin.ibm.com
Address: 9.3.240.2

```

```

Non-authoritative answer:
Name: mirror.prokom.pl
Address: 153.19.177.201
Aliases: www.prokom.pl

```

The data was captured by the `iptrace` command, similar to the following:

```
iptrace -a -P UDP -s 9.3.240.59 -b -d 9.3.240.2 /tmp/dns.query
```

The output from the `iptrace` command was formatted by the `ipreport` command, as follows:

```

TOK: ====(81 bytes transmitted on interface tr0)=== 17:14:26.406601066
TOK: 802.5 packet
TOK: 802.5 MAC header:

```

```

TOK: access control field = 0, frame control field = 40
TOK: [src = 00:04:ac:61:73:f7, dst = 00:20:35:29:0b:6d]
TOK: 802.2 LLC header:
TOK: dsap aa, ssap aa, ctrl 3, proto 0:0:0, type 800 (IP)
IP: < SRC = 9.3.240.59 > (server4f.itsc.austin.ibm.com)
IP: < DST = 9.3.240.2 > (dhcp240.itsc.austin.ibm.com)
IP: ip_v=4, ip_hl=20, ip_tos=0, ip_len=59, ip_id=64417, ip_off=0
IP: ip_ttl=30, ip_sum=aecc, ip_p = 17 (UDP)
UDP: <source port=49572, <destination port=53(domain) >
UDP: [udp length = 39 | udp checksum = 688d]
DNS Packet breakdown:
 QUESTIONS:
 www.prokom.pl, type = A, class = IN

TOK: ==== (246 bytes received on interface tr0)==== 17:14:26.407798799
TOK: 802.5 packet
TOK: 802.5 MAC header:
TOK: access control field = 18, frame control field = 40
TOK: [src = 80:20:35:29:0b:6d, dst = 00:04:ac:61:73:f7]
TOK: routing control field = 02c0, 0 routing segments
TOK: 802.2 LLC header:
TOK: dsap aa, ssap aa, ctrl 3, proto 0:0:0, type 800 (IP)
IP: < SRC = 9.3.240.2 > (dhcp240.itsc.austin.ibm.com)
IP: < DST = 9.3.240.59 > (server4f.itsc.austin.ibm.com)
IP: ip_v=4, ip_hl=20, ip_tos=0, ip_len=222, ip_id=2824, ip_off=0
IP: ip_ttl=64, ip_sum=7cc3, ip_p = 17 (UDP)
UDP: <source port=53(domain), <destination port=49572 >
UDP: [udp length = 202 | udp checksum = a7bf]
DNS Packet breakdown:
 QUESTIONS:
 www.prokom.pl, type = A, class = IN
 ANSWERS:
 -> www.prokom.plcanonical name = mirror.prokom.pl
 -> mirror.prokom.plinternet address = 153.19.177.201
 AUTHORITY RECORDS:
 -> prokom.plnameserver = phobos.prokom.pl
 -> prokom.plnameserver = alfa.nask.gda.pl
 -> prokom.plnameserver = amber.prokom.pl
 ADDITIONAL RECORDS:
 -> phobos.prokom.plinternet address = 195.164.165.56
 -> alfa.nask.gda.plinternet address = 193.59.200.187
 -> amber.prokom.plinternet address = 153.19.177.200

```

There are two packets shown in the **ipreport** output above (the key data is shown in boldface text). Every packet is divided into a few parts. Each part describes a different network protocol level. There are the token ring (TOK), IP, UDP, and the application (DNS) parts. The first packet is sent by host 9.3.240.59

and is a query about the IP address of the `www.prokom.pl` host. The second one is the answer.

## 8.4 NFS troubleshooting

Prior to starting any NFS debugging, it is necessary to ensure the underlying network is up and working correctly. It is also important to ensure that name resolution is functional and consistent across the network and that end-to-end routing is correct both ways.

### 8.4.1 General steps for NFS problem solving

The general steps for NFS problem solving are as follows:

1. Check for correct network connectivity and configuration as described in previous sections.
2. Check the following NFS configuration files on the client and server for content and permissions:
  - `/etc/exports` (servers only)
  - `/etc/rc.tcpip`
  - `/etc/rc.nfs`
  - `/etc/filesystems` (clients only)
  - `/etc/inittab`
3. Check that the following NFS daemons are active on the client and server.  
Server NFS daemons required:
  - `portmap`
  - `biod`
  - `nfsd`
  - `rpc.mountd`
  - `rpc.statd`
  - `rpc.lockd`Client NFS daemons required:
  - `portmap`
  - `biod` (these are dynamically created on AIX Version 4.2.1 and later)
  - `rpc.statd`
  - `rpc.lockd`

4. Initiate an **iptrace** (client, server, or network), reproduce the problem, then view the **ipreport** output to determine where the problem is.

## 8.4.2 NFS mount problems

Mount problems fall into one of the following categories:

- ▶ File system not exported, or not exported to a specific client.  
Correct server export list (/etc/exports)
- ▶ Name resolution different from the name in the export list. Normally, it is due to one of the following causes:
  - The export list uses a fully qualified name but the client host name is resolved without a network domain. Fully qualified names cannot be resolved. Mount permission is denied. Usually, this happens after upgrade activity and can be fixed by exporting to both forms of the name.
  - The client has two adapters using two different names and the export only specifies one. This problem can be fixed by exporting both names.
  - The server cannot do a lookuphostbyname or lookuphostbyaddr onto the client. To check, make sure the following commands both resolve to the same system:
    - **host name**
    - **host ip\_addr**
- ▶ The file system mounted on the server after **exportfs** was run. In this case, the **exportfs** command is exporting the mount point and not the mounted file system. To correct this problem run:  

```
/usr/etc/exportfs -ua; /usr/etc/exportfs -a
```

Then fix the /etc/filesystems file to mount the file system on boot, so it is already mounted when NFS starts from /etc/rc.nfs at system startup.
- ▶ Changes in the exports list, mounts, or somewhere else unexpectedly can sometimes lead to mountd getting confused. This usually happens following mounting, exporting, or because of mount point conflicts and similar errors. To correct this condition, mountd needs to be restarted by using the following commands:  

```
stopsrc -s rpc.mountd
startsrc -s rpc.mountd
```
- ▶ The system date being extremely off on one or both machines is another source of mount problems. To fix this, it is necessary to set the correct date and time, then reboot the system.

- ▶ Slow mounts from AIX Version 4.2.1 or later, clients running NFS Version 3 to AIX Version 4.1.5 or earlier and other non-AIX servers running NFS Version 2. NFS Version 3 uses TCP by default, while NFS Version 2 uses UDP only. This means that the initial client mount request using TCP will fail. To provide backwards compatibility, the mount is retried using UDP, but this only occurs after a timeout. To avoid this problem, NFS Version 3 provides the `proto` and `vers` parameters with the `mount` command. These parameters are used with the `-o` option to hard wire the protocol and version for a specific mount. The following example forces the use of UDP and NFS Version 2 for the mount request:

```
mount -o proto=udp,vers=2,soft,retry=1 platypus:/test /mnt
```

- ▶ Older non-AIX clients can also incur mount problems. If your environment has such clients, you need to start `mountd` with the `-n` option:

```
stopsrc -s rpc.mountd
startsrc -s rpc.mountd -n
```

- ▶ Another mount problem that can occur with older non-AIX clients is when a user who requests a mount is in more than eight groups. The only work-around for this is to decrease the number of groups the user is in or mount using a different user.

### 8.4.3 Increasing NFS Socket Buffer Size

You may find that the `netstat -s` command indicates a significant number of UDP socket buffer overflows. As with ordinary UDP tuning, increase the `sb_max` value. You also need to increase the value of `nfs_socketsize`, which specifies the size of the NFS socket buffer. The following is an example:

```
no -o sb_max=131072
nfsd -o nfs_socketsize=130972
```

**Note:** In AIX Version 4, the socket size is set dynamically. Configurations using the `no` and `nfsd` commands must be repeated every time the machine is booted. Add them in the `/etc/rc.net` or `/etc/rc.nfs` file immediately before the `nfsd` daemons are started and after the `biod` daemons are started. The position is crucial.

### 8.4.4 The `biod` and `nfsd` daemons

Starting with AIX 4.2.1, there is a single `nfsd` daemon and a single `biod` daemon, each of which is multi-threaded (multiple kernel threads within the process). Also, the number of threads is self-tuning in that it creates additional threads as needed. You can, however, tune the maximum number of `nfsd` threads by using

the `nfs_max_threads` parameter of the `nfsd` command. You can also tune the maximum number of biod threads per mount via the `biod mount` option.

## 8.5 Command summary

The following are commands discussed in this chapter and the flags most often used.

### 8.5.1 The `chdev` command

The `chdev` command changes the characteristics of a device. The command has the following syntax:

```
chdev -l Name [-a Attribute=Value ...]
```

The most commonly used flags are provided in Table 8-1.

Table 8-1 Commonly used flags of the `chdev` command

| Flag                      | Description                                                                                                                                         |
|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| -l <i>Name</i>            | Specifies the device logical name, specified by the name parameter, in the Customized Devices object class, whose characteristics are to be changed |
| -a <i>Attribute=Value</i> | Specifies the device attribute, value pairs used for changing specific attribute values                                                             |

### 8.5.2 The `exportfs` command

The `exportfs` command exports and unexports directories to NFS clients. The syntax of the `exportfs` command is:

```
exportfs [-a] [-v] [-u] [-i] [-fFile] [-oOption [,Option ...]]
[Directory]
```

The most commonly used flags are provided in Table 8-2.

Table 8-2 Commonly used flags of the `exportfs` command

| Flags            | Description                                                   |
|------------------|---------------------------------------------------------------|
| -a               | Exports all filesets defined in <code>/etc/exports</code>     |
| -u               | Unexports the directories you specify; can be used with -a    |
| -o <i>option</i> | Specifies optional characteristics for the exported directory |

### 8.5.3 The ifconfig command

The **ifconfig** command configures or displays network interface parameters for a network using TCP/IP. The command has the following syntax:

```
ifconfig Interface [AddressFamily [Address [DestinationAddress]]
[Parameters...]]
```

The most commonly used flags are provided in Table 8-3.

Table 8-3 Commonly used flags of the ifconfig command

| Flag                 | Description                                                                                                              |                                                                                                                        |
|----------------------|--------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| <i>AddressFamily</i> | Specifies which network address family to change.                                                                        |                                                                                                                        |
| <i>Parameters</i>    | alias                                                                                                                    | Establishes an additional network address for the interface.                                                           |
|                      | delete                                                                                                                   | Removes the specified network address.                                                                                 |
|                      | detach                                                                                                                   | Removes an interface from the network interface list.                                                                  |
|                      | down                                                                                                                     | Marks an interface as inactive (down), which keeps the system from trying to transmit messages through that interface. |
|                      | netmask <i>Mask</i>                                                                                                      | Specifies how much of the address to reserve for subdividing networks into subnetworks.                                |
| up                   | Marks an interface as active (up). This parameter is used automatically when setting the first address for an interface. |                                                                                                                        |
| <i>Address</i>       | Specifies the network address for the network interface.                                                                 |                                                                                                                        |

### 8.5.4 The iptrace command

The syntax of the **iptrace** command is:

```
iptrace [-a] [-e] [-PProtocol] [-iInterface] [-pPort]
[-sHost [-b]] [-dHost [-b]] LogFile
```

Some useful **iptrace** flags are provided in Table 8-4.

Table 8-4 Commonly used flags of the iptrace command

| Flags | Description            |
|-------|------------------------|
| -a    | Suppresses ARP packets |

| Flags          | Description                                                                |
|----------------|----------------------------------------------------------------------------|
| -s <i>host</i> | Records packets coming from the source host specified by the host variable |
| -b             | Changes the -d or -s flag to bidirectional mode                            |
| -e             | Enables promiscuous mode on network adapters that support this function    |

### 8.5.5 The **lsattr** command

The **lsattr** command displays attribute characteristics and possible values of attributes for devices in the system. The command has the following syntax:

```
lsattr -E -l Name [-a Attribute] ...
```

The most commonly used flags are provided in Table 8-5.

Table 8-5 Commonly used flags of the *lsattr* command

| Flag                | Description                                                                                                                  |
|---------------------|------------------------------------------------------------------------------------------------------------------------------|
| -E                  | Displays the attribute names, current values, descriptions, and user-settable flag values for a specific device              |
| -l <i>Name</i>      | Specifies the device logical name in the Customized Devices object class whose attribute names or values are to be displayed |
| -a <i>Attribute</i> | Displays information for the specified attributes of a specific device or kind of device                                     |

### 8.5.6 The **netstat** command

The **netstat** command shows network status. The command has the following syntax:

```
/bin/netstat [-n] [{ -r -i -I Interface }] [-f AddressFamily]
[-p Protocol] [Interval]
```

The most commonly used flags are provided in Table 8-6.

Table 8-6 Commonly used flags of the *netstat* command

| Flag | Description                                  |
|------|----------------------------------------------|
| -n   | Shows network addresses as numbers           |
| -r   | Shows the routing tables                     |
| -i   | Shows the state of all configured interfaces |

| Flag                       | Description                                                                                                   |
|----------------------------|---------------------------------------------------------------------------------------------------------------|
| -l <i>Interface</i>        | Shows the state of the configured interface specified by the interface variable                               |
| -f<br><i>AddressFamily</i> | Limits reports of statistics or address control blocks to those items specified by the AddressFamily variable |
| -p <i>Protocol</i>         | Shows statistics about the value specified by the protocol variable                                           |

## 8.5.7 The route command

The **route** command manually manipulates the routing tables. The command has the following syntax:

```
route Command [Family] [[-net | -host] Destination
[-netmask [Address]] Gateway] [Arguments]
```

The most commonly used flags are provided in Table 8-7.

Table 8-7 Commonly used flags of the route command

| Flag               | Description                                                                 |                                                   |
|--------------------|-----------------------------------------------------------------------------|---------------------------------------------------|
| <i>Command</i>     | <b>add</b>                                                                  | Adds a route                                      |
|                    | <b>flush</b> or <b>-f</b>                                                   | Removes all routes                                |
|                    | <b>delete</b>                                                               | Deletes a specific route                          |
|                    | <b>get</b>                                                                  | Looks up and displays the route for a destination |
| <i>-net</i>        | Indicates that the destination parameter should be interpreted as a network |                                                   |
| <i>-host</i>       | Indicates that the destination parameter should be interpreted as a host    |                                                   |
| <i>Destination</i> | Identifies the host or network to which you are directing the route         |                                                   |
| <i>-netmask</i>    | Specifies the network mask to the destination address                       |                                                   |
| <i>Gateway</i>     | Identifies the gateway to which packets are addressed                       |                                                   |

## 8.5.8 The tcpdump command

The syntax of the **tcpdump** command is:

```
tcpdump [-I] [-n] [-N] [-t] [-v] [-c Count] [-i Interface]
[-w File] [Expression]
```

The most useful **tcpdump** flags are provided in Table 8-8 on page 196.

Table 8-8 Commonly used flags of the tcpdump command

| Flags               | Description                                            |
|---------------------|--------------------------------------------------------|
| -c <i>Count</i>     | Exits after receiving <i>Count</i> packets             |
| -n                  | Omits conversion of addresses to names                 |
| -N                  | Omits printing domain name qualification of host names |
| -t                  | Omits the printing of a time stamp on each dump line   |
| -i <i>Interface</i> | Listens on <i>Interface</i>                            |

## 8.6 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

1. When a user tries to ping a particular machine on the network, she gets the error 0821-069 ping: sendto: Cannot reach the destination network. Which of the following procedures should be performed next to determine the cause of the problem?
  - A. **netstat -in**
  - B. **netstat -rn**
  - C. **route -nf**
  - D. **route refresh**
2. When a **ping -f** is executed, which of the following is represented when the periods are displayed?
  - A. The numbers of packets sent.
  - B. The number of packets returned.
  - C. The number of packets dropped.
  - D. There is no representation that indicates that the program is functioning.
3. Which of the following commands should be used to determine if a network interface is active?
  - A. **netstat -a**
  - B. **lsdev -Cc en0**
  - C. **lsdev -l adapter**
  - D. **lsdev -Cc if**

4. Which of the following commands will start the named daemon?
  - A. **refresh -a named**
  - B. **startsrc -a named**
  - C. **startsrc -s named**
  - D. **refresh -a named**
  
5. A second network adapter has been configured on a system and network connectivity is lost. Which of the following actions should be performed to fix the problem?
  - A. Check the routing.
  - B. Replace the second network adapter.
  - C. Check /etc/services for an incorrect entry.
  - D. Add the route from the first card to the second.
  
6. Which of the following procedures is most appropriate to secure a system from remote intruders?
  - A. Implement Network Information System (NIS).
  - B. Restrict remote root access.
  - C. Change the root password daily.
  - D. Place the system in a secure location.
  
7. Intermittent network delays are occurring on a system. Using the information provided in Figure 8-1, which of the following actions should be performed to determine the cause of the problem?

```
netstat -in
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
lo0 16896 <Link> 127.0.0.1 7219 0 7219 0 0
lo0 16896 127 127.0.0.1 7219 0 7219 0 0
en0 1500 <Link> 10.0.5a.4f.32.9a 132313 3543 89425 0 0
en0 1500 129.35.56 129.35.56.160 132313 3543 89425 0 0
```

Figure 8-1 Case study

- A. **arp**
- B. **entstat -d**
- C. **netstat -rn**
- D. **errpt**

8. Which of the following options will *not* show if the interface on the local host is active?
- A. **lsattr**
  - B. **lsdev**
  - C. **entstat**
  - D. **ifconfig**
9. A user is not able to ping any machines on the same network. Which of the following options is most likely causing the problem?
- A. The loopback is down.
  - B. The routing table is wrong.
  - C. The default gateway is wrong.
  - D. The network interface is down.
10. All of the following options suggest why a file system cannot be mounted from a NFS server, *except*:
- A. The server is not running the `rpc.mountd`.
  - B. The `mountd` subsystem must be refreshed.
  - C. The date is incorrect on one or both machines.
  - D. The file system was mounted after **exportfs** was run.
11. The network security team tells a system administrator to turn on promiscuous mode. Which command accomplishes this task?
- A. **iptrace -p**
  - B. **iptrace -e**
  - C. **iptrace -d**
  - D. **iptrace -a**
12. What limits the number of packets sent to and received from a remote host using the `ping` command?
- A. **ping -d**
  - B. **ping -D**
  - C. **ping -c**
  - D. **ping -q**

13. Assume that an AIX workstation has just been modified to be a DHCP client. Previously, the workstation had a static IP definition and was functioning on the network. A shutdown and restart of the workstation was performed which revealed that it did receive an address from the DHCP server. However, the workstation is unable to reach certain hosts on the network. Which of the following procedures should be performed next?
- A. Set a default route using `ifconfig`.
  - B. Run the SMIT fast path `smit route` to finish the configuration.
  - C. Refresh the `dhcpcd` daemon by sending it a `SIGHUP`.
  - D. Run the SMIT fast path `smit usedhcp` to finish the DHCP configuration.
14. Which of the following commands should be used to display collision errors on an Ethernet adapter?
- A. `errpt -a`
  - B. `netstat -i`
  - C. `entstat -d`
  - D. `lsattr -El en0`
15. Which of the following commands records IP packets received from configured interfaces?
- A. `iptrace`
  - B. `entstat`
  - C. `tcpdump`
  - D. `traceroute`

## 8.6.1 Answers

The following are the preferred answers to the questions provided in this section.

1. B
2. C
3. D
4. C
5. A
6. B
7. B
8. B
9. D
10. C
11. B
12. C
13. B
14. C
15. A

## 8.7 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

1. Check the settings of your network interface with the **lsattr** command.
2. Check name resolution ordering of your system.
3. Try to resolve a few host names to IP addresses using either **nslookup** or **host** command.



## System access and printer problem determination

The following topics are discussed in this chapter:

- ▶ User license problems
- ▶ Telnet problems
- ▶ System settings
- ▶ Tracing

It can be very frustrating when you cannot access a system. There are many reasons for this, despite a valid user account and the corresponding password. This chapter discusses some of the reasons why a system may have access problems and suggests some solutions for them.

## 9.1 User license problems

If it is not possible to log in to an AIX system because your session is disconnected after a login from the login prompt, an AIX license problem could exist.

The following are ways that a user can access the system, which requires an AIX Version 4 user license:

- ▶ Logins provided from a `getty` (from an active, local terminal)
- ▶ Logins provided using the `rlogin` or `rsh -l` command
- ▶ Logins provided using the `telnet` or `tn` command
- ▶ Logins provided through the Common Desktop Environment (visual login CDE)

All other ways of accessing a base AIX Version 4 system does not require AIX user licenses (for example, `ftp`, `rexec`, and `rsh` without the `-l` flag).

The `lslicense` command displays the number of fixed licenses and the status of the floating licensing, as in the following example:

```
lslicense
Maximum number of fixed licenses is 32.
Floating licensing is disabled.
```

To change the number of licenses, use the `smit chlicense` fast path. Figure 9-1 on page 203 shows the corresponding SMIT screen.

```

Change / Show Number of Licensed Users

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

Maximum number of FIXED licenses [Entry Fields] #
FLOATING licensing [32] +
 off

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

```

Figure 9-1 SMIT menu to change the number of licensed users

In order for the changes to take effect, a system reboot is required.

## 9.2 Telnet troubleshooting

If a telnet connection to an AIX system is not possible, there can be a number of causes, such as:

- ▶ No network connection.
- ▶ The inetd server is not running.
- ▶ The telnet subserver is not configured.
- ▶ There are slow login times because of name server problems.
- ▶ Telnet error.

In the following sections, these problem areas are discussed in further detail.

### 9.2.1 Network problems

If a telnet from a client shows the following error message, it is likely to be a network problem:

```
telnet server1
Trying...
```

```
telnet: connect: A remote host did not respond within the timeout period.
```

Try the **ping** command to see if the destination system can be reached. If you cannot ping the system, your problem is related to the network and it can either be the system itself or an access error to the network due to a router or gateway failure.

## 9.2.2 The telnet subserver

The telnet service is managed by a subserver controlled by the inetd super daemon. If a telnet session shows the following error message, use the following steps to analyze and recover the problem:

```
telnet server1
Trying...
telnet: connect: A remote host refused an attempted connect operation.
```

1. Check to see if the inetd subsystem is running by using the system resources controller (SRC) **lssrc** command.

```
lssrc -s inetd
Subsystem Group PID Status
inetd tcpip 7482 active
```

2. Check to see if the telnet subserver is defined.

```
grep telnet /etc/inetd.conf
telnet stream tcp6 nowait root /usr/sbin/telnetd telnetd
```

3. Start the telnet subserver using the SRC **startsrc** command with the **-t** option.

```
startsrc -t telnet
0513-124 The telnet subserver has been started.
```

Verify that the telnet subserver is running with the **lssrc** command.

```
lssrc -t telnet
Service Command Description Status
telnet /usr/sbin/telnetd telnetd -a active
```

When the telnet subserver is running, a login prompt similar to the following is presented:

```
telnet server1
Trying...
Connected to server1.
Escape character is '^]'.

telnet (server1)
```

AIX Version 4

(C) Copyrights by IBM and by others 1982, 1996.  
login:

If the **telnet** command displays the following error, the telnet problem is likely to be related to the `/etc/services` file:

```
telnet server1
telnet: tcp/telnet: unknown service
```

The file might be corrupt or the telnet entry is missing. The following stanza should be present in the `/etc/services` file, mapping the telnet service to port 23:

```
grep telnet /etc/services
telnet 23/tcp
```

### 9.2.3 Slow telnet login

If the login with **telnet** takes a long time, for example, over two minutes, it is likely that the problem is related to the domain name system (DNS) name server resolution. On the server the telnet daemon is running, check the `/etc/resolv.conf` file.

The `/etc/resolv.conf` file defines the DNS name server information for local resolver routines. If the `/etc/resolv.conf` file does not exist, the DNS is not available and the system will attempt name resolution using the default paths, the `/etc/netsvc.conf` file (if it exists), or the `NSORDER` environment variable (if it exists).

When a DNS server is specified during TCP/IP configuration, a `/etc/resolv.conf` file is generated. Further configuration of the `resolv.conf` file can be done using the **smit resolv.conf** fast path.

Determine the IP address of your name server from the `/etc/resolv.conf` file. Then test if name resolution is working correctly using the **nslookup** command (to determine the IP address of your telnet client machine), with the host name as input. If the DNS name server does not respond, contact the network administrator to fix the problem or, alternatively, provide you with another name server. Additionally, change the name resolution order by either editing or creating the `/etc/netsvc.conf` file. Change the search order to be the same as the following example:

```
hosts=local, bind
```

This will force the system to use the `/etc/hosts` file for name resolution first. Enter a stanza for your telnet client machine and your login time should improve significantly.

## 9.2.4 Telnet error

If when attempting a telnet session it fails with error code 0403-031, this indicates that a fork function has failed due to insufficient memory and/or paging space on the server. Memory is most likely overcommitted and working pages are being released from real memory and filling the paging space. To prevent this we suggest tuning `minfree`, `maxfree`, `maxperm`, and `minperm` using the `vm tune` command, after the server has been rebooted, to free up the memory.

## 9.3 FTP troubleshooting

The `ftp`, `rcp`, and `tftp` commands rely on TCP/IP to establish direct connections from your local host to a remote host. Basic Network Utilities (BNU) can also use TCP/IP to provide direct connections with remote hosts.

### 9.3.1 File limits

If you are transferring a file over 2 GB in size, the following should be considered:

- ▶ In the target server's `/etc/security/limits` file, the default stanza identifies global limit. The `fsize` has to be larger than the file you are transferring or the limit can be set to unlimited with the `-1` option. Limits can be set at the user level. For example, the global `fsize` is set to unlimited (`-1`), but `ftpuser` `fsize` is set to 3 GB.
- ▶ Even if the limit is set to unlimited, the target server's file system has to be a large file enabled journaled file system. You can use the `smitty crjfsbf` fast path to create a large file enabled file system. Figure 9-2 on page 207 shows an example.

```

Add a Large File Enabled Journaled File System

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

 [Entry Fields]
Volume group name rootvg
* SIZE of file system (in 512-byte blocks) [] #
* MOUNT POINT []
Mount AUTOMATICALLY at system restart? no +
PERMISSIONS read/write +
Mount OPTIONS [] +
Start Disk Accounting? no +
Number of bytes per inode 4096 +
Allocation Group Size (MBytes) 64 +

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

```

Figure 9-2 The smitty crjfsbf fast path

You can validate if a file system is large file enabled by checking the logical volume:

```

lsfs -q /dev/lv01
Name Nodename Mount Pt VFS Size Options
Auto
Accounting
/dev/lv01 -- /test jfs 32768 rw
no
(lv size: 32768, fs size: 32768, frag size: 4096, nbpi: 4096, compress:
no, bf: true, ag: 16)

```

Note that bf: true indicates that this is large-file enabled.

## 9.4 System settings

In some cases, specific system settings should be checked to resolve access problems. This section describes the most common cases.

### 9.4.1 Adjusting AIX kernel parameters

Some applications need to run as a certain type of user, such as database applications. Depending on the implementation, some of these applications may require a large set of running processes. However, the number of processes per

user is limited and defined as an AIX kernel parameter. If you see the following error message, it is likely that you have reached the maximum possible number of processes per user:

```
0403-030 fork function failed too many processes exist
```

This can be changed using the `smit chgsys` fast path. Figure 9-3 shows the corresponding SMIT screen.

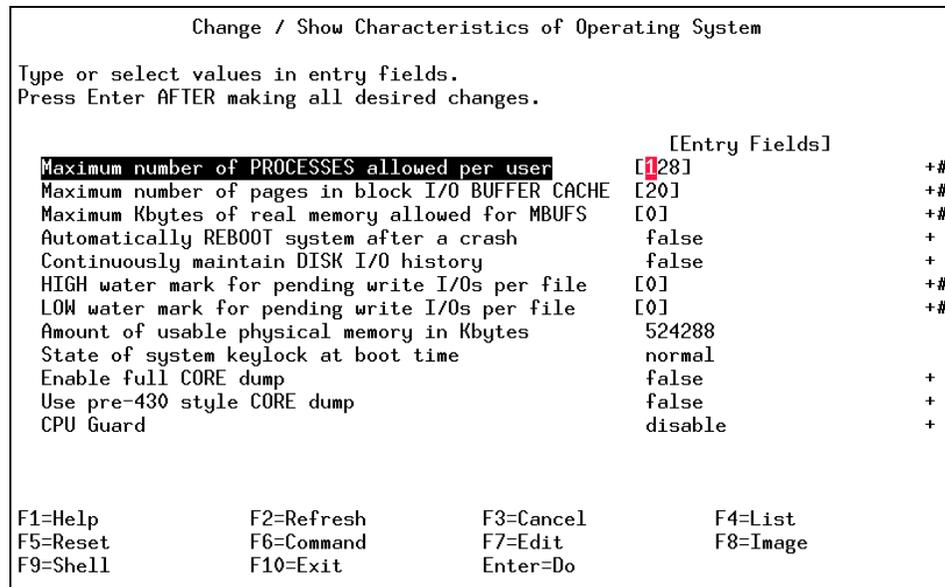


Figure 9-3 SMIT screen for changing AIX operating system characteristics

The same value can be changed using the `chdev` command on the device `sys0` by setting the attribute `maxuproc`. For example, type:

```
chdev -l sys0 -a maxuproc=200
sys0 changed
```

## 9.4.2 The `su` command

The `su` command changes user credentials to those of the root user or to the user specified by the name parameter, and then initiates a new session. The following functions are performed by the `su` command:

**Account checking** Validates that the user account that is enabled for the `su` command. Checks that the current user is in a group permitted to switch to this account with the `su` command, and that it can be used from the current controlling terminal.

|                                  |                                                                                                                                                                                                                                                               |
|----------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>User authentication</b>       | Validates the user's identity by using the system-defined primary authentication methods for the user. If a password has expired, the user must supply a new password.                                                                                        |
| <b>Credentials establishment</b> | Establishes initial user credentials by using the values in the user database. These credentials define the user's access rights and accountability on the system.                                                                                            |
| <b>Session initiation</b>        | If the minus (-) flag is specified, the <b>su</b> command initializes the user environment from the values in the user database and the <code>/etc/environment</code> file. When the - flag is not used, the <b>su</b> command does not change the directory. |

Examine the following example for the use of the - flag. The **su** command is used with the - flag and, as shown in the following, the user environment is set for the user `ostach`:

```
id
uid=0(root) gid=0(system) groups=2(bin),3(sys),7(security),8(cron),10(audit)
su - ostach
$ id
uid=201(ostach) gid=1(staff)
$ env
LOGIN=ostach
LOGNAME=ostach
MAIL=/usr/spool/mail/ostach
USER=ostach
HOME=/home/ostach
PWD=/home/ostach
```

In the following example, the **su** command is used with - flag, which specifies that the process environment is to be set as if the user had logged in to the system using the **login** command. Nothing in the current environment is propagated to the new shell.

```
$ id
uid=201(ftpuser) gid=1(staff)
$ su - root
root's Password:
id
uid=0(root) gid=0(system)
groups=2(bin),3(sys),7(security),8(cron),10(audit),11(lp)
```

In the next example, the **su** command is used without the - flag, so the user `ostach` has an environment set for the root user.

```
id
```

```
uid=0(root) gid=0(system) groups=2(bin),3(sys),7(security),8(cron),10(audit)
su ostach
$ id
uid=201(ostach) gid=1(staff)
$ env
LOGIN=ostach
LOGNAME=root
MAIL=/usr/spool/mail/root
USER=root
R_PORT=49213
HOME=/
PWD=/
```

Each time the **su** command is run, an entry is recorded in the `/var/adm/sulog` file. The `/var/adm/sulog` file records the following information: Date, time, system name, and login name. The `/var/adm/sulog` file also records whether the login attempt was successful: a plus sign (+) indicates a successful login, and a minus sign (-) indicates an unsuccessful login.

```
cat /var/adm/sulog
SU 06/30 09:36 + pts/1 root-thomasc
SU 06/30 10:25 + pts/3 root-thomasc
SU 07/18 11:56 + pts/3 root-ostach
SU 07/18 13:05 + pts/3 root-ostach
SU 07/18 13:05 - pts/3 ostach-root
SU 07/18 13:06 - pts/3 ostach-thomasc
```

### 9.4.3 A full file system

When a file system on your system becomes full, it can cause logins to the system (using telnet), a directly connected TTY, or the system console to fail. The following message is typically displayed:

```
telnet problem 004 - 004 you must exect "login from the lowest login shell"
```

Or, on the system console, the following error message may be displayed:

```
3004-004 you must 'exec' login from the lowest login shell
```

Check that your file systems are not full, especially the `/` (root) file system. Use the **df** command to verify the status of free disk space on your file systems. If a file system is full, enlarge the file system using the **chfs** command.

If your file systems are not full, check to see if the following files are okay:

- ▶ `/etc/utmp`
- ▶ `/etc/security/limits`

Check that the files exist, and that the permissions and ownerships are correct. If the problem persists, check to see if there is an APAR that addresses this or a similar problem.

## 9.5 Tracing

The trace system is a tool that allows you to capture the sequential flow of system activity or system events. Unlike a stand-alone kernel dump that provides a static snapshot of a system, the trace facility provides a more dynamic way to gather problem data.

Trace can be used to:

- ▶ Isolate, understand, and fix system or application problems.
- ▶ Monitor system performance.

The events that are traced are time stamped as they are written to a binary trace file named `/var/adm/ras/trcfile`.

There are trace events predefined in AIX and included in selected commands, libraries, kernel extensions, device drivers, and interrupt handlers. A user can also define his own trace events in application code.

The trace facility generates a large amount of data. For example, a trace session capturing one second of events from an idle system gathered four thousand events. The amount of data depends on what events you trace and the CPU performance of the system.

The trace facility and commands are provided as part of the Software Trace Service Aids fileset named `bos.sysmgt.trace`.

**Note:** Before tracing events, it is important to have a strategy for what to trace, and the time the tracing is to be done.

Follow these steps to gather a useful trace:

1. Select the trace hook IDs for tracing.
2. Start the trace.
3. Recreate the problem.
4. Stop the trace.
5. Generate the trace report.

## 9.5.1 Trace hook IDs

The events traced are referenced by hook identifiers. Each hook ID uniquely refers to a particular activity that can be traced.

Hook IDs are defined in the `/usr/include/sys/trchkid.h` file. When tracing, you can select the hook IDs of interest by using the `trace -j` command, and exclude others that are not relevant to your problem by using the `trace -k` command.

The following example is extracted from the `trchkid.h` file:

```
...
#define HKWD_SYSC_MKDIR 0x15600000
#define HKWD_SYSC_MKNOD 0x15700000
#define HKWD_SYSC_MNTCTL 0x15800000
#define HKWD_SYSC_MOUNT 0x15900000
#define HKWD_SYSC_NICE 0x15a00000
#define HKWD_SYSC_OPEN 0x15b00000
#define HKWD_SYSC_OPENX 0x15c00000
#define HKWD_SYSC_OUNAME 0x15d00000
#define HKWD_SYSC_PAUSE 0x15e00000
#define HKWD_SYSC_PIPE 0x15f00000
#define HKWD_SYSC_PLOCK 0x16000000
#define HKWD_SYSC_PROFIL 0x16100000
#define HKWD_SYSC_PTRACE 0x16200000
#define HKWD_SYSC_READ 0x16300000
#define HKWD_SYSC_READLINK 0x16400000
#define HKWD_SYSC_READX 0x16500000
#define HKWD_SYSC_REBOOT 0x16600000
#define HKWD_SYSC_RENAME 0x16700000
#define HKWD_SYSC_RMDIR 0x16800000
...
```

When specifying the hook ID to the `trace` command, only the three left-most digits should be specified. From the preceding example, when the open system call is traced, the ID 15b must be specified.

Specifying relevant (or irrelevant) hook IDs can be difficult at this time since you do not know the actual cause of the problem. If source code access to the application is available or the developer is known, then this can be helpful for specifying useful hook IDs.

**Note:** Specifying useful hook IDs can reduce the amount of data significantly and make the analysis part of the problem easier.

## 9.5.2 Starting a trace

A trace can be started in background mode or interactive mode.

The usual way to perform a trace is in the background using the `-a` flag. An ampersand (`&`) is not necessary at the end of the command, as the `trace` command will spawn the trace daemon and return to the shell prompt immediately. The trace is stopped using the `trcstop` command.

To perform a trace in interactive mode, invoke the `trace` command with a list of events you want to monitor and the name of the trace log output file. The events are assigned numbers that are called trace hooks.

A typical command sequence may be as follows:

```
trace -a -j 15b
myprogram
trcstop
```

This example traces only the open operating system that is made on the system.

Trace uses in-memory buffers to save the trace data. There are three methods of using the trace buffers:

- |                       |                                                                                                                                                  |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Alternate mode</b> | This is the default mode. All trace events will be recorded in the trace log file.                                                               |
| <b>Circular mode</b>  | The trace events wrap within the in-memory buffers and are not captured in the trace log file until the trace data collection is stopped.        |
| <b>Single mode</b>    | The collection of trace events stops when the in-memory trace buffer fills up and the contents of the buffer are captured in the trace log file. |

## 9.5.3 Trace reports

The binary `/var/adm/ras/trcfile` trace file contains all system events collected during the trace period. To obtain a readable format, this file needs to be translated using the `trcrpt` command, which generates an output report.

To write a formatted trace report to the `/tmp/trace.out` file, run the following command:

```
trcrpt -o /tmp/trace.out
```

The output of the trace report file is usually very large, depending on the trace parameters and the system activity. Despite selecting a narrow time period for



Since this system is a 4-way SMP, the overall CPU usage is only 25 percent. To analyze what is actually happening on this system, use the trace facility. Notice that the process ID of  **aixterm**  is 19436.

This  **aixterm**  command process seems to wait continuously; therefore, the tracing time is limited to one second.

Using the following command sequence, all system events for one second are traced:

```
trace -a; sleep 1; trcstop
```

It is not known what the  **aixterm**  process is doing since no event hook IDs can be specified at this point. The trace generates a raw trace file of the following size:

```
ls -l /var/adm/ras/trcfile
-rw-rw-rw- 1 root system 557152 Jul 17 14:27 /var/adm/ras/trcfile
```

Based on this file, a trace report can be generated with  **trcrpt** . Since the process ID is known, you can use this information as a filter and limit the output of the report using the following commands:

```
trcrpt -p 19436 > /tmp/trace.out
ls -l /tmp/trace.out
-rw-r--r-- 1 root system 201014 Jul 17 14:31 /tmp/trace.out
```

The contents of the trace report is provided in the following example as an extracted part of the complete report (limited due to space constraints):

```
Mon Jul 17 14:27:27 2000
System: AIX server1 Node: 4
Machine: 000BC6FD4C00
Internet Address: 0903F038 9.3.240.56
The system contains 4 cpus, of which 4 were traced.
Buffering: Kernel Heap
This is from a 32-bit kernel.
```

```
trace -a
```

| ID  | ELAPSED_SEC                                                                 | DELTA_MSEC | APPL | SYSCALL | KERNEL | INTERRUPT    |
|-----|-----------------------------------------------------------------------------|------------|------|---------|--------|--------------|
| 100 | 0.004256674                                                                 | 4.256674   |      |         |        | DECREMENTER  |
|     | INTERRUPT iar=D031EB                                                        |            |      |         |        |              |
|     | 60 cpuid=FFFFFFFF                                                           |            |      |         |        |              |
| 234 | 0.004258505                                                                 | 0.001831   |      | clock:  |        | iar=D031EB60 |
|     | lr=D036CA24 [2503 usec]                                                     |            |      |         |        |              |
| 112 | 0.004260143                                                                 | 0.001638   |      | lock:   |        | lock lock    |
|     | addr=352118 loc                                                             |            |      |         |        |              |
|     | k status=10000001 requested_mode=LOCK_READ return addr=2D80C name=0000.0000 |            |      |         |        |              |

```

113 0.004261661 0.001518 unlock: lock addr=352118
lock status=000
0 return addr=2D8D8 name=0000.0000
112 0.004270432 0.008771 lock: lock lock
addr=352118 loc
k status=10000001 requested_mode=LOCK_READ return addr=2D3C4 name=0000.0000
113 0.004271781 0.001349 unlock: lock addr=352118
lock status=000
0 return addr=2D5A4 name=0000.0000
112 0.004272503 0.000722 lock: lock lock
addr=352118 loc
k status=10000001 requested_mode=LOCK_READ return addr=2DEA8 name=0000.0000
113 0.004274213 0.001710 unlock: lock addr=352118
lock status=000
0 return addr=2E0EC name=0000.0000
112 0.004274863 0.000650 lock: lock lock
addr=352118 loc
k status=10000001 requested_mode=LOCK_READ return addr=2D90C name=0000.0000
113 0.004275706 0.000843 unlock: lock addr=352118
lock status=000
0 return addr=2D980 name=0000.0000
10E 0.004278910 0.003204 relock: lock addr=34DEA0
oldtid=12679
newtid=1033
10E 0.004279946 0.001036 relock: lock addr=34DEA0
oldtid=1033 n
ewtid=12679
106 0.004280644 0.000698 dispatch: cmd=aixterm
pid=19436 tid=12
679 priority=93 old_tid=12679 old_priority=93 CPUID=2 [3551 usec]
200 0.004283438 0.002794 resume aixterm
iar=D031EB60 cpuid=02
100 0.014254631 9.971193 DECREMENTER
INTERRUPT iar=D031EB
60 cpuid=02
234 0.014256414 0.001783 clock: iar=D031EB60
lr=D036CA24 [2497
usec]
112 0.014258004 0.001590 lock: lock lock
addr=352118 loc
k status=10000001 requested_mode=LOCK_READ return addr=2D80C name=0000.0000
...

```

The heading shows system information. The next section shows the parameters used to activate the **trace** command. In the next section, the actual report is provided, where each line is the event recorded. The first column shows the event hook IDs the system has performed.

From the output of this example, it appears that the  **aixterm**  process is hung up waiting for some kernel resources, as the only events the process is performing are lock and unlock operations. To go into deeper analysis of this problem, you would need to look into the program source code of the application you are tracing.

## 9.6 Managing mail logging

The  **sendmail**  command logs mail system activity through the syslogd daemon. The syslogd daemon must be configured and running for logging to occur. Specifically, the  `/etc/syslog.conf`  file should contain the uncommented line:

```
mail.debug /var/spool/mqueue/log
```

Refresh the syslogd daemon by typing the following at the command line:

```
refresh -s syslogd
0513-095 The request for subsystem refresh was completed successfully.
```

If the  `/var/spool/mqueue/log`  file does not exist, you must create it by typing the following command:

```
touch /var/spool/mqueue/log
```

The types of activities that the  **sendmail**  command puts into the log file are specified by the option in the  `/etc/mail/sendmail.cf`  file. (For versions earlier than AIX 5L Version 5.1, this file is  `/etc/sendmail.cf` .)

Because information is continually appended to the end of the log, the file can become very large. Run the  `/usr/lib/smdemon.cleau`  shell script to keep the file from growing too large.

If you suspect a problem, you can set traffic logging by using the  `-X`  flag. For example:

```
#/usr/sbin/sendmail -X /tmp/traffic -bd
```

This command logs all traffic in the  `/tmp/traffic`  file. This command logs a lot of data and should not be used during normal operations.

The  **sendmail**  command tracks the volume of mail being handled by each of the mailer programs that interface with it. These mailers are defined in the  `/etc/mail/sendmail.cf`  file. To start the accumulation of mailer statistics, create the  `/etc/mail/statistics`  file by typing the following:

```
touch /etc/mail/statistics
```

If the **sendmail** command encounters errors when trying to record statistics information, the command writes a message through the syslog subroutine. The **sendmail** command updates the information in the file each time it processes mail.

The statistics kept in the `/etc/mail/statistics` file are in a database format that cannot be read as a text file. To display the mailer statistics, type the following at the command line:

```
/usr/sbin/mailstats
```

This reads the information in the `/etc/mail/statistics` file, formats it, and writes it to standard output.

## 9.7 TTY troubleshooting

The following sections discuss various common TTY problems and their solutions.

The `/usr/bin/tty [-s]` command writes to standard output the full path name of your terminal. If your standard input is not a terminal and you do not specify the `-s` flag, you get the message **Standard input is not a tty**.

This chapter discusses some common TTY errors, recovery procedures, and also general TTY management commands.

### 9.7.1 Respawning too rapidly errors

The system records the number of getty processes spawned for a particular TTY in a short time period. If the number of getty processes spawned in this time frame exceeds five, then the Respawning Too Rapidly error is displayed on the console and the port is disabled by the system.

The TTY stays disabled for about 19 minutes or until the system administrator enables the port again. At the end of 19 minutes, the system automatically enables the port, resulting in the spawning of a new getty process.

#### Possible causes

Possible causes are:

- ▶ Incorrect modem configuration
- ▶ A port is defined and enabled but no cable or device is attached to it
- ▶ Bad cabling or loose connection
- ▶ Noise on communication line
- ▶ Corruption of `/etc/environment` or `/etc/inittab` files

- ▶ TTY configuration is corrupted
- ▶ Hardware is defective

## Recovery procedures

Recovery procedures are:

- ▶ Correct modem configuration as per your modem configuration documentation.
- ▶ Disable the TTY, remove the TTY definition, or attach a device to the port.

To disable the TTY definition use the **chdev** command as follows:

```
#chdev -l ttyName -a Login=disable
```

After running this command, the TTY does not become enabled after a system restart.

To remove the TTY definition, disable the TTY port using the **pdisable** command:

```
#pdisable ttyName
```

- ▶ Check cabling and that the correct part number is being used.
- ▶ Eliminate noise on communication line.
- ▶ In the /etc/inittab file, examine the TTY devices line. If the TTY is set to off, it is likely that the TTY port is not being used. If not, remove it or add a device to the port.
- ▶ Remove the corrupted TTY configuration.
- ▶ Locate the defective hardware using diag.

Refer to Table 9-1 for additional TTY commands.

*Table 9-1 Managing TTY devices*

| Task                      | SMIT fast path      | Command or file    | Web-based system manager management environment        |
|---------------------------|---------------------|--------------------|--------------------------------------------------------|
| List defined tty devices. | <b>smit lsdttty</b> | lsdev -C -c tty -H | <b>Software -&gt; Devices -&gt; All Devices</b>        |
| Add a tty.                | <b>smit mkttty</b>  | mkdev -t tty       | <b>Software -&gt; Devices -&gt; Overview and Tasks</b> |

| Task                     | SMIT fast path     | Command or file                                          | Web-based system manager management environment                           |
|--------------------------|--------------------|----------------------------------------------------------|---------------------------------------------------------------------------|
| Move a tty.              | <b>smit movtty</b> | <i>chdev -l Name -p Parentname -w ConnectionLocation</i> |                                                                           |
| Change a tty.            | <b>smit chtty</b>  | <i>lsattr -l Name -E; chdev -l Name</i>                  | <b>Software -&gt; Devices</b>                                             |
| Remove a tty.            | <b>smit rmtty</b>  | <i>rmdev -l Name</i>                                     | <b>Software &gt; Devices -&gt; All Devices -&gt; Selected &gt; Delete</b> |
| Configure a defined tty. | <b>smit mktty</b>  | <i>mkdev -l Name</i>                                     | <b>Software -&gt; Devices -&gt; Overview and Tasks</b>                    |

## 9.8 Printing

AIX PowerPC-based printing administrators must be assigned to the `printq` group to be able to manage printing subsystems. Users who are members of the `printq` administrative group can add print devices to the operating system, which can be used by the print subsystem. For detailed information on printing, see *Printing for Fun and Profit under AIX 5L*, SG24-6018.

### 9.8.1 Local printing problem

The `lpstat` command displays information about the current status of the line printer. To display the status for all print queues, enter:

```
lpstat
Queue Dev Status Job Files User PP % Blks Cp Rnk

IBM_Loc lp1 READY
```

If the queue is up, check to see if the job is missing, make sure the printer is set up properly, etc. The `qchk` command displays the current status information regarding specified print jobs, print queues, or users. The basic syntax of the `qchk` command is:

```
qchk -P QueueName -# JobNumber -u OwnerName
```

## 9.8.2 Remote printing problem

Make sure the lpd daemon is active by running:

```
lssrc -s lpd
Subsystem Group PID Status
lpd spooler 36550 active
```

Check to see if the host name can be resolved, if the client has permission (/etc/hosts.lpd file) to print to the remote server, etc.

## 9.8.3 The /var file system

You must have enough room in the /var file system for print spools. If the space in /var is smaller than the file you are trying to print, increase the file system size using the **smitty chjfs** fast path. Spooling files makes copies of the file before sending them to the printer. For PowerPC-based printing, files are spooled to the /var/spool/lpq/qdaemon directory.

## 9.8.4 Default printing subsystem

The print subsystems have a lot in common in terms of architecture and yet are quite different in their approach to that architecture. All three subsystems pass the file to be printed through a series of filters that can make changes to the data, add setup information, and pass the data to the printed device.

Table 9-2 shows some of the similarities and differences between these print subsystems.

Table 9-2 Comparison of print subsystem functions

| Action or function                                     | System V                                                                                                                                      | PowerPC                                                                                                                                                                                                                                      | Infoprint Manager for AIX                                                                                                                                                                                                    |
|--------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Printing from the command line and flags passed</b> | Primary command is <b>lp</b> with <b>lpr</b> also available. The <b>lp</b> command has <b>-o</b> option for print command line printer setup. | Primary commands are <b>enq</b> and <b>qprt</b> , but supports <b>lp</b> and <b>lpr</b> with slightly different commands. The <b>lp</b> and <b>enq</b> commands have <b>-o</b> options and <b>qprt</b> has direct options for printer setup. | Primary command is <b>pdpr</b> ; the <b>-x</b> flag is used to pass attributes or options. <b>enq</b> , <b>qprt</b> , <b>lp</b> , and <b>lpr</b> are also supported. Infoprint clients are also available on some platforms. |

| Action or function                 | System V                                                                                                                                                                            | PowerPC                                                                                                                                                                               | Infoprint Manager for AIX                                                                                                                                                                                                                     |
|------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Administration</b>              | Web-based System Manager and command line with <b>1padmin</b> .                                                                                                                     | Web-based System Manager, SMIT, and command line commands like <b>1svirprt</b> .                                                                                                      | Java administrator and operator GUIs on AIX, Windows NT, and Windows 2000, AIX VSM GUI, limited SMIT, and command line.                                                                                                                       |
| <b>Scheduler</b>                   | <b>1psched</b>                                                                                                                                                                      | <b>qdaemon</b>                                                                                                                                                                        | <b>pdserver</b>                                                                                                                                                                                                                               |
| <b>Printer or queue definition</b> | Subdirectory of files in /etc/lp/printers directory.                                                                                                                                | Stanza in /etc/qconfig.                                                                                                                                                               | Object orientated: Information stored in objects in the pdserver.                                                                                                                                                                             |
| <b>Job processing: stage one</b>   | Interface shell scripts: Information passed by parameters and three shell script arguments. The fifth parameter contains multiple print options passed by the <b>1p -o</b> options. | Backend programs: Information passed by libq subroutine calls and command line parameters. When piobe is the backend, a virtual printer is defined that fully supports printer setup. | Command line attributes are converted to Infoprint Manager attributes where appropriate; other attributes are passed to the backend.                                                                                                          |
| <b>Printer customization</b>       | Some is possible through the standard script and terminfo, but most often done through custom interface programs provided by printer manufacturers.                                 | Uses virtual printer definitions built on predefined printer-specific files and customized through SMIT for user requirements.                                                        | Customizable using customer interfaces. IPDS printers are predefined by attachment type with specialized templates available for selected printers. AIX printers use internal printer definitions built on predefined printer-specific files. |

| Action or function              | System V                                                                                                                                                                                                                           | PowerPC                                                                                                                                                                                                                                                                                                                                                                                                 | Infoprint Manager for AIX                                                                                                                                                                                                                                |
|---------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Print processing filters</b> | Called by interface script based on filter setup with <code>lpfilter</code> and by actually calling the filters in the shell script. Some filters are called by <code>lpsched</code> to convert files to PostScript automatically. | Formatting filters, such as <code>pioformat</code> , are called automatically based on the data type set up for the printer. Other filters are called by the <code>piobe</code> backend based on virtual printer attributes starting with <code>f</code> , and can be called by users from the command line. The <code>encrypt</code> filter can be called automatically to convert text to PostScript. | Depends on type of incoming data and destination. Data stream transforms are called automatically based on the type of input data if the destination supports it. AIX formatting filters if the destination is supported by the PowerPC print subsystem. |
| <b>Output filters</b>           | The shell script passes the setup and commands to output filters such as <code>lp.cat</code> and <code>postio</code> , which can talk directly with printers and handle errors.                                                    | The <code>piobe</code> backend passes data through <code>pioout</code> , which writes directly to the device. Errors are handled by communications through the <code>qdaemon</code> .                                                                                                                                                                                                                   | Secondary processes are provided to communicate with TCP/IP-attached printers using sockets, channel-attached printers, and PowerPC print filters.                                                                                                       |
| <b>LPD/LPR remote printing</b>  | Controlled by <code>lpsched</code> , which starts <code>lpNet</code> daemon.                                                                                                                                                       | Remote <code>lpr</code> jobs are sent by the <code>qdaemon</code> backend <code>rembak</code> and received by the <code>lpd</code> daemon. The <code>lpd</code> queues local jobs by calling <code>enq</code> .                                                                                                                                                                                         | In multi-server environments remote jobs are transferred between servers using sockets and RPCs.                                                                                                                                                         |

## Switching between subsystems

The option to change/show current print subsystems has been added to the top level Print Spooling menu in SMIT to allow switching between printing subsystems, as shown in Figure 9-5.

```
Print Spooling

Move cursor to desired item and press Enter.

AIX Print Mode Only:

Start a Print Job
Manage Print Jobs
List All Print Queues
Manage Print Queues
Add a Print Queue
Add an Additional Printer to an Existing Print Queue
Change / Show Print Queue Characteristics
Change / Show Printer Connection Characteristics
Remove a Print Queue
Manage Print Server
Programming Tools

AIX and System V Print Mode:

Change / Show Current Print Subsystem

F1=Help F2=Refresh F3=Cancel F8=Image
F9=Shell F10=Exit Enter=Do
```

Figure 9-5 Change/Show Current Print Subsystem option

The syntax for the command line option to switch between printer subsystems is:

```
switch.prt [-s print_subsystem] [-d]
```

The valid values for *print\_subsystem* are AIX and SystemV. Running the command with the -d flag will display the current print subsystem.

For security reasons, this command is a front-end to the script `/usr/aix/bin/switch.prt.subsystem`, which will do the following when switching from PowerPC to SystemV printing:

- ▶ Ensure that you are not trying to switch to the current system.
- ▶ Check to see if all print jobs are terminated.
- ▶ Stop `qdaemon`, `writesrv`, and `lpd` daemons, and modifies `/etc/inittab` so that they do not restart if the system is rebooted. The `writesrv` daemon allows users to send messages to users on a remote system and receive responses from users on a remote system. The `lpd` daemon is the remote print server.
- ▶ Disable the SMIT menus.
- ▶ Switch the Web-based System Manager plug-ins.

- ▶ Change the lock files from PowerPC to SystemV.
- ▶ Remove AIX links and adds SystemV links for the duplicate commands.
- ▶ Launch `/usr/lib/lp/lpsched` to start the SystemV print subsystem.

From time to time, you may need to remove old print files in order to free up space in `/var` file system. For example, execute the following command to remove print files older than seven days:

```
find /var/spool/lpd/qdir -mtime +7 -type f -exec rm -f{} \;
```

## 9.9 Command summary

The following sections provide the key commands discussed in this chapter.

### 9.9.1 The `lslicense` command

The `lslicense` command displays the number of fixed licenses and the status of the floating licensing. The command has the following syntax:

```
lslicense [-c]
```

### 9.9.2 The `lssrc` command

The `lssrc` command obtains the status of a subsystem, a group of subsystems, or a subserver. The command has the following syntax.

Subsystem status:

```
lssrc [-h Host] { -a | -g GroupName | [-l] -s Subsystem | [-l] -p SubsystemPID }
```

Subserver status:

```
lssrc [-h Host] [-l] -t Type [-p SubsystemPID] [-o Object] [-P SubserverPID]
```

Table 9-3 provides a list of commonly used flags and their descriptions.

*Table 9-3 Commonly used flags of the `lssrc` command*

| Flag            | Description                                                                                                                                                               |
|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -a              | Lists the current status of all defined subsystems.                                                                                                                       |
| -g <i>Group</i> | Specifies a group of subsystems to get the status for. The command is unsuccessful if the <code>GroupName</code> variable is not contained in the subsystem object class. |

| Flag                | Description                                                                                                                                                                                                                                       |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -s <i>Subsystem</i> | Specifies a subsystem to get the status for. The subsystem variable can be the actual subsystem name or the synonym name for the subsystem. The command is unsuccessful if the subsystem variable is not contained in the subsystem object class. |
| -t <i>Type</i>      | Requests that a subsystem send the current status of a subserver. The command is unsuccessful if the subserver type variable is not contained in the subserver object class.                                                                      |

### 9.9.3 The startsrc command

The **startsrc** command starts a subsystem, a group of subsystems, or a subserver. The command has the following syntax.

For subsystem:

```
startsrc [-a Argument] [-e Environment] [-h Host] {-s Subsystem | -g Group}
```

For subserver:

```
startsrc [-h Host] -t Type [-o Object] [-p SubsystemPID]
```

Table 9-4 provides a list of commonly used flags and their descriptions.

Table 9-4 Commonly used flags of the startsrc command

| Flag                | Description                                                                                                                                                                                                             |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -s <i>Subsystem</i> | Specifies a subsystem to be started. The subsystem can be the actual subsystem name or the synonym name for the subsystem. The command is unsuccessful if the subsystem is not contained in the subsystem object class. |
| -t <i>Type</i>      | Specifies that a subserver is to be started. The command is unsuccessful if type is not contained in the subserver object class.                                                                                        |

### 9.9.4 The trace command

The **trace** command records selected system events. The command has the following syntax:

```
trace [-a [-g]] [-f | -l] [-b | -B] [-c] [-d] [-h] [-j Event [
,Event]] [-k Event [,Event]] [-m Message] [-n] [-o Name] [-o-]
[-s] [-L Size] [-T Size]startsrc [-a Argument] [-e Environment]
[-h Host] {-s Subsystem | -g Group}
```

Table 9-5 provides a list of commonly used flags and their descriptions.

Table 9-5 Commonly used flags of the trace command

| Flag                                                     | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|----------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -a                                                       | The -a flag runs the trace daemon asynchronously (as a background task). Once trace has been started this way, you can use the <b>trcon</b> , <b>trcoff</b> , and <b>trcstop</b> commands to respectively start tracing, stop tracing, or exit the trace session. These commands are implemented as links to trace.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| -j <i>Event[,Event]</i><br>or<br>-k <i>Event[,Event]</i> | <p>Specifies the user-defined events for which you want to collect (-j) or exclude (-k) trace data. The event list items can be separated by commas, or enclosed in double quotation marks and separated by commas or blanks.</p> <p>Note: The following events are used to determine the PID, the CPU ID, and the exec path name in the <b>trcrpt</b> report:</p> <p>001 TRACE ON<br/>002 TRACE OFF<br/>106 DISPATCH<br/>10C DISPATCH IDLE PROCESS<br/>134 EXEC SYSTEM CALL<br/>139 FORK SYSTEM CALL<br/>465 KTHREAD CREATE</p> <p>If any of these events are missing, the information reported by the <b>trcrpt</b> command will be incomplete. Consequently, when using the -j flag, you should include all these events in the Event list; conversely, when using the -k flag, you should not include these events in the event list.</p> |

## 9.9.5 The trcrpt command

The **trcrpt** command formats a report from the trace log. The command has the following syntax:

```
trcrpt [-c] [-d List] [-e Date] [-h] [-j] [-n Name] [-o File]
[-p List] [-r] [-s Date] [-t File] [-T List] [-v] [-O Options]
[-x] [File]
```

Table 9-6 provides a list of commonly used flags and their descriptions.

Table 9-6 Commonly used flags of the trcrpt command

| Flag           | Description                                                |
|----------------|------------------------------------------------------------|
| -o <i>File</i> | Writes the report to a file instead of to standard output. |

| Flag              | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -O <i>Options</i> | <p>Specifies options that change the content and presentation of the <b>trcrpt</b> command. Arguments to the options must be separated by commas.</p> <p>Examples of options are:</p> <p><b>cpuid=[onloff]</b> - Displays the physical processor number in the trace report. The default value is off.</p> <p><b>endtime=Seconds</b> - Displays trace report data for events recorded before the seconds specified. Seconds can be given in either an integral or rational representation. If this option is used with the <b>starttime</b> option, a specific range can be displayed.</p> <p><b>exec=[onloff]</b> - Displays exec path names in the trace report. The default value is off.</p> <p><b>pid=[onloff]</b> - Displays the process IDs in the trace report. The default value is off.</p> <p><b>svc=[onloff]</b> - Displays the value of the system call in the trace report. The default value is off.</p> <p>For a complete list of options, refer to the manual page of <b>trcrpt</b>.</p> |

## 9.10 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

- Which of the following commands can be used to give more detailed information about a hung process?
  - od**
  - trace**
  - proc**
  - stream**

2. When users try to **telnet** to the system, they get the error message `connection refused`. Which of the following commands should be run to verify that users can **telnet** into the system properly?
- A. `lssrc -g tcp`
  - B. `lssrc -g tcpip`
  - C. `lssrc -s inetd`
  - D. `lssrc -t telnet`
3. When starting a system, a user does not get a login prompt at the console. However, the user can get a login using **telnet** from another machine. All of the following are probable solutions for this problem except:
- A. `chdev`
  - B. `chcons -a 'login=enable'`
  - C. Edit `/etc/inittab`.
  - D. Reconfigure the network.
4. While attempting to log in, the following message was received:
- ```
'All available login sessions are in use.'
```
- Which of the following procedures should be performed first?
- A. Check `/etc/security`.
 - B. Check `/etc/password`.
 - C. Increase the number of AIX license users.
 - D. Reboot the system into service mode and run `fsck`.
5. A customer is setting up the system as an Oracle DB server. While attempting to start all of the applications, the following error is received:
- ```
User getting 0403-030 Fork function failed, too many processes already exist.
```
- Which of the following commands should be used to correct the problem?
- A. `chps -s'13' paging00`
  - B. `lsattr -El sys0 I grep maxuproc`
  - C. `chdev -l sys0 -a maxuproc=600`
  - D. `mkps -s'20' -n'' -a'' oraclevg hdisk1`

6. A user has the ability to ping the IP address of the printer, but cannot print to the printer. The print queue indicates that everything is fine. Which of the following commands should be performed to determine how to fix the print problem?
  - A. Run `lpstat`.
  - B. Check `enq -a`.
  - C. Check `/etc/hosts`.
  - D. Check `/etc/qconfig`.
7. The system has just been migrated from AIX Version 4.3.3 to AIX 5L Version 5.1. What is the default printing subsystem after the migration?
  - A. AIX
  - B. System V
  - C. SCO UNIXWare 7
  - D. BSD
8. Operator-attention messages from print commands are routed through which subsystem daemon?
  - A. `qdaemon`
  - B. `lpd`
  - C. `errdemon`
  - D. `writesrv`

### 9.10.1 Answers

The following are the preferred answers to the questions provided in this section.

1. B
2. D
3. D
4. C
5. C
6. C
7. A
8. D

### 9.11 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

1. Verify the number of licenses available on your system.
2. List the AIX kernel parameters on your system by using the `lsattr` command on device `sys0`.
3. Perform a trace on your system to see what your system is actually doing. Limit the trace to only a few seconds.
4. Generate a report of the trace performed in the previous step, adding the option for showing the process ID in the report.





## Performance problem determination

In this chapter the following topics are covered:

- ▶ Performance tuning flowchart
- ▶ Tools

Performance tuning issues, from a problem determination perspective, are concentrated around the skills of interpreting output from various commands. For a well-structured approach to such problems, most problem solvers work according to the flowchart shown in Figure 10-1 on page 234.

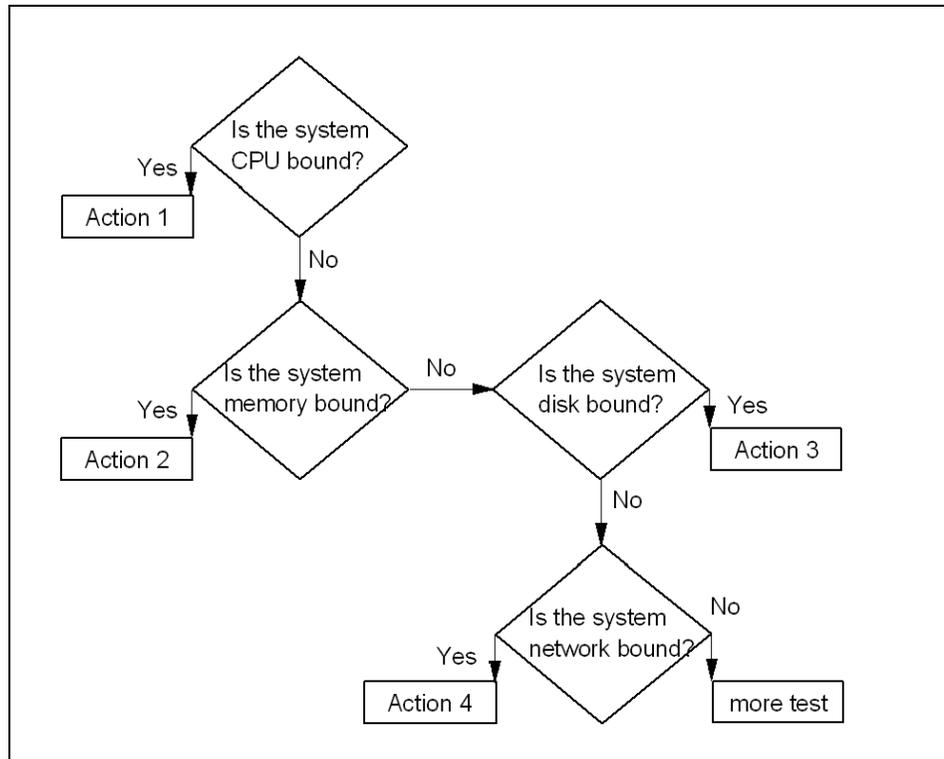


Figure 10-1 General performance tuning flowchart

When investigating a performance problem, CPU constraint is probably the easiest to find. That is why most performance analysts start with checking for CPU constraints.

## 10.1 CPU-bound system

CPU performance problems can be handled in different ways. For example:

- ▶ Reschedule tasks to a less active time of the day or week.
- ▶ Change the priority of processes.
- ▶ Manipulate the scheduler to prioritize foreground processes.
- ▶ Implement Workload Manager.
- ▶ Buy more CPU power.

Whatever the solution finally will be, the way to the solution is usually the same: Identify the process (or groups of processes) that constrains the CPU. When working with CPU performance tuning problems, historical performance information for comparison reasons is useful, if such is available. A very useful tool for this task is the `sar` command.

### 10.1.1 The `sar` command

The `sar` command gathers statistical data about the system. Though it can be used to gather useful data regarding system performance, the `sar` command can increase the system load, which will worsen a pre-existing performance problem. The system maintains a series of system activity counters, which record various activities, such as fork rates. The `sar` command does not cause these counters to be updated or used; this is done automatically, regardless of whether the `sar` command runs. It merely extracts the data in the counters and saves it, based on the sampling rate specified to the `sar` command. There are three situations in which to use the `sar` command; they are discussed in the following sections.

#### Real-time sampling and display

To collect and display system statistic reports immediately, use the following command:

```
sar -u 2 5
AIX texmex 3 4 000691854C00 01/27/00
17:58:15 %usr %sys %wio %idle
17:58:17 43 9 1 46
17:58:19 35 17 3 45
17:58:21 36 22 20 23
17:58:23 21 17 0 63
17:58:25 85 12 3 0
Average 44 15 5 35
```

This example is from a single user workstation and shows the CPU utilization.

## Display previously captured data

The `-o` and `-f` options (write and read to or from user given data files) allow you to visualize the behavior of your machine in two independent steps. This consumes less resources during the problem-reproduction period. You can move the binary to another machine, because the binary file contains all data the `sar` command needs.

```
sar -o /tmp/sar.out 2 5 > /dev/null
```

The previous command runs the `sar` command in the background, collects system activity data at two-second intervals for five intervals, and stores the (unformatted) `sar` data in the `/tmp/sar.out` file. The redirection of standard output is used to avoid a screen output.

The following command extracts CPU information from the file and outputs a formatted report to standard output:

```
sar -f/tmp/sar.out
AIX texmex 3 4 000691854C00 01/27/00
18:10:18 %usr %sys %wio %idle
18:10:20 9 2 0 88
18:10:22 13 10 0 76
18:10:24 37 4 0 59
18:10:26 8 2 0 90
18:10:28 20 3 0 77
Average 18 4 0 78
```

The captured binary data file keeps all information needed for the reports. Every possible `sar` report could, therefore, be investigated.

## System activity accounting and the cron daemon

The `sar` command calls a process named `sadc` to access system data. Two shell scripts (`/usr/lib/sa/sa1` and `/usr/lib/sa/sa2`) are structured to be run by the cron daemon and provide daily statistics and reports. Sample stanzas are included (but commented out) in the `/var/spool/cron/crontabs/adm` crontab file to specify when the cron daemon should run the shell scripts.

The following lines show a modified crontab for the `adm` user. Only the comment characters for the data collections were removed:

```
#####
SYSTEM ACTIVITY REPORTS
8am-5pm activity reports every 20 mins during weekdays.
activity reports every an hour on Saturday and Sunday.
6pm-7am activity reports every an hour during weekdays.
Daily summary prepared at 18:05.
#####
0 8-17 * * 1-5 /usr/lib/sa/sa1 1200 3 &
```

```

0 * * * 0,6 /usr/lib/sa/sa1 &
0 18-7 * * 1-5 /usr/lib/sa/sa1 &
5 18 * * 1-5 /usr/lib/sa/sa2 -s 8:00 -e 18:01 -i 3600 -ubcwyqvm &
#=====

```

Collection of data in this manner is useful to characterize system usage over a period of time and to determine peak usage hours.

Another useful feature of the `sar` command is that the output can be specific for the usage of each processor in a multiprocessor environment, as seen in the following output. The last line is an average output:

```

sar -P ALL 2 1

AIX client1 3 4 000BC6DD4C00 07/06/00

14:46:52 cpu %usr %sys %wio %idle
14:46:54 0 0 0 0 100
 1 0 1 0 99
 2 0 0 0 100
 3 0 0 0 100
 - 0 0 0 100

```

In general, if `%usr` plus `%sys` is consistently over 80 percent, then the system is CPU bound.

## System calls

The `sar -c` command reports system calls.

```

sar -c 1 10

AIX server2 3 4 000FA17D4C00 06/30/00

09:33:04 scall/s sread/s swrit/s fork/s exec/s rchar/s wchar/s
09:33:05 1050 279 118 0.00 0.00 911220 5376749
09:33:06 186 19 74 0.00 0.00 3272 3226417
09:33:07 221 19 79 0.00 0.00 3272 3277806
09:33:08 2996 132 400 0.00 0.00 314800 2284933
09:33:09 3304 237 294 0.00 0.00 167733 848174
09:33:10 4186 282 391 0.00 0.00 228196 509414
09:33:11 1938 109 182 1.00 1.00 153703 1297872
09:33:12 3263 179 303 0.00 0.00 242048 1003364
09:33:13 2751 172 258 0.00 0.00 155082 693801
09:33:14 2827 187 285 0.00 0.00 174059 1155239

```

## 10.1.2 The vmstat command

The `vmstat` command reports statistics about kernel threads, virtual memory, disks, traps, and CPU activity. Reports generated by the `vmstat` command can be used to balance system load activity. These system-wide statistics (among all processors) are calculated as averages for values expressed as percentages, and as sums otherwise. For a CPU point of view, the highlighted two left-hand columns and the four highlighted right-hand columns provide useful data, as shown in the following example. The columns are discussed in the following sections.

```
vmstat 2
kthr memory page faults cpu

 r b avm fre re pi po fr sr cy in sy cs us sy id wa
 0 0 16998 14612 0 0 0 0 0 0 101 10 8 55 0 44 0
 0 1 16998 14611 0 0 0 0 0 0 411 2199 54 0 0 99 0
 0 1 16784 14850 0 0 0 0 0 0 412 120 51 0 0 99 0
 0 1 16784 14850 0 0 0 0 0 0 412 88 50 0 0 99 0
```

### The kthr columns

The kthr columns show how kernel threads are placed on various queues per second over the sampling interval.

#### *The r column*

The r column shows the average number of kernel threads waiting on the run queue per second. This field indicates the number of threads that can be run. This value should be less than five for non-SMP systems. For SMP systems, this value should be less than:

$$5 \times (Ntotal - Nbind)$$

Where *Ntotal* stands for the total number of processors, and *Nbind* for the number of processors that have been bound to processes, for example, with the `bindprocessor` command.

If this number increases rapidly, examine the applications. However, some systems may be running normally with 10 to 15 threads on their run queue, depending on the thread tasks and the amount of time they run.

#### *The b column*

The b column shows the average number of kernel threads in the wait queue per second. These threads are waiting for resources or I/O. Threads are also located in the wait queue when waiting for one of their thread pages to be paged in. This value is usually near zero. But, if the run-queue value increases, the wait queue normally also increases. If processes are suspended due to memory load

control, the blocked column (b) in the `vmstat` command report indicates the increase in the number of threads rather than the run queue.

### **The cpu columns**

The four right-hand columns are a breakdown, in percentage of CPU time used, of user threads, system threads, CPU idle time (running the wait process), and CPU idle time when the system had outstanding disk or NFS I/O requests.

#### ***The us column***

The `us` column shows the percent of CPU time spent in user mode. A UNIX process can execute in either user mode or system (kernel) mode. When in user mode, a process executes within its application code and does not require kernel resources to perform computations, manage memory, or set variables.

#### ***The sy column***

The `sy` column details the percentage of time the CPU was executing a process in system mode. This includes CPU resource consumed by kernel processes (`kprocs`) and others that need access to kernel resources. If a process needs kernel resources, it must execute a system call, and is thereby switched to system mode to make that resource available. For example, reading or writing of a file requires kernel resources to open the file, seek a specific location, and read or write data, unless memory mapped files are used.

#### ***The id column***

The `id` column shows the percentage of time that the CPU is idle, or waiting, without pending local disk I/O. If there are no processes available for execution (the run queue is empty), the system dispatches a process called `wait`. On an SMP system, one `wait` process per processor can be dispatched. On a uniprocessor system, the process ID (PID) is usually 516. SMP systems will have an idle `kproc` for each processor. If the `ps` command report shows a high aggregate time for this process, it means there were significant periods of time when no other process was ready to run or waiting to be executed on the CPU. The system was therefore mostly idle and waiting for new tasks.

If there are no I/Os pending to a local disk, all time charged to `wait` is classified as idle time. In AIX Version 4.3.2 and earlier, an access to remote disks (NFS-mounted disks) is treated as idle time (with a small amount of `sy` time to execute the NFS requests) because there is no pending I/O request to a local disk. With AIX Version 4.3.3 and later, NFS goes through the buffer cache, and waits in those routines are accounted for in the `wa` statistics.

#### ***The wa column***

The `wa` column details the percentage of time the CPU was idle with pending local disk I/O (this is also true for NFS-mounted disks in AIX Version 4.3.3 and

later). The method used in AIX Version 4.3.2 and earlier versions of the operating system can, under certain circumstances, give an inflated view of wa time on SMPs. In AIX Version 4.3.2 and earlier, at each clock interrupt on each processor (100 times a second per processor), a determination is made as to in which of the four categories (usr/sys/wio/idle) to place the last 10 ms of time. If any disk I/O is in progress, the wa category is incremented. For example, systems with just one thread doing I/O could report over 90 percent wa time regardless of the number of CPUs it has.

The change in AIX Version 4.3.3 is to only mark an idle CPU as wa if an outstanding I/O was started on that CPU. This method can report much lower wa times when just a few threads are doing I/O and the system is otherwise idle. For example, a system with four CPUs and one thread doing I/O will report a maximum of 25 percent wa time. A system with 12 CPUs and one thread doing I/O will report a maximum of 8.3 percent wa time.

Also, NFS now goes through the buffer cache, and waits in those routines are accounted for in the wa statistics.

A wa value over 25 percent could indicate that the disk subsystem might not be balanced properly, or it might be the result of a disk-intensive workload.

## **The faults columns**

It may also be worthwhile to look at the faults columns, which provide information about process control, such as trap and interrupt rate.

### ***The in column***

In the in column is the number of device interrupts per second observed in the interval.

### ***The sy column***

In the sy column is the number of system calls per second observed in the interval. Resources are available to user processes through well-defined system calls. These calls instruct the kernel to perform operations for the calling process and exchange data between the kernel and the process. Because workloads and applications vary widely, and different calls perform different functions, it is impossible to define how many system calls per-second are too many. But typically, when the sy column raises over 10000 calls per second on a uniprocessor, further investigation is called for (on a SMP system, the number is 10000 calls per second per processor). One reason for this high number of calls per second could be *polling* subroutines like select(). For this column, it is advisable to have a baseline measurement that gives a count for a normal sy value.

### ***The cs column***

The `cs` column shows the number of context switches per second observed in the interval. The physical CPU resource is subdivided into logical time slices of 10 milliseconds each. Assuming a thread is scheduled for execution, it will run until its time slice expires, until it is preempted, or until it voluntarily gives up control of the CPU. When another thread is given control of the CPU, the context or working environment of the previous thread must be saved and the context of the current thread must be loaded. The operating system has a very efficient context switching procedure, so each switch is inexpensive in terms of resources. Any significant increase in context switches, such as when `cs` is a lot higher than the disk I/O and network packet rate, should be cause for further investigation.

If the system has low performance because of a lot of threads on the run queue or threads waiting for I/O, then `ps` command output will be useful in determining which process has used the most CPU resources.

## **10.1.3 The ps command**

The `ps` command is a flexible tool for identifying the programs that are running on the system and the resources they are using. It displays statistics and status information about processes on the system, such as process or thread ID, I/O activity, and CPU and memory utilization.

### **The ps command output and CPU monitoring**

Three of the possible `ps` command output columns report CPU usage, each in a different way, as provided in Table 10-1.

*Table 10-1 CPU-related ps output*

| <b>Column</b> | <b>Value</b>                                                                                                                                                        |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| C             | Recent CPU time used for a process.                                                                                                                                 |
| TIME          | Total CPU time used by the process since it started.                                                                                                                |
| %CPU          | Total CPU time used by the process since it started, divided by the elapsed time since the process started. This is a measure of the CPU dependence of the program. |

### ***The C column***

The `C` column can be generated by the `-l` flag and the `-f` flag. In this column, the CPU utilization of processes or threads is reported. The value is incremented each time the system clock ticks and the process or thread is found to be running. Therefore, it also can be said to be a process penalty for recent CPU usage. The value is decayed by the scheduler by dividing it by two once per

second. Large values indicate a CPU-intensive process and result in a lower process priority, while small values indicate an I/O-intensive process and result in a more favorable priority. In the following example, `tctestprog` is running, which is a CPU-intensive program. The `vmstat` command output shows that about 25 percent of the CPU is used by `usr` processes.

```
vmstat 2 3
kthr memory page faults cpu

 r b avm fre re pi po fr sr cy in sy cs us sy id wa
 0 0 26468 51691 0 0 0 0 0 0 100 91 6 47 0 53 0
 1 1 26468 51691 0 0 0 0 0 0 415 35918 237 26 2 71 0
 1 1 26468 51691 0 0 0 0 0 0 405 70 26 25 0 75 0
```

The `ps` command is useful in this situation. The following formatting sorts the output so that the third column has the biggest value at the top, and shows only five lines from the total output.

```
ps -ef | sort +3 -r | head -n 5
 UID PID PPID C STIME TTY TIME CMD
 root 22656 27028 101 15:18:31 pts/11 7:43 ./tctestprog
 root 14718 24618 5 15:26:15 pts/17 0:00 ps -ef
 root 4170 1 3 Jun 15 - 12:00 /usr/sbin/syncd 60
 root 21442 24618 2 15:26:15 pts/17 0:00 sort +3 -r
```

From the previous example, you can tell that `tctestprog` is the process with the most-used CPU in recent history.

### **The TIME column**

The second value mentioned is the `TIME` value. This value is generated with all flags, and it shows the total execution time for the process. This calculation does not take into account when the process was started, as seen in the following output. The same test program is used again, and event, though the `C` column shows that the process gets a lot of CPU time, is not yet on top in the `TIME` column:

```
ps -ef | sort +3 -r | head -n 5
 UID PID PPID C STIME TTY TIME CMD
 root 18802 27028 120 15:40:28 pts/11 1:10 ./tctestprog
 root 9298 24618 3 15:41:38 pts/17 0:00 ps -ef
 root 15782 24618 2 15:41:38 pts/17 0:00 head -n 5
 root 24618 26172 2 Jun 21 pts/17 0:03 ksh

ps -e | head -n 1 ; ps -e|egrep -v "TIME|0:"|sort +2b -3 -n -r|head -n 10
 PID TTY TIME CMD
 4170 - 12:01 syncd
 4460 - 2:07 X
 3398 - 1:48 dtssession
 18802 pts/11 1:14 tctestprog
```

The `syncd`, `X`, and `dtsession` are all processes that have been active since IPL; that is why they have accumulated more total TIME than the test program.

### The %CPU column

The %CPU column, generated by the `-u` or `-v` flag, shows the percentage of time that the process has used the CPU since the process started. The value is computed by dividing the time the process uses the CPU by the elapsed time of the process. In a multi-processor environment, the value is further divided by the number of available CPUs, since several threads in the same process can run on different CPUs at the same time. Because the time base over which this data is computed varies, the sum of all %CPU fields can exceed 100 percent. In the following example, there are two ways to sort the extracted output from a system. The first example includes `kprocs`, for example, PID 516, which is a wait process. The other, more complex command syntax excludes such `kprocs`:

```
ps auxwww |head -n 5
USER PID %CPU %MEM SZ RSS TTY STAT STIME TIME COMMAND
root 18802 25.0 1.0 4140 4160 pts/11 A 15:40:28 5:44 ./tctestprog
root 516 25.0 5.0 8 15136 - A Jun 15 17246:34 kproc
root 774 20.6 5.0 8 15136 - A Jun 15 14210:30 kproc
root 1290 5.9 5.0 8 15136 - A Jun 15 4077:38 kproc
```

```
ps gu|head -n1; ps gu|egrep -v "CPU|kproc"|sort +2b -3 -n -r |head -n 5
USER PID %CPU %MEM SZ RSS TTY STAT STIME TIME COMMAND
root 18802 25.0 1.0 4140 4160 pts/11 A 15:40:28 7:11
./tctestprog
imnadm 12900 0.0 0.0 264 332 - A Jun 15 0:00 /usr/IMNSearch/ht
root 0 0.0 5.0 12 15140 - A Jun 15 4:11 swapper
root 1 0.0 0.0 692 764 - A Jun 15 0:28 /etc/init
root 3398 0.0 1.0 1692 2032 - A Jun 15 1:48 /usr/dt/bin/dtses
```

From the output, you can see that the test program, `tctestprog`, uses about 25 percent of available CPU resources since the process started.

Upon finding a run-away process, the next step in the analysis is to find out what exactly in the process uses the CPU. For this, a profiler is needed. The AIX profiler of preference is `tprof`.

## 10.1.4 The `tprof` command

The `tprof` command can be used for application tuning and for information collection of overall CPU utilization. The `tprof` command can be run over a time period to trace the activity of the CPU.

In the AIX operating system, an interrupt occurs periodically to allow a *housekeeping* kernel routine to run. This occurs 100 times per second. When the `tprof` command is invoked, it counts every such kernel interrupt as a *tick*. This

kernel routine records the process ID and the address of the instruction executing when the interrupt occurred and this information is used by the **tprof** command. The **tprof** command also records whether the process counter is in the kernel address space, the user address space, or the shared library address space.

In AIX 5L Version 5.1, the **tprof** command has been enhanced to do subroutine or method level profiling for Java applications. The Java Virtual Machine Profiling Interface (JVMPi), a new feature supported by Java 1.2 or later, has been enhanced to do class and method level profiling for Java applications. The `-j` flag was added to **tprof** to enable profiling for Java applications.

### The tprof summary CPU utilization report

A summary ASCII report with the suffix `.all` is always produced. If no program is specified, the report is named `__prof.all`. If a program is specified, the report is named `__program.all`. This report contains an estimate of the amount of CPU time spent in each process that was executing while the **tprof** command was monitoring the system. This report also contains an estimate of the amount of CPU time spent in each of the three address spaces.

The files containing the reports are left in the working directory. All files created by the **tprof** command are prefixed by two underscores (`__`).

In the following example, a generic report generated an output of:

```
tprof -x sleep 30
Starting Trace now
Starting sleep 30
Wed Jun 28 14:58:58 2000
System: AIX server3 Node: 4 Machine: 000BC6DD4C00

Trace is done now
30.907 secs in measured interval
* Samples from __trc_rpt2
* Reached second section of __trc_rpt2
```

In this case, `sleep 30` points out to the **tprof** command to run for 30 seconds.

### The total column

The total column in `__prof.all` is of interest. The first section indicates the use of ticks on a per-process basis.

| Process   | PID   | TID   | Total       | Kernel | User  | Shared | Other |
|-----------|-------|-------|-------------|--------|-------|--------|-------|
| =====     | ===   | ===   | =====       | =====  | ===== | =====  | ===== |
| wait      | 516   | 517   | <b>3237</b> | 3237   | 0     | 0      | 0     |
| tctestprg | 14746 | 13783 | <b>3207</b> | 1      | 3206  | 0      | 0     |
| tctestprg | 13730 | 17293 | <b>3195</b> | 0      | 3195  | 0      | 0     |

|              |       |       |              |       |       |       |       |
|--------------|-------|-------|--------------|-------|-------|-------|-------|
| wait         | 1032  | 1033  | <b>3105</b>  | 3105  | 0     | 0     |       |
| wait         | 1290  | 1291  | <b>138</b>   | 138   | 0     | 0     | 0     |
| swapper      | 0     | 3     | <b>10</b>    | 7     | 3     | 0     | 0     |
| tprof        | 14156 | 5443  | <b>6</b>     | 3     | 3     | 0     | 0     |
| trace        | 16000 | 14269 | <b>3</b>     | 3     | 0     | 0     | 0     |
| syncd        | 3158  | 4735  | <b>2</b>     | 2     | 0     | 0     | 0     |
| tprof        | 5236  | 16061 | <b>2</b>     | 2     | 0     | 0     | 0     |
| gil          | 2064  | 2839  | <b>1</b>     | 1     | 0     | 0     | 0     |
| gil          | 2064  | 3097  | <b>1</b>     | 1     | 0     | 0     |       |
| trace        | 15536 | 14847 | <b>1</b>     | 1     | 0     | 0     | 0     |
| sh           | 14002 | 16905 | <b>1</b>     | 1     | 0     | 0     | 0     |
| sleep        | 14002 | 16905 | <b>1</b>     | 1     | 0     | 0     | 0     |
| =====        | ===   | ===   | =====        | ===== | ===== | ===== | ===== |
| <b>Total</b> |       |       | <b>12910</b> | 6503  | 6407  | 0     | 0     |

Each tick is 1/100 second. You can calculate the total amount of available ticks; for example, 30 seconds, times 100 ticks, make a total of 3000 ticks. This is according to theory, but when looking at the output, there are over 12000 total ticks. This is because the test system is a 4-way F50, so the available ticks are calculated in the following way:

Time (in seconds) x Number of available CPUs x 100

### ***The user column***

If the user column shows high values, application tuning might be necessary. In the output, you can see that both `tctestprgs` used about 3200 ticks, which is around 25 percent of the total number of available ticks. This is confirmed with `ps auxwww` output:

```
ps auxwww
USER PID %CPU %MEM SZ RSS TTY STAT STIME TIME COMMAND
root 14020 25.0 0.0 300 320 pts/1 A 15:23:55 16:45 ./tctestprg
root 12280 25.0 0.0 300 320 pts/1 A 15:23:57 16:43 ./tctestprg
```

### ***The freq column***

The second section of the report includes the `freq` column, which has the total amount of ticks used by a specified type of process. Here the ticks used by all three `wait` processes are added together, and the two `tctestprgs` are added together. The total workload produced by one type of process is shown below (as well as the number of instances of the processes that are running):

| Process   | <b>FREQ</b> | Total | Kernel | User  | Shared | Other |
|-----------|-------------|-------|--------|-------|--------|-------|
| =====     | ===         | ===== | =====  | ===== | =====  | ===== |
| wait      | <b>3</b>    | 6480  | 6480   | 0     | 0      | 0     |
| tctestprg | <b>2</b>    | 6402  | 1      | 6401  | 0      | 0     |
| swapper   | <b>1</b>    | 10    | 7      | 3     | 0      | 0     |
| tprof     | <b>2</b>    | 8     | 5      | 3     | 0      | 0     |
| trace     | <b>2</b>    | 4     | 4      | 0     | 0      | 0     |
| gil       | <b>2</b>    | 2     | 2      | 0     | 0      | 0     |

|       |     |       |       |       |       |       |
|-------|-----|-------|-------|-------|-------|-------|
| syncd | 1   | 2     | 2     | 0     | 0     | 0     |
| sh    | 1   | 1     | 1     | 0     | 0     | 0     |
| sleep | 1   | 1     | 1     | 0     | 0     | 0     |
| ===== | === | ===== | ===== | ===== | ===== | ===== |
| Total | 15  | 12910 | 6503  | 6407  | 0     | 0     |

## 10.2 Memory-bound system

Memory in AIX is handled by the Virtual Memory Manager (VMM). The VMM provides a method by which real memory appears larger than its true size. The virtual memory system is composed of real memory plus physical disk space, where portions of a memory region that are not currently in use are stored.

VMM maintains a list of free page frames that are used to accommodate pages that must be brought into memory. In memory-constrained environments, the VMM must occasionally replenish the free list by moving some of the current data from real memory. This is called page stealing. A page fault is a request to load a 4-KB data page from disk. A number of places are searched in order to find data.

First the data and instruction caches are searched. Next, the Translation Lookaside Buffer (TLB) is searched; this is an index of recently used virtual addresses with their page frame IDs. If the data is not in the TLB, the Page Frame Table (PTF) is consulted; this is an index for all real memory pages, and it is held in pinned memory. Since the table is large, there are indexes to this index. The Hash Anchor Table (HAT) links pages of related segments to get a faster entry point to the main PTF.

From the page stealer perspective, the memory is divided into computational memory and file memory. The page stealer tries to balance these two types of memory usage when stealing pages.

- ▶ Computational memory consists of pages that belong to the working segment or program text segment.
- ▶ File memory consists of the remaining pages. These are usually pages from the permanent data file in persistent memory.

When starting a process, a slot is assigned. When a process references a virtual memory page that is on the disk, the referenced page must be paged in, and probably one or more pages must be paged out, creating I/O traffic and delaying the startup of the process. AIX attempts to steal real memory pages that are unlikely to be referenced in the near future, using the page replacement algorithm. The page replacement algorithm can be manipulated.

If the system has too little memory, no RAM pages are good candidates to be paged out, as they will be reused in the near future. When this happens, continuous pagein and pageout occurs. This condition is called thrashing.

The `vmstat` command can help you recognize memory-bound systems.

## 10.2.1 The `vmstat` command

The `vmstat` command summarizes the total active virtual memory used by all of the processes in the system, as well as the number of real-memory page frames on the free list. Active virtual memory is defined as the number of virtual-memory working segment pages that have actually been touched. This number can be larger than the number of real page frames in the machine, because some of the active virtual-memory pages may have been written out to paging space.

When determining if a system might be short on memory or if some memory tuning needs to be done, run the `vmstat` command over a set interval and examine the `pi` and `po` columns on the resulting report. These columns indicate the number of paging space page-ins per second and the number of paging space page-outs per second, respectively. If the values are constantly non-zero, there might be a memory bottleneck. Having occasional non-zero values is not a concern, because paging is the main activity of virtual memory.

```
vmstat 2 10
kthr memory page faults cpu
----- -
r b avm fre re pi po fr sr cy in sy cs us sy id wa
1 3 113726 124 0 14 6 151 600 0 521 5533 816 23 13 7 57
0 3 113643 346 0 2 14 208 690 0 585 2201 866 16 9 2 73
0 3 113659 135 0 2 2 108 323 0 516 1563 797 25 7 2 66
0 2 113661 122 0 3 2 120 375 0 527 1622 871 13 7 2 79
0 3 113662 128 0 10 3 134 432 0 644 1434 948 22 7 4 67
1 5 113858 238 0 35 1 146 422 0 599 5103 903 40 16 0 44
0 3 113969 127 0 5 10 153 529 0 565 2006 823 19 8 3 70
0 3 113983 125 0 33 5 153 424 0 559 2165 921 25 8 4 63
0 3 113682 121 0 20 9 154 470 0 608 1569 1007 15 8 0 77
0 4 113701 124 0 3 29 228 635 0 674 1730 1086 18 9 0 73
```

Notice the high I/O wait and the number of threads on the blocked queue. Most likely, the I/O wait is due to the paging in/out from paging space.

To determine if the system has performance problems with its VMM, examine the columns under `memory` and `page`.

## The memory columns

The memory columns provide information about the real and virtual memory. Under memory for the **vmstat** command are two additional columns, avm and fre.

### *The avm column*

The avm (Active Virtual Memory) column gives the average number of 4-KB pages that are allocated to paging space. The avm value can be used to calculate the amount of paging space assigned to executing processes.

**Note:** The **vmstat** command (avm column), **ps** command (SIZE, SZ), and other utilities report the amount of virtual memory actually accessed, but with delayed page space algorithm, the paging space may not get touched. The **svmon** command (up through AIX Version 4.3.2) shows the amount of paging space being used, so this value may be much smaller than the avm value of the **vmstat** command.

For more information on DPSA, see the *IBM @server Certification Study Guide - AIX 5L Performance and System Tuning*, SG24-6184.

The number in the avm field divided by 256 will yield the approximate number of megabytes (MB) allocated to the paging space system wide. Prior to AIX Version 4.3.2, the same information is reflected in the Percent Used column of the **1sps -s** command output or with the **svmon -G** command under the page space inuse field, pg.

### *The fre column*

The fre column shows the average number of free memory pages. A page is a 4-KB area of real memory. The system maintains a buffer of memory pages, called the free list, that will be readily accessible when the VMM needs space. The minimum number of pages that the VMM keeps on the free list is determined by the minfree parameter of the **vmtune** command. When an application terminates, all of its working pages are immediately returned to the free list. Its persistent pages (files), however, remain in RAM and are not added back to the free list until they are stolen by the VMM for other programs. Persistent pages are also freed if the corresponding file is deleted.

For this reason, the fre value may not indicate all the real memory that can be readily available for use by processes. If a page frame is needed, then persistent pages related to terminated applications are among the first to be handed over to another program.

## The page columns

The page columns show information about page faults and paging activity. These are averaged over the interval and given in units per second.

### ***The pi column***

The pi column details the number (rate) of pages paged in from paging space. Paging space is the part of virtual memory that resides on disk. It is used as an overflow when memory is over committed. Paging space consists of logical volumes dedicated to the storage of working set pages that have been stolen from real memory. When a stolen page is referenced by the process, a page fault occurs, and the page must be read into memory from paging space.

Due to the variety of configurations of hardware, software, and applications, there is no absolute number to look out for, but five page-ins per second per paging space should be the upper limit. This guideline should not be rigidly adhered to, but used as a reference. This field is important as a key indicator of paging-space activity. If a page-in occurs, there must have been a previous page-out for that page. It is also likely, in a memory-constrained environment, that each page-in will force a different page to be stolen and, therefore, paged out. But systems could also work fine when they have close to 10 pi per second for one minute and then work without any page-ins.

### ***The po column***

The po column shows the number (rate) of pages paged out to paging space. Whenever a page of working storage is stolen, it is written to paging space (if it does not yet reside in paging space or if it was modified). If not referenced again, it will remain on the paging device until the process terminates or disclaims the space. Subsequent references to addresses contained within the faulted-out pages results in page faults, and the pages are paged in individually by the system. When a process terminates normally, any paging space allocated to that process is freed. If the system is reading in a significant number of persistent pages (files), you might see an increase in po without corresponding increases in pi. This does not necessarily indicate thrashing, but may warrant investigation into data-access patterns of the applications.

### ***The fr column***

The fr column shows the number of pages that were freed per second by the page-replacement algorithm during the interval. As the VMM page-replacement routine scans the Page Frame Table (PFT), it uses criteria to select which pages are to be stolen to replenish the free list of available memory frames. The criteria includes both kinds of pages, working (computational) and file (persistent) pages. Just because a page has been freed, it does not mean that any I/O has taken place. For example, if a persistent storage (file) page has not been modified, it will not be written back to the disk. If I/O is not necessary, minimal system resources are required to free a page.

If the ratio of po to fr is greater than 1 to 6, this could indicate a thrashing system.

### ***The sr column***

The `sr` column shows the number of pages that were examined per second by the page-replacement algorithm during the interval. The VMM page-replacement code scans the PFT and steals pages until the number of frames on the free list is at least the `maxfree` value. The page-replacement code might have to scan many entries in the PFT before it can steal enough to satisfy the free list requirements. With stable, unfragmented memory, the scan rate and free rate might be nearly equal. On systems with multiple processes using many different pages, the pages are more volatile and disjointed. In this scenario, the scan rate might greatly exceed the free rate.

Memory is over committed when the ratio of `fr` to `sr` (`fr:sr`) is high. A `fr:sr` ratio of 1:4 means that for every page freed, four pages had to be examined. It is difficult to determine a memory constraint based on this ratio alone, and what constitutes a high ratio is workload/application dependent.

### ***The cy column***

The `cy` column shows the number of cycles per second of the clock algorithm. The VMM uses a technique known as the clock algorithm to select pages to be replaced. This technique takes advantage of a referenced bit for each page as an indication of what pages have been recently used (referenced). When the page-stealer routine is called, it cycles through the PFT, examining each page's referenced bit. The `cy` column shows how many times per second the page-replacement code has scanned the PFT. Because the free list can be replenished without a complete scan of the PFT, and because all of the `vmstat` command fields are reported as integers, this field is usually zero. If not, it indicates a complete scan of the PFT, and the stealer has to scan the PFT again, because `fre` is still under the `maxfree` value.

One way to determine the appropriate amount of RAM for a system is to look at the largest value for `avm` reported by the `vmstat` command. Multiply that by 4 KB to get the number of bytes, and then compare that to the number of bytes of RAM on the system. Ideally, `avm` should be smaller than the total RAM. If not, some amount of virtual memory paging will occur. How much paging occurs will depend on the difference between the two values. Remember, the idea of virtual memory is that it gives us the capability of addressing more memory than we have (some of the memory is in RAM and the rest is in paging space). If there is far more virtual memory than real memory, this could cause excessive paging, which then results in delays. If `avm` is lower than RAM, then paging-space paging could be caused by RAM being filled up with file pages. In that case, tuning the `minperm` or `maxperm` values could reduce the amount of paging-space paging. This can be done with the `vm tune` command.

Another useful command for memory performance problem determination is the `ps` command.

## 10.2.2 The ps command

The `ps` command is a flexible tool for identifying the programs that are running on the system and the resources they are using. It displays statistics and status information about processes on the system, such as process or thread ID, I/O activity, CPU, and memory utilization.

### The ps command output and memory usage monitoring

The `ps` command gives useful information on memory usage. The most useful output is presented in Table 10-2.

Table 10-2 Memory-related ps output

| Column | Value                                                             |
|--------|-------------------------------------------------------------------|
| SIZE   | The virtual size of the data section of the process in 1-KB units |
| RSS    | The real-memory size of the process in 1-KB units                 |
| %MEM   | The percentage of real memory used by this process                |

#### The SIZE column

The `v` flag generates the `SIZE` column. This is the virtual size (in paging space) in kilobytes of the data section of the process (displayed as `SZ` by other flags). This number is equal to the number of working segment pages of the process that have been touched times four. If some working segment pages are currently paged out, this number is larger than the amount of real memory being used. `SIZE` includes pages in the private segment and the shared-library data segment of the process, as shown in the following example:

```
ps av |sort +5 -r |head -n 5
 PID TTY STAT TIME PGIN SIZE RSS LIM TSIZ TRS %CPU %MEM COMMAND
25298 pts/10 A 0:00 0 2924 12 32768 159 0 0.0 0.0 smitty
13160 1ft0 A 0:00 17 368 72 32768 40 60 0.0 0.0/usr/sbin
27028 pts/11 A 0:00 90 292 416 32768 198 232 0.0 1.0 ksh
24618 pts/17 A 0:04 318 292 408 32768 198 232 0.0 1.0 ksh
```

#### The RSS column

The `v` flag also produces the `RSS` column, as seen in the previous example. This is the real-memory (resident set) size, in kilobytes, of the process. This number is equal to the sum of the number of working segment and code segment pages in memory times four. Remember that code segment pages are shared among all of the currently running instances of the program. If 26 `ksh` processes are running, only one copy of any given page of the `ksh` executable program would be in memory, but the `ps` command would report that code segment size as part of the `RSS` of each instance of the `ksh` program.

If you want to sort to the sixth column, you will get the output ordered using the RSS column, as shown in the following example:

```
ps av |sort +6 -r |head -n 5
PID TTY STAT TIME PGIN SIZE RSS LIM TSIZ TRS %CPU %MEM COMMAND
21720 pts/1 A 0:00 1 288 568 32768 198 232 0.0 1.0 ksh
27028 pts/11 A 0:00 90 292 416 32768 198 232 0.0 1.0 ksh
24618 pts/17 A 0:04 318 292 408 32768 198 232 0.0 1.0 ksh
15698 pts/1 A 0:00 0 196 292 32768 52 60 0.0 0.0 ps av
```

### **The %MEM column**

The %MEM column is generated by the u and v flags. This is calculated as the sum of the number of working segment and code segment pages in memory times four (that is, the RSS value), divided by the size of the real memory of the machine in KB, times 100, rounded to the nearest full percentage point. This value attempts to convey the percentage of real memory being used by the process. Unfortunately, like RSS, it tends to exaggerate the cost of a process that is sharing program text with other processes. Further, the rounding to the nearest percentage point causes all of the processes in the system that have RSS values under .005 times real memory size to have a %MEM of 0.0. For example:

```
ps au |head -n 1; ps au |egrep -v "RSS"|sort +3 -r |head -n 5
USER PID %CPU %MEM SZ RSS TTY STAT STIME TIME COMMAND
root 22750 0.0 21.0 20752 20812 pts/11 A 17:55:51 0:00 ./tctestprog2 root
root 21720 0.0 1.0 484 568 pts/1 A 17:16:14 0:00 ksh
root 25298 0.0 0.0 3080 12 pts/10 A Jun 16 0:00 smitty
root 27028 0.0 0.0 488 416 pts/11 A 14:53:27 0:00 ksh
root 24618 0.0 0.0 488 408 pts/17 A Jun 21 0:04 ksh
```

You can combine all these columns into one output by using the gv flags. For example:

```
ps gv|head -n 1; ps gv|egrep -v "RSS" | sort +6b -7 -n -r |head -n 5
PID TTY STAT TIME PGIN SIZE RSS LIM TSIZ TRS %CPU %MEM COMMAND
15674 pts/11 A 0:01 0 36108 36172 32768 5 24 0.6 24.0 ./tctestp
22742 pts/11 A 0:00 0 20748 20812 32768 5 24 0.0 14.0 ./backups
10256 pts/1 A 0:00 0 15628 15692 32768 5 24 0.0 11.0 ./tctestp
2064 - A 2:13 5 64 6448 xx 0 6392 0.0 4.0 kproc
1806 - A 0:20 0 16 6408 xx 0 6392 0.0 4.0 kproc
```

The columns from the previous output that are described in the following sections are also of interest.

### **The PGIN column**

Number of page-ins caused by page faults. Since all I/O is classified as page faults, this is basically a measure of I/O volume.

### ***The TSIZ column***

Size of text (shared-program) image. This is the size of the text section of the executable file. Pages of the text section of the executable program are only brought into memory when they are touched, that is, branched to or loaded from. This number represents only an upper bound on the amount of text that could be loaded. The TSIZ value does not reflect actual memory usage.

### ***The TRS column***

Size of the resident set (real memory) of text. This is the number of code segment pages times four. This number exaggerates the memory usage of programs that have multiple instances running.

## **10.2.3 The svmon command**

The **svmon** command provides a more in-depth analysis of memory usage. It is more informative, but also more intrusive, than the **vmstat** and **ps** commands. The **svmon** command captures a snapshot of the current state of memory. There are some significant changes in the flags and in the output from the **svmon** command between AIX Version 4.3.2 and AIX Version 4.3.3.

You can use four different reports to analyze the displayed information:

- ▶ Global (-G)  
Displays statistics describing the real memory and paging space in use for the whole system
- ▶ Process (-P)  
Displays memory usage statistics for active processes
- ▶ Segment (-S)  
Displays memory usage for a specified number of segments, or the top ten highest memory-usage processes, in descending order
- ▶ Detailed Segment (-D)  
Displays detailed information on specified segments

Additional reports are available in AIX Version 4.3.3 and later, as follows:

- ▶ User (-U)  
Displays memory usage statistics for the specified login names. If no list of login names is supplied, memory usage statistics display all defined login names.
- ▶ Command (-C)  
Displays memory usage statistics for the processes specified by the command name.

► Workload Management Class (-W)

Displays memory usage statistics for the specified workload management classes. If no classes are supplied, memory usage statistics display all defined classes.

To support 64-bit applications, the output format of the **svmon** command was modified in AIX Version 4.3.3 and later. Additional reports are available in operating system versions later than AIX Version 4.3.3, as follows:

► Frame (-F)

Displays information about frames. When no frame number is specified, the percentage of used memory is reported. When a frame number is specified, information about that frame is reported.

► Tier (-T)

Displays information about tiers, such as the tier number, the superclass name when the -a flag is used, and the total number of pages in real memory from segments belonging to the tier.

## 10.2.4 The **schedtune** command

The **schedtune** command sets parameters for the CPU scheduler and virtual memory manager processing. With the **schedtune** command you can change the scheduler parameters at run time, and these changes will be cleared at the next reboot. Edit `rc.local` with any preferred **schedtune** settings and run the command from the `rc.local` file to use the changes. The `rc.local` file should be executed from `/etc/inittab`. When using this command, be sure you are comfortable with the appropriate tuning sections before using **schedtune** to change system parameters, as misuse of this command can cause performance degradation or operating system failure. When executing the **schedtune** command without flags, the current values will be shown:

```
/usr/samples/kernel/schedtune

 THRASH SUSP FORK SCHED
-h -p -m -w -e -f -d -r -t -s
SYS PROC MULTI WAIT GRACE TICKS SCHED_D SCHED_R TIMESLICE MAXSPIN
0 4 2 1 2 10 16 16 1 16384

 CLOCK
 -c
%usDELTA
 100
```

## Priority calculation parameters

The priority of most user processes varies with the amount of CPU time the process has used recently. The CPU scheduler's priority calculations are based on two flags that are set with **schedtune**, **-r** and **-d**. The formula used by the scheduler to calculate the amount to be added to a process's priority value as a penalty for recent CPU use is:

$$\text{CPU penalty} = (\text{recently used CPU value of the process}) * (r/32)$$

The once-per-second recalculation of the recently used CPU value of each process is:

$$\text{new recently used CPU value} = (\text{old recently used CPU value of the process}) * (d/32)$$

Both **-r** and **-d** have default values of 16. This maintains the CPU scheduling behavior of previous versions of the operating system.

## Memory load control parameters

The operating system scheduler performs memory load control by suspending when memory is over committed. Memory is considered over committed when the following condition is met:

$$p * h > s$$

Where  $p$  is the number of pages written to paging space in the last second,  $h$  is an integer specified by the **-h** flag, and  $s$  is the number of page steals that have occurred in the last second.

A process is suspended when memory is over committed and the following condition is met:

$$r * p > f$$

Where  $r$  is the number of repages that the process has accumulated in the last second,  $p$  is an integer specified by the **-p** flag, and  $f$  is the number of page faults that the process has experienced in the last second.

## Time slice increment parameter

The **schedtune** command can also be used to change the amount the operating system allows a given process to run before the dispatcher is called to choose another process to run (*time slice*). The default value for this interval is a single clock tick (*10 milliseconds*). The **-t** flag of the **schedtune** command allows the user to specify the number of clock ticks by which the time length is to be increased.

## **schedtune limitations**

The **schedtune** command can only be executed by root. Changes made by the **schedtune** command last until the next reboot of the system. If a permanent change in VMM or time-slice parameters is needed, an appropriate **schedtune** command should be put in the `/etc/inittab` file.

### ***schedtune example one***

The following command will result in `SCHED_R = 0` and `SCHED_D = 0.5`:

```
/usr/samples/kernel/schedtune -r 0
```

This would mean that the CPU penalty was always 0, making priority absolute. No background process would get any CPU time unless there were no dispatchable foreground processes at all, as background processes in **ksh** are started with adding 4 to the nice value of the parent shell. The priority values of the threads would effectively be constant, although they would not technically be fixed-priority threads.

### ***schedtune example two***

The following command would result in an effective `SCHED_R = 1` and `SCHED_D = 1`:

```
/usr/samples/kernel/schedtune -r 32 -d 32
```

This would mean that long-running threads would reach a C value of 120 and remain there, contending on the basis of their nice values. New threads would have priority, regardless of their nice value, until they had accumulated enough time slices to bring them within the priority value range of the existing threads.

### ***schedtune example three***

The most likely reason to manipulate the values would be to make sure that background processes do not compete with foreground processes. By making `SCHED_R` smaller, you can restrict the range of possible priority values. For example:

```
/usr/samples/kernel/schedtune -r 5
```

(`SCHED_R = 0.15625`, `SCHED_D = 0.5`) would mean that a foreground process would never have to compete with a background process started with the command **nice -n 20**. The limit of 120 CPU time slices accumulated would mean that the maximum CPU penalty for the foreground process would be 18. In Figure 10-2 on page 257, this relationship is graphically shown. Because the CPU penalty will get a maximum value of 18, the foreground process with a nice value of 20 will always, when it needs, get CPU. On the other hand, the background process, with a nice value of 40, will use CPU only when the foreground process does not need the CPU.

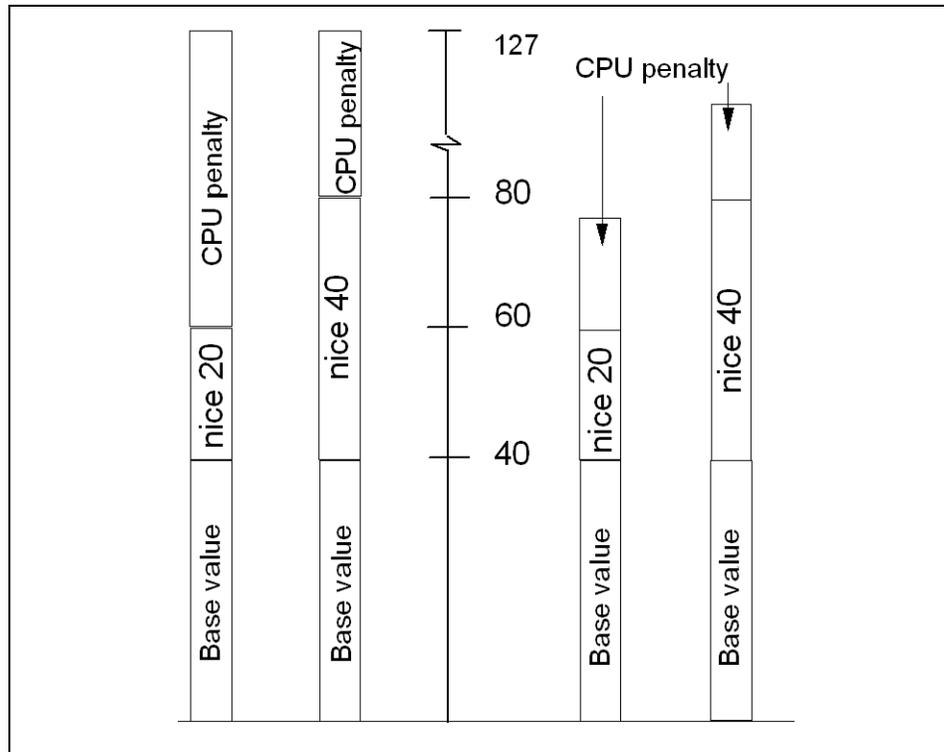


Figure 10-2 CPU penalty example

### **SCHED\_R and SCHED\_D guidelines**

The following discusses guidelines for tuning performance using SCHED\_R and SCHED\_D:

- ▶ Smaller values of SCHED\_R narrow the priority range and the nice value has more of an impact on the priority.
- ▶ Larger values of SCHED\_R widen the priority range and the nice value has less of an impact on the priority.
- ▶ Smaller values of SCHED\_D decay CPU usage at a faster rate and can cause CPU-intensive threads to be scheduled sooner.
- ▶ Larger values of SCHED\_D decay CPU usage at a slower rate and penalize CPU-intensive threads more (thus favoring interactive-type threads).

If you conclude that one or both parameters need to be modified to accommodate your workload, you can enter the **schedtune** command while logged on as root user. The changed values will persist until the next **schedtune** command that modifies them, or until the next system boot. Values can be reset to their defaults with the command **schedtune -D**, but remember that all

**schedtune** parameters are reset by that command, including VMM memory load control parameters. To make a change to the parameters that will persist across boots, add an appropriate line at the end of the `/etc/inittab` file.

*Table 10-3 Some commonly used flags for the schedtune command*

| Flag | Description                                                                                                  |
|------|--------------------------------------------------------------------------------------------------------------|
| -d   | Each process's recently used CPU value is multiplied by $d/32$ once a second.                                |
| -r   | A process's recently used CPU value is multiplied by $r/32$ when the process's priority value is calculated. |
| -t   | Increases the duration of the time slice.                                                                    |
| -m   | Sets minimum multiprogramming level.                                                                         |
| -p   | Specifies the per-process criterion for determining which process to suspend.                                |

## 10.3 Disk I/O bound system

The set of operating system commands, library subroutines, and other tools that allow you to establish and control logical volume storage is called the Logical Volume Manager (LVM). The Logical Volume Manager controls disk resources by mapping data between a more simple and flexible logical view of storage space and the actual physical disks. The LVM does this using a layer of device driver code that runs above traditional disk device drivers.

While an operating system's file is conceptually a sequential and contiguous string of bytes, the physical reality might be very different. Fragmentation may arise from multiple extensions to logical volumes as well as allocation/release/reallocation activity within a file system. A file system is fragmented when its available space consists of large numbers of small clusters of space, making it impossible to write out a new file in contiguous blocks.

Access to files in a highly fragmented file system may result in a large number of seeks and longer I/O response times (seek latency dominates I/O response time). For example, if the file is accessed sequentially, a file placement that consists of many widely separated clusters requires more seeks than a placement that consists of one or a few large contiguous clusters.

If the file is accessed randomly, a placement that is widely dispersed requires longer seeks than a placement in which the file's blocks are close together.

The VMM tries to anticipate the future need for pages of a sequential file by observing the pattern in which a program is accessing the file. When the program accesses two successive pages of the file, the VMM assumes that the program will continue to access the file sequentially, and the VMM schedules additional sequential reads of the file. This is called *Sequential-Access Read Ahead*. These reads are overlapped with the program processing, and will make the data available to the program sooner than if the VMM had waited for the program to access the next page before initiating the I/O. The number of pages to be read ahead is determined by two VMM thresholds:

|                   |                                                                                                                                                                                                                                                                                             |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>minpgahead</b> | Number of pages read ahead when the VMM first detects the sequential access pattern. If the program continues to access the file sequentially, the next read ahead will be for 2 times minpgahead, the next for 4 times minpgahead, and so on until the number of pages reaches maxpgahead. |
| <b>maxpgahead</b> | Maximum number of pages the VMM will read ahead in a sequential file.                                                                                                                                                                                                                       |

If the program deviates from the sequential-access pattern and accesses a page of the file out of order, sequential read ahead is terminated. It will be resumed with minpgahead pages if the VMM detects a resumption of sequential access by the program. The values of minpgahead and maxpgahead can be set with the **vm tune** command.

To increase write performance, limit the number of dirty file pages in memory, reduce system overhead, and minimize disk fragmentation. The file system divides each file into 16-KB partitions. The pages of a given partition are not written to disk until the program writes the first byte of the next 16-KB partition. At that point, the file system forces the four dirty pages of the first partition to be written to disk. The pages of data remain in memory until their frames are reused, at which point no additional I/O is required. If a program accesses any of the pages before their frames are reused, no I/O is required.

If a large number of dirty file pages remains in memory and does not get reused, the sync daemon writes them to disk, which might result in abnormal disk utilization. To distribute the I/O activity more efficiently across the workload, *write-behind* can be turned on to tell the system how many pages to keep in memory before writing them to disk. The write-behind threshold is on a per-file basis, which causes pages to be written to disk before the sync daemon runs. The I/O is spread more evenly throughout the workload.

There are two types of write-behind: *Sequential* and *random*. The size of the write-behind partitions and the write-behind threshold can be changed with the **vm tune** command.

Normal files are automatically mapped to segments to provide mapped files. This means that normal file access bypasses traditional kernel buffers and block I/O routines, allowing files to use more memory when the extra memory is available (file caching is not limited to the declared kernel buffer area).

Because most writes are asynchronous, FIFO I/O queues of several megabytes can build up, which can take several seconds to complete. The performance of an interactive process is severely impacted if every disk read spends several seconds working its way through the queue. In response to this problem, the VMM has an option called *I/O pacing* that controls writes.

I/O pacing does not change the interface or processing logic of I/O. It simply limits the number of I/Os that can be outstanding against a file. When a process tries to exceed that limit, it is suspended until enough outstanding requests have been processed to reach a lower threshold.

Disk-I/O pacing is intended to prevent programs that generate very large amounts of output from saturating the system's I/O facilities and causing the response times of less-demanding programs to deteriorate. Disk-I/O pacing enforces per-segment (which effectively means per-file) *high* and *low water marks* on the sum of all pending I/Os. When a process tries to write to a file that already has high water mark pending writes, the process is put to sleep until enough I/Os have completed to make the number of pending writes less than or equal to the low water mark. The logic of I/O-request handling does not change. The output from high-volume processes is slowed down somewhat.

When gathering information on I/O performance, the first command to use is normally the **iostat** command.

### 10.3.1 The **iostat** command

The **iostat** command is used for monitoring system input/output device loading by observing the time the physical disks are active in relation to their average transfer rates. The **iostat** command generates reports that can be used to change system configuration in order to better balance the input/output load between physical disks and adapters. The **iostat** command gathers its information on the protocol layer.

AIX Version 4.3.3 and AIX 5L Version 5.x have some significant changes to the output of the **iostat** command. These changes are similar to the changes described for the **vmstat** command found in “The wa column” on page 239.

#### **The TTY columns**

The two columns of TTY information (tin and tout) in the **iostat** output show the number of characters read and written by all TTY devices. This includes both real

and pseudo TTY devices. Real TTY devices are those connected to an asynchronous port. Some pseudo TTY devices are shells, **telnet** sessions, and **aixterm** windows. Because the processing of input and output characters consumes CPU resources, look for a correlation between increased TTY activity and CPU utilization. If such a relationship exists, evaluate ways to improve the performance of the TTY subsystem. Steps that could be taken include changing the application program, modifying TTY port parameters during file transfer, or perhaps upgrading to a faster or more efficient asynchronous communications adapter.

## The CPU columns

The CPU statistics columns (%user, %sys, %idle, and %iowait) provide a breakdown of CPU usage. This information is also reported in the **vmstat** command output in the columns labeled us, sy, id, and wa. For a detailed explanation for the values, see “The us column” on page 239, “The sy column” on page 239, “The id column” on page 239, and “The wa column” on page 239.

On systems running one application, a high I/O wait percentage might be related to the workload. On systems with many processes, some will be running while others wait for I/O. In this case, the % iowait can be small or zero because running processes *hide* some wait time. Although % iowait is low, a bottleneck can still limit application performance.

If the **iostat** command indicates that a CPU-bound situation does not exist, and % iowait time is greater than 20 percent, you might have an I/O or disk-bound situation. This situation could be caused by excessive paging due to a lack of real memory. It could also be due to unbalanced disk load, fragmented data, or usage patterns. For resolving such problems, a reorganization of logical volumes or a defragmentation of file systems might be necessary. For an unbalanced disk load, the same **iostat** report provides the necessary information. However, for information about file systems or logical volumes, which are logical resources, you must use tools such as the **filemon** or **fileplace** commands.

## The drive reports

When you suspect a disk I/O performance problem, use the **iostat** command. To avoid the information about the TTY and CPU statistics, use the -d option. In addition, the disk statistics can be limited to certain disks by specifying the disk names. Remember that the first set of data represents all activity since system startup. In the following example, the data is collected between intervals:

```
iostat 1 2

tty: tin tout avg-cpu: % user % sys % idle % iowait
 0.0 6.2 16.3 0.0 83.6 0.0

" Disk history since boot not available. "
```

```

tty: tin tout avg-cpu: % user % sys % idle % iowait
 0.0 192.7 100.0 0.0 0.0 0.0

```

```

Disks: % tm_act Kbps tps Kb_read Kb_wrtn
hdisk1 0.0 0.0 0.0 0 0
hdisk3 0.0 0.0 0.0 0 0
hdisk2 0.0 0.0 0.0 0 0
cd0 0.0 0.0 0.0 0 0

```

In such a case, statistics can be turned on with the following command:

```

chdev -l sys0 -a iostat=true
sys0 changed

```

### ***The disks column***

The disks column shows the names of the physical volumes. They are either `hdisk` or `cd`, followed by a number. If physical volume names are specified with the `iostat` command, only those names specified are displayed.

### ***The %tm\_act column***

The `%tm_act` column indicates the percentage of time that the physical disk was active (bandwidth utilization for the drive) or, in other words, the total time disk requests are outstanding. A drive is active during data transfer and command processing, such as seeking a new location. The *disk active time* percentage is directly proportional to resource contention and inversely proportional to performance. As disk use increases, performance decreases, and response time increases.

In general, when the utilization exceeds 70 percent, processes are waiting longer than necessary for I/O to complete, because most UNIX processes block (or sleep) while waiting for their I/O requests to complete. Look for busy versus idle drives. Moving data from busy to idle drives can help alleviate a disk bottleneck. Paging to and from disk will contribute to the I/O load.

### ***The Kbps column***

The Kbps column indicates the amount of data transferred (read or written) to the drive in KB per second. This is the sum of `Kb_read` plus `Kb_wrtn`, divided by the seconds in the reporting interval.

### ***The tps column***

The `tps` column indicates the number of transfers per second that were issued to the physical disk. A transfer is an I/O request through the device driver level to the physical disk. Multiple logical requests can be combined into a single I/O request to the disk. A transfer is of indeterminate size.

### ***The Kb\_read column***

The Kb\_read column reports the total data (in KB) read from the physical volume during the measured interval.

### ***The Kb\_wrtn column***

The Kb\_wrtn column shows the amount of data (in KB) written to the physical volume during the measured interval.

Taken alone, there is no unacceptable value for any of the above fields, because statistics are too closely related to application characteristics, system configuration, and type of physical disk drives and adapters. Therefore, when you are evaluating data, look for patterns and relationships. The most common relationship is between disk utilization (%tm\_act) and data transfer rate (tps).

To draw any valid conclusions from this data, you have to understand the application's disk data access patterns, such as sequential, random, or combination, as well as the type of physical disk drives and adapters on the system. For example, if an application reads/writes sequentially, you should expect a high disk transfer rate (Kbps) when you have a high disk busy rate (%tm\_act). Columns Kb\_read and Kb\_wrtn can confirm an understanding of an application's read/write behavior. However, these columns provide no information on the data access patterns.

Generally, you do not need to be concerned about a high disk busy rate (%tm\_act) as long as the disk transfer rate (Kbps) is also high. However, if you get a high disk busy rate and a low disk transfer rate, you may have a fragmented logical volume, file system, or individual file.

Discussions of disk, logical volume, and file system performance sometimes lead to the conclusion that the more drives you have on your system, the better the disk I/O performance. This is not always true because there is a limit to the amount of data that can be handled by a disk adapter. The disk adapter can also become a bottleneck. If all your disk drives are on one disk adapter, and your hot file systems are on separate physical volumes, you might benefit from using multiple disk adapters. Performance improvement will depend on the type of access.

To see if a particular adapter is saturated, use the **iostat** command and add up all the Kbps amounts for the disks attached to a particular disk adapter. For maximum aggregate performance, the total of the transfer rates (Kbps) must be below the disk adapter throughput rating. In most cases, use 70 percent of the throughput rate. In operating system versions later than AIX Version 4.3.3, the **-a** or **-A** option will display this information.

When looking for performance problems due to disk I/O, the next step is to find the file system causing the problem. This can be done with the **filemon** command.

### 10.3.2 The filemon command

The **filemon** command uses the trace facility to obtain a detailed picture of I/O activity during a time interval on the various layers of file system utilization, including the logical file system, virtual memory segments, LVM, and physical disk layers. Both summary and detailed reports are generated. Tracing is started by the **filemon** command, optionally suspended with the **trcoff** subcommand, resumed with the **trcon** subcommand, and terminated with the **trcstop** subcommand. As soon as tracing is terminated, the **filemon** command writes its report to stdout.

The following example shows the **filemon** command startup, writing the output report to a log file **fmon.out**, and then stopping the **filemon** command with the **trcstop** command.

```
filemon -o fmon.out -0 all
```

Enter the "trcstop" command to complete filemon processing

```
trcstop
[filemon: Reporting started]
[filemon: Reporting completed]

[filemon: 11.440 secs in measured interval]
```

If a file is identified as the problem, the **fileplace** command can be used to see how the file is stored.

### 10.3.3 The fileplace command

The **fileplace** command displays the placement of a specified file within the logical or physical volumes containing the file. By default, the **fileplace** command lists, to standard output, the ranges of logical volume fragments allocated to the specified file.

### 10.3.4 The sar command

The **sar** command collects, reports, or saves system activity. The **sar** command writes to standard output the contents of selected cumulative activity counters in the operating system. The accounting system, based on the values in the number and interval parameters, writes information about the specified number

of times spaced at the specified intervals in seconds. If you suspect an I/O-bound system, run the command in the example:

```
sar -d 1 1
```

```
AIX server4 1 5 00015F8F4C00 08/28/02
```

| 10:00:42 | device | %busy       | avque | r+w/s       | blks/s      | await | avserv |
|----------|--------|-------------|-------|-------------|-------------|-------|--------|
| 10:00:43 |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |
| 0.0      |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |
| 0.0      |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |
| 0.0      |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |
| 0.0      |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |
| 0.0      |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |
| 0.0      |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |
| 0.0      |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |
| 0.0      |        | -2147483648 | 0.0   | -2147483648 | -2147483648 | 0.0   |        |

The -d flag reports activity for each block device (for example, disk or tape drive), with the exception of XDC disks and tape drives. When data is displayed, the device specification disk is generally used to represent a disk drive. The device specification used to represent a tape drive is machine dependent. The activity data reported is as follows. The following reports the portion of time the device was busy servicing a transfer request, and the average number of requests outstanding during that time.

```
%busy, avque
```

The following reports the number of read/write transfers from or to a device. The number of bytes is transferred in 512-byte units.

```
read/s, write/s, blks/s
```

The following reports the number of milliseconds per average seek:

```
avseek
```

## 10.4 Network I/O bound system

When performance problems arise, your system might be totally innocent, while the real culprit is buildings away. An easy way to tell if the network is affecting overall performance is to compare those operations that involve the network with

those that do not. If you are running a program that does a considerable amount of remote reads and writes and it is running slowly, but everything else seems to be running normally, then it is probably a network problem. Some of the potential network bottlenecks can be caused by the following:

- ▶ Client-network interface
- ▶ Network bandwidth
- ▶ Network topology
- ▶ Server network interface
- ▶ Server CPU load
- ▶ Server memory usage
- ▶ Server bandwidth
- ▶ Inefficient configuration

A large part of network tuning involves tuning TCP/IP to achieve maximum throughput. With the new high bandwidth interfaces such as Gigabit Ethernet, this has become even more important.

The first command to use for gathering information on network performance is the **netstat** command.

### 10.4.1 The netstat command

The **netstat** command is used to show network status. Traditionally, it is used more for problem determination than for performance measurement. However, the **netstat** command can be used to determine the amount of traffic on the network, which can help determine whether performance problems are due to network congestion.

#### The netstat -i command

The **netstat -i** command shows the state of all configured interfaces.

The following example shows the statistics for a workstation with an integrated Ethernet and a token-ring adapter:

```
netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
lo0 16896 <Link> 144834 0 144946 0 0
lo0 16896 127 localhost 144834 0 144946 0 0
tr0 1492 <Link>10.0.5a.4f.3f.61 658339 0 247355 0 0
tr0 1492 9.3.1 ah6000d 658339 0 247355 0 0
en0 1500 <Link>8.0.5a.d.a2.d5 0 0 112 0 0
en0 1500 1.2.3 1.2.3.4 0 0 112 0 0
```

The count values are summarized since system startup:

|                         |                                                                                                                                  |
|-------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| <b>The Mtu column</b>   | Maximum transmission unit; the maximum size of packets in bytes that are transmitted using the interface                         |
| <b>The Ipkts column</b> | Total number of packets received                                                                                                 |
| <b>The Ierrs column</b> | Total number of input errors; for example, malformed packets, checksum errors, or insufficient buffer space in the device driver |
| <b>The Opkts column</b> | Total number of packets transmitted                                                                                              |
| <b>The Oerrs column</b> | Total number of output errors; for example, a fault in the local host connection or adapter output queue overrun                 |
| <b>The Coll column</b>  | Number of packet collisions detected                                                                                             |

### Tuning guidelines based on `netstat -i`

If the number of errors on input packets is greater than one percent of the total number of input packets (from the command `netstat -i`), that is:

$$Ierrs > 0.01 \times Ipkts$$

Then run the `netstat -m` command to check for a lack of memory.

If the number of errors during output packets is greater than one percent of the total number of output packets (from the command `netstat -i`), that is:

$$Oerrs > 0.01 \times Opkts$$

Then increase the send queue size (`tx_que_size`) for that interface. The size of the `tx_que_size` could be checked with the following command:

```
lsattr -El adapter
```

If the collision rate is greater than 10 percent, that is:

$$Coll / Opkts > 0.1$$

Then there is a high network utilization, and a reorganization or partitioning may be necessary. Use the `netstat -v` or `entstat` command to determine the collision rate.

### The `netstat -i -Z` command

The `netstat -i -Z` command clears all the statistic counters for the `netstat -i` command to zero.

## The netstat -m command

The **netstat -m** command displays the statistics recorded by the mbuf memory-management routines. The most useful statistics in the output of the **netstat -m** command are the counters that show the requests for mbufs denied and non-zero values in the failed column. If the requests for mbufs denied is not displayed, then this must be an SMP system running AIX Version 4.3.2 or later; for performance reasons, global statistics are turned off by default. To enable the global statistics, set the no parameter extended\_netstats to 1. This can be done by changing the /etc/rc.net file and rebooting the system.

The following example shows the first part of the **netstat -m** output with extended\_netstats set to 1:

```
netstat -m
426 mbufs in use:
384 mbuf cluster pages in use
1642 Kbytes allocated to mbufs
0 requests for mbufs denied
0 calls to protocol drain routines
0 sockets not created because sockthresh was reached
```

Kernel malloc statistics:

```
***** CPU 0 *****
```

| By size | inuse | calls | failed | delayed | free | hiwat | freed |
|---------|-------|-------|--------|---------|------|-------|-------|
| 32      | 92    | 187   | 0      | 0       | 36   | 1440  | 0     |
| 64      | 33    | 57    | 0      | 0       | 31   | 720   | 0     |
| 128     | 21    | 227   | 0      | 0       | 11   | 360   | 0     |
| 256     | 13    | 1602  | 0      | 0       | 19   | 864   | 0     |
| 512     | 30    | 125   | 0      | 0       | 2    | 90    | 0     |
| 1024    | 12    | 129   | 0      | 0       | 4    | 225   | 0     |
| 2048    | 1     | 3     | 0      | 0       | 1    | 225   | 0     |
| 4096    | 1     | 7     | 0      | 0       | 3    | 270   | 0     |
| 8192    | 2     | 43    | 0      | 0       | 1    | 22    | 0     |
| 16384   | 0     | 0     | 0      | 0       | 40   | 54    | 0     |
| 65536   | 1     | 1     | 0      | 0       | 0    | 2047  | 0     |

```
***** CPU 1 *****
```

| By size | inuse | calls | failed | delayed | free | hiwat | freed |
|---------|-------|-------|--------|---------|------|-------|-------|
| 32      | 87    | 313   | 0      | 0       | 41   | 1440  | 0     |
| 64      | 17    | 191   | 0      | 0       | 47   | 720   | 0     |
| 128     | 17    | 311   | 0      | 0       | 15   | 360   | 0     |
| 256     | 60    | 2497  | 0      | 0       | 20   | 864   | 0     |
| 512     | 62    | 406   | 0      | 0       | 10   | 90    | 0     |
| 1024    | 24    | 139   | 0      | 0       | 4    | 225   | 0     |
| 2048    | 1     | 3     | 0      | 0       | 1    | 225   | 0     |

|       |   |    |   |   |    |     |   |
|-------|---|----|---|---|----|-----|---|
| 4096  | 1 | 69 | 0 | 0 | 50 | 270 | 0 |
| 8192  | 0 | 42 | 0 | 0 | 2  | 22  | 0 |
| 16384 | 1 | 1  | 0 | 0 | 40 | 54  | 0 |

\*\*\*\*\* CPU 2 \*\*\*\*\*

| By size | inuse | calls | failed | delayed | free | hiwat | freed |
|---------|-------|-------|--------|---------|------|-------|-------|
| 32      | 31    | 335   | 0      | 0       | 97   | 1440  | 0     |
| 64      | 23    | 157   | 0      | 0       | 41   | 720   | 0     |
| 128     | 21    | 180   | 0      | 0       | 11   | 360   | 0     |
| 256     | 26    | 2593  | 0      | 0       | 22   | 864   | 0     |
| 512     | 39    | 350   | 0      | 0       | 1    | 90    | 0     |
| 1024    | 18    | 150   | 0      | 0       | 10   | 225   | 0     |
| 2048    | 1     | 7     | 0      | 0       | 1    | 225   | 0     |
| 4096    | 0     | 7     | 0      | 0       | 1    | 270   | 0     |
| 8192    | 1     | 48    | 0      | 0       | 2    | 22    | 0     |
| 16384   | 0     | 0     | 0      | 0       | 40   | 54    | 0     |

\*\*\*\*\* CPU 3 \*\*\*\*\*

| By size | inuse | calls | failed | delayed | free | hiwat | freed |
|---------|-------|-------|--------|---------|------|-------|-------|
| 32      | 43    | 188   | 0      | 0       | 85   | 1440  | 0     |
| 64      | 49    | 109   | 0      | 0       | 15   | 720   | 0     |
| 128     | 30    | 370   | 0      | 0       | 34   | 360   | 0     |
| 256     | 396   | 2483  | 0      | 0       | 84   | 864   | 0     |
| 512     | 26    | 186   | 0      | 0       | 6    | 90    | 0     |
| 1024    | 9     | 151   | 0      | 0       | 7    | 225   | 0     |
| 2048    | 386   | 390   | 0      | 0       | 2    | 225   | 0     |
| 4096    | 1     | 135   | 0      | 0       | 63   | 270   | 0     |
| 8192    | 0     | 32    | 0      | 0       | 2    | 22    | 0     |
| 16384   | 0     | 0     | 0      | 0       | 40   | 54    | 0     |

| By type          | inuse | calls | failed | delayed | memuse | memmax  | mapb |
|------------------|-------|-------|--------|---------|--------|---------|------|
| mbuf             | 426   | 6738  | 0      | 0       | 109056 | 125440  | 0    |
| mcluster         | 384   | 520   | 0      | 0       | 786432 | 1068032 | 0    |
| socket           | 280   | 1767  | 0      | 0       | 106272 | 121664  | 0    |
| pcb              | 93    | 713   | 0      | 0       | 15552  | 15808   | 0    |
| routetbl         | 19    | 0     | 0      | 0       | 2336   | 3488    | 0    |
| ifaddr           | 16    | 0     | 0      | 0       | 1792   | 1792    | 0    |
| mblk             | 9     | 2168  | 0      | 0       | 1280   | 3584    | 0    |
| mbldata          | 31    | 233   | 0      | 0       | 38912  | 41216   | 0    |
| strhead          | 19    | 64    | 0      | 0       | 6368   | 6368    | 0    |
| strqueue         | 35    | 165   | 0      | 0       | 17920  | 17920   | 0    |
| strmodsw         | 24    | 6     | 0      | 0       | 1536   | 1536    | 0    |
| strosr           | 0     | 13    | 0      | 0       | 0      | 256     | 0    |
| strsyncq         | 42    | 432   | 0      | 0       | 4864   | 4928    | 0    |
| streams          | 167   | 211   | 0      | 0       | 25312  | 25312   | 0    |
| devbuf           | 1     | 0     | 0      | 0       | 256    | 256     | 0    |
| kernel tablemoun | 16    | 15    | 0      | 0       | 85344  | 87392   | 0    |

|            |   |   |   |   |       |       |   |
|------------|---|---|---|---|-------|-------|---|
| spec buf   | 1 | 0 | 0 | 0 | 128   | 128   | 0 |
| locking    | 2 | 0 | 0 | 0 | 256   | 256   | 0 |
| temp       | 8 | 6 | 0 | 0 | 10560 | 15168 | 0 |
| mcast opts | 0 | 0 | 0 | 0 | 0     | 128   | 0 |
| mcast addr | 3 | 0 | 0 | 0 | 192   | 192   | 0 |

Streams mblk statistic failures:

0 high priority mblk failures

0 medium priority mblk failures

0 low priority mblk failures

If global statistics are not on, and you want to determine the total number of requests for mbufs denied, add up the values under the failed columns for each CPU. If the **netstat -m** command indicates that requests for mbufs or clusters have failed or been denied, then you may want to increase the value of thewall by using the **no -o thewall=NewValue** command.

Beginning with AIX Version 4.3.3, a delayed column was added. If the requester of an mbuf specified the M\_WAIT flag, then if an mbuf was not available, the thread is put to sleep until an mbuf is freed and can be used by this thread. The failed counter is not incremented in this case; instead, the delayed column will be incremented. Prior to AIX Version 4.3.3, the failed counter was also not incremented, but there was no delayed column.

If the currently allocated amount of network memory is within 85 percent of thewall, you may want to increase thewall. If the value of thewall is increased, use the **vmstat** command to monitor total memory use to determine if the increase has had a negative impact on overall memory performance.

## The netstat -v command

The **netstat -v** command displays the statistics for each Common Data Link Interface (CDLI) based device driver that is in operation. Interface-specific reports can be requested using the **tokstat**, **entstat**, **fddistat**, or **atmstat** commands.

Every interface has its own specific information and some general information. The most important output fields' descriptions are provided in the following sections.

### ***Transmit and Receive Errors***

The Transmit and Receive Errors statistic provides the number of output/input errors encountered on this device. This field counts unsuccessful transmissions due to hardware or network errors. These unsuccessful transmissions could also slow down the performance of the system.

### ***Max Packets on S/W Transmit Queue***

The Max Packets on S/W Transmit Queue statistic shows the maximum number of outgoing packets ever queued to the software transmit queue. An indication of an inadequate queue size is if the maximum transmits queued equals the current queue size (tx\_que\_size). This indicates that the queue was full at some point.

To check the current size of the queue, use the `lsattr -El adapter` command (where *adapter* is, for example, tok0 or ent0). Because the queue is associated with the device driver and adapter for the interface, use the adapter name, not the interface name. Use the SMIT or the `chdev` command to change the queue size.

### ***S/W Transmit Queue Overflow***

The S/W Transmit Queue Overflow statistic shows the number of outgoing packets that have overflowed the software transmit queue. A value other than zero requires the same actions as when the Max Packets on S/W Transmit Queue reaches the tx\_que\_size. The transmit queue size must be increased.

### ***Broadcast Packets***

The Broadcast Packets statistic shows the number of broadcast packets received without any error. If the value for broadcast packets is high, compare it with the total received packets. The received broadcast packets should be less than 20 percent of the total received packets. If it is high, this could be an indication of a high network load; a solution would be to use multicasting. The use of IP multicasting enables a message to be transmitted to a group of hosts, instead of having to address and send the message to each group member individually.

### ***DMA Overrun***

The DMA Overrun statistic is incremented when the adapter is using DMA to put a packet into system memory and the transfer is not completed. There are system buffers available for the packet to be placed into, but the DMA operation failed to complete. This occurs when the MCA bus is too busy for the adapter to be able to use DMA for the packets. The location of the adapter on the bus is crucial in a heavily loaded system. Typically, an adapter in a lower slot number on the bus, by having the higher bus priority, is using so much of the bus that adapters in higher slot numbers are not being served. This is particularly true if the adapters in a lower slot number are ATM or SSA adapters.

### ***Max Collision Errors***

The Max Collision Errors statistic shows the number of unsuccessful transmissions due to too many collisions. The number of collisions encountered exceeded the number of retries on the adapter.

### ***Late Collision Errors***

The Late Collision Errors statistic shows the number of unsuccessful transmissions due to the late collision error.

### ***Timeout Errors***

The Timeout Errors statistic shows the number of unsuccessful transmissions due to adapter-reported timeout errors.

### ***Single Collision Count***

The Single Collision Count statistic shows the number of outgoing packets with a single (only one) collision encountered during transmission.

### ***Multiple Collision Count***

The Multiple Collision Count statistic shows the number of outgoing packets with multiple (2–15) collisions encountered during transmission.

### ***Receive Collision Errors***

The Receive Collision Errors statistic shows the number of incoming packets with collision errors during reception.

### ***No mbuf Errors***

The No mbuf Errors statistic shows the number of times that mbufs were not available to the device driver. This usually occurs during receive operations when the driver must obtain memory buffers to process inbound packets. If the mbuf pool for the requested size is empty, the packet will be discarded. Use the **netstat -m** command to confirm this, and increase the **thewall** parameter.

The No mbuf Errors value is interface specific and not identical to the requests for mbufs denied from the **netstat -m** output. Compare the values of the example for the commands **netstat -m** and **netstat -v** (the Ethernet and token-ring part).

## **Tuning guidelines based on netstat -v**

To check for an overloaded Ethernet network, calculate (from the **netstat -v** command):

$(\text{Max Collision Errors} + \text{Timeout Errors}) / \text{Transmit Packets}$

If the result is greater than five percent, reorganize the network to balance the load.

Another indication for a high network load is found in the output of the command **netstat -v**.

If the total number of collisions from the **netstat -v** output (for Ethernet) is greater than 10 percent of the total transmitted packets, using the following formula, the system may have a high network load:

$$\text{Number of collisions} / \text{Number of Transmit Packets} > 0.1$$

If the system suffers from extensive NFS load, the **nfsstat** command provides useful information.

## 10.4.2 The **nfsstat** command

NFS gathers statistics on types of NFS operations performed, along with error information and performance indicators. You can use the **nfsstat** command to identify network problems and observe the type of NFS operations taking place on your system. The **nfsstat** command displays statistical information about the NFS and Remote Procedure Call (RPC) interfaces to the kernel. You can also use this command to reinitialize this information. The **nfsstat** command splits its information into server and client parts. The following commands can be used to match a particular need:

- ▶ **nfsstat -r** (to see the application NFS statistics)

The output is divided into server connection oriented and connectionless, as well as client connection oriented and connectionless.

- ▶ **nfsstat -s** (to see the server statistics)

The NFS server displays the number of NFS calls received (calls) and rejected (badcalls) due to authentication, as well as the counts and percentages for the various kinds of calls made.

- ▶ **nfsstat -c** (to see the client statistics)

The NFS client displays the number of calls sent and rejected, as well as the number of times a client handle was received (clgets) and a count of the various kinds of calls and their respective percentages. For performance monitoring, the **nfsstat -c** command provides information on whether the network is dropping UDP packets. A network may drop a packet if it cannot handle it. Dropped packets can be the result of the response time of the network hardware or software, or an overloaded CPU on the server. Dropped packets are not actually lost, because a replacement request is issued for them.

A high badxid count implies that requests are reaching the various NFS servers, but the servers are too loaded to send replies before the client's RPC calls time out and are retransmitted. The badxid value is incremented each time a duplicate reply is received for a transmitted request (an RPC request retains its XID through all transmission cycles). Excessive retransmissions place an additional strain on the server, further degrading response time.

The retrans column displays the number of times requests were retransmitted due to a timeout in waiting for a response. This situation is related to dropped UDP packets. If the retrans number consistently exceeds five percent of the total calls in column one, it indicates a problem with the server keeping up with demand.

When going into more detailed output, the **netpmon** command, using a trace facility, is useful.

### 10.4.3 The netpmon command

The **netpmon** command monitors a trace of system events and reports on network activity and performance during the monitored interval. By default, the **netpmon** command runs in the background while one or more application programs or system commands are being executed and monitored. The **netpmon** command automatically starts and monitors a trace of network-related system events in real time.

## 10.5 Workload Manager (WLM)

WLM is designed to give system administrators more control over how the scheduler and the virtual memory manager (VMM) allocate resources to processes. This can be used to prevent different classes of jobs from interfering with each other and to allocate resources based on the requirements of different groups of users.

With WLM, you can create different classes of service for jobs, as well as specify attributes for those classes. These attributes specify minimum and maximum amounts of CPU, physical memory, and disk I/O throughput to be allocated to a class. WLM then assigns jobs automatically to classes using class assignment rules provided by a system administrator. These assignment rules are based on the values of a set of attributes for a process. Either the system administrator or a privileged user can also manually assign jobs to classes, overriding the automatic assignment.

Class definitions, class attributes, the shares and limits, and the automatic class assignment rules, can be entered using Web-based System Manager, SMIT, or the WLM command-line interface. These definitions and rules are kept in plain text files that can also be created or modified using a text editor.

The system administrator can specify the properties for the WLM subsystem by using either the Web-based System Manager graphical user interface, SMIT ASCII-oriented interface, the WLM command line interface, or by creating flat

ASCII files. The Web-based System Manager and SMIT interfaces use the WLM commands to record the information in the same flat ASCII files.

These files are named as follows:

|                    |                                |
|--------------------|--------------------------------|
| <b>classes</b>     | Class definitions              |
| <b>description</b> | Configuration description text |
| <b>limits</b>      | Class limits                   |
| <b>shares</b>      | Class target shares            |
| <b>rules</b>       | Class assignment rules         |

These files are called the WLM property files. A set of WLM property files defines a WLM configuration. You can create multiple sets of property files, defining different configurations of workload management. These configurations are located in subdirectories of /etc/wlm. The WLM property files describing the superclasses of the Config configuration are the file's classes, description, limits, shares, and rules in /etc/wlm/Config. Then, the property files describing the subclasses of the superclass Super of this configuration are the file's classes, limits, shares, and rules in the directory /etc/wlm/Config/Super. Only the root user can start or stop WLM, or switch from one configuration to another.

The command to submit the WLM property files, **wlmcntr1**, and the other WLM commands allow users to specify an alternate directory name for the WLM properties files. This allows you to change the WLM properties without altering the default WLM property files.

A symbolic link, /etc/wlm/current, points to the directory containing the current configuration files. Update this link with the **wlmcntr1** command when you start WLM with a specified set of configuration files. The sample configuration files shipped with the operating system are in /etc/wlm/standard.

WLM configuration is performed through the preferred interface, the Web-based System Manager (Figure 10-3 on page 276), through a text editor and AIX commands, or through the AIX administration tool SMIT.

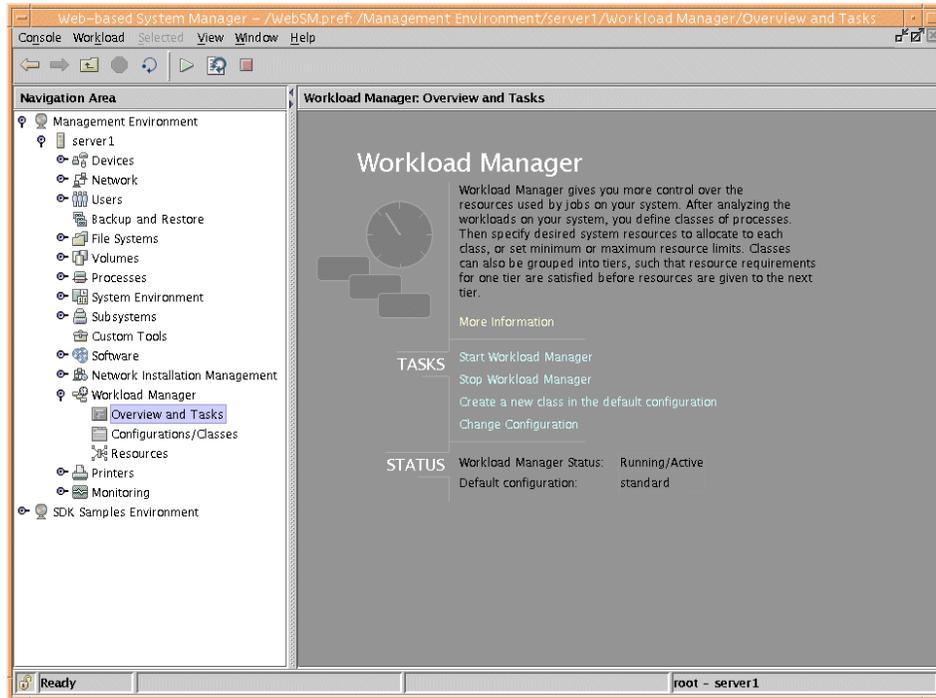


Figure 10-3 Web-based System Manager Overview and Tasks dialog

## 10.5.1 WLM concepts and architecture

The following sections outline the concepts provided with WLM on AIX 5L.

### Classes

The central concept of WLM is the class. A class is a collection of processes (jobs) that has a single set of resource limits applied to it. WLM assigns processes to the various classes and controls the allocation of system resources among the different classes. For this purpose, WLM uses class assignment rules and per-class resource shares and limits set by the system administrator. The resource entitlements and limits are enforced at the class level. This is a way of defining classes of service and regulates the resource utilization of each class of applications to prevent applications with very different resource utilization patterns from interfering with each other when they are sharing a single server.

### *Hierarchy of classes*

WLM allows system administrators to set up a hierarchy of classes with two levels by defining superclasses and subclasses. In other words, a class can

either be a *superclass* or a *subclass*. The main difference between superclasses and subclasses is the resource control (shares and limits):

- ▶ At the superclass level, the determination of resource entitlement (based on the resource shares and limits) is based on the total amount of each resource managed by WLM available on the machine.
- ▶ At the subclass level, the resource shares and limits are based on the amount of each resource allocated to the parent superclass.

The system administrator (the root user) can delegate the administration of the subclasses of each superclass to a *superclass administrator* (a non-root user), thus allocating a portion of the system resources to each superclass and then letting superclass administrators distribute the allocated resources among the users and applications they manage.

WLM supports 32 superclasses (27 user-defined plus five predefined). In turn, each superclass can have 12 subclasses (10 user-defined and two predefined, as shown in Figure 10-4). Depending on the needs of the organization, a system administrator can decide to use only superclasses or both superclasses and subclasses. An administrator can also use subclasses only for some of the superclasses.

Each class is given a name by the WLM administrator who creates it. A class name can be up to 16 characters long and can only contain uppercase and lowercase letters, numbers, and underscores (\_). For a given WLM configuration, the names of all the superclasses must be different from one another, and the names of the subclasses of a given superclass must be different from one another. Subclasses of different superclasses can have the same name. The fully qualified name of a subclass is *superclass\_name.subclass\_name*.

In the remainder of this section, whenever the term *class* is used, it is applicable to both subclasses and superclasses. The following subsections describe both super and subclasses in greater detail, as well as the backward compatibility WLM provides to configurations of its first release.

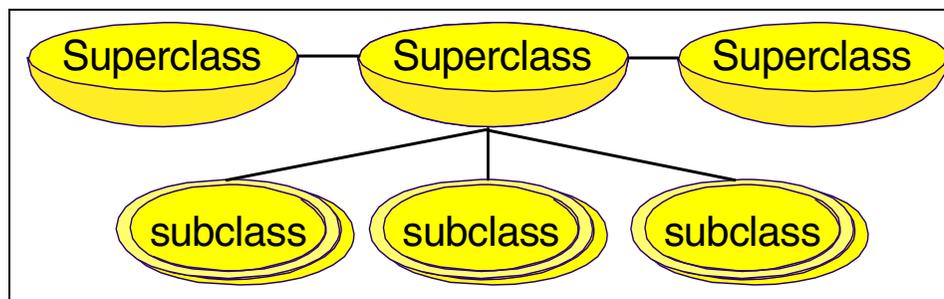


Figure 10-4 Hierarchy of classes

## **Superclasses**

A superclass is a class with subclasses associated with it. No process can belong to the superclass without also belonging to a subclass, either predefined or user defined. A superclass has a set of class assignment rules that determines which processes will be assigned to it. A superclass also has a set of resource limitation values and resource target shares that determine the amount of resources that can be used by processes belonging to it. These resources will be divided among the subclasses based on the resource limitation values and resource target shares of the subclasses.

Up to 27 superclasses can be defined by the system administrator. In addition, five superclasses are automatically created to deal with processes, memory, and CPU allocation, as follows:

- ▶ *Default* superclass: The default superclass is named Default and is always defined. All non-root processes that are not automatically assigned to a specific superclass will be assigned to the Default superclass. Other processes can also be assigned to the Default superclass by providing specific assignment rules.
- ▶ *System* superclass: This superclass has all privileged (root) processes assigned to it if they are not assigned by rules to a specific class, plus the pages belonging to all system memory segments, kernel processes, and kernel threads. Other processes can also be assigned to the System superclass. The default is for this superclass to have a memory minimum limit of one percent.
- ▶ *Shared* superclass: This superclass receives all the memory pages that are shared by processes in more than one superclass. This includes pages in shared memory regions and pages in files that are used by processes in more than one superclass (or in subclasses of different superclasses). Shared memory and files used by multiple processes that belong to a single superclass (or subclasses of the same superclass) are associated with that superclass. The pages are placed in the Shared superclass only when a process from a different superclass accesses the shared memory region or file. This superclass can have only physical memory shares and limits applied to it. It cannot have shares or limits for the other resource types, subclasses, or assignment rules specified. Whether a memory segment shared by the processes in the different superclasses is classified into the Shared superclass, or remains in the superclass it was initially classified into depends on the value of the localshm attribute of the superclass the segment was initially classified into.
- ▶ *Unclassified* superclass: The processes in existence at the time WLM is started are classified according to the assignment rules of the WLM configuration being loaded. During this initial classification, all the memory pages attached to each process are charged either to the superclass the

process belongs to (when not shared, or shared by processes in the same superclass) or to the Shared superclass, when shared by processes in different superclasses. However, there are a few pages that cannot be directly tied to any processes (and thus to any class) at the time of this classification, and this memory is charged to the Unclassified superclass; for example, pages from a file that has been closed. The file pages will remain in memory, but no process *owns* these pages; therefore, they cannot be charged to a specific class. Most of this memory will end up being correctly reclassified over time, when it is either accessed by a process, or freed and reallocated to a process after WLM is started. There are a few kernel processes, such as wait or Irud, in the Unclassified superclass. Even though this superclass can have physical memory shares and limits applied to it, WLM commands do not allow you to set shares and limits or specify subclasses or assignment rules on this superclass.

- ▶ *Unmanaged* superclass: A special superclass named Unmanaged will always be defined. No processes will be assigned to this class. This class will be used to accumulate the memory usage for all pinned pages in the system that are not managed by WLM. The CPU utilization for the waitprocs is not accumulated in any class. This is deliberate; otherwise, the system would always seem to be at 100 percent CPU utilization, which could be misleading for users when looking at the WLM or system statistics. This superclass cannot have shares or limits for any other resource types, subclasses, or assignment rules specified.

### **Subclasses**

A subclass is a class associated with exactly one superclass. Every process in the subclass is also a member of the superclass. Subclasses only have access to resources that are available to the superclass. A subclass has a set of class assignment rules that determine which of the processes assigned to the superclass will belong to it. A subclass also has a set of resource limitation values and resource target shares that determine the resources that can be used by processes in the subclass. These resource limitation values and resource target shares indicate how much of the superclass's target (the resources available to the superclass) can be used by processes in the subclass.

Up to 10 out of a total of 12 subclasses can be defined by the system administrator or by the superclass administrator for each superclass. In addition, two special subclasses, Default and Shared, are always defined in each superclass as follows:

- ▶ *Default* subclass: The default subclass is named Default and is always defined. All processes that are not automatically assigned to a specific subclass of the superclass will be assigned to the Default subclass. You can also assign other processes to the Default subclass by providing specific assignment rules.

- ▶ *Shared* subclass: This subclass receives all the memory pages used by processes in more than one subclass of the superclass. This includes pages in shared memory regions and pages in files that are used by processes in more than one subclass of the same superclass. Shared memory and files used by multiple processes that belong to a single subclass are associated with that subclass. The pages are placed in the Shared subclass of the superclass only when a process from a different subclass of the same superclass accesses the shared memory region or file. There are no processes in the Shared subclass. This subclass can only have physical memory shares and limits applied to it. It cannot have shares or limits for the other resource types or assignment rules specified.

## Tiers

Tier configuration is based on the importance of a class relative to other classes in WLM. There are 10 available tiers, from 0 through to 9. Tier value 0 is the most important and value 9 is the least important. As a result, classes belonging to tier 0 will get resource allocation priority over classes in tier 1; classes in tier 1 will have priority over classes in tier 2; and so on. The default tier number, if the attribute is not specified, is 0.

The tier applies at both the superclass and subclass levels. Superclass tiers are used to specify resource allocation priority between superclasses, and subclass tiers are used to specify resource allocation priority between subclasses of the same superclass. There is no relationship between tier numbers of subclasses of different superclasses.

Tier separation, in terms of prioritization, is much more enforced in AIX 5L than in the previous release. A process in tier 1 will never have priority over a process in tier 0, since there is no overlapping of priorities in tiers. It is unlikely that classes in tier 1 will acquire any resources if the processes in tier 0 are consuming all the resources. This occurs because the control of leftover resources is much more restricted than in the AIX Version 4.3.3 release of WLM, as shown in Figure 10-5 on page 281.

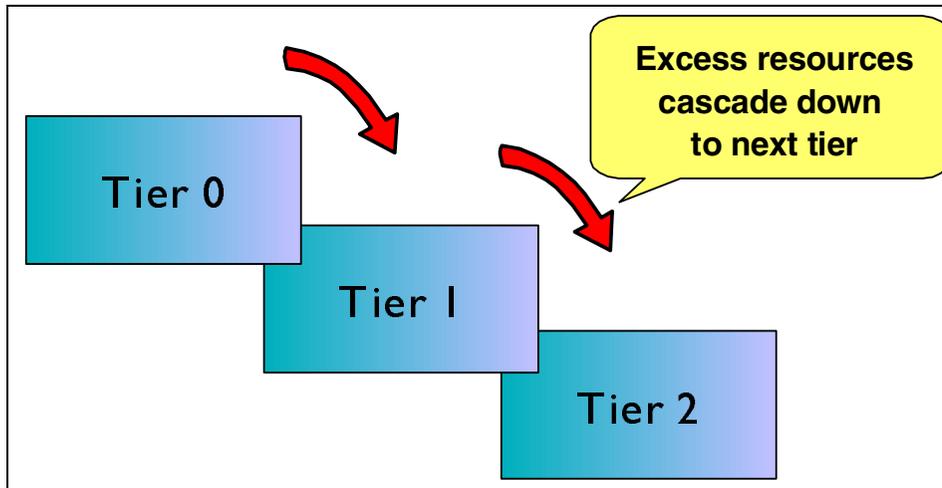


Figure 10-5 Resources cascading through tiers

### Class attributes

In order to create a class, there are different attributes that are needed to have an accurate and well-organized group of classes. Figure 10-6 shows the SMIT panel for class attributes.

```

 General characteristics of a class

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

 [Entry Fields]
* Class name []
Description []
Tier [0] + #
Resource Set +
Inheritance [No] +
User authorized to assign its processes to this cl [] +
ass
Group authorized to assign its processes to this c [] +
lass
User authorized to administrate this class [] +
(Superclass only)
Group authorized to administrate this class [] +
(Superclass only)

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

```

Figure 10-6 SMIT with the class creation attributes screen

The sequence of attributes within a class (as shown in Figure 10-6 on page 281) is outlined below:

- ▶ **Class name**  
A unique class name with up to 16 characters. It can contain uppercase and lowercase letters, numbers, and underscores (\_).
- ▶ **Description**  
An optional brief description about this class.
- ▶ **Tier**  
A number between 0 and 9, for class priority ranking. It will be the tier that this class will belong to. An explanation about tiers can be found in “Tiers” on page 280.
- ▶ **Resource Set**  
This attribute is used to limit the set of resources a given class has access to in terms of CPUs (processor set). The default, if unspecified, is *system*, which gives access to all the CPU resources available on the system.
- ▶ **Inheritance**  
The inheritance attribute indicates whether a child process should inherit its parent’s class or get classified according to the automatic assignment rules upon exec. The possible values are *yes* or *no*; the default is *no*. This attribute can be specified at both superclass and subclass level.
- ▶ **User and Group authorized to assign its processes to this class**  
These attributes are valid for all the classes. They are used to specify the user name and the group name of the user or group authorized to manually assign processes to the class. When manually assigning a process (or a group of processes) to a superclass, the assignment rules for the superclass are used to determine which subclass of the superclass each process will be assigned to.
- ▶ **User and Group authorized to administer this class**  
These attributes are valid only for superclasses. They are used to delegate the superclass administration to a user and group of users.
- ▶ **Localchm**  
Specifies whether memory segments that are accessed by processes in different classes remain local to the class they were initially assigned to, or if they go to the Shared class.

### ***Segment authorization to migrate to the Shared class***

With Workload Manager in earlier versions of AIX, whenever a memory segment is accessed by processes from different classes, the segment is reclassified as

Shared. This occurs because one of the classes sharing the memory segment would otherwise be penalized as the user of this resource while the others are not. The consequence of the segment moving to Shared is that users partially lose control of it. In AIX 5L, an attribute has been added at the class level to avert the automatic reclassification of the class. This attribute, `localshm`, if set to `no`, allows the segment to be reclassified to the Shared class. If it is set to `yes`, then it is not reclassified. From the command line, the command will be similar to that shown in the example below:

```
mkclass -a tier=2 -a adminuser=wlm06 -a localshm=yes -c shares=2\
-m shares=3 -d new_config super3
```

From the SMIT panels, general characteristics of a class panel will have the `localshm` option, as in the example shown in Figure 10-7.

General Characteristics of a class

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

|                                                    | [Entry Fields]    |     |
|----------------------------------------------------|-------------------|-----|
| Class name                                         | bob               |     |
| Description                                        | [bob the builder] |     |
| Tier                                               | [0]               | + # |
| Resource Set                                       |                   | +   |
| Inheritance                                        | [No]              | +   |
| User authorized to assign its processes to this cl | [ ]               | +   |
| ass                                                |                   |     |
| Group authorized to assign its processes to this c | [ ]               | +   |
| lass                                               |                   |     |
| User authorized to administrate this class         | [ ]               | +   |
| (Superclass only)                                  |                   |     |
| Group authorized to administrate this class        | [ ]               | +   |
| (Superclass only)                                  |                   |     |
| <b>Localshm</b>                                    | [Yes]             | +   |

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

Figure 10-7 SMIT panel shows the additional `localshm` attribute

## Classification process

There are two ways to classify processes in WLM:

- ▶ Automatic assignment when a process calls the system call `exec`, using assignment rules specified by a WLM administrator. This automatic assignment is always in effect (cannot be turned off) when WLM is active. This is the most common method of assigning processes to the different classes.
- ▶ Manual assignment of a selected process or group of processes to a class by a user with the required authority on both the process and the target class.

This manual assignment can be done either by a WLM command, which could be invoked directly or through SMIT or Web-based System Manager, or by an application, using a function of the WLM Application Programming Interface. Manual assignment overrides automatic assignment.

## 10.5.2 Automatic assignment

The automatic assignment of processes to classes uses a set of class assignment rules specified by a WLM administrator. There are two levels of assignment rules:

- ▶ A set of assignment rules at the WLM configuration level used to determine which superclass a given process should be assigned to.
- ▶ A set of assignment rules at the superclass level used to determine which subclass of the superclass the process should be assigned to.

The assignment rules at both levels have exactly the same format.

When a process is created by fork, it remains in the same class as its parent. Usually, reclassification happens when the new process calls the system call `exec`. In order to classify the process, WLM starts by examining the top level rules list for the active configuration to find out which superclass the process should belong to. For this purpose, WLM takes the rules one at a time, in the order they appear in the file, and checks the current values for the process attributes against the values and lists of values specified in the rule. When a match is found, the process will be assigned to the superclass named in the first field of the rule. Then the rules list for the superclass is examined in the same way to determine which subclass of the superclass the process should be assigned to. For a process to match one of the rules, each of its attributes must match the corresponding field in the rule. The rules to determine whether the value of a process attribute matches the values in the field of the rules list are as follows:

- ▶ If the field in the rule has a value of hyphen (-), any value of the corresponding process attribute is a match.
- ▶ If the value of the process attribute (for all the attributes except *type*) matches one of the values in the list in a rule, and it is not excluded (prefaced by an exclamation point (!)), it is considered a match.
- ▶ When one of the values for the *type* attribute in the rule is comprised of two or more values separated by a plus sign (+), a process will be a match for this value only if its characteristics match all the values mentioned above.

As previously mentioned, at both superclass and subclass levels, WLM goes through the rules in the order in which they appear in the rules list, and classifies the process in the class corresponding to the first rule for which the process is a

match. This means that the order of the rules in the rules list is extremely important, and caution must be applied when modifying it in any way.

### 10.5.3 Manual assignment

Manual assignment is a feature introduced in AIX 5L WLM. It allows system administrators and applications to override, at any time, the traditional WLM automatic assignment (processes' automatic classification based on class assignment rules) and force a process to be classified in a specific class.

The manual assignment can be made or canceled separately at the superclass level, the subclass level, or both. In order to manually assign processes to a class or cancel an existing manual assignment, a user must have the right level of privilege (that is, must be the root user, adminuser, or admingroup for the superclass or authuser and authgroup for the superclass or subclass). A process can be manually assigned to a superclass only, a subclass only, or to a superclass and a subclass of the superclass. In the latter case, the dual assignment can be done simultaneously (with a single command or API call) or at different times, possibly by different users.

A manual assignment will remain in effect (and a process will remain in its manually assigned class) until:

- ▶ The process terminates.
- ▶ WLM is stopped. When WLM is restarted, the manual assignments in effect when WLM was stopped are lost.
- ▶ The class the process has been assigned to is deleted.
- ▶ A new manual assignment overrides a prior one.
- ▶ The manual assignment for the process is canceled.

In order to assign a process to a class or cancel a prior manual assignment, the user must have authority both on the process and on the target class. These constraints translate into the following:

- ▶ The root user can assign any process to any class.
- ▶ A user with administration privileges on the subclasses of a given superclass (that is, the user or group name matches the attributes adminuser or admingroup of the superclass) can manually reassign any process from one of the subclasses of this superclass to another subclass of the superclass.
- ▶ A user can manually assign her own processes (same real or effective user ID) to a superclass or a subclass for which they have manual assignment privileges (that is, the user or group name matches the attributes authuser or authgroup of the superclass or subclass).

This defines three levels of privilege among the persons who can manually assign processes to classes, root being the highest. In order for a user to modify or cancel a manual assignment, the user must be at the same or a higher level of privilege as the person who issued the last manual assignment.

## **Class assignment rules**

After the definition of a class, it is time to set up the class assignment rules so that WLM can perform its automatic assignment. The assignment rules are used by WLM to assign a process to a class based on the user, group, application path name, type of process, and application tag, or a combination of these five attributes.

The next sections describe the attributes that constitute a class assignment rule. All these attributes can contain a hyphen (-), which means that this field will not be considered when assigning classes to a process.

### ***Class name***

This field must contain the name of a class, which is defined in the class file corresponding to the level of the rules file we are configuring (either superclass or subclass). Class names can contain only uppercase and lowercase letters, numbers, and underscores (\_), and can be up to 16 characters in length. No assignment rule can be specified for the system-defined classes *Unclassified*, *Unmanaged*, and *Shared*.

### ***Reserved***

Reserved for future use. Its value *must* be a hyphen (-), and it must be present in the rule.

### ***User***

The user name (as specified in the */etc/passwd* file, LDAP, or in NIS) of the user owning a process can be used to determine the class to which the process belongs. This attribute is a list of one or more user names, separated by a comma (,). Users can be excluded by using an exclamation point (!) prefix. Patterns can be specified to match a set of user names using full Korn shell pattern matching syntax.

Applications that use the *setuid* permission to change the *effective* user ID they run under are still classified according to the user that invoked them. The processes are only reclassified if the change is done to the *real* user ID (UID).

### ***Group***

The group name (as specified in the */etc/group* file, LDAP, or in NIS) of a process can be used to determine the class to which the process belongs. This attribute is a list composed of one or more groups, separated by a comma (,). Groups can

be excluded by using an exclamation point (!) prefix. Patterns can be specified to match a set of group names using full Korn shell pattern matching syntax.

Applications that use the **setgid** permission to change the *effective* group ID they run under are still classified according to the group that invoked them. The processes are only reclassified if the change is done to the *real* group ID (GID).

### **Application path names**

The full path name of the application for a process can be used to determine the class to which a process belongs. This attribute is a list composed of one or more applications, separated by a comma (.). The application path names will be either full path names or Korn shell patterns that match path names. Application path names can be excluded by using an exclamation point (!) prefix.

### **Process type**

In AIX 5L, the process type attribute is introduced as one of the ways to determine the class to which a process belongs. This attribute consists of a comma-separated list, with one or more combination of values, separated by a plus sign (+). A plus sign provides a logical *and* function, and a comma provides a logical *or* function. Table 10-4 provides a list of process types that can be used. (Note: *32bit* and *64bit* are mutually exclusive.)

Table 10-4 List of process types

| Attribute value | Process type                                               |
|-----------------|------------------------------------------------------------|
| 32bit           | The process is a 32-bit process.                           |
| 64bit           | The process is a 64-bit process.                           |
| plock           | The process called plock() to pin memory.                  |
| fixed           | The process has a fixed priority (SCHED_FIFO or SCHED_RR). |

### **Application tags**

In AIX 5L, the application tag attribute is introduced as one of the forms of determining the class to which a process belongs. This is an attribute meant to be set by WLM's API, as a way to further extend the process classification possibilities. This process was created to allow differentiated classification for different instances of the same application. This attribute can have one or more application tags, separated by commas (.). An application tag is a string of up to 30 alphanumeric characters.

The classification is done by comparing the value of the attributes of the process at exec time against the lists of class assignment rules to determine which rule is a match for the current value of the process attributes.

The class assignment is done by WLM:

- ▶ When WLM is started for all the processes existing at that time
- ▶ Every time a process calls the system calls `exec`, `setuid` (and related calls), `setgid` (and related calls), `setpri`, and `plock`, once WLM is started

There are two *default* rules that are always defined (that is, hardwired in WLM). These are the default rules that assign all processes started by the user `root` to the System class, and all other processes to the Default class. If WLM does not find a match in the assignment rules list for a process, these two rules will be applied (the rule for System first), and the process will go to either System (UID `root`) or Default. These default rules are the only assignment rules in the standard configuration installed with AIX.

Table 10-5 is an example of classes with their respective attributes for assignment rules.

Table 10-5 Examples of class assignment rules

| Class  | Reserved | User    | Group | Application         | Type  | Tag  |
|--------|----------|---------|-------|---------------------|-------|------|
| System | -        | root    | -     | -                   | -     | -    |
| db1    | -        | -       | -     | /usr/oracle/bin/db* | -     | _db1 |
| db2    | -        | -       | -     | /usr/oracle/bin/db* | -     | _db2 |
| devlt  | -        | -       | dev   | -                   | 32bit | -    |
| VPs    | -        | bob,ted | -     | -                   | -     | -    |
| acctg  | -        | -       | acct* | -                   | -     | -    |

In Table 10-5, the rule for Default class is omitted from the display, though this class's rule is always present in the configuration. The rule for System is explicit, and has been put first in the file. This is deliberate so that all processes started by `root` will be assigned to the System superclass. By moving the rule for the System superclass further down in the rules file, the system administrator could have chosen to assign the root processes that would not be assigned to another class (because of the application executed, for example) to System only. In Table 10-5, with the rule for System on top, if `root` executes a program in `/usr/oracle/bin/db*` set, the process will be classified as System. If the rule for the System class was after the rule for the `db2` class, the same process would be classified as `db1` or `db2`, depending on the tag.

These examples show that the order of the rules in the assignment rules file is very important. The more specific assignment rules should appear first in the rules file, and the more general rules should appear last. An extreme example would be putting the default assignment rule for the Default class, for which every

process is a match, first in the rules file. That would cause every process to be assigned to the Default class (the other rules would, in effect, be ignored).

You can define multiple assignment rules for any given class. You can also define your own specific assignment rules for the System or Default classes. The default rules mentioned previously for these classes would still be applied to processes that would not be classified using any of the explicit rules.

## 10.5.4 Backward compatibility

As mentioned earlier, in the first release of WLM, the system default for the resource shares was one share. In AIX 5L, it is a hyphen (-), which means that the resource consumption of the class for this particular resource is not regulated by WLM. This changes the semantics quite a bit, and it is advisable that system administrators review their existing configurations and consider if the new default is good for their classes, or if they would be better off either setting up a default of one share (going back to the previous behavior) or setting explicit values for some of the classes.

In terms of limits, the first release of WLM only had one maximum, not two. This maximum limit was in fact a *soft* limit for CPU and a *hard* limit for memory. Limits specified for the old format, *min percent-max percent*, will have, in AIX 5L, the max interpreted as a softmax for CPU and both values of hardmax and softmax for memory. All interfaces (SMIT, AIX commands, and Web-based System Manager) will convert all data existing from its old format to the new one.

The disk I/O resource is new for the current version, so when activating the AIX 5L WLM with the configuration files of the first WLM release, the values for the shares and the limits will be the default ones for this resource. The system defaults are:

- ▶ shares = -
- ▶ min = 0 percent, softmax = 100 percent, hardmax = 100 percent

For existing WLM configurations, the disk I/O resource will not be regulated by WLM, which should lead to the same behavior for the class as with the first version.

## 10.5.5 Resource sets

WLM uses the concept of resource sets (or rsets) to restrict the processes in a given class to a subset of the system's physical resources. In AIX 5L, the physical resources managed are the memory and the processors. A valid resource set is composed of memory and at least one processor.

Figure 10-8 shows the SMIT panel where a resource set can be specified for a specific class.

```

 General characteristics of a class

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

 [Entry Fields]
Class name Redbook
Description [Redbook example]
Tier [0] + #
Resource Set sys/cpu.00003 +
Inheritance [Yes] +
User authorized to assign its processes to this class [user_s] +
ass
Group authorized to assign its processes to this class [system] +
lass
User authorized to administrate this class [user_s] +
(Superclass only)
Group authorized to administrate this class [system] +
(Superclass only)

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

```

Figure 10-8 Resource set definition to a specific class

By default, the system creates one resource set for all physical memory, one for all CPUs, and one separate set for each individual CPU in the system. The **lsrset** command lists all resource sets defined. A sample output for the **lsrset** command follows:

```

lsrset -av
T Name Owner Group Mode CPU Memory Resources
r sys/sys0 root system r----- 4 511 sys/sys0
sys/node.00000 sys/mem.00000 sys/cpu.00003 sys/cpu.00002 sys/cpu.00001
sys/cpu.00000
r sys/node.00000 root system r----- 4 511 sys/sys0
sys/node.00000 sys/mem.00000 sys/cpu.00003 sys/cpu.00002 sys/cpu.00001
sys/cpu.00000
r sys/mem.00000 root system r----- 0 511 sys/mem.00000
r sys/cpu.00003 root system r----- 1 0 sys/cpu.00003
r sys/cpu.00002 root system r----- 1 0 sys/cpu.00002
r sys/cpu.00001 root system r----- 1 0 sys/cpu.00001
r sys/cpu.00000 root system r----- 1 0 sys/cpu.00000

```

## 10.5.6 rset registry

As mentioned previously, some resource sets in AIX 5L are created, by default, for memory and CPU. It is possible to create different resource sets by grouping two or more resource sets and storing the definition in the rset registry.

The rset registry services enable system administrators to define and name resource sets so that they can then be used by other users or applications. In order to alleviate the risks of name collisions, the registry supports a two-level naming scheme. The name of a resource set takes the form *name\_space/rset\_name*. Both the name space and rset\_name may each be 255 characters in size, are case sensitive, and may contain only upper and lower case letters, numbers, underscores, and periods. The name space of sys is reserved by the operating system and used for rset definitions that represent the resources of the system.

The **SMIT rset** command has options to list, remove, or show a specific resource set used by a process and the management tools, as shown in Figure 10-9.

```
Resource Set Management

Move cursor to desired item and press Enter.

List All Resource Sets
List All Resource Sets in a given namespace
List All System RADs
List Application-defined Resource Sets
Remove Application-defined Resource Sets
Show a Process Partition
Manage Resource Set Database

F1=Help F2=Refresh F3=Cancel F8=Image
F9=Shell F10=Exit Enter=Do
```

Figure 10-9 SMIT main panel for resource set management

To create, delete, or change a resource set in the rset registry, you must select the **Manage Resource Set Database** item in the SMIT panel. In this panel, it is also possible to reload the rset registry definitions to make all changes available to the system. Figure 10-10 on page 292 shows the SMIT panel for rset registry management.

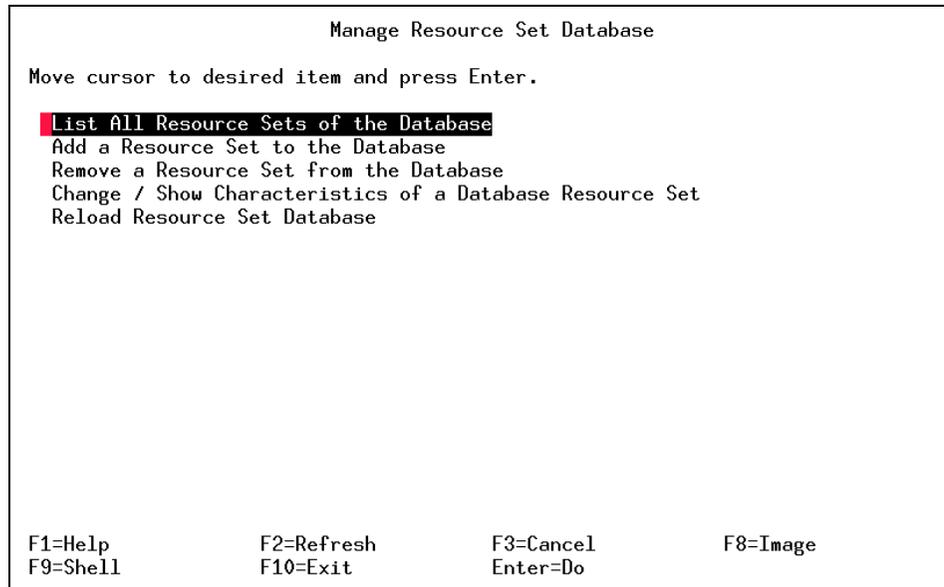


Figure 10-10 SMIT panel for rset registry management

To add a new resource set, you must specify a name space, a resource set name, and the list of resources. It is also possible to change the permissions for the owner and group of this rset. In addition, permissions for the owner, groups, and others can also be specified. Figure 10-11 on page 293 shows the SMIT panel for this task.

```

 Add a Resource Set to the Database

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

 [Entry Fields]
* Name Space [Redbook] +
* Resource Set Name [CPU0and1] +
* Owner root +
* Group system +
* Owner Permissions rw +
* Group Permissions r- +
* Others Permissions r- +
* Resources sys/cpu.00001,sys/cpu.> +

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

```

Figure 10-11 SMIT panel to add a new resource set

Whenever a new rset is created, deleted, or modified, a reload in the rset database is needed in order to make the changes effective.

After a WLM configuration has been defined by the system administrator, it can be made the active configuration using the `wlmcntrl` command. For example, to start WLM in active mode, enter:

```
wlmcntrl -a
```

To validate if WLM is running, enter:

```
wlmcntrl -q
1495-052 WLM is running
```

To stop WLM, enter:

```
wlmcntrl -o
```

Please see *AIX Version 4.3 System Management Concepts: Operating System and Devices*, SC23-4311, for more information.

## 10.6 System debuggers

This section highlights some of the tools used as system debuggers.

## 10.6.1 The dbx command

The **dbx** command provides an environment to debug and run programs. The **dbx** command provides a symbolic debug program for C, C++, and FORTRAN programs, allowing you to carry out operations such as the following:

- ▶ Examine object and core files.
- ▶ Provide an environment for running a program.
- ▶ Set breakpoints at selected statements or run the program one line at a time.
- ▶ Debug using symbolic variables and display them in their correct format.
- ▶ Hot-keying into the system debugger is a way to check for hung or busy systems.

An example of using the **dbx** command would be to identify which program has core dumped and where exactly it ended. Using **dbx** on the core file will tell you which program created the core file but may report that there is no program matching the core image. If you have an executable compiled with -g you might find out where the program crashed with the where subcommand. Copy the executable and the core file into one directory and run **dbx** on the executable and use the where subcommand. To display the full qualification of the given identifier use the which subcommand.

Table 10-6 lists further **dbx** subcommands.

*Table 10-6 Sample dbx subcommands*

| Subcommand | Description                                                          |
|------------|----------------------------------------------------------------------|
| /          | Searches forward in the current source file for a pattern            |
| ?          | Searches backward in the current source file for a pattern           |
| alias      | Creates aliases for <b>dbx</b> subcommands                           |
| assign     | Assigns a value to a variable                                        |
| attribute  | Displays information about all or selected attributes objects        |
| call       | Runs the object code associated with the named procedure or function |
| move       | Changes the next line to be displayed                                |
| multproc   | Enables or disables multiprocess debugging                           |
| mutex      | Displays information about all or selected mutexes                   |
| case       | Changes how the <b>dbx</b> debug program interprets symbols          |

| Subcommand     | Description                                                                                                                                    |
|----------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| catch          | Starts trapping a signal before that signal is sent to the application program                                                                 |
| clear          | Removes all stops at a given source line                                                                                                       |
| cleari         | Removes all breakpoints at an address                                                                                                          |
| condition      | Displays information about all or selected condition variables                                                                                 |
| cont           | Continues application program execution from the current stopping point until the program finishes or another breakpoint is encountered        |
| delete         | Removes the traces and stops corresponding to the specified event numbers                                                                      |
| detach         | Continues execution of application and exits the debug program                                                                                 |
| display memory | Displays the contents of memory                                                                                                                |
| down           | Moves the current function down the stack                                                                                                      |
| dump           | Displays the names and values of variables in the specified procedure                                                                          |
| edit           | Starts an editor on the specified file                                                                                                         |
| file           | Changes the current source file to the specified file                                                                                          |
| func           | Changes the current function to the specified procedure or function                                                                            |
| goto           | Causes the specified source line to be the next line run                                                                                       |
| map            | Displays information about load characteristics of the application                                                                             |
| next           | Runs the application program up to the next source line                                                                                        |
| nexti          | Runs the application program up to the next machine instruction                                                                                |
| print          | Prints the value of an expression or runs a procedure and prints the return code of that procedure                                             |
| prompt         | Changes the <b>dbx</b> command prompt                                                                                                          |
| quit           | Stops the <b>dbx</b> debug program                                                                                                             |
| registers      | Displays the values of all general-purpose registers, system-control registers, floating-point registers, and the current instruction register |
| rerun          | Begins execution of an application with the previous arguments                                                                                 |

| Subcommand | Description                                                                                    |
|------------|------------------------------------------------------------------------------------------------|
| return     | Continues running the application program until a return to the specified procedure is reached |
| where      | Displays a list of active procedures and functions                                             |
| which      | Displays the full qualification of the given identifier                                        |
| help       | Displays help information for <b>dbx</b> subcommands or topics                                 |
| ignore     | Stops trapping a signal before that signal is sent to the application program                  |
| list       | Displays lines of the current source file                                                      |

## 10.6.2 The **kdb** command

The **kdb** command is an interactive utility for examining an operating system image or the running kernel (KDB). The **kdb** command interprets and formats control structures in the system and provides miscellaneous functions for examining a dump. Root permissions are required for execution of the **kdb** command on the active system. This is required because the special file `/dev/mem` is used. To invoke the **kdb** command on a system image file, enter:

```
kdb SystemImageFile
```

Where *SystemImageFile* is either a file name or the name of the dump device. The **kdb** command can also be used to determine if the system is hanging or busy by hot-keying into the **kdb** system debugger.

Table 10-7 on page 296 lists some of the sample **kdb** subcommands.

Table 10-7 Sample **kdb** subcommands

| Subcommand | Description                                                                                                                                                                                                                     |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| h          | Displays the list of valid subcommands. The help subcommand can be reduced at only one topic.                                                                                                                                   |
| hist       | The hist subcommand prints a history of user input. An argument can be used to specify the number of historical entries to display. Each historical entry can be recalled and edited for use with the usual control characters. |

| Subcommand | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| e          | The exit subcommand exits the <b>kdb</b> command and the KDB Kernel Debugger. For the KDB Kernel Debugger, this subcommand exits the debugger with all breakpoints installed in memory. To exit the KDB Kernel Debugger without breakpoints, the ca subcommand should be invoked to clear all breakpoints prior to leaving the debugger. The exit subcommand leaves KDB session and returns to the system; all breakpoints are installed in memory. To leave KDB without breakpoints, the clear all subcommand must be invoked.                                                                                                                        |
| set        | The set subcommand can be used to list and set kdb toggles.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| f          | The stack subcommand displays all the stack frames from the current instruction as deep as possible. Interrupts and system calls are crossed and the user stack is also displayed, but KDB cannot directly access these symbols. Use the +x toggle to have hex addresses displayed (for example, to put a breakpoint on one of these addresses). The amount of data displayed may be controlled through the mst_wanted and display_stack_wanted options of the set subcommand. If invoked with no argument the stack for the current thread is displayed. The stack for a particular thread may be displayed by specifying its slot number or address. |
| trb        | Displays timer request blocks.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| trace      | Displays data from kernel trace buffers.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |

In the following example, the **kdb** command is used to look at the system dump.

```
kdb /dev/dumplv
The specified kernel file is a MP kernel.
Preserving 981564 bytes of symbol table
First symbol __mulh
Component Names:
 1) dmp_minimal [5 entries]
 2) proc [295 entries]
 3) thrd [757 entries]
 4) ldr [2 entries]
 5) errlg [3 entries]
 6) bos [7 entries]
 7) ipc [7 entries]
 8) vmm [20 entries]
 9) rtastrc [8 entries]
10) sscsidd [4 entries]
11) scdisk [10 entries]
12) lvm [2 entries]
13) tty [4 entries]
14) netstat [10 entries]
```

```

15) phxent_dd [5 entries]
16) kbddd [2 entries]
17) mousedd [2 entries]
Component Dump Table has 1143 entries
 START END <name>
0000000000003500 0000000001D7CEC8 _system_configuration+000020
000000002FF3B400 000000002FF80A70 __ublock+000000
000000002FF22FF4 000000002FF22FF8 environ+000000
000000002FF22FF8 000000002FF22FFC errno+000000
00000000E0000000 00000000F0000000 ameseg+10000000
PFT:
id.....0007
raddr....000000002000000 eaddr....000000002000000
size.....00000000 align.....00000000
valid..1 ros....0 holes..0 io.....0 seg....1 wimg...2

PVT:
id.....0008
raddr....0000000007D4000 eaddr....0000000000000000
size.....00000000 align.....00000000
valid..1 ros....0 holes..0 io.....0 seg....1 wimg...2
Dump analysis on CHRP_SMP_PCI POWER_PC POWER_RS64III machine with 4 cpu(s)
(64-
bit registers)
Processing symbol table...
.....done
(0)> trace
Trace channel[0 - 7]: 7
Trace Channel 7 (0 entries)
(0)> f
pvthread+000100 STACK:
[0002A880]waitproc_find_run_queue+000264 (00000000 [??])
[0002BA94]waitproc+0000E8 ()
[0006AC3C]procentry+000010 (??, ??, ??, ??)
(0)> trb
Usage: trb [CPU selector] [1-9]
 CPU selector is '*' for all CPUs, 'cpu n' for CPU n, default is current CPU

Timer Request Block Information Menu
 1. TRB Maintenance Structure - Routine Addresses
 2. System TRB
 3. Thread Specified TRB
 4. Current Thread TRB's
 5. Address Specified TRB
 6. Active TRB Chain
 7. Free TRB Chain
 8. Clock Interrupt Handler Information
 9. Current System Time - System Timer Constants
Please enter an option number: 8

```

```

Clock Interrupt Handler Information:
 intr->next.....00000000 intr->handler (clock()).00000000
 intr->bus_type.....00000000 intr->flags.....00000000
 intr->level.....00000000 intr->priority.....00000000
 intr->bid.....00000000 intr->i_count.....00000000

(0)> his
his
trace
7
f
trb
8

```

## 10.7 Summary

The flowchart shown at the start of this chapter (Figure 10-1 on page 234) is used in the summary, now with some suggestions included (Figure 10-12).

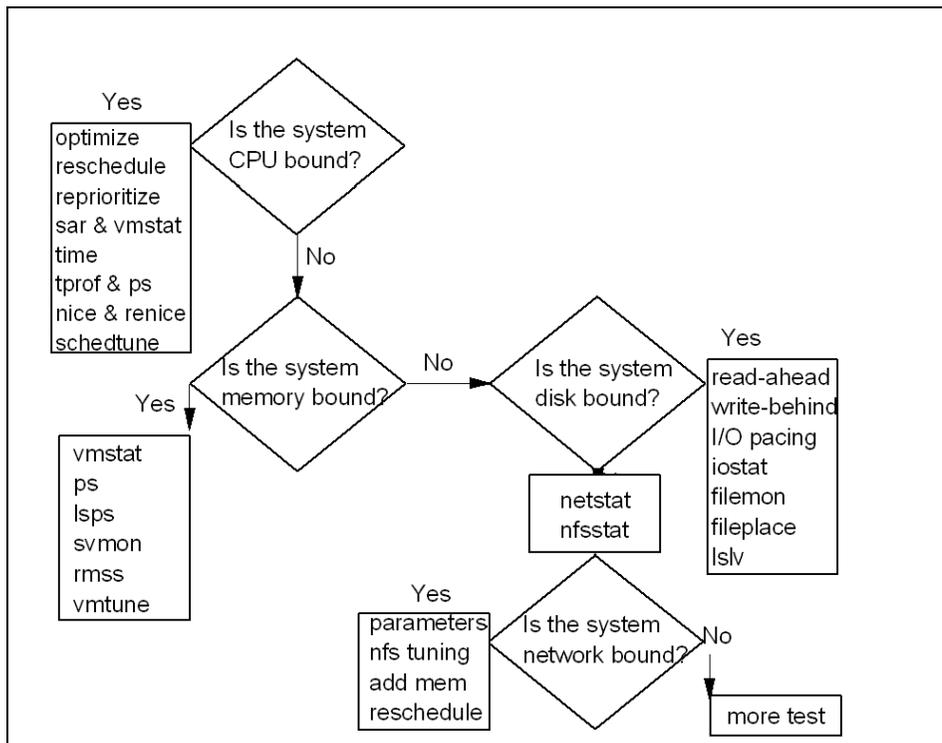


Figure 10-12 Performance tuning flowchart

## 10.8 Command summary

The following section shows a summary of some of the commands and their flags used for CPU performance problem determination.

### 10.8.1 The sar command

The **sar** command collects, reports, or saves system activity information.

The syntax of the **sar** command:

```
/usr/sbin/sar [{ -A | [-a] [-b] [-c] [-d] [-k] [-m] [-q] [-r]
[-u] [-V] [-v] [-w] [-y] }] [-P ProcessorIdentifier, ... | ALL] [-ehh [:mm [:ss]]]
[-fFile] [-iSeconds] [-oFile] [-shh [:mm [:ss]]] [Interval [Number]]
```

The most commonly used flags are provided in Table 10-8.

Table 10-8 Commonly used flags of the sar command

| Flags  | Description                                                                                     |
|--------|-------------------------------------------------------------------------------------------------|
| -u     | Displays %idle, %sys, %usr, and %wio                                                            |
| -P ALL | Reports per-processor statistics for each individual processor, and globally for all processors |

### 10.8.2 The ps command

The **ps** command shows the current status of the processes.

The syntax of the **ps** command is (X/Open, then Berkeley):

```
ps [-A] [-N] [-a] [-d] [-e] [-f] [-k] [-l] [-F format]
[-o Format] [-c Clist] [-G Glist] [-g Glist] [-m] [-n NameList] [-p Plist] [-t Tlist]
[-U Ulist] [-u Ulist]

ps [a] [c] [e] [ew] [eww] [g] [n] [U] [w] [x]
[l | s | u | v] [t Tty] [ProcessNumber]
```

The most commonly used flags are provided in Table 10-9.

Table 10-9 Commonly used flags of the ps command

| Flags | Description   |
|-------|---------------|
| -f    | Full listing. |

| Flags | Description                                                                                                                  |
|-------|------------------------------------------------------------------------------------------------------------------------------|
| -l    | Long listing.                                                                                                                |
| u     | Displays user-oriented output. This includes the USER, PID, %CPU, %MEM, SZ, RSS, TTY, STAT, STIME, TIME, and COMMAND fields. |
| v     | Displays the PGIN, SIZE, RSS, LIM, TSIZ, TRS, %CPU, and %MEM fields.                                                         |

### 10.8.3 The netstat command

The **netstat** command shows the network status.

The syntax for the **netstat** command is as follows.

To display active sockets for each protocol or routing table information:

```
/bin/netstat [-n] [{ -A -a } | { -r -C -i -I Interface }]
[-f AddressFamily] [-p
Protocol] [Interval] [System]
```

To display the contents of a network data structure:

```
/bin/netstat [-m | -s | -ss | -u | -v] [-f AddressFamily] [-p
Protocol] [Interval] [System]
```

To display the packet counts throughout the communications subsystem:

```
/bin/netstat -D
```

To display the network buffer cache statistics:

```
/bin/netstat -c
```

To display the data link provider interface statistics:

```
/bin/netstat -P
```

To clear the associated statistics:

```
/bin/netstat [-Zc | -Zi | -Zm | -Zs]
```

The most commonly used flags are provided in Table 10-10.

*Table 10-10 Commonly used flags of the netstat command*

| Flags | Description      |
|-------|------------------|
| -i    | Interface status |
| -m    | Mbuf information |

| Flags                | Description                                          |
|----------------------|------------------------------------------------------|
| -Z { c   l   m   s } | Clears the statistics defined by the additional flag |
| -v                   | Statistics for each CDLI                             |

## 10.8.4 The nfsstat command

The **nfsstat** command displays statistical information about the Network File System (NFS) and Remote Procedure Call (RPC) calls.

The syntax of the **nfsstat** command is:

```
/usr/sbin/nfsstat [-c] [-s] [-n] [-r] [-z] [-m]
```

The most commonly used flags are provided in Table 10-11.

Table 10-11 Commonly used flags of the nfsstat command

| Flags | Description                 |
|-------|-----------------------------|
| -r    | Displays RPC information    |
| -s    | Displays server information |
| -c    | Displays client information |

## 10.9 Quiz

The following assessment questions help verify your understanding of the topics discussed in this chapter.

1. A system administrator is experiencing some performance problems. After running the **vmstat** command, the following output appeared:

```
kthr memory page faults cpu

r b avm fre re pi po fr sr cy in sy cs us sy id wa
2 0 44298 340 0 0 0 1 2 0 138 360 64 65 30 0 5
2 0 44298 358 0 0 0 1 2 0 138 360 64 65 30 0 5
2 0 44298 358 0 0 0 1 2 0 138 360 64 65 30 0 5
2 0 44298 358 0 0 0 1 2 0 138 360 64 65 30 0 5
2 0 44298 358 0 0 0 1 2 0 138 360 64 65 30 0 5
2 0 44245 358 0 0 0 1 2 0 138 360 64 65 30 0 5
```

Which of the following outputs best describes the cause of the problem?

- A. The machine is CPU bound.
  - B. The machine needs memory optimized.
  - C. The machine requires more paging space.
  - D. A user program is causing unnecessary paging.
2. Which of the following commands should be used to observe the number of threads in the run queue?
- A. **bf**
  - B. **iostat**
  - C. **filemon**
  - D. **vmstat**
3. Which of the following commands should be used to show the percentage of time that any given disk was busy?
- A. **ps**
  - B. **tprof**
  - C. **iostat**
  - D. **vmstat**
4. Which WLM command will reload the assignment rules of the current running configuration into the kernel *without* reloading the class definitions?
- A. **wlmcntrl -u -d config\_dir**
  - B. **wlmcntrl -u -d ""**
  - C. **wlmcntrl -u config\_dir**
  - D. **wlmcntrl -u -a**
5. In general, when the disk utilization exceeds a certain percentage, processes are waiting longer than necessary for I/O to complete because most UNIX processes block (or sleep) while waiting for their I/O requests to complete or be serviced by the interrupt handler. What is that percentage threshold?
- A. 45 percent
  - B. 60 percent
  - C. 70 percent
  - D. 85 percent

6. Which command will show system calls and fork rates?
- A. **sar -c**
  - B. **vmtime**
  - C. **vmstat**
  - D. **svmon**
7. A system appears to be experiencing network performance problems on the Ethernet adapter. A network engineer reports seeing an abnormally large number of collisions on the network. However, a **netstat -i** is run and shows 0 collisions. What does this statistic indicate?
- A. There are no collisions in the collision domain.
  - B. There were no packet errors caused by collisions.
  - C. The system did not generate any collisions.
  - D. This statistic is not meaningful for detecting collisions.
8. Which of the following pieces of information is not included in the **tprof** report?
- A. Idle
  - B. PID
  - C. TID
  - D. Kernel
9. Which WLM flag/option allows all classes to have access to the whole resource set of the system, regardless of whether they use a restricted resource set?
- A. **wlmcntrl -d**
  - B. **wlmcntrl -p**
  - C. **wlmcntrl -g**
  - D. **wlmcntrl -o**

## 10.9.1 Answers

The following are the preferred answers to the questions provided in this section.

1. A
2. D
3. C
4. B
5. C
6. A
7. D
8. A
9. C





# Software updates

This chapter covers the AIX software update procedures, including the following topics:

- ▶ An overview of the process
- ▶ Installing a software patch
- ▶ Software inventory

## 11.1 Overview

The biggest goal of all system administrators is to have an operating system running well with current software installed on it. Installation of software fixes is one of the actions an administrator must perform to keep a system error free. Software problems most often occur when changes have been made to the system, and either the prerequisites have not been met (for example, system firmware is not at the minimum required level) or instructions have not been followed exactly in order. You, as a system administrator, should carefully choose the methodology to maintain your software inventory.

### 11.1.1 Terminology

The following terms are useful for understanding software packaging:

|                       |                                                                                                                                                                                                                               |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Fileset</b>        | The smallest individually installable unit. It is a collection of files that provides a specific function. An example of a fileset is bos.net.tcp.nfs 4.3.3.0.                                                                |
| <b>Fileset update</b> | An individually installable update. Fileset updates either enhance or correct a defect in a previously installed fileset. An example of a fileset update is bos.net.tcp.nfs 4.3.3.10.                                         |
| <b>Package</b>        | Contains a group of filesets with a common function. It is a single, installable image. An example of a package is bos.net.                                                                                                   |
| <b>LPP</b>            | Licensed Program Product (LPP) is a complete software product collection, including all packages and filesets. For example, the Base Operating System (BOS) itself is an LPP, which is a collection of packages and filesets. |
| <b>PTF</b>            | Program Temporary Fix (PTF). The PTF is an updated or fixed fileset (or group of filesets). Each fix has an Authorized Program Analysis Report (APAR) number.                                                                 |

### 11.1.2 Software layout

Each software component is divided into three parts that support code serving and diskless workstation:

|             |                                                                                                                                                                                                                                                                                          |
|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>root</b> | The root part of a software product contains the part of the product that cannot be shared. In a client/server environment, these are the files for which there must be a unique copy for each client of a server. Most of the root software is associated with the configuration of the |
|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

machine or product. In a standard system, the root parts of a product are stored in the root (/) file tree. The /etc/objrepos directory contains the root part of an installable software product.

**usr** The usr part of a software product contains the part of the product that can be shared by machines that have the same hardware architecture. Most of the software that is part of a product usually falls into this category. In a standard system, the usr parts of products are stored in the /usr file tree.

**share** The share part of a software product contains the part of the product that can be shared among machines, even if they have different hardware architectures (this would include nonexecutable text or data files). For example, the share part of a product might contain documentation written in ASCII text or data files containing special fonts.

To verify that the root (/), /usr, and /usr/share parts of the system are valid with each other, use the following command:

```
lppchk -v
```

This command verifies that all software products installed on the / (root) file system are also installed on the /usr file system and, conversely, all the software products installed in the /usr file system are also installed on the / (root) file system.

### 11.1.3 Software states

The installed software or software update can stay in one of the following states:

- ▶ Applied
- ▶ Committed

If the service update was not committed during installation, then you must commit it after installation once you have decided that you will not be returning to the previous version of the software. Committing the updated version of the service deletes all previous versions from the system and recovers the disk space that was used to store the previous version. When you are sure that you want to keep the updated version of the software, you should commit it. To commit the fileset bos.sysmgt.trace that is currently applied but not committed, use:

```
installp -c bos.sysmgt.trace
```

**Note:** Before installing a new set of updates, you should consider committing any previous updates that have not yet been committed.

If you decide to return to the previous version of the software, you must reject the updated version that was installed. Rejecting a service update deletes the update from the system and returns the system to its former state. A service update can only be rejected if it has not yet been committed. Once committed, there is no way to delete an update except by removing the entire fileset, or by force-installing the fileset back to a previous level.

When you install a base level fileset, it is automatically committed during installation. If you want to delete a fileset, it must be removed (as opposed to rejected) from the system. A fileset is always removed with all of its updates.

To display the installation and update history information for the `bos.sysmgt.trace` fileset, use:

```
ls1pp -h bos.sysmgt.trace
Fileset Level Action Status Date Time

Path: /usr/lib/objrepos
bos.sysmgt.trace
 4.3.3.0 COMMIT COMPLETE 06/15/00 09:57:28
 4.3.3.11 COMMIT COMPLETE 06/16/00 11:19:13

Path: /etc/objrepos
bos.sysmgt.trace
 4.3.3.0 COMMIT COMPLETE 06/15/00 09:57:33
 4.3.3.11 COMMIT COMPLETE 06/16/00 11:19:14
```

As shown, the fileset `bos.sysmgt.trace` was once updated. It is now in the committed state at the fix level 4.3.3.11.

If something goes wrong during the software installation that causes the installation to be prematurely canceled or interrupted, a cleanup must be done. To do this, use `smitty maintain_software` or use the `installp` command:

```
installp -C
```

Figure 11-1 on page 311 shows how to clean up after an interrupted installation using SMIT.

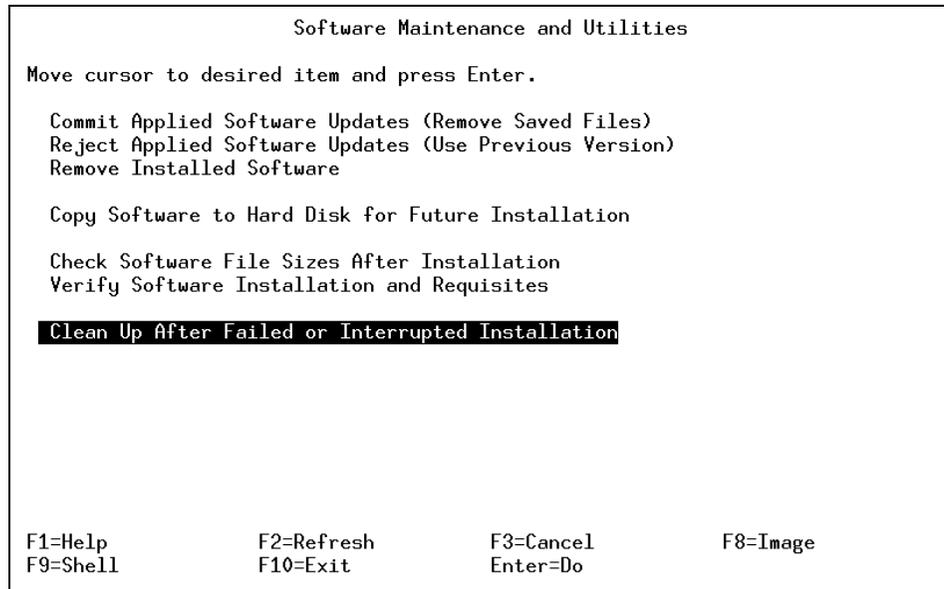


Figure 11-1 SMIT Software Maintenance and Utilities panel

## 11.2 Installing a software patch

Once you have AIX installed, you may want to upgrade or enhance the software on your system. To do this, there are two special bundles:

### Update bundle

Collection of fixes and enhancements that update software products on the system. This will include updated filesets. For example, a fileset may be updated from 4.3.3.0 to 4.3.3.10. Applying an updated bundle will not change the level of the operating system.

**Maintenance level bundle** Collection of fixes and enhancements that upgrade the operating system to the latest level. For example, a maintenance level bundle can upgrade the operating system from AIX Version 4.3.2 to AIX Version 4.3.3.

Software fixes are identified using one of the following conventions:

- ▶ *fileset.version.release.modification.fix*. Modification level is used to describe functional support. Fix levels describe a fix change.
- ▶ PTF number, such as U469083.

- ▶ APAR number, such as IY00301.

It is simple to obtain software updates for AIX. Check the following Web page and download what is required:

<http://techsupport.services.ibm.com/support/rs6000.support/databases>

## 11.2.1 Software patch installation procedure

Before installing optional software or service updates, complete the following prerequisites:

1. AIX BOS must be installed on your system.
2. The software you are installing is available on either CD-ROM, tape, or diskette; it is located in a directory on your system; or, if your computer is a configured Network Installation Management (NIM) client, it is in an available lpp\_source resource.
3. If you are installing service updates and do not have a current backup of your system, back up your system before any installation.
4. If the file system has been modified, it is a good idea to back it up separately before updates are applied, since it is possible that the update process may replace configuration files.
5. Check if there is enough space in the file system.
6. Log in as a root user.

The easiest way to install software updates is SMIT. Use **smitty install\_update** to access the installation menu. The appropriate menu is shown in Figure 11-2 on page 313.

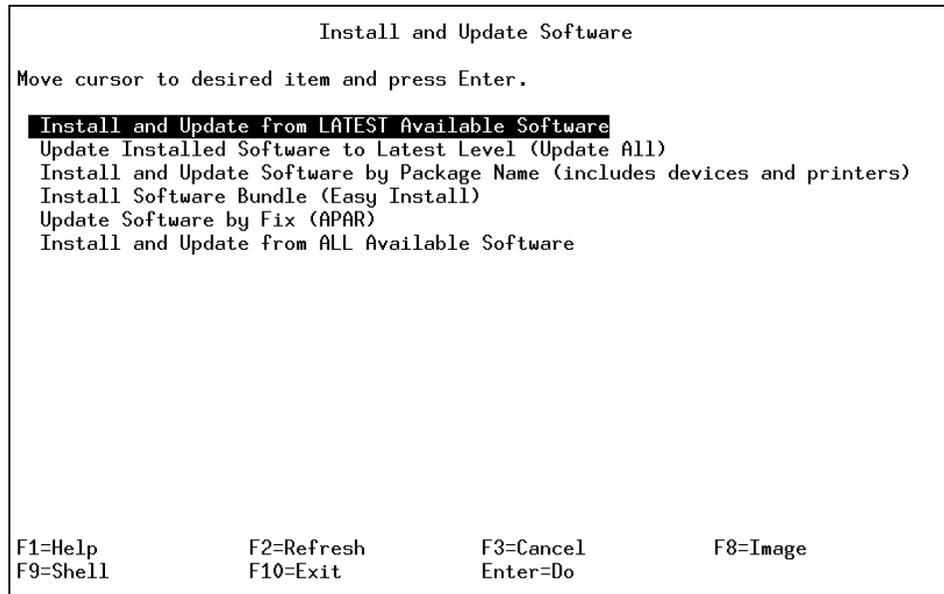


Figure 11-2 Install and Update Software panel

The major menu options are as follows:

- ▶ Install and Update from LATEST Available Software  
This option allows you to install or update software from the latest level software available on the installation media.
- ▶ Update Installed Software to the Latest Level  
Enables you to update all currently installed filesets to the latest level available on the installation media. This option is also used to update currently installed software to a new maintenance level.
- ▶ Update Software by Fix (APAR)  
Enables you to install fileset updates that are grouped by some relationship and identified by a unique APAR. A fix to an APAR can consist of one or more fileset updates.

If you are more comfortable with a shell, all of this can be done using the **installp** or **instfix** command.

- ▶ To install all filesets within the bos.net software package (located in the /tmp/install.images directory) and expand file systems if necessary, enter:

```
installp -aX -d/tmp/install.images bos.net
```

- To install all filesets associated with fix IX38794 from the CD-ROM, enter:

```
instfix -k IX38794 -d /dev/cd0
```

**Note:** If you choose to apply the updates during installation (rather than committing them at installation time), you can still reject those updates later. If a particular update is causing problems on your system, you can reject that update without having to reject all the other updates that you installed. Once you are convinced that the updates cause no problems, you may want to commit those updates to retrieve the disk space that is used to save the previous levels of that software.

After you have installed a new fix, use the **lppchk** command to check if the installation was successful. The **lppchk** command verifies that files for an installable software product (fileset) match the Software Vital Product Data database information for file sizes, checksum values, or symbolic links. The useful flags are shown in Table 11-1.

*Table 11-1 Commonly used flags of the lppchk command*

| Flag | Description                                                                                                                                        |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| -c   | Performs a checksum operation on the input file list items and verifies that the checksum and the file size are consistent with the SWVPD database |
| -f   | Checks that the file list items are present and that the file size matches the SWVPD database                                                      |
| -l   | Verifies symbolic links for files, as specified in the SWVPD database                                                                              |

If you have been installing software using SMIT, the screen returns to the top of the list of messages that are displayed during installation. You can review the message list as described in the next step, or you can exit SMIT and review the `$HOME/smit.log` file.

After you check that the installation is successful, you should create a new boot image using the **bosboot** command:

```
bosboot -ad /dev/hdiskX
```

## 11.3 Software inventory

After all the software installations, you can check what is really installed with the `instfix` and `lslpp` commands.

1. To display the most recent level, state, description and all updates of the `bos.sysmgmt.trace` fileset, run the following command:

```
lslpp -La bos.sysmgmt.trace
Fileset Level State Description

bos.sysmgmt.trace 4.3.3.0 C Software Trace Service Aids
 4.3.3.11 C Software Trace Service Aids
...
```

2. To see whether fix `IX78215` is installed or information about each fileset associated with it, run the following command:

```
instfix -ik IX78215 -v
IX78215 Abstract: trace allocates too much memory

Fileset bos.sysmgmt.trace:4.3.1.1 is applied on the system.
All filesets for IX78215 were found.
```

3. To list maintenance level updates, enter:

```
instfix -i -tp
All filesets for 4.3.1.0_AIX_ML were found.
All filesets for 4.3.2.0_AIX_ML were found.
All filesets for 4.3.1.0_AIX_ML were found.
All filesets for 4.3.2.0_AIX_ML were found.
All filesets for 4.3.3.0_AIX_ML were found.
```

Or run `instfix -i | grep ML`.

## 11.4 Command summary

This section shows a summary of some of the commands and their flags that are used for CPU performance problem determination.

### 11.4.1 The `lslpp` command

The `lslpp` command displays information about installed filesets or fileset updates. The command has the following syntax:

```
lslpp { -f | -h | -i | -L } [-a] [FilessetName ... | FixID ... | all]
```

The most commonly used flags are provided in Table 11-2 on page 316.

Table 11-2 Commonly used flags of the `lspp` command

| Flag | Description                                                                                                                                                           |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -a   | Displays all the information about filesets specified when combined with other flags.                                                                                 |
| -f   | Displays all the information about filesets specified when combined with other flags.                                                                                 |
| -h   | Displays the installation and update history information for the specified fileset.                                                                                   |
| -i   | Displays the product information for the specified fileset.                                                                                                           |
| -L   | Displays the name, most recent level, state, and description of the specified fileset. Part information (usr, root, and share) is consolidated into the same listing. |
| -w   | Lists the fileset that owns this file.                                                                                                                                |

## 11.4.2 The `installp` command

The `installp` command installs available software products in a compatible installation package.

The most commonly used flags are provided in Table 11-3.

Table 11-3 Commonly used flags of the `installp` command

| Flag | Description                                      |
|------|--------------------------------------------------|
| -ac  | Commits                                          |
| -g   | Includes requisites                              |
| -N   | Overrides saving of existing files               |
| -q   | Quiet mode                                       |
| -w   | Does not place a wildcard at end of fileset name |
| -X   | Attempts to expand file system size if needed    |
| -d   | Inputs device                                    |
| -l   | List of installable filesets                     |
| -c   | Commits an applied fileset                       |
| -C   | Cleans up after a failed installation            |
| -u   | Uninstalls                                       |

| Flag | Description                               |
|------|-------------------------------------------|
| -r   | Rejects an applied fileset                |
| -p   | Preview of installation                   |
| -e   | Defines an installation log               |
| -F   | Forces overwrite of same or newer version |

### 11.4.3 The instfix command

The **instfix** command installs filesets associated with keywords or fixes. The command has the following syntax:

```
instfix [-T] [-s String] [-k Keyword] [-d Device] [-i]
```

The most commonly used flags are provided in Table 11-4.

Table 11-4 Commonly used flags of the instfix command

| Flag              | Description                                                            |
|-------------------|------------------------------------------------------------------------|
| -d <i>device</i>  | Specifies the input device                                             |
| -i                | Displays whether fixes or keywords are installed                       |
| -k <i>keyword</i> | Specifies an APAR number or keyword to be installed                    |
| -s <i>string</i>  | Searches for and displays fixes on media containing a specified string |
| -T                | Displays the entire list of fixes present on the media                 |

### 11.4.4 The lppchk command

The **lppchk** command verifies files of an installable software product. The command has the following syntax:

```
lppchk { -c | -f | -l | -v } [-0 { [r] [s] [u] }]
[ProductName [FileList ...]]
```

The most commonly used flags are provided in Table 11-5.

Table 11-5 Commonly used flags of the lppchk command

| Flag | Description                                                                                                                                   |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| -c   | Performs a checksum operation on the file list items and verifies that the checksum and the file size are consistent with the SWVPD database. |

| Flag           | Description                                                                                               |
|----------------|-----------------------------------------------------------------------------------------------------------|
| -f             | Checks that the file list items are present and the file size matches the SWVPD database.                 |
| -l             | Verifies symbolic links for files, as specified in the SWVPD database.                                    |
| -O {[r][s][u]} | Verifies the specified parts of the program. The flags specify the following parts: root, share, and usr. |

## 11.5 Quiz

The following assessment question helps verify your understanding of the topics discussed in this chapter.

1. A system administrator must determine if the operating system is in a consistent state or if it does not have the fileset installed correctly. Which of the following commands should be used?
  - A. `ls1pp -v`
  - B. `ls1v -v`
  - C. `lsvg -v`
  - D. `lppchk -v`

## 11.5.1 Answers

The following is the preferred answer to the question provided in this section.

1. D

## 11.6 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

1. Use the various flags of the **lppchk** command to verify the checksum, the file sizes, symbolic links, and requisites of the software products installed.
2. Use the **ls1pp** command to find out which fileset is used to package a given command.
3. Use the **instfix** command to list fixes installed on your system.
4. Use the FixDist utility to download AIX fixes.
5. Use the **ls1pp** command to display state, description, and all updates of the filesets.





## Online documentation

AIX 5L provides an optionally installable component for Web-based documentation: The Documentation Library Service. It allows you to search online HTML documents. It provides a search form that appears in your Web browser. When you type words into the search form, it searches for the words and then presents a search results page that contains links that lead to the documents that contain the target words.

You can set up one of your AIX systems to be the documentation server and all other systems as documentation clients. This will allow documentation to be installed on only one system, and all other systems can access this system without needing the documentation installed locally.

You need the following products and components installed for a complete set of services:

- ▶ For the client:
  - A Web browser
  - The bos.docsearch.client.\* filesets (for AIX integration)
- ▶ For the documentation server (which may also act as a client):
  - The entire bos.docsearch package
  - The documentation libraries
  - A Web browser

- A Web server

The browser must be a forms-capable browser, and the Web server must be CGI compliant.

If you are planning on integrating your own documentation on the documentation server, you will also need to build the documents' indexes.

Except for the end-user tasks described in 12.6, "Invoking the Documentation Library Service" on page 326, you need root authority to perform the installation and configuration tasks.

There are a variety of ways to install the documentation, Web server, and Document Library Service. You can use the Configuration Assistant TaskGuide, Web-Based Systems Management, or SMIT.

The easiest way for a non-technical user to install and configure Documentation Library Services is by using the Configuration Assistant TaskGuide. To run the Configuration Assistant TaskGuide, use the `configassist` command, then select the item titled **Configure Online Documentation and Search**.

If you would rather install Documentation Library Services manually, you can use SMIT.

## 12.1 Installing the Web browser

Use `smitty install` to install Netscape supplied on the AIX Version 5.1 Expansion Pack CD-ROM, as shown in Figure 12-1.

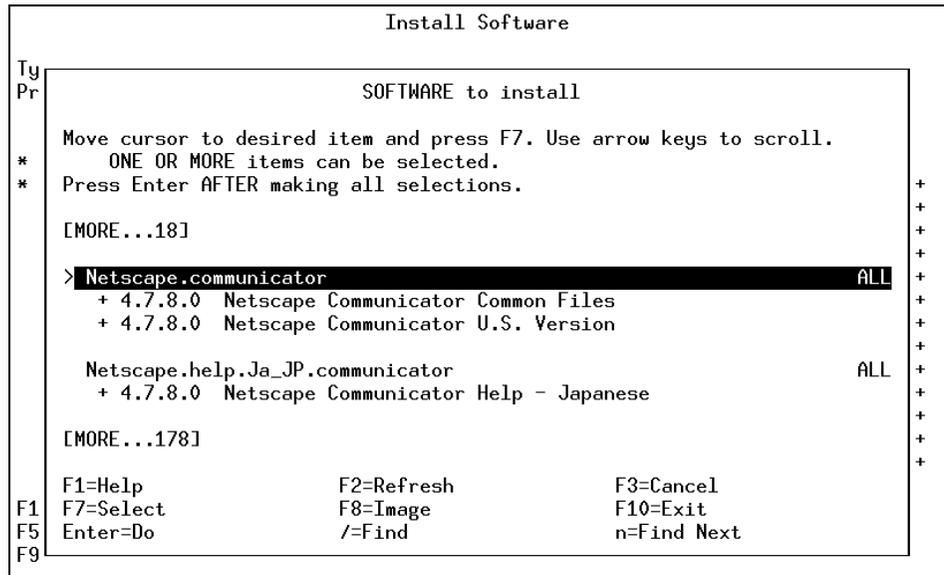


Figure 12-1 Netscape filesets

If you are installing the Netscape browser from other sources, or you are installing other Web browsers, follow the installation instructions that come with the software. Note that there will not be any records in the ODM if your product source is not in `installp` format.

## 12.2 Installing the Web server

You may install any CGI-compliant Web Server. The IBM HTTP Web server used here is supplied on the AIX Version 5.1 Expansion Pack CD-ROM. Figure 12-2 on page 324 shows the filesets installed.

```

Install Software

Ty
Pr
 SOFTWARE to install
* Move cursor to desired item and press F7. Use arrow keys to scroll.
* ONE OR MORE items can be selected.
* Press Enter AFTER making all selections.
[MORE...91]
> http_server.admin ALL
 + 1.3.12.4 HTTP Server Administration (Run-time)
> http_server.base ALL
 + 1.3.12.4 HTTP Server Base Run-Time
 + 1.3.12.4 HTTP Server Source Code
[MORE...105]

F1=Help F2=Refresh F3=Cancel
F7=Select F8=Image F10=Exit
F5=Enter=Do /=Find n=Find Next
F9

```

Figure 12-2 HTTP filesets

If you are installing another Web server, follow the installation instructions that come with the software. Note that there will not be any records in the ODM if your product source is not in installp format.

## 12.3 Installing the Documentation Library Service

The Documentation Library Service is on AIX Version 5.1 installation CD-ROMs. Install the client portions for a client AIX image or install the entire bos.docsearch package for a documentation server. The following filesets are the prerequisites for other Documentation Library Service filesets (such as IMNSearch):

- ▶ bos.docsearch.client.Dt
- ▶ bos.docsearch.client.com
- ▶ bos.docsearch.rte

For the documentation clients, you need only a Web browser. Installation of the bos.docsearch.client fileset will give you the CDE desktop icon and the **docsearch** command. Refer to 12.6, “Invoking the Documentation Library Service” on page 326, for further details.

Use **smitty list\_installed** to check whether you have the document library filesets installed, as shown in Figure 12-3 on page 325.

```

COMMAND STATUS
Command: OK stdout: yes stderr: no
Before command completion, additional instructions may appear below.

[TOP]
█ Fileset

bos.docsearch.client.Dt 5.1.0.0 C F DocSearch Client CDE Applicat
ion
bos.docsearch.client.com 5.1.0.0 C F DocSearch Client Common Files
Integration
bos.docsearch.rte 5.1.0.25 C F DocSearch Runtime

State codes:
A -- Applied.
[MORE...12]

F1=Help F2=Refresh F3=Cancel F6=Command
F8=Image F9=Shell F10=Exit /=Find
n=Find Next

```

Figure 12-3 Documentation Library Service filesets

## 12.4 Configuring the Documentation Library Service

Use either **wsm** or **smitty** to configure the Documentation Library Service. If you used the Configuration Assistant TaskGuide to install and configure the Documentation Library Service, you will not need to perform any further configuration.

For **wsm**, double-click the **Internet Environment** icon, or you can use **smitty web\_configure** to configure the following:

- ▶ Default browser
 

Type into the field the command that launches the browser that you want to be the default browser for all users on this computer, for example, `/usr/prod/bin/netscape`. This will set the `/etc/environment` variable `DEFAULT_BROWSER` to the string you type in.
- ▶ Documentation and search server
 

You can define the Documentation Library Server location to be:

  - None (disabled)

- Remote computer

Type the remote documentation server name. The default TCP/IP port address is 80. Change it to the port address used by the documentation server.

- Local (this computer)

If you are using Lotus Domino Go Webserver or IBM HTTP Server in the default location, all the default settings of the cgi-bin directory and HTML directory will have been filled in for you. If you are using other Web servers, or you are not using the default location, you have to fill in your cgi-bin directory and the HTML directory that the Web server requires. You may change the port address used by the server. If you change the port address, you have to use the same address for all of your documentation clients.

## 12.5 Installing online manuals

You can either install the documentation information onto the hard disk or mount the documentation CD-ROM in the CD-ROM drive. Mounting the CD-ROM will save some amount of hard disk space, but it requires the CD-ROM to be kept in the CD-ROM drive at all times. Also, searching the documentation from the CD-ROM drive can be significantly slower (in some cases, up to 10 times slower) than searching the information if it is installed on a hard disk.

Use `smitty install_latest` to install the online manuals onto the hard disk. The fileset `bos.docregister` is a prerequisite for all online manuals. It will be automatically installed the first time you install any online manuals, even if you have not selected this fileset.

## 12.6 Invoking the Documentation Library Service

You must log out and log in again after the Documentation Library Service has been configured so that you will pick up the environment variables set up during the configuration.

If you are running the CDE desktop environment, double-click the **Documentation Library** icon in the Application Manager window.

Alternatively, you can use the command `docsearch` to invoke the Documentation Library Service. Your Web browser will start, and you should see the Documentation Library Service page. Netscape is used as the default Web browser for this discussion.

You can invoke the Documentation Library Service without installing the docsearch client component. In fact, you do not even need to invoke the Documentation Library Service from an AIX machine. You can do this by first invoking the browser and entering the following URL:

```
http://server_name[:port_number]/cgi-bin/ds_form
```

This URL points to a global search form on the document server where the name of the remote server is given in *server\_name*. The *port\_number* only needs to be entered if the port is not 80.

If you have not run Netscape previously, a series of informational messages and windows will be shown while Netscape is setting up the environment in your home directory. This is standard behavior for the first execution of Netscape. The messages will not be shown the next time you start Netscape.

The top part of the Documentation Library Service page allows you to specify your search criteria, and the bottom part shows what online manuals have been installed. Figure 12-4 on page 328 shows the Documentation Library Service page with only the command reference manuals and the programming guide manuals installed.

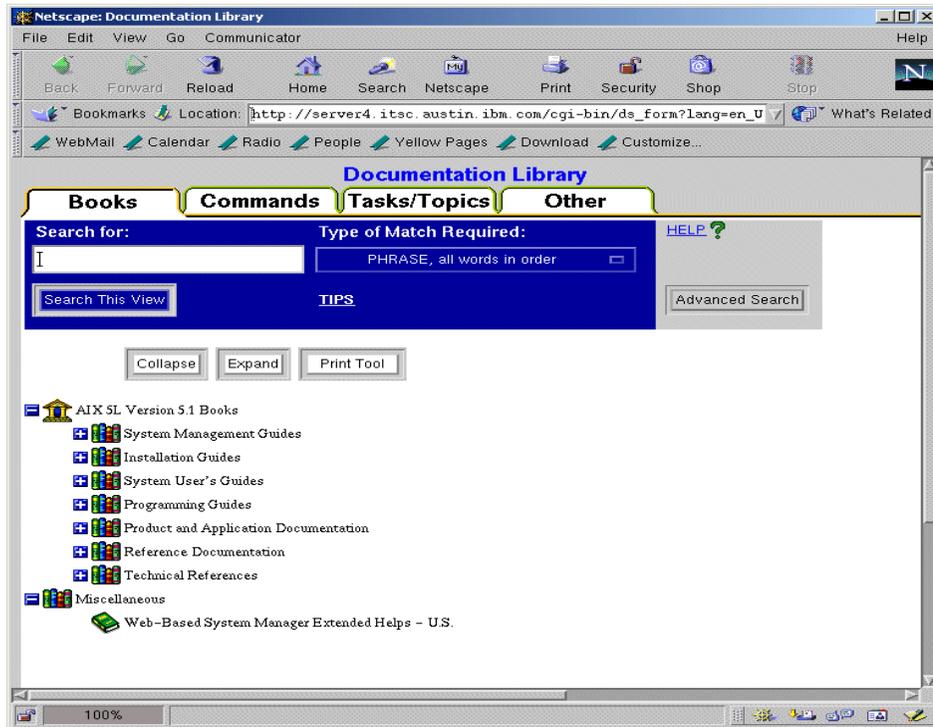


Figure 12-4 Documentation Library Service

If you have a problem starting the Documentation Library Service, check the following environment variables. These environment variables may be set, displayed, and changed using SMIT. Start **smitty**, select **System Environments**, then select **Internet and Documentation Services**.

- ▶ On the client machine:
  - a. Invoke the Web browser manually and enter the URL `http://server_name[:port_number]/cgi-bin/ds_form` to ensure that the server is up and running.
  - b. Ensure that the `DEFAULT_BROWSER` variable is set to the command for starting your Web browser. Use the command `echo $DEFAULT_BROWSER` to find out the command used in starting the browser. Test whether that command can bring up the browser by manually entering it on the command line.
  - c. Ensure that the `DOCUMENT_SERVER_MACHINE_NAME` variable is set to the document server's host name or IP address.

- d. Ensure that the `DOCUMENT_SERVER_PORT` variable is set to the port address used by the document server's port address.
- ▶ On the server machine:
  - Ensure that the `DEFAULT_BROWSER` variable is set to the command for starting your Web browser.

Use the command `echo $DEFAULT_BROWSER` to find out the command used in starting the browser. Test whether that command can bring up the browser by manually entering it on the command line.

1. Ensure that the `DOCUMENT_SERVER_MACHINE_NAME` variable is set to the local host name.
2. Ensure that the `DOCUMENT_SERVER_PORT` variable is set to the port address used by the local Web server.
3. Ensure that the `CGI_DIRECTORY` variable is set to the correct cgi-bin directory used by the local Web server.
4. Ensure that the `DOCUMENT_DIRECTORY` is set to the directory where the symbolic links `doc_link` and `ds_images` reside.
5. If you are not using the default directory, ensure that you have defined the necessary directory mapping in your Web server configuration file so that the directory can be resolved.

## 12.7 Exercises

The following exercises provide sample topics for self study. They will help ensure comprehension of this chapter.

1. Install a Web browser.
2. Install a Web server.
3. Install the Document Library Services fileset.
4. Install some online manuals.
5. Configure Document Library Services.
6. Access the online manuals using the `docsearch` command and from a Web browser on other systems.



# Abbreviations and acronyms

|                |                                                              |               |                                                                   |
|----------------|--------------------------------------------------------------|---------------|-------------------------------------------------------------------|
| <b>ABI</b>     | Application Binary Interface                                 | <b>BIST</b>   | Built-In Self-Test                                                |
| <b>AC</b>      | Alternating Current                                          | <b>BLAS</b>   | Basic Linear Algebra Subprograms                                  |
| <b>ACL</b>     | Access Control List                                          | <b>BLOB</b>   | Binary Large Object                                               |
| <b>ADSM</b>    | ADSTAR Distributed Storage Manager                           | <b>BLV</b>    | Boot Logical Volume                                               |
| <b>ADSTAR</b>  | Advanced Storage and Retrieval                               | <b>BOOTP</b>  | Boot Protocol                                                     |
| <b>AFPA</b>    | Adaptive Fast Path Architecture                              | <b>BOS</b>    | Base Operating System                                             |
| <b>AFS</b>     | Andrew File System                                           | <b>BSC</b>    | Binary Synchronous Communications                                 |
| <b>AH</b>      | Authentication Header                                        | <b>CAD</b>    | Computer-Aided Design                                             |
| <b>AIX</b>     | Advanced Interactive Executive                               | <b>CAE</b>    | Computer-Aided Engineering                                        |
| <b>ANSI</b>    | American National Standards Institute                        | <b>CAM</b>    | Computer-Aided Manufacturing                                      |
| <b>APAR</b>    | Authorized Program Analysis Report                           | <b>CATE</b>   | Certified Advanced Technical Expert                               |
| <b>API</b>     | Application Programming Interface                            | <b>CATIA</b>  | Computer-Graphics Aided Three-Dimensional Interactive Application |
| <b>ARP</b>     | Address Resolution Protocol                                  | <b>CCM</b>    | Common Character Mode                                             |
| <b>ASCI</b>    | Accelerated Strategic Computing Initiative                   | <b>CD</b>     | Compact Disk                                                      |
| <b>ASCII</b>   | American National Standards Code for Information Interchange | <b>CDE</b>    | Common Desktop Environment                                        |
| <b>ASR</b>     | Address Space Register                                       | <b>CDLI</b>   | Common Data Link Interface                                        |
| <b>ATM</b>     | Asynchronous Transfer Mode                                   | <b>CD-R</b>   | CD Recordable                                                     |
| <b>AuditRM</b> | Audit Log Resource Manager                                   | <b>CD-ROM</b> | Compact Disk-Read Only Memory                                     |
| <b>AUI</b>     | Attached Unit Interface                                      | <b>CE</b>     | Customer Engineer                                                 |
| <b>AWT</b>     | Abstract Window Toolkit                                      | <b>CEC</b>    | Central Electronics Complex                                       |
| <b>BCT</b>     | Branch on CounT                                              | <b>CFD</b>    | Computational Fluid Dynamics                                      |
| <b>BFF</b>     | Backup File Format                                           | <b>CGE</b>    | Common Graphics Environment                                       |
| <b>BI</b>      | Business Intelligence                                        | <b>CHRP</b>   | Common Hardware Reference Platform                                |
| <b>BIND</b>    | Berkeley Internet Name Daemon                                |               |                                                                   |

|               |                                                       |               |                                                     |
|---------------|-------------------------------------------------------|---------------|-----------------------------------------------------|
| <b>CISPR</b>  | International Special Committee on Radio Interference | <b>DHCP</b>   | Dynamic Host Configuration Protocol                 |
| <b>CLIO/S</b> | Client Input/Output Sockets                           | <b>DIMM</b>   | Dual In-Line Memory Module                          |
| <b>CLVM</b>   | Concurrent LVM                                        | <b>DIP</b>    | Direct Insertion Probe                              |
| <b>CMOS</b>   | Complimentary Metal-Oxide Semiconductor               | <b>DIT</b>    | Directory Information Tree                          |
| <b>CMP</b>    | Certificate Management Protocol                       | <b>DIVA</b>   | Digital Inquiry Voice Answer                        |
| <b>COFF</b>   | Common Object File Format                             | <b>DLT</b>    | Digital Linear Tape                                 |
| <b>COLD</b>   | Computer Output to Laser Disk                         | <b>DMA</b>    | Direct Memory Access                                |
| <b>CPU</b>    | Central Processing Unit                               | <b>DMT</b>    | Directory Management Tool                           |
| <b>CRC</b>    | Cyclic Redundancy Check                               | <b>DN</b>     | Distinguished Name                                  |
| <b>CSID</b>   | Character Set ID                                      | <b>DNS</b>    | Domain Naming System                                |
| <b>CSR</b>    | Customer Service Representative                       | <b>DOE</b>    | Department of Energy                                |
| <b>CSS</b>    | Communication Subsystems Support                      | <b>DOI</b>    | Domain of Interpretation                            |
| <b>CSU</b>    | Customer Set-Up                                       | <b>DOS</b>    | Disk Operating System                               |
| <b>CWS</b>    | Control Workstation                                   | <b>DPCL</b>   | Dynamic Probe Class Library                         |
| <b>DAD</b>    | Duplicate Address Detection                           | <b>DRAM</b>   | Dynamic Random Access Memory                        |
| <b>DAS</b>    | Dual Attach Station                                   | <b>DS</b>     | Differentiated Service                              |
| <b>DASD</b>   | Direct Access Storage Device                          | <b>DSA</b>    | Dynamic Segment Allocation                          |
| <b>DAT</b>    | Digital Audio Tape                                    | <b>DSE</b>    | Diagnostic System Exerciser                         |
| <b>DBCS</b>   | Double Byte Character Set                             | <b>DSMIT</b>  | Distributed SMIT                                    |
| <b>DBE</b>    | Double Buffer Extension                               | <b>DSU</b>    | Data Service Unit                                   |
| <b>DC</b>     | Direct Current                                        | <b>DTE</b>    | Data Terminating Equipment                          |
| <b>DCE</b>    | Distributed Computing Environment                     | <b>DW</b>     | Data Warehouse                                      |
| <b>DDC</b>    | Display Data Channel                                  | <b>EA</b>     | Effective Address                                   |
| <b>DDS</b>    | Digital Data Storage                                  | <b>EC</b>     | Engineering Change                                  |
| <b>DE</b>     | Dual-Ended                                            | <b>ECC</b>    | Error Checking and Correcting                       |
| <b>DES</b>    | Data Encryption Standard                              | <b>EEPROM</b> | Electrically Erasable Programmable Read Only Memory |
| <b>DFL</b>    | Divide Float                                          | <b>EFI</b>    | Extensible Firmware Interface                       |
| <b>DFP</b>    | Dynamic Feedback Protocol                             | <b>EHD</b>    | Extended Hardware Drivers                           |
| <b>DFS</b>    | Distributed File System                               | <b>EIA</b>    | Electronic Industries Association                   |
|               |                                                       | <b>EISA</b>   | Extended Industry Standard Architecture             |

|                    |                                                |                  |                                                               |
|--------------------|------------------------------------------------|------------------|---------------------------------------------------------------|
| <b>ELA</b>         | Error Log Analysis                             | <b>FRU</b>       | Field Replaceable Unit                                        |
| <b>ELF</b>         | Executable and Linking Format                  | <b>FSRM</b>      | File System Resource Manager                                  |
| <b>EMU</b>         | European Monetary Union                        | <b>FTP</b>       | File Transfer Protocol                                        |
| <b>EOF</b>         | End of File                                    | <b>FTP</b>       | File Transfer Protocol                                        |
| <b>EPOW</b>        | Environmental and Power Warning                | <b>GAI</b>       | Graphic Adapter Interface                                     |
| <b>ERRM</b>        | Event Response resource manager                | <b>GAMESS</b>    | General Atomic and Molecular Electronic Structure System      |
| <b>ESID</b>        | Effective Segment ID                           | <b>GPFS</b>      | General Parallel File System                                  |
| <b>ESP</b>         | Encapsulating Security Payload                 | <b>GPR</b>       | General-Purpose Register                                      |
| <b>ESSL</b>        | Engineering and Scientific Subroutine Library  | <b>GUI</b>       | Graphical User Interface                                      |
| <b>ETML</b>        | Extract, Transformation, Movement, and Loading | <b>GUID</b>      | Globally Unique Identifier                                    |
| <b>F/C</b>         | Feature Code                                   | <b>HACMP</b>     | High Availability Cluster Multi Processing                    |
| <b>F/W</b>         | Fast and Wide                                  | <b>HACWS</b>     | High Availability Control Workstation                         |
| <b>FC</b>          | Fibre Channel                                  | <b>HCON</b>      | IBM AIX Host Connection Program/6000                          |
| <b>FCAL</b>        | Fibre Channel Arbitrated Loop                  | <b>HDX</b>       | Half Duplex                                                   |
| <b>FCC</b>         | Federal Communication Commission               | <b>HFT</b>       | High Function Terminal                                        |
| <b>FCP</b>         | Fibre Channel Protocol                         | <b>HIPPI</b>     | High Performance Parallel Interface                           |
| <b>FDDI</b>        | Fiber Distributed Data Interface               | <b>HiPS</b>      | High Performance Switch                                       |
| <b>FDPR</b>        | Feedback Directed Program Restructuring        | <b>HiPS LC-8</b> | Low-Cost Eight-Port High Performance Switch                   |
| <b>FDX</b>         | Full Duplex                                    | <b>HMC</b>       | Hardware Management Console                                   |
| <b>FIFO</b>        | First In/First Out                             | <b>HostRM</b>    | Host Resource Manager                                         |
| <b>FLASH EPROM</b> | Flash Erasable Programmable Read-Only Memory   | <b>HP</b>        | Hewlett-Packard                                               |
| <b>FLIH</b>        | First Level Interrupt Handler                  | <b>HPF</b>       | High Performance FORTRAN                                      |
| <b>FMA</b>         | Floating point Multiply Add operation          | <b>HPSSDL</b>    | High Performance Supercomputer Systems Development Laboratory |
| <b>FPR</b>         | Floating Point Register                        | <b>HP-UX</b>     | Hewlett-Packard UNIX                                          |
| <b>FPU</b>         | Floating Point Unit                            | <b>HTML</b>      | Hyper-text Markup Language                                    |
| <b>FRCA</b>        | Fast Response Cache Architecture               | <b>HTTP</b>      | Hypertext Transfer Protocol                                   |
|                    |                                                | <b>Hz</b>        | Hertz                                                         |

|                       |                                                                                                                |               |                                                              |
|-----------------------|----------------------------------------------------------------------------------------------------------------|---------------|--------------------------------------------------------------|
| <b>I/O</b>            | Input/Output                                                                                                   | <b>IS</b>     | Integrated Service                                           |
| <b>I<sup>2</sup>C</b> | Inter Integrated-Circuit Communications                                                                        | <b>ISA</b>    | Industry Standard Architecture, Instruction Set Architecture |
| <b>IAR</b>            | Instruction Address Register                                                                                   | <b>ISAKMP</b> | Internet Security Association Management Protocol            |
| <b>IBM</b>            | International Business Machines                                                                                | <b>ISB</b>    | Intermediate Switch Board                                    |
| <b>ICCCM</b>          | Inter-Client Communications Conventions Manual                                                                 | <b>ISDN</b>   | Integrated-Services Digital Network                          |
| <b>ICE</b>            | Inter-Client Exchange                                                                                          | <b>ISMP</b>   | InstallShield Multi-Platform                                 |
| <b>ICElib</b>         | Inter-Client Exchange library                                                                                  | <b>ISNO</b>   | Interface Specific Network Options                           |
| <b>ICMP</b>           | Internet Control Message Protocol                                                                              | <b>ISO</b>    | International Organization for Standardization               |
| <b>ID</b>             | Identification                                                                                                 | <b>ISV</b>    | Independent Software Vendor                                  |
| <b>IDE</b>            | Integrated Device Electronics                                                                                  | <b>ITSO</b>   | International Technical Support Organization                 |
| <b>IDS</b>            | Intelligent Decision Server                                                                                    | <b>JBOD</b>   | Just a Bunch of Disks                                        |
| <b>IEEE</b>           | Institute of Electrical and Electronics Engineers                                                              | <b>JDBC</b>   | Java Database Connectivity                                   |
| <b>IETF</b>           | Internet Engineering Task Force                                                                                | <b>JFC</b>    | Java Foundation Classes                                      |
| <b>IHV</b>            | Independent Hardware Vendor                                                                                    | <b>JFS</b>    | Journaled File System                                        |
| <b>IIOB</b>           | Internet Inter-ORB Protocol                                                                                    | <b>JTAG</b>   | Joint Test Action Group                                      |
| <b>IJG</b>            | Independent JPEG Group                                                                                         | <b>KDC</b>    | Key Distribution Center                                      |
| <b>IKE</b>            | Internet Key Exchange                                                                                          | <b>L1</b>     | Level 1                                                      |
| <b>ILS</b>            | International Language Support                                                                                 | <b>L2</b>     | Level 2                                                      |
| <b>IM</b>             | Input Method                                                                                                   | <b>L2</b>     | Level 2                                                      |
| <b>INRIA</b>          | Institut National de Recherche en Informatique et en Automatique                                               | <b>LAN</b>    | Local Area Network                                           |
| <b>IP</b>             | Internetwork Protocol (OSI)                                                                                    | <b>LANE</b>   | Local Area Network Emulation                                 |
| <b>IPL</b>            | Initial Program Load                                                                                           | <b>LAPI</b>   | Low-Level Application Programming Interface                  |
| <b>IPSec</b>          | IP Security                                                                                                    | <b>LDAP</b>   | Lightweight Directory Access Protocol                        |
| <b>IrDA</b>           | Infrared Data Association (which sets standards for infrared support including protocols for data interchange) | <b>LDIF</b>   | LDAP Directory Interchange Format                            |
| <b>IRQ</b>            | Interrupt Request                                                                                              | <b>LED</b>    | Light Emitting Diode                                         |
|                       |                                                                                                                | <b>LFD</b>    | Load Float Double                                            |
|                       |                                                                                                                | <b>LFT</b>    | Low Function Terminal                                        |

|                |                                                 |              |                                   |
|----------------|-------------------------------------------------|--------------|-----------------------------------|
| <b>LID</b>     | Load ID                                         | <b>MP</b>    | Multiprocessor                    |
| <b>LLNL</b>    | Lawrence Livermore National Laboratory          | <b>MPC-3</b> | Multimedia PC-3                   |
| <b>LP</b>      | Logical Partition                               | <b>MPI</b>   | Message Passing Interface         |
| <b>LP64</b>    | Long-Pointer 64                                 | <b>MPOA</b>  | Multiprotocol over ATM            |
| <b>LPI</b>     | Lines Per Inch                                  | <b>MPP</b>   | Massively Parallel Processing     |
| <b>LPP</b>     | Licensed Program Product                        | <b>MPS</b>   | Mathematical Programming System   |
| <b>LPR/LPD</b> | Line Printer/Line Printer Daemon                | <b>MST</b>   | Machine State                     |
| <b>LRU</b>     | Least Recently Used                             | <b>MTU</b>   | Maximum Transmission Unit         |
| <b>LTG</b>     | Logical Track Group                             | <b>MWCC</b>  | Mirror Write Consistency Check    |
| <b>LV</b>      | Logical Volume                                  | <b>MX</b>    | Mezzanine Bus                     |
| <b>LVCB</b>    | Logical Volume Control Block                    | <b>NBC</b>   | Network Buffer Cache              |
| <b>LVD</b>     | Low Voltage Differential                        | <b>NCP</b>   | Network Control Point             |
| <b>LVM</b>     | Logical Volume Manager                          | <b>ND</b>    | Neighbor Discovery                |
| <b>MAP</b>     | Maintenance Analysis Procedure                  | <b>NDP</b>   | Neighbor Discovery Protocol       |
| <b>MASS</b>    | Mathematical Acceleration Subsystem             | <b>NFB</b>   | No Frame Buffer                   |
| <b>MAU</b>     | Multiple Access Unit                            | <b>NFS</b>   | Network File System               |
| <b>MBCS</b>    | Multi-Byte Character Support                    | <b>NHRP</b>  | Next Hop Resolution Protocol      |
| <b>Mbps</b>    | Megabits Per Second                             | <b>NIM</b>   | Network Installation Management   |
| <b>MBps</b>    | Megabytes Per Second                            | <b>NIS</b>   | Network Information System        |
| <b>MCA</b>     | Micro Channel Architecture                      | <b>NL</b>    | National Language                 |
| <b>MCAD</b>    | Mechanical Computer-Aided Design                | <b>NLS</b>   | National Language Support         |
| <b>MDI</b>     | Media Dependent Interface                       | <b>NT-1</b>  | Network Terminator-1              |
| <b>MES</b>     | Miscellaneous Equipment Specification           | <b>NTF</b>   | No Trouble Found                  |
| <b>MFLOPS</b>  | Million of Floating point Operations Per Second | <b>NTP</b>   | Network Time Protocol             |
| <b>MII</b>     | Media Independent Interface                     | <b>NUMA</b>  | Non-Uniform Memory Access         |
| <b>MIP</b>     | Mixed-Integer Programming                       | <b>NUS</b>   | Numerical Aerodynamic Simulation  |
| <b>MLR1</b>    | Multi-Channel Linear Recording 1                | <b>NVRAM</b> | Non-Volatile Random Access Memory |
| <b>MMF</b>     | Multi-Mode Fibre                                | <b>NWP</b>   | Numerical Weather Prediction      |
| <b>MODS</b>    | Memory Overlay Detection Subsystem              | <b>OACK</b>  | Option Acknowledgment             |
|                |                                                 | <b>OCS</b>   | Online Customer Support           |
|                |                                                 | <b>ODBC</b>  | Open DataBase Connectivity        |
|                |                                                 | <b>ODM</b>   | Object Data Manager               |

|              |                                           |              |                                                            |
|--------------|-------------------------------------------|--------------|------------------------------------------------------------|
| <b>OEM</b>   | Original Equipment Manufacturer           | <b>POE</b>   | Parallel Operating Environment                             |
| <b>OLAP</b>  | Online Analytical Processing              | <b>POP</b>   | Power-On Password                                          |
| <b>OLTP</b>  | Online Transaction Processing             | <b>POSIX</b> | Portable Operating Interface for Computing Environments    |
| <b>ONC+</b>  | Open Network Computing                    | <b>POST</b>  | Power-On Self-test                                         |
| <b>OOUI</b>  | Object-Oriented User Interface            | <b>POWER</b> | Performance Optimization with Enhanced Risc (Architecture) |
| <b>OSF</b>   | Open Software Foundation, Inc.            | <b>PPC</b>   | PowerPC                                                    |
| <b>OSL</b>   | Optimization Subroutine Library           | <b>PPM</b>   | Piecewise Parabolic Method                                 |
| <b>OSLp</b>  | Parallel Optimization Subroutine Library  | <b>PPP</b>   | Point-to-Point Protocol                                    |
| <b>P2SC</b>  | POWER2 Single/Super Chip                  | <b>PREP</b>  | PowerPC Reference Platform                                 |
| <b>PAM</b>   | Pluggable Authentication Mechanism        | <b>PSE</b>   | Portable Streams Environment                               |
| <b>PAP</b>   | Privileged Access Password                | <b>PSSP</b>  | Parallel System Support Program                            |
| <b>PBLAS</b> | Parallel Basic Linear Algebra Subprograms | <b>PTF</b>   | Program Temporary Fix                                      |
| <b>PCI</b>   | Peripheral Component Interconnect         | <b>PTPE</b>  | Performance Toolbox Parallel Extensions                    |
| <b>PDT</b>   | Paging Device Table                       | <b>PTX</b>   | Performance Toolbox                                        |
| <b>PDU</b>   | Power Distribution Unit                   | <b>PV</b>    | Physical Volume                                            |
| <b>PE</b>    | Parallel Environment                      | <b>PVC</b>   | Permanent Virtual Circuit                                  |
| <b>PEDB</b>  | Parallel Environment Debugging            | <b>PVID</b>  | Physical Volume Identifier                                 |
| <b>PEX</b>   | PHIGS Extension to X                      | <b>QMF</b>   | Query Management Facility                                  |
| <b>PFS</b>   | Perfect Forward Security                  | <b>QoS</b>   | Quality of Service                                         |
| <b>PGID</b>  | Process Group ID                          | <b>QP</b>    | Quadratic Programming                                      |
| <b>PHB</b>   | Processor Host Bridges                    | <b>RAID</b>  | Redundant Array of Independent Disks                       |
| <b>PHY</b>   | Physical Layer                            | <b>RAM</b>   | Random Access Memory                                       |
| <b>PID</b>   | Process ID                                | <b>RAN</b>   | Remote Asynchronous Node                                   |
| <b>PID</b>   | Process ID                                | <b>RAS</b>   | Reliability, Availability, and Serviceability              |
| <b>PIOFS</b> | Parallel Input Output File System         | <b>RDB</b>   | Relational DataBase                                        |
| <b>PKR</b>   | Protection Key Registers                  | <b>RDBMS</b> | Relational Database Management System                      |
| <b>PMTU</b>  | Path MTU                                  | <b>RDISC</b> | ICMP Router Discovery                                      |
|              |                                           | <b>RDN</b>   | Relative Distinguished Name                                |

|                  |                                        |              |                                             |
|------------------|----------------------------------------|--------------|---------------------------------------------|
| <b>RDP</b>       | Router Discovery Protocol              | <b>SDLC</b>  | Synchronous Data Link Control               |
| <b>RFC</b>       | Request for Comments                   | <b>SDR</b>   | System Data Repository                      |
| <b>RIO</b>       | Remote I/O                             | <b>SDRAM</b> | Synchronous Dynamic Random Access Memory    |
| <b>RIP</b>       | Routing Information Protocol           | <b>SE</b>    | Single Ended                                |
| <b>RIPL</b>      | Remote Initial Program Load            | <b>SEPBU</b> | Scalable Electrical Power Base Unit         |
| <b>RISC</b>      | Reduced Instruction-Set Computer       | <b>SGI</b>   | Silicon Graphics Incorporated               |
| <b>RMC</b>       | Resource Monitoring and Control        | <b>SGID</b>  | Set Group ID                                |
| <b>ROLTP</b>     | Relative Online Transaction Processing | <b>SHLAP</b> | Shared Library Assistant Process            |
| <b>RPA</b>       | RS/6000 Platform Architecture          | <b>SID</b>   | Segment ID                                  |
| <b>RPC</b>       | Remote Procedure Call                  | <b>SIT</b>   | Simple Internet Transition                  |
| <b>RPL</b>       | Remote Program Loader                  | <b>SKIP</b>  | Simple Key Management for IP                |
| <b>RPM</b>       | Redhat Package Manager                 | <b>SLB</b>   | Segment Lookaside Buffer                    |
| <b>RSC</b>       | RISC Single Chip                       | <b>SLIH</b>  | Second Level Interrupt Handler              |
| <b>RSCT</b>      | Reliable Scalable Cluster Technology   | <b>SLIP</b>  | Serial Line Internet Protocol               |
| <b>RSE</b>       | Register Stack Engine                  | <b>SLR1</b>  | Single-Channel Linear Recording 1           |
| <b>RSVP</b>      | Resource Reservation Protocol          | <b>SM</b>    | Session Management                          |
| <b>RTC</b>       | Real-Time Clock                        | <b>SMB</b>   | Server Message Block                        |
| <b>RVSD</b>      | Recoverable Virtual Shared Disk        | <b>SMIT</b>  | System Management Interface Tool            |
| <b>SA</b>        | Secure Association                     | <b>SMP</b>   | Symmetric Multiprocessor                    |
| <b>SACK</b>      | Selective Acknowledgments              | <b>SMS</b>   | System Management Services                  |
| <b>SAN</b>       | Storage Area Network                   | <b>SNG</b>   | Secured Network Gateway                     |
| <b>SAR</b>       | Solutions Assurance Review             | <b>SOI</b>   | Silicon-on-Insulator                        |
| <b>SAS</b>       | Single Attach Station                  | <b>SP</b>    | IBM RS/6000 Scalable POWER parallel Systems |
| <b>SBCS</b>      | Single-Byte Character Support          | <b>SP</b>    | Service Processor                           |
| <b>ScaLAPACK</b> | Scalable Linear Algebra Package        | <b>SPCN</b>  | System Power Control Network                |
| <b>SCB</b>       | Segment Control Block                  | <b>SPEC</b>  | System Performance Evaluation Cooperative   |
| <b>SCSI</b>      | Small Computer System Interface        | <b>SPI</b>   | Security Parameter Index                    |
| <b>SCSI-SE</b>   | SCSI-Single Ended                      |              |                                             |

|                |                                                    |              |                                         |
|----------------|----------------------------------------------------|--------------|-----------------------------------------|
| <b>SPM</b>     | System Performance Measurement                     | <b>UDI</b>   | Uniform Device Interface                |
| <b>SPOT</b>    | Shared Product Object Tree                         | <b>UIL</b>   | User Interface Language                 |
| <b>SPS</b>     | SP Switch                                          | <b>ULS</b>   | Universal Language Support              |
| <b>SPS-8</b>   | Eight-Port SP Switch                               | <b>UP</b>    | Uniprocessor                            |
| <b>SRC</b>     | System Resource Controller                         | <b>USB</b>   | Universal Serial Bus                    |
| <b>SRN</b>     | Service Request Number                             | <b>USLA</b>  | User-Space Loader Assistant             |
| <b>SSA</b>     | Serial Storage Architecture                        | <b>UTF</b>   | UCS Transformation Format               |
| <b>SSC</b>     | System Support Controller                          | <b>UTM</b>   | Uniform Transfer Model                  |
| <b>SSL</b>     | Secure Socket Layer                                | <b>UTP</b>   | Unshielded Twisted Pair                 |
| <b>STFDU</b>   | Store Float Double with Update                     | <b>UUCP</b>  | UNIX-to-UNIX Communication Protocol     |
| <b>STP</b>     | Shielded Twisted Pair                              | <b>VESA</b>  | Video Electronics Standards Association |
| <b>SUID</b>    | Set User ID                                        | <b>VFB</b>   | Virtual Frame Buffer                    |
| <b>SUP</b>     | Software Update Protocol                           | <b>VG</b>    | Volume Group                            |
| <b>SVC</b>     | Switch Virtual Circuit                             | <b>VGDA</b>  | Volume Group Descriptor Area            |
| <b>SVC</b>     | Supervisor or System Call                          | <b>VGSA</b>  | Volume Group Status Area                |
| <b>SWVPD</b>   | Software Vital Product Data                        | <b>VHDCI</b> | Very High Density Cable Interconnect    |
| <b>SYNC</b>    | Synchronization                                    | <b>VLAN</b>  | Virtual Local Area Network              |
| <b>TCE</b>     | Translate Control Entry                            | <b>VMM</b>   | Virtual Memory Manager                  |
| <b>Tcl</b>     | Tool Command Language                              | <b>VP</b>    | Virtual Processor                       |
| <b>TCP/IP</b>  | Transmission Control Protocol/Internet Protocol    | <b>VPD</b>   | Vital Product Data                      |
| <b>TCQ</b>     | Tagged Command Queuing                             | <b>VPN</b>   | Virtual Private Network                 |
| <b>TGT</b>     | Ticket Granting Ticket                             | <b>VSD</b>   | Virtual Shared Disk                     |
| <b>TLB</b>     | Translation Lookaside Buffer                       | <b>VSM</b>   | Visual System Manager                   |
| <b>TOS</b>     | Type Of Service                                    | <b>VSS</b>   | Versatile Storage Server                |
| <b>TPC</b>     | Transaction Processing Council                     | <b>VT</b>    | Visualization Tool                      |
| <b>TPP</b>     | Toward Peak Performance                            | <b>WAN</b>   | Wide Area Network                       |
| <b>TSE</b>     | Text Search Engine                                 | <b>WLM</b>   | Workload Manager                        |
| <b>TSE</b>     | Text Search Engine                                 | <b>WTE</b>   | Web Traffic Express                     |
| <b>TTL</b>     | Time To Live                                       | <b>XCOFF</b> | Extended Common Object File Format      |
| <b>UCS</b>     | Universal Coded Character Set                      | <b>XIE</b>   | X Image Extension                       |
| <b>UDB EEE</b> | Universal Database and Enterprise Extended Edition | <b>XIM</b>   | X Input Method                          |
|                |                                                    | <b>XKB</b>   | X Keyboard Extension                    |

|             |                             |
|-------------|-----------------------------|
| <b>XLF</b>  | XL Fortran                  |
| <b>XOM</b>  | X Output Method             |
| <b>XPM</b>  | X Pixmap                    |
| <b>XSSO</b> | Open Single Sign-on Service |
| <b>XTF</b>  | Extended Distance Feature   |
| <b>XVFB</b> | X Virtual Frame Buffer      |



# Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

## IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 343.

- ▶ *AIX 5L Differences Guide Version 5.1 Edition*, SG24-5765
- ▶ *AIX 5L Performance Tools Handbook*, SG24-6039
- ▶ *AIX Logical Volume Manager From A to Z: Introduction and Concepts*, SG24-5432
- ▶ *IBM @server Certification Study Guide - AIX 5L Communications*, SG24-6186
- ▶ *IBM @server Certification Study Guide - AIX 5L Installation and System Recovery*, SG24-6183
- ▶ *IBM @server Certification Study Guide - AIX 5L Performance and System Tuning*, SG24-6184
- ▶ *IBM @server Certification Study Guide - pSeries AIX System Administration*, SG24-6191
- ▶ *IBM @server Certification Study Guide - pSeries AIX System Support*, SG24-6199
- ▶ *IBM @server Certification Study Guide - pSeries HACMP for AIX*, SG24-6187
- ▶ *IBM @server Certification Study Guide - RS/6000 SP*, SG24-5348
- ▶ *Managing AIX Server Farms*, SG24-6606
- ▶ *Printing for Fun and Profit under AIX 5L*, SG24-6018
- ▶ *Problem Solving and Troubleshooting in AIX 5L*, SG24-5496
- ▶ *TCP/IP Tutorial and Technical Overview*, GG24-3376

## Other resources

These publications are also relevant as further information sources:

- ▶ *AIX Version 4.3 System Management Concepts: Operating System and Devices*, SC23-4311
- ▶ *AIX Version 5.0 Installation Guide*, SC23-4112
- ▶ *RS/6000 and pSeries PCI Adapter Placement Reference*, SA38-0538
- ▶ *SSA Adapters: User's Guide and Maintenance Information*, SA33-3272
- ▶ The following types of documentation are located through the Internet at the following URL:

<http://www-1.ibm.com/servers/eserver/pseries/library>

- User guides
- System management guides
- Application programmer guides
- All commands reference volumes
- Files reference
- Technical reference volumes used by application programmers

## Referenced Web sites

These Web sites are also relevant as further information sources:

- ▶ AIX software updates  
<http://techsupport.services.ibm.com/support/rs6000.support/databases>
- ▶ IBM Certification tests by number  
<http://www.ibm.com/certify/tests/info.shtml>
- ▶ IBM eServer pSeries support  
<http://techsupport.services.ibm.com/server/support?view=pSeries>
- ▶ IBM hardware documentation  
[http://www-1.ibm.com/servers/eserver/pseries/library/hardware\\_docs/index.html](http://www-1.ibm.com/servers/eserver/pseries/library/hardware_docs/index.html)
- ▶ IBM Professional Certification Program Web site  
<http://www.ibm.com/certify>
- ▶ IBM Redbooks Web site  
<http://www.redbooks.ibm.com>

- ▶ IBM TotalStorage  
<http://www.storage.ibm.com>
- ▶ Open Group Technical Standard Protocols for Interworking: XNFS, Version 3W  
<http://www.opengroup.org/onlinepubs/9629799/toc.htm>
- ▶ SSA adapters - comprehensive and up-to-date information  
<http://www.storage.ibm.com/hardsoft/products/ssa>
- ▶ UNIX servers (pSeries) information  
[http://www-132.ibm.com/content/home/store\\_IBMPublicUSA/en\\_US/eServer/pSeries/pSeries.html](http://www-132.ibm.com/content/home/store_IBMPublicUSA/en_US/eServer/pSeries/pSeries.html)

## How to get IBM Redbooks

You can order hardcopy Redbooks, as well as view, download, or search for Redbooks at the following Web site:

[ibm.com/redbooks](http://ibm.com/redbooks)

You can also download additional materials (code samples or diskette/CD-ROM images) from that site.

## IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.



# Index

## Symbols

/dev/mem 83  
/etc/exports 189  
/etc/filesystems 128–129, 152, 189  
/etc/filesystems file 165  
/etc/gated.conf 184  
/etc/gateways 184  
/etc/hosts 185  
/etc/inittab 189  
/etc/mail/sendmail.cf 217  
/etc/netsvc.conf 185  
/etc/qconfig 128  
/etc/rc.nfs 189  
/etc/rc.tcpip 189  
/etc/resolv.conf 186, 205  
/etc/security/limits 206, 210  
/etc/services 205  
/etc/syslog.pid 130  
/etc/utmp 210  
/proc  
    see also proc pseudo file system 162  
/unix 79  
/usr/include/sys/syslog.h 130  
/usr/include/sys/trchkid.h 212  
/var/adm/ras/trcfile 211  
/var/spool/cron/crontabs/root 128

## Numerics

32bit, WLM process type 287  
64bit  
    WLM process type 287  
64-bit kernel, JFS2 160  
7020-40P 47  
7248-43P 47

## A

abend code 80  
Accounting 128  
acctdisk 129  
adding a new disk 147  
addressing exception 95  
administration

    workload manager 277  
aixterm command 214  
alog 25, 28  
APAR 312–313  
application path names (WLM) 287  
application tags (WLM) 287  
as pseudo file 163  
assign disk to volume group 148  
assignment rules (WLM) 284  
ATMLE 34  
attributes  
    classes 281  
attributes, localshm 283  
automatic assignment (WLM) 283  
automatic error log analysis - diagela 144

## B

backup data 140  
bigfile file system 151  
bindprocessor 238  
biod 189, 191  
BIST 14–15  
    LED 200 19  
    LED 299 19  
BLV 14  
    how to recreate BLV 19  
boot 25  
    /etc/inittab figure 31  
    /mnt 24  
    alog 28  
    BIST 14–15  
    BLV 14  
    bootlist 16  
    Config\_Rules 23  
    error log 35  
    general boot order figure 14  
    general overview 14  
    generic device names 16  
    how to recreate BLV 19  
    magic number 26  
    maintenance 16  
    normal boot 21  
    phase1 15, 22, 36

- phase1 figure 23
- phase2 15, 24, 36
- phase2 figure1 24
- phase2 figure2 25
- phase3 15, 29, 37
- POST 14–15
- runlevel 32
- service 16
- service boot 21
- service mode 50
- SMS main menu figure 21
- superblock 27
- boot logical volume 14
- bootinfo 23, 44
- bootinfo command 161
- bootlist 16, 22, 26
  - generic device names 16
- bosboot 20, 26, 314
- built in self test 15
- bundle
  - maintenance level bundle 311
  - update bundle 311

## C

- cfgmgr 23, 30
- cfgmgr command 147
- changes to the system 9
- changing the bootlist on PCI systems 20
- chdev 179
- chfs command 152, 165, 210
- chps command 167
- CHRP 44
- chrp 44
- chvg command 143, 170
- ckpacct 128
- class
  - inheritance 282
- class assignment rules 286
- class name (WLM) 286
- classes 276
- classification process 283
- commands
  - aixterm 214
  - alog 25, 28
  - bindprocessor 238
  - bootinfo 23, 44, 161
  - bootlist 16, 22, 26
  - bosboot 20, 26, 314

- cfgmgr 23, 30, 147
- chdev 179
- chfs 152, 165, 210
- chps 167
- chvg 143, 170
- crash 79, 83
- date 119
- dd 28, 138, 153
- df 156, 210
- diag 47
- du 156
- edquota 164–165
- entstat 180
- errclear 108, 127
- errdemon 108–109
- errpt 35, 38, 81, 117
- exportfs 190
- extendvg 148
- filemon 264
- fileplace 264
- find 82
- fsck 26, 152
- genkex 155
- getconf 161
- host 185, 190
- ifconfig 178
- importvg 138, 141, 171
- installp 309–310, 313
- instfix 313, 315
- iostat 260
- ipl\_varyon 24
- ipreport 33
- iptrace 33, 186
- logform 26
- lppchk 309, 314
- lpstat 220
- lquerypv 83
- lsattr 46, 179, 271
- lscfg 44
- lsdev 44, 178
- lsjfs 151
- lslicense 202, 225
- lslpp 310, 315
- lsps 76
- lspv 142
- lsrset 290
- lssrc 183, 204, 225
- lsvg 145, 169
- mergedev 25

- migratepv 53
- mkclass 283
- mkivcopy 148
- mount 191
- netpmon 274
- netstat 179, 181, 266, 301
- nfsstat 273, 302
- no 185, 268
  - ipforwarding 185
- nslookup 185, 205
- ping 181–182
- ps 241, 251, 300
- qchk 220
- quotacheck 166
- quotaon 166
- redefinevg 141
- reducevg 172
- restbase 23
- rmdev 172
- rmlvcopy 146, 172
- rmps 168–169
- route 185
- sar 235, 300
- savebase 27, 31
- shconf 74
- snap 77, 95
- startsrc 183, 204, 226
- stopsrc 190
- strings 83, 97
- su 208
- svmon 248, 253
- swapoff 169
- synclvodm 141, 150
- syncvg 30, 141, 148, 173
- sysdumpdev 66–68, 70–71, 76–77, 98, 168
- sysdumpstart 68, 100
- tar 78
- tcpdump 186
- tee 94
- topas 214
- tprof 243
- trace 211–212, 226
- traceroute 183
- trcoff 264
- trcon 264
- trcrpt 213, 227
- trcstop 213, 264
- uptime 39
- varyonvg 138–139
- vmstat 238, 247
- vmtune 248
- w 39
- wlmcntrl 293
- compatibility
  - workload manager 289
- computational memory 246
- Config\_Rules 23
- configuration, ODM 144
- connection oriented 273
- connectionless 273
- core dump
  - checking error report 81
  - determine program responsible 82
  - locating core file 82
- core dumps 81
- CPU bound system 235
- crash command 79, 83
  - uses 83
- crash subcommands 85
  - errpt 93
  - le 90
  - od 92
  - proc 90
    - output fields 91
  - set 94
  - stat 86
  - symptom 93
  - thread 91
    - output fields 92
  - trace 87
    - exception structure 89
- creating JFS 150
- cred pseudo file 163
- crfs 157
- ctl pseudo file 163
- customer relations 7

## D

- daemon
  - biod 189
  - gated 184
  - mountd 190
  - nfsd 189
  - portmap 189
  - routed 184
  - rpc.lockd 189
  - rpc.mountd 189

- rpc.statd 189
- daemons
  - shdaemon 73
  - telnetd 204
  - ypbind 33
  - ypserv 33
- data
  - logical volume manager 138
  - logical volumes 139
  - relocation 140
  - volume group 138
- Data fragmentation 160
- data storage interrupt 95
- database, ODM 144
- date command 119
- dbx 294
- dd 28
- dd command 138, 153
- default route 33
- default system error log 108
- define the problem 8
- deleting error log entries 125
- determination process of a problem 8
- device
  - state
    - available 45
    - define 45
- df command 156, 210
- diag 47
  - advanced diagnostics 48
  - alter bootlist menu 18
  - diagnostic routines 48
  - function selection menu 17
  - task selection 48
    - disk maintenance 53
    - SSA 53
  - task selection menu 18
- diagela - automatic error log analysis 144
- diagnostic
  - concurrent mode 47, 49
  - CPU 50
  - memory 50
  - stand-alone
    - from CD 51
    - from disk 50
    - MCA machines 50–51
    - PCI machines 50–51
- disk
  - adding a new disk 147
  - recovering an incorrectly removed disk 148
  - remove the disk with rmdev 146
  - removing bad disk 146
  - replacement 144
- disk bound 258
- disk problems 137
- diskusg 129
- DNS 33–34
  - server 185
- dodisk 128–129
- du 156
- du command 156
- dumpfile namelist 80

**E**

- E1F1 LEDs 52
- edquota command 164
- edquota commands 165
- entstat 180
- environment variable
  - NSORDER 205
- errclear command 108, 127
  - flags 125
  - syntax 125
- errdemon command 108–109
  - flags 108
  - syntax 108
- error daemon 108
- error log configuration database 110
- error notification database 108
- error report 107, 110
- errpt 35, 38
  - flag table 38
- errpt command 81, 117
  - commands
    - errpt 145
  - error log report 111
  - error record template report 112
  - flags 112
  - syntax 111–112
- exec() system call 284, 288
- exportfs 190
- extended\_netstats 268
- extending number of max PPs 143
- extendvg command 148
- Extents 159

## F

### figures

- boot phase1 23
- boot phase2 figure1 24
- boot phase2 figure2 25
- changing AIX operating system parameters 208
- CPU penalty 257
- diag alter bootlist menu 18
- diag function selection menu 17
- diag task selection menu 18
- disk problem report form diagela 145
- general boot order 14
- licensed users 203
- SMS main menu 21
- topas display for trace example 214

### file memory 246

### file system

- fixing bad superblock 153
- full file system 210
- problems 137
- removing file systems 155
- unmount problems 154
- verification and recovery 152

### file table 84

### filemon 264

### fileplace 264

### files

- /dev/error 108
- /etc/exports 189
- /etc/filesystems 33, 152, 165, 189
- /etc/gated.conf 184
- /etc/gateways 184
- /etc/hosts 185
- /etc/inittab 15, 30–31, 189
- /etc/netsvc.conf 34, 185
- /etc/objrepos/errnotify 108
- /etc/rc.boot 15, 23–24
- /etc/rc.net 32–33, 191, 268
- /etc/rc.nfs 189, 191
- /etc/rc.tcpip 189
- /etc/resolv.conf 186, 205
- /etc/security/limits 206, 210
- /etc/services 205
- /etc/syslog.conf 130
- /etc/syslog.pid 130
- /etc/utmp 210
- /usr/include/sys/trchkid.h 212
- /usr/lib/sa/sa1 236

/usr/lib/sa/sa2 236

/var/adm/ras/conslog 31

/var/adm/ras/errlog 108, 110

/var/adm/ras/trcfile 211

\_\_prof.all 244

quota.group 164

quota.user 164

fileset update 308

find command 82

fixed 287

flag table 302

flood ping 182

fork 284

fork function failed

adjusting kernel parameters 208

Format of the 103 code message 60

fragmentation 258

FRU Number 46

fsck 26

fsck command 152

## G

gated 184

gateway 184

general boot overview 14

generic device names 16

genkex command 155

getconf command 161

getty 218

group (WLM) 286

## H

handling crash output 94

### hardware

diagnostic 46

inventory 43

platform

CHRP 44

chrp 44

PREP 44

RPA 44

rs6k 44

rs6ksmp 44

rspc 44

HAT 246

high-water mark 260

hook IDs for trace 212

host 185, 190

- host name resolution
  - telnet login problem 205
- how to recreate BLV 19
- hung process tracing 211

## I

- I/O pacing 260
- IAR 88
- ifconfig 178
- importvg command 138, 141, 171
- importvg problems
  - disk change 142
  - shared disk environment 143
- increasing the file system size 152
- information
  - collecting information from the user 8
  - collection information about the system 10
  - questions you should ask 9
- i-node 150
- inode table 84
- installing
  - online manuals 326
- installp 309–310, 313
- instfix 313, 315
- internal fragmentation 159
- invalid dump 80
- iostat 260
  - the %tm\_act column 262
  - the CPU columns 261
  - the disks column 262
  - the Drive reports 261
  - the Kb\_read column 263
  - the Kb\_wrtn column 263
  - the Kbps column 262
  - the tps column 262
  - the TTY columns 260
- ipforwarding 185
- ipl\_varyon 24
- ipreport command 33
- iptrace 186
- iptrace command 33

## J

- JFS
  - bigfile file system 151
  - creating file system 150
  - fixing bad superblock 153
  - fsck command 152

- increasing the file system size 152
- i-node 150
- removing file system 155
- unmount problems 154
- verification and recovery 152
- JFS file system 150

## K

- kdb 296
- kernel address 90
- kernel description 84
- kernel extension 90
- kernel panic 95
- kernel parameters
  - adjusting number of processes per user 207
- kernel stack traceback 84
- kernel trap 95

## L

- LED 58
- LED codes
  - 100 - 195 16, 58
  - 200 19, 57
  - 200 - 2E7 16
  - 201 19
  - 221 19
  - 221 - 229 19
  - 223 - 229 19
  - 225 - 229 19
  - 233 - 235 19
  - 299 19, 58
  - 511 23
  - 518 24, 28
  - 548 23
  - 551 26
  - 552 24, 26
  - 553 30, 32
  - 554 24, 26
  - 555 24, 26
  - 556 24, 26
  - 557 24, 26
  - 581 32
  - 721 19
  - 888 58
  - 102 58
  - 103 59–60
  - 105 59
  - c31 32

- common MCA LED code table 19
- common MCA LED codes 58
- MCA POST LED code table 37
- phase2 LED code table 37
- phase3 LED code table 37
- LEDs E1F1 52
- license problems 202
- locked volume group 139
- logform 26
- logical volume
  - mirrored copy with mklvcopy 148
  - stale LV 148
- logical volume manager 258
  - AIX version level 142
  - data 138
    - logical volumes 139
    - physical volume 138
    - volume group 138
  - importvg 141
  - problems 137
  - VGDA 138
  - VGID 139
  - VGSA 139
- logical volumes
  - data 139
  - LVCB 139
  - LVID 139
- login
  - problems with full file system 210
  - problems with telnet 203
  - problems with user license 202
- low-water mark 260
- LPP 308
- lppchk 309, 314
- lpstat command 220
- lquerypv command 83
- LR 88
- lsattr 46, 179, 271
  - tx\_que\_size 271
- lscfg 44
- lsdev 44, 178
- lsfs 157
- lsjfs command 151
- lslicense command 202, 225
- lslpp 310, 315
- lsps command 76
- lspv command 142
- lsrset command 290
- lssrc 183

- lssrc command 204, 225
- lsvg command 145, 169
- LVCB - logical volume control block 139
  - rebuild 141
- LVDD 146
- LVID - logical volume identifier 139
- LVM 258
  - fragmentation 258
- LVM problem determination 139

## M

- magic number 26
- mail 217
- mail.debug 217
- maintenance 16
- maintenance level bundle 311
- maintenance mode 47
- manual assignment (WLM) 283
- map pseudo file 163
- maxfree 250
- maxperm 250
- maxpgahead 259
- maxuproc 208
- mbufs 268
- MCA 15
  - BIST 14–15
  - bootlist 16
  - common MCA LED code table 19
  - LED codes 100 - 195 16
  - LED codes 200 - 2E7 16
  - POST 14–15
- memory bound 246
- mergedev 25
- microcode download 52
- migratepv command 53
- minfree 248
- minperm 250
- minpgahead 259
- mirrored environment 144
- mkclass command 283
- mkfs 157
- mklvcopy command 148
- mksysb tape 47
- monacct 128
- mount 191
- mountd 190
- MST 87
- multiboot 21

## N

- name resolution
  - diagnostic 185
  - problems 185
- namelist 80
- netpmo 274
- netstat 179, 181, 266, 301
  - broadcast packets 271
  - DMA overrun 271
  - flag table 301
  - late collision errors 272
  - max collision errors 271
  - max packets on S/W transmit queue 271
  - mbufs 268
  - multiple collision count 272
  - no mbuf errors 272
  - receive collision errors 272
  - receive errors 270
  - S/W transmit queue overflow 271
  - single collision count 272
  - the Coll column 267
  - the lerrs column 267
  - the lpkts column 267
  - the Mtu column 267
  - the Oerrs column 267
  - the Opkts column 267
  - timeout errors 272
  - transmit errors 270
  - tuning guidelines 267, 272
- network bound 265
  - thewall 270
  - tuning guidelines 267
- network interface
  - collision count 180
  - lo0 183
  - loopback 183
  - problems 178
  - setup 179
  - state
    - detach 178
    - down 178
    - up 178
- NFS
  - problems
    - mount 190
    - troubleshooting 189
  - nfs\_socketsize 191
  - nfsd 189, 191
  - nfsstat 273, 302

- badcalls 273
- badxid 273
- clgets 273
- retrans 274
- NIS 33
  - server 185
- no 185, 268
  - extended\_netstats 268
  - ipforwarding 185
  - thewall 270
- normal boot 21
- nslookup 185
- nslookup command 205
- NSORDER 34, 185
- NSORDER environment variable 205
- nulladm 128

## O

- ODM 33, 144
- ODM - Object Data Manager 138–139
  - corruption 141
  - re-synchronization 140

## P

- package 308
- page stealing 246
- page-replacement routine 249
- paging space 166
  - determining need for more paging space 167
  - recommendations 166
  - removing 168
- panic string 86
- PCI 20
  - changing bootlist 20
  - Normal boot 21
  - Service boot 21
  - SMS main menu figure 21
- phase1 15, 22, 36
  - /etc/rc.boot 23
  - Conifg\_Rules 23
  - LED code 511 23
  - LED code 548 23
  - phase1 figure 23
- phase2 15, 24, 36
  - /etc/rc.boot 24
  - /mnt 24
  - alog 28
  - LED code 518 24, 28

- LED code 551 26
- LED code 552 24, 26
- LED code 554 24, 26
- LED code 555 24, 26
- LED code 556 24, 26
- LED code 557 24, 26
- LED code table 37
- phase2 figure1 24
- phase2 figure2 25
- phase3 15, 29, 37
  - /etc/inittab 31
  - LED c31 32
  - LED code 553 30, 32
  - LED code 581 32
  - LED code table 37
  - rc.boot 3 in inittab 31
  - runlevel 32
- physical volume
  - adding a new disk 147
  - data 138
  - extending max PPs 143
  - PVID 138
  - recovering an incorrectly removed disk 148
- ping 181–182
- ping -f command 182
- pinned memory 90
- plock 287
- plock() system call 288
- port for telnet 205
- portmap 189
- POST 14–15
- power on self test 15
- PPs - extending the maximum number 143
- PREP 44
- Printing 220
- problem
  - definition 8
  - determination process 8
  - questions you should ask 9
- problem determination
  - LVM 139
- problems
  - disk 137
  - importvg 141
  - system access with telnet 203
  - user license 202
- proc pseudo file system
  - as pseudo file 163
  - cred pseudo file 163
  - ctl pseudo file 163
  - map pseudo file 163
  - psinfo pseudo file 163
  - sigact pseudo file 163
  - status pseudo file 164
  - sysent pseudo file 164
  - vfs entry 162
- process
  - adjusting number of processes per user 207
  - hung process tracing 211
- process penalty 241
- process priority 242
- process table 84, 90
- process type (WLM) 287
- ps 241, 251, 300
  - CPU related table 241
  - flag table 300
  - memory related output table 251
  - penalty 241
  - process priority 242
  - the %CPU column 243
  - the %MEM column 252
  - the C column 241
  - the fre column 248
  - the PGIN column 252
  - the RSS column 251
  - the SIZE column 251
  - the SZ column 251
  - the TIME column 242
  - the TRS column 253
  - the TSIZ column 253
- pseudo files 163–164
- psinfo pseudo file 163
- PTF 246, 308, 311
- PVID - physical volume identifier 138
- PVID invalid 142

**Q**

- qchk command 220
- questions you should ask 9
- quorum 138
- quota.group file 164
- quota.user file 164
- quotacheck command 166
- quotaon command 166

**R**

- receive errors 270

- recent CPU usage 241
- recovering an incorrectly removed disk 148
- Redbooks Web site 343
  - Contact us xx
- redefinevg command 141
- reducevg command 172
  - commands
  - reducevg 146
- relocation of data 140
- remote procedure call 273
- remote reboot 72
- remove the disk with rmdev 146
- removing bad disk 146
- removing file systems 155
- removing paging space 168
- replacement of disk 144
- replacing disk 144
- report
  - filtering trace reports 214
- reports
  - trace reports 213
- resource sets (WLM) 289
- restbase 23
- RIP 185
- rmdev command 172
- rmlvcopy command 146, 172
- rmpps command 168–169
- ROS Level 46
- route 185
- routed 184
- routing
  - problems 181
  - tables 181
    - default 182
- RPA 44
- RPC 273
- rpc.lockd 189
- rpc.mountd 189
- rpc.statd 189
- rs6k 44
- rs6ksmp 44
- rset 291
- rset registry 291
- rspc 44
- runacct 128
- runlevel 32

## S

- sar 235, 300
  - /usr/lib/sa/sa1 236
  - /usr/lib/sa/sa2 236
  - display previously captured data 236
  - flag table 300
  - real-time sampling and display 235
  - system activity accounting via cron daemon 236
- savebase 31
- savebase command 27
- scenario
  - replacing a disk 144
- schedtune 254
  - SCHED\_R and SCHED\_D guidelines 257
- scheduler
  - CPU penalty figure 257
  - SCHED\_R and SCHED\_D guidelines 257
- SCSI
  - adapter 49
  - bus analyzer 52
- sendmail 217
- sequential-access read ahead 259
- Serial Storage Architecture (SSA) 54
- service boot 16, 21
- Service Request Number (SRN) 49
- setgid() system call 288
- setpri() system calls 288
- setuid() system call 288
- shconf command 74
- shdaemon daemon 73
- sigact pseudo file 163
- single-user mode 47
- smdemon.cleanu 217
- smit dump 69
- SMIT fast path
  - smit chgsys 208
  - smit chlicense 202
  - smit jfs 151
  - smit resolv.conf 205
- smitty
  - install\_update 312
  - maintain\_software 310
  - ssaraid 55
- snap command 77, 95
  - flags 96
  - syntax 96
- software
  - updates 307

- software patch
  - installation 312
- software problems 9
- software states
  - applied 309
  - committed 309–310
  - reject 310
- sparse file allocation 154
- SRC 130
- SRN 49–50, 59
- SSA
  - adapter
    - information 55
    - loop 54
    - speed 54
  - devices 55
  - divide
    - disk 55
    - SSA RAID 55
  - setup 54
- SSA disk 146
- stale logical volume 148
- startsrc 183, 190
- startsrc command 204, 226
- startup 128
- statts 157
- static routes 184
- status pseudo file 164
- stopsrc 190
- strings command 83, 97
  - flags 97
  - syntax 97
- su command 208
- subclass 277
- subserver
  - telnetd 204
- superblock 27
  - fixing bad superblock 153
- superclass 277
- svmon 248, 253
  - command report 253
  - detailed segment report 253
  - frame report 254
  - global report 253
  - process report 253
  - segment report 253
  - tier report 254
  - user report 253
  - workload management class report 254
- swapoff command 169
- synclvodm command 141, 150
- syncvg 30
- syncvg command 141, 148, 173
- sysdumpdev 76
- sysdumpdev command 66–68, 70–71, 76–77, 98, 168
  - flags 99
  - syntax 98
- sysdumpstart command 68, 100
  - flags 101
  - syntax 100
- sysent pseudo file 164
- syslogd 130, 217
- system access
  - adjusting number of processes per user 207
  - full file system 210
  - telnet problems 203
- system calls
  - exec() 284, 288
  - plock() 288
  - setgid() 288
  - setpri() 288
  - setuid() 288
- system dump 65
  - copy 77
  - dump device configuration 66
    - increase size 76
    - pre-requisites 66
  - examine with crash 84
  - panic string 86
  - routines 79
  - starting dump 68
    - command line 68
    - key sequences 71
    - reset button 70
    - smit interface 69
  - status check 74
  - status codes 74, 100
- system dumps
  - reading 79
- system events
  - tracing 211
- system hang 95
- system resource controller - SRC
  - lssrc 225
  - lssrc command 204
  - startsrc 226
  - startsrc command 204

System Resource Controller (SRC) 130

## T

tables

- CPU related ps output 241
- errpt flags 38
- MCA LED codes 37
- memory related ps output table 251
- netstat flags 301
- nfsstat flags 302
- phase2 LED codes 37
- phase3 LED codes 37
- ps flags 300
- sar flags 300
- w flags 39

tar command 78

tcpdump 186

tee command 94

telnet

- name server resolution 205
- network problem 203
- port 23 205
- session 186
- slow telnet login 205
- start the telnet subserver 204
- system access problems 203
- telnet subserver 204

thewall 270

thread table 84

three-digit display 57

tick 243

TLB 246

topas command 214

tprof 243

\_\_prof.all 244

summary report 244

the freq column 245

the total column 244

the user column 245

trace

buffers

alternate mode 213

circular mode 213

single mode 213

example 214

filtering trace reports 214

hook IDs 212

reports 213

start of trace 213

trcrpt command 213

trace command 211–212, 226

traceroute 183

tracing hung process 211

transmit errors 270

trcoff 264

trcon 264

trcrpt command 213

trcrpt commands 227

trcstop 264

trcstop command 213

TTY 218

## U

understanding the problem 8

unmount problems 154

update bundle 311

uptime 39

user (WLM) 286

user license 202

users view of a problem 8

## V

varyonvg command 138–139

VGDA - volume group descriptor area 138

VGID - volume group identifier 139

VGSA - volume group status area 139

view of a problem 8

Virtual Memory Manager 246

vital product data (VPD) 45

VMM 246

computational memory 246

DPSA 248

file memory 246

hash anchor table 246

high-water mark 260

I/O pacing 260

low-water mark 260

maxfree 250

maxperm 250

maxpgahead 259

minfree 248

minperm 250

minpgahead 259

page frame table 246

page stealing 246

page-replacement routine 249

- sequential-access read ahead 259
- translation lookaside buffer 246
- write-behind 259
- vmmerrlog structure 93
- vmstat 238, 247
  - the avm column 248
  - the b column 238
  - the cpu columns 239
  - the cs column 241
  - the cy column 250
  - the fault columns 240
  - the fr column 249
  - the id column 239
  - the in column 240
  - the kthr columns 238
  - the memory columns 248
  - the page columns 248
  - the pi column 249
  - the po column 249
  - the r column 238
  - the sr column 250
  - the sy column 239–240
  - the us column 239
  - the wa column 239
- vmtune 248
  - maxfree 250
  - maxperm 250
  - maxpgahead 259
  - minfree 248
  - minperm 250
  - minpgahead 259
  - random write-behind 259
  - sequential write-behind 259
- volume group
  - data 138
  - importvg 141
  - lock 139
  - redefinevg 141
  - VGDA 138
  - VGID 139
  - VGSA 139

## W

- w 39
  - flag table 39
- Web-based System Manager 10
- WLM
  - 32bit 287

- 64bit 287
  - fixed 287
  - plock 287
- wlm, localshm 283
- wlm, shared memory segment 282
- wlmcntrl command 293
- Workload Manager
  - class attributes 281
  - classes 276
  - inheritance 282
  - subclass 277
  - superclass 277
- write-behind 259
  - random 259
  - sequential 259

## Y

- ypbind daemon 33
- ypserv daemon 33





**Redbooks**

**IBM @server Certification Study Guide AIX 5L Problem Determination Tools and Techniques**







# IBM server Certification Study Guide - AIX 5L Problem Determination Tools and Techniques



**Developed specifically for the purpose of preparing for AIX certification**

**Makes an excellent companion to classroom education**

**For experienced AIX professionals**

This IBM Redbook is designed as a study guide for professionals wishing to prepare for the AIX 5L Problem Determination Tools and Techniques certification exam as a selected course of study in order to achieve the IBM server Certified Advanced Technical Expert - pSeries and AIX 5L certification.

This IBM Redbook is designed to provide a combination of theory and practical experience needed for a general understanding of the subject matter. It also provides sample questions that will help in the evaluation of personal progress and provide familiarity with the types of questions that will be encountered on the exam.

This publication does not replace practical experience, nor is it designed to be a stand-alone guide for any subject. Instead, it is an effective tool that, when combined with education activities and experience, can be a very useful preparation guide for the exam. Whether you are planning to take the AIX 5L Problem Determination Tools and Techniques certification exam, or if you just want to validate your AIX skills, this redbook is for you.

This publication was updated to include the new content included in Test 235, which is based on AIX 5L Version 5.1.

## **INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION**

### **BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE**

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

**For more information:  
[ibm.com/redbooks](http://ibm.com/redbooks)**