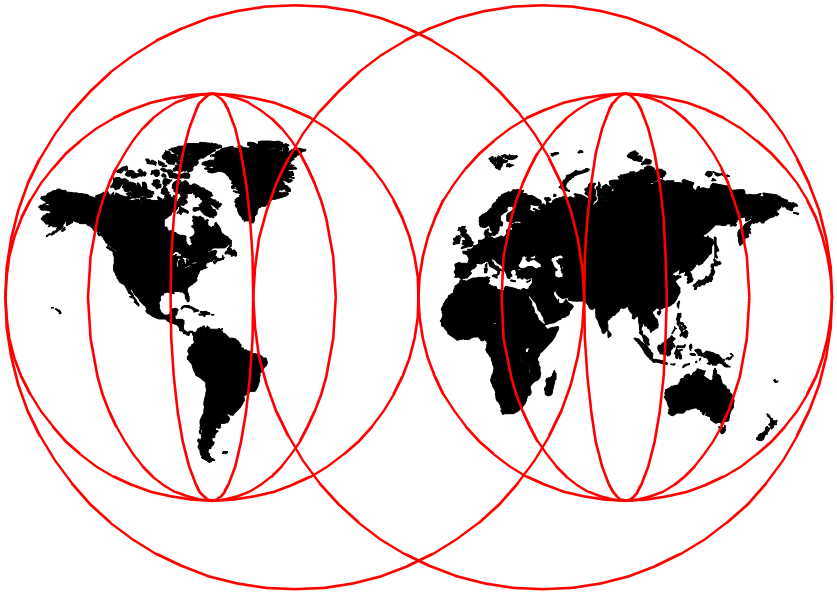
Problem Solving and Troubleshooting in AIX Version 4.3

Richard Cutler, Jaeyong An, Derrick Daines, John Hance,
SangSig Lee, Ma Jun, Yuan Er Fang



International Technical Support Organization

www.redbooks.ibm.com

SG24-5496-00

SG24-5496-00

International Technical Support Organization

Problem Solving and Troubleshooting in AIX Version 4.3

October 1999

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix A, “Special notices” on page 279.

First Edition (October 1999)

This edition applies to Version 4, Release 3 of the AIX Operating System, Program Number 5765-C34.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. JN9B Building 003 Internal Zip 2834
11400 Burnet Road
Austin, Texas 78758-3493

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

Contents

Figures

Tables

Preface

This redbook covers problem determination and troubleshooting on the RS/6000 platform. It is intended as a guide to the approach that should be adopted when attempting to resolve a problem on an RS/6000 system running AIX Version 4.3. It is not intended to replace the comprehensive documentation available for the AIX operating system or the System Installation and Service Guides available for RS/6000 system hardware. Instead, it should be considered as a useful supplement to guide you through the problem determination process.

A problem can manifest itself in many ways, and very often the root cause of the problem is not obvious. This redbook describes an approach to problem determination that will guide you through the initial steps in problem diagnosis and to narrow down and identify the component or subsystem that is experiencing a problem.

In addition to helping you determine the likely cause of your problem, this redbook describes some of the most common problems that can occur with AIX systems and, where possible, describes the actions you can take to resolve them.

This redbook is a valuable tool for system administrators and other technical support personnel who deal with RS/6000 and AIX problems.

The information in this redbook is not intended to completely cover problem determination on RS/6000 SP systems or RS/6000 systems configured as nodes in an HACMP cluster.

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.

Richard Cutler is an AIX and RS/6000 Technical Specialist at the ITSO, Austin Center. Before joining the ITSO, he worked in the RS/6000 Technical Center in the UK where he assisted customers and independent software vendors to port their applications to AIX.

Jaeyong An is an RS/6000 and AIX Technical Specialist at the Technical Support Center in IBM Korea. He currently works on AIX-related problem solving for customers, Business Partners, and IBM internal support.

Derrick Daines is a Senior Field Hardware Engineer in the United Kingdom. He has 10 years of experience working on RS/6000 products. He has worked at IBM for 31years. In the UK, he provides support to other engineers within the branch office and also provides backup help to the country support center.

John Hance is an Advisory I/T Specialist in Australia. He has worked with RS/6000 and AIX as a support professional since the products' inception. He has worked for IBM for 35 years on a wide variety of hardware and software products and currently provides level 2 product support in the IBM Support Center. His areas of expertise include dump analysis and TCP/IP.

SangSig Lee is an RS/6000 hardware, AIX, and AIX-related IBM software field support specialist in IBM Korea. He has five years of experience in RS/6000 and AIX. His areas of expertise include RS/6000 SP hardware, software, and Enterprise Storage products.

Ma Jun is an Advisory Technical Specialist in IBM China. He works on RS/6000, AIX, and storage product problem solving for customers, Business Partners, and IBM internal support. He joined IBM in 1995 and has a Bachelors degree in computer science from HeFei Industrial University.

Yuan Er Fang is a Technical Specialist in IBM China. He works on RS/6000, SP, and AIX problem determination for customers as well as field support. He joined IBM in 1994, and has five years experience in AIX and related products.

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

Kristina Aldred
IBM Atlanta

Jim Babka
IBM Austin

Dean Krahn
IBM Austin

Dennis Lee
IBM Austin

Jim Shaffer
IBM Austin

Comments welcome

Your comments are important to us!

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

- Fax the evaluation form found in “ITSO redbook evaluation” on page 305 to the fax number shown on the form.
- Use the online evaluation form found at `http://www.redbooks.ibm.com/`
- Send your comments in an Internet note to `redbook@us.ibm.com`

Chapter 1. Problem determination introduction

This redbook is intended to help system administrators and other support personnel have a better understanding of problem solving on RS/6000 systems using AIX Version 4.3. It is not the intention of this book to provide an exhaustive list of all the solutions to all of the possible problems that can be encountered on RS/6000 systems running AIX. Instead, the book is intended to guide you through the problem determination process and to narrow down and identify the component or subsystem that is experiencing a problem.

The approach detailed in this book will help you to determine the cause of a problem in a component or subsystem that was previously functioning correctly. It is not intended to help you diagnose problems when you are installing and configuring new hardware or software. You should refer to the appropriate documentation supplied with the new component for installation and configuration information.

Note that some of the Web addresses referenced in this book are IBM intranet addresses, and are, therefore, not available for general public access. Where possible, Internet sites offering similar information are referenced.

Although the title of this redbook refers specifically to AIX Version 4.3, much of the information supplied applies to other releases of AIX Version 4. You may find that some of the commands mentioned may not be present on your system, or may not support the exact syntax given. In this case, refer to the specific documentation supplied with the AIX level you have on the machine in question.

1.1 Problem determination process

For the purposes of this redbook, a problem can be considered as any situation where something that was previously working correctly is now not working as expected.

1.1.1 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. First of all, it can give you a hint as to the cause of the problem. Secondly, 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.

Take, for example, 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, 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 example described above, 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 users perspective is so important. In this example, a better way of proving that the problem has been resolved is to get the user to print their document.

1.1.2 Gathering information from the user

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

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

- What is the problem?

  Try to get the user 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 different 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 to 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 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 so, 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 they start talking, they invariably tell you things that will enable you to build a picture to help you to decide the starting point for problem resolution.

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

1.1.3 Gathering information from the system

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

It is not only the user of the machine that can provide information on 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`, 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 normally 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 often other files are viewed or extracted for sending to your IBM support function for analysis, such as system dumps or checkstop files.

1.1.4 Resolving the problem

Once you have defined the problem, you should use the flowcharts in Figure 1 on page 8, Figure 2 on page 9, and Figure 3 on page 10 and the symptom index cross reference chart in Table 3 on page 11 to direct you to the most relevant part of this book to perform basic troubleshooting of the suspected device or software component. During the investigative process, you should keep a log of the actions you perform in trying to determine the cause of the problem, and any actions you perform to correct the problem.

In some cases, you will encounter a problem that you can not resolve using the basic troubleshooting techniques described in this redbook, or the more detailed troubleshooting techniques described in the complete AIX or system hardware documentation. In this situation, you should report the problem to IBM. The information you have collected as part of the problem determination process should be supplied to IBM. If possible, you should create a simple test case to replicate the problem. In some cases, a test case may not be possible, for example, if a particular problem appears on a random basis and there is no apparent sequence of actions that trigger the problem.

1.1.5 Obtaining software fixes

Software fixes for AIX and many LPPs are available on the Internet from the following URL:

http://service.software.ibm.com/support/rs6000/

For a more customized approach to downloading AIX fixes, you can use the AIX application called FixDist instead of the Web. As a Web-alternative application, FixDist provides more discrete downloads and transparently delivers all required images with just one click. It can also keep track of fixes you have already downloaded, so you can download smaller fix packages the next time you need them. FixDist can be downloaded from the Web site mentioned above.

Once you have determined the nature of your problem, you should try searching this Web site or using FixDist to see if you are experiencing a known problem for which a fix has already been made available.

1.1.6 Other relevant documentation

Each RS/6000 machine has a specific set of documentation that should be used in the problem determination process when a hardware problem is suspected.

Every RS/6000 machine has a dedicated System Installation and Service Guide, which is supplemented by a number of additional documents, depending on whether the machine is a Micro Channel Bus system, or a PCI machine, otherwise known as a Multiple Bus system. Each bus type (Micro Channel and Multiple Bus) has a set of two manuals, one covering common diagnostic procedures on machines of that type, and the other describing the adapters, devices, and cables that can be used on systems of that type. In addition, the Multiple Bus systems have a guide detailing the placement rules for PCI adapters.

Most of the hardware documentation can be viewed online at the IBM RS/6000 Library Internet site. The URL for the site is:

http://www.rs6000.ibm.com/library/

Additionally, hardcopy versions of the manuals can be ordered from your IBM marketing representative. Table 1 shows details of the RS/6000 General Service Documentation. Table 2 on page 6 shows the details of the model-specific System Installation and Service Guides.

Table 1. RS/6000 General Service Documentation

Document	Form number
Adapters, Devices, and Cable Information for Micro Channel Bus Systems	SA38-0533

Document	Form number
Adapters, Devices, and Cable Information for Multiple Bus Systems	SA38-0516
Diagnostics Information for Micro Channel Bus Systems	SA38-0532
Diagnostics Information for Multiple Bus Systems	SA38-0509
PCI Adapter Placement Reference	SA38-0538
Site & Hardware Planning Information	SA38-0508

Table 2. RS/6000 System Installation and Service Guides

RS/6000 Model	Form number
7006 Graphics Workstation	SA23-2719
7009 Compact Server	SA23-2716
7012 300 Series	SA38-0545
7012 G Series	SA23-2741
7013 500 Series	SA38-0531
7013 J Series	SA23-2725
7014 Model S00 Rack	SA38-0550
7015 Model R00 Rack	SA23-2744
7015 R10/R20/R21 CPU Drawer	SA23-2708
7015 R30, R40, & R50 CPU Enclosure	SA23-2743
7017 S Series	SA38-0548
7024 E Series	SA38-0502
7025 F30 Series	SA38-0505
7025 F40 Series	SA38-0515
7025 F50 CPU Drawer	SA38-0541
7026 CPU Drawer	SA38-0535
7043 43P Series	SA38-0512
7043 Model 260	SA38-0554

RS/6000 Model	Form number
Enterprise Server H Series CPU Drawer	SA38-0547

1.2 Troubleshooting starting point

This section contains flowcharts and a table of symptoms of common problems. Based on the problem described by the user, you can either:

- Go directly to the chapter most relevant to your problem. You can use the table of contents at the front of the book or the index at the back of the book to decide which section is most relevant.
- If you have a boot problem, use the boot path flowchart starting in Figure 1 on page 8 in this section to decide on a course of action.
- Refer to the symptom index shown in Table 3 on page 11 to decide your next action.

1.2.1 Boot path flowchart

The boot path flowchart is designed to assist you in the diagnosis of boot problems. The chart takes you through the stages of the boot process, listing some of the LED codes that are signposts for either things completing successfully or indicating you have a problem. As you follow the chart, you will eventually come to an exit point. This exit point will direct you to the relevant chapter or section within the book. The section you are referred to will then aim to either provide you with suggestions for a solution, or direct you to a publication or Web site that will provide you with additional information to fix the problem. If neither of these solutions provide a solution, contact the organization that provides you with AIX software support.

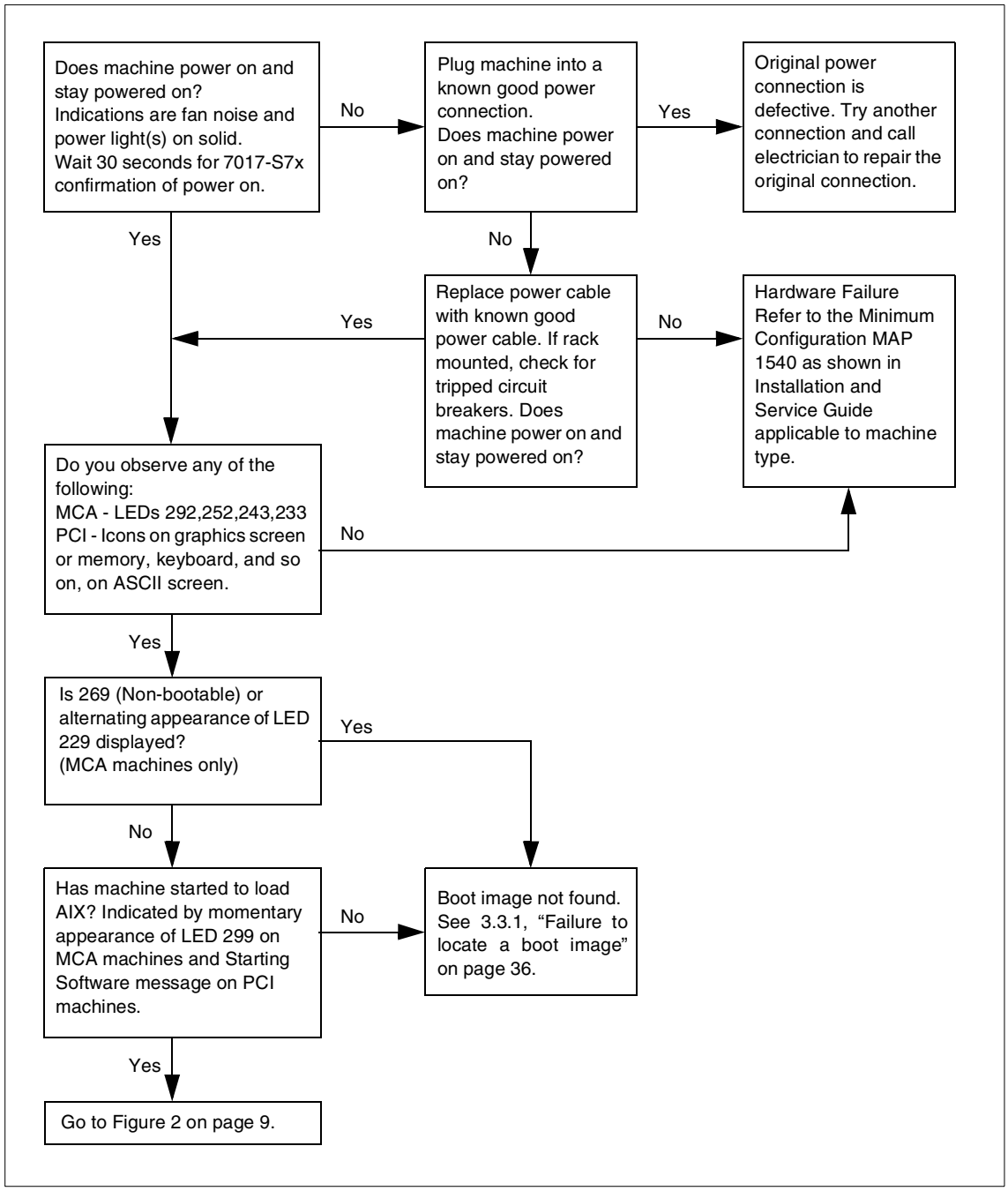


Figure 1. Boot path flowchart: Part one

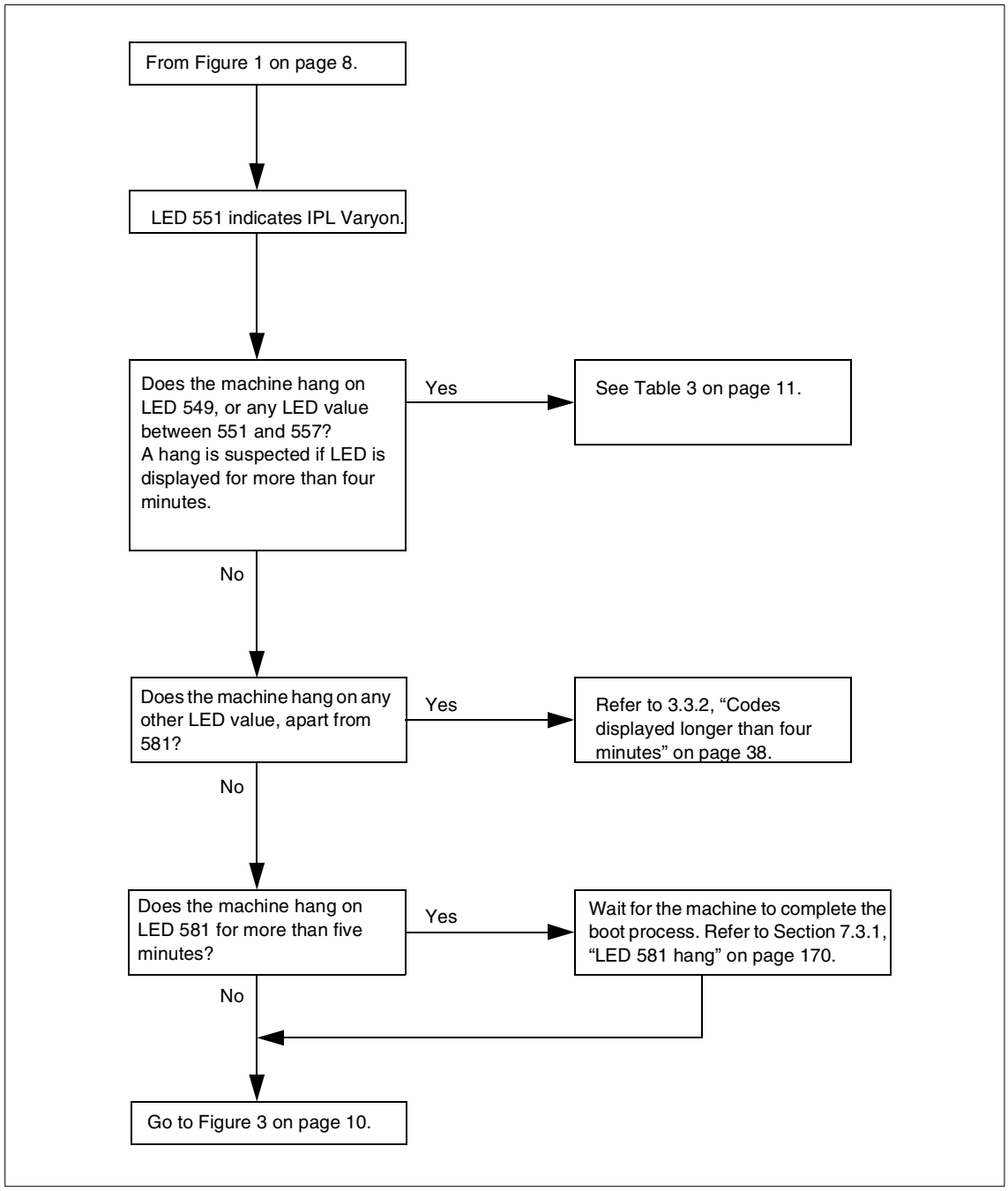


Figure 2. Boot path flowchart: Part two

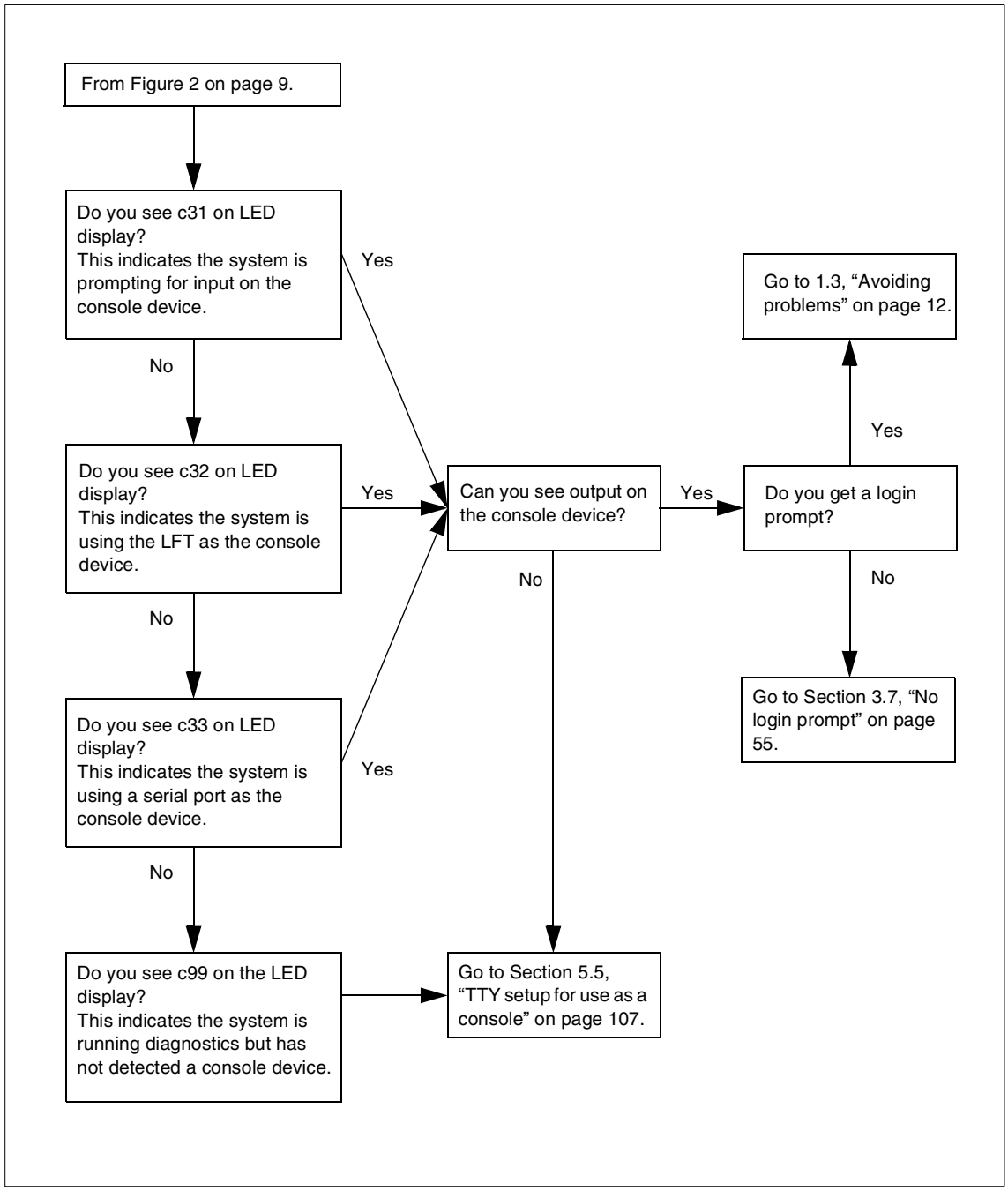


Figure 3. Boot path flowchart: Part three

1.2.2 Symptom index

Use this table as an alternative method of finding the correct section in the book that will help you deal with your problem.

Table 3. Symptom cross reference chart

Symptom	Possible Cause	Refer to
SCSI devices missing or duplicated	Address conflicts on the bus	Section 5.7.1.1, “SCSI devices missing or duplicated” on page 111
Checkstops or machine checks in error log	CPU or memory	Section 5.6, “Checkstops and machine checks” on page 108
LED 553	inittab problem, AIX corruption	Section 3.6.3, “LED 553 halt” on page 52
LED 551,555,557	A corrupted file system, JFS log, or defective disk	Section 3.6.1, “LED 551, 555, or 557 halt” on page 45
LED 552,554,556	A corrupted file system, JFS log bad IPL record, Corrupted Boot LV Defective Disk	Section 3.6.2, “LED 552, 554, or 556 halt” on page 48
LED 549	No system console or console problem with unsaved dump	Section 4.2.1, “LED 549” on page 58
LED 888-102	AIX kernel or kernel extension problem	Chapter 4, “System dumps” on page 57
System hang	AIX system resource	Section 4.8.6.3, “System hang” on page 87
Tape stuck in drive	Back up not completed, tape or tape drive problem	Section 5.9.4, “4 mm, 8 mm, and DLT tape drives” on page 119
Tape media errors	Defective media or drive needs cleaning	Section 5.9.4, “4 mm, 8 mm, and DLT tape drives” on page 119
TAPE_ERR in error log		Section 5.9.5.1, “Tape error log entries” on page 121
hdisk - pdisk mismatch	Disk from RAID set	Section 5.8.3, “SSA disk does not configure as hdisk” on page 114

Symptom	Possible Cause	Refer to
SSA errors in error log	SSA drive or cable	Section 5.8, “Serial Storage Architecture (SSA) disks” on page 112
Cannot log in to specific remote machine	Network problem Remote machine down	Section 7.2.1, “Selective host network problems” on page 156
Cannot log in to any remote machine	Network problem Adapter problem	Section 7.2.2, “No network access” on page 156
LED 581	TCP/IP problem	Section 7.3.1, “LED 581 hang” on page 170
Error log entries	Hardware or software	Section 2.3, “Viewing the error log” on page 18
Core dump error log entries	Software failure	Section 2.4, “Finding a core dump” on page 26
Machine fails to boot	Hardware or software	Section 1.2.1, “Boot path flowchart” on page 7
Corrupt boot list		Section 3.3.1, “Failure to locate a boot image” on page 36
3 Digit LED starting 0c	System dump in progress	Section 4.5.2, “Dump status codes” on page 66

1.3 Avoiding problems

The Reliability, Availability and Serviceability (RAS) features of AIX are designed to fulfil many functions. As well as helping you to determine the cause of a problem once it has actually occurred, such as the diagnostics subsystem, many of the RAS features are designed to provide information on potential problems before they occur. By default, RS/6000 systems are configured to run periodic automatic diagnostic checks. Any errors or warnings reported by this system will appear in the system error log.

Good system administration is not only fixing problems when they occur, but managing a system in such a way as to minimize the chances of a problem having an impact on the users of the system.

Periodic system maintenance can help reduce the number of problems experienced on a machine. Simple tasks such as cleaning the tape drive as required can prevent tape errors. Examining the system error log on a regular

basis can help you spot a potential problem when the related error log entries are still warnings rather than errors.

1.3.1 System healthcheck

The following section lists some simple commands that can be run on a regular basis to monitor a system. They will help you to become aware of how the system is operating.

- Use the `errpt` command to look at a summary error log report. Be on the lookout for recent additions to the log. Use the `errpt -a` command to examine any suspicious error log entries.
- Check disk space availability on the system with the `df -k` command. A full file system can cause a number of problems, so it's best to avoid the situation if at all possible. The two solutions to a full file system are to either delete some files to free up space, or use the LVM to add additional resources to the file system. The option you take will depend on the nature of the data in the file system, and whether there is any available space in the volume group.
- Check volume groups for stale partitions with the `lsvg` command. If stale partitions, logical volumes or physical volumes are reported, try and repair the situation with the `syncvg` command.
- Check system paging space with the `lsps -s` command. A system will not perform very well if it is low on paging space. In extreme circumstances, the system can terminate processes in order to solve the problem. Obviously, it is better that the system administrator ensures that there is enough paging space. You can either increase the size of existing paging space volumes, or add a new paging space volume. Again, the option you take will depend on the available space in the volume groups on the system.
- Check if all expected subsystems are running with the `lssrc -a` command.
- Check the networking by trying to `ping` a well-known address. If you cannot `ping` the address, refer to Chapter 7, "TCP/IP networking problems" on page 155.

Chapter 2. Error logging

This chapter describes the error logging subsystem. It shows how you can use various commands to view the error log and interpret the information it contains.

2.1 Error logging overview

The error logging process begins when an operating system module detects an error. The error-detecting segment of code then sends error information to either the errsave and errlast kernel services or the errlog application subroutine, where the information is, in turn, written to the /dev/error special file. This process then adds a time stamp to the collected data. The errdemon daemon constantly checks the /dev/error file for new entries, and when new data is written, the daemon conducts a series of operations.

Before an entry is written to the error log, the errdemon daemon compares the label sent by the kernel or application code to the contents of the *error record template repository*. If the label matches an item in the repository, the daemon collects additional data from other parts of the system.

To create an entry in the error log, the errdemon daemon retrieves the appropriate template from the repository, the resource name of the unit that detected the error, and detailed data. Also, if the error signifies a hardware-related problem and hardware Vital Product Data (VPD) exists, the daemon retrieves the VPD from the Object Data Manager (ODM). When you access the error log, either through SMIT or with the `errpt` command, the error log is formatted according to the error template in the error template repository and presented in either a summary or detailed report. Most entries in the error log are attributable to hardware and software problems, but informational messages can also be logged.

The system administrator can look at the error log to determine what caused a failure, or to periodically check the health of the system when it is running. Service personnel can also examine the log to help them service the machine.

The software components that allow the AIX kernel and commands to log errors to the error log are contained in the fileset bos.rte.serv_aid. This fileset is automatically installed as part of the AIX installation process.

The commands that allow you to view and manipulate the error log, such as the `errpt` and `errclear` commands, are contained in the fileset called

bos.sysmgt.serv_aid. This fileset is not automatically installed by earlier releases of AIX Version 4. Use the following command to check whether the package is installed on your system:

```
# lslpp -h bos.sysmgt.serv_aid
```

2.2 Error log file processing

The error log is used by system administrators and IBM service personnel to diagnose system problems. The error log contains error IDs, timestamp, error type, error class, and resource names associated with each error. Some error log entries also contain detailed data passed by the errlog or errsave routines, such as information about the slot number of the failing card or the name of the file that could not be opened.

Figure 4 shows how the error log is processed to allow users to view information about the system errors.

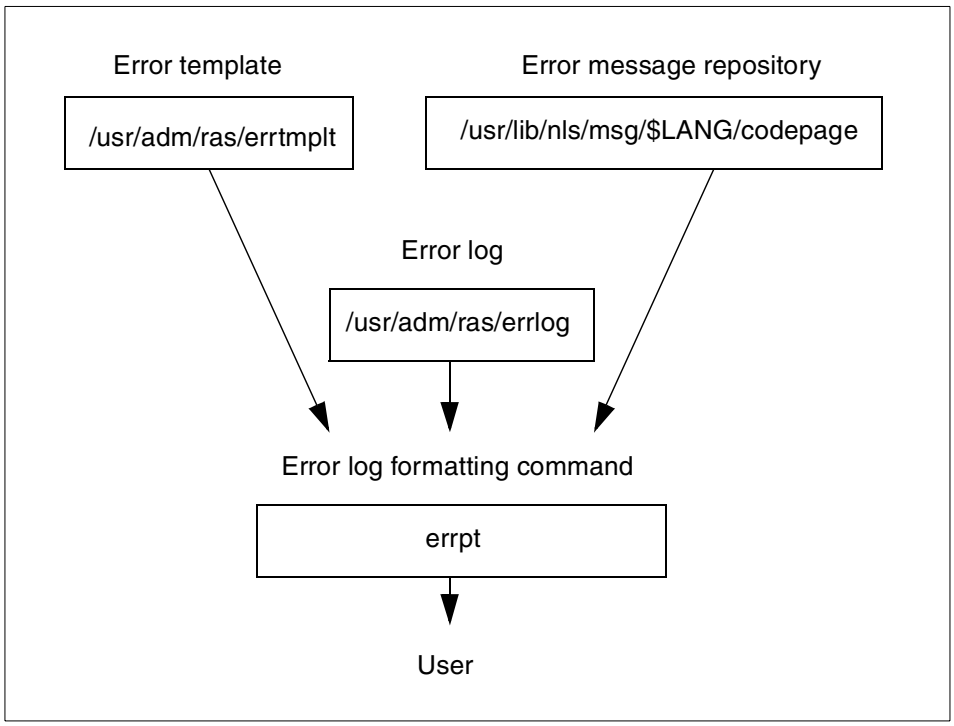


Figure 4. Error log file processing

2.2.1 Error templates

The error template contains numbers that correspond to error messages in the codepoint catalogue. These are sometimes referred to as codepoint messages. These error messages are used to communicate possible causes and to recommend actions for an error. They are also used to explain the detailed data that may accompany the error.

AIX Version 4.3.2 added the facility to define a template that uses regular XPG/4 NLS messages that are stored in a normal message catalog.

The template also is used to indicate whether or not the error should be reportable, loggable, or alertable. If it is not reportable, it will not appear when the `errpt` command is run. If an error is not loggable, it will not be put in the error log. For example, if a developer just wanted concurrent error notification to kick off an action, he or she may not want the error to show up in an error report.

Each template in the template file contains unique information that corresponds to a unique error. The contents of the error template are used to calculate the error ID of the error. The error numbers correspond to error messages indicating causes and recommendations for action.

The templates in the errtmplt file can be viewed by invoking the `errpt` command with the `-t` flag.

Templates are installed into the errtmplt file using the `errupdate` command. This is carried out automatically as part of the installation process of any filesets that contain new error templates.

2.2.2 Error messages

Error messages are numbered and placed in a separate file, called the codepoint catalogue. Similar to message catalogues, the codepoint catalogue is a translated file that exists in the different language directories. The user's LANG environment variable determines which codepoint catalogue is accessed when the `errinstall`, `errmsg`, or `errpt` commands are run.

The codepoint catalogue can be viewed by using the `errmsg` command with the `-w` flag.

The error message source is installed in the codepoint catalogue using the `errinstall` command. This is carried out automatically as part of the installation process of any filesets that contain new error messages.

Note

If the LANG variable is set to C or to some non-existent language, the default codepoint catalogue in the /var/adm/ras directory is accessed.

2.3 Viewing the error log

You can generate an error report from entries in an error log. The `errpt` command allows flags for selecting errors that match specific criteria. By using the default condition, you can display error log entries in the reverse order they occurred and were recorded. In other words, the latest error log entry appears first.

There are two main ways of viewing the error log:

- You can use the System Management Interface Tool (SMIT) with a fast path to run the errpt command. To use the SMIT fast path, enter:

  ```
  # smit errpt
  ```

  After completing a dialog about the destination of the output, you will see a screen similar to that shown in Figure 5.

```
                        Generate an Error Report

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

[TOP]                                                  [Entry Fields]
  CONCURRENT error reporting?                         no
  SUMMARY or DETAILED error report                    summary                +
  Error CLASSES (default is all)                      []                     +
  Error TYPES   (default is all)                      []                     +
  Error LABELS (default is all)                       []                     +
  Error ID's      (default is all)                    []                     +X
  Resource CLASSES (default is all)                   []
  Resource TYPES   (default is all)                   []
  Resource NAMES  (default is all)                    []
  SEQUENCE numbers (default is all)                   []
  STARTING time interval                              []
  ENDING time interval                                []
  LOGFILE                                             [/var/adm/ras/errlog]
[MORE...3]

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

Figure 5. SMIT Generate an Error Report screen

- You can also view the error log from the command line using the `errpt` command. When used from the command line, considerable amounts of output can often be generated, so it is best to control the command by piping the output to either the `more` or `pg` commands, which allow it to be viewed one screen at a time. When invoked with no options, `errpt` will display a summary report, listing one line of information about each error log entry. An example of this is shown in Figure 6.

```
 itsosrv1:/> errpt
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
C60BB505   0525162099 P S SYSPROC        SOFTWARE PROGRAM ABNORMALLY TERMINATED
610BDA5E   0525122899 P S ssa0           UNABLE TO LOG AN ERROR AGAINST A PDISK
C60BB505   0521153099 P S SYSPROC        SOFTWARE PROGRAM ABNORMALLY TERMINATED
AA8AB241   0521152899 T O CMDCRASH_EXIT  OPERATOR NOTIFICATION
AA8AB241   0521152899 T O CMDCRASH_ENTRY OPERATOR NOTIFICATION
369D049B   0521152799 I O SYSPFS         UNABLE TO ALLOCATE SPACE IN FILE SYSTEM
AA8AB241   0521140199 T O CMDCRASH_EXIT  OPERATOR NOTIFICATION
AA8AB241   0521140199 T O CMDCRASH_ENTRY OPERATOR NOTIFICATION
AA8AB241   0521140099 T O CMDCRASH_EXIT  OPERATOR NOTIFICATION
AA8AB241   0521140099 T O CMDCRASH_ENTRY OPERATOR NOTIFICATION
AA8AB241   0521114999 T O CMDCRASH_EXIT  OPERATOR NOTIFICATION
AA8AB241   0521114899 T O CMDCRASH_ENTRY OPERATOR NOTIFICATION
2712BEF2   0519140199 T H ent0           TRANSMIT FAILURE
9DBCFDEE   0519141999 T O errdemon       ERROR LOGGING TURNED ON
192AC071   0519132499 T O errdemon       ERROR LOGGING TURNED OFF
itsosrv1:/>
```

Figure 6. Summary error report

In addition to the summary report, the `errpt` command can be used with various flags to generate a customized report detailing the error log entries you are interested in:

- To display information about errors in the error log file in detailed format, enter:

  ```
  # errpt -a
  ```

- To display a detailed report of all errors logged for a particular error identifier, enter:

  ```
  # errpt -a -j identifier
  ```

  Where `identifier` is the eight digit hexadecimal error identifier.

- To clear all entries from the error log, enter:

  ```
  # errclear 0
  ```

- To stop error logging, enter:

  ```
  # /usr/lib/errstop
  ```

- To start error logging, enter:

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

Software service aids configuration information is stored in the /etc/objrepos/SWservAt ODM database. This ODM class is used to store information about the location and size of various log files used by the system. It is also used to hold information about trace hooks available for use by the trace subsystem, described in Chapter 11, “Event tracing” on page 259.

2.3.1 Error log management

By default, AIX runs a cron job that deletes all hardware error log entries older than 90 days and all software and operator message error log entries older than 30 days. The cron job simply uses the `errclear` command to delete the old entries. If you are investigating a software problem that has been on a machine for a long time, do not assume that the first instance of the error in the error log was caused the first time the software problem occurred. Previous entries may have been deleted if older than 30 days.

2.3.2 Reading a summary error log

A summary error log report, obtained by using the `errpt` command with no flags, contains the following information for each error log entry.

2.3.2.1 Identifier

The error identifier is a 32-bit CRC hexadecimal code that determines which error record template is used to interpret the information contained in the error log entry.

2.3.2.2 Timestamp

This is the time and date when the error occurred, formatted as MMDDhhmmYY.

2.3.2.3 Type

The type indicates the severity of the reason for logging the error. There are five different types of errors, along with a sixth type for when the severity can not be determined:

PEND The loss of availability of a device or component is imminent.

PERF The performance of the device or component has degraded to below an acceptable level.

PERM Condition that could not be recovered from. Error types with this value are usually the most severe errors and are more likely to mean that you have a defective hardware device or software

module. Error types other than PERM usually do not indicate a defect, but they are recorded so that they can be analyzed by the diagnostics programs.

TEMP Condition that was recovered from after a number of unsuccessful attempts. This error type is also used to record informational entries, such as data transfer statistics for DASD devices.

UNKN It is not possible to determine the severity of the error.

INFO The error log entry is informational and was not the result of an error.

The summary error log report only displays the first letter of the error type.

2.3.2.4 Class

This indicates the general source of the error. The classes are:

H This means hardware device failures or media errors. When you receive a hardware error, refer to your system operator guide for information about performing diagnostics on the problem device or other piece of equipment. The diagnostics program tests the device and analyzes the error log entries related to it to determine the state of the device. Refer to Chapter 5, “Hardware problem determination” on page 89 for more information.

S This means software application program failures, system program failures, and kernel problems, such as low paging space.

O This indicates an operator notification error that gets logged when the `errorlogger` command is used.

U This means the source of the error is undetermined.

2.3.2.5 Resource name

Name of the resource that has detected the error. For software errors, this is the name of a software component or an executable program. For hardware errors, this is the name of a device or system component. It does not indicate that the component is faulty or needs replacement. Instead, it is used to determine the appropriate diagnostic modules to be used to analyze the error.

2.3.2.6 Description

Gives a description of the logged information.

2.3.3 Reading error logs in detail

To obtain a detailed report of all errors logged by the system, enter:

```
# errpt -a | pg
```

A detailed error log report may contain multiple entries. Each entry contains multiple fields. The exact fields that are displayed depend upon the nature of the error. The following fields are always included in every error entry:

LABEL Predefined name for the event.

ID Numerical identifier for the event.

Date/Time Date and time of the event.

Sequence Number Unique number for the event.

Machine ID Identification number of your system processor unit.

Node ID Mnemonic name of your system.

Class General source of the error. See Section 2.3.2.4, “Classes of Error Log” on page 27 for a description of the class types.

Type Severity of the error that has occurred. Five types of errors are possible. See Section 2.3.2.3, “Type” on page 20 for a description of the error log types.

Resource Name Name of the resource that has detected the error. For software errors, this is the name of a software component or an executable program. For hardware errors, this is the name of a device or system component. It does not indicate that the component is faulty or needs replacement. Instead, it is used to determine the appropriate diagnostic modules to be used to analyze the error.

Description Summary of the error.

The following fields may or may not be present, depending on the nature of the error:

Resource Class General class of the resource that detected the failure.

Resource Type Type of the resource that detected the failure.

Location Code Path to the device. There may be up to four fields that refer to drawer, slot, connector, and port, respectively.

VPD Vital product data. The contents of this field, if any, vary. Error log entries for devices typically return information concerning the device manufacturer, serial number, Engineering Change (EC) levels, and Read Only Storage (ROS) levels.

Probable Causes List of some of the possible sources of the error.

User Causes List of possible reasons for errors due to user mistakes. An improperly inserted disk or external devices (such as modems and printers) that are not turned on are examples of user-caused errors.

Install Causes List of possible reasons for errors due to incorrect installation or configuration procedures. Examples of this type of error include hardware and software mismatches, incorrect installation of cables or cable connections becoming loose, and improperly configured systems.

Failure Causes List of possible defects in hardware or software.

> **Note**
>
> A failure causes section in a software error log usually indicates a software defect. Logs that list user or install causes or both, but not failure causes, usually indicate that the problem is not a software defect. If you suspect a software defect, or are unable to correct user or install causes, report the problem to your software service department.

Recommended Actions Description of actions for correcting the failure. For hardware errors, PERFROM PROBLEM DETERMINATION PROCEDURES is one of the recommended actions listed. For hardware errors, this will lead to running the diagnostic programs. Refer to Chapter 5, “Hardware problem determination” on page 89 for more information.

Detailed Data Failure data that is unique for each error log entry, such as device sense data.

2.3.4 Examples of detailed error reports

The following are sample error report entries that are generated by issuing the `errpt -a` command.

The error log entry in Figure 7 is taken from a system with a hardware problem. The entry shows that this system has a problem related to the SCSI adapter card with logical device name scsi0 located in slot number 08.

```
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 7. Sample hardware error log entry

The error indicates the termination of the SCSI bus failed, probably due to a defective cable or terminator. The error could not be recovered from, indicated by the Error Class of PERM. The VPD section displays information that is required if you need to replace the adapter card. Usually, you need to use the FRU number for replacement action, *not* the part number. For more information about hardware diagnostics, refer to Chapter 5, “Hardware problem determination” on page 89.

```
LABEL:          CORE_DUMP
IDENTIFIER:     C60BB505

Date/Time:       Tue Aug  3 21:19:46
Sequence Number: 159
Machine Id:      000420DCA000
Node Id:         asterix
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
          11
USER'S PROCESS ID:
       29182
FILE SYSTEM SERIAL NUMBER
12
INODE NUMBER
          25
PROGRAM NAME
a.out
ADDITIONAL INFORMATION
bcopy AC
process_a 2C
main 1C
__start 8C

Symptom Data
REPORTABLE
1
INTERNAL ERROR
0
SYMPTOM CODE
PCSS/SPI2 FLDS/a.out SIG/11 FLDS/bcopy VALU/ac FLDS/process_a
```

Figure 8. Sample software error log entry

The example shown in Figure 8 is of an error logged when a software program core dumped. The first thing we need to know is if it was an AIX executable that dumped core. This can be found by looking at the name of the problem executable, given in the PROGRAM NAME data field. If this program is a user application or a third-party application, it is up to the developer of the application to fix it.

If the core dumping executable is an AIX executable, proceed to debug the problem. One clue to start with is the SIGNAL NUMBER in the error log. In this example, the program dumped core due to a SIGSEGV (Segmentation Violation). You can build a debug version of the executable and proceed to debug it with the `dbx` command. To use the `dbx` command, you need to install the bos.adt.debug fileset.

If you have an AIX command that repeatedly terminates and leaves a coredump, you should send the core dump to IBM for further analysis.

The last section of the error log entry produced when an application core dumps normally shows a stack trace of the executable at the time of the failure.

2.4 Finding a core dump

Lots of old core dumps over time can create a problem by filling up file systems. Sometimes, you will want to find the core dumps, examine them, and remove them to save space. This is particularly useful if you are administering a multiuser system; you will not know about all the core dumps created because they were most probably created by other users' applications.

To find the core file corresponding to an error log entry, use the `corepath` command located in the /usr/samples/findcore directory.

This facility is part of the bos.sysmgt.serv_aid fileset.

Chapter 3. Boot problems

The aim of this chapter is to guide you through a series of steps that are aimed at diagnosing most problems you will encounter with the boot process. In the first section, we show how the boot process normally works, because in any fault finding, it is always best to first understand how the process should work. This makes the process of determining what is going wrong much easier.

3.1 Types of machines

Certain procedures, particularly those related to the boot process, are implemented in a different way depending on the type of machine. This section describes the two main types of RS/6000 machines, and the terminology used to distinguish between them for the remainder of this redbook.

The RS/6000 family of machines was launched in 1990 and, over the years, has changed to adopt new technology as it becomes available. The first RS/6000 machines were based around the Micro Channel Architecture (MCA) and had a number of features common to each machine in the range, in particular, a three digit LED and a three position key mode switch. Throughout the rest of this redbook, machines of this type are referred to as *MCA machines*.

In recent years, the RS/6000 family has migrated to Peripheral Component Interconnect (PCI) bus technology. Throughout the rest of this redbook, machines of this type are referred to as *PCI machines*. Initial machines of this type (7040 and 7248) did not have the three digit LED or three position key mode switch of the previous MCA machines. Subsequent PCI machines have an LED display, but none have the three position key mode switch.

3.2 The boot process

The boot sequence, otherwise known as the Initial Program Load (IPL), is, in principle, similar on all computer systems. The basic steps carried out by the hardware are:

Boot Stage One	Check the power, CPU, and memory systems.
Boot Stage Two	Look for the operating system image.
Boot Stage Three	Load and run the operating system image, and configure devices and subsystems.

The RS/6000 range of machines are no different and so adhere to the above model, though some machines in the range have a pre-boot state. This pre-boot state is when the service processor is accessible to alter machine settings and obtain error information. Because of variations in models, such as some machines having service processors, the method by which the boot model is actually achieved varies.

You may find it useful while using this section to refer to the flowchart in Section 1.2.1, “Boot path flowchart” on page 7 to determine at which stage your boot process fails.

3.2.1 Pre-boot state

The Micro Channel Architecture (MCA) machines 7012 G series, 7013 J series, and 7015 R series and the PCI bus machines 7025-F50, 7026-H50, 7026-H70, and 7043-260 have a stage prior to starting the boot process. These machines have a mini processor, often referred to as the service processor, that can be accessed prior to the start of the boot sequence. Dependent on machine type, the service processor can be used to determine the amount of hardware testing to be performed during the first stage of the boot process and record any errors generated. The service processor can also govern system monitoring while the machine is running AIX.

This rest of this section will explain some of the more often used functions of the service processors.

3.2.2 BUMP program

The Bring Up Multiprocessor (BUMP) function is present within the firmware on the MCA Symmetrical Multiprocessor (SMP) family of machines, consisting of 7012 G series, 7013 J series, and 7015 R series. The program controls the initial stages of the boot process up to the point where the NVRAM starts looking for a boot image. The program has two stages. The initial stage is where the machine is powered off. The second stage is during a Service mode boot after all CPU and memory testing has been completed.

3.2.2.1 Standby menu

The standby menu is available only on an ASCII terminal attached to the S1 serial port when the machine is not running AIX but is still powered on. The LCD control panel will be displaying the word `Standby` at this point and continues to do so while the BUMP program is being accessed. To access this menu, you need to have the key in the Service position and the greater than sign prompt (`>`) displayed on the console. Type `sbb` and then press the

Return key. You will get a menu that displays a number of options. The two most frequently used options are Display Configuration and Set Flags:

Display Configuration This option enables you to view the configuration detected when the last IPL was attempted. If the power has been completely removed from the machine upon reconnection, this selection will only show the CPU and memory configuration as `C` for configured but mark other adapter positions as `A` for absent.

Set Flags This option enables you to set parameters before the machine boots. A full boot of an eight processor machine with 4 GB of memory can take a substantial length of time. The time can be decreased by setting option six, Fast IPL, to enabled. It is also possible to set this Fast IPL parameter from the AIX command line prior to running the `shutdown` command. When running, the following command will configure the machine so that the next time you shut the machine down the subsequent boot will be the fast one:

```
mpcfg -cf 11 1
```

To get the machine to boot after altering any of the parameters, exit from the menus out to the `>` prompt. If you want a Normal mode boot, turn the key to Normal, type `power`, then press the **Enter** key. If you require a Service mode boot, leave the key in the Service position, type `power`, then press the **Enter** key. You may need to press the Power button on the front of the machine at this time, depending upon how the machine was last powered off.

It is recommended that the BUMP menus relating to system test are only accessed by trained service personnel, since some of the messages that can be displayed are not actually problems and can lead to hours of needless troubleshooting.

3.2.2.2 PCI service processor access

Machine types 7025-F50, 7026-H50, 7026-H70, and 7043-260 have service processors as standard equipment. Certain other PCI models, such as the 7025-F40, can have them as an optional feature. The service processor provides a level of error detection as well as some control over certain settings on the machine. To access the service processor on all of the PCI machines, except for 7017-S70 and 7017-S7A machines, attach an ASCII terminal to the S1 serial port and press the **Enter** key when the machine is

not running AIX and the power is still connected. With the service processor menu displayed, you will see E075 displayed on the LED panel.

Once you have exited the service processor menus, you can power on the machine.

3.2.3 Boot stage one

This is the portion of the boot sequence where the power supply, CPU, and memory systems are checked. All checks are performed by hardware; no operating system is loaded at this time.

3.2.3.1 Micro Channel Architecture machines

The Micro Channel Architecture (MCA) range of machines are 7011, 7012, 7013, 7015, 7030, 7009, and 7006. All of these machines have a display on the front of the machine that uses three digits to indicate status of the machine during the boot.

Built-In Self-Test (BIST) LED 100-199

The routines run during this LED sequence include testing of CPU, memory, and the initial bringing up of the power supply. Since it is not possible to test a CPU or any portion of memory while it is in use, it is performed at this time by the On-board Chip Sequencer (OCS). This is a piece of hardware that runs tests on the memory and CPUs, and if it finds a problem, writes an entry in the NVRAM. These tests are usually a quick process, but if the machine is a fully configured eight processor SMP machine, you can expect it to take ten to twenty minutes or more to test all components in the CPU complex.

Stationary LEDs of 185, 111, 112, or 113 indicate that the machine has found a problem that it can not recover from.

SMP Maintenance Menu

This menu is displayed on 7012 G series, 7013 J series, and 7015 R series machines after the tests have been run to the CPU and memory and the adapter cards and the buses have been initialized. This menu will only appear when the flags shown in Table 4 have been set from the Set Flags option of the standby menu:

Table 4. Flag settings to obtain maintenance menu

Flag	Value
BUMP Console present	Enabled
Auto service IPL	Disabled

See Section 3.2.2.1, “Standby menu” on page 28 for details on how to set these flags.

If the maintenance menu fails to display, and the key is in Service, and flags are set as shown above, the most likely cause is that the system ID has been corrupted or lost. This most often happens when both the part containing the master copy of the system ID and the part containing the backup copy are changed together. On the 7013 J series, the master copy is held on the backplane, and on the 7015 R series, the master copy is held on the L2 planar. On both the 7013 and 7015 machines, the backup copy is held on the System Interface Board (SIB). The 7012 G series, being much smaller, has the master copy in the control panel and the backup copy on the system planar.

The maintenance menu will include another view of the configuration giving more detailed information, BUMP error log entries, power off facility, and a boot aid.

3.2.3.2 PCI Bus machines except 7017

Boot stage one on 7020, 7024, 7025, 7027, 7043, and 7248 machines differs from the 7017 due to firmware differences. These machines also perform the first two stages differently from the MCA bus machines. The difference being that the machines have either a system planar with firmware on it, or they have a service processor with code on it plus the system planar with firmware on it. The firmware on the system planar and service processor can be refreshed when updates become available. The updates are usually to accommodate new hardware releases or cure known problems. Most of the MCA machines have a permanent ROS embedded on the system planar to which no changes can be made. The MCA G, J, and R series machines, however, do use a version of reloadable firmware but this only controls CPU and memory functions.

When the machine is under the control of the firmware, you will see either a three digit Fxx code or a four digit Exxx code displayed in the LED panel. The codes you see depend on the machine type. The firmware controls the testing and verification of the memory, CPUs, and the initialization of the PCI bus components.

3.2.3.3 PCI machine icons

After a period of time, the machine will either display a series of icons depicting keyboard, memory, network, SCSI, network, and speaker on the graphical screen. As the speaker icon appears, if the machine has a speaker, it will emit a beep. Or, you are using an ASCI console, instead of seeing

graphical icons, you will see the words keyboard, memory, network, SCSI, network, and speaker. Throughout the rest of this redbook, this startup screen will be referred to as the *PCI Icons screen.*

Before the speaker icon or word appears, a keyboard response must be received if you want to do anything other than boot the machine in Normal mode. During this phase of the boot process, additional error codes of eight hexadecimal characters can be displayed on the screen. If the machine has an eight digit LED panel, the codes can also be displayed on it. These codes are generally not fatal but are signalling to you that something requires your attention. The meaning of these codes can be found in the relevant service guide for the machine. System Installation And Service Guides for most RS/6000 machines can be viewed on the Internet at the IBM RS/6000 Library:

`http://www.rs6000.ibm.com/library/`

See Table 1 on page 5 and Table 2 on page 6 for a full list of current documentation.

3.2.3.4 7017 Models S70 and S7A

Although these machines are PCI-based, they differ substantially in the way that they boot and access service processor menus. The reason for this is that the CPU complex has a large portion of code that is used to run the CPU complex and the Serial Power Control Network (SPCN). This portion of the firmware is unique to this machine and, when it is running, displays eight digit codes in the LED panel. Additionally, when an error code is produced, an orange attention light is illuminated and the first portion of the error code displayed on the panel. There is a possibility of a further eight additional panels of error information being displayed on the control panel. To access these, you need to use the scroll and Enter buttons on the operator panel. These additional fields are quite often used in the process of fault determination and so should be collected any time that an error code is displayed.

Boot Process Differences

On the S70 and S7A, you can set the speed of the boot process using the control panel buttons. There are three settings for the speed, plus a special condition setting:

F This gives a fast boot path through the CPU memory checks. Basic testing of the CPU complex takes place

S This is the slow boot option. This fully tests the CPU and the memory and also checks all data paths between each CPU and the memory

subsystem. This will take a substantial length of time if you have the maximum 12 processors and 32 GB of memory.

SE This setting should only be used when a part has been replaced and there is a need to get the machine up and running so that firmware code can be loaded to support the EC level of the new part. This setting ignores EC checking during the slow boot process it runs. This setting will probably be used when being directed by IBM support personnel.

V This setting means variable. This enables the machine to decide the speed it will use during the boot process. Normally, the fast option is used. However, anytime the machine has detected an error either during a previous boot sequence, or the machine has instigated a reboot, for example, after a checkstop has been detected, then a slow boot will be used.

To power on the machine, you need to press and hold the Power button for approximately one second. The machine will then at first not appear to do anything for approximately 30 seconds, then the power LED will start to flash. Once the power supply has fully started, the LED panel will have progress codes made up of eight characters, quite often with one of the characters changing rapidly. This will carry on for a few minutes. You should observe E043 displayed in the LED panel momentarily followed by E07A accompanied by three beeps from the speaker in the primary I/O drawer. If you wish to enter the service processor menus, press any key on the ASCII terminal after the beeps but within the 10 second window before the E07A code changes on the LED panel. Once the service processor menus have been entered, it is possible to check the service processor error logs, read Vital Product Data (VPD) of components, and perform many other functions.

The service processor is situated in the primary I/O drawer in the rack adjacent to the CPU rack, and all communication between the service processor and CPU complex is carried out using the JTAG and operator panel cables.

When you exit from the service processor menus, or if you did not enter them, the machine will continue the boot process in a similar way to the other PCI-based machines. However, the time taken to get to a login prompt may be considerable since it is possible to have a configuration having up to 56 adapters in machines with four I/O drawers configured.

3.2.4 Boot stage two

At this point in the boot process, the CPU and memory have been tested. The machine now starts to look for an operating system image.

3.2.4.1 MCA machines LEDs 200-299

This LED sequence on all MCA machines is known as the Power-On Self-Test (POST) phase. The LEDs indicate the sequence up to the point where the machine has read a valid boot image and has just handed over the boot process to AIX. At this point, 299 is momentarily displayed on the LED panel on MCA machines.

The POST sequence will only start once it has been established that the hardware present within the machine is capable of running a system, in other words, it has a CPU and some memory. The next step is to look for a boot image. The next series of numbers displayed indicate the adapters and devices it is starting up in the quest to find the boot image. Important numbers to look out for are:

292 Initializing a SCSI adapter. Needed to run the disk containing AIX.

252 Locating the diskette drive or reading from a bootable diskette media.

243 or 233 Booting from a device listed in the NVRAM boot list. Usually hdisk0, a bootable CD, or a mksysb tape.

Non Volatile Random Access Memory (NVRAM) is the device that holds the boot list, the list of the places to look for a boot image.

3.2.4.2 All PCI machines

On PCI machines, the distinction between stage one and stage two of the boot process is not as clear as it is on MCA machines. The point at which the display attached to the machine first displays the PCI Icon screen, as described in Section 3.2.3.3, “PCI machine icons” on page 31, nearly matches the start of the 200-299 LED codes of a MCA machine.

Once the icons have disappeared from the screen, the machine then uses the boot list in the NVRAM to look for a boot image. The boot image is then read and AIX is started to be loaded. When the machine finds a valid boot image, it displays the message `Software starting please wait`, followed by some information detailing the SCSI ID of the device it loaded the boot image from. Unlike the MCA machines, you do not get 299 displayed on the LED panel to confirm that a boot image has been found, instead you will eventually see displayed the three digit AIX LED codes you are familiar with. On some

machines, the three digit code is actually four digits but the first digit is always a zero. From this point on, the PCI machines behave in a similar way to the MCA-based machines.

3.2.4.3 Boot stage three

Boot stage three is common across all platforms and commences at the point when the initial RAM image of AIX is loaded from the boot device. From this point, all LED codes displayed are generated by AIX. AIX generates codes in the range from 300 to 999 and cxx. Some machine types will place a leading zero to give 0300 to 0999 and 0cxx.

Machine type 7024, when running AIX levels below 4.2.1, will not show the stage three LED sequence. The driver for the LED panel was not added until AIX 4.2.1

AIX start to login prompt LEDs 300-999

This is generally the longest part of any boot sequence. It is the section where all of the disks are started, tested, and then configured. While this is happening, 570 or 80c will be displayed on the LED panel, interspersed with 538, which indicates that a configuration method is being run. During the whole of the boot sequence, no LED should be displayed for longer than four minutes. However, if you have a large disk configuration or a complex TCP/IP setup, you may see 570, 80c, or 581 displayed for longer than the four minute period. In the case of the 570 and 80c codes, the LED panel is actually alternating between the two numbers very quickly. You may not see them change.

Important LEDs to look out for in this sequence are:

551 This is an indication that all devices in the machine are configured correctly and the machine is ready to varyon the root volume group.

517, 553 Once these two LEDs have been displayed, any problem experienced after this point is more than likely going to be AIX-related as opposed to hardware-related.

581 TCP/IP configuration is taking place. If this number stays on the LED panel for a very long time, you should perhaps look at your TCP/IP settings and routing information once you are able to login to the system. See Section 7.3.1, “LED 581 hang” on page 170 for more information.

c31 This code indicates the system is awaiting input from you on the keyboard. This is usually encountered when booting from CD or

mksysb tape. This is normally the dialogue to select the system console.

c32, c33 These codes tell you that the boot process is nearly complete. Shortly afterwards, you should see output on the screen from the AIX boot process starting various software subsystems.

3.3 Boot problem determination

Boot problems fall into a number of categories:

- Power on or start up problems.
- Failure to find a boot image.
- Hang somewhere in the boot process.
- Machine gets some or all of the way up and cannot be used or contacted.

3.3.1 Failure to locate a boot image

As described earlier in this chapter, the boot list held in the NVRAM stores the list of the locations where the machine will look for a boot image. In some situations, the boot list can be become corrupted, be erased, or list non-existent devices.

Erasure of the boot list is the most usual reason for machines not being able to locate a boot image. Other reasons for the machine not actually finding a boot image are:

- Hardware problem on the boot device or on the SCSI bus to which the boot device is connected. See Section 5.7, “Diagnosis of SCSI problems” on page 109 for more information.
- Devices listed in the boot list either non-existent or do not contain a valid boot image.

The recovery procedure depends upon the reason for the failure. The first thing to establish is where the boot list tells the machine to look for a boot image. Once you have established what is in the boot list, then you can add to it or correct it as appropriate.

3.3.1.1 Access boot list on PCI machine

To access the boot list on PCI machines, use the following procedure:

1. Bring the machine up to SMS. See Section 5.4, “System Management Services (SMS)” on page 106.
2. Select **Multiboot** (option 6).

3. Select **Select Boot Devices** (option 4).
4. Select **Display current settings** (option 1).

   If the current settings show a device that you know to be present and in the position that you know it should be, run diagnostics from CD against that device. If the diagnostics run clean, the boot image on the disk may be corrupted. To fix this problem, refer to Section 3.5, “Accessing rootvg from bootable media” on page 44. Once you have accessed rootvg, you need to create a new boot image with the `bosboot` command.

   If the current settings show no devices, return to the Select Boot Devices menu. Select **Configure first boot device** (option 3). You will now see a list of all SCSI devices and network adapters. Choose from the list and enter the number of the device that you want to boot from and press **Enter**. The screen will now display the current boot list. If you wish to add additional devices, repeat the process.

   Once you have devices listed in the boot list, exit from SMS and continue the boot process. Sometimes, you may find that at this point it is better to power the machine off and then power up again.

3.3.1.2 Access boot list on MCA machine

To access the boot list on MCA machines, perform the following:

1. Turn the key to Service and boot from CD diagnostics or diskette diagnostics. See Section 5.3.3, “Stand-alone diagnostics from CD or diskette” on page 99.
2. Select **Task Selection** from the Diagnostics screen
3. Select the **Display or Change Bootlist entry**.
4. Use the menus to view and change both the Normal and Service mode boot lists. Be aware when the boot list is viewed from CD or diskettes, the hdisk(x) entry may have a comment such as `Not Available Device`. You need not worry about this as it is a by-product of booting from CD. Use the generic entries to construct the boot list to get the machine to boot successfully. It will be possible once AIX is running to reconfigure the boot list using AIX logical device resource names.

Note

Take care when deciding if the boot disk is present or not. Do not decide solely by hdisk numbers. Hdisk numbering when booting from CD is governed by the configuration manager walking the bus. When AIX is running on the machine in question, the numbering is governed by how and when disks were added to the system and ODM entries available at that time. The disk you want may have a totally different number from the one you are expecting.

3.3.2 Codes displayed longer than four minutes

Trying to decide if a LED code is in a hung state is difficult. The generally accepted convention is that a constant display of four minutes duration or longer is liable to indicate a stalled boot process. To be on the safe side if a code is displayed longer than four minutes, then it is normally better to be absolutely sure there is not a problem before you start to turn the machine off or attempt other remedial actions.

If you are working on a 7017-S70 or 7017-S7A, then an error condition is indicated by the attention light being permanently illuminated as well as the eight digit code being displayed. However, the attention light will not be turned on when the S70 or S7A stops on a code beginning with Exxx or 0xxx.

There are a number of reasons why a progress LED will hang after AIX has started to load. The most common cause is that a device that is being configured is defective and so the boot process is unable to proceed. At this point, take note of the LED details and then make a Service Request Number (SRN) by adding a 101- to the code. For example, 570 displayed in the LED will give a SRN 101-570. Go to the Common Service and Diagnostic book for your machine type (either Micro Channel Systems or Multiple Bus Systems), look up the SRN, and it will give you the part number of the part or parts causing the problem.

If the error code is not an AIX system code (300-999 or 0300-0999), the machine is still performing either the BIST or POST phase of the boot. It is easy to decide which phase the machine is in on a MCA machine, because the LED numbering is very specific. On PCI machines, the firmware is less specific and so does not differentiate between BIST and POST. Sometimes, if you look at the error description in the service manual, you can make an educated guess as to if the machine is in the BIST or POST phase of the boot process. If the machine is in the BIST part of the boot process, see Section 3.4, “Minimum configuration” on page 43 for information on how to proceed.

3.3.2.1 Machine hang shortly after power on

The condition here is that the machine is powered on and either almost immediately produces a LED hang condition or the machine is powered on and the LED display cycles through a number of checkpoints but hangs prior to the point that SMS can be accessed or LED codes 252, 243, or 233 are seen on MCA machines. Due to the limited intelligence on the machine at this point, the best method to diagnose the problem is to bring the machine down to minimum configuration. See Section 3.4, “Minimum configuration” on page 43 for details of the minimum configuration process.

3.3.2.2 Machine attempts to boot

Try and determine whether the machine is trying to find a boot image, often signified by alternating 223, 229 on MCA machine. The codes E1F7, E1FB, or E174 may be displayed on PCI CHRP machines. RS/6000 7024-E20, 7024-F30, 7043-140, and 7043-240 may display FCE, FD2, or FDO. If the machine is failing to boot from disk, restart the machine and attempt a boot from a Diagnostic CD, explained in Section 5.3.3, “Stand-alone diagnostics from CD or diskette” on page 99, or an AIX CD at the same AIX version as loaded on the machine you are working on, or a mksysb image, explained in Section 3.5, “Accessing rootvg from bootable media” on page 44. If the machine can boot to diagnostics, look at the diagnostic list and see if all of the disks that should be present on the system are detected. If the disk that should contain the boot image (logical volume hd5) is not seen, you need to investigate further.

Note

Take care when deciding if the boot disk is present or not. Do not decide solely by hdisk numbers. Hdisk numbering when booting from CD is governed by the configuration manager walking the bus. When AIX is running on the machine in question, the numbering is governed by how and when disks were added to the system and ODM entries available at that time. The disk you want may have a totally different number from the one you are expecting.

3.3.3 LED code 269 on SMP machines

The 269 Non Bootable message can be displayed both on the ASCII console and the LED panel of the system unit. This message is displayed when the machine can not find a valid boot image. This situation will often occur after the machine has had work carried out on it that involved splitting the CPU complex from the rest of the machine for a repair or changes to the machine

CPUs or memory. This action will cause the boot list held in the NVRAM to be erased.

3.3.3.1 Recovery

To recover from the `269` message, perform the following actions:

1. Power the machine off.
2. Set the machine into Service mode.
3. Type `sbb` when the `>` prompt is displayed on the ASCII console.
4. Set the following flags to the values shown:
   - Bump Console present set to `enabled`.
   - Auto service IPL set to `disabled`.
   - Fast IPL set to `enabled`
5. Power on the machine.

   You will see all of the tests being run to the CPU and memory. The adapters will be polled and the corresponding LED number will be displayed on the ASCII screen and on the LCD panel. You should see at least one 292 displayed. This indicates that a SCSI adapter has been initialized. You will eventually be presented by a menu, headed MAINTENANCE MENU (Rev. xx).
6. Select **SYSTEM BOOT** (option 6).
7. Select **BOOT FROM SCSI DEVICE** (option 2).

   What you select now depends upon the model of machine. However, what you are attempting to achieve is to get the machine to boot from a particular SCSI device by specifying the SCSI adapter and SCSI ID of the device.
8. Select **CHANGE SLOT** (option 1). Enter the number of the slot that contains the SCSI adapter that the boot disk is connected to.
9. Select **CHANGE SCSI ID** (option 2). Enter the address of the device that you know to contain the boot image. If you do not know the address, pick any address. If you do not pick the correct ID the first time, this process will eventually drop you back to the Maintenance menu when 269 is displayed. At this point, you will be able to repeat the process using different values.
10. Select **BOOT FROM SELECTED DEVICE** (option 5).

    The machine will now attempt to find a boot image on the device you have selected.

If you get LED 269 again, go to step 12.

If the machine boots into diagnostic mode, and the source of the diagnostic mode is the AIX system on the boot disk, you can then use the AIX Prompt selection in the diagnostics Task Selection menu to log on as root. Once you have logged on, you can reconstruct the boot list using the following command:

```
bootlist -m normal hdiskx
```

Where `hdiskx` is the hdisk number of the disk containing the boot logical volume hd5. This will set the Normal mode boot list. Use the following command to set the Service mode boot list:

```
bootlist -m service rmtx cdx hdiskx
```

11. Turn the key to the Normal position and reboot the machine. This is the end of the procedure.

12. You have arrived at this point because the disk that you have selected either did not contain the boot image hd5 or it did contain an image but this image was corrupted in some way. If you are positive that you picked the disk containing hd5, return to step 5 and in the selections pick the address of the CD-ROM device and use the Diagnostic CD or AIX install CD to boot from. Once the machine has booted successfully, continue to step 13.
    If you were not sure you picked the correct disk, go to Step 5 and use the address of another disk. Continue this method until you get the machine to boot.

13. If you get the machine to boot from the Diagnostics CD, test all disks. You can also use the Diagnostics Task Selection to check the boot list. However, since you have used the CD-ROM, you will get a message saying some devices are not available. If you chose to erase the Normal or Service mode boot lists, you will only be able to select the Generic options. These selections cause all devices to be started and the machine searches all of them for a bootable image.

    If you got the machine to boot from an AIX CD, go to Section 3.5, “Accessing rootvg from bootable media” on page 44. Return to this point once you have a root `#` prompt. At this point, locate the disk containing hd5 by using the following command:

    ```
    lspv
    ```

    This displays information detailing the volume group each disk belongs to. For each disk that is part of rootvg, run the following command:

    ```
    lspv -l hdiskX
    ```

Where `hdiskX` is one of the disks in rootvg. Look for hd5 in the output. The disk containing hd5 will be the boot disk.

Use the following command sequence to set the NVRAM boot list and create a new boot image on the boot disk:

```
bootlist -m normal hdiskX
bosboot -d /dev/hdiskX -a
savebase -d /dev/hdiskX
```

The `savebase` command saves the boot customization details to be used on the next boot.

Reboot the machine after turning the key to the *Normal* position.

3.3.4 LED 549 hang

This LED hang is not listed in the system Common Diagnostic Information books, *Diagnostics Information for Micro Channel Bus Systems*, SA38-0532 or *Diagnostics Information for Multiple Bus Systems*, SA38-0509. It occurs when the machine has a system dump to be copied off, but because the graphics adapter has not been configured or cannot be configured at this time, it cannot inform you via the screen.

3.3.4.1 Recovery

Complete the following recovery steps:

1. Power off the machine.
2. Obtain an ASCII terminal or a laptop with a terminal emulator. Set the speed of the screen to 9600 baud, 8 bit, no parity, and 1 stop bit. Set the screen to emulate a vt100 terminal.
3. Connect the ASCII terminal to the S1 serial port on the machine using a RS232 terminal cable.
4. Power on the machine.
5. You will eventually see a message on the screen informing you that a dump is on the paging space. The system is prompting for a device to copy the system dump information.
6. If you have a tape, copy it off and send it to your normal AIX support function. If you do not have a tape drive on the machine, you will have to discard this dump by continuing the boot process.
7. When you get the login prompt, use SMIT to configure another dump device on your machine. See Chapter 4, “System dumps” on page 57 for more information.

3.4 Minimum configuration

Removing components from a machine to bring it down to minimum configuration is the best way to diagnose problems when there is not enough intelligence on the machine to enable the diagnostics or firmware to accurately indicate the failing part. The principle behind this method is that every machine has a minimum number of parts that will display a predetermined stable LED code. The code that is displayed is a positive identification that the parts left in the machine are working correctly and have been tested as good by the firmware.

The Installation and Service Guide for each machine will have a specific MAP for this procedure. Every one of the guides assign the same number to the MAP title. The MAP you need to find is MAP1540 Minimum Configuration MAP. If you are in a situation where you do not have this MAP available, the following directions may be used, but be aware that since you do not have the documentation, you may make incorrect assumptions. If you have the Installation and Service Guide for your machine, it is *essential* that you refer to MAP1540 to ensure you correctly follow the minimum configuration procedure.

Note

Power cables must be removed from the machine prior to any removal or insertion of CPUs, memory, or adapters. Components within the machine are susceptible to static electricity damage, so take ESD precautions. Failure to observe these rules will result in severe damage to the machine. This procedure should only be attempted by authorized service personnel.

Before proceeding with the minimum configuration procedure, you should check any maintenance contract covering the machine. In many cases, opening of the machine by the customer may void the service contract.

A typical minimum configuration setup consists of:

- Minimum supported number of CPUs. Check that you leave the remaining CPUs in the correct slot that the system expects.
- Minimum amount of memory. Check the memory configuration rules for the machine you are working on. It varies greatly between machine types. In some MCA machines, the minimum configuration for the MAP1540 is just the memory base card or card pair with no SIMMs or DIMMs plugged into them. The 7025-F50 and 7026-H50 minimum is one memory base card with one pair of DIMMs in the first bank position. See MAP1540 for other machine types.

- System planar with no adapters in place except the graphics adapter, assuming the graphics adapter is used for the console screen.
- No cables attached externally except keyboard, mouse, graphics screen, (assuming graphics screen is used for console) or tty terminal cable.
- Base power supply.
- All internal SCSI cables, diskette cables, and DC power cables to internal media including disks and active backplanes are disconnected.

When you plug in the power and press the Power button, you should get the check code as described in the minimum configuration MAP. If you do not have the service guide for the machine, then the only test that can be applied is whether the code displayed is a different error code to the one you started with.

If you do not get the expected code from the MAP1540, or you get the same code as you started with, you now have to order and change singularly each of the components in the machine at minimum configuration until you get the expected code. If you do not have the MAP1540, your best indication that you may have found the failing part is when the code displayed changes from the one you started with.

If at minimum configuration, you get the code indicated in MAP1540, one of the components that you removed is defective. Add each component singularly until you reproduce the problem. The last part you added is probably the defective item and should be replaced. Quite often you will find that you will refit all of the original parts and not reproduce the fault again. This can be due to either bad seating of a component or a bad connection.

3.5 Accessing rootvg from bootable media

If you used either an AIX CD or a mksysb image to boot the system, perform the following actions:

- Select option 3 **Start Maintenance Mode for System Recovery**.
- Select option 1 **Access a Root Volume group**.
- Select **0** to **Continue**.

You will now be presented with a list of volume group information. Look through the list of disks configured and decide from what you know about the system which is the root volume group. Enter the selection corresponding to the volume group you wish to select. The next screen will list all of the logical volumes within that volume group. If hd5 is listed, carry on with the next step.

However, if hd5 is not listed in the logical volumes of the selected volume group, select option **99** to return to the previous menu and select another listed volume group. If you have correctly chosen rootvg, select option 1 **Access this Volume group and start a shell**. You are now attempting to find all disks in rootvg and mount them on to the RAM image you have just loaded. Look at the messages produced carefully as they are your best clue as to the sort of problem you are dealing with.

If you finally get a message indicating that the operation failed, you need to run diagnostics from CD to test the disk and also certify it. From the tests of the diagnostics, you can decide if the disk is defective or it has for some reason been so badly corrupted that AIX is not capable of being started by any means.

However, if the messages produced indicate that rootvg has been varied on successfully, but the system fails to mount the file systems, refer to the procedure documented in Section 3.6.1, “LED 551, 555, or 557 halt” on page 45 to correct the file systems.

3.6 LED 551, 552, 554, 555, 556, and 557 halts

If the machine stops on any of these LEDs, the first phase of the boot that tests the hardware has completed successfully. It is generally accepted, therefore, that these LED halts are related to problems with AIX. In a small minority of cases though, the final resolution will be the replacement of a piece of hardware. To assist in your diagnosis, the following sections contain information that fix the most common causes of the halt.

3.6.1 LED 551, 555, or 557 halt

Recovery from LED 551, 555, or 557 in AIX V4. This section describes the causes for LED 551, 555, or 557 during IPL on a RS/6000. Also outlined is a recovery procedure.

The known causes of an LED 551, 555, or 557 during IPL on an RS/6000 are:

- A corrupted file system
- A corrupted Journaled-file-system (JFS) log device
- A failing fsck (file-system check) caused by a bad file system helper
- A bad disk in the machine that is a member of the rootvg

To diagnose and fix the problem, boot from bootable media, run `logform` on /dev/hd8, and run `fsck` to fix any file systems that may be corrupted.

Note

Do not use this procedure if the system is a /usr client, diskless client, or dataless client.

1. Turn the key to the Service position (MCA machines) then power on. On PCI machines, power on and press **F5** or **5** when the Icon screen appears. This will start the Service mode boot from CD or mksysb.
2. With bootable media of the same version and level as the system, boot the system. The bootable media can be any one of the following:
   - Bootable CD-ROM
   - NON_AUTOINSTALL mksysb
   - Bootable Install Tape
3. Follow the screen prompts to the following menu:

   ```
   Welcome to Base Operating System
   Installation and Maintenance
   ```

   Choose option 3, **Start Maintenance Mode for System Recovery**.

   The next screen has the maintenance menu.

   Choose option 1, **Access a Root Volume Group**.

   The next screen displays a warning that indicates you will not be able to return to the Base OS menu without rebooting.

   Choose **0** to continue.

   The next screen displays information about all volume groups on the system.

   Select the root volume group by number.

   Choose option 2, **Access this volume group and start a shell before mounting the file systems**.

   If you get errors indicating that a physical volume is missing from the rootvg, run diagnostics on the physical volumes to find out if you have a bad disk. Do not continue with the rest of the steps in this process.

   If you get other errors from the above option, do not continue with the rest of the steps in this process. Correct the problem causing the error. If you need assistance correcting the problem, contact your AIX support function.
4. Format the default jfslog for the rootvg JFS file systems:

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

Answer yes when asked if you want to destroy the log.

5. Next, run the following commands to check and repair file systems. The `-y` option gives the `fsck` command permission to repair file systems when necessary.

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

   Type `exit`. The file systems will automatically mount after you type exit.

6. If you are running the Andrew File System (AFS), use the following commands to find out whether you have more than one version of the v3fshelper file (otherwise, skip to step 8):

   ```
   cd /sbin/helpers
   ls -l v3fshelper*
   ```

   If you have only one version of the v3fshelper file (for example, v3fshelper), proceed to step 8.

7. If there is a version of v3fshelper marked as original (for example, v3fshelper.orig), run the following commands:

   ```
   copy v3fshelper v3fshelper.afs
   copy v3fshelper.orig v3fshelper
   ```

8. Determine which disk is the boot disk with the `lslv` command. The boot disk will be shown in the PV1 column of the lslv output.

   ```
   lslv -m hd5
   ```

   Note

   Do not proceed further if the system is a /usr client, diskless client, or dataless client.

9. Recreate the boot image and alter the boot list (hdisk# is the boot disk determined in step 8):

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

10. If you copied the v3fshelper file in step 7, copy AFS file-system helper back to v3fshelper:

    ```
    copy v3fshelper.afs v3fshelper
    ```

11. With the key in Normal position (MCA machine), run:

    ```
    shutdown -Fr
    ```

If you followed all of the preceding steps, and the system still stops at LED 551, 555, or 557 during a reboot in Normal mode, you may want to pursue further system recovery assistance from your AIX support function.

For reasons of time and the integrity of your AIX operating system, the best alternative at this point may be to reinstall AIX.

3.6.2 LED 552, 554, or 556 halt

This section discusses the known causes of LED 552, 554, and 556, including a procedure for recovery from these errors.

An LED code of 552, 554, or 556 during a standard disk based boot indicates a failure occurred during the varyon of the rootvg volume group.

The known causes of an LED 552, 554, or 556 are:

- A corrupted file system.
- A corrupted 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 corrupted 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.

To diagnose and fix the problem, you will need to boot from BOOTABLE MEDIA and run the `fsck` command on each file system. If the file system check fails, you may need to perform other steps.

Note

Do not use this procedure if the system is a /usr client, diskless client, or dataless client.

To recover form an LED 552, 554, or 557, complete the following steps:

1. Turn the key to the Service position (MCA machines) then power on. On PCI machines, power on, press **F5** or **5** when the Icon screen appears. This will start the Service mode boot from CD or mksysb.
2. With bootable media of the same version and level as the system, boot the system. The bootable media can be any one of the following:
   - Bootable CD-ROM

- mksysb
- Bootable Install Tape

Follow the prompts to the Welcome to Base OS menu.

3. Choose **Start Maintenance Mode for System Recovery** (Option 3). The next screen displays prompts for the Maintenance menu.

   Choose **Access a Root Volume Group** (Option 1).

   At this stage, the console will display information about rootvg and a menu with two options.

   Choose **Access this volume group and start a shell before mounting the filesystems** (Option 2).

   If you get errors from the preceding option, do not continue with the rest of this procedure. Correct the problem causing the error. If you need assistance correcting the problem causing the error, contact your AIX support function.

4. Run the following commands to check and repair file systems. The `-y` option gives the `fsck` command permission to repair file systems when necessary.

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

   If any of the following conditions occur, proceed accordingly.

   If `fsck` indicates that block 8 could not be read, the file system is probably unrecoverable. See step 5 for information on unrecoverable file systems.

5. If `fsck` indicates that block 8 could be read, but one of the following errors is given:

   - `fsck: Not an AIXV3 file system.`
   - `fsck: Not a recognized file system type.`

   Then go to step 6.

   If `fsck` indicates that a file system has an unknown log record type, or if `fsck` fails in the logredo process, then go to step 7.

   If the file system checks were successful, skip to step 9.

   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 hd2 and hd3 can be recreated but hd4 cannot be recreated. If hd4 is

unrecoverable, you must reinstall AIX. For assistance with unrecoverable file systems, contact your AIX support function.

Do *not* follow the rest of the steps in this procedure.

Attempt to repair the file system with this command:

```
fsck -p /dev/hd#
```

Replace `hd#` with the appropriate file system logical volume name.

Now skip to step 8.

6. A corruption of the JFS log logical volume has been detected. Use the `logform` command to reformat it:

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

   Answer yes when asked if you want to destroy the log.

7. Repeat step 4 for all file systems that did not successfully complete fsck the first time.

   If step 4 fails a second time, the file system is almost always unrecoverable. See step 5 for an explanation of the options at this point. In most cases, step 4 will be successful. If step 4 is successful, continue to step 9.

8. With the key in the Normal position, run the following commands to reboot the system:

   ```
   exit
   sync;sync;sync
   shutdown -Fr
   ```

   As you reboot in Normal mode, notice how many times LED 551 appears. If LED 551 appears twice, fsck is probably failing because of a bad fshelper file. If this is the case, and you are running AFS, see step 12.

   The majority of instances of LED 552, 554, and 556 will be resolved at this point. If you still have an LED 552, 554, or 556, repeat step 1 through step 3.

9. Run the following commands that remove much of the system's configuration and save it to a backup directory:

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

10. Determine which disk is the boot disk with the `lslv` command. The boot disk will be shown in the PV1 column of the lslv output.

    ```
    lslv -m hd5
    ```

11. Save the clean ODM database to the boot logical volume. (`#` is the number of the fixed disk, determined with the previous command.)

    ```
    savebase -d /dev/hdisk#
    ```

    If you are running AFS, go to step 12;

    otherwise, go to step 13.

12. If you are running the Andrew File System (AFS), use the following commands to find out whether you have more than one version of the v3fshelper file:

    ```
    cd /sbin/helpers
    ls -l v3fshelper*
    ```

    If you have only one version of the v3fshelper file (for example, v3fshelper), proceed to step 13.

    If there is a version of v3fshelper marked as original (for example, v3fshelper.orig), run the following commands:

    ```
    cp v3fshelper v3fshelper.afs
    cp v3fshelper.orig v3fshelper
    ```

    Note

    Do not proceed further if the system is a /usr client, diskless client, or dataless client.

    Recreate the boot image (`hdisk#` is the fixed disk determined in step 11):

    ```
    bosboot -a -d /dev/hdisk#
    ```

13. If you copied files in step 12, copy the AFS file-system helper back to v3fshelper:

    ```
    cp v3fshelper.afs v3fshelper
    ```

    Turn the key to Normal position and run:

    ```
    shutdown -Fr
    ```

If you followed all of the preceding steps, and the system still stops at an LED 552, 554, or 556 during a reboot in Normal mode, you may want to pursue further system recovery assistance from your AIX support function.

For reasons of time and the integrity of your AIX operating system, the best alternative at this point may be to reinstall AIX.

3.6.3 LED 553 halt

This section sets out a procedure to recover from an LED 553 and applies to AIX V4.

An LED 553 occurs during IPL on a RS/6000 if the system cannot read or run the /etc/inittab file.

To recover from an LED 553, check /dev/hd3 and /dev/hd4 for space problems and erase files if necessary. Check the /etc/inittab file for corruption and fix it if necessary. If the inittab file was not corrupted, you will need to check the shell profiles, the /bin/bsh file, and some other files:

1. Turn the key to the Service position (MCA machines) then power on. On PCI machines, power on and press **F5** or **5** when the Icon screen appears. This will start the Service mode boot from CD or mksysb.
2. With bootable media of the same version and level as the system, boot the system. The bootable media can be any one of the following:
   - Bootable CD-ROM
   - mksysb
   - Bootable Install Tape

   Follow the prompts to the Welcome to Base OS menu.
3. Choose **Start Maintenance Mode for System Recovery** (Option 3). The next screen contains prompts for the Maintenance menu.
   - Choose **Access a Root Volume Group** (Option 1). At this stage, the console displays information about rootvg and a menu with two options.
   - Choose **Access this volume group and start a shell** (Option 1).

   If you get errors from the preceding option, do not continue with the rest of this procedure. Correct the problem causing the error. If you need assistance correcting the problem causing the error, contact your AIX support function.
4. Use the `df` command to check for free space in /dev/hd3 and /dev/hd4:

   ```
   df /dev/hd3
   df /dev/hd4
   ```

   If `df` shows that either file system is out of space, erase some files from that file system. Three files you may want to erase are /smit.log, /smit.script, and /.sh_history.
5. Check the /etc/inittab file for corruption. It may be empty or missing, or it may have an incorrect entry. For comparison, see the sample inittab file

shown in Figure 9 on page 55. Set your terminal type in preparation for editing the file by setting the TERM environment variable. For example:

```
TERM=xxx
export TERM
```

Where xxx stands for a terminal type, such as lft, ibm3151, or vt100.

Now, use an editor to create the /etc/inittab file. It may be possible for you to correct a corrupt inittab. If the file is missing, or it is not possible to repair the corruption, it is simpler to create a new file. If your /etc/inittab file was corrupt, and you recreated it, you may not need to perform any of the following steps.

6. Use the following command to check for any modifications or problems with permission:

   ```
   ls -al /.profile /etc/enviroment /etc/profile
   ```

   Example output:

   ```
   -rw-r--r--  1 root  system  158 Dec 14 1993  /.profile
   -rw-rw-r--  1 root  system 1389 Oct 26 1993  /etc/environment
   -rw-r--r--  1 root  system 1214 Jan 22 1993  /etc/profile
   ```

7. /etc/profile or .profile may contain a command that is valid only in the Korn shell. Change the command to something that is also valid in the Bourne shell. For example, change the following:

   ```
   export PATH=/bin:/usr/bin/:/etc:/usr/ucb:.
   ```

   To the following:

   ```
   PATH=/bin:/usr/bin/:/etc:/usr/ucb:.
   export PATH
   ```

8. /etc/environment is a special case. The only commands it may contain are simple variable assignments, such as statements of the form var=value. Check this file with an editor to verify the format.

9. Check for missing or moved files with the following command:

   ```
   ls -al /bin /bin/bsh /bin/sh /lib /u /unix
   ```

   Example output:

   ```
   lrwxrwxrwx 1 root sys      8 Aug 5 1994 /bin -> /usr/bin
   -r-xr-xr-x 3 bin  bin 256224 Jun 4 1993 /bin/bsh
   -r-xr-xr-x 3 bin  bin 256224 Jun 4 1993 /bin/sh
   lrwxrwxrwx 1 root sys      8 Aug 5 1994 /lib -> /usr/lib
   lrwxrwxrwx 1 root sys      5 Aug 5 1994 /u -> /home
   lrwxrwxrwx 1 root sys     18 Aug 5 1994 /unix -> /usr/lib/boot/unix
   ```

If any of these files are missing, the problem may be a missing symbolic link. Use the commands from the following list that correspond to the missing links:

```
ln -s /usr/bin /bin
ln -s /usr/lib/boot/unix /unix
ln -s /usr/lib /lib
ln -s /home /u
```

10. Use the following command to make sure that fsck and rc.boot are not missing or corrupt:

    ```
    ls -l /etc/fsck /sbin/rc.boot
    ```

    Example output:

    ```
    lrwxrwxrwx 1 root system    14 Aug  5 1994 /etc/fsck -> /usr/sbin/fsck
    -rwxrwxr-- 1 root system 33760 Aug 30 1993 /sbin/rc.boot
    ```

11. Make sure the /etc/inittab file is the AIX V4. For that version, the line that begins with `brc` is:

    ```
    brc::sysinit:/sbin/rc.boot 3 >/dev/console 2>&1
    ```

    See Figure 9 on page 55 for an example.

12. If you have not found any obvious problems, try substituting `ksh` for `bsh` with the following commands (the first command saves your `bsh` before you copy over it):

    ```
    cp /bin/bsh /bin/bsh.orig
    cp /bin/ksh /bin/bsh
    ```

    If you can then reboot successfully, you know that one of the profiles was causing problems for bsh. Check the profiles again by running the following:

    ```
    /bin/bsh.orig /.profile
    /bin/bsh.orig /etc/profile
    /bin/bsh.orig /etc/environment
    ```

    If you receive errors with any of these commands, you know there is a command in that profile that bsh cannot handle.

If you followed all of the preceding steps, and the system still stops at an LED 553 during a reboot in Normal mode, you may want to pursue further system recovery assistance from your AIX support function.

For reasons of time and the integrity of your AIX operating system, the best alternative at this point may be to reinstall AIX.

```
:@(#)49 1.28.2.7 src/bos/etc/inittab,cmdoper,bos411,
: 9430C411a 7/26/94 16.27.45
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
rc:2:wait:/etc/rc > alog -tboot > /dev/console 2>&1# Multi-User checks
fbcheck:2:wait:/usr/lib/dwm/fbcheck > alog -tboot >/dev/console 2>&1
srcmstr:2:respawn:/etc/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:/etc/cron
cons:0123456789:respawn:/etc/getty /dev/console
piobe:2:wait:/usr/lib/lpd/pio/etc/pioinit > /dev/null 2>&1 # pb cleanup
qdaemon:2:wait:/bin/startsrc -sqdaemon
writesrv:2:wait:/bin/startsrc -swritesrv
uprintfd:2:respawn:/usr/sbin/uprintfd
dt:2:wait:/etc/rc.dt
```

Figure 9. Sample /etc/inittab file

3.7 No login prompt

If the system boots, and you see output on the console but do not get a login prompt, then there are a number of things that you need to investigate.

If the console is a graphics console, refer to Section 8.6.2, “Login screen does not appear” on page 222.

If the console device is an ASCII terminal, perform the following steps:

- Try to remotely login to the machine. If you succeed, use the `ps` command to check to see if there is a getty process running against /dev/console. For example:

  ```
  # ps ax|grep console
   10846      0 A     0:00 /usr/sbin/getty /dev/console
  ```

  If the process is running, ensure that the console output is directed to the serial port the terminal is connected to. This can be done with the `lscons` command. For example:

  ```
  # lscons
  /dev/tty0
  ```

  Use the `lsdev` command to determine the serial port being used for the console terminal. For example:

```
# lsdev -C -l tty0
tty0 Available 01-S1-00-00 Asynchronous Terminal
```

Ensure that the terminal is connected to the serial port listed by the `lsdev` command.

- If you cannot login to the machine, then there are two possible problems. The first is that the machine has hung at the end of the boot process, which is why there is no login prompt. The second is that there is no inittab entry to start a getty process on the console. To check the second situation, boot from AIX CD media and enter the maintenance shell as described in 3.6.3, “LED 553 halt” on page 52. Once you have mounted the file systems, look at the /etc/inittab file. Ensure that there is an entry in the file as follows:

  ```
  cons:0123456789:respawn:/etc/getty /dev/console
  ```

  Once you have done this, try rebooting the machine.

Chapter 4. System dumps

This chapter describes system dumps and how to handle them.

4.1 Introduction

Your system generates a system dump when a severe error occurs, such as a system halt with an 888 number flashing. It can also be initiated by the system administrator when the system has hung.

The system dump is a copy of the contents of all or part of the physical memory of your system. It is obtained from memory locations used by kernel components. Actually, a system dump is a snapshot of the operating system state at the time of the crash or manually initiated dump.

The software tool `crash` command is used to examine a system dump. Using `crash`, you can examine kernel data structures. These data structures tell what state the system was in when it crashed or became hung. Please refer to Section 4.8, “The crash command” on page 74 for more information.

4.2 Saving a system dump when system is booting

Since AIX 4.1, the default dump device is /dev/hd6, which is also the default paging device. If you have not added a dedicated default dump device (for example, /dev/hd7), then on reboot, the system will attempt to copy the dump image from /dev/hd6 to a file (vmcore.X) in a directory in rootvg (the default is /var/adm/ras). This is because the /dev/hd6 device needs to be used as paging space when the AIX system starts. If the copy fails, usually because there is not enough space, it will prompt you to copy off the dump to a tape device or to diskettes.

Note

If you do not copy the dump out of the paging space, it will be overwritten, and the dump will be lost.

The screen shown in Figure 10 on page 58 will be displayed on reboot if the default dump device is /dev/hd6 and there was not enough space to copy the dump to a directory.

```
Copy a System Dump to Removable Media

        The system dump is 5604352 bytes and will be copied from /dev/hd6
        to media inserted into the device from the list below.

        Please make sure that you have sufficient blank, formatted
        media before you continue.

        Step One:   Insert blank media into the chosen device.
        Step Two:   Type the number for that device and press Enter.

                Device Type             Path Name

     >>>  1  tape/scsi/8mm              /dev/rmt0
          2  diskette/siofd/fd          /dev/fd0

        88  Help?
        99  Exit -- Waring, the dump will be lost
```

Figure 10. Save a system dump screen

The dump and the /unix file will be copied to tape when you select option **1**.

If this step fails for some reason, such as there is no system console, or the system console has been redirected to a file, the system will hang with LED 549.

4.2.1 LED 549

If the system hangs at LED 549 when booting, you need to boot from the install media or mksysb tape. Next, go into maintenance mode and then use the `sysdumpdev` command to either ignore the dump if the copy fails or specify a different location where there is enough space to copy the dump.

If you got a login prompt without seeing the screen shown in Figure 10, the dump was successfully saved to the selected directory or device. You will be able to collect this dump file with the `snap` command. Refer to Section 4.6, “Collecting the dump and related information” on page 70.

4.3 Preparing for the dump

The system needs to be initially configured to capture a dump successfully. This can be done by following steps detailed in this section.

4.3.1 Estimate the size of the dump

First of all, you need to determine how large the dump device for your machine needs to be. This can be done through smit by following the fast path:

```
# smit dump_estimate
```

Or, run the following command:

```
# sysdumpdev -e
0453-041 Estimated dump size in bytes: 57671680
```

This value can change based on the activity of the system. It is best to run this command when the machine is under its heaviest workload. You should make the dump device slightly larger than the value reported by the `sysdumpdev` command in order to handle a system dump during peak system activity.

4.3.2 Selecting the dump device

When you install the operating system, the dump device is automatically configured for you. By default, the primary dump device is /dev/hd6, which is also the default paging logical volume. The secondary dump device is configured by default to be /dev/sysdumpnull. You can also select these dump devices to be disk, tape, or remote.

To view information about the current dump devices, use the following command：

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

In this example, the primary dump device is the logical volume `hd6`, and the secondary dump device is `/dev/sysdumpnull`, meaning that there is no secondary.

The original usage of the secondary dump device was limited to when the user specifically requested a dump to the secondary. In practice, secondary dump devices were rarely used. However, the use of the secondary dump device changed in AIX 4.2.1 with the addition of dump failover.

Note

Dump failover

The secondary dump device is now used to back up the primary dump device. If an error occurs during a system dump to the primary dump device, the system attempts to dump to the secondary device (if it is defined). If that fails as well, AIX will refer to whichever device accepted more data as the real dump. This is useful if your system is unable to successfully perform a system dump to the primary dump device.

Bear the following rules in mind when selecting dump devices:

- Do not use a mirrored logical volume as the active dump device. System dump error messages will not be displayed, and any subsequent dumps to a mirrored logical volume will fail.
- This limitation has been removed in AIX Version 4.3.3 and subsequent releases.
- Do not use a diskette drive as your dump device.
- The primary paging device hd6 is the only paging device that should be used as the primary dump device. Using any other paging device as the primary dump device results in that paging space not being used by the AIX virtual memory manager, effectively deactivating that paging space.
- AIX Version 4.2.1 or later supports using any paging device in the root volume group (rootvg) as the secondary dump device.

Note

Dumping to a mirrored logical volume

AIX releases prior to Version 4.3.3 do not support dumping to a mirrored logical volume. This is because the dump bypasses the LVM mechanism and writes directly to one copy of the logical volume. In other words, only one of the mirrors will contain the dump; the other mirrors will contain whatever the logical volume had before the dump started. When crash tries to read the dump, it uses the normal LVM read mechanism, so it can get data from any of the mirrors, only one of which actually contains the dump. In other words, crash sees good dump data mixed with garbage data and will not read the dump. This limitation is removed with AIX Version 4.3.3, which does support using a mirrored logical volume as the primary dump device.

To specify the primary dump device, use the command:

```
# sysdumpdev -P -p /dev/hd7
primary              /dev/hd7
secondary            /dev/sysdumpnull
copy directory       /var/adm/ras
forced copy flag     TRUE
always allow dump    FALSE
```

To specify the secondary dump device, use the command:

```
# sysdumpdev -P -s /dev/hd7
primary              /dev/hd6
secondary            /dev/hd7
copy directory       /var/adm/ras
forced copy flag     TRUE
always allow dump    FALSE
```

4.3.3 Create a dump device

If you want to create a standard dump logical volume, perform the following steps.

1. Estimate the size of a dump on the system. Refer to Section 4.3.1, “Estimate the size of the dump” on page 59.

   ```
   # sysdumpdev -e
   0453-041 Estimated dump size in bytes: 54525952
   ```

   Remember to make the dump device slightly larger than that reported by the `sysdumpdev` command to handle dumps during peak workloads.

2. Calculate the required number of PPs for the dump device.

   Get the PP size of the volume group by using the `lsvg` command:

   ```
   # lsvg rootvg
   VOLUME GROUP:   rootvg                  VG IDENTIFIER:  00017d37a4155bbc
   VG STATE:       active                  PP SIZE:        8 megabyte(s)
   VG PERMISSION:  read/write              TOTAL PPs:      537 (4296 megabytes)
   MAX LVs:        256                     FREE PPs:       218 (1744 megabytes)
   LVs:            10                      USED PPs:       319 (2552 megabytes)
   OPEN LVs:       8                       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
   ```

   Determine the necessary number of PPs by dividing the estimated size of the dump by the PP size. For example:

   `54525952 / 8 megabytes = 6.5` (required number is 7)

3. Create a logical volume of the required size, for example:

   ```
   # mklv -y hd7 -t sysdump rootvg 7
   ```

hd7

4.3.4 Change the size of dump device

If the system already has a dump device, make sure the estimated dump size can fit into this dump device. The `lslv` command displays the size of the logical volume. For example:

```
# lslv hd6
LOGICAL VOLUME:     hd6                    VOLUME GROUP:   rootvg
LV IDENTIFIER:      00017d37a4155bbc.2     PERMISSION:     read/write
VG STATE:           active/complete        LV STATE:       opened/syncd
TYPE:               paging                 WRITE VERIFY:   off
MAX LPs:            512                    PP SIZE:        8 megabyte(s)
COPIES:             1                      SCHED POLICY:   parallel
LPs:                256                    PPs:            256
STALE PPs:          0                      BB POLICY:      non-relocatable
INTER-POLICY:       minimum                RELOCATABLE:    yes
INTRA-POLICY:       middle                 UPPER BOUND:    32
MOUNT POINT:        N/A                    LABEL:          None
MIRROR WRITE CONSISTENCY: off
EACH LP COPY ON A SEPARATE PV ?: yes
```

Note the values for LPs and PP SIZE. Multiply these two values together to get the size of the dump device in megabytes.

If the dump device is a standard dump logical volume, such as `hd7`, then the command to use to increase its size is `extendlv`. For example:

```
# extendlv hd7 1 hdisk
```

If it is the primary paging space `hd6`, then use the `chps` command. For example:

```
# chps -s'1' hd6
```

If the dump device is the paging space, ensure the forced copy flag value, as shown by the `sysdumpdev` command, is set to `true` and the size of copy directory is enough. The default copy directory is /var/adm/ras. You can check the size of copy directory and modify it by running the following commands:

```
# df -k /var
Filesystem    1024-blocks      Free %Used    Iused %Iused Mounted on
/dev/hd9var          8192      4776   42%      102     5% /var
# chfs -asize=+200000 /var
File System size changed to 229376
```

4.3.5 Optional setup

This section details other AIX and machine settings that can have an impact on the ability to access a system dump.

- Autorestart

A useful system attribute is autorestart. If autorestart is true, the system will automatically reboot after a crash. This is useful if the machine is physically distant or often unattended.

To list these value, use: `lsattr -Dl sys0`

To set `autorestart` to `true`, use SMIT by following the fast path:

```
smit chgsys
```

Or use the command:

```
# chdev -l sys0 -a autorestart=true
sys0 changed
```

- Always allow system dump

  If this item is set to `true`, the Reset button or key sequences start a system dump, even when the key mode switch is in the Normal position or when no key mode switch is present. You can change this by `smit dump_allow`, or with the `sysdumpdev` command.

- Surveillance timeout interval

  Most of the newer server machines, such as 7017-S7A and 7025-F50, have a service processor, that has a surveillance function of hardware and software. Sometimes, you fail to get a successful AIX dump due to the surveillance timeout interfering with the dump process. This is due to an insufficient time setting of the surveillance timeout interval value. If your system has a service processor, make sure the value of the surveillance timeout is sufficient. Generally, more than five minutes is suitable. To view this information, use the `diag` command, and select the **Task Selection** menu then select the **Configure Surveillance Policy** menu. You will see the information in a menu similar to that shown in Figure 11 on page 64.

```
CONFIGURE SURVEILLANCE POLICY                                        802583

The following system configuration parameters are available for
Surveillance Policy. Any data in brackets [] may be changed
or added at this time.

When finished, use 'Commit' to accept the data.


  Surveillance                                                [off]       +
  Surveillance Time Interval, in minutes                      [5]          #
  Surveillance Delay, in minutes                              [10]         #
  Changes are to take effect immediately                      [yes]       +



F1=Help            F2=Refresh          F3=Cancel          F4=List
F5=Reset           F7=Commit           F10=Exit
```

Figure 11. Surveillance policy service aid

There is no need to change the surveillance timeout value on systems running AIX Version 4.3.3, since the value is automatically changed to 60 minutes at the start of the dump process.

4.4 The sysdumpdev command

The `sysdumpdev` command changes the primary or secondary dump device designation in a system that is running. It can also be used to manage dump devices. For example:

- `sysdumpdev -d` <Directory >

  Sets the directory to copy the dump into at system boot, but if copy fails, it will not prompt the user at boot time and dump is lost. The forced copy flag is set to `false`.

- `sysdumpdev -D` <Directory >

  Same as above, but if copy fails, it will prompt the user during boot to copy the dump to external media. The forced copy flag is set to `true`.

- `sysdumpdev -l`

  Lists current dump devices, copy directory, and state of forced copy flag.

- `sysdumpdev -L`

  Lists information about the most recent system dump.

- `sysdumpdev -e`

Estimates the size of the dump in bytes.

- `sysdumpdev -P -p` <device>

  Defines the device as the permanent, primary dump device.

- `sysdumpdev -P -s` <device>

  Defines the device as the permanent, secondary dump device.

- `sysdumpdev -K`

  Allows a system without a key mode switch to use the Reset button or dump key sequence to force a dump.

Other options of the `sysdumpdev` command are:

-r Frees up space used by a remote dump file on a server, Host. Usage: sysdumpdev -r Host:Path flag, where Path is the location of the dump file on the host.

-z Determines if a new system dump is present and, if yes, writes its size and the name of the dump device to stdout.

-q Suppresses messages to standard out. Ignored if used with -l, -L, or -r flags.

-k Requires the key mode switch to be in Service mode before a system dump can be initiated with the dump key sequence or the Reset button. This is the default behavior.

-K Does not require the key mode switch to be in Service mode before a system dump can be initiated with the reset key sequence or the Reset button. This is non-default behavior and is useful on systems that do not have a key mode switch.

4.5 Checking the dump status

Checking that the system dump is valid and readable before submitting it to IBM for analysis saves time in case the dump is not valid.

4.5.1 Get the last dump information

If you run the command `sysdumpdev -L`, it will display statistical information about the most recent system dump. This includes date and time of last dump, number of bytes written, and completion status. For example:

```
# sysdumpdev -L
0453-039

Device name:         /dev/hd6
```

```
Major device number: 10
Minor device number: 2
Size:                68197888 bytes
Date/Time:           Fri Mar 12 14:43:52 CST 1999
Dump status:         0
dump completed successfully
0481-195 Failed to copy the dump from /dev/hd6 to /var/adm/ras.
0481-198 Allowed the customer to copy the dump to external media.
```

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

4.5.2 Dump status codes

Once you see a flashing 888 in the LEDs, the system has crashed. You may see a 0c9 for a short time, indicating a system dump is in progress. When the dump is completed, the dump status code will change to 0c0 if the system was able to dump successfully.

Table 5 shows the possible dump status codes along with brief descriptions. Use them to check the status and result of your dump.

Depending on the level of AIX, some of these codes may not be available on your system.

Table 5. Dump status codes

LED code	sysdumpdev status	Description
0c0	0	Dump successful.
0c1	-4	I/O error during dump. This code was added in AIX 4.1.5 and AIX 4.2.1.
0c4	-2	Dump device is too small. Partial dump taken. This code was changed at AIX 4.1.5 and AIX 4.2.1. It is now more specific. It used to include other types of internal errors.
0c5	-3	Internal dump error. This code changed in AIX 4.1.5 and AIX 4.2.1. It now only shows when the dump facility itself fails. This does not include the failure of dump component routines.
0c8	-1	No dump device defined.
0c2	N/A	User-initiated dump in progress. This code was changed in AIX 4.1.5 and AIX 4.2.1 In AIX 4.2.1 and above, this code will appear regardless of whether the dump is going to the primary or secondary dump device.

LED code	sysdumpdev status	Description
0c6	N/A	User-initiated dump to secondary device in progress. Note that AIX 4.2.1 and above support dump failover, so this code is obsolete.
0c7	N/A	Dump waiting for acknowledgment from NFS server.
0c9	N/A	System-initiated dump in progress.
0cc	N/A	Switched to secondary dump device. Indicates dump has performed failover to secondary device. Supported on AIX 4.2.1 and above.

4.5.3 Error log

If you lost the dump or did not save it during system boot, the error log can help determine the nature of problem that caused the dump. Starting with AIX 4.1.3, the error log may include a symptom string as a SYSDUMP_SYMP error if the dump was readable by AIX during the next reboot after the crash. The error log entry also includes a stack traceback.

To check the error log, use the `errpt` command. Figure 12 on page 68 shows a sample SYSDUMP_SYMP error log entry.

```
LABEL:          SYSDUMP_SYMP
IDENTIFIER:     3573A829

Date/Time:       Tue May  4 19:17:15
Sequence Number: 205
Machine Id:      000126774C00
Node Id:         sp5i
Class:           S
Type:            UNKN
Resource Name:   CMDCRASH

Description
SYSTEM DUMP

Probable Causes
UNEXPECTED SYSTEM HALT

User Causes
SYSTEM DUMP REQUESTED BY USER

        Recommended Actions
        PERFORM PROBLEM DETERMINATION PROCEDURES

Failure Causes
UNEXPECTED SYSTEM HALT

        Recommended Actions
        PERFORM PROBLEM DETERMINATION PROCEDURES

Detail Data
DUMP STATUS
LED:300
csa:2ff3b400
soo_select 2c8
soo_select 198
selpoll 120
select 4f0
sys_call_ret 0

Symptom Data
REPORTABLE
1
INTERNAL ERROR
1
SYMPTOM CODE
PIDS/5765C3403 LVLS/430 PCSS/SPI1 MS/300 FLDS/soo_selec VALU/90150008 FLDS/selpoll
VALU/120
```

Figure 12. Sample dump error log entry

In this example, you can see the dump status and some stack traceback information in the `Detail Data` section of the entry. The `User Causes` section always says the dump was requested by the user even if the dump is system initiated.

4.5.4 Verifying the dump

To check that the dump is readable, start the `crash` command 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 any message from `crash` about dump routines failing, you probably have a valid dump file. Then, run the `stat` subcommand at the `>` prompt. For 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 reasonable for the dump, then perform some initial analysis. Refer to Section 4.8, “The crash command” on page 74 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.

4.6 Collecting the dump and related information

The easiest way to copy a dump and other system information to be used in analyzing the problem is by using the `snap` command. The `snap` command gathers system configuration information and compresses the information into a tar file that can then be downloaded to some other media. The `snap` command automatically creates the /tmp/ibmsupt directory, and several subdirectories are created below this.

4.6.1 The snap command

The `snap` command is a general purpose utility for gathering information about a system.

In general, it is best to run `snap -a` when building a snap image for sending to IBM. Also, The `-o` option is useful for writing the information collected by `snap` to removable media, such as a tape, for example:

```
snap -o /dev/rmt0.
```

The snap command supports the following information gathering options:

-D Dump and /unix information

-g General information

-k Kernel information

-f File system information

-S Security information

-L LVM information

-l Product levels for installed compilers

-n NFS information

-p Printer information

-s SNA information

-t TCP/IP information

-A TTY (async) information

-b SSA information

To gather all of the above data, use the `-a` option.

The snap command supports the following output control options:

-c Creates a compressed tar image of /tmp/ibmsupt. It produces snap.tar.Z images in /tmp/ibmsupt.

-o device Creates tar image on device.

-d directory Uses directory for snap operations, it replaces the /tmp/ibmsupt directory.

-r Removes all directories created by snap under /tmp/ibmsupt. Please make sure cleaning up the previous snap output. The `snap` command appends information to some snap output files.

4.7 Initiating a system dump

Normally, a system dump will occur automatically when the system crashes. When a system has hung, the system administrator can force a dump to determine the cause of the hang.

If you see the LED code 0c9, then a system dump is in progress. When the dump is successfully completed, this code will change to 0c0. Please refer to Section 4.5.2, “Dump status codes” on page 66 for other dump related LED codes. If you press the Reset button, the LED code will cycle through a set of codes, which will be something like 888-102-700-0c0-888.

If the Low-Level Debugger (LLDB) is enabled, a c20 will appear in the LEDs, and an ASCII terminal connected to the s1 or s2 serial port will show an LLDB screen. Typing `quit dump` will initiate a dump.

LED code description:

- 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.

4.7.1 LED reason codes

The reason code is the second value 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

  A page fault always begins as a DSI, which is handled in the exception processing of the VMM. However, if a page fault can not 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

  Instruction Access Interrupt. 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 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.

4.7.2 How to force a dump

You only force a dump on a machine that is completely hung. If the machine can be accessed at the console or remotely and commands can be run, then the machine is not hung. There are several ways of initiating a dump. You can choose one of these methods depending on the status of your machine.

4.7.2.1 Forcing a dump on MCA systems

To force a dump:

1. Turn the key mode switch to the Service position.
2. Press the Reset button once.
3. The system will start a dump and the LED panel will display LED 0c2.

4.7.2.2 Forcing a dump on PCI systems

There are a few different methods of forcing system dumps on PCI machines. The following information details briefly how to accomplish this for the different systems:

- Set the system option Always Allow System Dump to `true` using `smit`. Refer to Section 4.3.5, “Optional setup” on page 62, or run the `sysdumpdev -K` command.
- 6015/6050/6070/7020/7248 All Models

  Press the **Ctrl-Alt-Numpad 1** key sequence.
- 7017-Sxx

  1. Select **Function 22** on the operator panel and press **Enter**. The response `A1003022` displays on the operator panel.
  2. Select **Function 22** again and press **Enter**. The response `D1823080` displays on the operator panel. This indicates the dump request has been accepted by AIX, and the process continues, displaying four-character progress messages.

  Please refer to the *7017 S Series Installation and Service Guide*, SA38-0548 for more information.
- 7024-E Series

  Press and hold the Power button for longer than two seconds.

  The power LED will blink quickly to indicated that the dump has been started.

  Please refer to pages 1-8 in Chapter 1 of the *RS/6000 7024 E Series Service Guide*, SA38-0502.
- 7025-F30

  Press and hold the Power button, located on the front of the system, for approximately three to five seconds. A system dump is in progress when you see the disk activity light flash rapidly.
- 7025-F40/50

  Press and hold Reset button until the dump completion code is displayed. Alternatively, press the soft-power button for about five seconds or until 0c2 is displayed.
- Other PCI-based machines

  Press the Reset button.

> **Note**
>
> You can start a system dump by **Ctrl-Alt-Numpad 1** key sequence *only* on the native console keyboard.

4.8 The crash command

This section allows you to recognize some common problems using the `crash` command, and to make a basic determination as to what caused the problem.

4.8.1 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 `crash` on the live system.

The `crash` can also be used on a system dump. It is the primary tool used to analyze a dump resulting from a system failure. Invoking `crash` with a parameter specifying a dumpfile allows you to examine a dumpfile 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

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

4.8.2 What is the kernel?

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
- Provides synchronization and communication tools for processes

If the kernel has an error, the machine will crash. A user program will only create a core dump and halt.

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

4.8.3 Examining a system dump

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

```
itsosrv1:/dumptest> crash dumpfile unix
>
```

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

```
itsosrv1:/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`.

4.8.4 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 `?` character. To exit, type `q`. You can run any shell command from within the crash command by preceding it with an exclamation mark (`!`).

Please refer to the online documentation of *AIX Version 4.3 Kernel Extensions and Device Support Programming Concepts* for more information of the crash utility and all 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 uthread structure (uthread.h, mstsave.h).

- ds [addr]

  Finds the data symbol closest to the given address.

- knlist [symbol]

  Displays address of symbol name given. Opposite of ds.

- trace [-k][-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 addr 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).

4.8.4.1 stat subcommand

The `stat` subcommand gives 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.

4.8.4.2 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 gives you information on what was happening in the kernel when the crash occurred. The `trace -m` subcommand gives you a history of function calls and what interrupt processing was going on in the system. If the crash occurred while interrupt processing was going on, this is the command to use. This command traces the linked list of mstsave areas. The mstsave areas basically contain a history of what interrupt processing was going on in the system.

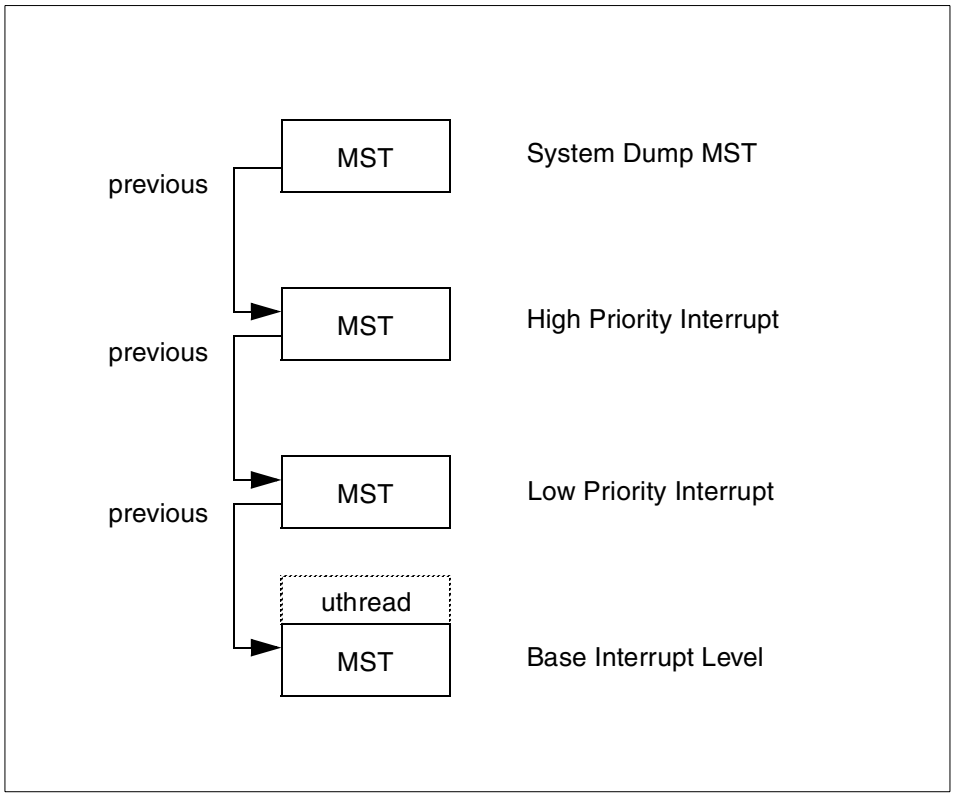


Figure 13. Machine save state area

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`.

IAR - Instruction Address Register. It has a address which caused the crash.
LR - Link Register who 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. We can also see that the interrupt priority is 3 (`intpri=3`). In this case, we can see that interrupt processing was occurring when the crash happened because the interrupt priority was less than 11 or 0xB. The base interrupt priority is indicated by 0xB or 11. This is the level at which a normal process runs.

When looking at a stack traceback, realize that the first thing on the stack 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. So, in the case of the middle stack traceback in our example, we see that `i_softmod` called `i_poll_soft`, which called some functions in `atmdd` and `atm_demux` module, which called `atmle_receive_ether_data`, which called `atmle_ready_ind`, and an assert was hit in `atmle_ready_ind`. You would have to look at the code for this to try to find out the cause of the assert action. Anyway, you can be sure that the `atmle_dd module` did something wrong.

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 `lslpp -w` command to find the fileset that contains the specific module. Refer to Section 4.8.4.7, “Finding addresses in kernel extensions” on page 84 for more information. You can get the latest fileset information from the Internet at:

```
http://service.software.ibm.com/support/rs6000
```

Use the `le` subcommand with an argument of the address listed in the IAR of the topmost MST area. 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 `lslpp -w` command to determine the fileset that contains the module. For example:

```
itsosrv1:/> lslpp -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.

Looking at the line:

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

You can see in the first column the address of the entry on the stack (not really important). 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 `1ec` from the beginning of the module `atmle_dd`.

The last of the stack tracebacks indicates the user level process (intpri=b) and the running process is `wait`. If you run the `user` subcommand, you will see 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, as we have shown, 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:

DAR Data Address Register

DSISR Data Storage Interrupt Status Register

SRV Segment Register Value

DAR2 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.

4.8.4.3 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  *uidl:0xe3001fa0
    *ganchor:0x00000000  *pgrpl:0x00000000  *ttyl: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
0xfff7ffff
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's position in the process table. You use this number to tell the `crash` command which specific process block or u-block to display. Note that the slot numbers are in decimal.

ST This is a 1-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.

4.8.4.4 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    CPUID  POLICY PRI CPU    EVENT  PROCNAME
  0 s       3        0  unbound    FIFO  10  78            swapper
        t_flags:  wakeonsig 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(hashed):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:

SLT Slot number.

ST Status. This may be i=idle, r=running, s=sleeping, w=swapped out, t=stopped, or z=zombie.

TID Thread ID.

PID Process id of the associated process. There may be multiple threads per process, but only one process per thread.

CPUID CPU ID of the CPU running the thread. On a uniprocessor system, this will always be 0.

POLICY This is the scheduling policy used for the thread and may have the values FIFIO, RR, or other.

PRI Dispatch priority. This is not the *nice* value.

CPU CPU utilization. This value is used for scheduling.

PROCNAME The name of the process for this thread.

EVENTS This is the wait channel if not zero.

FLAGS Status flags.

4.8.4.5 Display memory with od

You can display and examine memory areas from the dump using 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|
```

4.8.4.6 Looking for the error log

You can examine the last few error log entries from the dump using 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 fffffffa
>
```

4.8.4.7 Finding addresses in kernel extensions

The `le` subcommand can indicate what kernel extension 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, we can see 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.

4.8.4.8 VMM error log

When the Dump Status code indicates a DSI or an ISI, you need to look at the VMM error log. This is done using the `od` subcommand and looking at the `vmmerrlog` structure. See Table 6. For example:

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

Table 6. vmmerrlog structure components

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 shown in below:

0000000E This return code indicates an EFAULT. It comes from errno.h (14) and is returned if you attempt to access an invalid address.

FFFFFFFA This return code indicates 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 either 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.

4.8.5 Handling crash output

Some crash subcommands generate many more lines than can fit on one screen. Also, crash does not pause its output after each screen full. You will want to have some way of seeing scrolled-off 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 4.1.5, 4.2.1, 4.3, and above.

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:

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

4.8.6 Types of crashes

Common problems requiring crash dump analysis include the following:

4.8.6.1 Kernel panic or trap

This 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.

4.8.6.2 Addressing exception or data storage interrupt

This type of crash 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. Refer to Section 4.8.4.8, “VMM error log” on page 85 for more information.

4.8.6.3 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 Section 4.7.2, “How to force a dump” on page 72 for more information.

4.8.7 Data required by IBM support

In any type of system crash (Trap or DSI), the following data is required by IBM support to perform problem determination.

Ideally, the output of the `snap` command, collected as follows:

```
/usr/bin/snap -a -o /dev/rmt#
```

This collects the system dump, /unix, and other required information and puts it onto a tape drive.

In the event that a dump can not be sent, the following minimum information is a mandatory requirement for IBM support to analyze the problem:

- A kernel stack trace obtained by using the `trace -m` subcommand at the `crash` prompt. For example:

  ```
  # crash <dump> <unix>
  > trace -m
  ```

  This is usually sufficient unless the crash occurred in a kernel extension or device driver.

- The error log from the dump obtained by using the `errpt` subcommand at the `crash` prompt. For example:

  ```
  > errpt
  ```

- A full stack dump obtained by using the `fs` subcommand at the `crash` prompt. For example:

  ```
  > fs
  ```

  If this subcommand returns `Frame pointer not valid`, this output will not be useful.

Chapter 5. Hardware problem determination

This chapter guides you through the process of running diagnostics. There are various modes of running diagnostics, each with some limitations. This chapter will help you decide which mode is best for you. The process of running diagnostics enables you to confirm whether or not the problem you are experiencing is hardware related.

5.1 General advice

Where possible, run diagnostics concurrently while AIX is running. If it is not possible to test the suspect device concurrently, or there is a doubt about the integrity of the AIX system, then run stand-alone diagnostics from CD or diskette to the suspected device using the correct additional parts requested, such as wrap plugs or test media. If you run diagnostics and get a `No Trouble Found` report, you will probably be more successful in resolving the problem by concentrating on investigating software issues.

It is important that you use the exact additional parts requested by the diagnostic system. The diagnostics system specifies each required part by part number. The use of a similar, but incorrect, part can cause the diagnostics system to report a failure when none exists, or to report `No Trouble Found`, when in fact there is a problem.

5.1.1 Device location notation

You will see, both in this chapter and other chapters, output from various commands showing the location of devices or adapters in the system. This section explains the notation used to describe device location.

The format of the output normally consists of four fields, as shown below:

```
<logical_name> <status> <location> <description>
```

The first field is the logical name of the device or adapter. This is the name normally used as input to other commands that display or alter the configuration of the device. The second field shows the status of the device. Normally, this is either `Available` or `Defined`. The third field is the device location information. The fourth field is a description of the device.

The format of the device location field is as follows:

- For non-SCSI devices/drives

  AB-CD-EF-GH

- For SCSI devices/drives

  AB-CD-EF-G,H

  - The AB value identifies a bus type, or PCI parent bus, as assigned by the firmware.
  - The CD value identifies the adapter number or physical location.
  - The EF value identifies a connector.
  - The GH value identifies a port, address, device, or FRU.

The following example is the output from a command run on an MCA machine. The command has produced a list of all the SCSI devices:

```
hdisk0 Available 00-08-00-0,0 670 MB SCSI Disk Drive
hdisk1 Available 00-08-00-1,0 2.0 GB SCSI Disk Drive
```

The device location is the third field on each line. In this example, the format of the location code is as follows:

- The first pair of zeros denotes a system unit.
- 08 is the slot number on the I/O Planar on bus 0.
- The third pair of digits on some adapters would indicate the port or connector number. For example, some SCSI adapters have an external connector and an internal edge connector.
- In this case, 1,0 is the SCSI address of the disk but could be the address of a controller followed by the child device address number.

The format for PCI adapters shown below is slightly different. The first two pairs of numbers are read together and they correspond to a particular PCI bus and slot. You will, however, need to refer to the system documentation for your machine type to correlate this to a particular slot number in the machine, since most PCI-based machines have more than one PCI bus. The last two pairs of numbers have the same function as the MCA example above. For example:

```
hdisk0 Available 30-68-00-8,0  16 Bit SCSI Disk Drive
```

5.2 Problem diagnosis

Diagnosis of hardware problems depends upon observation, error information collection, and the results of running diagnostics. Generally, the first indication of a problem is an entry in the error log. The first entry for a problem can be many days or even weeks before you notice a problem. Sometimes, there may be multiple similar entries showing a degradation in

device performance before a failure. In other cases, there may only be a single error log entry.

There are two ways of accessing the error log. You can either use SMIT or the command line. Unless you are very proficient in AIX and can remember the syntax for the `errpt` command, then, probably SMIT is your best option since this will enable you to easily filter out unwanted entries. Refer to Section 2.2, “Error log file processing” on page 16 for more information.

Observation of a how a problem manifests itself will quite often give you an indication as to what may be the cause.

One of the more powerful tools to help you resolve a fault is the Diagnostics system, available through AIX when loaded on the machine and on a separate CD or diskette. Additionally, included in the diagnostics is a set of utilities in the Task Selection or Service Aid section. Included in this section are aids to help further diagnose SCSI, LAN, and disk subsystem faults.

5.2.1 Making sense of the error log

The easiest and most flexible way to access the error log is to use SMIT. For information on how to access the error log, refer to Section 2.2, “Error log file processing” on page 16.

The error log can contain many hundreds of entries, so it is always best to start with the summary format. This format will give you a chronological list of events starting with the latest event at the top of the screen. The following is an example of the summary output of the `errpt` command:

```
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
C60BB505   0511122999 P S SYSPROC        SOFTWARE PROGRAM ABNORMALLY TERMINATED
C60BB505   0511122999 P S SYSPROC        SOFTWARE PROGRAM ABNORMALLY TERMINATED
C60BB505   0511122999 P S SYSPROC        SOFTWARE PROGRAM ABNORMALLY TERMINATED
74533D1A   0510182999 U H SYSIOS         LOSS OF ELECTRICAL POWER
9DBCFDEE   0511083999 T O errdemon       ERROR LOGGING TURNED ON
192AC071   0510182599 T O errdemon       ERROR LOGGING TURNED OFF
0734DA1D   0503110499 P H fd0            DISKETTE MEDIA ERROR
C60BB505   0422105399 P S SYSPROC        SOFTWARE PROGRAM ABNORMALLY TERMINATED
C60BB505   0422105399 P S SYSPROC        SOFTWARE PROGRAM ABNORMALLY TERMINATED
C60BB505   0422105399 P S SYSPROC        SOFTWARE PROGRAM ABNORMALLY TERMINATED
3573A829   0422105099 U S CMDCRASH       SYSTEM DUMP
AD331440   0422104599 U S SYSDUMP        SYSTEM DUMP
AE26DD07   0422083099 P S SYSSPECFS      DRIVER RETURNED WITH INTERRUPTS DISABLED
9DBCFDEE   0422104799 T O errdemon       ERROR LOGGING TURNED ON
```

Look through the first couple of screens to see the sort of errors being produced and the frequency. Also check whether the errors fall into an obvious sequence. For example, a disk error followed by a SCSI error. Look at the time stamps of the errors for any pattern, for example, if they occur at or near the same time each day. The timestamp format is Month Month, Day

Day, Hour Hour, Minute Minute, Year Year. So, the first error in the example output above occurred on May 11th, 1999 at 12:29. The column marked C denotes the class of error. Class types are H for hardware, S for software, and O for operator message. See Section 2.3.2.4, “Class” on page 21 for a full description of the class entries. Once you have decided which of the errors interest you the most, expand the error log into the detail format. The example below shows a complete detailed error log entry:

```
LABEL:DISK_ERR4
IDENTIFIER:1581762B

Date/Time:       Thu Apr 29 08:08:04
Sequence Number: 604
Machine Id:      00018367A400
Node Id:         test1
Class:           H
Type:            TEMP
Resource Name:   hdisk0
Resource Class:  disk
Resource Type:   2000mb
Location:        00-08-00-0,0
VPD:
        Manufacturer.................IBMRISC
        Machine Type and Model......0664M1H
        Part Number.................86F0101
        ROS Level and ID............5 5A
        Serial Number...............00438487
        EC Level....................895186
        FRU Number..................86F0118
        Device Specific.(Z0)........000002029F00001E
        Device Specific.(Z1)........75G3644
        Device Specific.(Z2)........0983
        Device Specific.(Z3)........95333
        Device Specific.(Z4)........0002
        Device Specific.(Z5)........22
        Device Specific.(Z6)........895180

Description
DISK OPERATION ERROR

Probable Causes
MEDIA
DASD DEVICE

User Causes
MEDIA DEFECTIVE

Recommended Actions
FOR REMOVABLE MEDIA, CHANGE MEDIA AND RETRY
PERFORM PROBLEM DETERMINATION PROCEDURES

Failure Causes
MEDIA
DISK DRIVE

Recommended Actions
FOR REMOVABLE MEDIA, CHANGE MEDIA AND RETRY
PERFORM PROBLEM DETERMINATION PROCEDURES

Detail Data
```

```
SENSE DATA
0600 0000 0800 10F3 0100 0000 0000 0000 0102 0000 7000 0100 0000 0018 0000 0000
1500 0180 0001 0000 01A6 0000 000E 02FF 0000 0000 0000 0000 0000 0000 0A06 0000
0000 0000 100E 0801 0000 0032 4000 1800 0000 0000 1106 0100 0000 0000 0F0E 8080
0000 0001 0000 0001 0000 0000 2020 2020 2020 2020 2020 4C31 2020 2020 2020 2020
2020 2020 2020 3539 4833 3438 3120 2020 2020 4533 0000 0028 0002 7600
```

If you look at the sample error output for the disk error above, you will see a number of points of interest. At the top of each entry is the error label and identifier. Take a note of this ID and then use the Error Log Entry Analysis program accessible at the following URL:

`http://rshelp.austin.ibm.com/hardware/tools.html`

This IBM intranet Web interface will give you a description of the error, an explanation of the cause, and a recommended action plan.

The next area of the error log entry to look at is the area starting at the resource name and finishing at the end of the VPD section. The information given here will enable you to identify the type of device that is giving the problem, its size, and, more importantly on a complex system, its location. Additionally, the VPD will often give you the part number and FRU number to enable you to arrange for a replacement part to be ordered. The VPD can also indicate the microcode level of the device at the time the error occurred.

The next area that can provide additional information, especially when working with disk drives, 8mm tape drives, and 4mm tape drives, is the sense data. For IBM manufactured SCSI disk drives, the four digits in the fifth column from the left on the second row gives the UEC reported back to the system by the disk, and the data in the third column from the left in the same row is the retry count. Cutting and pasting the sense data from an error report into the form presented at the following URL will enable you to get a detailed description of the error generated:

`http://rshelp.austin.ibm.com/cgi-bin/dsense/dsense_form.sh`

Some diagnostic routines, except when running diagnostics from CD, use the sense data in the error log when run in Problem Determination mode. This use of the error log data by the diagnostics is the main way of deciding what is the cause of machine checks or checkstops in machines such as 7026. To run the diagnostics in this way, the date set on the machine must be within seven days of the error log timestamp on machines running AIX 4.3.1 and above (using Diagnostic Run Time options within Task Selection under AIX 4.3.1 and above lets you set a figure of 1 to 60 days). Machines running AIX levels 4.3.0 and below must have the date set within 24 hours of the timestamp for the error log analysis diagnostic to give correct results.

5.2.2 Physical inspection

The most obvious thing to look for when starting a diagnosis is physical damage, such as impact damage. Is anything obviously broken or does anything look incorrect to you? Look at cables going into the machine: Are any of them showing damage? Is the cable securely fixed to the adapter and to the device that is at the other end? Most installations end up with a tangle of cables either at the rear of the machine or under the floor. Try and look to see if there is any additional cabling nearby or intertwined with the cables of the machine you are looking at that might be a source of electrical interference. Power cables carrying heavy current are a prime source of electrical noise.

Look at the cabling of devices attached to the machine especially their routing, the tightness of the fixing of cabling, damage to cabling, and the proximity of heavy current carrying power cables. The positioning of adapters and devices can also influence problems.

If the machine you are working on is a PCI Bus machine, check the adapter positions in the machine against the recommended positions shown in the *PCI Adapter Placement Reference,* SA38-0538. The latest copy of this publication can be found at the following URL:

`http://www.rs6000.ibm.com/resource/aix_resource/Pubs/`

The placement of the adapters should also be checked against a RETAIN search to ensure that you have the most recent information. This especially applies to high-bandwidth adapters, such as SP Switch, SSA RAID, and ATM. PCI SP Switch adapters must not have any adapters in slots on either side of them. While you are looking at the adapter placement, make sure that the adapters are securely clamped to the chassis and are as deep into the card slot as possible. The correct seating of adapters is most important, especially in the case of J and R series SMP machines. An improperly seated adapter will, sometimes, not have a problem itself but will cause another adapter elsewhere on the bus to cause strange or intermittent problems.

5.3 Running diagnostics

Diagnostics on hardware can be run in three different ways. The first way of running the diagnostics is in concurrent mode, that is to say the system is up and running with users on and 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 of processes started and only rootvg varied on. Finally, the third way is stand-alone diagnostics from CD.

The CD-based diagnostics are a completely isolated version of AIX and so any diagnostics run are totally independent of the AIX setup on the machine being tested.

Which of these methods you use 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 or diskette are independent of the machine operating system. Advanced diagnostics run using the diagnostic CD or diskettes and completing successfully should be taken as proof of no hardware problem.

Note

If you are going to boot from CD or a mksysb tape on a machine that is in any configuration that has 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 in the machine being booted and will cause severe problems on any other machines sharing the same SCSI bus that have address 7 IDs set on their adapters.

The method by which you run diagnostics varies between machine type. The next three sections describe in detail how to run all the diagnostic modes on all machine types.

There are two RS/6000 models that do not have the capability to run AIX-based diagnostics. These machines are 7020-40P and 7248-43P. To run diagnostics on these models requires you to have the SMS diskette for the machine.

5.3.1 Concurrent mode

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

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

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

  This command will enable 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.

- To go directly to the main diagnostics menu, use the `diag` command.
- Using SMIT take the following menu route:
  - Problem Determination
  - Hardware Diagnostics
  - Current shell

Methods 2 and 3 will get you to the entry screen of the diagnostics menu. If you press **Enter** to continue from the entry screen, you will be presented with a menu as shown in Figure 14.

```
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 14. Main Diagnostics menu

The menu options shown in Figure 14 are explained in the following paragraphs:

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

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 Section 5.3.4, “Task selection or service aids” on page 104.

After you have selected the level of diagnostics you wish to run, you will then be presented with a menu for you to decide to use either the Problem Determination method or the System Verification method.

Problem Determination

This selection will make the diagnostic routine 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 machine checks and checkstops on 7025 and 7026 machine types. If you are performing diagnostics more than seven days since the machine check occurred then you will need to set the system date and time to within seven days of the machine check timestamp. The seven day period is required when using AIX 4.3.1 and later. If you are using AIX 4.3.0 or below, then the system date and time must be within 24 hours of the checkstop entry. See Section 5.2.1, “Making sense of the error log” on page 91 for more information.

System Verification

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

Concurrent mode provides a way to run Online Diagnostics on the 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
- The disk drives used for paging
- Memory
- Processor

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

- Minimal testing is used when the device being tested 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 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 to the CPU or memory, the diagnostics refer to an entry in the NVRAM that records any CPU or memory errors generated during initial testing at system power on time. By analyzing these entries, the diagnostics produce any relevant SRNs.

5.3.2 Stand-alone diagnostics from disk

This 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 disks containing paging space or rootvg. Stand-alone diagnostics from disk is 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.

5.3.2.1 MCA machines

To start a Service mode boot, power off the machine, then:

1. Set the key mode switch of the machine to the Service position.
2. Power on the machine without a CD, 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 you will then get to the screen giving you the choice of diagnostics to run.

5.3.2.2 PCI machines

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

To start a Service mode boot, power off the machine, then:

1. Turn on the machine power.
2. After a short period of time you will see the Icons screen. The Icons screen is described in Section 3.2.3.3, “PCI machine icons” on page 31. At this point, press **F6** if using a graphics console, or **6** if using an ASCII terminal. If you are using the graphics console, sometimes the display device will have power saving enabled, and so will take time to warm up and display images. This can lead 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, simply press the **F6** key.

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

5.3.3 Stand-alone diagnostics from CD or diskette

Stand-alone diagnostics run from CD or diskettes is a good way of proving if the problem is a hardware fault or an AIX problem. The CD 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.

Note

If you are going to boot from CD or a mksysb tape on a machine that is in any configuration that has 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 in the machine being booted, and so will cause severe problems on any other machines sharing the same SCSI bus that have address 7 IDs set on their adapters

5.3.3.1 MCA machines

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

Boot from CD

To boot from CD, complete the following steps:

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

   For the machine to boot from the Diagnostic CD, there must be an entry in the boot list including the CD. Using the code on the CD, the machine will boot, eventually pausing when displaying c31 in the LED panel. The code c31 is an indication to you that you need to select a system console. After selecting a console at the prompt, you will get the Diagnostic Entry screen 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 whole process again.

Boot from diskettes

The booting of a machine from diskette will only be possible on uniprocessor MCA machines. It is also a more complex method to use than the Diagnostic CD.

The machine types that support booting Diagnostic diskettes are:

- 7011 Models 220 - 250
- 7012 Models 320-390
- 7013 Models 520-59H

To boot a machine from diskette, the procedure is as follows:

1. Power off the machine.
2. Place diskette 1A or 1B into the drive and then power on the machine. The difference between diskette 1A and diskette 1B is the minimum size of memory required to run the diagnostics. If your machine is running with more than 8 MB of memory, either diskette will work satisfactorily.
3. The diskette is being read when 252 is displayed on the LED. When diskette 1 has been read and processed, c07 will be flashing on the LED. Insert diskette 2. The LED will then display c09 as it reads the diskette.

4. When c07 displays with diskette 2 in the drive, select diskette 3A, 3B, or 3C. The number 3 diskettes each correspond to different sorts of graphics cards. Select the diskette that corresponds to the graphics cards in your machine and place it in the drive. This is then followed by diskette 4 after c07 is displayed when diskette 3 is in the drive. However, if you only have an ASCII terminal and no graphics card, you can skip diskette 3 and place diskette 4 in the drive.
5. When diskette 4 has finished processing the LED will display c31. This means the machine is prompting for input on the console. Make your console selection as prompted on the screen. If you are using a vt100 emulator on a PC, you may have no text displayed on the screen. This is due to the boot process resetting the serial port. If the emulator is connected to serial port one, then type a `1` and then press **Enter**. The c31 should then disappear, and you should then get the Diagnostics screen displayed. If this does not work then try typing `2` then **Enter**. You will then be able to test the system unit. Should you wish to test any other devices, such as disks, you will need to load the corresponding diskette associated with the device. In the case of disks, this will be diskette 5.

5.3.3.2 PCI Bus machines

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

To start a CD boot:

1. Power off the machine.
2. Turn on machine power.
3. Place the CD 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 so take time to warm up before anything can be seen on the screen. This can lead 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 then press the **F5** key.

After doing the above, 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 on from this screen, you will then have the Diagnostic Entry screen displayed.

If your system contains ISA adapters, you will not be able to automatically test them. You will have to configure the adapters to the system using the ISA Configuration menu found in Service Aids or Task Selection.

The ISA Configuration Service Aids will display all the ISA adapters supported by the diagnostics that you are running. Diagnostic support for ISA adapters not shown in the list may be supported from a supplemental diskette supplied with the adapter.

Configuring ISA adapters from diagnostics

As mentioned above, when booting into diagnostics from CD, you can only test the ISA adapters in your system after they have been configured. The configuration is only for diagnostics purposes and will not affect the adapters configuration within AIX.

The following procedure explains how to configure an adapter using diagnostics. For the purpose of the explanation, we have used an 8-port adapter as an example.

1. Before you boot into diagnostics, note the ISA adapter parameters that are set for the adapter(s) you wish to test. These will either be the parameters you entered to install the card or parameters you have set on the card if the card has not yet been configured by AIX. For the example, the only parameter we had to record was the Bus I/O Address.
2. Boot into diagnostics from the CD-ROM as explained above.
3. Select Service Aids or Task Selection.
   The item that you select is dependent upon the level of diagnostics you are using. AIX 4.3.2 uses Task Selection.
4. Select **ISA Device Configuration Service Aid**. This will provide you with a menu like the one shown in Figure 15 on page 103.

```
            ISA DEVICE CONFIGURATION SERVICE AID

Move cursor to the selection, then press Enter.

Add an ISA Adapter
Remove an ISA Adapter
List all Defined ISA Adapters
List all Supported ISA Adapters









F3=Cancel                F10=Exit
```

Figure 15. ISA device configuration screen

5. From the menu, select **Add an ISA Adapter**.
6. Choose the desired adapter.
7. Enter the parameters, as requested, as shown in Figure 16. For our 8-port card, we only had to enter the Bus I/O Address.

```
ISA ADAPTER ATTRIBUTE SELECTION

Press 'List' to select additional values. If a
range is indicated below the description, then enter a
value within the range. Press 'Commit' when finished.



Bus I/O Address                          [0x104]
Bus Memory Address                       [0xd0000]
ISA Interrupt Level                      [   ]









F1=Help      F2=Refresh     F3=Cancel     F4=List
F5=Reset     F7=Commit      F10=Exit
```

Figure 16. ISA adapter attribute screen

8. After selecting **Commit**, a message similar to the following will appear:

```
Adding ISA Adapter 'IBM 8-Port Async, EIA-232 (ISA)'
Please Standby
```

If the addition was successful, you should see a message similar to:

```
ISA Adapter device 'sa2'
Configuration complete
```

9. Return to the main Diagnostics menu.

You should now be able to test the ISA adapter in addition to other devices in your system.

5.3.4 Task selection or service aids

This section is known by two names *service aids* or *task selection* dependent upon the level of diagnostics you are using. Task selection is the name used by AIX 4.3.2; however, in AIX 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.

5.3.4.1 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 sometimes help in furthering fault diagnosis.

5.3.4.2 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. This enables the down loading of microcode to devices that are not capable of being updated when AIX is running.

5.3.4.3 SCSI bus analyzer

This is probably 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:

1. From the Task Selection menu, select **SCSI Bus Analyzer**.
2. At the next screen, 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, the response saying so 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, you need to **Control-C** out of the service aid. Refer to Section 5.7, “Diagnosis of SCSI problems” on page 109 if you get this condition,

5.3.4.4 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 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 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 than the source disk 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.

5.3.4.5 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.

5.4 System Management Services (SMS)

System Management Services (SMS) is the diagnostic and configuration portion of the firmware. It is used on all PCI-based machines. The method of starting SMS varies between machines and is explained below. In the 7020 and 7248 machine types, SMS is very important since it is the only way of diagnostically testing the machine due to the fact that AIX diagnostics are not supported.

5.4.1 Firmware-based SMS

SMS on machine types 7017, 7024, 7025, 7026, and 7043 is contained within the firmware and can be accessed any time the machine is booted. SMS on these machines can be accessed by pressing **F1** (graphics display) or **1** (ASCII terminal) when the Icons screen is displayed on the console during the boot sequence.

You will then drop into a menu enabling you, among other things, to change the boot list or update system firmware.

5.4.2 Diskette-based SMS

As previously mentioned in this chapter, there are two types of machines on which it is not possible to run any form of AIX diagnostics. These machines are 7020-40P and 7248-43P.

The only method by which hardware tests can be run is by using the SMS diskette that was shipped with the machine. If, however, you no longer have this diskette, or your diskette has been corrupted, you can download a replacement image at the following URL:

`http://service.boulder.ibm.com/support/rs6000.support/downloads`

Note

The 7248 43P SMS microcode is for the 7248 machine only. Do not attempt to load it onto any 7043 43P machine. This will irreparably corrupt the firmware and permanently damage the machine.

5.4.3 Using SMS on 7020 and 7248

To start SMS do the following:

1. Power off the machine.
2. Place the SMS diskette in to the diskette drive.
3. Power on the machine.
4. When the icons appear on the screen, press **F1** anytime after the second icon but before the last.

You should now have a menu that gives you four options. Select the option marked **Test the Computer**. You can now run the test to the component you wish to test. Be aware that the level of code on the machine is designed to test IBM components only and so any other non-IBM devices may give incorrect results.

Additionally, within SMS, you can change the boot list, manage configuration changes, and load or refresh machine microcode.

5.5 TTY setup for use as a console

The tty console is primarily used when booting from CD, diskettes, or mksysb. At this point, output is displayed on the screen using spaces and characters. Therefore, any terminal will work providing the following settings are used:

- 9600 Baud
- 8 bit
- 1 Stop bit
- No Parity
- Xon/Xoff

Once AIX is running or Diagnostics screens are in use, screen addressability is used and a definite screen type must be configured.

The supported screen types are as follows

- IBM models 3101, 3151, 3161, 3162, 3163, 3164, ibmpc, lft
- tvi912, tvi920, tvi925, tvi950
- vs100
- vt100, vt320, tv330, tv340
- wyse30, wyse50, wyse60, wyse100, wyse350

- Sun

If your terminal type is not listed, you can use an emulator, often supplied with Windows family operating systems. Set the emulation mode to vt100 and use the communication settings as shown above.

5.6 Checkstops and machine checks

Most machine checkstops and machine checks are caused by problems in the CPU complex. The most common symptom seen when a checkstop occurs is that the machine will be up and running quite satisfactorily, then, all of a sudden, the machine halts unexpectedly and undertakes a reboot. The reason for a checkstop rebooting the machine is that, if the processor logic detects a problem within the processor or memory, the only way that it can do any problem determination itself is to shut the machine down. On the way down, the processor writes the error data from the problem to the NVRAM. Once the machine has shutdown completely, it will reboot and then the processor and memory can be fully tested. If there is a problem at this stage, a LED code will be displayed. If no problem is found, then during the boot process, the error information from the NVRAM is transferred to the error log.

5.6.1 How to use the data

The method of processing the checkstop or machine check data depends on the technology of the machine.

MCA machines will have an error log entry indicating a checkstop. At the end of the error log entry, you will see the full path name of the checkstop file. These files require a special program to decode the cause of the problem. Open a support call with your software support and send them the latest checkstop file plus a copy of the detailed error report and a full VPD listing. Software support will then provide you with an action plan.

The CHRP machines models 7017-S70, 7017-S7A, 7025-F50, 7026-H50, 7026-H70, 7043-150, and 7043-260 transfer the machine check details to the error log. It is then possible to run advanced diagnostics in problem determination mode to get the cause of the machine check. See Section 5.2.1, “Making sense of the error log” on page 91 for time stamp parameters. This is possible because as you select problem determination mode, the diagnostics then read the error log entry for the machine check, take the sense data from the entry, and perform an analysis of it. This will then produce an SRN and a list of failing FRUs.

All of the other PCI machines when running AIX 4.1.5 or above produce an error log entry with sense data that can be decoded by your software support organization.

5.7 Diagnosis of SCSI problems

Diagnosis of problems within the SCSI subsystem requires you, in most cases, to have the machine to yourself so that you can perform tests that require having the power removed from the machine.

There are two distinct classes of SCSI devices and adapters each having different properties. The two types are:

Single-ended SCSI Over the years this implementation of SCSI has progressed from SCSI 1 to SCSI 2 then onto SCSI 2 Fast and Wide with Ultra SCSI being the latest at the time of this writing. Each of them is different in speed and bus width. Interchanging of devices between the bus types can have an effect on bus performance. Differential SCSI devices and terminators should not be attached to a single-ended SCSI bus because they will cause unpredictable results.

Differential SCSI This is another implementation of the SCSI bus. The speeds and bus widths are the same as the corresponding single-ended implementation, but the permitted cable lengths are between 18 to 25 meters, which is much longer than single-ended SCSI. Because of the technology used to achieve a greater bus length, no differential device should be attached to a single-ended SCSI bus; likewise, single-ended devices should not be attached on a differential SCSI bus.

It should be noted, however, that some differential SCSI adapters will have a single-ended SCSI connector enabled on them, so they can run internal single-ended devices on a system. It should also be noted that differential adapters have a set of terminator resistors on the adapter. These will be removed when the adapter is used in a shared SCSI bus environment, and therefore, if you remove an adapter from such a system, you should remember to replace the resistors or use a terminator on a pigtail Y cable.

SCSI problems can be caused by either a device, cables, or terminator. You can sometimes end up diagnosing a SCSI problem when the original call was,

for example, a disk drive. There are a number of ways to diagnose the various faults, some of which are listed in the following sections.

5.7.1 Basic SCSI checks

Check the cabling thoroughly for the security of the cables connected to the devices; also cross reference the cables used with the diagrams in the publications *Adapter Devices And Cable Information For Multiple Bus Systems*, SA38-0516, and *Adapter Devices And Cable Information For Micro Channel Bus Systems*, SA38-0533. In each of these publications, you will find wiring diagrams for many different device scenarios. Each diagram will give you the feature code and part number of all components in the diagram; cross check with your setup paying particular attention to termination of the bus. In addition, make sure that all devices are compatible and that you do not have a mixture of differential and single-ended cables or terminators.

Next, run the following command:

```
lsdev -Cs scsi.
```

You should then get similar output to what is shown below:

```
hdisk0 Available 00-08-00-0,0 670 MB SCSI Disk Drive
hdisk1 Available 00-08-00-1,0 2.0 GB SCSI Disk Drive
hdisk3 Available 00-08-00-3,0 2.0 GB SCSI Disk Drive
cd0    Available 00-08-00-4,0 CD-ROM Drive
rmt0   Available 00-08-00-5,0 2.3 GB 8mm Tape Drive
rmt1   Available 00-08-00-6,0 7332 4mm Auto Loader SCSI Tape Device
hdisk2 Available 00-08-00-2,0 2.0 GB SCSI Disk Drive
```

The output you get should show all of the SCSI devices listed in the ODM both available and defined. Are all of the devices you would expect to see shown as available? If not, check that the device has power turned on, check cabling for bent pins, improper termination such as two terminators on the bus, or missing terminators. Some non-IBM devices have jumpers inside the device to provide termination with that device. If you find that you are having difficulty finding the problem, then, if possible, start with one device on the bus and keep adding devices until the fault occurs. However, *never* hot plug onto a live SCSI bus; it has been known to seriously corrupt rootvg or customer data. The exception to this, of course, are hot plug SCSI disks available in machine types 7017, 7025, and 7026.

If you can not get the SCSI bus to work even with just one device on the bus, then before replacing the adapter, refer to *Diagnostics Information for Multiple Bus Systems*, SA38-0509, or *Diagnostic Information for Micro Channel Bus Systems*, SA38-0532. These two documents contain procedures for testing the adapters PTC thermal cutout and also a test for correct bus termination.

5.7.1.1 SCSI devices missing or duplicated

Quite often, when you add devices or change the layout of a SCSI bus, you will get one of the following scenarios:

- SCSI device missing. The most common cause of this is that you set the SCSI ID or address to the same as another device. This can easily be done accidently if the device has the push button address switch on the rear of the device. The only effective way to sort this situation out is to physically recheck the address of each device connected to the bus. Remember that active SCSI backplanes on machines will often have SCSI IDs 14 or 15 assigned to them. Also, check to see that terminators are in the correct place, especially if you have just added an OEM device. Some OEM devices are sent out with a default setting of termination enabled. A device after a terminator may well not be seen. Check to see that all cables are seated squarely and firmly in their sockets. A misplaced cable or terminator can also cause an incorrect SCSI ID to be reported. SCSI address probing is accomplished by the adapter monitoring data lines for a response during bus initialization. The responses received are used to determine the addresses used. For example, a response received on line 0 equates to address 0, a response on line 5 equates to address 5, and so on. Therefore, anything that can result in the signal being misrouted or misinterpreted can give address misreporting.
- Multiple instances of a device. For example, `lsdev -Cc tape` reports back rmt1, rmt2, rmt3 through to rmt7. The cause of this is that one of the tape drives has the same SCSI ID as the SCSI adapter, generally 7. Again, this can be done during installation of the device or by accident. If no device duplicating the address of the SCSI adapter is found, see the previous paragraph for instructions on how to check SCSI cabling.
- The SCSI bus service aid, available from diagnostics, is most useful for the diagnosis of missing SCSI devices. It will give you an indication as to the state of the device at a specific SCSI address. The three possible replies 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.*

    This message will be produced when there is no device using the SCSI ID, or the device using the ID is powered off for some reason, for example, the power cable not being plugged in or there is a power problem on the device itself.
  - *The exerciser transmitted a SCSI Inquiry command and received a valid response back without any errors being detected.*

This response indicates that the device at the polled ID is capable of responding to some instructions from the adapter and, therefore, is connected to the SCSI bus. If the device you tested is not being seen correctly or not working as expected, run diagnostics to see if it finds a problem.

- *A check condition was returned from the device.*

  If a check condition is returned by the device, it will usually fail diagnostics as well and, so, should be replaced. Should the failing device be a disk, and, due to the lack of a good backup, you are desperate to save the data, then in a small number of cases, this may be possible by changing the electronics. The success of this will depend on what is causing the check condition. If the electronics change is successful, it is strongly advised that you immediately take a backup of the data followed by the complete replacement of the disk assembly. Should all efforts be unsuccessful in saving the data, professional assistance should be sought from a reputable data recovery service.

5.8 Serial Storage Architecture (SSA) disks

This disk subsystem is capable of being externally connected to one or more RS/6000 systems. Certain models of RS/6000 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/sec or 160 MB/sec. As you can see, the number of possible combinations is almost endless and changes at each product announcement. Therefore, the SSA configuration rules detailed below cover basic considerations.

5.8.1 General SSA setup rules

The following rules must be followed when connecting a 7133 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 a 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 a 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.
- Cabling 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 shown in Table 7.

Table 7. 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 1 and 8 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	

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

`http://www.hursley.ibm.com/~ssa/`

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

5.8.2 SSA devices

SSA subsystem components use microcode to control their functionality. When dealing with any SSA problem, you should ensure that the microcode level and any drivers on all devices in the loop are at the latest published level. The latest code can be obtained from the following URL:

`http://www.hursley.ibm.com/~ssa/`

5.8.3 SSA disk does not configure as hdisk

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. Type `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.

Obviously, 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 some time if the disk is 9 GB or larger.

5.8.3.1 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 the *SSA Adapters: User's Guide and Maintenance Information*, SA33-3272.

If you propose to do any actions 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.

5.8.3.2 Changing SSA disks

SSA disk changing activity is hot swapable. In the work preparing AIX for the removal of an SSA disk, do *not* rmdev the pdisk prior to physically removing the disk from the enclosure. You will need the pdisk to do the following steps. Remove the pdisk only when all steps are completed. 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. When the replacement disk or blanking module is inserted, you use the same service aid to reset Service mode. This will start up the new disk and take the other disks out of string mode. At this point, you can now *rmdev* the pdisk allocated to the disk you removed. The disk change procedures will tell you to run the `cfgmgr` command.

Note

The `cfgmgr` command should *not* be executed on any system that is in an HACMP cluster. To do so will seriously damage the configuration of the machine possibly resulting in the crashing of the complete cluster.

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: Users Guide and Maintenance Information,* SA33-3272. Read these procedures carefully because some of the earlier editions of this book 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 disks are:

1. Addition of the replacement disk to the system using `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.

5.9 Peripheral devices

Peripheral devices can cover a multitude of equipment that can be directly connected to RS/6000 system units. The more obvious of these are disk drives, tape drives, tape libraries, and CD juke boxes. Most of these devices are usually connected using an SCSI bus of some description. Any problem shown by a peripheral device could be ultimately caused by the SCSI bus. Therefore, always include SCSI bus checking in any problem determination

on a peripheral device that is connected to a SCSI bus. See Section 5.7, “Diagnosis of SCSI problems” on page 109.

5.9.1 Disk drawers or towers

Most disk drawers and towers contain one or more power supplies and cooling fans. The power supply can sometimes be fuse protected and so the fuse should be checked prior to replacement of the power supply.

Note

Refer to the documentation supplied with the equipment prior to performing any service actions. Always use the manufacturers recommended replacement parts.

Most disk enclosures have fans to keep the internal components cool. The power supplies will often use internal logic to check for fan rotation speed before they allow power to stay up. Therefore, check that no debris has blocked the fan from rotating or is decreasing the speed at which the fan rotates.

The inside of a drawer or tower of disks can contain from one to sixteen disks or more. The disks can be accessed individually or grouped together into RAID sets. If the disks are RAID or controlled by a logic card, then problem determination should only be carried out using the supplied documentation, so data integrity is protected.

External disk enclosures containing no logic can be treated as just a straight SCSI bus and diagnosed accordingly. This means that diagnostics should be run from either Service mode boot or from CD. If the diagnostics fail, all connections on the data bus within the enclosure should be checked for tightness and bent or deformed connections prior to changing components. Some enclosures can provide bus termination, so always check to see that the enclosure is supplying bus termination only at the end of the bus and without any other external terminator being used. Two terminators on the same end of the bus will cause unpredictable results.

For more information about SSA disk drawers and towers, see Section 5.8, “Serial Storage Architecture (SSA) disks” on page 112.

5.9.2 External tapes and tape libraries

Tape drives and libraries range from a QIC tape drive capable of saving 150 MB, to tape silos containing tens of thousands of 1/2 inch tape cartridges. Physical connections to the RS/6000 can range from SCSI to ESCON.

Due to the diversity of sizes and methods of connection, this redbook will only cover QIC, 4 mm, 8 mm, and DLT devices, and their desktop libraries. The larger drives require specific training to enable repairs to be carried out. The tape silos can also be dangerous to enter as the robots are under automatic control and can move without warning.

There are, however, a number of things that can cause problems on tape drives that will apply to all drives and libraries regardless of size or complexity.

5.9.3 General tape troubleshooting

Tape drives are very expensive and often are not the reason a customer cannot read or write a tape. Often the problem is caused by one or more of the following:

- Bad blocksize. The customer attempts to read at one blocksize when the tape is written at another. Or the blocksize the customer is attempting to write at is preceded by a 0 (ZERO); 0512 is not the same as 512. The zero is recognized by the system as Octal and some applications do not support Octal values. The `tcopy` command can be used to determine the blocksize of a tape. Invoking the `tcopy` command with the name of the tape drive as the only argument will cause it to attempt to read the tape in the drive and display information about the blocksize and tape files. For example:

  ```
  # tcopy /dev/rmt0
  tcopy: Tape File: 1; Records 1 to 7537; Size 512.
  tcopy: File: 1; End of File after 7537 Records, 3858944 Bytes.
  tcopy: Tape File: 2; Records 1 to 2900; Size 512.
  tcopy: File: 2; End of File after 2900 Records, 1484800 Bytes.
  tcopy: Tape File: 3; Record 1; Size 512.
  tcopy: File: 3; End of File after 1 Records, 512 Bytes.
  tcopy: Tape File: 4; Records 1 to 514048; Size 1024.
  tcopy: File: 4; End of File after 514048 Records, 526385152 Bytes.
  ```

  This may take a considerable amount of time to run, since it will attempt to read the tape at many different blocksizes until it succeeds.

- Device buffering turned off causing the drive to write every block of data as it is received. Normal operation is to buffer the data, allowing the drive to stream the data to the tape as a continuous operation.

Device buffering being turned off will create a sort of shoe shining effect that will greatly accelerate head wear on the drive.
Check this parameter using the `smit tape` command.

- Writing a tape with one command then attempting to read the tape with a non-compatible command, such as writing the tape with a `tar` command and attempting to read the tape with a `restore` command.

  Rather than attempting to read the tape directly with various backup commands, it can be quicker to determine the format of the tape by reading a small portion of the start of the tape using the `dd` command, then using the `file` command to attempt to determine the format of the archive. For example:

  ```
  # dd if=/dev/rmt0 of=/tmp/tape.data bs=512 count=4
  4 + 0 records in.
  4 + 0 records out.
  # file /tmp/tape.data
  /tmp/tape.data: tar archive
  ```

  When attempting to read tapes written on other systems, the data format may be incompatible depending on the options used to write the tape. The `cpio` command is particularly prone to this problem. The GNU version of the `cpio` command, available at many sites on the Internet, can very often read the non-standard format archives.

- Poor quality media or worn out media being used. Clean the drive and retest the customer operation with a new IBM data cartridge recommended for the drive.
- Misunderstanding of the meaning of the LEDs on the tape drive.
- Not cleaning the tape drive, using cleaning cartridges that are used up, or not following the recommended cleaning frequency or cleaning instructions.

  Check that you are using the correct cleaning cartridge. The part numbers of the IBM cleaning tapes are given in Section 5.9.4, “4 mm, 8 mm, and DLT tape drives” on page 119, and Section 5.9.5, “QIC tape drives” on page 120.
- Not cleaning the tape drive after an I/O error prior to retrying the failing operation.
- Improper SCSI bus termination. See Section 5.7, “Diagnosis of SCSI problems” on page 109 for further information.
- Conflict of tape drives SCSI address with another device on the bus.
- Another device causing noise on the bus.

- Environment not suitable for tape drive operation, for example, the area around the machine has deposits of dirt or abrasive dust. Ideally, the operating temperature of the drive and the tapes should also be the same. Using tapes that are much colder than the drive itself will promote condensation on the tape surface that will accelerate dirt depositing on the read/write head.
- Software level of the system does not support the tape drive. When adding a new type of tape drive, you may need to add PTFs to the system to fully support the new device.
- Device microcode not at the latest level, even on new devices or replacement parts direct from stock. Check the following RS/6000 microcode download Web site for the latest levels of microcode:

  `http://www.rs6000.ibm.com/support/micro`
- Trying to read or write a data density not supported on the drive. Check the documentation supplied with the drive. For example, trying use a 2.3 GB 8 mm tape drive to read a tape written on a 5 GB 8 mm drive.
- Inconsistencies in the backup script. Use a new data tape and `tar` a large portion of data onto the tape. If this copies without error, use the `tar -tvf` command to read the tape back. /var is often a sufficiently large enough mount point to use to test the drive.
- Incorrect positioning of tape labels. The use of sticky labels anywhere on the tape cartridge except on the proscribed indentations will cause the tape to position itself incorrectly within the tape path resulting in excessive errors. The use of labels on the metal reference surfaces of QIC and 3570 tapes will cause errors while reading or writing to the tape and may result in possible damage to the drive itself. In some cases, the tape drive will fail to eject a tape cartridge due to incorrect positioning of labels on the tape cartridge or multiple labels stuck on top of one another. To correct this problem, the tape device will have to be disassembled to remove the tape cartridge. This should only be attempted by trained service personnel.

5.9.4 4 mm, 8 mm, and DLT tape drives

These families of drives are the most common of the drives used on RS/6000 machines. Their capacities range from 2 GB uncompressed to as much as 70 GB using data compression. For example:

- Removal of stuck tapes can sometimes be effected by resetting the drive. These devices can be reset by either pressing the **Eject** button for 20 seconds or by using the following command:

  `diag -c -d /dev/rmt(x)`

If the above fails, then, when possible, power off the drive or, if it is an internal device in a system, power off the system.

- Media errors can occur if the customer has not been cleaning the drive regularly or has been using an incorrect cleaning tape. Take an IBM cleaning tape and clean the drive three times in succession. Be sure to mark off the three uses on the label of the cleaning tape.

  At the time of this writing, the cleaning tapes are:

  - 4 mm Tape Drives Cleaning Cartridge P/N 21F8763. Maximum use is fifty cleanings.
  - 8 mm 2.3 GB, 5 GB, and 7 GB Tape Drives Cleaning Cartridge P/N 16G8467 (do not use this cleaning cartridge on the 20 GB 8 mm drive). Twelve cleanings on 2.3 GB drive and 22 cleanings on 5 GB and 7 GB drives.
  - 8 mm 20 GB tape drives. These will request a cleaning tape each time a 5 GB or 7 GB data tape is read. Writing to tapes, other than 20 GB capacity, is not possible with this drive. Cleaning Cartridge P/N 59H2898 (only use this cleaning cartridge on the 20 GB 8 mm drive). Minimum number of cleanings per tape is 18. Cleaning tape usage depends on the reason the tape drive is being cleaned, whether due to normal usage or after a 5 GB or 7 GB tape has been read.
  - DLT Cleaning Cartridge P/N 59H3092. Twenty cleanings per tape.

If you want to return a tape drive to the default configuration, the quickest way is normally to `rmdev` the device, then run the `cfgmgr` command to add it to the system again.

5.9.5 QIC tape drives

The Quarter Inch Cartridge (QIC) drive comes in many capacities from 150 MB to 13 GB. All cartridges are dimensionally the same but the densities are not universally interchangeable. When conducting diagnostics, make sure that you are using the correct part number test tape as asked for in the dialog panel. If you use the wrong test tape, you will get an incorrect test result.

Cleaning cartridges for the QIC drives are:

- 1/4 inch Tape Drives Cleaning Cartridge P/N 16G8572 (except 7207-122/315). 50 cleanings.
- 1/4 inch 7207-122/315 Tape Drive Cleaning Cartridge P/N 59H4366 (use only for 7207-122/315). 50 cleanings.

When checking the settings for these drives, make sure that the retension setting is enabled. This setting is sometimes changed to save the 5 to 10 minutes taken to retension a tape when it is inserted into the drive. If the setting is changed, then for the first couple of times, the tape will perform correctly, but as the tension decreases, loop back errors will occur.

5.9.5.1 Tape error log entries

The error log entries for tapes can be most useful in deciding if it is one particular tape cartridge that is causing the problem or the tape drive itself. The type of error logged, plus whether the tape operation completed or not, will give an indication of where the problem may lie.

The following is an abbreviated list of the possible tape errors and their meanings:

TAPE_ERR1 This is likely to be a media failure. The process using the tape drive normally terminates on encountering this error.

TAPE_ERR2 This is normally a drive failure, but can be caused by a media failure. The process using the tape drive normally terminates on encountering this error.

TAPE_ERR3 This is likely to indicate that the tape media is starting to fail. The process using the tape drive will not normally terminate on encountering this error.

TAPE_ERR4 Unknown cause, requires further analysis. The process using the tape drive normally terminates on encountering this error.

TAPE_ERR5 Unknown cause, requires further analysis. The process using the tape drive normally terminates on encountering this error.

TAPE_ERR6 The 5 GB, 7 GB, and 20 GB 8 mm tape drives log this error when the drive exceeds the preset limit of tape motion hours since the drive was last cleaned. The process using the tape drive does not normally terminate on encountering this error.

Note

The 20 GB 8 mm tape drive will also set the cleaning indicator LED, and log TAPE_ERR6 on most media errors.

5.9.5.2 Tape libraries

Tape libraries are tape drives with a controller card that controls and operates an auto tape feeder mechanism of some description. The tape load and feed mechanism is controlled by a software product called Atape. This product

includes the `tapeutil` command that provides a command line interface that can be used to move the tapes within the library.

If you are experiencing tape movement problems during backup or restores, you may get a better idea of what is happening if you use `tapeutil` commands and watch the library at the same time. This diagnostic procedure can not be used with the model 7337 or model 7332 tape libraries because the auto feed mechanism is hidden from view.

Tape movement problems can often be eliminated by performing a calibration of the library using the control panel of the device. This will require the service documentation for the tape library device, since the method used varies between machine types. Additionally, for persistent problems where hardware has been eliminated as the cause of the problem, or all of the hardware components have been changed, install the latest version of the Atape product. This product can be downloaded from the Storage Systems Web site at:

`ftp://index.storsys.ibm.com/devdrvr/AIX/`

When installing or upgrading the Atape product, examine the README files carefully because the installation process requires certain conditions to be met prior to installation.

If the Atape product is removed from the system, you will need to reload any device drivers supplied with the tape drive if you wish to use them instead of Atape. If this is not done, the drivers will not work even though the `lslpp` command shows them as correctly installed.

5.9.6 CD jukeboxes

These machines are usually connected to the system using an SCSI Differential Fast Wide bus. Most devices require special training to set them up, so really the only problem determination you can do is regarding the SCSI bus. See Section 5.7, “Diagnosis of SCSI problems” on page 109.

Chapter 6. LVM and JFS

This chapter explains the basic concepts of the LVM and JFS systems and how to repair some of the more common problems that can occur.

6.1 Logical Volume Manager (LVM)

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 LVM 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 by using a layer of device driver code that runs above traditional disk device drivers.

The LVM consists of:

- High-level Commands
- Intermediate Commands
- Library Commands
- LVM Device Driver
- Disk Device Driver
- SCSI Device Driver
- Actual Disk

6.1.1 High-level commands

The LVM high level commands are the commands that provide the user interface to the LVM. The commands are used by users from the command line or shell scripts and are called from the LVM screens in SMIT.

The high-level commands are a mixture of shell scripts and binaries and provide a high degree of error checking on input arguments. Most of the LVM commands reside in /usr/sbin. There are exceptions, such as the library commands that reside in /usr/lib/liblvm.a, /usr/lib/libsm.a, and /etc/cfgvg.

The high-level commands are:

```
chlv, chpv, chvg, cplv, exporting, extendlv, extendvg, importvg, lslv,
lsvg, lsvgfs, migratepv, mklv, mkvg, reducevg, reorgvg, rmlv, rmlvcopy,
syncvg, synclvodm, varyonvg, and varyonvg.
```

6.1.2 Intermediate commands

The intermediate commands are called by the high-level commands. These commands have been designed to only be called from the high-level LVM commands, so little or no error checking is performed on the input arguments. This is because the arguments have been generated by the high-level commands that would terminate before calling the intermediate commands if they were invoked with erroneous input.

Sometimes, it is necessary to use an intermediate command during the problem determination process. Care should be taken to invoke the command with the correct arguments.

The intermediate commands are:

```
getlvcb, getlvname, getlvodm, getvgname, lvgenmajor, lvgenminor,
lvrelmajor, lvrelminor, putlvcb, putlvodm, lchangelv, lcreatelv, ldeletelv,
lextendlv, lquerylv, lreducelv, lresynclv, lchangepv, ldeletepv,
installpv, lquerypv, lresyncpv, lcreatevg, lqueryvg, lqueryvgs, lvaryonvg,
lvaryoffvg, lresynclp, and lmigratepp.
```

6.1.3 Library calls

The LVM library calls are part of the LVM Applications Programmer Interface (API) that lets programmers access the LVM layer directly through the library layer provided by /usr/lib/liblvm.a. It is unusual for this API to be used because of the complexity of issues within LVM. Users mostly rely on shell scripts that call the high-level commands. The API library functions begin with lvm_ and are documented within the AIX documentation. These library calls are only used in situations where the user is writing C code and does not want to access the high-level commands from within a C program, or they require a faster response from LVM than they can achieve with the high-level commands. Writing an LVM specific program using the API certainly does give a performance improvement to some applications. Note, that the LVM library API is not a thread-safe library. Thus, threaded user programs that reference the LVM API may not work correctly.

6.1.4 LVM device driver

The LVM device driver comes in two portions: the pinned portion /usr/lib/drivers/hd_pin_bot and the non-pinned portion /usr/lib/drivers/hd_pin. The non-pinned portion is called hd_pin for backwards compatibility with AIX V3 where the driver was just called hd_pin and the entire driver was pinned into memory (not pageable). With AIX V4, the driver's true non-pageable portion is in hd_pin_bot, and the hd_pin is now pageable. The LVM device driver is called directly by the jfs file system and the lvm library routines.

When a request is received by the LVM device driver, it calls the disk device driver.

6.1.5 Disk device driver

There are various disk device drivers in AIX. The most common disk device drivers are the ones for SCSI device drivers, /usr/lib/drivers/scdisk and /usr/lib/drivers/scdiskpin. The second most common disk device driver is probably the one for the 7133 SSA dasd device. This is contained in the binaries /usr/lib/drivers/ssadisk and /usr/lib/drivers/ssadiskpin. In both device *pin* is that which is pinned into memory and cannot be paged out of memory. When a request from the LVM device driver is received by the disk device driver, the request is packaged and then transferred down to the SCSI device driver

6.1.6 SCSI device driver

There are several SCSI device drivers for AIX and they have the common name string of scsi residing as part of their name. The most common of them is the original SCSI-1 device driver /usr/lib/drivers/hscsidd. The SCSI device driver takes a command for a SCSI device, such as tape or disk, and processes it to be sent to the SCSI device connected onto the SCSI bus. The SCSI device is device neutral; it does not know or care which device sent it a command or even what that command is. It treats all requests the same and puts them into the same type of SCSI command packaging required by all SCSI devices.

6.2 LVM data

The different pieces of data required by the LVM system to operate correctly are stored in a number of places.

6.2.1 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 disk itself and is also stored in the 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.

6.2.2 Volume groups

Each volume group has a Volume Group Descriptor Area (VGDA). There are multiple copes 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 in the volume group 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 if a volume group should be varied on.

For a single disk volume group, there are two VGDAs on the disk. When a second disk is added to make a two disk volume group, the original disk retains two VGDAs 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 each additional disk adds one new VGDA per disk.

A volume group with quorum checking enabled (the default) must have at least 51 percent of the VGDAs in the volume group available before it can be varied on. Once varied on, if the number of VGDAs 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 VGDAs 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.

6.2.3 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 Journalled 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.

6.2.4 Object Data Manager (ODM)

The Object Data Manger (ODM) 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.

6.3 LVM problem determination

By far, the most common LVM-related problems are to do with disk failures. Depending on the extent of the failure, you may be able to recover the situation with little or no data loss.

Depending on the exact nature of the problem you are experiencing, a recovery attempt that fails may leave the system in a worse condition than before. In some situations, the only way to recover is to restore from a backup.

6.3.1 Data relocation

When a problem occurs with a disk drive, sometimes data relocation takes place. There are three types of data relocation:

- Internal to the disk
- Hardware relocate 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, so 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 a 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.

6.3.2 Step one

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.

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 /) becomes full during the process of importing a volume group.

6.3.3 Resyncronize the ODM

If you suspect the ODM entries for a particular volume group have been corrupted, a simple way to resynchronize the entries is to vary off and export the volume group from the system, then import and vary on to refresh the ODM. This process can only be performed for non-rootvg volume groups.

For the rootvg volume group, you can try using the `redefinevg` command that 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 the LVM information stored on disk has become corrupted, use the `synclvodm` command to synchronizes and rebuild the LVCB, the device configuration database, and the VGDAs on the physical volumes. For example:

```
# synclvodm -v myvg
```

If you have a volume group in which one or more logical volumes is mirrored, use the `syncvg` command if you suspect that one or more mirror 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
```

Will synchronize the mirror copies of the logical volume `lv02`.

```
# syncvg -v myvg
```

Will synchronize all of the logical volumes in the volume group `myvg`.

6.3.4 Collecting data

If your LVM problem can not be traced to a faulty disk or corrupted ODM data, you need to collect information to determine the cause of the problem.

Since many of the high-level LVM commands are shell scripts, you can alter the script to add the `-x` option that causes the commands executed by the script to be displayed on standard out. The first line in an executable shell script is normally similar to:

```
#!/bin/ksh
```

Edit the command you are having a problem with, and add `-x` to the end of the first line, as follows:

```
#!/bin/ksh -x
```

Now invoke the script with the correct arguments to recreate the problem. Redirect the standard output and standard error to a file to generate a log trace of the command. For example:

```
# synclvodm rich > /tmp/synclvodm.log 2>&1
```

The output in the log file will contain a record of each command executed by the script, along with the arguments used. A small portion of the output is shown below:

```
+ PATH=/usr/bin:/etc:/usr/sbin:/usr/ucb:/usr/bin/X11:/sbin:/usr/bin:/etc:/usr/sb
in:/usr/ucb:/usr/bin/X11:/sbin:/usr/sbin/perf/pmr:.:/usr/samples/kernel:/usr/loc
al/bin
+ EXIT_CODE=1
+ + basename /usr/sbin/synclvodm
PROG=synclvodm
+ + odmget -q attribute=TCB_STATE and deflt=tcb_enabled PdAt
TCB_ON=
+ trap cleanup 0 1 2 15
+ getopt vPDL:FckR rich
+ set -- -- rich
```

You can use this log file to determine the exact point in the high-level command that is failing.

6.4 Problems with importvg

If you experience problems with the `importvg` command when importing a volume group into a system, you should check 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 required a change to the format of the VGDA stored on the disk, and hence, will not be understood by previous levels of AIX.

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, the PVID for `hdisk5` is not shown by the `lspv` command for some reason:

```
# lspv
hdisk0          00000000003bedca    rootvg
hdisk1          00000000da5aaa02    rootvg
hdisk2          00000108b4dff0f9    nimvg
hdisk3          00000259b2faeefe    nimvg
hdisk4          00000108b4d688c5    none
hdisk5          none                none
```

This may be resolved by running the `chdev` command as follows:

```
# chdev -l hdisk5 -a pv=yes
```

This will read the PVID from the disk and place it in the ODM if the disk is accessible. It will only write a new PVID if there truly is no PVID on the disk. Alternatively, you can use the `rmdev` command to remove the disk, for example:

```
# rmdev -l hdisk5 -d
```

Then run the cfgmgr command to configure the disk. If the disk is accessible, it will be configured into the system with the correct PVID, and therefore, the `importvg` command should work.

6.5 JFS problems

As with 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 is imported on an earlier version of AIX. Such features include large file enabled file systems and file systems with non-default allocation group size.

6.5.1 Mounting file systems

Mounting makes file systems, files, directories, devices, and special files available for use at a particular location. It is the only way a file system is made accessible. The `mount` command instructs the operating system to attach a file system at a specified directory.

You can mount a file or directory if you have access to the file or directory being mounted and write permission for the mount point. Members of the system group can also perform device mounts (in which devices or file systems are mounted over directories) and the mounts described in the /etc/filesystems file. A user operating with root user authority can mount a file system arbitrarily by naming both the device and the directory on the command line. The /etc/filestems file is used to define mounts to be automatic at system initialization.

When mounting a file system defined in the /etc/filesystem file, the specified log device is also used if the file system is mounted in read/write mode.

Problems can occur when mounting a file system if it was not previously unmounted correctly as can happen if a system experiences kernel crash. The status of the file system structure may be inconsistent and require repair actions before being mounted.

6.5.2 File system recovery

In the AIX system, the `fsck` command is used to check file system consistency. It can also be used to interactively perform repairs the file system.

File system inconsistencies can stem from the following:

- Stopping the system with file systems mounted.
- Physical disk deterioration or damage. This procedure should be used before mounting any file system.

The `fsck` command checks for the following inconsistencies:

- Blocks or fragments allocated to multiple files.
- i-nodes containing block or fragment numbers that overlap.
- i-nodes containing block or fragment numbers out of range.
- Discrepancies between the number of directory references to a file and the link count of the file.
- Illegally allocated blocks or fragments.
- i-nodes containing block or fragment numbers that are marked free in the disk map.
- i-nodes containing corrupt block or fragment numbers.
- A fragment that is not the last disk address in an i-node. This check does not apply to compressed file systems.
- Files larger than 32 KB containing a fragment. This check does not apply to compressed file systems.
- Size checks:

  -Incorrect number of blocks.
  -Directory size not a multiple of 512 bytes.
  -These checks do not apply to compressed file systems.

- Directory checks:

  -Directory entry containing an i-node number marked free in the i-node map.
  -i-node number out of range.
  -Dot (.) link missing or not pointing to itself.
  -Dot dot (..) link missing or not pointing to the parent directory.
  -Files that are not referenced or directories that are not reachable.

- Inconsistent disk map.
- Inconsistent i-node map.

The `fsck` command does not make changes to a mounted file system. You should unmount the file system before invoking `fsck`.

If the / or /usr file systems require consistency checking, you should only invoke fsck on them from the maintenance shell obtained after performing a Service mode boot of the system from bootable media. Through this route you can access the root volume group and start a shell before mounting file systems. This will then allow you to perform the `fsck` operation on the / and /usr file systems, since at this point they are not mounted.

You can invoke the `fsck` command with the `-y` parameter, which instructs the command to take whatever actions it deems necessary to repair the file system. Only use this option on severely damaged file systems.

6.5.3 JFS log problems

The jfslog device used with JFS file systems can become overloaded in some situations. This can occur when a large amount of JFS I/O is taking place on file systems that share the same log device.

The problem can be avoided by creating an individual jfslog device for each file system that has a high amount of file system structure changes.

The size of the jfslog device should be in proportion to the amount of JFS file system data it is being used to log. A general rule of thumb is to use one PP of jfslog space for every 512 PPs of JFS space. This rule holds true regardless of the PP size.

To create a new jfslog device for an existing file system, use the following procedure:

1. Create a logical volume to be used for the jfslog device. The logical volume must exist in the same volume group as the file system that will use it. The name of the logical volume does not matter, although you may wish to indicate that it is a jfslog device. For example:

   ```
   mklv rootvg loglv01 1
   ```

2. Format the logical volume for use as a jfslog device:

   ```
   # logform /dev/loglv01
   ```

   The command will warn you that the action will destroy all data on the logical volume. Enter `y` to continue.

3. Unmount the file system that will use the new jfslog device.

4. Edit the /etc/filesystems entry for the file system that will use the new jfslog device. Change the log= stanza to reference the new jfslog device. For example:

   ```
   /export:
   ```

```
dev             = /dev/lv00
vfs             = jfs
log             = /dev/loglv01
mount           = true
options         = rw
account         = false
```

5. Mount the file system that will use the new jfslog device. For example:

   ```
   # mount /export
   ```

6. Check that the file system is now using the new jfslog device by examining the output of the command. For example:

   ```
   # mount | grep /export
           /dev/lv00       /export          jfs    Jul 01 17:06 rw,log=/dev/loglv01
   ```

If the jfslog device for a file system becomes corrupted, you may have to reformat the jfslog device. To do this, use the following procedure:

1. Unmount any file systems using the jfslog device.
2. Format the jfslog device. For example:

   ```
   # logform /dev/loglv01
   ```

   The command will warn you that the action will destroy all data on the logical volume. Enter `y` to continue.
3. Mount the file systems that use the jfslog device.

If the jfslog used by the /, /usr, /tmp, and /var file systems, normally called /dev/hd8, becomes corrupted, you will need to perform the `logform` operation from the maintenance shell obtained by booting the machine in Service mode from installable media.

If you need to increase the size of an existing jfslog device, use the following procedure:

1. Unmount any file systems using the jfslog device.
2. Increase the size of the logical volume used for the jfslog. For example:

   ```
   # extendlv loglv01 1
   ```

3. Format the jfslog device. For example:

   ```
   # logform /dev/loglv01
   ```

   The command will warn you that the action will destroy all data on the logical volume. Enter `y` to continue.
4. Mount the file systems that use the jfslog device.

As described previously, if you need to change the size of the jfslog used by the /, /usr, /tmp, and /var file systems, normally called /dev/hd8, you will need to perform the `logform` operation from the maintenance shell obtained by booting the machine in Service mode from installable media.

6.6 Disk replacement aid

AIX, like all operating systems, can be problematic when you have to change a disk. It is, therefore, not advisable to change a faulty disk, restore the data, and just walk away.

Within AIX, you have 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.

The procedures in the following sections provide a number of methods of using AIX commands to remove a disk and its associated ODM entries to enable its replacement. The procedures then incorporate the replacement disk back into the system to the point where data, if any, needs to be restored. Two of the procedures, however, do not involve the use of AIX commands but merely do either a disk copy or a complete restore of an AIX mksysb. One of these six procedures should be capable of providing a solution to 95 percent of disk replacement situations.

6.6.1 Disk replacement procedure summary

The disk you are dealing with is probably in an unstable condition and liable to go totally defective at anytime, possibly before the replacement procedure you are attempting has completed. Because of this possibility, you should only attempt to use the first four methods when you have a full backup of the data.

Method One This method is the least complicated but requires you to have not only the replacement disk but also another spare disk of your own and space within the machine to accommodate it. The spare disk simplifies the procedures used to keep the ODM clean. If dealing with non-hot swap disks, this procedure may require the machine to be powered down and rebooted three times to allow for the disks to be added and removed.
This procedure can also be used if you have only the replacement disk and a space for it be attached to the system. This method will require using only the first half of the procedure and has the possible disadvantage that the hdisk numbering will either be out of sequence or have an additional hdisk added to the ODM.

Method Two This method assumes that there is no physical space for any additional disks to be added to the system at all. You do, however, have to have sufficient spare space on other disks within the failing disks volume group to enable all of the data on the failing disk to be moved. If you are sufficiently proficient in the use of AIX, you possibly could use a combination of spare space on other disks within the volume group and saving files to tape to enable the defective disk to be emptied of data prior to removal.

Method Three If the failing disk contains any logical volumes that are not mirrored, a backup of the data on the disk is required. The backup will be used during the recovery phase, since this method assumes there is no free space within the volume group to move data to. All non-mirrored logical volumes that have data residing on the failing disk will be deleted prior to disk replacement. If the disk being replaced has all of its constituent logical volumes mirrored, this method can be used successfully. It involves removing the mirrored copies of all the logical volumes residing on the disk.

Method Four Method four is most often used when none of the other methods are suitable. This is normally when the defective disk is not usable at all by AIX. The affected volume group is then removed from the ODM to enable entries for the defective disk to be removed from the ODM. When the disk has been replaced, it will be

necessary to recreate the volume group from scratch and completely restore from a previous backup.

Disk-to-Disk Copy Disk-to-disk copy is only available for SCSI disks. This method is best used on disks containing raw logical volumes, striped data, or where the failing disk is within rootvg and the contents of the disk are best copied rather than trying to use any of the recovery methods. See Section 5.3.4.4, “Disk maintenance” on page 105 for a full description.

mksysb restore This is the only solution when the failed disk is either the only disk within rootvg or one of the vital disks within rootvg and not usable by AIX.

6.6.2 Method one

This procedure allows you to perform a fast data restore operation after a disk replacement by using a spare hard disk. In most circumstances, this will shorten the down time of the system and increase customer satisfaction.

This procedure should *not* be used as a replacement for the standard backup procedure. You should ensure that a valid backup exists prior to using this procedure.

The procedure assumes that a spare hard disk exists in addition to the replacement disk and that the spare disk is big enough to hold all data currently residing on the disk to be replaced.

It is assumed that both the spare disk and the replacement disk are either new or newly formatted, and if they are not, that they will be formatted in step 1 or step 13, respectively.

This procedure will restore all logical volumes or parts of logical volumes on the replacement disk, but the logical volumes will not necessarily have the same physical partitions allocated to them as on the original disk.

Make sure no other users are using the system while this procedure is being followed. Complete the following steps:

1. Connect a spare disk to the system. If the disk is not a hot swap disk, you will need to shut the machine down to do this. If it is a SCSI disk, make sure you use a unique valid SCSI address. Use the command `lsdev -C -sscsi` to find out which SCSI addresses are currently being used. Format the disk if necessary.

2. Configure the spare disk into the system. This is done by the configuration manager during Normal mode IPL. If you added a hot swap disk, you should run the `cfgmgr` command manually.
3. Use the `lspv` command to get the name of the spare disk. Also, identify the volume group containing the disk you want to replace. The spare disk will have `none` in the volume group membership column, and if it is a new or newly formatted disk, it will also have `none` in the physical volume ID column. For example:

```
# lspv
hdisk0          00000036960cbdd1     rootvg
hdisk1          00000036960c31de     rootvg
hdisk2          00000036960d3007     myvg
hdisk3          none                 none
```

4. Make the spare disk a physical volume. For example:

```
# chdev -l hdisk3 -a pv=yes
```

   Where `hdisk3` is the name of the spare disk.

5. Add the spare disk to the volume group from which you want to replace another disk. For example:

```
# extendvg -f rootvg hdisk3
```

6. Identify which logical volumes reside on the disk you want to replace. Make a note of their names, listed in the `LV ID` column, and their types. For example:

```
# lspv -p hdisk0
```

7. What happens now depends on whether the disk you want to replace contains the boot logical volume (hd5) or not. Check the output from the command run in step 6 to check if the disk contains the boot logical volume.

> **Note**
>
> It is assumed that the name of the boot logical volume is hd5, as originally created during the operating system installation. If it is not, the commands listed in the following steps will have to be modified to reflect the different name used.

   If the disk to be replaced does not belong to rootvg, or does not contain the boot logical volume (hd5), go to step 8.

   If the disk to be replaced belongs to rootvg and contains the boot logical volume (hd5), do the following:

a. Migrate the boot logical volume. For example:

```
# migratepv -l hd5 hdisk0 hdisk3
```

b. Verify that the physical partitions that have been allocated to the migrated boot logical volume are contiguous with the following command:

```
# lslv -m hd5
```

c. Execute the following commands to recreate the boot image, update the boot record on the spare disk, update the IPL list for Normal mode boot in NVRAM, save information about base customized devices in the ODM onto the boot device, and erase the old boot image from the disk you want to replace. For example:

```
# bosboot -d /dev/hdisk3 -a -u -l /dev/hd5
# savebase -d /dev/hdisk3
# mkboot -c -d /dev/hdisk0
```

d. Continue with step 8.

8. What you do now depends on whether the disk you want to replace contains the primary system dump logical volume or not. Use the `sysdumpdev` command to determine which logical volume is the primary dump logical volume. Check the output from the command run in step 6 to see if the disk contains part of the system dump logical volume.

   If the disk to be replaced does not contain the primary system dump logical volume, or any part of it, do the following:

   a. Move all logical volumes from the disk you want to replace to the spare disk. For example:

      ```
      # migratepv hdisk0 hdisk3
      ```

      This command will take time to complete. It is *very* important to *not* interrupt the command while it is running.

   b. Continue to step 9.

   If the disk to be replaced belongs to rootvg and contains the primary system dump logical volume, or any part of it, do the following:

   a. Make a system dump logical volume on the spare disk. For example:

      ```
      # mklv -y newdumplv -t sysdump rootvg 2 hdisk3
      ```

   b. Temporarily change the primary dump device to `newdumplv` using the `sysdumpdev` command. For example:

      ```
      # sysdumpdev -p /dev/newdumplv
      ```

c. Move all logical volumes from the disk you want to replace to the spare disk. For example:

```
# migratepv hdisk0 hdisk3
```

This command will take time to complete. It is *very* important to *not* interrupt the command while it is running.

d. Change the primary system dump device back to the logical volume identified in step 8. For example:

```
# sysdumpdev -p /dev/hd7
```

e. Remove the temporary dump logical volume `newdumplv` using the `rmlv` command. For example:

```
# rmlv -f newdumplv
```

f. Continue with step 9.

9. Check that the disk you want to replace does not contain any physical partitions that are allocated to any logical volumes. For example:

```
# lsvg -p rootvg
```

The output from the command should confirm that `TOTAL PPs` and `FREE PPs` are identical for the disk you want to replace. For example:

```
PV_NAME           PV STATE    TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk0            active      76          76          16..15..15..15..15
```

10. Reduce the volume group by removing the disk you want to replace from its volume group. For example:

```
# reducevg rootvg hdisk0
```

11. Remove the disk as a device from the system and from the ODM database using the `rmdev` command. For example:

```
# rmdev -l hdisk0 -d
```

12. Shut the system down.

13. Power the system off, remove the old disk, and replace it with a new disk. Make sure the new disk has the same SCSI address as the old one and that any terminators on the disk are removed. Format the new disk if necessary.

14. Go to step 2 and repeat the same procedure up to and including step 12 for migrating logical volumes from the spare disk to the new disk, unconfiguring the spare disk, and removing it from the system. After this procedure is completed, the system will appear as it did before the disk failed. In other words, the hdisk numbers will be identical.

6.6.3 Method two

This procedure allows you to move data from the failing disk to other disks in the same volume group and save time in restoring the data from a backup tape. In most circumstances, this procedure will shorten the down time of the system and, therefore, increase customer satisfaction.

This procedure *should not* be used as a replacement for the standard backup procedure. It is the responsibility of the customer to ensure that a valid backup exists prior to the CE using this procedure.

It is assumed that the replacement disk is either new or newly formatted, and if it is not, that it will be formatted in step 12.

This procedure will not restore logical volumes to their original places on the replacement disk. You should decide which, if any, logical volumes should be moved back to the replacement disk and perform that operation.

Make sure no other users are using the system while this procedure is being followed. Complete the following steps:

1. Identify which disks belong to which volume groups in the system by using the `lspv` command.

   The output from the command may look something like the following:

   ```
   hdisk0          00000036960cbdd1    rootvg
   hdisk1          00000036960c31de    rootvg
   hdisk2          00000036960d3007    vg01
   hdisk3          000003870001328f    vg01
   hdisk4          00000360ebf34660    vg01
   hdisk5          00000360d7c1f19f    vg01
   hdisk6          00000036628b9724    vg02
   ```

   This indicates that `hdisk0` and `hdisk1` belong to `rootvg`, `hdisk2`, `hdisk3`, `hdisk4`, and `hdisk5` belongs to `vg01`, and `hdisk6` belongs to `vg02`.

2. Identify the number of used physical partitions on the disk you want to replace. For example:

   ```
   # lspv  hdisk0 | grep "USED PPs"
   ```

3. Identify the number of free physical partitions on all other disks belonging to the same volume group as the disk you want to replace. For example:

   ```
   # lspv hdisk1 | grep "FREE PPs"
   ```

   Repeat this command for all disks belonging to the volume group.

4. Check that the number of Used Physical Partitions (UPP) on the disk you want to replace is smaller than the Sum of Free Physical Partitions (SFPP)

on the remaining disks in the same volume group. If UPP is larger than SFPP, this procedure can *not* be used to replace the disk. In this case, you should select an alternative procedure.

5. Identify the names of all logical volumes that fully or partially reside on the disk you want to replace. This step involves the use of multiple commands. Use the `lsvg` command to determine which logical volumes belong to the volume group and which of them span multiple disks. For example:

   ```
   # lsvg -l vg01
   ```

   The command may produce output similar to the following:

   ```
   vg01:
   LV NAME             TYPE       LPs PPs PVs  LV STATE      MOUNT POINT
   loglv01             jfslog     1   1   1    open/syncd    N/A
   lv04                jfs        31  31  1    open/syncd    /vol/x11r4
   lv27                jfs        53  53  1    open/syncd    /vol/X11R5
   lv08                jfs        26  26  1    open/syncd    /vol/tex
   lv10                jfs        63  63  1    open/syncd    /vol/abc
   lv21                jfs        84  84  2    open/syncd    /var/spool/news
   lv12                jfs        99  99  1    open/syncd    /vol/lpp/3005
   lv23                jfs        66  66  2    open/syncd    /vol/src
   lv07                jfs        92  92  2    open/syncd    /vol/misc
   ```

   This means that of all logical volumes in the `vg01` volume group only three of them (`lv21`, `lv23`, and `lv07`) span two disks (`PVs`), and that they are `84`, `66`, and `92` physical partitions in size respectively.

   Use the `lspv` command to tell you which logical volumes fully or partially reside on the disk you want to replace. Make a note of their names and mark those that only partially reside on the disk. For example:

   ```
   # lspv -l hdisk3
   ```

   The command may produce output similar to the following:

   ```
   hdisk3:
   LV NAME               LPs   PPs   DISTRIBUTION          MOUNT POINT
   lv08                  26    26    02..10..10..04..00    /vol/tex
   lv21                  16    16    02..00..12..00..02    /var/spool/news
   lv04                  31    31    01..08..05..15..02    /vol/x11r4
   ```

   If you compare this output with the output from `lsvg -l vg01`, you can conclude that `lv08` and `lv04` logical volumes reside entirely on the disk `hdisk3`, and that `lv21` only partially resides on the disk `hdisk3`, since only 16 out of 84 physical partitions for `lv21` are placed on `hdisk3`. In the subsequent steps, `lv08`, `lv04`, and 16 physical partitions belonging to `lv21` will be migrated from `hdisk3` to other disks in the same volume group.

6. What happens now depends on whether the disk you want to replace contains the boot logical volume (hd5) or not. See the output from the command run in step 5 to check if the disk contains the boot logical volume.

> **Note**
>
> It is assumed that the name of the boot logical volume is hd5, as originally created during the operating system installation. If it is not, the commands listed in the following steps will have to be modified to reflect the different name used.

   If the disk to be replaced does not belong to rootvg, or does not contain the boot logical volume (hd5), continue to step 7.

   If the disk to be replaced belongs to rootvg and contains the boot logical volume (hd5), do the following:

   a. Find a disk belonging to the same volume group that has two contiguous free physical partitions. Use the `lspv` command to find out which physical partitions on the disk are free. For example:

      ```
      # lspv -p hdisk3
      ```

      Migrate the boot logical volume. For example:

      ```
      # migratepv -l hd5 hdisk0 hdisk3
      ```

   b. Verify that the physical partitions that have been allocated to the migrated boot logical volume are contiguous with the following command:

      ```
      # lslv -m hd5
      ```

   c. Execute the following commands to recreate the boot image, update the boot record on the spare disk, update the IPL list for Normal mode boot in NVRAM, save information about base customized devices in the ODM onto the boot device, and erase the old boot image from the disk you want to replace. For example:

      ```
      # bosboot -d /dev/hdisk3 -a -u -l /dev/hd5
      # savebase -d /dev/hdisk3
      # mkboot -c -d /dev/hdisk0
      ```

   d. Continue with step 7.

7. What you do now depends on whether the disk you want to replace contains the primary system dump logical volume or not. Use the `sysdumpdev` command to determine which logical volume is the primary

dump logical volume. Check the output from the command run in step 5 to see if the disk contains part of the system dump logical volume.

If the disk to be replaced does not contain the primary system dump logical volume, or any part of it, do the following:

a. Move all logical volumes from the disk you want to replace to other disks in the same volume group that have enough free physical partitions. You can select some, or all, of the remaining disks provided the criterion described in step 4 is met. For example:

   ```
   # migratepv hdisk0 hdisk1
   ```

   This command will take time to complete. It is *very* important to *not* interrupt the command while it is running.

b. Continue with step 8.

If the disk to be replaced belongs to rootvg and contains the primary system dump logical volume, or any part of it, do the following:

a. Make a system dump logical volume on one of the remaining disks in the same volume group that has at least two free physical partitions. For example:

   ```
   # mklv -y newdumplv -t sysdump rootvg 2 hdisk1
   ```

b. Temporarily change the primary dump device to `newdumplv` using the `sysdumpdev` command. For example:

   ```
   # sysdumpdev -p /dev/newdumplv
   ```

c. Move all logical volumes from the disk you want to replace to other disks in the same volume group that have enough free physical partitions. You can select some, or all, of the remaining disks provided the criterion described in step 4 is met. For example:

   ```
   # migratepv hdisk3 hdisk4 hdisk5
   ```

   This command will take time to complete. It is *very* important to *not* interrupt the command while it is running.

d. Change the primary system dump device back to the primary device identified above. For example:

   ```
   # sysdumpdev -p /dev/hd7
   ```

e. Remove the temporary dump logical volume `newdumplv` with the `rmlv` command. For example:

   ```
   # rmlv -f newdumplv
   ```

f. Continue with step 8.

8. Check that the disk you want to replace contains no physical partitions that are allocated to any logical volumes. For example:

   ```
   # lsvg -p rootvg
   ```

   The output from the command should confirm that the `TOTAL PPs` and `FREE PPs` for the disk you want to replace are identical. For example:

   ```
   PV_NAME           PV STATE    TOTAL PPs   FREE PPs    FREE DISTRIBUTION
   hdisk0            active      159         159         32..32..31..32..32
   ```

9. Reduce the volume group by removing the disk you want to replace from its volume group. For example:

   ```
   # reducevg rootvg hdisk0
   ```

10. Remove the disk as a device from the system and from the ODM database using the `rmdev` command. For example:

    ```
    # rmdev -l hdisk0 -d
    ```

11. Shut the system down.

12. Power the system off, remove the old disk, and replace it with a new disk. Make sure the new disk has the same SCSI address as the old one, and that any terminators on the disk are removed. Format the disk if necessary.

13. Configure the new disk into the system (done by the configuration manager cfgmgr during Normal mode IPL).

14. Confirm that the new disk has the same name as the disk that was replaced.

15. Make the new disk a physical volume with the `chdev` command. For example:

    ```
    # chdev -l hdisk0 -a pv=yes
    ```

Once the procedure is completed, you have a number of options available for using the replacement disk:

- Leave the new disk empty, and use it later on when needed.
- Add the disk to the original volume group, and move some or all of the logical volumes that originally resided on the replaced disk, or any other logical volumes in that volume group, to the new disk.

As a general guideline, you may wish to move only those logical volumes that fully resided on the replaced disk. If you decide to move logical volumes that partially resided on the replaced disk as well, there is a chance that not only the physical partitions that were on the replaced disk, but all physical

partitions belonging to these logical volumes, will be moved to the new disk. In some cases, the migration process might fail due to lack of disk space.

6.6.4 Method three

This procedure describes the proper way of unconfiguring and replacing a hard disk. It should be used when neither method one or method two as described above can be applied due to lack of free disk space or spare disk availability, or because these two procedures are simply not applicable.

This procedure ensures that the system will be left in a sound state after the disk is replaced, that is, no ODM database or VGDA corruptions will occur.

The procedure assumes that a valid backup of the data residing on the disk exists. That backup *will be used* to restore the data once the disk is unconfigured and replaced.

It is assumed that the replacement disk is either new or newly formatted, and if it is not, that it will be formatted in step 12.

Make sure no other users are using the system while this procedure is being followed. Complete the following steps:

1. Identify to which volume group the disk you want to replace belongs using the `lspv` command. For example, the output from the command may look like the following:

   ```
   hdisk0          00000036960cbdd1    rootvg
   hdisk1          00000036960c31de    rootvg
   hdisk2          00000036960d3007    vg01
   hdisk3          000003870001328f    vg01
   hdisk4          00000360ebf34660    vg01
   hdisk5          00000360d7c1f19f    vg01
   hdisk6          00000036628b9724    vg02
   ```

   This indicates that `hdisk0` and `hdisk1` belong to `rootvg`; `hdisk2`, `hdisk3`, `hdisk4`, and `hdisk5` belong to `vg01`; and `hdisk6` belongs to `vg02`.

2. Identify the names of all logical volumes that fully or partially reside on the disk you want to replace. This step involves the use of multiple commands. Use the `lsvg` command to determine which logical volumes belong to the volume group, which of them span multiple disks, and which of them are mirrored. For example:

   ```
   # lsvg -l vg01
   ```

   The command may produce output similar to the following:

   ```
   vg01:
   ```

```
LV NAME             TYPE       LPs PPs PVs  LV STATE      MOUNT POINT
loglv01             jfslog     1   1   1    open/syncd    N/A
lv04                jfs        31  31  1    open/syncd    /vol/x11r4
lv27                jfs        53  53  1    open/syncd    /vol/X11R5
lv08                jfs        26  26  1    open/syncd    /vol/tex
lv10                jfs        63  63  1    open/syncd    /vol/abc
lv21                jfs        84  84  2    open/syncd    /var/spool/news
lv12                jfs        99  99  1    open/syncd    /vol/lpp/3005
lv23                jfs        66  66  2    open/syncd    /vol/src
lv07                jfs        92  92  2    open/syncd    /vol/misc
```

This means that of all logical volumes in the `vg01` volume group only three of them (`lv21`, `lv23`, and `lv07`) span two disks (PVs), and that they are `84`, `66`, and `92` physical partitions in size, respectively. In this example, no logical volumes are mirrored since the number of LPs and PPs is always the same. When the number of PPs is twice or three times bigger, the number of LPs then a logical volume is singly or doubly mirrored.

If the disk you want to replace is the only disk in its volume group, this command will simply tell you which logical volumes reside on the disk you want to replace. In this case, you need not run any other commands described in this step.

Use the `lspv` command to tell you which logical volumes fully or partially reside on the disk you want to replace. Make a note of their names and mark those that only partially reside on the disk, or have mirror copies on it. For example:

```
# lspv -l hdisk3
```

The command may produce output similar to the following:

```
hdisk3:
LV NAME             LPs   PPs   DISTRIBUTION          MOUNT POINT
lv08                26    26    02..10..10..04..00    /vol/tex
lv21                16    16    02..00..12..00..02    /var/spool/news
lv04                31    31    01..08..05..15..02    /vol/x11r4
```

If you compare this output with the output from `lsvg -l vg01`, you can conclude that `lv08` and `lv04` logical volumes reside entirely on the disk `hdisk3`, and that `lv21` only partially resides on the disk `hdisk3`, since only 16 out of 84 physical partitions for `lv21` are placed on `hdisk3`. In the subsequent steps, lv08, lv04, and 16 physical partitions belonging to lv21 will be migrated from hdisk3 to other disks in the same volume group.

3. Check that you have a backup of all file systems (logical volumes of jfs type) that either fully or partially reside on the hard disk that will be replaced, or have mirror copies on it. If not, perform such a backup by using your favorite command, for example `backup`, `tar`, or `cpio`. You should

also ensure that you have a backup of all data on raw logical volumes. For example, database packages, such as Oracle, Informix, and Sybase, can be configured to use data on logical volumes that are not of jfs type.

4. Unmount all single-copy (non-mirrored) file systems that either fully or partially reside on the hard disk that is to be replaced. For example:

   ```
   # umount /vol/tex
   ```

5. Delete all single-copy file systems from the disk that is to be replaced. For example:

   ```
   # rmfs /vol/tex
   ```

6. Remove physical partition copies of mirrored logical volumes from the disk that will be replaced. For example:

   ```
   # rmlvcopy lv100 1 hdisk3
   ```

7. Close all other logical volumes residing on the disk to be replaced.

   If the disk contains a raw logical volume with data on it, it will probably be closed when the application that is using it is stopped. The customer should know how to close such a logical volume.

   If the disk contains a paging space, it will have to be inactivated after the next system restart. The following steps will achieve this:

   a. Change the paging space so it will not be used at the next system restart. For example:

      ```
      chps -a n pagingXX
      ```

      Where `pagingXX` is the name of the logical volume containing the paging space.

   b. Shut the system down and reboot the machine.

   c. Confirm that `pagingXX` is not active after the reboot with the following command:

      ```
      # lsps -a
      ```

   If the disk contains a non-mirrored journaled file system log (jfslog), and it is *not* the only disk in its volume group, you will have to migrate the journaled file system log to another disk in the volume group. This will allow file systems residing on other disks in the volume group to remain mounted and available. The following command will achieve this:

   ```
   # migratepv -l jfslogname pvname1 pvname2
   ```

   Where `jfslogname` is the name of the log logical volume, `pvname1` is the name of the disk to be replaced, and `pvname2` is the name of the target disk that has a free physical partition to accept the jfslog.

8. Reduce the volume group by removing the disk you want to replace from its volume group. If the disk you want to replace is the only disk in its volume group, the volume group will also be deleted. For example:

   ```
   # reducevg -df vg01 hdisk3
   ```

9. Check that the disk you want to replace does not belong to any volume group by using the `lspv` command.
10. Remove the disk as a device from the system and from the ODM database with the `rmdev` command. For example:

    ```
    #rmdev -l hdisk3 -d
    ```

11. Shut the system down.
12. Power the system off, remove the old disk, and replace it with a new disk. Make sure the new disk has the same SCSI address as the old one, and that any terminators on the new disk are removed. Format the disk if necessary.
13. Configure the new disk into the system.
14. Confirm that the new disk has the same name as the disk that was replaced with the `lspv` command.
15. Make the new disk a physical volume with the `chdev` command. For example:

    ```
    # chdev -l hdisk3 -a pv=yes
    ```

16. Depending on whether the disk that was replaced was the only disk in its volume group or not, run one of the following:
    - If the disk was the only disk in its volume group, create the volume group again, and activate it:

      ```
      # mkvg -y vgname newpvname
      # varyonvg vgname
      ```

      Where `vgname` is the name of the volume group, and `newpvname` is the name of the new disk.
    - If the disk was one of several disks in its volume group, add the new disk to the same volume group from which the old one was removed:

      ```
      # extendvg -f vgname newpvname
      ```

      Where `vgname` is the name of the volume group, and `newpvname` is the name of the new disk.
17. Recreate the logical volumes and file systems removed in the previous steps, and then restore the data from backup tape(s) or diskettes into the file systems and raw logical volumes if there were any.

If a mirror copy was removed in step 6, it can be recreated now by using SMIT, and then synchronizing the mirror copies by running the following command:

```
# syncvg -p newpvname
```

Where `newpvname` is the name of the new hard disk.

6.6.5 Method four

This procedure describes the proper way of unconfiguring and replacing a hard disk when it belongs to a non-rootvg volume group and the disk or data on it cannot be accessed.

This procedure ensures that the system will be left in a sound state after the disk is replaced, that is, no ODM database or VGDA corruptions will occur, and the disks will have consecutive numbers.

This procedure assumes that a valid backup of the data residing on *all* disks in the volume group exists. This backup *will be used* to restore the data once the disk is unconfigured and replaced, and the volume group is recreated.

It is assumed that the replacement disk is either new or newly formatted, and if it is not, that it will be formatted in step 12.

Make sure no other users are using the system while this procedure is being followed. Complete the following steps:

1. If the system is still up and running, identify volume group membership for all disks in the system by using the `lspv` command. If the system is not running, ask the customer to provide this information.

   For example, the output from the command may look like the following:

   ```
   hdisk0          00000036960cbdd1    rootvg
   hdisk1          00000036960c31de    rootvg
   hdisk2          00000036960d3007    vg01
   hdisk3          000003870001328f    vg01
   hdisk5          00000360d7c1f19f    vg01
   hdisk6          00000036628b9724    vg02
   ```

   This indicates that `hdisk0` and `hdisk1` belong to `rootvg`; `hdisk2`, `hdisk3`, `hdisk4`, and `hdisk5` belong to `vg01`; and `hdisk6` belongs to `vg02`.

2. Shut the system down.
3. Power the system off. Disconnect the disk that has failed. Disconnect all other disks belonging to the same volume group, if any.

Leave everything else unchanged. For example, make sure the remaining disks still have the same SCSI addresses as before.

4. Power the system on.
5. Confirm the failed disk and other disks you disconnected is step 3 are in the DEFINED state, rather than AVAILABLE by using the following command:

   ```
   # lsdev -Cc disk
   ```

6. Confirm the volume group that contains the failed disk that was disconnected in step 3 is still known to the system by using the `lsvg` command.
7. Export the volume group that contains the failed disk using the following command:

   ```
   # exportvg vgname
   ```

   Where `vgname` is the name of the volume group. This will remove all references to the volume group and all logical volumes and file systems belonging to it from the ODM database, /dev directory, and /etc/filesystems file.
8. Check that the /etc/filesystems file does not have any stanzas left for the file systems that belonged to the exported volume group. If there are any stanzas left, delete them by using your favorite editor.
9. Remove the failed disk as a device from the system and from the ODM database using the `rmdev` command:

   ```
   # rmdev -l oldpvname -d
   ```

   Where `oldpvname` is the name of the failed disk.
10. Confirm the failed disk is not known to the system any more using the `lsdev` command:

    ```
    # lsdev -Cc disk
    ```

    The output should not list the failed disk.
11. Shut the system down.
12. Power the system off and replace the failed disk with a new disk. Make sure the new disk has the same SCSI address as the old one, and that any terminators on the new disk are removed. Format the disk if necessary. Reconnect all other disks belonging to the same volume group that were disconnected in step 3. Make sure you do not change the position of the disks on the SCSI chain or their SCSI addresses.
13. Configure the new disk into the system. This is done by the configuration manager cfgmgr during Normal mode IPL.

14. Confirm that the new disk has the same name as the disk that was replaced by using the `lspv` command.

15. Make the new disk a physical volume:

```
# chdev -l newpvname -a pv=yes
```

    Where `newpvname` is the name of the new disk.

16. Depending on whether the disk that was replaced was the only disk in its volume group or not, run one of the following:

    - If the disk was the only disk in its volume group, recreate the volume group on the replacement disk, and activate it:

```
# mkvg -y vgname newpvname
# varyonvg vgname
```

      Where `vgname` is the name of the volume group, and `newpvname` is the name of the new disk.

    - If the disk was one of several disks in its volume group, recreate the volume group on the replacement disk and all other disks that used to belong to the volume group and activate the volume group:

```
# mkvg -y vgname newpvname pvname2 pvname3 ...
# varyonvg vgname
```

      Where `vgname` is the name of the volume group, `newpvname` is the name of the new disk, and `pvname2`, `pvname3`, and so on are the names of other disks that used to belong to the same volume group.

> **Note**
>
> Since the disks that previously belonged to the volume group still contain data, you will be warned that the data on them will be destroyed. You will be asked if you want to continue, to which you should say yes provided you have confirmed that you have a good backup.

17. You should now recreate the logical volumes and file systems in the volume group, and then restore the data from backup tape(s) or diskettes.

6.6.6 Disk-to-disk copy

This method can only be employed to copy SCSI disks. The disk to be copied to (target disk) must *not* be a smaller capacity than the disk to be copied from (source disk) or *more* than 10 percent larger capacity than the source disk. Complete the following steps:

1. Power the machine off.

2. Connect the target disk to a SCSI bus using a free SCSI ID.
3. Boot the machine using a Diagnostic CD. Refer to Section 5.3.3, “Stand-alone diagnostics from CD or diskette” on page 99 for instructions.
4. At the Diagnostics screen select the **Task Selection** menu option, or the **Service Aids** option if using an AIX Diagnostics CD older than Version 4.1.5.
5. Select **Disk Maintenance**.
6. Select **Disk to Disk copy**.
7. Enter the addresses of the target and source disks. Care should be taken not to get the addresses confused. A mistake at this point will damage data irreparably.
8. Start the copying.
9. When the copy has completed, power off the machine and immediately remove the defective disk from the machine. Both the disks carry the same PVID, which would cause problems if you started AIX with both disks left in the machine. Since both disks carry the same PVID, you might also forget which is the defective disk.

Note

This procedure will only complete successfully if the defective disk passes diagnostics with minimal errors. If the error rate is too high, the service aid will terminate.

This process is not suitable for use on 7020 model 40P and 7248 model 43P machines, since these machines are not capable of loading AIX diagnostics.

6.6.7 mksysb restore

Doing a restore of a mksysb (image backup of rootvg) requires a bootable mksysb tape or an AIX install CD and a non-bootable mksysb.

6.6.7.1 MCA machines

Booting from a mksysb tape or AIX install CD requires the machine to perform a Service mode boot. Complete the following steps:

1. Power off the machine.
2. Turn the key to Service.
3. Power on the machine.

4. Immediately place bootable media in the CD or tape drive.
5. Answer any screen prompts.
6. When the Installation Screen appears select **Start Maintenance Mode for System Recovery** (option 3).
7. Select **Install from a System Backup** (option 4).

6.6.7.2 All PCI machines

To start a Service mode boot:

1. Power off the machine.
2. Turn on machine power.
3. Place the bootable media in the CD or tape 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 console. If you are using the graphics console, they will sometimes have power saving on them and so take time to restart. This can lead you to miss the symbols being displayed. In this situation, observe the power LED on the display monitor, and when it changes from orange to green, press the **F5** key.
   After doing the above, you will get various screens displayed, one of which will indicate to you the SCSI address of the device that it is booting from.
5. Answer all of the screen prompts.
6. When the Installation and Screen appears select **Start Maintenance Mode for System Recovery** (option 3).
7. Select **Install from a System Backup** (option 4).

Chapter 7. TCP/IP networking problems

This chapter deals with network problem source identification and resolution on different types of networks. The following problem determination assumes that the underlying network has been wired in accordance with the applicable networking standards.

7.1 General network problem isolation

Prior to beginning any network debugging, a series of questions need to be asked in order to isolate the problem:

- Is the problem with your own machine?
  - The physical network?
  - The network device?
  - Routing tables?
- Is the problem with the target host?
- Is the problem with name resolution?
- Is the problem with a router?

7.2 Problem isolation steps for TCP/IP network problems

The problems listed here are the most commonly encountered. Sometimes, it is very difficult to discover exactly where the problem lies – within your machine or outside of it.

To determine the problem, you must use status tools and commands and a process of elimination. If none of the following steps lead to resolution of the problem, you will need to open a PMR, gather the required testcase documentation, and send this in for analysis.

If you are unable to connect to any host, or if you have a general network problem, go to Section 7.2.2, “No network access” on page 156.

If the problem is only with specific hosts, go to Section 7.2.1, “Selective host network problems” on page 156.

7.2.1 Selective host network problems

If networking is generally working, but you are unable to access a particular host, try to `ping` by IP address and host name. Refer to Section 7.2.2, “No network access” on page 156 if necessary.

If this is a name resolution problem, refer to Section 7.2.3, “Name resolution problems” on page 156 to verify the correct setup.

Try the `traceroute` command to see the attempted route to the host:

```
traceroute <hostname>
```

The `traceroute` 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 going astray.

7.2.2 No network access

This could mean you have a problem with your specific host or the network.

First, attempt to `ping` a known host name. See Section 7.4.4, “ping” on page 176 for examples if necessary. For example:

```
ping startrek
```

If the `ping` is successful, go back to Section 7.2.1, “Selective host network problems” on page 156. If not, try using the IP address for the host instead of the name. For example:

```
ping 9.3.45.89
```

If you can `ping` by IP address but not by name, you may have a name resolution problem. Go to Section 7.2.3, “Name resolution problems” on page 156.

If you cannot `ping` an external host by IP address or by name, you may have a routing problem. Go to Section 7.2.4, “Routing problem debugging” on page 157.

7.2.3 Name resolution problems

Resolver routines on hosts running TCP/IP are used to attempt to resolve a host name to an IP address.

If you think you have a name resolution problem, first, check the ordering of name resolution and determine if you are using a name server.

The default order for name resolution in AIX V4 is:

1. Domain Name Server (DNS)
2. Network Information Service (NIS)
3. Local /etc/hosts file

The order can be changed in the /etc/netsvc.conf file or by using the NSORDER environment variable; check these for your particular name resolution ordering.

If the file /etc/resolv.conf contains an entry for a name server, you are using a name server.

Try to `ping` the address given for the name server. If you can `ping` it, then it is up and reachable.

If the /etc/resolv.conf file exists but has no entries, remove the file. Then, retry the name resolution with ping.

If the local name server is up, check with the network administrator and verify that the named daemon on your local name server is active with the `lssrc -s named` command (if using an AIX name server). Also check that the target host is correctly identified in the name server configuration database.

If you are using NIS, go to Section 7.5, “NIS troubleshooting” on page 184.

You can tell if you are using Network Information Services (NIS) by using the `ps -ef` command and looking for the ypserv and ypbind processes. Use the `ypwhich` command to display the name of the NIS server being used.

If you are resolving some or all host names locally, check the /etc/hosts file for the correct target host name and IP address.

Name resolution is working if you can translate host names to addresses and addresses to host names.

7.2.4 Routing problem debugging

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:

- Use the `netstat -rn` command to show you the content of your local routing table using IP addresses.
- Check the netmask displayed and ensure that it is correct (ask the network administrator what it should be if you are unsure).
- If there is a default route, attempt to `ping` it.
- If you have more than one network interface, attempt to determine if any interfaces are working.

If you cannot `ping` your default route, either it is down, or your local network connection may be down. Attempt to `ping` all of the other gateways listed in the routing table to see if any portion of your network is functioning:

```
# netstat -rn
Routing tables
Destination      Gateway           Flags   Refs     Use  If   PMTU  Exp  Groups

Route Tree for Protocol Family 2 (Internet):
default          9.185.112.254     UG        1     1656  tr0     -   -
9.185.112/23     9.185.113.7       U        26     4642  tr0     -   -
127/8            127.0.0.1         U         5      424  lo0     -   -
192.168.1/24     192.168.1.10      U         0        0  en0     -   -

Route Tree for Protocol Family 24 (Internet v6):
::1              ::1               UH        0        0  lo0 16896   -
#
```

If you cannot `ping` any host or router interface from among those listed in the routing table, try to `ping` your local loopback interface (lo0) with the following command:

```
ping localhost
```

If the `ping` is successful you have either an adapter or network hardware problem or a routing problem. Continue in this section for routing problem determination or go to Section 7.2.6, “Network interface problems” on page 161.

If the `ping` is not successful, you need to:

- Ensure that the inetd process is active using the `lssrc -g tcpip` command. If inetd is not active issue the `startsrc -s inetd` or `startsrc -g tcpip` commands.
- 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:

  ```
  127.0.0.1 loopback localhost # loopback (lo0) name/address
  ```

An asterix (*) 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
```

7.2.5 Dynamic or static routing

Are you using dynamic or static routing?

You are using dynamic routing if you see either the gated or routed program running when you execute the `ps -ef` command.

In dynamic routing, the gated program runs on a server that broadcasts routing information. The routed program runs on clients. It listens to the broadcast and updates the routing information on the client.

If you are using dynamic routing, verify that the gateway is listed and correct in the kernel routing tables by issuing the `netstat -r` command.

If you are using dynamic routing with the routed daemon:

- 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.
- 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.
- Run the routed daemon using the debug option to log such information as bad packets received. Invoke the daemon from the command line using the following command:

  ```
  startsrc -s routed -a "-d"
  ```

- 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 you are using dynamic routing with the gated daemon:

- Verify that the /etc/gated.conf file is configured correctly and that you are running the correct protocols.
- Make sure the gateway on the source network is using the same protocol as the gateway on the destination network.

- Make sure that the machine with which you are trying to communicate has a return route back to your host machine.
- Verify that the gateway names in the gated.conf file correspond to the gateway names listed in the /etc/networks file.

If you are using the RIP or HELLO protocols, and routes to the destination cannot be identified through routing queries, check the gated.conf file to verify that a route to the target host is defined. You should set static routes under either of 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.

Refer to the gated information in the AIX Version 4 Files Reference on how to modify the gated.conf file if you do wish to add static routes. Also note that AIX Versions 4.3.2 and later run gated Version 3.5.9. The syntax of the gated.conf file has changed slightly from earlier versions. Read the gated.conf documentation or use the sample file that is shipped in the /usr/samples/tcpip directory for the correct syntax.

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 demons. This is a last case resort, since any networking that was functioning before will be temporarily cut off once the routes are removed. Be sure no other users will be interrupted by this.

If you are not using dynamic routing, you are probably using default static routing, which may have some direct routes set up, but all other routes are directed to a gateway machine that handles forwarding of the data. In this case, you can manually add and delete routes. If there are routes in the routing table that look obviously wrong, delete them with the `route` command as shown:

```
route delete <target network> <route address>
```

Routes can be added the same way:

```
route add net <target network> <route address>
```

For example, to add a default route:

```
# route add default 129.35.128.1
```

To delete a network route:

```
# route delete net 9.3.199 9.3.189.45
```

7.2.6 Network interface problems

If you still cannot establish communications, the following sections contain tips on debugging specific network interface types.

7.2.6.1 General interface debugging

This section applies to all TCP/IP interface types and should be checked prior to using the following interface specific sections.

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 should be to check the installed adapter types and states using the `lsdev -Cc adapter` and `lsdev -Cc if` commands.

If all adapters and interfaces you are using are listed and available, use the `netstat -i` command and check the output for `Ierrs` (input errors), `Oerrs` (output errors), and `Coll` (collisions).

Also check the `Ipkts` and `Opkts` (input and output packets) to see if there has been any successful network traffic since the last reboot. For example:

```
# netstat -in
Name  Mtu   Network     Address              Ipkts Ierrs    Opkts Oerrs  Coll
lo0   16896 link#1                           2107     0     2109     0     0
lo0   16896 127         127.0.0.1            2107     0     2109     0     0
lo0   16896 ::1                             2107     0     2109     0     0
en0*  1500  link#2      2.60.8c.2e.e1.c9        0     0        9     0     0
en0*  1500  192.168.1   192.168.1.10            0     0        9     0     0
tr0   1492  link#3      10.0.5a.a8.70.67   1102168     0    84137   406     0
tr0   1492  9.185.112   9.185.113.7        1102168     0    84137   406     0
at0   9180  link#4      8.0.5a.75.20.f6       2788     0     2661     0     0
at0   9180  192.168.84  192.15.15.2           2788     0     2661     0     0
#
```

Check that all listed interfaces have unique network addresses, and if these are correct, use the `ifconfig` command to check the state of any interface that has an asterisk next to the interface name. In the preceding example, the `ifconfig en0` command would show `en0` in a down state.

Verify the interface state with the `ifconfig <interface>` command, and if it is shown as down, detach it, bring it up, and verify its status again, as shown in the following terminal session example:

```
# ifconfig en0
en0: flags=e080862<BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT>
```

```
        inet 192.168.1.10 netmask 0xffffff00 broadcast 192.168.1.255
# ifconfig en0 detach
# ifconfig en0 inet 192.168.1.10 up
# ifconfig en0
en0: flags=e080863<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT>
        inet 192.168.1.10 netmask 0xffffff00 broadcast 192.168.1.255
```

If the interface shows `UP` and `RUNNING` entries, try to `ping` the interface again. If this is successful, try to `ping` another address on the same subnet. A `ping` failure here indicates an external network problem.

If the interface is not showing `UP` and `RUNNING`, or if the `ping` to the interface fails, you need to run diagnostics against the adapter to verify the adapter is good before proceeding. Refer to Section 5.3, “Running diagnostics” on page 94 for information on running diagnostics.

Also verify that the adapter option settings match the network and ensure that all cables are correctly terminated and fully seated at both system and network ends.

7.2.6.2 Ethernet interface problems

Some Ethernet adapters may be used with either the transceiver that is on the card or with an external transceiver. The Ethernet adapter riser card found in some 3 series MCA machines (indicated by the designation 2-8) has jumpers to specify which physical interface on the card is to be used. Verify that the jumpers are set correctly (see your adapter manual for instructions). The remainder simply require the appropriate connector type setting as described in the following paragraph.

Verify that you are using the proper Ethernet connector type (thin is BNC; thick is DIX). If you change this connector type, use the Web-based System Manager fast path, wsm devices, or the SMIT fast path, `smit chgenet`, to set the Apply Change to Database Only field. (The field should be checked in Web-based System Manager or set to `yes` in SMIT.) Reboot the machine to apply the configuration change.

Note

If a microchannel system with a previously operational Ethernet network fails following reboot, verify the connector type setting using SMIT as above.

7.2.6.3 Token-ring interface problems

If you cannot communicate with some of the machines on your network even though the network interface has been initialized, the addresses correctly specified, and you have verified that the adapter card is good:

- Check to see if the hosts with whom you cannot communicate are on a different ring. If they are, use the Web-based System Manager fast path, wsm devices, or the SMIT fast path, `smit chinet`, to check the Confine BROADCAST to Local Token-Ring field. The field should not be checked in Web-based System Manager and should be set to `no` in SMIT.
- Check to see whether the token-ring adapter is configured to run at the correct ring speed. If it is configured incorrectly, use the Web-based System Manager Network application or SMIT to change the adapter ring speed attribute. When TCP/IP is restarted, the token-ring adapter will have the same ring speed as the rest of the network.

Note

Running at the wrong speed will cause problems for all machines on the same ring.

7.2.6.4 ATM or ATMLE interface problems

If you are unable to communicate with any network devices through an ATM interface:

- Check the interfaces for the presence of both ATM LAN emulation and classic IP interfaces with the `netstat -in` command. If both are configured and are required, you need to ensure that each configured interface has a unique IP address. A common problem is for users to configure the at0 interface and the ATMLE interfaces en0 or tr0 with the same address.
- If you are using an ATM switch, check to make sure there is a proper connection to the switch (usually indicated by a port light on the switch). To identify cable problems, look for wire fault errors in the error log with the `errpt |grep atm` command. If the connectors on your fibre cables are not keyed or connected with a plastic cover, try reversing the connectors on one end.
- Verify the switch configuration (if you are using one) with the switch administrator since many varied problems occur when the switch is incorrectly configured.
- Make sure you have the latest ATM and ATMLE fileset PTFs installed on your RS/6000 client as many switch ping problems are defects and can be solved by upgrading the filesets to their latest PTF levels.

The filesets consist of the ATM device driver, the ATMLE LAN device driver, and the common ATM filesets.

ATM LAN Emulation Problem Debugging Steps

If you are unable to `ping` to the switch or local ATMLE host, run the `entstat -d entx` or `tokstat -d tokx` command. These commands check if the ATMLE interface has registered with the ATM switch and also what Emulated LAN (ELAN) Name and MAC Address is being used.

In the `General Statistics:` section of the `entstat` or `tokstat` command output, check the driver flags for correct operation:

- The `Driver Flags:` should be similar to the following:

```
Driver Flags: Up Broadcast Running
        Simplex AlternateAddress 64BitSupport
```

If the `Up` or `Running` flag is missing, there is no contact with the ATM switch.

If the `Limbo` flag is present, that means the client lost contact with one or more ATMLE servers and is attempting network recovery.

If the `Dead` flag is present, a hard failure has occurred, and the ATMLE client is no longer operational.

- Check the error log for additional error messages.

The following examples show the `entstat` command outputs for both a failure to register and a successful switch registration.

The following is an example of the `entstat -d ent0` command output following the failure of the RS/6000 ATMLE client to register with the ATM switch:

```
ATM LAN Emulation Specific Statistics:
--------------------------------------
Emulated LAN Name: ZTrans_Lab_Eth
Local ATM Device Name: atm0
Local LAN MAC Address: 08.00.5a.99.89.c4
Local ATM Address:
47.00.91.81.00.00.00.00.00.04.15.00.00.40.00.00.03.75.43.00
Auto Config With LECS: Yes
LECS ATM Address: 00.00.00.11.00.00.00.00.00.00.00.00.00.00.00.00.00.00.00.00
LES ATM Address: 00.00.00.00.00.00.00.00.00.00.00.00.00.00.00.00.00.00.00.00
General Errors: 287                         Address Deregistrations: 0
Control Timeout (sec): 120                  LE_ARP Rsp Timeout (sec): 1
Max Unknown Frame Count: 1                  Flush Timeout (sec): 4
Max Unknown Frame Time (sec): 1             Path Switch Delay (sec): 6
VCC Activity Timeout (sec): 1200            VCC Avg Rate (Kbps): 25600
```

The following ATM Addresses should be listed:

- Local ATM Address of the RS/6000 ATMLE client. Bytes 14 through 19 contain the MAC address that was entered through SMIT.
- The Lan Emulation Configuration Server (LECS) ATM Address if using Auto Config. This will contain either an ATM address with a different switch prefix (bytes 1 through 13), or it may contain the following *well-known* ATM address as defined by the ATM Forum:

  `47.00.79.00.00.00.00.00.00.00.00.00.00.A0.3E.00.00.01.00`

- The Lan Emulation Server (LES) ATM Address. The thirteen byte prefix (bytes 1 through 13) should match that of the local ATM address that is assigned to the RS/6000 Lan Emulation Client (LEC).

If the Local ATM Address and/or the LES ATM Address contain mostly zeros, the ATM switch registration has failed as in the preceding `entstat` example.

Failure to register can be caused by any one of the following:

- The UNI version of the ATM adapter does not match that of the switch.
- The PDU size of the ATM adapter is different to that of the switch.
- The RS/6000 interface to the ATM switch connection is broken.
- The LES/LECS ATM Address is bogus or entered incorrectly, or if the well-known address was used (ATM Address field is blank), the ATM switch does not support the well-known address.
- The Emulated LAN Name is not recognized at the ATM switch or there was no Emulated LAN Name entered in the Add ATM LE Client SMIT panel, and the switch expected one.

The following is an example of the `entstat -d ent0` command output following a successful registration of RS/6000 ATMLE client to the ATM switch:

```
ATM LAN Emulation Specific Statistics:
--------------------------------------
Emulated LAN Name: ZTrans_Lab_Eth
Local ATM Device Name: atm0
Local LAN MAC Address: e0.00.5a.99.89.c4
Local ATM Address: 47.00.05.80.ff.e1.00.00.00.f2.0f.28.f7.08.00.5a.99.89.c4.00
Auto Config With LECS: Yes
LECS ATM Address: 47.00.79.00.00.00.00.00.00.00.00.00.00.a0.3e.00.00.01.00
LES ATM Address: 47.00.05.80.ff.e1.00.00.00.f2.0f.28.f7.00.20.48.0f.28.f7.e0
```

If the ATMLE client registers, but still fails to `ping`:

- Check to see that the ELAN name listed in the `entstat` or `tokstat` command matches the correct network address.

  This problem can occur if the ELAN name is recognized by the ATM switch that allows the client to register, but the name is assigned to a different

subnet. The ELAN names are listed in the LECS/LES depending on how the ATM switch vendor implements ATMLE configuration.

If you have another ATMLE client on the same subnet, `ping` this and check whether the ELAN name is different from the problem ATMLE.

- Check the ARP table for remote IP address and MAC address.
- Try to `ping` the switch port instead of the ATMLE client. If the `ping` to the switch works, the switch may have a problem.
- Check whether the ATM adapter and device driver are working as follows:

  ```
  # atmstat -d atm0 > /tmp/atmstat.out
  ```

  Wait 10 seconds, then run the command again, and append the output to the same file:

  ```
  # atmstat -d atm0 >> /tmp/atmstat.out
  ```

  Edit the /tmp/atmstat.out file and check the following parameters under the `Receive Statistics` column:

  - Packets and Cells received – These two parameters should show an increase in the second `atmstat` which means the adapter and device driver are working.

  No increase in either parameter means the adapter is not able to receive ATM cells.

  If either number is not increasing, the adapter and/or the device driver may be hung. Try powering down and rebooting, if this is possible, to clear any hang condition.

ATM Classic IP Problem Debugging Steps

The debugging steps for classic IP problems are dependent on whether the connection is a Permanent Virtual Circuit (PVC) or Switched Virtual Circuit (SVC).

For PVC connections, verify that you are using the correct Virtual Path Indicator (VPI):Virtual Channel Indicator (VCI) pair, VPI:VCI, with the `smit lsatmpvc` SMIT fastpath.

For SVC connections, ensure that the ARP server address is correctly entered for the ARP client. The following three examples of the `arp -t atm -a` command show typical ARP server registration failures and a good registration output.

The first entry in the `arp` output is the local at0 IP address and its 20 byte ATM address.

- The IP address should match that of at0 and the ATM address should contain the MAC address of the atm0 (or atmx) in bytes 14 through 19.
- If the MAC address, the entire ATM address, or the IP address is zeros or incorrect, registration with the switch did not take place.

  Check the configuration of at0 with the `smit chinet` SMIT fastpath and ensure the ATM address has not been entered incorrectly.

The second entry in the `arp` output is the IP address and the ATM address of the ARP server.

Example 1: The ATM client did not register with the switch.

```
# arp -t atm -a
at0(198.179.228.31)     0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0
  IP Addr        VPI:VCI Handle ATM Address
  ?(198.179.228.6)          N/A     N/A
0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0
#
```

Possible causes of a failure to register are: hardware, wrong or incorrectly entered ATM address, or switch configuration.

Example 2: The ATM client registered with switch, but was unable to contact the ARP server.

```
# arp -t atm -a
at0(146.146.75.239)
  39.9.85.11.11.11.11.11.11.11.11.1.1.0.4.ac.ad.28.6a.0
IP Addr        VPI:VCI Handle ATM Address
  ?(0.0.0.0)      N/A     N/A
  39.9.85.11.11.11.11.11.11.11.11.1.1.0.20.35.99.7.33.0
#
```

Possible causes of being unable to contact the ARP server are: UNI version incorrect or ARP server not recognizing auto-detect, ARP server down or having problems, or different PDU sizes.

Example 3: The ATM client has registered with the switch and made contact with the ARP server.

```
# arp -t atm -a
at0(9.3.35.157) 47.0.5.80.ff.e1.0.0.0.f2.f.28.f7.8.0.5a.99.89.c4.0
  IP Addr        VPI:VCI Handle ATM Address
  flute.austin.ibm.com(9.3.35.150)        0:41   3
47.0.5.80.ff.e1.0.0.0.f2.f.28.f7.0.20.48.f.28.f7.0
  horn.austin.ibm.com(9.3.35.154)         0:43   4
47.0.5.80.ff.e1.0.0.0.f2.f.28.f7.8.0.5a.99.88.cc.0
#
```

If you are receiving output similar to Example 3, try a different port on the ARP server/ATM switch (check the LED on port); also check the UNI version and PDU size using `smit chg_atm` to ensure they match those of switch.

If the remote host is able to `ping` the RS/6000 ATM client through the ARP server, try manually adding the remote client ARP entry into the ARP table as shown below.

For SVC clients:

```
arp -t atm -s svc  <20 byte atm address of remote host>
```

If the ATM client `arp` command output continues to show no registration or no ARP server contact, then run the `atmstat` command as follows to verify the adapter and device driver:

```
# atmstat -d atm0 > /tmp/atmstat.out
```

Wait 10 seconds, then run the command again, and append the output to the same file:

```
# atmstat -d atm0 >> /tmp/atmstat.out
```

Edit the /tmp/atmstat.out file and check the following parameters under the Receive Statistics column:

- Packets and Cells received – These two parameters should show an increase in the second `atmstat` which means the adapter and device driver are working.

No increase in either parameter means the adapter is not able to receive ATM cells.

If either number is not increasing, the adapter and/or the device driver may be hung. Try powering down and rebooting, if this is possible, to clear any hang condition.

If the problem cannot be determined at this point, you may need to place a support call using local reporting procedures.

7.2.6.5 X.25 interface problems

This section contains specific debugging information for X.25 TCP/IP interfaces and assumes that the basic TCP/IP network debugging in Section 7.2.1, “Selective host network problems” on page 156 through Section 7.2.4, “Routing problem debugging” on page 157 has been done.

To simplify the X.25 TCP/IP problem determination, you may find it useful to disable any external name resolution and suppress any default gateways. This will avoid any discrepancies that there might be with /etc/hosts.

Note that /etc/hosts is always used by the `x25ip` command when it is executed by /etc/rc.net during the initialization of the system, even if you have a name server.

If you are unable to `ping` your remote X.25 host (see Section 7.4.4, “ping” on page 176 for `ping` command examples), proceed with the following steps:

1. Use the `lsx25` command to see if your TCP/IP X.25 interface (xs0) and your X.25 port (sx25a0) are configured and available. If they are not, you need to configure them. If either the X.25 port or TCP/IP X.25 interface will not configure, see the “X.25 Problem Determination” section in the *AIXLink/X.25 LPP Cookbook,* SG24-4475, for more information on debugging the X.25 LPP.
2. Verify with `smit chinetsx25` SMIT fastpath that the IP-to-NUA translation table entry for the remote host is correct, then do the same verification on the remote host.
3. Refresh the X.25/IP translation table using the `x25ip` command without any arguments, then run the `ping` command again.

If you are still unable to `ping` the remote system, you need to make sure you have routes set up for the X.25 interface (for example, xs0). If the routes seem correct, you need to use the `x25mon` command to monitor the packet level flow as follows:

```
# x25mon -f -p -n sx25a0
```

In another window or session, `ping` the remote host. If you do not get the expected result, there are four possible scenarios:

1. No error message, nothing transmitted.

   You have a gateway that `ping` used to send your packets. Verify with `netstat -r` (not with `ifconfig`) that the X.25 connection is active. Recheck the static routes and, if necessary, remove and replace the X.25 routes that have a `G` flag in the `netstat` output.

2. Error message, nothing transmitted.

   If the error message is:

   ```
   sendto: Network is unreachable
   ```

   Or:

   ```
   sendto: No route for this host
   ```

   You have a problem either with the attachment definition or with the route.

   If the error message is:

```
ping: sendto: Can't assign requested address
Sendto: The socket name is not available
```

There is a problem with the x25ip translation table. Change it with `smit chinetsx25` or add a correct entry if necessary with `smit mksx25`.

3. No error message, the trace shows attempts to establish a virtual circuit.

   If you see in the trace a series of call packets showing unsuccessful attempts to open a virtual circuit, you have an X.25 protocol problem. Refer to X.25 LPP documentation for the CCITT Cause and Diagnostic code or the “X.25 Protocol Problems” section in the *AIXLink/X.25 LPP Cookbook*, SG24-4475.

   If the incoming call is rejected by the remote system with the diagnostic 241, calling address missing, the calling address was not in the translation table. Add it and use `x25ip` on the remote host to regenerate it.

4. No error message, data transmitted but nothing received.

   In this case, a virtual circuit is established and data packets are sent to the remote host but the remote host does not reply. The packets are correctly transmitted to the remote host, but the remote host IP address they carry does not match the actual IP address of the remote host. Change the IP address of the remote host so it is the same on the remote host as on the local host.

7.3 Common TCP/IP problems

The following sections contain some of the more commonly encountered problems that occur with TCP/IP and other related protocols.

7.3.1 LED 581 hang

During the machine boot, 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.

This problem can be either a system or a network problem that happens because TCP/IP waits for replies over some interfaces (token ring, for example). If there are no replies, it eventually times out on the attempt and marks the interface as down. This timeout 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:

   ```
   mv /etc/rc.net /etc/rc.net.save
   ```

3. Reboot in Normal mode boot to see if the system continues past the LED 581 and allows you to log in.

Note

The above steps assume that neither DNS or NIS is 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 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, ypbind/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.

- Bad LPPs

  Reinstall the LPP.

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

```
######################################################################
# 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

###################################################################
```

7.3.2 Telnet problems

If you can `ping` both ways between the source and target systems but cannot `telnet` or `rlogin`, use the following checklist:

1. Ensure that reverse name resolution is functioning on both the Telnet client and server systems with the `host` command as follows:

   ```
   # host <telnet server system host name>
   # host <IP address returned from previous command>
   ```

   If both commands return identical output, run the commands from the Telnet server to the Telnet client. If there are discrepancies in the output pairs, you have name resolution problems, refer to Section 7.2.3, “Name resolution problems” on page 156 to resolve these.

2. Check to make sure you have the following line in the /etc/inetd.conf file:

   ```
   telnet  stream  tcp6    nowait  root    /usr/sbin/telnetd      telnetd -a
   ```

   Note

   The `tcp6` in the above line appears in AIX V4.3.0 and later.

3. Check to make sure you have the following line in the /etc/services file:

   ```
   telnet          23/tcp
   ```

4. Check to see that inetd is running as follows:

   ```
   # ps -ef | grep inetd | grep -v grep
       root  7746  5714   0 04:56:38      -  0:00 /usr/sbin/inetd
   ```

   If inetd is running, you will receive a single line of output similar to the line above. If inetd is not running, start it with the `startsrc -s inetd` command.

5. Try to `telnet` to a different port number:

   ```
   # telnet <hostname> chargen
   ```

   If the connection to the chargen port is successful, you will see streaming output similar to the following:

   ```
   vwxyz{|}~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^
   wxyz{|}~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
   xyz{|}~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`
   ```

```
yz{|}~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`a
z{|}~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`ab
{|}~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abc
|}~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcd
}~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcde
~ !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdef
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefg
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefgh
"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghi
#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghij
$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijk
%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijkl
&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklm
```

Terminate the `telnet` command with the **Ctrl-C** sequence.

7.3.3 Login delays from AIX 4.3.x systems

If your network is using a DNS running AIX 4.2.1 or earlier (or is a non-AIX system), and you are experiencing login delays of around two minutes when establishing remote sessions, for example, telnet or ftp, from the 4.3.x systems, you may have a problem with name resolution.

AIX 4.3.2 attempts to use IPv6 by default for name resolution; therefore, if you are experiencing the above problem, you need to select IPv4 for the DNS name lookup. This can be achieved by one of the following methods:

- Create a file called /etc/netsvc.conf that contains the entry:

  ```
  hosts=local,bind4
  ```

- Export the korn shell variable `NSORDER` as follows:

  ```
  export NSORDER=local,bind4
  ```

The above examples simply illustrate the use of the `bind4` parameter to force the use of the IPv4 protocol stack, your order of name resolution may vary from the order shown.

7.3.4 Dynamic Host Configuration Protocol (DHCP) problems

If clients are unable to obtain an IP address or other configuration parameters:

- Check to see that you have specified an interface to be configured. This can be done through the Web-based System Manager Network application, by editing the /etc/dhcpcd.ini file, or by using the SMIT fast path `smit dhcp`.
- Check to see that there is a server on the local network or a relay agent configured to get your requests off the local network.

- Check to see that the dhcpcd daemon program is running. If it is not, use the `startsrc -s dhcpcd` command.

The dhcpsd daemon not running:

- This error is common when bootp is running under inetd; bootpd and dhcpsd cannot run at the same time. The dhcpsd process uses the same service port as bootps; however, dhcpsd is not an inetd subserver and is started in the /etc/rc.tcpip file rather than the /etc/inetd.conf file. The bootps line in /etc/inetd.conf must be commented out with `#`, then inetd must be refreshed using the command:

  ```
  refresh -s inetd
  ```

  The bootps line in /etc/services remains as is:

  ```
  bootps        67/udp      # bootp server port
  ```

7.3.5 X.25 function keys not working properly

When using telnet or rlogin over X.25, the remote system does not always respond correctly to use of the function keys, especially if using Esc-1 instead of F1.

The ESCDELAY environment variable controls the timeout period, after which the screen handling applications will consider the Esc key to be separate from the following 1.

The ESCDELAY variable only works if the application is written to Curses or Extended Curses. When running over X.25, the Esc key is often put by IP into a separate data packet from the following key. The gap between the two data packets on the receiving system is often further apart than the default timeout. The default value of ESCDELAY (or if it is not set at all) is 500.

To see if this is the reason your function keys are not working correctly, try setting ESCDELAY to a very high value as follows:

1. Insert the following in the /etc/environment file:

   ```
   ESCDELAY=5000
   ```

2. Log completely out of the system, then back in to make sure the new ESCDELAY value is set.
3. Verify using the following command:

   ```
   echo $ESCDELAY
   ```

Once your test is completed, remember to either adjust the value down to a more reasonable value (between 1000 and 2000) or, if your problem is not

resolved, remove this line completely from the /etc/environment file. If setting this ESCDELAY parameter does not fix your problem, then your application is not written to extended curses and the application must be modified to await the entire escape sequence. For example, the AIX SMIT application is written to curses, but it will work over a slow network regardless of ESCDELAY because SMIT was written to wait indefinitely for the entire escape sequence. Thus, even though the `iptrace` command output clearly shows the separation of the Esc and the remainder of the escape sequence, the Escape keys works correctly in SMIT.

7.4 TCP/IP network configuration issues

The following sections discuss a selection of configuration issues that can effect network performance.

7.4.1 Maximum Transmission Unit (MTU)

Each device has an MTU and will not transmit packets larger than this without fragmenting. For example:

- If packets are larger than this, they will be fragmented before they are sent.
- Packets crossing a router may be fragmented for the destination network.
- Machines with different MTU sizes may have difficulty communicating.
- If traffic is predominantly local, use the largest MTU supported by your LAN type. This minimizes fragmentation of packets, which is costly.

The largest MTU for Ethernet is 1500 bytes.

TCP establishes a Maximum Segment Size (MSS) that the two hosts agree on during the handshaking protocol. This MSS will not be greater than the smallest MTU, and defaults to 512 bytes (the tcp_mssdflt value can be displayed from the network options).

UDP does not establish a connection between two hosts.

You also need to satisfy the sometimes conflicting goals of setting the MTU size to the maximize packet size while at the same time trying to minimize fragmentation.

7.4.2 Mbufs

To check that enough memory (mbufs) is available for networking activities, use the `netstat -m` and `netstat -i` commands:

- A non-zero `requests for mbufs denied` value in the `netstat -m` output indicates a possible lack of network memory; use the `no` command to increase `thewall` value. Some possible causes of non-zero in this field are:
  - Heavy network activity is going on and packets are being dropped.
  - A network application with a memory leak can exhaust the supply of mbufs.
  - More memory may need to be allocated for mbufs or more physical memory added to the machine.
- Check the `Oerrs` column of the `netstat -i` command output:
  - If the number in the `Oerrs` column exceeds 1 percent of that in the `Opkts` column for any interface, the TRANSMIT queue size (`xmt_que_size`) for the adapter should be increased.
  - If the number of `Oerrs` exceeds 1 percent of the `Ipkts`, use the `netstat -m` command as above to check for a possible lack of memory problem.

The network options parameter `thewall` specifies the maximum amount of memory that can be allocated to the network buffer pools of mbufs and mclusters. To check the value, use the `no -a | grep thewall` command; the value shown is in Kbytes. To change the value, refer to Section 7.4.10, “no” on page 181.

7.4.3 TCP/IP problem isolation commands

The following sections contain summaries and usage examples of some useful TCP/IP problem isolation commands.

7.4.4 ping

The `ping` command is used for investigating basic point-to-point network connectivity problems, answering questions about whether the remote host is attached to the network, and whether the network between the hosts is reliable. Additionally, `ping` can indicate whether a host name and IP address is consistent across several machines. See the following example for usage:

`ping <host name>` or `<address>`

```
# ping bern
PING bern: (89.86.41.183): 56 data bytes
64 bytes from 89.86.41.183: icmp_seq=0 ttl=255 time=1 ms
64 bytes from 89.86.41.183: icmp_seq=1 ttl=255 time=2 ms
^C
----bern PING Statistics----
4 packets transmitted, 4 packets received, 0% packet loss
round-trip min/avg/max = 1/1/2 ms
```

`ping -R` reports the path a packet travels – use in conjunction with the `traceroute` command.

`ping -c <num>` sends num packets.

Control size of packets sent with `ping <host> <size>`.

The `ping` command normally sends 56 bytes of data plus the IP header, for a total of 64 bytes. Using different values can help determine if there are problems with fragmentation. Try values slightly less than, equal to, and greater than the MTU size.

The `ping` command uses echo request and echo reply ICMP messages, which is part of the IP layer, in the data portion of IP datagrams.

ICMP messages are also used to indicate if a destination for a packet is unreachable, if the packet should be redirected to another router (for more efficient routing called ICMP redirect), and can indicate other types of errors as well.

7.4.5 rup

The `rup` command queries a host for its up time. The `rup` command:

- Uses broadcast on all interfaces by default.
- Uses RPC over UDP.

The `rup` command is useful to find out whether name service and interfaces work in one command. It also tells you how long the machines have been up and whether they are busy.

A remote host responds only if it is running the rstatd daemon, which is defined in the /etc/inetd.conf file and started from the inetd daemon.

7.4.6 netstat

The `netstat` command with various flags is useful in the debugging of most network problems. `netstat` collects and displays a large amount of network related statistical and configuration information including:

- Active sockets

  Displayed with the `netstat -f <family>` command.

  Where family is:

  - `inet` for Internet sockets

- `unix` for Unix sockets
- `ns` for Xerox Network System sockets

The output format is family specific.

The default display (no flags) shows active sockets for all families.

- Interfaces

  The state and usage of all configured interfaces is displayed with the `netstat -i` command as shown in the following example:

```
# netstat -i
Name  Mtu   Network     Address            Ipkts Ierrs    Opkts Oerrs  Coll
lo0   16896 link#1                         8114     0     8116     0     0
lo0   16896 127         loopback           8114     0     8116     0     0
lo0   16896 ::1                            8114     0     8116     0     0
en0   1500  link#2      2.60.8c.2e.e1.c9      0     0       19     0     0
en0   1500  192.168.1   jumbuck               0     0       19     0     0
tr0   1492  link#3      10.0.5a.a8.70.67   3340825     0   202253     0     0
tr0   1492  9.185.112   msahanc.au.ibm.co  3340825     0   202253     0     0
sl2*  1006  link#4                            0     0        0     0     0
sl2*  1006  130.130.130 130.130.130.1         0     0        0     0     0
#
```

  For each interface, `netstat -i` shows:

  - Name (name* indicates the interface is down)
  - MTU
  - Link layer address (<link# line>)
  - IP network
  - IP name or address
  - Input packets and errors
  - Output packets and errors

- Routing tables

  Displayed with the `netstat -r` command as shown in the following example:

```
# netstat -r
Routing tables
Destination      Gateway           Flags   Refs     Use  If   PMTU  Exp  Groups

Route Tree for Protocol Family 2 (Internet):
default          9.185.112.254     UG        2   124847  tr0     -   -
9.185.112/23     msahanc.au.ibm.co U        29    82305  tr0     -   -
127/8            loopback          U         5     2011  lo0     -   -
130.130.130.2    130.130.130.1     UH        0        0  sl2     -   -
192.168.1/24     jumbuck           U         0       62  en0     -   -

Route Tree for Protocol Family 24 (Internet v6):
::1              ::1               UH        0        0  lo0 16896   -
#
```

The route tree for each protocol family is shown.

- Protocol statistics

  The `netstat -s` command displays protocol statistics for IP, ICMP, IGMP, TCP, and UDP. The data shown includes information on errors, fragmentation, redirects, packets sent and received, and packets dropped and discarded.

  It also shows TCP connections established, timed out, and UDP datagrams received and sent.

  To display a single protocol, use `netstat -p <protocol>` as shown in the following terminal output:

```
# netstat -p udp
udp:
        2507220 datagrams received
        0 incomplete headers
        0 bad data length fields
        1 bad checksum
        297 dropped due to no socket
        2498325 broadcast/multicast datagrams dropped due to no socket
        0 socket buffer overflows
        8597 delivered
        3233 datagrams output
```

- Network adapter device driver statistics

  The `netstat -v` command shows device driver statistics for all installed network adapters.

  If you only want the output of a specific adapter type you can use the `stat` commands instead of `netstat -v`, refer to Section 7.4.11, “Stat commands” on page 181 for details.

- Network memory usage statistics

  Network memory usage is displayed using the `netstat -m` command. AIX V4.1 greatly expanded the collection of network memory usage statistics. To reduce unnecessary system load, AIX V4.3.2 has now made the collection of extended network memory services statistics a configurable option. By default, the extended statistics collection is set to off. To enable extended statistics collection, use the command:

```
/usr/sbin/no -o extendednetstats=1
```

  For for this change to survive reboot, comment out the following lines in /etc/rc.net as shown below:

```
######################################################################
#if [ -f /usr/sbin/no ] ; then
#        /usr/sbin/no -o extendednetstats=0 >>/dev/null 2>&1
#fi
```

- Protocol Control Block (PCB) addresses

The `netstat -A` command is used to show the address of any PCB associated with the sockets. The `-A` flag is normally only used for low-level debugging purposes.

7.4.7 arp

Address Resolution Protocol (ARP) maps IP addresses to hardware addresses. For example:

- To display the local machine's cache of ARP translations, use the `arp -a` command.
- IP addresses only have meaning to IP. The data-link level of the Internet network layer use the hardware address of the adapter itself. Each hardware address is unique.
- ARP translates IP addresses to hardware addresses and vice-versa.
- It works by using a broadcast asking for the owner of a particular IP address.

7.4.8 iptrace and ipreport

The `iptrace` command traces all packets to and from a machine. For example:

- This command generates very detailed output that is viewed by using the `ipreport` command.
- It can be a useful tool for finding lower level communication errors needed for detailed debugging of network problems. It is not really needed for general use since interpretation of the output requires a detailed knowledge of TCP/IP.

Specify source and or destination host with `-s` <source> `-d` <dest> flags to narrow down the problem. The following flags are also useful in reducing the amount of captured data, greatly simplifying analysis of the output:

- -a excludes ARP packets.
- -P <protocol> gathers data only for a specified protocol.
- -i <interface> gathers data for a specific interface.

The following example illustrates how to trace all traffic sent to host `columbia` through the `en0` interface and place formatted output into a file:

```
# iptrace -i en0 -d columbia /tmp/trace.raw
```

Reproduce the problem:

```
# ps -ef | grep iptrace
# kill <iptrace PID>
# ipreport -rns /tmp/trace.raw > /tmp/trace.out
```

7.4.9 tcpdump

The `tcpdump` command prints out detailed packet headers for TCP:

- It allows very detailed analysis, again useful for expert network debugging.
- It allows you to specify pattern matching to narrow the information shown.

The following example prints out all packets arriving or departing from `host firefly`:

```
# tcpdump host firefly
```

7.4.10 no

The `no` command changes or displays network options:

- The `no -a` command displays the values of all the network parameters defined in the system.
- The `no -o` command allows you to change selected settings or values within the kernel. It allows the changes to take effect immediately but the changes only last until the next system boot. For example, to change `thewall` value to 64 MB, the following command syntax is used:

  ```
  # no -o thewall=65536
  ```

  To maintain the settings across system boots, the specific `no` commands need to be appended to the /etc/rc.net file as shown below:

  ```
  ############################################################
  if [ -f /usr/sbin/no ] ; then
          /usr/sbin/no -o tcp_sendspace = 16384
          /usr/sbin/no -o tcp_recvspace = 16384
          /usr/sbin/no -o thewall=32768
  fi
  ```

System values should only be changed after referring to tuning information in AIX documentation or under the direction of support personnel.

7.4.11 Stat commands

Instead of using the `netstat -v` command, you can use a `stat` command for the particular interface. These commands reside in the /usr/sbin directory and include the following:

- `entstat <ent#>`
- `tokstat <tok#>`

- `atmstat <at#>`

> **Note**
>
> Use the `-d` flag for detailed device specific data.

7.4.12 dadmin

The `dadmin` command was added to AIX V4.3.1. This man page information is included here, because at the time of writing this book, the man page information was not available online using the `man` command. The `dadmin` command is a tool that lets DHCP administrators query and modify the state of their DHCP servers' databases. It gives the administrator the ability to locally or remotely query the DHCP server for the status of an IP address, query for a pool of IP addresses, query for a client, delete an IP address mapping, refresh the server, and change the server's tracing level. The `dadmin` command is backwards compatible with previous AIX release DHCP servers to list their IP address status and refresh.

When querying for IP address information, the `dadmin` command returns the IP address status and, depending on the status, may return the lease duration, start lease time, last leased time, whether the server supports DNS A record updates for this IP address, and the client identifier that is mapped to this IP address.

When querying for client information, the `dadmin` command returns the client's IP address and IP address status, whether the server supports DNS A record updates for this IP address, the last time the client was given any IP address, and the host name and domain name used by the client.

When modifying the server tracing level, the `dadmin` command sets and returns the server tracing level in the form of a tracing mask. This mask represents a bitstring where each bit represents whether a specific log item is being traced by the server (see DHCP Server Configuration in the online documentation). In order from least significant to most significant, these log items are LOG_NONE, LOG_SYSERR, LOG_OBJERR, LOG_PROTOCOL and LOG_PROTERR (same value), LOG_WARN and LOG_CONFIG (same value), LOG_EVENT and LOG_PARSEERR (same value), LOG_ACTION, LOG_INFM, LOG_ACNTING, LOG_STAT, LOG_TRACE, LOG_START, and LOG_RTRACE.

Note

LOG_START cannot be disabled. This implies a mask range from 0x0800 through 0x1FFF.

The syntax is as follows:

```
dadmin [ -? ] [ -v ] [ -h Hostname ] [ -f ] -d IpAddress | [ -x ] -i | [ -x
] -s | -t on|off|Value | -q IpAddress | -p IpAddress | -c ClientId
```

Flags:

-v Execute the command in verbose mode.

-h Used to specify the destination DHCP server. Hostname can either be a name or IP address.

-f To be used with the -d flag, this forces the deletion of the address without any prompting. Deletes the lease information associated with IP.

-x Deletes the lease information associated with IP address IpAddress. As a result, the address will be moved to the FREE state and be available for binding once again.

-x Uses Version 1 of the dadmin protocol. This flag is used to connect to previous AIX release DHCP servers and is only valid for the -i and -s flags.

-i Reinitializes the DHCP server. This signals the server to sync its databases and restart by rereading the configuration file.

-s Returns the status of each address in the DHCP server's configured pools.

-t Changes the tracing level of the DHCP server. Trace values are reported in a hexadecimal format representing the tracing mask in use on the server. Value can be specified as either a decimal or hexadecimal format. The keywords on and off enable or disable a single bit at a time in the tracing mask.

-q Returns the status of a specific IP address.

-p Returns the status of each address in a subnet. IpAddress is used to identify the subnet to list.

-c Returns the status for a specific client that may be known to the DHCP server. ClientId represents the client identifier that a DHCP client used to identify itself, or the hexadecimal client hardware address that BOOTP clients use to identify themselves. This field can either be

specified as hexadecimal characters only, or in the TYPE-STRING representation used by the DHCP server.

The following example shows typical output when the `dadmin -v -s` command is used to query the DHCP server's configured pools for the status of each IP address:

```
# dadmin -v -s
Connecting to the DHCP server: 192.168.10.227
Got a socket, attempting to connect.
Connected to 192.168.10.227 successfully.
Send of header completed.
PLEASE WAIT....Gathering Information From the Server....PLEASE WAIT
Receive of header completed.
IP Address      Status  Lease Time Start Time  Last Leased Proxy ClientID
192.168.10.150   Used
192.168.10.151   Leased  INFINITE   06/18 10:32 06/18 10:32 FALSE 1-080009d6ae59
192.168.10.152   Reserved  48:00:00             06/18 10:37 FALSE 1-00c04fdc5e22
192.168.10.153   Free
192.168.10.154   Free
192.168.10.155   Free
Received 6 data records.
DADMIN completed successfully.
#
```

7.5 NIS troubleshooting

This section covers the tools and methods used to troubleshoot NIS problems. Additional information on RPC debugging can be found in Section 7.6, "NFS troubleshooting" on page 191.

7.5.1 Troubleshooting tools for NIS

The following sections contain a selection of networking commands that can be used as NIS troubleshooting tools.

7.5.1.1 ping

The `ping` command is a general purpose tool for investigating point-to-point connectivity problems.

If the `ping` command hangs or has long response times, then there is a low-level connectivity problem. This should be resolved before debugging NIS problems any further.

For more detail on the `ping` command, refer to Section 7.4.4, "ping" on page 176.

7.5.1.2 rpcinfo

The `rpcinfo` command is an analog of `ping` that queries RPC servers and their registration with the portmapper, thereby verifying that a remote machine is capable of replying to an RPC request.

The `rpcinfo` command can be used to detect:

- Dead or hung servers caused by improper configuration or a failed daemon.
- RPC program version number mismatches between client and server.
- Renegade RPC servers (NIS servers that do not have valid maps for the domain they are serving).
- Broadcast-related problems.

The `rpcinfo -p` command takes a remote host name and queries the portmapper on the host for all registered RPC services:

- Output from the `rpcinfo` command shows the RPC program and version numbers, the protocols supported, the IP port used by the RPC server, and the name of the RPC service.
- Service names come from the rpc.bynumber NIS map. If no name is printed next to the registration information, the RPC program number does not appear in the map.
- Missing RPC service names could indicate a corrupted or incomplete rpc.bynumber NIS map.

When working toward diagnosing any RPC-related problem, verify that the remote portmapper is alive and returning valid RPC registrations.

- If the portmapper on the remote machine has died or is not accepting connections for any reason, `rpcinfo` times out attempting to reach it and reports the error.

  This indicates a low-level problem in the network.

If you are debugging in a heterogeneous environment and running multiple versions of each RPC service, it is possible to get RPC version number mismatch errors.

These problems affect NIS and diskless client booting; they are best sorted out by using the `rpcinfo` command to emulate an RPC call and by observing server responses.

For example, perform a broadcast and then watch the order in which responses are received. The `rpcinfo -b` command sends a broadcast request to the specified RPC program and version number:

```
# rpcinfo -b ypserv 1
89.86.41.194 austin
89.86.41.195 newyork
89.86.41.196 boston
```

In this example, all NIS servers on the local network answer the `rpcinfo` broadcast request to the null procedure of the `ypserv` daemon. If `austin` should not be the NIS server and clients bind to it, the network will be prone to periods of intermittent failure.

A renegade NIS server may be the first to answer a ypbind broadcast for NIS service. Its lack of information about the domain makes the client machine unusable.

Failure to fully decommission a host as an NIS server (leaving empty NIS map directories, for example) may cause this problem.

The `rpcinfo` command helps to determine why a particular client cannot start the NIS service. If no host replies to the `rpcinfo` command request, then the broadcast packet is failing to reach any NIS servers. If the NIS domain name and the broadcast address are correct, then it may be necessary to override the broadcast-based search and give ypbind the name and address of a valid NIS server by issuing a `ypset` command.

Sometimes, just looking at the list of servers that respond to a request may indicate a problem if you notice that one of the servers should not be answering the broadcast.

Like `ping`, the `rpcinfo` command provides a measure of basic connectivity at the session layer in the network protocol stack. Pinging a remote machine ensures that the underlying physical network and IP address handling are correct: using the `rpcinfo` command to perform a similar test verifies that the remote machine is capable of replying to an RPC request. This includes validation of the integrity of the network and that there is an RPC service registered on the other machine.

7.5.1.3 ypmatch

As a diagnostic tool, the `ypmatch` command can be used to see if a change to a map has taken effect and identify NIS maps that are out of synchronization after a map transfer has been requested or scheduled.

Generally, building a new map will push the map to other servers with the `yppush` command.

To check for map consistency, issue the `ypmatch` command on several clients and then the server. Verify that the same results are returned.

If not, a map inconsistency exists. Try pushing the maps to servers with the `yppush` command.

Often, NIS map changes may not propagate as quickly as desired even though the change has been pushed with the `yppush` command.

7.5.1.4 ypwhich

The `ypwhich` command is used to verify an NIS server for a client as shown in the following example:

```
# ypwhich
godzilla
```

If a host name is passed as a parameter, the `ypwhich` command queries the named host for its current binding. If the `ypwhich` command cannot resolve the host name into an IP address, it reports an error as follows:

```
# ypwhich kingkong
ypwhich: can't find kingkong
```

Note

An IP address may be used in place of a host name.

The `ypwhich -x` command prints the table of nicknames:

```
stepiii $ ypwhich -x
Use passwd for map passwd.byname
Use group for map group.byname
Use networks for map networks.byaddr
Use hosts for map hosts.byaddr
Use protocols for map protocols.bynumber
Use services for map services.byname
Use aliases for map mail.aliases
Use ethers for map ethers.byname
```

The `ypwhich -m` command examines the NIS master server name embedded in the NIS map DBM file:

```
# ypwhich -m
auto.master stargate
auto.src stargate
rpc.bynumber stargate
protocols.bynumber stargate
auto.projects stargate
auto.mail stargate
auto.home stargate
```

If you have concerns about data disappearing from NIS maps, dump the entire map (including keys) using the `makedbm -u` command:

```
stargate $ /usr/etc/yp/makedbm -u ypservers
stargate
lazerus
delli
YP_LAST_MODIFIED 0706832572
YP_MASTER_NAME stargate
stargate $
```

The master map information is useful if you have changed NIS master servers and need to verify that client maps are built correctly and synchronized with the new server.

Querying client bindings individually is useful for debugging client problems.

In addition to providing NIS server binding information, the `ypwhich` command examines the NIS map information, the master server for a map, the list of all maps, and map nickname translations.

7.5.2 Troubleshooting examples with NIS

When debugging a network problem:

- Think about potential causes of the problem.
- Work your way through the protocol layers to ensure you do not miss a low-level problem that is posing as a high-level failure.

For example, if your attempts to bind to an NIS server are failing:

- Test the network using the `ping` command.
- Test the ypserv processes using the `rpcinfo` command.
- Finally, check the binding itself with the `ypset` or `ypwhich` commands.

7.5.2.1 yppush hangs making new maps

If you find the `yppush` command hangs when adding new NIS maps after initial maps have been transferred to a slave server, you need to:

1. Make the new map(s) with the `NOPUSH` option set to `1`.

   ```
   make NOPUSH=1 <new_mapname>
   ```

   Failure to use the NOPUSH option will cause the command to hang.

2. Update the slave servers by running `ypxfr` on each one.

   This only applies if the new map does not already exist on the slave server.

7.5.2.2 Client broadcasts do not bind

The most common problem occurring at an NIS client node is for a command to hang.

Sometimes a command appears to hang, and a message like the following appears on the console:

```
NIS: server not responding for domain <wigwam>. Still trying
```

This error message indicates that the ypbind daemon on the local machine is unable to communicate with a ypserv daemon serving the wigwam domain.

Normally, this condition is temporary; the messages go away when the NIS server machine reboots and the ypserv daemon restarts or when the load on the NIS server and the network decreases.

If you are getting a `server not responding for domain` message, one of the following situations might exist:

- The domain name on the NIS client machine is not set or is set incorrectly.

  Clients must use a domain name that the NIS servers know.

- Your local network may not have an NIS server machine.

  In this circumstance, all other NIS clients on your network show the same or similar problems.

- The NIS server may not be up and running or it may be overloaded.

  If there is not a backup server in the broadcast subnet, all clients will hang on NIS query until the server recovers or the load reduces.

- The broadcast mask may be set incorrectly.

  Configuration errors in setting up the network or broadcast mask can also keep the broadcasts that ypbind uses from reaching a server.

If a server is unreachable due to some catastrophic failure, it may be necessary to temporarily start up a new server for the domain so the clients can be entered with logins.

7.5.2.3 The ypwhich command is inconsistent

When you use the `ypwhich` command repeatedly at the same client node, the response varies because the status of the NIS server changes. This is normal.

- The binding of NIS client to NIS server changes over time on a busy network. Whenever possible, the system stabilizes so that all clients get acceptable response time from the NIS servers.

- The source of an NIS service is not important because an NIS server machine often gets its own NIS services from another NIS server on the network.

A common cause of concern is when an NIS server that is also running as an NIS client is bound to some other server. It seems intuitive that the machine would bind to itself, but this is often not the case. The client is indiscriminate as to the server it binds to.

7.5.2.4 Very large maps are not supported

The NIS architecture does not scale well. 2 GB maps are the absolute maximum.

Large DBM databases are very slow and inefficient. When map sizes approach 2 GB, the DBM routines can fail with fseek() errors that cause map creation to fail. Prior to reaching this size however, the performance has usually degraded to the point where using the maps is no longer practicable.

7.5.2.5 Routers, network devices, and NIS binding

NIS binding is normally done through broadcasts.

Broadcasting is the primary way for NIS clients to bind to a server and for NIS slaves to bind to a master for map transfers. If the clients fail to bind, and there are bridges or routers between the machines seeking to bind, the problem may be that the hardware cannot pass through the broadcast packets necessary for the binding.

The `ypset` command can help get around this, but has the disadvantage that it is unreliable and not secure. The best solutions in these situations are to have a server machine bridging the two networks or to break the network into two separate domains (perhaps duplicates of one another).

7.5.2.6 Group membership wrong for new group changes

Usually, this problem occurs because the netid map is incorrect.

Since it contains a mapping of groups by user, the netid map will yield incorrect results if it is not rebuilt whenever the passwd and group maps are changed.

To avoid this problem, do not make the group map by itself.

7.5.2.7 mkdbm fails with large NIS map entries

The design of DBM limits entries to 1024 bytes or less as follows:

The sum of DBM key + data must be < = 1024 bytes.

The most common problem caused by this is when a large group entry is rejected by dbm because there are too many users in one group, causing mkdbm to fail.

The problem is usually seen in group map entries and has two workarounds:

- Make two or more groups with the same gid.
- Add users with the *large group* as their primary group then delete the user names from the group entry in /etc/group.

7.6 NFS troubleshooting

Prior to starting any NFS debugging, it is necessary to ensure the underlying network is up and working correctly. It is also most important to ensure that name resolution is functional and consistent across the network and that end-to-end routing is correct both ways.

7.6.1 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 above. Refer to Section 7.2, “Problem isolation steps for TCP/IP network problems” on page 155 if necessary.
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 dymanically created on AIX Version 4.2.1 and later)
- rpc.statd
- rpc.lockd

4. Initiate an `iptrace` (client or server or network), reproduce the problem, then view the ipreport output to determine where the problem is. Refer to Section 7.4.8, “iptrace and ipreport” on page 180 for usage details.

7.6.2 NFS mount problems

Mount problems fall into one of the categories below:

- 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 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 and two different names for the two adapters and the export only specifies one. This problem can be fixed by exporting both names.
  - Server cannot do a lookuphostbyname or lookuphostbyaddr onto the client. To check, make sure the following commands both resolve to the same thing:
    - `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 the like. To correct this condition, `mountd` needs to be restarted:

```
# stopsrc -s rpc.mountd
# startsrc -s rpc.mountd
```

- System date being wildly 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 V4.2.1 or later clients running NFS Version 3 to AIX V4.1.5 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 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 of some minutes. To avoid this problem, NFSV3 provided the `proto` and `vers` parameters with the `mount` command. These parameters are used with the `-o` option to hardwire the protocol and version for a specific mount. The following example forces the use of UDP and NFSV2 for the mount request:

```
# mount -o proto=udp,vers=2,soft,retry=1 platypus:/test /mnt
```

Note

If the `proto` and the `vers` do not match the server, the mount will fail altogether.

- 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 that requests a mount is in more than eight groups. The only workaround for this is to decrease the number of groups the user is in or mount via a different user.

7.6.2.1 Hard versus soft mounts

When the network or server has problems, programs that access hard-mounted remote file systems fail differently from those that access soft-mounted remote file systems.

If a server fails to respond to a hard-mount request, NFS prints the message:

```
NFS server <hostname> not responding, still trying
```

Hard-mounted remote file systems cause programs to hang until the server responds because the client retries the mount request until it succeeds. You should use the `-bg` flag with the `mount` command when performing a hard mount so that if the server does not respond, the client will retry the mount in the background.

If a server fails to respond to a soft-mount request, NFS prints the message:

```
Connection timed out
```

Soft-mounted remote file systems return an error after trying unsuccessfully for a while. Unfortunately, many programs do not check return conditions on file system operations, so you do not see this error message when accessing soft-mounted files. However, this NFS error message will print on the console.

7.6.3 NFS performance problems

The first thing to determine when you have an NFS performance problem is whether this is a problem of not enough resource or too much resource. The following questions need to be answered to assist with this determination:

- What is the process load on the client and/or server?
  - Are there a lot of users or processes using NFS?
  - Are there a lot of read/write operations taking place over NFS simultaneously?

  If either, or both, of the above are true, the probable cause is that the machine is running out of resources somewhere.
- Are other machines on the same subnet having problems?

  If all the clients on a certain subnet are having problems, and the server is on another subnet, the likely suspects are the network hardware and configurations between the subnets.
- Are other machines *not* on the same subnet having problems?

  The finger of suspicion tends to point at the server in this case, but this could still be a network problem as in the previous paragraph.
- Is the server having problems?

  If the machine is only running one test that copies a file and the copy takes an order of magnitude longer than it should, it is likely that there is something being overrun in either the network or on the server.

When you suspect that the machine is overrunning a network or server resource, it is often helpful to try the *one biod* test. For AIX 4.2.1 and later, the test can be run by unmounting and remounting the file system using the `mount -o biods=1` command. For earlier systems, simply run the `stopsrc -s biod` command; this leaves just one kbio running.

- If the test runs faster with only one biod or kbio, something is being overrun.
- If the standard mount is using UDP, try using a TCP mount.

  TCP mount is the default from AIX V4.3.0 onwards.

If you have determined from the above steps that your system needs to be tuned to provide more resources for adequate NFS performance, the next step is to put in place a controlled environment to evaluate what effects your tuning efforts are having on the system. This test environment should be as close as possible to your production environment as variations in system or network load can completely invalidate the results.

You then need to create an easily repeatable test case that shows the problem. For example, if the problem is read and write performance, the best way to test is to simply use the `cp` command to copy a file over NFS. If copying a file to the client (testing reads), it is best to copy to /dev/null in order to eliminate the disk write time from the test.

When you have established a test for benchmarking, you need to evaluate how your performance tuning affects the system. The simplest method is to track the elapsed time for completion of the test. The simple criteria is that if it runs faster, you have helped, and if it runs slower, you have done something wrong. The easiest method of timing execution is to use the time command as shown in the following example:

```
time cp /servera/testdir/1MBfile /dev/null
```

Another way of monitoring your test results is to use the `nfsstat` command and look for any dropped packets as follows:

1. Use the `nfsstat -z` command to zero the stats.
2. Run the test.
3. Use the `nfsstat -cr` command to see how many packets were dropped.

The following example shows that twelve packets were dropped somewhere on a UDP mount. Often retrans and timeout will not be the same number, since all timeouts do not necessarily result in a retransmit.

```
# nfsstat -cr
```

```
Client rpc:
Connection oriented
calls      badcalls   badxids    timeouts   newcreds   badverfs   timers
61663      0          0          0          0          0          0
nomem      cantconn   interrupts
0          0          0
Connectionless
calls      badcalls   retrans    badxids    timeouts   newcreds   badverfs
2851       0          12         0          12         0          0
timers     nomem      cantsend
0          0          0
#
```

In general, you will see that as the number of dropped packets go down, then test completion time will decrease.

Note that in AIX V4.2.1 and later, you need to look at both Connectionless (UDP) and Connection oriented (TCP) stats to see if there are timeouts and retransmits.

If there are dropped packets, you should first be sure that there are no resource problems on the client. This is done by checking the `netstat` command output.

If there are no dropped packets, you should suspect problems with the application, or that more biods are required.

The `nfsstat` command allows monitoring of statistics on NFS and RPC operation and performance.

To find where packets are being dropped, use the `netstat` command with various options as shown in the following examples:

- Check the `netstat -in` command output for `Oerrs`:

  ```
  # netstat -in
  Name  Mtu   Network     Address              Ipkts Ierrs    Opkts Oerrs  Coll
  lo0   16896 link#1                           2107     0     2109     0     0
  lo0   16896 127         127.0.0.1            2107     0     2109     0     0
  lo0   16896 ::1                              2107     0     2109     0     0
  en0   1500  link#2      2.60.8c.2e.e1.c9     24068     0      819     0     0
  en0   1500  192.168.1   192.168.1.10         24068     0      819     0     0
  tr0   1492  link#3      10.0.5a.a8.70.67   1102168     0    84137   406     0
  tr0   1492  9.185.112   9.185.113.7        1102168     0    84137   406     0
  #
  ```

  `Oerrs` normally indicate that you need to increase the device transmit queue size. In the case of the above example, since there are `Oerrs` showing against the tr0 interface, you would need to change the tok0 xmt_que_size using the following command syntax:

  ```
  # chdev -P -l tok0 -a xmt_que_size=120
  ```

- Check the `netstat -s` output for *non-zero* counts:

The `netstat -s` command reports all of the networking statistics for IP, UDP, TCP, ICMP, and so on. There are just a few of the counters or statistics that can be helpful in determining if there are any underlying network or device configuration problems that may lead to a performance problem. In the IP statistics section, these are the `bad header checksums` and the `fragments dropped` counters. In the UDP section statistics, the `bad checksum` counter and the `dropped due to no socket` counters are the ones to check for non-zero.

- To display only the non-zero statistics of the `ip:` section, use the `netstat -s -s` form of the command:

  ```
  # netstat -s -s -p ip
  ```

  Check the output for the following fields:

  - `bad header checksums` (network problems)
  - `fragments dropped` (device driver queues not large enough)

- To display only the non-zero statistics of the `udp:` section, use the `netstat -s -s` form of the command:

  ```
  netstat -s -s -p udp
  ```

  Check the output for the following fields:

  - `bad checksums` (network problems)
  - `dropped due to no socket` (receive space too small)

  Non-zero counts in the above fields can be an indication of dropped packets.

Note

Using the `netstat -s -s` form of the command as above greatly reduces the volume of output by showing only non-zero statistics counters.

- Check the `netstat -v` command output for network interface statistics.

  The `netstat -s` command reports statistics for the ip level and above. For specific network device statistics, the `netstat -v` command can be used. To help determine if a specific network interface is dropping packets, look for the following statistics to be non-zero:

  - transmit/receive errors
  - transmit/receive packets dropped

  Non-zero counts in the above fields indicate either queue sizes too small or adapter device problems.

The `Max Packets on S/W Transmit Queue` and `S/W Transmit Queue Overflow` statistics can also be helpful in determining appropriate queue sizes for your particular configuration.

- Check the `netstat -m` command output for your systems mbuf/network memory usage.

  There is a limited number of mbufs or memory allocated for use by the network sub-system. If the system is low on this memory, and packets are received, it is possible that the packets will be dropped because there is not enough room for them.

  The `netstat -m` command reports these mbuf memory allocation failures for the system. This statistic is `requests for memory denied`.

  Non-zero counters for `requests for mbufs denied` or `failed` counts in `Kernel malloc statistics` section can indicate a condition that can lead to dropped packets.

  If the `netstat -m` output shows non-zero `requests for mbufs denied`, the `no` command can be used to increase memory allocated to general mbuf use. Refer to Section 7.4.10, “no” on page 181 for command syntax.

- If after going through the preceding steps there is no indication of dropped packets, you can try increasing the number of client biod processes.

  There is no set of rules on how many biods to run; normally, the default settings give adequate performance. However, there are situations where many biods/kbios can be consumed in reading one file. If this is the case, and it is likely that this system will have multiple reads or writes occurring concurrently, performance may be improved by increasing the number of biods from the default number.

  The default number of kbio or biods allowed on one file system varies depending on the AIX and NFS levels:

  - NFS V2 on AIX versions prior to 4.2.1 had six biods.
  - AIX Version 4.2.1 and later has six kbios on NFS V2 and four on V3.

  Any increase in the number of biods/kbios needs to be done and monitored on a trial and error basis. You need to increase the number of biods until performance levels off or packets start to be dropped:

  - To increase the number of biod daemons on the client to 16 when running AIX Version 4.2.0 and earlier, use the following command syntax:

    ```
    # /usr/sbin/chnfs -b 16 -B
    ```

- To increase the number of kbio threads on a client mount point to 16 when using AIX Version 4.2.1 and later, include `-o biods=16` in the mount command:

```
# mount -o rw,intr,bg,timeo=2,biods=16 thor:/usr/tools /usr/tools
```

Note

If you are seeing dropped packets on the machine, increasing the number of biods is likely to cause dramatic performance degradation.

7.6.3.1 Detuning machines for slow servers or networks

When the `nfsstat` command shows that the machine is having to retransmit packets, and you suspect that the machine is overrunning a network or server resource, there are two things you should initially look into:

1. If the mount is over UDP, try using TCP if both the server and client support it.
2. If that does not work, it is often helpful to try the 1 biod test.

To do this for versions of AIX prior to V4.2.1, you just execute the `stopsrc -s biod` command. That will leave the 1 kproc biod running and will slow down the client read/write speed. For AIX V4.2.1 and later, unmount the remote file system and then remount the file system using the `-o biods=1` option on the `mount` command.

If the NFS throughput increases when this is done, you have run into a problem where the machine is configured to run too fast for the network or server.

Sometimes, this can be fixed by locating the problem (network device or server) and correcting the problem. The optimum solution is *not* to de-tune the client, but to fix whatever the bottleneck is in either the network or on the NFS server. If this cannot be done, then there are two major ways of de-tuning a client:

1. The first is to reduce the number of biods, as is done in the 1 biod test.
2. The second is to reduce the read/write sizes.

Both of these will place less load on the server and network, but there are subtle differences.

Decreasing the number of biods can be done by reducing the number of biods that are available to run on any one file system, thus limiting the number that can be working on any one file. The number allowed per file system is

changed using the `mount` command with the `biods=X` option on the file system(s) in question.

A better option is often to change the read/write sizes on the machines rather than changing the number of biods. The default size of an NFS read or write in NFS Version 2 is 8192 bytes per RPC read/write request. Therefore, when 8 K of read or write data is sent over the wire, with a typical MTU of around 1500 bytes, the read or write will be fragmented into six packets. On NFS V3, the problem is magnified because the read/write size is 32 K and number of packets goes up to 23. The loss of any single packet causes a timeout and retransmission, causing all packets to be resent. The idea behind changing read/write sizes is to reduce the number of packets necessary to satisfy a read or write request and increase the chances of success of the initial call. The read/write sizes are specified for the mount command for the file system(s) in question.

The best general strategy is to start with two or three kbios and read/write sizes of 1024. Then the numbers can be adjusted up and down to achieve best performance.

7.6.4 Locking hangs

Use the following checklist to investigate NFS locking problems:

- Check name resolution. Refer to Section 7.2.3, “Name resolution problems” on page 156 if necessary.
- Use the `rpcinfo` command to make sure daemons are running. The required daemons are listed in Section 7.6.1, “General steps for NFS problem solving” on page 191.
- Start the lock daemon with debugging turned on. This would normally only be done under the direction of support personnel and has been included here for information purposes.

7.6.4.1 Starting rpc.lockd with debug

The method for capturing lockd debug output changed at AIX 4.2.1. Prior to AIX 4.2.1, the output went directly to the system console. For AIX 4.2.1 and later, the lockd output is captured with syslogd, therefore syslog needs to be configured by editing the /etc/syslog.conf file to send the debug output to a file. The /etc/syslog.conf file needs an entry similar to the following:

```
*.debug /tmp/cons.out
```

This will send all messages of severity debug and higher to the /tmp/cons.out file. The /tmp/cons.out file must exist first because syslogd will not create it.

Refresh the syslog so it starts using the new file with the command:

```
# refresh -s syslogd
```

When you have done the above, restart `lockd` with the `-d1` flag using one of the following commands and it will log debug messages to the cons.out file.

```
# rpc.lockd -d1
```

```
# startsrc -s rpc.lockd -a -d1
```

Note

The `-d` flag is not documented. Values 1 to 5 will give adequate debug information.

7.6.5 NFS client system boot hangs

When an NFS client system hangs during system restart, there is probably an automatic mount specified in the /etc/filesystems file that is not done in the background.

If you are experiencing boot hangs, check the NFS stanzas in /etc/filesystems to make sure that the `bg` option is specified with the `mount` command.

All automatic NFS mounts, those mounts that are performed automatically at system startup, whether from /etc/filesystems or startup scripts, must have the background option set.

The following example shows an incorrectly setup automatic mount that will cause the system to hang upon system restart:

```
# vi /etc/filesystems
/usr/local:
     dev             = /usr/local
     vfs             = nfs
     nodename        = thor
     mount           = true
     type            = bsd
     options         = hard,intr,fg
```

The `mount = true` entry indicates an automatic mount; therefore, the options entry should be changed to read: `options = hard,intr,bg`

7.7 Serial Line Internet Protocol (SLIP) debugging

The initial step in any SLIP debugging is to test the physical link. This can be done using either the `ate` or `cu` commands; the `ate` command needs the bos.net.ate fileset installed. To test using `ate`:

1. Physically connect the modems or direct connect cables to serial ports on both systems.
2. Create ttys on the above ports, set `Enable LOGIN` on the calling port to `disable`, the called port to `enable`, and `FLOW CONTROL to be used` to `rts`.
3. Enter the `ate` command on the calling system, then:
   a. At the Unconnected main menu, select the `Alter` subcommand. Set the Rate to the baud rate of your modem and the Device to the tty used.
   b. At the Unconnected main menu, select the `Connect` subcommand. When ATE prompts you for a phone number, enter the phone number of the target system modem and press **Enter** (for leased line or direct connection, just press **Enter**).
   c. At this point, you should receive a login prompt. If this is the case, press **Ctrl-v** to return to the connected screen, press `t` to logout, and press `q` to exit ATE.

> **Note**
>
> If you do not receive a login prompt, return to the beginning of this section and verify that your configuration is correct. Do not proceed with any further SLIP debugging until you can login to the target system.

4. Use `smit chgtty` to change the Enable LOGIN setting to disable on the target system tty.

When you have verified the connection is good, as above, you need to check the tty entry in the /etc/uucp/Devices file on both systems to ensure that it reflects the correct tty and connection details. The following example shows the required entry for a 9600 baud connection on tty1:

```
Direct tty1 - 9600 direct
```

You then need to use the `smit chinet` SMIT fastpath to check the SLIP network details on both systems. Pay particular attention to the following fields:

- Source and destination Internet addresses
- Baud rates selected

Check the /etc/hosts file for the above addresses and unique interface names.

If you are using sliplogin, be sure that the user exists in the /etc/slip.hosts configuration file and that the remote IP address, the local IP address, and the netmask are supplied correctly.

If you are calling into a sliplogin server, try using the /usr/sbin/slipcall sample shell script to dial and set up the SLIP connection.

Useful information can also be obtained using the Debug_Level parameter with the `slattach` command. Debug levels 0 through 9 may be used to produce debug output.

7.8 Asynchronous Point-to-Point Protocol (PPP) debugging

This section contains debugging information for AIX PPP client and server connections.

7.8.1 AIX as a PPP client (outgoing calls)

In order to dial out with PPP, there must be a tty defined for the modem port with `Enable LOGIN` set to `disable` and `FLOW CONTROL to be used` set to `rts`. To test the tty, you first of all need to ensure that BNU utilities are installed, check for bos.net.uucp fileset on the system, and then execute the `cu -ml ttyXX` command. Refer to the following `cu` example for the session flow:

```
john@rios:/home/john > cu -ml tty8
Connected
at
OK
~[rios].
The connection is ended.
```

Note

Do not proceed with further PPP debugging until the modem responds as above. If additional debug output is needed, use `cu -dml ttyXX`.

Another useful tool in PPP debugging is syslogd. If this has not already been set up as part of the initial PPP configuration, you need to set it up now. The following setup example can be used as a guide:

1. Add the following line into the /etc/syslog.conf file:

   ```
   *.debug  /tmp/ppp
   ```

2. Create the file /tmp/ppp and set the permissions so it can be written to:

```
>/tmp/ppp
chmod +w /tmp/ppp
```

3. Tell syslogd that /etc/syslog.conf has been updated. Execute the following command:

```
refresh -s syslogd
```

Also check that the /etc/uucp/Devices file contains the following line:

```
Direct tty## - baud_rate direct
```

Where `tty##` is the tty created above and `baud_rate` is the baud rate set for that tty.

Try to start the PPP subsystem either via the smit ppp screen or execute the following command from the command line:

```
startsrc -s pppcontrold
```

Note

Any changes to the link control configuration require that the PPP subsystem (pppcontrold) be stopped and restarted. Use either SMIT or `stopsrc -cs pppcontrold` to stop the subsystem.

Use the following commands to check that the PPP subsystem started:

- `netstat -in` (to see whether pp# network interfaces have been created). Before a connection is established, the IP address of PPP interfaces will be 0.0.0.0.
- `lssrc -s pppcontrold` (to see whether PPP is running)

If PPP is not running, there may be problems in the link control configuration (check with `smit ppp`), or PPP fileset updates are needed. To provide more detailed debugging output from the `pppcontrold` command, you can send signal 30 as follows:

```
lssrc -s pppcontrold
```

Note the Process ID (PID):

```
kill -30 <pppcontrold_PID>
```

Where `pppcontrold_PID` is the PID number of `pppcontrold` returned by the `lssrc` command.

This will append a message indicating that debugging has been activated to the /tmp/ppp file that was set up with syslogd configuration. Diagnostic output

can later be switched off by using `kill -31 <pppcontrold_PID>`. This option can be used when the PPP connection is being established and thereby provide protocol, addressing, PAP/CHAP, and other information.

If the link is not started when the `pppattachd` command runs:

- Check the route tables on both systems.
- Ensure that the dial string from the Chat Script matches the modem requirements.

The `pppattachd` command calls the pppdial program which uses the chat script to do the actual dial-out. Refer to the following for a sample dial-out command:

```
/usr/sbin/pppattachd client tty8 connect "/usr/sbin/pppdial -v -f
CHAT_SCRIPT_FILE"
```

- After issuing the `pppattachd` command, the progression of the dial-out can be watched by executing the following command:

  ```
  tail -f /tmp/ppp
  ```

  Where `/tmp/ppp` is the file setup with syslogd to which debug output has been directed.

The following configuration files also contain useful data for connection problem determination:

- /etc/ppp/if_conf

  Contains PPP IP interface configuration.

- /etc/ppp/lcp.conf

  Contains configuration information set in the LCP config panels.

- /etc/ppp/if_link.map

  Contains the correlation of interfaces to LCP network connection blocks.

- /etc/ppp/attXXX.pid

  Contains process ID files for the async attachments.

7.8.1.1 Chat scripts

The pppdial program called by the `pppattachd` command uses a UUCP chat dialog to establish connection with the remote system. The following example shows a simplified chat script file with explanatory notes to assist with PPP troubleshooting:

```
''
ATDT555-5555
```

```
CONNECT
''
in:
myuserid
word:
mypassword
```

On a line-by-line basis the above script means:

- Expect nothing.
- Send the modem `ATDT555-5555` (to make the modem dial this number).
- Expect `CONNECT` from modem.
- Send nothing.
- Expect [log]in: (the login prompt sent by the remote system).
- Send my userid.
- Expect [pass]word: (the password prompt sent by the remote system).
- Send my password.

7.8.1.2 Authentication problems

The specific reason for failure to authenticate when using PAP or CHAP security authentication protocol is normally shown in the syslogd log file (/tmp/ppp in section examples).

If you are using CHAP authentication, note that Microsoft uses a CHAP algorithm that differs from that of AIX. The Windows 95/Windows NT CHAP protocol is incompatible with AIX.

If using PAP, use the following simplified authentication setup procedure to see if authentication works. In the PAP Authentication SMIT panel:

- The remote host name is the name of the authenticator from the peer's perspective.
- A * for remote host name indicates any authenticator.

On the server machine:

```
smitty ppp
  PAP  Authentication
   Add a User
            User name             [david]
            Remote host name      [*]
            Password              [david]
```

To verify the pap user has been created:

```
vi /etc/ppp/pap-secrets
```

Look for david * david in the file:

```
cd /home/goliath
```

Where `goliath` is the user created for login on the server

```
vi .profile
```

Add the following line:

```
exec /usr/sbin/pppattachd server authenticate pap 2>/dev/null
```

On the client machine:

```
smitty ppp
PAP  Authentication
   Add a User
          User name             [david]
          Remote host name      [*]
          Password              [david]
```

Use the following command to connect as the client:

```
/usr/sbin/pppattachd client /dev/tty# peer pap USERNAME connect
"/usr/sbin/pppdial -v -f CHAT_SCRIPT_FILE"
```

Where `USERNAME` is the name of the PAP user and `CHAT_SCRIPT_FILE` is the full path name of your chat script file.

Dial from the client, get the link established, and the authentication should work. If not check the syslogd output file for the reason.

7.8.2 AIX as a PPP server (incoming calls)

In order to accept incoming calls with PPP, there must be a tty defined for the modem port with `Enable LOGIN` set to `enable`, `delay`, or `share`, and `FLOW CONTROL to be used` set to `rts`. Refer to Section 7.8.1, “AIX as a PPP client (outgoing calls)” on page 203 for tty test details.

Also, set up syslogd as was done in the client section.

Verify that the PPP subsystem is active on this server system with the following commands:

- `lssrc -s pppcontrold`

  This should show as active.

- `ps -ef | grep ppp`

  There should be a pppcontrold process running.

- `ifconfig pp0`

The pp0 interface should display the correct IP address.

If you are unable to login after establishing the connection, check that the user is a member of the UUCP group.

Also check the users .profile for the first line as follows (for ksh):

```
exec /usr/sbin/pppattachd server 2>/dev/null
```

Chapter 8. X11 and graphics

This chapter describes graphics components and how to handle problems related to them.

8.1 Handling graphics devices

Before running any graphics software or application, the underlying hardware components must be in place.

These components must be correctly configured and the capabilities of the hardware understood. Often a graphics problem may be isolated to the fact some underlying hardware does not support a graphics feature called at some higher level.

If you suspect a problem at this level, use the following checklist:

1. Configure the hardware correctly.

   If you have moved the hardware, ensure you do not have conflicting ODM entries for the same hardware in different slots or the adapter is mis-seated. Graphics hardware often occupies multiple slots, especially the high function 3D adapters.

   Refer to Section 8.1.1, “Removing devices” on page 210, and 8.1.2, “Hot plugging” on page 210 for more information.

2. Install the correct device drivers.

   Without the required software support filesets, your hardware will not function.

   Install these as required for your hardware via the command `smit devinst` to install device support filesets. Select from your installation media the applicable filesets from the list.

   A fresh installation will automatically install required filesets; however, the addition of new devices not present at initial AIX installation will require additional device support filesets to be installed.

   If the device support software requires modules to be loaded into the kernel, often a system reboot will be required.

3. Run advanced diagnostics against the hardware.

   If you have installed the required diagnostics software for your hardware, run the `diag` command and select **Advanced Diagnostics** then **Problem Determination**. If your devices support a diagnostic module, test it from

diagnostics. Ensure you are not currently running the X server as this often conflicts with the diagnostic routines.

You may also boot the machine in Service mode boot and load diagnostics from a Diagnostic CD-ROM. Refer to Chapter 5, “Hardware problem determination” on page 89 for more information.

4. Compare with another workstation.

   Try to isolate a problem with a specific graphics device by substituting it with another device or similar device, if possible.

   Note in your problem report if the problem appears on all hardware used or only a specific adapter or other graphics hardware.

8.1.1 Removing devices

Often if an adapter is removed or moved to other location in the machine, the ODM may not reflect the correct status if the `diag -a` command is not run. This command prompts you to remove missing devices from ODM or add newly detected devices. Generally, use this to remove old devices in old locations and indicate the new device is valid for a new location.

Failure to do this may require a manual check of the output of the `lscfg` command against what is presently in each slot in the machine. If conflicting or overlapping devices are present, often the case with multi-slot graphics adapters, the results are unpredictable.

8.1.2 Hot plugging

Physically disconnecting and reconnecting the keyboard, mouse, tablet, LPFKs, and Dials to a workstation after they have been configured with power turned on may cause problems, even though the devices may appear to function correctly. This is especially true for the keyboard.

If you physically detach a device from the workstation, then either reboot the machine, or:

1. Reconnect the device.
2. Unconfigure the device using the `rmdev` command.
3. Reconfigure the device using the `mkdev` command.

8.2 X11 component isolation

Using X11 implies the use of two main components: an X server and an X client. The X server drives the screen or display output and input from devices such as a keyboard and mouse.

Therefore, the task to isolate graphic related problem starts with isolating the problem to one of the following applicable components:

- X server
- X client or application
- Transport between the X client and X server

8.3 X server

Before you can start an X client, the client must have an X server to connect to. The X server may execute on the same machine as the client or on another machine.

8.3.1 Start X server

Locating the source of the problem depends on how the X server is started.

For a non-Xstation, the process /usr/lpp/X11/bin/X is started together with some initial clients as shown below:

- CDE

  AIX Version 4.3 uses graphical boot by default when a graphics monitor is detected on a machine.

  You can disable graphical boot without disabling CDE:

  ```
  /usr/dt/bin/dtconfig -enograph
  ```

  If CDE is not working properly, it is a good idea to boot-up your workstation without CDE.

  To enable graphical boot, run:

  ```
  /usr/dt/bin/dtconfig -e
  ```

  The default startup script for the X server is /usr/lpp/x11/defaults/xserverrc.

- XDM

XDM is also known as X display manager. Its major function is to provide X login. The default startup configuration file to start the X server is /etc/dt/config/Xservers.

- Startup script

  User can start the X server with the `startx` or `xinit` commands.

- Xstation

  Once an Xstation boots, it is already running the X server. X clients can begin to connect straight away.

8.3.2 Failure to start

Prior to starting the X server process, it is assumed the graphics screen can produce a character login or prompt. If you are not at this stage, perform hardware diagnostics and general problem reporting first.

Before attempting to perform problem diagnosis on an X11 startup problem, try starting the X server using the default configuration files. The files used will depend on how the X server is configured to start. This will isolate the problem to a genuine X server problem, not a configuration problem.

If the X server fails to start, check these points:

1. The AIX windows Runtime Environment fileset, X11.base.rte, is installed and at the correct level:

   ```
   # lslpp -l X11.base.rte
   ```

   Verify the X server file exists and has execute permissions:

   ```
   # ls -al /usr/lpp/X11/bin/X
   -rxwr-xr-x 1 bin bin 2611202 OCT 21 1998 /usr/lpp/X11/bin/X
   ```

   Use the `lppchk` command to check the fileset for missing files or incorrect size files:

   ```
   # lppchk -c X11.base.rte
   lppchk:0504-206 File /usr/lpp/X11/bin/X could not be located.
   ```

   The easiest and fastest way to recover the situation mentioned above is to back up all your graphic configuration files if you have made any change on it, then reinstall the relevant filesets.

2. Libraries and device drivers installed.

   The X server may call other libraries or device drivers when executed.

   Try to execute the X server manually to check for error messages. If it does not load, you may see errors similar to the following example:

```
# /usr/lpp/X11/bin/X
exec():0509-036 Cannot load program /usr/lpp/X11/bin/X because of the
following errors:
0509-023 Symbol aixgsc in ksh is not defined
0509-026 System error:cannot run a file that does not have a valid
format
```

In our example, a symbol, `aixgsc`, cannot be found in any libraries on the system. Knowledge of the symbols within libraries is required to locate the symbol.

3. Wrong display.

   The X server maybe running correctly but on the wrong display. Use the `lsdisp` command to display all available adapters and the `chdisp` command to modify the current display.

   For example, to permanently change the display to use the other graphics adapter, first list the available devices, then change it, and then reboot the system:

```
# lsdisp
DEV_NAME SLOT BUS ADPT_NAME DESCRIPTION
======== ==== === ========= ===========
wga0     0J   sys POWER_Gt1x POWER Gt1x Graphics Adapter
bbl0     0J   buc GXT150    GXT150 Graphics Adapter
Default display = wga0
# chdisp -p bbl0
# shutdown -Fr
```

   If you have only one display, remember to change the cabling so the output is shown correctly.

4. Logging errors.

   Examine the system error log as shown in Section 2.3, “Viewing the error log” on page 18, and examine the specific /tmp/xlogfile X log file for messages that appear when an attempt to start the X server is made.

8.3.3 Hung or stopped server

If the X server starts executing but then terminates or stops abnormally, check against the following symptoms.

You may see the following symptoms indicating an X server problem:

- The screen blanks out and returns to a login prompt.
- A core file is created.
- The cursor appears frozen.

- All windows are no longer refreshed.

  Attempting to click on icons does not open any X clients, such as a window. If you can still select windows, obscured parts when revealed are no longer redrawn for all windows.

- You are unable to bring up another aixterm.

The following indicates that the problem is something other than an X server problem:

- You can still start X clients remotely.
- Power management.

  Check the power indicator LED on the physical monitor you are viewing. If the power led has switched to standby mode, typically indicated by a non-green led, this is merely the power management subsystem conserving power.

- The cursor responds to the mouse.

8.3.3.1 Server terminates

When the X server core dumps, it creates a core file. If there is sufficient space and you have write authority, then the core file exists in the directory where you started the X server.

If a core file exists, the following steps can help locate which function caused the core dump:

1. Use `cd` to change to the directory that contains the core file.
2. Invoke:

   ```
   dbx /usr/lpp/X11/bin/X core
   where > coredump_x.log
   quit
   ```

   Ensure the core file actually came from the X process. When you run `dbx`, it will display the executable name like this: (X). If it is not (X), you have the wrong core file.

8.3.3.2 Server hangs

If you are experiencing a hang, execute the following:

1. Telnet into the hung machine from another machine.
2. Invoke:

   ```
   ps -ef | grep X
   ```

   This generates, for example:

```
userid 11634 11377 0 09:35:30 pts/1 0:33 /usr/lpp/X11/bin/X -x dps -D /usr/lib/X11/rgb
```

The first number, in this case 11634, is the process ID.

3. Invoke:

   ```
   dbx -a pid
   ```

   Where `pid` is the process ID associated with the X process. Therefore, in this example, you would issue:

   ```
   dbx -a 11634
   ```

4. To find the location of the hang, type:

   ```
   where > hang_x.log
   detach
   pg hang_x.log
   ```

   To kill the X server when it is hung, you can issue:

   ```
   where > hang_x.log
   quit
   pg hang_x.log
   ```

   However, ensure you have collected all of the necessary information for reporting the problem before killing the X server.

5. With applications that also use the X server (such as graPHIGS remote nucleus), the dbx trace can be valuable input to IBM as well. Follow the same procedures listed above to obtain a dbx trace for the application you were running at the time the hang occurred. Instead of grep X, grep the name of the application. Also, to avoid overwriting your hang_x.log file, issue the `dbx` subcommand:

   ```
   where > hang_app.log
   ```

   Ensure that the time stamp on the file(s) you are including with your problem report have the appropriate time stamp (that is the time at which the failure occurred):

   ```
   ls -l file
   ```

   Where `file` is xlogfile, core, or any other pertinent file.

8.4 Connecting X clients to the X server

The DISPLAY environment variable indicates what server a client will connect to and the transport method it will use.

If this is incorrectly specified, this can often stop an application from starting or achieve less than optimal performance.

8.4.1 The DISPLAY variable

An X client knows what X server to connect to by the value of the DISPLAY environment variable.

This variable has the general form:

```
TCP/IP host : display# [ . screen# ]
```

Where `TCP/IP host` is the TCP/IP host name defined for the X server, `display#` is a number representing the display number, and `[ . screen# ]` optionally indicates a screen within the display.

For example, to display X clients on the first display on the TCP/IP host mercury, set DISPLAY as follows:

```
mercury:0
```

For the Korn shell (/bin/ksh), the command to set the variable for your current process is:

```
export DISPLAY=mercury:0
```

For the C shell (/bin/csh), the syntax to set the variable is:

```
setenv DISPLAY mercury:0
```

For the Bourne shell (/bin/bsh), the syntax to set the variable is:

```
DISPLAY=mercury:0
export DISPLAY
```

The remaining examples in this section assume the use of the Korn shell.

The current setting of the variable can be found with the command:

```
echo $DISPLAY
```

Before resetting the variable, write down the current setting.

8.4.2 X transport

The transport mechanism between an X client and X server allows X clients to execute on remote hosts or on the same machine as the X server.

The transport method is implicitly defined by the value of the DISPLAY environment variable, or the X client code specifying the parameter directly to the OpenDisplay routine.

This flexible mechanism has three forms:

- Shared Memory Transport (SMT)

  This is the fastest transport method and only available for clients running on the same host as the X server process. To set the DISPLAY variable, omit the host name portion, for example:

  ```
  export DISPLAY=:0
  ```

- UNIX Sockets

  This is only applicable to for clients running on the same host as the X server process. To set the DISPLAY variable, use the following syntax:

  ```
  export DISPLAY=unix:0
  ```

- TCP Sockets

  This is applicable to clients executing remotely, for example, on a Xstation. To set the DISPLAY variable, use a valid TCP/IP host name, for example:

  ```
  export DISPLAY=ted:0
  ```

8.4.3 Testing client/server connectivity

Test the validity of your DISPLAY variable by executing a simple well-known client.

For example, if your application does not seem to connect to your X server, execute:

```
xsetroot -solid white
xsetroot -solid red
```

Does your X server screen change color?

If it does not, check each of the following:

- DISPLAY not set or malformed

  ```
  # echo $DISPLAY
  ted2
  # xsetroot -solid white
  1356-265 xsetroot: Unable to open display: ted2.
  ```

  In this example, the DISPLAY variable is not set correctly. It is missing :0 at the end.

- Invalid DISPLAY host name

  ```
  # echo $DISPLAY
  bogus:0
  # xsetroot -solid white
  1356-265 xsetroot: Unable to open display: bogus:0.
  # ping bogus
  ```

```
0821-062 ping: host name bogus NOT FOUND
```

In this example, the TCP/IP host name component of the `DISPLAY` variable refers to a host that does not exist.

- Client hangs

```
# xsetroot -solid white
1356-265 xsetroot: Unable to open display: rally:0.
```

In this example, there is a long pause or hang after running the `xsetroot` command. Attempt to `ping` the host and check if the X server is running on that host. Perhaps your client does not have a route to the X server.

- Connection refused

```
# xsetroot -solid white
Xlib: connection to "mossad:0.0" refused by server
Xlib: Client is not authorized to connect to Server
1356-265 xsetroot: Unable to open display: mossad:0.
```

The X server maintains an access control list, allowing certain hosts to be denied access. For a machine running an X server on AIX, the default access control list disables all other hosts.

To enable all hosts to connect, use the `xhost` command. For example, run the following command on the same host as the X server is running:

```
xhost +
```

The `xhost` command is part of the X11.apps.config fileset, which, depending on the level of AIX you are running, may not be installed by default.

- Missing output

```
# xsetroot -solid white
```

If the command executes without error and without delay, yet the result is not visible, ensure you are viewing the correct screen. If you are using multiple graphics adapters in your machine and/or multiple X servers, you may be looking at the wrong screen.

Note

For CDE environment users, this command will only affect the first workspace. Ensure you select the first workspace from the switch area.

8.4.4 Failed X client start

Now that you have verified the functionality of the X server, run your application in place of the well-known X client above.

If this fails to start, check the following symptoms:

- X protocol error messages

```
# ./my_app
X Error of failed request: BadWindow (invalid Window parameter)
Major opcode of failed request: 1 (X_CreateWindow)
Resource id in failed request: 0x200003e8
Serial number of failed request: 3
Current serial number in output stream: 5
```

  This indicates a run-time error received by the client application.

  For example, an inappropriate operation is performed on the X resource or a requested resource is not available.

  The above example shows the output from the default X error handler. A user application may call a custom error handler by using the function XSetErrorHandler().

  Function isolation is aided by the client application calling XSynchronize() to enable the X server's synchronization of an X protocol request and the X server's completion of the request. This ensures that the application will terminate at the point of the X call causing the problem, rather that after the event due to the X servers buffering of requests.

- exec format errors

```
# ./my_app
exec(): 0509-036 Cannot load program ./my_app because of the following errors:
0509-022 Cannot load library libXm.a[shr4.o].
0509-026 System error: A file or directory in the path name does not exist.
```

  Or

```
# ./my_app
exec(): 0509-036 Cannot load program ./my_app because of the following errors:
0509-023 Symbol _XmStrings in ksh is not defined.
0509-023 Symbol XmStringLtoRCreate in ksh is not defined.
0509-026 System error: Cannot run a file that does not have a valid form
```

  These examples indicate an installation problem, either the libraries referenced by the executable at load time are missing or the version of the libraries is downlevel.

  In the first case, the libXm.a library is missing. It must be installed for the application to run.

  In the second case, symbols cannot be located in the libraries found on the machine. Check with the supplier of the client software to find out what libraries the symbols are expected to be in and what version should be installed. Generally, specific knowledge of the location of the symbols within libraries is required. Refer to Section 9.2.4, "Library problems" on page 229 for more information on determining the missing libraries.

8.5 X kernel extension

The functionality of the X server is extended with the addition of X extensions.

8.5.1 Loading an X extension

The method to load an X extension will depend on the method used to start the X server. Please refer to Section 8.3.1, “Start X server” on page 211 for more information.

8.5.2 Installing verification commands

The commands `xrdb -symbols` and `xdpyinfo` enable you to check for loaded X extensions. The `xrdb` command is part of the X11.apps.rte fileset. The `xdpyinfo` command is part of the X11.samples.apps.clients fileset.

8.5.3 Checking X extensions

With `xrdb`, check for the -DEXT_ entries. For example:

```
# xrdb -symbols
-DHOST=ted2 -DSERVERHOST=ted2 -DSRVR_ted2 -DDISPLAY_NUM=0 -DCLIENTHOST=bob3 -DCLNT_bob3
-DVERSION=11 -DREVISION=0 -DVENDOR="International Business Machines"
-DVNDR_International_Business_Machines -DRELEASE=5 -DNUM_SCREENS=1
-DEXT_SCREEN_SAVER -DEXT_SHAPE
-DEXT_XTestExtension1 -DEXT_xColormapExtension
-DEXT_aixCursorExtension -DEXT_XAixExtension
-DEXT_xDirectAccessExtension -DEXT_XInputExtension
-DSCREEN_NUM=0 -DWIDTH=1280 -DHEIGHT=1024 -DX_RESOLUTION=3596
-DY_RESOLUTION=3606 -DPLANES=8 -DBITS_PER_RGB=8 -DCLASS=PseudoColor
-DCLASS_PseudoColor=33 -DCOLOR -DCLASS_PseudoColor_8=33
-DCLASS_DirectColor_8=34
```

If you are using `xdpyinfo`, check the loaded extensions list. For example:

```
# xdpyinfo | pg
.
.
number of extensions: 8
XInputExtension
xDirectAccessExtension
XAixExtension
aixCursorExtension
xColormapExtension
XTestExtension1
SHAPE
SCREEN-SAVER
.
```

Some APIs may require that extensions to the X server are loaded before running an application.

Check your API below and verify that the extension is loaded:

- OpenGL

- abx (xAncillaryBufferExtension)
- mbx (Multi-Buffering)
- GLX (GLX)

- graPHIGS
  - xgpshm (GP-MIT-SHM)
- PEX
  - mbx (Multi-Buffering)
  - pex (X3D-PEX)

8.6 CDE problem determination

CDE is the default graphic interface on AIX 4.3.

8.6.1 CDE log files

If you run into a problem with CDE, the first place you should look is the log files:

- /var/dt/Xerrors
- $HOME/.dt/startlog
- $HOME/.dt/errorlog
- /tmp/xlogfile

/var/dt/Xerrors contains errors that occur when the Login Server is started. The location of this log file can be changed using the Dtlogin.errorLogFile resource in /usr/dt/config/Xconfig.

Session Manager logs problems with session startup in the $HOME/.dt/startlog file.

Various desktop applications log errors to $HOME/.dt/errorlog. For example, errors that occur when an action is loaded will be logged in $HOME/.dt/errorlog.

X logs errors to /tmp/xlogfile.

It is important that you look at these log files to diagnose problems**.**

We list common problems you may meet and their solutions.

8.6.2 Login screen does not appear

This can happen when customizations to the X servers file were made, but the command does not execute properly.

Check the log files and confirm that X or xinit will start. If X or xinit will not start, go to Section 8.3.1, “Start X server” on page 211.

Next, look for /etc/dt/config/Xservers and see if it was customized. If it was, verify that the command will execute from an lft. If not, look for customizations made directly to /usr/dt/config/Xservers. Customizations of the /usr/dt files are discouraged because they may be overwritten at the next upgrade of CDE. The default /usr/dt/config/Xservers file contains the following line:

```
:0 Local local@console /usr/lpp/X11/defaults/xserverrc -T -force :0
```

To start the X server, Login Manager then executes:

```
/usr/lpp/X11/defaults/xserverrc -T -force :0
```

The `xserverrc` executable is a shell script, so customizations to the start of X could be made directly to the xserverrc file. If no changes to either of the Xservers files are detected, then check to make sure the /usr/lpp/X11/defaults/xserverrc file is executable by all users. File mode 755 is recommended. Also, you can run the script manually to see if there are errors in the script:

```
ksh -x /usr/lpp/X11/defaults/xserverrc -T -force :0
```

8.6.3 Problem with login

The first step here is to check the log files for information. If you can not get enough information from log files, you may try to run Login Manager in debug mode. If the problem is on the console, start dtlogin remotely:

```
/usr/dt/bin/dtlogin -debug 10|tee /tmp/<logfile>
```

If you have to run dtlogin in debug mode, you must first stop all current dtlogin processes. Be sure all CDE users are logged out before killing the dtlogin process, or you may cause your users to lose data.

If X terminates, then you must kill dtlogin then issue the following command to restart the desktop:

```
dtlogin -daemon; exit
```

8.6.4 CDE hangs

Network interface parameters (such as host name and IP address) of a machine should never be changed while CDE is running.

Always exit from the desktop and stop desktop processes to modify network interfaces. If CDE is not halted before something like the host name is changed, exit from CDE and see if you can run `smit mktcpip` to stop and restart the network connection. If this still fails, reboot the machine.

8.6.5 The DT messaging system could not be started

You may get this message in a dialog box when attempting to login to CDE:

```
The DT messaging system could not be started. Login using failsafe, check
hostname at these locations:
/etc/src.sh
/etc/hosts
/usr/adm/inetd.sec
```

AIX only uses the /etc/hosts file, but this message is generic for all of the CDE platforms, so do not be surprised to see files mentioned that do not exist.

The message may indicate:

- Host name changed while still running CDE.
- Multiple machines on the network have the same host name
- If you do not use DNS, you may
  - Set /etc/netsvc.conf to read /etc/hosts:

    hosts = local,bind,nis
  - Set /etc/hosts to recognize both the short name and the fully qualified name of your machine:

    ```
    9.3.187.211 itsosrv1.austin.ibm.com itsosrv1
    ```
  - Make sure `smit mktcpip` and `hostname` both show the same fully qualified name for your machine.
- If you update /etc/netsvc.conf, /etc/hosts, check your host name. If the problem still occurs, modify /etc/hosts by adding your host name as an alias to the loopback interface line and commenting out the original name to address mapping as below:

  ```
  127.0.0.1 loopback localhost itsosrv1 # loopback (lo0)
  #9.3.187.211 itsosrv1.austin.ibm.com itsosrv1
  ```

  Then, rename /etc/netsvc.conf to /etc/netsvc.conf.old and reboot the machine.

8.7 Client libraries

X releases contain two numbers: The version number indicating major protocol or standards revisions and a release number indicating minor changes.

8.7.1 X11R6

At the time of writing, the latest version is X11 Release 6, also known as X11R6. The latest release of OSF/MOTIF is V2.1 (based on X11R5). Major revisions of X are incompatible, but there is backward compatibility with minor releases within major revision categories. AIX 4.3 runs on X11R6 and Motif 2.1.

8.7.2 X11R6 enhancements

Compared to X11R5, X11R6 has a lot of enhancements. Refer to the *AIX Version 4.3 Differences Guide*, SG24-2014, for more details.

8.7.3 X11/Motif compatibility fileset

Users installing AIX Version 4.3 who are concerned about binary compatibility with prior versions of AIX 4 should install the compatibility filesets offered on the installation media.

These filesets offer commands, library versions, symbolic links, and other items that, when added to the system, make it look more like old version system from an application point of view.

While some of these filesets increase disk requirements (substantially, in the case of the AIXwindows X11R3 and X11R4 compatibility packages) and contain obsolete function, the compatibility filesets increase portability in an environment with machines running mixed levels of AIX. Installing the compatibility filesets is highly recommended.

If you performed a Migration Installation, these filesets will already be installed.

The X11 and Motif compatibility filesets are shown in Table 8.

Table 8. X11 and Motif compatibility filesets

Fileset name	Description
X11.compat.lib.Motif10	Motif 1.0 Libraries Compatibility
X11.compat.lib.Motif114	Motif 1.1.4 Libraries Compatibility

Fileset name	Description
X11.compat.lib.X11R3	X11R3 Libraries Compatibility
X11.compat.lib.X11R4	X11R4 Libraries Compatibility
X11.compat.lib.X11R5	X11R5 Libraries Compatibility

Chapter 9. User applications

This chapter attempts to help you perform problem determination on user applications. The problems that can affect user applications are almost as varied and numerous as the applications themselves. There are a huge number of third party applications available for the AIX platform from many different vendors. In addition, some systems are running bespoke applications, developed by the user of a system to perform a customized task. This chapter cannot detail every single problem with every application but, instead, will mention some of the most common problems that can be encountered.

Since most users interact with applications, rather than the AIX operating system itself, when something goes wrong, it is most often first noticed as strange behavior of an application.

In general, the causes of user application problems can be divided into three distinct categories:

- Problems with the application

  It may be that the problem is entirely within the user application. In this case, the AIX operating system is functioning correctly.

- Problems with the system

  In some cases, there may be an underlying problem with the hardware or operating system that affects the user application.

- User error

  In some cases, a user may perceive a problem that does not, in fact, exist, for example, when trying to get the application or AIX operating system to perform a task for which it is not designed.

We will assume that user error has been eliminated as one of the steps in the problem determination process so far. In other words, the user is having a problem doing something that was previously working correctly.

9.1 Problem determination approach

The basic approach that can be taken when looking at user application problems is as follows:

- Can you start the application?
- Are you starting the correct application?

- Is the application configured correctly?
- Does the user running the application have the correct permissions?
- Does the system have enough free resources?
- Are any other processes required, such as background daemons, already running?

9.2 Application startup problems

The reasons for failure of an application to start are many and varied. Some of the most common problems are listed here and should be investigated as part of the problem determination process.

9.2.1 PATH problems

There are three methods by which a user normally starts an application from a shell. The first is to use an absolute path name to specify the executable or shell script to run, for example:

```
# /usr/bin/ls
```

The second is to use a relative path to invoke the application, for example:

```
# ../../bin/ls
```

The third is to just enter or use the name of the application. In this case, the users command shell will search the list of directories specified in the PATH environment variable for the named application. If it finds an executable that matches, it will invoke it.

If you use the PATH environment variable as the method of finding the application, you can encounter problems if the variable is configured incorrectly.

9.2.2 Permissions problems

All of the three methods described above will only work if the user invoking the command has appropriate permissions. In each case, the user must have execute permission for the application itself and all directories on the path specified to the application. If the application is a shell script, the user must also have read permission for the application itself.

9.2.3 Name conflict

Depending on how your system is configured, you may have multiple versions of the same user application command. To determine which

application will be started when you use the PATH environment variable, use the `whence` shell command. For example:

```
# whence ls
/usr/bin/ls
```

This indicates that the shell has parsed the PATH environment variable, and the first instance of the command `ls` that it can execute is /usr/bin/ls. Make sure you are starting the correct application.

You can search for any other instances of an application with the same name by using the `find` command. For example:

```
# find / -name ls -type f -print
/usr/bin/ls
```

9.2.4 Library problems

Once you have determined that you are invoking the correct application, you need to make sure that it can actually start. In order to start, executables need to be able to resolve all of their internal symbols. All executables need to resolve symbols from the kernel when they are started. Most also need to resolve symbols from one or more shared objects. You have symbol resolution problems if you get a message similar to either of the following:

```
exec(): 0509-036 Cannot load program clear because of the following errors:
        0509-022 Cannot load library libdb1.a[shr.o].
       0509-026 System error: A file or directory in the path name does not
exist.
```

Or

```
exec(): 0509-036 Cannot load program myapp because of the following errors:
0509-023 Symbol _XmStrings in ksh is not defined.
0509-023 Symbol XmStringLtoRCreate in ksh is not defined.
0509-026 System error: Cannot run a file that does not have a valid form
```

To determine which shared objects the executable will look for at invocation, use the `dump` command with the `-H` option, as shown in Figure 17 on page 230.

```
# dump -H /usr/bin/app1
/usr/bin/app1:

                        ***Loader Section***
                      Loader Header Information
VERSION#         #SYMtableENT     #RELOCent        LENidSTR
0x00000001       0x00000031       0x00000047       0x0000002f

#IMPfilID        OFFidSTR         LENstrTBL        OFFstrTBL
0x00000002       0x0000080c       0x00000083       0x0000083b


                        ***Import File Strings***
INDEX  PATH                              BASE               MEMBER
0      /usr/lib:/lib:/usr/lpp/xlC/lib
1                                        libc.a             shr.o
2                                        libapp.a           shr1.o
3                                        libapp2.so
```

Figure 17. Sample dump -H output

This command prints the header section of the executable. The important part of the output is contained in the `Import File Strings` section at the end. The entry for `index 0` shows the library search path that was used to compile the executable. This is the path that the system loader will use to search for the specified shared objects. This built in path can be extended by the use of the LIBPATH environment variable. Similar in format to the PATH environment variable, it specifies the list of directories to be searched for shared objects when trying to start an executable.

Subsequent entries list the archive libraries and shared objects that the executable requires to be available at load time to complete the symbol resolution process.

If you get an error message similar to the one described above, you should check to see that the shared objects the executable will look for at start time exist in one of the directories specified in the built in library path.

Some of the reasons for symbol resolution problems are described below.

9.2.4.1 Binary compatibility

The AIX operating system provides upwards binary compatibility between releases. That is, an executable produced on a particular version of AIX will also work on subsequent versions of AIX, providing some other criteria are met. The full binary compatibility statement is as follows:

Applications written using earlier releases of AIX Version 4 (AIX Version 4.1 or AIX Version 4.2) for RS/6000 POWER-, POWER2-, POWER3-, and PowerPC-based models can be executed on AIX Version 4.3 without recompilation for same and newer models in that processor family (POWER, POWER2, POWER3, or PowerPC). The exceptions to this statement are:

- Applications using non-shared compiles of AIX shared libraries
- Applications using features explicitly described as non-portable by IBM in the AIX Version 4 reference manuals
- Applications using non-documented AIX internal features
- Applications using X11R5 Server Extensions (AIX Version 4.3 Only)
- Applications compiled using POWER2, POWER3, or PowerPC-specific compiler options but executed on models other than POWER2, POWER3, or PowerPC, respectively.

Applications compiled on a given release level of AIX Version 4 may not operate properly on systems running an earlier release level of AIX Version 4.

Any program that must run in all environments – POWER, POWER2, POWER3, and PowerPC (601 and newer PowerPC processors) – must be compiled using the common mode option of the compiler. Programs compiled to exploit POWER2 technology must be run on POWER2-based processors. Programs compiled to exploit PowerPC-based technology must be run on PowerPC-based processors. Existing binaries need not be recompiled to operate on the target processors.

64-bit applications produced using AIX Version 4.3 on any of the 32-bit processor models or the 64-bit processor models will execute without recompilation on the 64-bit processor models. 32-bit applications produced using AIX Version 4.3 on either 32- or 64-bit processor models will execute without recompilation on both models.

You can use the `dump -Tv` command to examine the symbol resolution requirements of the executable. Using this command, you can determine the shared object in which the executable expects to find the missing symbols. The output from the command can be quite extensive, depending on the number of symbols that need to be resolved at load time. Pipe the output through the `grep` command, and search for the symbol name mentioned in the original error message.

If the source of the missing symbol is listed as /unix in the output of the `dump -Tv` command, then it is likely that the executable is using an undocumented

AIX interface that has been removed after applying a software patch or was compiled on a newer version of AIX than the one you are trying to run it on.

9.2.4.2 Missing shared object

If you receive an error message stating that a particular library could not be loaded, for example:

```
exec(): 0509-036 Cannot load program clear because of the following errors:
        0509-022 Cannot load library libapp.a[shr1.o].
       0509-026 System error: A file or directory in the path name does not
exist.
```

Then there are a number of possible explanations for the problem. The first step for resolution is to determine if the archive library mentioned exists on the system. Take as an example, the output shown in Figure 17 on page 230. The first step is to search the system for the missing library. If the library exists, but it is not in one of the directories listed in the `INDEX 0` entry, then you may need to set the LIBPATH environment variable to include the directory name. If the executable has the setuid bit enabled, then it will only look in directories listed in the `INDEX 0` entry. It will ignore the LIBPATH environment variable. In this case, you should either copy the library to one of the directories listed in the `INDEX 0` entry, or create a symbolic link.

If you find the library, you should check that it contains the particular shared object that the executable requires. This can be done with the `ar` command. For example:

```
# ar t libapp.a | grep shr1.o
shr1.o
```

If you can find the library on the system, you should check that the user invoking the executable has read permission on the library.

If you can not find the library on the system, then you may need to reinstall the user application or install additional AIX components to solve the problem.

9.3 Application problems

Once you have managed to start the correct executable or shell script, does it manage to keep going, or does it terminate straight away?

9.3.1 Configuration problems

You should ensure that the application is configured correctly for the environment and that the user invoking the application has the correct permissions.

Ensure that any other processes required, for example, background daemons, are also started in the correct sequence.

9.3.1.1 Licensing problems

Many applications use a licensing system to control their usage. Ensure that you have an appropriate number of application licenses for the machine in question.

If the application uses a nodelock license, ensure that the user invoking the application has permission to read the nodelock license file. If the application uses the NetLS or iFOR/LS license system, the nodelock file is normally /usr/lib/netls/conf/nodelock. If the system is using the LUM license system, the nodelock file is /usr /var/ifor/nodelock.

Ensure that the system date and time configuration is accurate. Sometimes, the system date and time can drift, or become out of date, if the system lithium battery fails. The correct system date and time can be significant if a remote license server is being used.

9.3.1.2 Environment problems

You should ensure that the user starting the application is doing so from the correct environment. Many applications require particular environment variables to be configured before working correctly. Check that any configuration scripts run by the user at login time are present and contain the correct information. The main AIX files to check are /etc/profile, /etc/environment, and the $HOME/.profile of the individual user. Depending on the application, there may be many more configuration files.

Ensure that the system running the application meets all the prerequisites, in other words, that the correct system software is installed, and any required hardware components are present.

9.3.2 Permissions problems

Ensure that the user invoking the application has the correct permissions to read or write to any required files.

9.3.3 Resource problems

You may run into resource limits when trying to run the application. There are two types of resource limits that can be encountered. You may encounter system resource limits, where the resources of the system itself become exhausted by the application. The main system resources that can become exhausted are memory and disk space.

If the application makes large memory requests, then the system may eventually run out of paging space. If this condition occurs, then an entry is made in the error log recording the problem. If the condition continues, the system will eventually start terminating processes in an attempt to alleviate the situation.

The other system resource that can become exhausted is disk space. If an application produces output to files, then it may eventually cause a file system to become full.

In addition to experiencing system resource problems, the user invoking the application may have restricted access to the system resources. For example, the user may have a disk quota on a particular file system that, if exceeded, may cause the application to terminate. In this case, it is not the system resource that has become exhausted, but the users allowed usage of the system resource.

As with disk space, a user's memory usage can be restricted. There are many different uses for system memory, and each type of usage can be controlled on an individual basis with the `ulimit` command. Depending on the version of AIX you are running, you may see different output from the `ulimit -a` command, which lists the resources that can be controlled, and the limits for the current user. For example:

```
# ulimit -a
time(seconds)        unlimited
file(blocks)         2097151
data(kbytes)         131072
stack(kbytes)        32768
memory(kbytes)       32768
coredump(blocks)     2097151
nofiles(descriptors) 2000
```

The main limits that can have an impact on the normal running of an application are file, data, stack, memory, and nofiles. Exceeding any of these limits may cause an application to terminate abnormally. The hard limits for each parameter can be set on a per-user basis by the root user and are stored in the configuration file /etc/security/limits.

If you suspect a resource problem of this nature, try changing the ulimit values for the user to unlimited, indicated with the value `-1`, and retrying the application.

Chapter 10. Performance problems

This chapter describes what to do if you suspect a performance problem. The causes of performance problems can generally be divided into three types:

- Resource constraint problems

  A resource constraint problem arises when the applications running on a system try to consume more resources than are available. The three resources that can become exhausted are:

  - CPU
  - Memory
  - I/O (Disk and network)

- Software problems

  A performance problem can be caused by an underlying problem with the application or operating system. The problems can be one of two types:

  - Incorrect configuration

    This is where the application or operating system is functioning correctly as configured, but the configuration is inappropriate for the system. Generally, this tends to lead to a resource constraint problem.

  - Software bug

    This is where the application or operating system does not function as it should.

- Hardware problems

  Problems with hardware devices can cause a degradation in application and system performance.

Most common performance problems are caused by resource constraints. The first step in examining a potential performance problem is to monitor the resources in use by the system. If you can determine that there is no resource constraint, then you should perform hardware diagnostics to determine if you have a hardware problem that is affecting system performance. Refer to Chapter 5, “Hardware problem determination” on page 89 for more information.

If after checking the system resources and hardware you still have a problem, then you will need to collect some data before reporting the problem to IBM.

10.1 Performance bottlenecks

A performance bottleneck is the slowest component in a computer environment. This can either be a system resource, such as CPU, memory, or disk, or it could be the network. There is always a bottleneck because some resource will always be the slowest. The question is whether this bottleneck is a problem on a daily basis.

The sequence of measuring system performance is extremely important. You should always follow the specified path, which is: CPU, Memory, I/O, and Network.

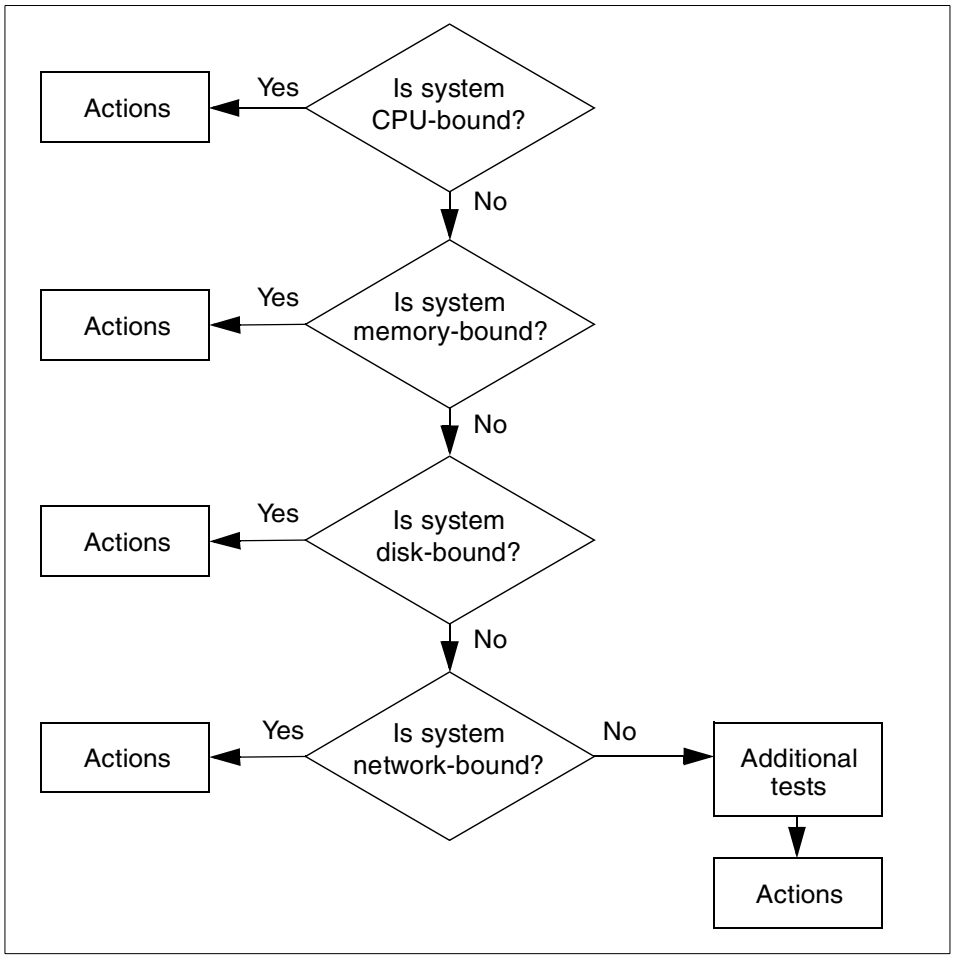


Figure 18. Bottleneck determining process

10.2 Monitoring performance

Due to the heritage of UNIX, AIX provides many tools for monitoring CPU, memory performance usage, and disk usage. If the system appears to be having performance problems, you must identify the bottleneck. Monitoring tools can also be used to detect cases where you are not aware of a problem. Making changes could result in overall better system performance. However, sometimes it could degrade the system performance. Therefore, changing should be carefully done over a long period of time.

In this section, we describe some monitoring tools, how they work, and what they mean.

Before using some commands, you need to install the performance agent package named perfagent.tools and the base accounting package named bos.acct. Use the following commands to check whether these filesets are installed:

```
# lslpp -h perfagent.tools
# lslpp -h bos.acct
```

10.2.1 Monitoring CPU

This section shows the CPU monitoring tools and how to interpret their output.

10.2.1.1 High CPU percentage for kproc processes

When you examine the output of the `ps` command and look at the CPU usage numbers in the column header %CPU, these figures are the percentage of CPU time consumed by that process since it was started. The CPU percentage reported for kproc processes may seem high. The kproc processes are part of the kernel and are started when the AIX system boots. Some kprocs are the idle processes that run when the system is idle. Since this process just indicates the idle time of the system, the high CPU percentage does not indicate a problem and no action is required. Depending on the number of CPUs and the exact software configuration of your system, you may see between two and thirty kproc processes.

The formula for computing the %CPU value is:

%CPU = CPUTime(pid) / (currentTime - startTime(pid)) * 100.0

This equation gives the true percentage of CPU that a process has consumed over its lifetime. On unburdened systems, kproc processes will use 50

percent or more of CPU. Since this computation is based on lifetime statistics, it will be slow to rise and slow to fall.

Additionally, you should see that adding up the numbers in the %CPU column does not yield 100 percent. For example, process A may have been using 80 percent of the CPU for its first hour. Then process B is started at a higher priority and consumes 80 percent. Process A still runs, but now only gets 5 percent of the CPU slices. The %CPU as shown by `ps ug` will gradually decline for process A, starting from 80 percent and eventually leveling off near 5 percent. For a while, both process A and B will show %CPU of 80 percent.

10.2.1.2 Using the vmstat command

The `vmstat` command reports statistics about processes, virtual memory, paging activity, faults, CPU activity, and disk transfers. Options and parameters recognized by this tool are indicated by the usage prompt:

```
# vmstat [-fs] [Drives] [Interval] [Count]
```

This command below displays one line of output every five seconds. Sample output is shown is Figure 19.

```
itsosrv1:/> vmstat 5
kthr     memory             page              faults        cpu
----- ----------- ------------------------ ------------ -----------
 r  b   avm   fre  re  pi  po  fr   sr  cy  in   sy  cs us sy id wa
 0  0 14765 231059   0   0   0   0    0   0 100    3   1 0 0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1207   42  23  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1206   46  26  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1206   38  25  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1207   46  24  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1205   37  23  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1205   23  22  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1206   37  23  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1207   43  24  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1205   28  23  0  0 99  0
 0  1 14765 231059   0   0   0   0    0   0 1210   37  26  0  0 96  4
```

Figure 19. Output of vmstat command

Columns under kthr

The columns under the `kthr` heading in the output provide information about the average number of threads on various queues.

r The `r` column indicates the average number of kernel threads on the run queue at one-second intervals. This field indicated the number of run-able threads. If each one executes for a complete or partial time slice, the number of run-able threads could easily exceed 100.

b The `b` column shows the average number of kernel threads on the wait queue (blocked) at one-second intervals (awaiting resource, awaiting input/output).

Columns under cpu

You need to pay attention to the numbers under the `cpu` label (last four columns) for CPU statistics. These columns are:

us User CPU utilization percentage

sy System CPU utilization percentage

id Percentage CPU idle

wa I/O wait percentage

The sum of all four of these columns will equal 100 percent. If `id` (idle) is consistently zero, or user + system is greater than 80 percent, then the workload may be CPU bound.

If the `wa` (percent I/O wait) is non-zero, some of the workload may be I/O limited. Refer to 10.2.3, “Monitoring I/O” on page 249. Also note the `pi` and `po` activity under page. Paging space page in and page out activity can indicate thrashing.

The other columns under the memory, page, and faults headings will be described in Section 10.2.2, “Monitoring memory” on page 243.

10.2.1.3 Starvation

If the system appears to be having performance problems and responding very slowly, check for threads showing high CPU usage and high priority. These threads could be hogging the CPU.

To see if you have that kind of threads, use the `ps -el` command, as shown in Figure 20 on page 242.

```
# ps -el
F S UID   PID  PPID   C PRI NI ADDR    SZ    WCHAN    TTY  TIME CMD
  200003 A   0     1     0   0  60 20 a00a   372                -  0:45 init
  200001 A   0  5112  7498   0  60 20 acea   708                -  0:00 telnetd
  240401 A   0  5434     1   0  60 20 150f5    36                -  0:00 ssa_daem
on
  240001 A   0  5722     1   2  61 20 b14b    96 511ef298       - 11:09 syncd
   40401 A   0  5978     1   0  60 20 18178   496    cedfc       -  0:00 errdemon
   40001 A   0  6226     1   0  60 20 5165  1596                -  8:18 dtlogin
  240001 A   0  6400  6978   0  60 20 191d9   760                -  0:01 sendmail
  240001 A   0  6826  6978   0  60 20 111d1   236                -  0:00 syslogd
  240001 A   0  6978     1   0  60 20 8188   576                 -  0:00 srcmstr
  240001 A   0  7498  6978   0  60 20 21e2   288                 -  0:00 inetd
  240001 A   0  7740     1   0  60 20 11191    72    d1c74       -  0:00 shlap
  240001 A   0  8268  6978   0  60 20 1a1da   684                -  0:00 portmap
  240001 A   0  8514  6978   0  60 20 81e8  1064                -  0:08 snmpd
  240001 A   0  8772  6978   0  60 20 131f3   576                -  0:00 dpid2
  240001 A   0  9030  6978   0  60 20 1a1fa   636                -  0:00 muxatmd
  240001 A   0  9328     1   0  60 20 17177   260 50fadd1c       -  0:03 cron
  200001 A   0  9806  7498   0  60 20 a20a   588                - 21:36 nmbd
  240001 A   0 10066  6978   0  60 20  200   100                -  0:00 biod
  240001 A   0 10328  6978   0  60 20 16216   156        *       -  0:00 nfsd
  240001 A   0 10586  6978   0  60 20 121d2  7028 f0240f88       -  0:00 rpc.moun
```

Figure 20. Output from ps -el command

In the output of the `ps -el` command, look at the `TIME` column (total CPU used), the `C` column (processor utilization used for scheduling), and the `PRI` column (priority). To get real-time execution for a critical application, use the `nice` command. The application can run at a high fixed priority (low number). This is a trade-off decision to make based on workload and priorities.

10.2.1.4 Monitoring CPU usage

Use the `time` command to understand the CPU usage of a particular program:

```
# time my_program
real 0m11.44s
user 0m0.04s
sys 0m0.03s
```

The `time` command shows the amount of real time (user-perceived clock time) and the amount of CPU time (user + sys) that a program consumes.

The difference between the real time that a program takes to finish and the CPU time it consumes can be explained by either other programs running and taking the CPU resource, or by I/O that the program has to wait for before it can continue.

10.2.2 Monitoring memory

When you monitor memory related resources, you need to know some of the following factors.

10.2.2.1 Real size of memory

Before using any other memory measuring tools, it is important to know how much memory you have. To check how much memory you have, use the following command:

```
# lsattr -El sys0 -a realmem
realmem 393216 Amount of usable physical memory in Kbytes False
```

The output displays the amount of usable physical memory in the system. Divide the answer by 1024 to get the answer in MB. The word `False` at the end of the line indicates that the attribute can not be changed by the `chdev` command. Unfortunately, the only way to increase the value is to insert more memory.

10.2.2.2 Paging space

To see the usage of paging space, enter:

```
# lsps -s
```

The output will be similar to the following.

```
Total Paging Space   Percent Used
      128MB              21%
```

Rules for Calculating Paging Space Requirements

There is no right amount of paging space for a system. Paging space requirements are unique for each system, depending on the applications that are running, the number of active users, and other factors.

There are several general rules that can help determine how much paging space is needed:

Rule 1

For systems with less than 64 MB of RAM, the installation process creates paging space equal to two times (2X) the memory.

Page space = 2 * RAM

For systems with 64 MB to 256 MB of RAM, the following paging space equation applies:

Page Space = RAM size + 16 MB

For systems with more than 256 MB of RAM, use the following equation:

Page Space = 512 + (RAM - 256) * 1.25

Note

The above guidelines apply to some computing environments, but may provide too much or too little paging space for others:

Rule 2
Systems with large amounts of memory typically do not need such large amounts of paging space. In a persistent storage environment, in which the machine hosts a few small programs and a large amount of data, the system may need less than one times (1X) its RAM size for paging space. For example, a 1 GB database server that runs on a RS/6000 with 256 MB of RAM and uses only 50 MB of working storage does not need 256 MB of paging space, or even 512 MB of paging space. It needs only the amount of paging space that allows all the working storage to be paged out to disk, because the 1 GB database is mostly persistent storage and requires little or no paging space.

Rule 3
If a disk drive containing an active paging space logical volume is removed from the system, the system will crash.

Considerations when 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 chunks. 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 onto 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.

Determining if more paging space is needed

Allocating more paging space than necessary results in unused paging space that is simply wasted disk space. But if you allocate too little paging space, a variety of unpleasant 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 `lsps -s` output is greater than 80.

Use the following commands to determine if you need to make changes regarding paging space logical volumes:

```
#iostat
#vmstat
#lsps
```

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 `chps` command, for example:

```
# chps -a n paging00
```

Once this has been done, 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.

If you wanted 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 created activated without having to reboot the machine using the `mkps` command.

10.2.2.3 Using the svmon command

The `svmon` command can be used to determine roughly how much memory the system is using. For example:

```
# svmon
               m e m o r y          i n  u s e          p i n       p g  s p a c e
    size    inuse     free      pin   work  pers  clnt   work  pers  clnt     size   inuse
   98304    53434    44870     3840  32701 20505   228   3840     0     0   102400     383
```

Columns under memory

The columns grouped under the `memory` heading have the following meanings:

size This shows total size of memory in 4 K pages.

inuse This shows the number of pages in RAM that are in use by a process plus the number of persistent pages that belonged to a terminated process and are still resident in RAM. This value is the total size of memory minus the number of pages on the free list.

free This shows the number of pages on the free list.

pin This shows the number of pages pinned in RAM (a pinned page is a page that is always resident in RAM and cannot be paged out).

Columns under in use

The columns grouped under the `in use` heading have the following meanings:

work This shows the number of working pages in RAM.

pers This shows the number of persistent pages in RAM.

clnt This shows the number of client pages in RAM (client page is a remote file page).

Columns under pin

The columns grouped under the `pin` heading have the following meanings:

work This shows the number of working pages pinned in RAM.

pers This shows the number of persistent pages pinned in RAM.

clnt This shows the number of client pages pinned in RAM.

Columns under pg space

The columns grouped under the `pg space` heading have the following meanings:

size This is total size of paging space in 4 K pages.

inuse This is total number of allocated slots.

To find out how much memory a process is using, enter:

```
# svmon -P PID
```

Or for information on all processes:

```
# svmon -Pau | more
```

To see the number of working pages unique to this process' private stack and data use in all of virtual memory, look at the work type and description `private`. The `svmon` output may also list several shared segments. For a complete picture, determine which segments are unique to an individual process and which are shared with other programs. Multiply the values by 4096 to get the number of bytes in memory the process is using. The number 4096 comes from the fact that each page is 4 KB in size. You can also divide the number of pages by 256 in order to get megabytes.

10.2.2.4 Using the vmstat command

To see how much of memory used, use `vmstat` again.

For sample output, refer to Figure 19 on page 240.

Look at the columns under memory, page, and faults.

Columns under memory

The information under the memory heading provides information about real and virtual memory:

avm The `avm` column gives the average number of pages allocated to paging space. In AIX, a page contains 4096 bytes of data. This stands for Active Virtual Memory and not Available Memory. The `avm` is the number of 4 K pages that are in use in paging space. The same idea is reflected in the `PERCENT USED` column of the `lsps -s` command**.**

fre The `fre` column shows the average number of free memory frames. A frame is a 4096-byte area of real memory. If the `fre` value is substantially above the `MAXFREE` value (which is defined as `MINFREE` plus 8), then it is unlikely that the system is thrashing (continuously paging in and out). However, if the system is thrashing, be assured

that the fre value is small. Most UNIX and AIX operating systems will use nearly all available memory for disk caching, so you need not be alarmed if the fre value oscillates between MINFREE and MAXFREE.

Columns under page

The information under the page heading includes information about page faults and paging activity:

re The re column shows the number (rate) of pages reclaimed.

pi The pi column details the number (rate) of pages paged in from paging space. One theory is that five page-ins per second should be the upper limit. Use this theoretical maximum as a reference but do not adhere to it rigidly. This field is important as a key indicator of paging space activity. If a page-in occurs, then 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.

po The po column shows the number (rate) of pages paged out to paging space.

fr The fr column details the number (rate) of pages freed. If fr / po is less than h (system wide thrashing parameter modified by schedtune), then you are probably thrashing.

sr The sr column details the number (rate) of pages scanned by the page-placement algorithm. Lots of page in and outs and high CPU wa (wait on I/O) times indicated you could be thrashing. Thrashing means that you might have a lack of memory.

cy The cy column provides the rate of complete scans of the Page Frame Table. The cy shows how many times (per second) the page-replacement code has scanned the Page Frame Table. Since the free list can be replenished without a complete scan of the PFT, and because all of the vmstat fields are reported as integers, this field is usually zero.

Columns under faults

The information under the faults heading in the vmstat output provides information about process control:

in The in column shows the number (rate) of device interrupts. This column shows the number of hardware or device interrupts (per second) observed over the measurement interval. Examples of interrupts are disk request completions and the 10 millisecond clock

interrupt. Since the latter occurs 100 times per second, the `in` field is always greater than 100.

sy The `sy` column details the number (rate) of system calls. 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. Since workloads and applications vary, and different calls perform different functions, it is impossible to say how many system calls per second are too many.

cs The `cs` column shows the number (rate) of context switches.

10.2.2.5 Memory related problem

Processes requesting additional memory are killed once the system runs low on paging space. The system appears hung as new processes and telnet connections are terminated. The error messages like `Not enough memory` or `Fork function failed` are generated. There are three ways to resolve this situation:

1. Add additional paging space.
2. Memory leak?

   Systems often have plenty of paging space (sometimes 3 to 4 times RAM), and can still run out. This could be due to a memory leak. The question then is which process is causing the memory leak.
3. Maximum process?

   The system may be reaching its `Maximum number of PROCESSES allowed per user`, or *maxuproc*. Depending on what maxuproc is set to (default is 40), if a user already has maxuproc number of processes, the system will not allow that user to fork anymore processes.

10.2.3 Monitoring I/O

The I/O subsystem is a very slow resource compared with CPU or memory. For better performance of the I/O subsystem, you need to consider the following factors.

10.2.3.1 Hard disk management consideration

Performance can be improved by putting more than one physical volume in a volume group. The idea is to get the volume group's journaled logical volume on one physical disk, and then place high activity file systems on a separate disk in the volume group. Low activity file systems should be placed on the hard disk containing the journaled logical volume.

High usage logical volumes can be spread across multiple physical disks. This allows parallel reading of the disks and speed I/O access to the logical volume. Large files that are heavily used and not accessed sequentially, such as database files, should be spread across more than one physical volume.

Depending on whether speed or reliability is your priority, there can be several ways to define hard disk. If high reliability is your priority, use mirroring. If sequential access speed is your priority, use striping. Striping allows a sequential read of a file to read from more than one disk at a time, speeding the access. For random access speed, distribute files across multiple disks.

10.2.3.2 Disk I/O pacing

Disk I/O pacing allows the system administrator to balance I/O intensive workloads with other system activity. Pacing will slow down the rate of an I/O intensive task, allowing I/O to be more quickly available to other tasks. This can be controlled by High and Low Water Marks. The default value of I/O pacing is 0. It means no pacing. A process cannot exceed the High Water Mark pending writes to a file. If it tries to do more than this, it is put to sleep until the number of writes is less than or equal to the Low Water Mark.

The values can be changed using with Smit under System Environments -> Change/Show Characteristics of Operating System.

10.2.3.3 Logical volume fragmentation

To look at the placement of logical volumes within a physical volume, use the `lslv` command. For example:

```
# lslv -p hdisk0 hd1
```

The above command will output the numbered partitions of logical volume hd1 that are contained in hdisk0.

The partitions shown are either USED, FREE, or numbered as part of hd1.

If the numbered partitions are scattered around, it may indicate fragmentation has occurred.

To reorganize a logical volume or volume group, use SMIT to run the `reorgvg` command.

```
Smit->Physical & Logical Storage -> Logical Volume Manager-> Volume Groups
-> Set Characteristics of a Volume Group-> Reorganize a volume group.
```

Or fastpath with `smit reorgvg`.

To reorganize a file system, back up the files to another file system, unmount the mount point, remake the file system, remount the mount point, and then restore the files.

Note

Do this very carefully, or you can lose data.

The `defragfs` command can also be used to increase a file systems contiguous free space.

10.2.3.4 File fragmentation

To see if a particular file is fragmented, use the `fileplace` command. For example:

```
# fileplace -pv /home/my_file
```

This command indicates the number of separate sections that make up the file. It also indicates the space efficiency of the file placement.

10.2.3.5 Disk I/O activity

The `iostat` command will provide data on the activity of physical volumes, but not file systems or logical volumes. Remember that the first set of data displayed by `iostat` represents a summary of all activity since system startup. If activity for a disk is greater than 70 percent, then that disk will suffer performance problems.

To see how much disk I/O activity is taking place, enter:

```
# iostat 2
```

This shows the output every 2 seconds.

```
tty:      tin         tout   avg-cpu:  % user    % sys     % idle    % iowait
          0.0          1.0               0.6      6.8       38.3      54.4

Disks:        % tm_act     Kbps      tps    Kb_read   Kb_wrtn
hdisk0          56.5      93.7      23.1      78282  11814223
hdisk1           0.1       1.4       0.1     182068      1892

tty:      tin         tout   avg-cpu:  % user    % sys     % idle    % iowait
          0.0        172.2               1.5     13.7        0.0      84.9

Disks:        % tm_act     Kbps      tps    Kb_read   Kb_wrtn
hdisk0          90.7     150.2      37.6          4       304
hdisk1           0.0       0.0       0.0          0         0

tty:      tin         tout   avg-cpu:  % user    % sys     % idle    % iowait
          0.0        175.6               0.5     10.4        0.0      89.1

Disks:        % tm_act     Kbps      tps    Kb_read   Kb_wrtn
hdisk0          90.0     155.2      38.3          0       312
hdisk1           0.0       0.0       0.0          0         0
```

Figure 21. Output of iostat command

The first set of output is cumulative since boot time.

Look at the `%iowait` numbers to see if a large amount of waiting on I/O is going on. Look at each individual disk listed to see if there is a balance of disk activity between the different disks. The columns `KB_read` and `KB_wrtn` show the kilobytes read and written to the disk. Also look at the `%tm_act` (time active or percentage utilization of the disk) to see if a particular disk is a bottleneck. In that case, some data from the disk should be redistributed. If any single disk averages more than 35 percent busy, it is a sign that it is a bottleneck to the system and should be tuned. An average physical volume utilization greater than 25 percent across all drives indicates need for more physical disks.

We also see the CPU numbers, `%user` (cpu time spent in user mode), `%sys` (percent of CPU time spent in kernel mode), `%idle` (percent of time CPU spent idle with no outstanding local disk requests), and `%iowait` (percent of time CPU spent idle but with outstanding I/O requests), should be less than 20 percent. On multiprocessor systems, CPU statistics are calculated as averages among all processors.

10.2.3.6 Using the filemon command

The `filemon` tool collects and presents trace data on the various layers of file system utilization including the logical file system, virtual memory segments, LVM, and physical disk layers. Data can be collected on all the layers, or

layers can be specified with the -O layer option. The default is to collect data on the VM, LVM, and physical layers. Both summary and detailed reports are generated.

The usage of the filemon command is as follows:

```
# filemon [-i file] [-o file] [-d] [-v] [-u] [-O opt]
   -i file: open input file (default is real-time trace)
   -o file: open output file (default is stdout)
   -d: deferred trace (until `trcon')
   -T num: set trace kernel buf sz (default 32000 bytes)
   -P: pin monitor process in memory
   -v: verbose mode (print extra details)
   -u: print unnamed file activity via pid
   -O opt: other monitor-specific options.
      valid -O options: lf,vm,lv,pv,all
      lf: monitor logical file I/O
      vm: monitor virtual memory I/O
      lv: monitor logical volume I/O
      pv: monitor physical volume I/O
      all: short for lf,vm,lv,pv (default is: vm,lv,pv)
```

Normally, filemon runs in the background while other applications are running and being monitored. When the trcstop command is issued, filemon stops and generates its report. You may want to issue nice -n -20 trcstop to stop filemon since filemon is currently running at priority 40.

Note

filemon will only collect data for those files opened after filemon was started unless you specify the -u flag. Normally, only the top 20 logical files and segments are reported unless the -v (verbose) flag is used.

10.2.3.7 Creating additional log logical volumes

Placing the log logical volume on a physical volume different from your most active file system logical volume will increase parallel resource usage. You are allowed to have a separate log for each file system.

When creating your logical volumes, remember that the performance of drives differs.

Try to create a logical volume for a hot file system on a fast drive.

10.2.3.8 Disk and SCSI performance issues

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. Since there is a limit to the amount of data that can be handled by the SCSI adapter, the SCSI adapter can become a bottleneck.

If all your disk drives are connected to one SCSI adapter, and your hot file systems are on separate physical volumes, you may benefit from using multiple SCSI adapters. The 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 SCSI adapter. For maximum aggregate performance, the total of the transfer rates (Kbps) must be below the SCSI adapter throughput rating. In most cases, use 70 percent SCSI adapter throughput capacity.

10.2.3.9 Asynchronous I/O

An application's processing cannot continue until the synchronous I/O operation complete. In contrast, asynchronous I/O operations run in the background and do not block the user application.This improves performance, because I/O operations and applications processing can run simultaneously.

Many applications, such as databases and file servers, take advantage of the ability to overlap processing and I/O. These asynchronous I/O operations use various kinds of devices and files.

You can change attributes relating to asynchronous I/O using the `chdev` command or SMIT. Likewise, you can use SMIT to configure and remove (unconfigure) asynchronous I/O. Alternatively, you can use the `mkdev` and `rmdev` commands to configure and remove asynchronous I/O. To start SMIT at the main menu for asynchronous I/O, enter `smit aio`.

The default minimum number of servers configured when async I/O is enabled is 1. This is the minservers attribute. There is also a maximum number of async I/O servers that can get created that is controlled by the maxservers attribute. This has a default value of 10. If the number of async I/O requests is high, then the recommendation is to set maxservers to at least 10*(number of disks accessed asynchronously), and minservers should be maxservers /2.

10.3 Performance diagnostic tool

Performance Diagnostic Tool (PDT) attempts to identify performance problems automatically by collecting and integrating a wide range of performance, configuration, and availability data. The data is regularly evaluated to identify and anticipate common AIX and RS/6000 performance problems.

You can install PDT from your AIX installation media via SMIT or the `installp` command. When you list the contents of the install media, the package containing PDT is bos.perf.

Once the package has been installed, PDT must be enabled in order to begin data collection and report writing. PDT is enabled by executing the script /usr/sbin/perf/diag_tool/pdt_config. Only root may execute this script. For example:

```
# /usr/sbin/perf/diag_tool/pdt_config
_________________PDT customization menu_____________________
1) show current  PDT report recipient and severity level
2) modify/enable PDT reporting
3) disable       PDT reporting
4) modify/enable PDT collection
5) disable       PDT collection
6) de-install    PDT
7) exit pdt_config
Please enter a number: 4
```

From the PDT menu, option `4` enables default data collection functions. Actual collection occurs via cron jobs run by the cron daemon.

When PDT is enabled, by default, it adds entries to the crontab file for the adm user**.** The messages are delivered to the adm user by mail.

10.4 Reporting performance problems to IBM

If after examining the system resources, there does not appear to be a resource constraint, and successfully testing the system hardware, then you may need to report the performance problem to IBM.

To report performance related problems, you need to gather some data using the AIX `perfpmr` command. Depending on the version of AIX, this is installed as part of the bos.perf.pmr fileset or the perfagent.tools fileset. The bos.perf.pmr fileset is no longer included on the AIX product media as of AIX Version 4.3.3.

To track a system for one hour, first add /usr/sbin/perf/pmr to your PATH variable. Then enter:

```
# perfpmr 3600
```

The output will go to the directory /var/perf/tmp. If you do not have at least 5 MB of free space, edit /usr/sbin/perf/perfpmr and change the directory name to someplace that has enough space.

When collection is completed, `tar` the contents of the collection directory to a file. If the performance problem has been reported, and a PMR number has been assigned, use the PMR number as part of the filename:

```
# cd /var/perf/tmp
# tar -cfg PMRnum.tar perfdata
# compress PMRnum.tar
```

Then send the collected data to IBM.

The pmr script files are available from the `http://www.ibm.com` Web site. To download these files, do the following:

1. Open your browser, and go to the `http://www.ibm.com` Web site.
2. In the Search window, enter `perfpmr`, and select **Go** or press the **Enter** key.

   The query will return the Search results: AIX Performance PMR Data Collection Scripts - perfpmr.
3. Select **Download** to go to the perfpmr download site.

Download the fileset that relates to the version of AIX you are running. See the README file contained in the package for further instructions.

10.5 Other tools

There are many other tools that can be used to determine what is causing the system performance degradation. Some of them are in the fileset perfagent.tools:

acctcms	Produces command usage summaries from accounting records.
acctcom	Displays selected process accounting record summaries.
accton	Performs process-accounting procedures.
bf, bfrpt	Reports on memory access by applications
fdpr	Optimizes executable modules.
lockstat	Reports on kernel lock contention.

mmtu	Displaying, adding, and deleting Maximum Transfer Unit (MTU) values used for path MTU discovery.
netpmon	With a moderate, network-oriented workload, `netpmon` increases overall CPU utilization by 3 to 5 percent. In a CPU-saturated environment with little I/O of any kind, netpmon slowed a large compile by about 3.5 percent.
netstat	Most of the variations of this command use fewer than 0.2 seconds of CPU time.
nfsstat	Most of the variations of this command use fewer than 0.1 seconds of CPU time.
no	Configures network options.
PDT	Performance Diagnostics Tool. This tools is installed as part of bos.perf.diag_tool and configured to send daily reports about system parameters and performance data to help monitor the system. Configure PDT by running /usr/sbin/perf/diag_tool/pdt_config as root. Select option 4, **modify/enable PDT collection**, then select **7** to exit.
renice	Alters priority of running processes.
sar	Collects, reports, or saves system activity information.
stem	Allows insertion of user-supplied instrumentation code at the entry and exit points of existing program and library subroutines.
stripnm	Displays the symbol information of a specified object file.
syscalls	Records and counts system calls
timex	Reports, in seconds, the elapsed time, user time, and system execution time for a command.
tprof	Since `tprof` uses `trace`, it causes some system overhead. `tprof` only enables one trace hook, however, so its overhead is less than that of a full trace. For example, `tprof` degraded the performance of a large compile by less than 2 percent.
trace	The overhead added by `trace` varies widely, depending on the workload and the number of hook IDs being collected. As an extreme case, a long-running, CPU-intensive job in an otherwise idle system took 3.2 percent longer when `trace` was running with all hooks enabled.
trcnm	Generates a kernel name list.

Chapter 11. Event tracing

This chapter describes the concept of trace and how to use it for gathering and reporting system events.

11.1 Introduction

The trace system is a tool allowing 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 problems
- Monitor system performance

The events that are traced are timestamped as they are written to a binary trace file named /var/adm/ras/trcfile.

There are events pre-defined in AIX and included in selected commands, libraries, kernel extensions, devices drivers, and interrupt handlers. A user can define their 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 in the trace. This value depends on what events you trace and the CPU performance of the system.

11.2 Installing trace

The trace facility and commands are provided as part of the Software Trace Service Aids fileset named bos.sysmgt.trace.

To check if trace has been installed on your system, run the command:

```
# type trace
trace is in /usr/bin/trace
```

If it is not installed, install it from your installation media.

11.3 Taking a trace

Before tracing events, a strategy for what to trace, and when to trace is important.

Follow these steps to gather a useful trace:

1. Select the required hook IDs for tracing.
2. Enable trace.
3. Recreate the problem.
4. Disable trace.
5. Generate the trace report.

11.3.1 Hook IDs

The events are traced referenced by hook identifiers.

Each hook ID uniquely refers to a particular activity that can be traced. When tracing you can select the hook IDs of interest and exclude others that are not relevant to your problem. See Figure 22.

```
25 DEVICE DRIVER:ETHERNET - HIGH PERFORMANCE LAN ADAPTER
26 DEVICE DRIVER:TOKEN RING - HIGH PERFROMANCE ADAPTER
27 DEVICE DRIVER:C3270
28 DEVICE DRIVER:FLOPPY DISK
29 DEVICE DRIVER:SCSI
30 DEVICE DRIVER:DISK
31 DEVICE DRIVER:MUTIL-PROTOCAL ADAPTER
32 DEVICE DRIVER:GRAPHICS
33 DEVICE DRIVER:pty
34 DEVICE DRIVER:rs232
```

Figure 22. Sample trace hook IDs

11.3.2 Selecting trace events

The specific hook IDs selected for a trace will generally be indicated by the developer or maintainer of the source code that has a problem. Specific calls recording an event are present.

Refer to the person supporting the specific code with a problem for a list of useful hook IDs to use.

Doing a blanket method of selecting all events is not useful because the trace log files will quickly fill and overwrite previous entries. Additionally, this generates lots of unwanted data that becomes difficult to manage.

11.3.3 Timing the trace

Carefully selecting the time to trace is important. You may only have a short period of time when you can gather a useful trace. Selecting this time wisely will ensure the problem is not missed during the trace session.

If tracing is stopped prematurely, the problem may not show up in the trace. If stopped too late, the problem may have been recorded, but then overwritten as the trace log file fills up.

The recommended method to trace only the data required is to start and stop the trace on the command line.

For example, to trace a user program named `penknife` and look at files it opened the (hook ID is 15b), run the commands:

```
# trace -a -j 15b; ./penknife; trstop
```

By limiting the period of the trace to include only the period of execution of the problem command, you are much more likely to capture the sequences of events leading up to the problem.

Ideally, you may write a shell script to trace your problem. This allows you to easily reproduce the trace session, so you can concentrate on reproducing the problem, not managing the timing of the trace.

```
 #!/bin/ksh
# This is a script to trace the problem
# now we start trace
trace -a -j 15b
# Now we run the command
./penknife
# Now we stop trace
trcstop
exit
```

Figure 23. Sample shell script

11.3.4 Starting a trace

Trace can be started in background mode or interactive mode.

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 have been assigned numbers that are called trace hooks.

The `trace` command acts like a specialized shell in that there are some direct commands that alter the activity of the trace daemon.

For example, `trcon` starts tracing events and `trcoff` stops tracing events. You can also run system commands (such as the command you want to trace) by preceding them with an exclamation mark (`!`). Press **<q>** to exit the specialized shell and stop the trace.

To `trace` the `file_open` event when running `xinit`, issue following commands:

```
trace -j 15b
->trcon
->!/usr/bin/X11/xinit
->trcstop
->quit
```

There is normally no need to run `trace` in the interactive mode, since the entire system is being traced, not just the commands started from that shell.

To perform a `trace` in the background, use the `-a` parameter. 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 daemon accepts no commands while the trace is in progress. The trace is stopped using the `trcstop` command.

To trace the same command in background mode, issue:

```
trace -a -j 15b
/usr/bin/X11/xinit
trcstop
```

Trace uses in-memory buffers to save the trace data. There are three methods of using the trace buffers:

- Alternate mode

  This is the default mode. All trace events will be recorded in the trace log file.

- Circular mode

  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.

For example:

```
trace -a -l -j 15b
```

- Single mode

  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.

  For example:

  ```
  trace -a -f -j 15b
  ```

11.3.5 Collecting trace data for analysis

To gather the required files for analysis and write to the tape device /dev/rmt0, run the command:

```
snap -Dg -o /dev/rmt0
```

You can ignore any messages regarding the dump device if your trace data does not relate to a system dump.

Send the tape to your support organization.

11.3.6 Tracing fatal problems

If the problem you are tracing crashes the system, increase the size of the trace buffer and log files. The trace buffer remains in kernel memory and can be extracted from the kernel dump.

This is the general outline:

1. Check the size of the dump device.

   The dump device will require sufficient space for the extra trace buffer that will be dumped. For example, if you specify a 6 MB trace buffer, the dump device must cater for an extra 6 MB.

2. Start trace with large buffers.

   For example, to have a 6 MB trace buffer in kernel memory and a 12 MB log file to trace some networking hook IDs, run:

   ```
   trace -a -j 251,252,253,254 -L 12000000 -T 6000000
   ```

   Ensure enough space is available for a 12 MB /var/adm/ras/trcfile.

3. Recreate your problem that crashes the system.

4. Extract the trace buffer from the kernel dump.

For example, with the kernel dump file in the current directory, run the following command to extract the trace buffer into a file trcfile.dump:

```
# trcdead -o trcfile.dump dump
```

You can run a report against this extracted file, for example:

```
# trcrpt trcfile.dump
```

11.4 Generating a trace report

The raw /var/adm/ras/trcfile trace file containing system events needs to be translated into a usable format for viewing. The `trcrpt` command formats the output.

For example, to output a formatted trace report to the /tmp/trcrpt.out file, run the following command:

```
# trcrpt >/tmp/trcrpt.out
```

Or

```
# trcrpt -o /tmp/trcrpt.out
```

Alternatively, execute the following command to display the SMIT panel for trace reporting:

```
# smitty trcrpt
```

If you want to see process names in the trace report, you should ensure that trace hooks 106, 107, 10C, 134, 139, and 465 are enabled when gathering the trace.

11.4.1 Filtering the trace report

Often the output of the trace report file is very large.

Despite selecting a narrow time period to do your trace, the system may be tracing other unrelated events from the execution of other threads or interrupt handlers.

Reducing the size of the output to the events of interest is useful for problem management. The output can be reduced by specifying the process ID of the application having the problems.

For example, to create a trace report for the process ID 8002, run the command:

```
# trcrpt -O'svc=y timestamp=3' -p 8002 > /tmp/trcrpt.out
```

11.4.2 Examining a client trace file

To examine the /var/adm/ras/trcfile trace file on another machine, like dump analysis, you must also include the kernel namelist file, /unix, to interpret address of kernel symbols correctly.

For example, copy the following files from the client machine to a temporary directory on the analysis machine:

```
/var/adm/ras/trcfile
/unix
```

If you are reading a tape written with the `snap -Dg` command, restore it from tape device /dev/rmt0 with the command:

```
# tar -xvf /dev/rmt0
```

Starting with the version of the `trace` command supplied in the fileset bos.sysmgt.trace.4.3.2.2 supplied when ordering APAR IX85220, the `-n` option generates symbol name data as part of the trace file, so it is no longer required to have the /unix file from the traced system.

11.4.3 Analysis of trace report

The analysis of the trace data is done by the person viewing the source code to the problem that was traced. The special points in the source code that create a trace entry can be examined along with any arguments that have been saved.

Outside of the source code, the trace facility is useful for gauging performance characteristics of a program by examine the timestamps of system calls. It is also useful for examining the precise sequential flow of system activity in a multithreaded and possibly multiprocessor environment.

Chapter 12. Printing problems

Printers are perhaps the most complex mechanical component of modern computer systems. Normally, they demand more time and attention than terminals and processors. In this chapter, we refer only to the problem determination of printing components related to AIX.

12.1 Local printing

There are two main classes of printers available on AIX. Local printers are physically connected to the RS/6000 machine, normally to the parallel port or serial port. Remote printers are not connected to the local RS/6000 but are either connected to another machine on the network, or connected directly to the network itself.

12.1.1 Adding a local printer

You may use the command line or SMIT to set up a local printer. Since there are so many parameters required, SMIT is recommended:

```
# smit makprt
```

Choose the printer/plotter type of the device you are adding. When you first install AIX 4, it does not automatically install the required files for all supported printer devices. If you are going to install different printers, check if these printers device drivers are already installed, and if they are not installed, install them from the AIX 4.3 installation CD.

If you can not find the device driver for your printer on the installation CD, you should ask the printer manufacturer if they have the device driver running on AIX; otherwise you should choose an existing device driver that serves you best.

```
                                Add a Printer/Plotter

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

[TOP]                                                  [Entry Fields]
  Printer/Plotter type                                 osp
  Printer/Plotter interface                            rs232
  Description                                          Other serial printer
  Parent adapter                                       sa0
* PORT number                                          [0]                    +
  BAUD rate                                            [9600]                 +
  PARITY                                               [none]                 +
  BITS per character                                   [8]                    +
  Number of STOP BITS                                  [1]                    +
  FLOW CONTROL to be used                              [dtr]                  +
  OPEN DISCIPLINE to be used                           [dtropen]              +
[MORE...21]

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

Figure 24. SMIT add a local printer menu

The actual menu shown may be different for individual printers. See Figure 24. Pay special attention to parameters of serial printer, including baud rate, parity, bits per character, number of stop bits, flow control mode, and so on. These parameters should be compatible with the parameters on the printer.

12.1.2 Troubleshooting communication

Testing communications ensures that the data can reach the printer and that the printer will print the data.

Assume we have a configured device driver designated by `lp0`, which points to a printer attached to the parallel port, and the configured device driver designated by `lp1` points to printer attached to the second serial port.

To check the system for configured local printers, issue following command:

```
# lsdev -Cc printer
lp0 Available 00-00-0P-00 Hewlett-Packard laserJet 4,4M
lp1 Available 00-00-S2-00 IBM 2380 Personal printer II
```

With the exception of certain specialized printers, working printers are capable of printing ASCII text. Issue the following command to test ASCII printing:

```
# lptest 40 5 > /dev/lp0
```

This will send an ASCII test pattern 40 characters wide and 20 lines long to the file /dev/lp0 and then to the device driver in the kernel and out the port for which lp0 was configured (in this case, the parallel port). The test pattern will look like the following:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGH
"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJK
%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKL
```

A printer that accepts data in the postscript format can be tested by writing a postscript file, such as the postscript header page to the device file. Issue the following command to test postscript printing:

```
# cd /usr/lib/lpd/pio/burst
# cat H.ps > /dev/lp0
```

Failures of the preceding tests (`lptest` or `cat`) will be one of three kinds:

1. A prompt is returned, but there is no output at the printer. This can be caused by one of the following:
   - Wrong data type

     The system has sent the data out of the port. In most cases where there is a printer buffer (or online) light, this light will flash indicating receipt of data. The printer most likely was not set to receive data in the format sent. For example, ASCII text was sent to a printer set to receive postscript data. While the data reached the printer buffer, the buffer was flushed. Check the printer mode and set it to match the data type sent, or send the appropriate data type.
   - Wrong cabling

     The system has sent the data out the port, but the printer has not accepted it. Some printers do not accept data on their RxD line unless the printer DSR signal is high. Wire the cable so that DSR goes high when the system DTR signal goes high (DTR goes high when the system sends data).
2. A prompt is returned, but one of the following error messages is returned: `cannot open the specified file` or `cannot create the specified file`. This can be caused by one of the following:
   - The system is unable to send data out of the port due to wrong cabling or a wrong printer setting.
   - The system is not getting the CTS signal. CTS indicates that the printer is present and switched on. This signal is normally sent along the

printer RTS line. CTS is expected to be high at all times. In some cases, the printer will have a setting for this (possibly listed as an RTS option which should be set to `true` or `high`).

3. The prompt does not return.

   Try the **Ctrl-C** key combination. If the prompt returns with a message, see the previous step. If the prompt returns without a message, check the use of the lp device with the `fuser` command. For example:

   ```
   # fuser /dev/lp1
   ```

   The following should be returned:

   ```
   /dev/lp1:
   ```

   If there are any numbers after the colon, processes associated with the lp have been identified. To find more information about the process(s), issue the following command:

   ```
   ps -ef|grep PID
   ```

   Where `PID` is the process ID shown by the `fuser` command.

   If this shows the process to be `pioout` or `piobe`, and there is no printed output from the printer, there is probably a flow control problem. If Xon/Xoff flow control is used, then the system is waiting for an Xon character from the printer. If DTR (hardware) flow control is in use, then the system is waiting for the CD signal from the printer DTR line. If proper flow control can be established, printed output will resume where it left off. Try the following:

   a. Switch printer power off and back on.

   b. Switch printer offline and back online.

   c. Disconnect the cable from the printer, plug it into a terminal, and type **Ctrl-Q** on the terminal.

   If method 3 works, then the printer is not sending an Xon character to the system. Investigate the use of Xon/Xoff flow control by the printer.

   If the flow control condition cannot be rectified or flow control is not the issue, kill the processes by issuing the following command:

   ```
   # fuser -k /dev/lp1
   ```

   Verify that the processes were killed by running the `fuser` command again:

   ```
   # fuser /dev/lp1
   ```

   If unsuccessful, detach the printer cable from the system.

   If still unsuccessful, try terminating the processes with the `kill` command:

```
# kill -9 PID
```

If this is unsuccessful, a system reboot is needed. If upon reboot the processes reappear, then you should deactivate the print queue:

```
# disable /dev/lp1q
```

Where `lp1q` is the printer queue name for printer lp1.

12.1.3 Troubleshooting the queue

If a queue is used for printing, and there is printed output from using the `cat` or `lptest` tests detailed in Section 12.1.2, “Troubleshooting communication” on page 268, then there is a problem with the queue.

Let us assume, lp0q and lp1q are the queues that point to lp0 and lp1 respectively. To view the status of the queues you may use, issue the following command:

```
# lpstat
```

Which in the case of the examples might return:

```
Queue Dev Status
----- ---- ----
lp0q lp0 READY
lp1q lp1 DOWN
```

Other possible states are DEV_BUSY or DEV_WAIT.

To enable the queue lp1q, you can try:

```
# enable qlp1
```

If the `lpstat` command subsequently shows the queue to be DOWN and nothing was printed, check your file permissions and file owners and groups. Refer to Section 12.1.5, “Printer file permissions” on page 274 for more information.

If repeated invocations of the `lpstat` command show the queue hanging for several minutes in a DEV_BUSY or DEV_WAIT state, check the `file=` field for that queue in the `qconfig` file by entering following command:

```
# pg /etc/qconfig
```

In the example queue and device, this would show:

```
lp1: file = /dev/lp1
```

If the `file=` field does not have `/dev/lp#`, contact your AIX support function. Another method to get the device name is to issue:

```
# lsque -qlp1q
```

In the example, the following would be returned:

```
# lp1q: device = lp1
```

With the device name, you could run:

```
# lsquedev -qlp1q -dlp1
```

Which would return information including the `file=` field.

If the queue is stuck in a READY state with jobs QUEUED, the qdaemon must be cycled. This is the case where queues have been added, removed, or changed while the queuing system was in use.

To cycle the `qdaemon`, use following commands:

```
# stopsrc -cs qdaemon
# startsrc -s qdaemon
```

12.1.4 Temporary spool files

AIX will write temporary files to the /var file system. Normally, AIX will erase temporary spool files after them being printed, but if the printing system encounters a problem, and it fails to print them, there maybe a large amount of rubbish files left behind. Thus, you have to erase them manually.

The directories used in the AIX V4 spooling system are shown in Table 9.

Table 9. Temporary files used in spooling system

Directory name	Files in the directory
/var/spool/lpd/qdir	Print job description files
/var/spool/qdaemon	Print job copies
/var/spool/stat	Printer queue description files
/tmp	System temporary files

The directory /var/spool/qdaemon is only used when we want the files to be printed to have a local copy first before they are send to the print queue. The `lpr` command will do the copy by default while `lp` and `enq` will not.

For example, to save a copy of /etc/passwd in /var/spool/qdaemon and print it on the default queue:

```
# lp -c /etc/passwd
```

The following steps should be used to manually clear the temporary files:

1. Make sure you are logged in as root.
2. Let all current jobs finish printing, or cancel them. You may use the `enq` command to cancel each job:

   ```
   # enq -x your_job_name
   ```

3. Issue the following command to stop the qdaemon:

   ```
   # stopsrc -s qdaemon
   ```

4. Issue the following commands to verify that the qdaemon did not fork other processes:

   ```
   # ps -ef|grep qdaemon
   # ps -ef|grep pio
   ```

   If you get more than one line back from the above `grep` commands, issue the following command to kill the process:

   ```
   # kill -9 pid
   ```

5. Perform this step only if it is necessary to save the current print jobs from being deleted. Otherwise, proceed to step 6.

   Make a copy of the files in the following directories. You can print it when the queuing system is running again.

   ```
   # /var/spool/qdaemon
   # /var/spool/lpd
   ```

6. Use the `df` command to check the free space in /var file system. If the /var file system gets too full, you may have a problem with the qdaemon or the spooler. In this case, a system reboot may not clear out the files or restart the qdaemon. Thus, you may either expand the /var file system or erase files unnecessary under /var file system. Candidate files for freeing space in /var are shown in Table 10.

Table 10. Files in /var that can be removed

File name	Description	Command to run
/var/asm/ras/trcfile	Trace file	`rm /var/adm/ras/trcfile`
/var/adm/ras/errlog	System errorlog	`errclear 0`
/var/adm/wtmp	User login history file	`cp /dev/null /var/adm/wtmp`
/var/adm/sulog	User su history file	`rm /var/adm/sulog`

7. Remove files in the spool directory:

   ```
   # cd /var/spool/lpd/qdir
   # rm *
   ```

```
# cd /var/spool/lpd/stat
# rm *
# cd /var/spool/qdaemon
# rm *
# cd /var/spool/lpd
# rm *
```

8. Restart the qdaemon:

   ```
   # startsrc -s qdaemon
   ```

   Enable printer queues if they do not automatically return to the READY state:

   ```
   # enable queue_name
   ```

9. Copy back any files saved in step 5.

12.1.5 Printer file permissions

To ensure a user can have the privilege to print, the owner, group, and permissions of the files shown in Table 11 should be checked.

Table 11. Print files permission

Permission	Owner	Group	File name
-r-sr-s---	root	printq	/usr/sbin/qdaemon
-r-sr-s---	root	printq	/usr/sbin/lpd
-r-sr-sr-x	root	printq	/bin/enq
-rw-rw-r--	root	printq	/etc/qconfig
drwxr-xr-x	bin	bin	/var
drwxrwxr-x	bin	bin	/var/spool
drwxrwxr-x	bin	printq	/var/spool/lpd
drwxrwxr-x	bin	printq	/var/spool/qdaemon
drwxrwxr-x	root	prinq	/var/spool/lpd/qdir
drwxrwxr-x	root	printq	/var/spool/lpd/stat

12.2 Printer in network environment

Normally, there is no need for every machine in a network environment to have a dedicated printer, so they are very often shared as a network device available to many machines.

Therefore, a correctly functioning network plays an important role in printer setup and maintenance.

12.2.1 Manipulate lpd

The lpd daemon is a member of the TCP/IP group, and must be running to support remote print requests.

Issue the following commands to manipulate the lpd daemon:

1. To start lpd in debug mode:

   ```
   # startsrc -s lpd -a "-d"
   ```

   The `-d` flag sends debugging information to the syslogd daemon.

2. To stop lpd:

   ```
   # stopsrc -s lpd
   ```

3. To refresh lpd:

   ```
   refresh -s lpd
   ```

You may use SMIT to start the lpd with the `both` option. This will start lpd automatically when the system restarts.

12.2.2 Printing privilege

The lpd daemon is the server-end program used in AIX to handle remote printer requests. It is a UNIX socket program that monitors a TCP/IP port for remote print requests. Each request is placed upon the specified queue or, if no queue is specified, the default queue. The files to be printed are also placed in the /var/spool/lpd directory.

Since any TCP/IP system can place a request on any host for printing services, as a security feature, the lpd daemon accepts printer requests only from remote hosts that are listed in the local /etc/hosts.equiv or /etc/hosts.lpd file.

/etc/hosts.equiv will grant remote user privilege with both remote printing and remote shell, so if you just want a client system to able to print on the server printer, edit /etc/hosts.lpd instead of /etc/hosts.equiv.

12.2.3 Remote printing

Let us assume we have one printer server named *psrv* with a local printer queue *s_queue* and a printer client named *pclt* with a remote queue *c_queue*. Both machines are on the same subnet.

1. Check printer queue status.

   On pclt issue the following command to print, then check printer queue status:

   ```
   # lp -d /dev/c_queueu /etc/hosts
   # lpstat
   ```

   Which in the case of the example may return:

   ```
   Queue   Dev   Status       Job  Files      User
   ------- ----- ---------    ---  -----      ----
   c_queue @psrv READY
   c_queue s_que READY
                 QUEUED        23   /etc/hosts root@pclt
   ```

   The name of the remote queue is `c_queue` on the client machine `pclt`. All the print requests sent to this queue will be redirected to `s_queue` on the printer server `psrv` through the rembak backend program.

   On the server machine `psrv`, you may issue the following command to check the queue status:

   ```
   # lpstat
   ```

   Which in the case of the example may return:

   ```
   Queue   Dev   Status       Job  Files      User
   ------- ----- ---------    ---  -----      ----
   s_queue lp0   READY
                 QUEUED        23   /etc/hosts root@pclt
   ```

2. If the `lpstat` command running on the client machine hangs or times out, check the following issues:

   a. Check the network connectivity between the client and server:

      ```
      # ping p_srv
      ```

      If the command hangs, you may have trouble with your network. Contact your network administrator for help, or refer to Chapter 7, “TCP/IP networking problems” on page 155.

   b. Check that the lpd daemon is running on the server machine. On `psrv`, issue the command:

      ```
      # lssrc -s lpd
      ```

      Refer to Section 12.2.1, “Manipulate lpd” on page 275 for details about lpd.

   c. Check if the server grants the privilege for the client to print. Refer to Section 12.2.2, “Printing privilege” on page 275 for details about printing privilege.

3. If lpstat running on both client and server seems similar as our example, except the queue status is DOWN, DEV_WAIT, or DEV_BUSY, it may be caused by the print queue on the server machine. On the server, issue the following command to enable the queue:

   ```
   # enable s_queue
   ```

   If the above command can not resolve the problem, refer to Section 12.1.2, “Troubleshooting communication” on page 268 and Section 12.1.3, “Troubleshooting the queue” on page 271 for local printing problem determination.

4. If you can find the job on the client machine, but are unable to find it on the server machine:

   a. Refresh the qdaemon on both server and client. Remove all print temporary files that reside in /var/spool directory and have not been cleaned by AIX. Refer to section 12.1.4, “Temporary spool files” on page 272 for details.

   b. Remove and recreate the remote print queue.

Appendix A. Special notices

This publication is intended to help system administrators and IBM service personnel to perform basic problem determination on RS/6000 machines running AIX V4.3. The information in this publication is not intended as the specification of any programming interfaces that are provided by AIX Version 4.3.2. See the PUBLICATIONS section of the IBM Programming Announcement for AIX Version 4.3.2, Program Number 5765-C34 for more information about what publications are considered to be product documentation.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

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

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

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

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The information about non-IBM ("vendor") products in this manual has been supplied by the vendor and IBM assumes no responsibility for its accuracy or completeness. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate

them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

Any performance data contained in this document was determined in a controlled environment, and therefore, the results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.

Reference to PTF numbers that have not been released through the normal distribution process does not imply general availability. The purpose of including these reference numbers is to alert IBM customers to specific information relative to the implementation of the PTF when it becomes available to each customer according to the normal IBM PTF distribution process.

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

AIX
AIXwindows
AS/400
CT
ESCON
IBM ®
Micro Channel
Netfinity
POWER Gt1
RETAIN
RS/6000
SP
System/390

The following terms are trademarks of other companies:

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

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

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

PC Direct is a trademark of Ziff Communications Company in the United

States and/or other countries and is used by IBM Corporation under license.

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

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

SET 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.

Appendix B. Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

B.1 International Technical Support Organization publications

For information on ordering these ITSO publications see “How to get ITSO redbooks” on page 287.

- *AIXLink/X.25 LPP Cookbook*, SG24-4475
- *AIX Version 4.3 Differences Guide*, SG24-2014
- *RS/6000 SP: Problem Determination Guide*, SG24-4778

B.2 Redbooks on CD-ROMs

Redbooks are also available on the following CD-ROMs. Click the CD-ROMs button at http://www.redbooks.ibm.com/ for information about all the CD-ROMs offered, updates and formats.

CD-ROM Title	Collection Kit Number
System/390 Redbooks Collection	SK2T-2177
Networking and Systems Management Redbooks Collection	SK2T-6022
Transaction Processing and Data Management Redbooks Collection	SK2T-8038
Lotus Redbooks Collection	SK2T-8039
Tivoli Redbooks Collection	SK2T-8044
AS/400 Redbooks Collection	SK2T-2849
Netfinity Hardware and Software Redbooks Collection	SK2T-8046
RS/6000 Redbooks Collection (BkMgr)	SK2T-8040
RS/6000 Redbooks Collection (PDF Format)	SK2T-8043
Application Development Redbooks Collection	SK2T-8037
IBM Enterprise Storage and Systems Management Solutions	SK3T-3694

B.3 Other publications

These publications are also relevant as further information sources:

- *7006 Graphics Workstation Service Guide*, SA23-2719
- *7009 Compact Server Service Guide*, SA23-2716
- *7012 300 Series Service Guide*, SA38-0545

- *7012 G Series Service Guide*, SA23-2741
- *7013 500 Series Installation and Service Guide*, SA38-0531
- *7013 J Series Service Guide*, SA23-2725
- *7014 Model S00 Rack*, SA38-0550
- *7015 Model R00 Rack Installation and Service Guide*, SA23-2744
- *7015 Model R10/R20/R21 CPU Drawer Service Guide*, SA23-2708
- *7015 R30, R40, and R50 CPU Enclosure Installation and Service Guide*, SA23-2743
- *7017 S Series Installation and Service Guide*, SA38-0548
- *7024 E Series Service Guide*, SA38-0502.
- *7025 F30 Series Service Guide*, SA38-0505
- *7025 F40 Series Service Guide*, SA38-0515
- *7025 F50 CPU Drawer*, SA38-0541
- *7026 CPU Drawer Installation and Service Guide*, SA38-0535
- *7043 Service Guide*, SA38-0512
- *7043 Model 260 Service Guide*, SA38-0554
- *Enterprise Server H Series CPU Drawer Installation and Service Guide*, SA38-0547
- *PCI Adapter Placement Reference*, SA38-0538
- *RS/6000 Adapters, Devices, and Cable Information for Multiple Bus Systems*, SA38-0516
- *RS/6000 Adapters, Devices, and Cable Information for Micro Channel Bus Systems*, SA38-0533
- *RS/6000 Diagnostics Information for Multiple Bus Systems*, SA38-0509
- *RS/6000 Diagnostics Information for Micro Channel Bus Systems*, SA38-0532
- *Site & Hardware Planning Information*, SA38-0508
- *SSA Adapters: User's Guide and Maintenance Information*, SA33-3272

B.4 Internet sites

The following are valuable resources located on the Internet:

- `fpt://index.storsys.ibm.com/devdrvr/AIX/`

- http://rshelp.austin.ibm.com/cgi-bin/dsense/dsense_form.sh
- http://rshelp.austin.ibm.com/hardware/tools.html
- http://service.boulder.ibm.com/support/rs6000.support/downloads
- http://service.software.ibm.com/support/rs6000/
- http://www.hursley.ibm.com/~ssa/
- http://www.ibm.com
- http://www.rs6000.ibm.com/library
- http://www.rs6000.ibm.com/resource/aix_resource/Pubs/
- http://www.rs6000.ibm.com/resource/hardware_docs/
- http://www.rs6000.ibm.com/support/micro

How to get ITSO redbooks

This section explains how both customers and IBM employees can find out about ITSO redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** http://www.redbooks.ibm.com/

  Search for, view, download, or order hardcopy/CD-ROM redbooks from the redbooks Web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this redbooks site.

  Redpieces are redbooks in progress; not all redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

  Send orders by e-mail including information from the redbooks fax order form to:

	e-mail address
In United States	usib6fpl@ibmmail.com
Outside North America	Contact information is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

This information was current at the time of publication, but is continually subject to change. The latest information may be found at the redbooks Web site.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and redbooks by accessing the IBM Intranet Web site at http://w3.itso.ibm.com/ and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may access MyNews at http://w3.ibm.com/ for redbook, residency, and workshop announcements.

IBM redbook fax order form

Please send me the following:

Title	Order Number	Quantity

First name	Last name	
Company		
Address		
City	Postal code	Country
Telephone number	Telefax number	VAT number

☐ Invoice to customer number _____

☐ Credit card number _____

Credit card expiration date	Card issued to	Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

List of abbreviations

ARP	Address Resolution Protocol
ASCII	American National Standards Code for Information Interchange
ATE	Asynchronous Terminal Emulator
ATM	Asynchronous Transfer Mode
ATMLE	ATM LAN Emulation
BIND	Berkeley Internet Name Daemon
BIST	Built-In Self-Test
BUMP	Bring-Up Multiprocessor
CD	Compact Disc
CD-ROM	Compact Disc Read Only Memory
CDE	Common Desktop Environment
CHRP	Common Hardware Reference Platform
CRC	Cyclic Redundancy Check
DAR	Data Address Register
DAR2	Secondary Data Address Register
DBM	Database Manager
DHCP	Dynamic Host Configuration Protocol
DIMM	Dual Inline Memory Module
DLT	Digital Linear Tape
DNS	Domain Name Server
DSI	Data Storage Interrupt
DSIRR	Data Storage Interrupt Reason Register
DSISR	Data Storage Interrupt Status Register
ELAN	Emulated LAN
ESD	Electro-static Discharge
FRU	Field Replacable Unit
IAR	Instruction Address Register
IBM	International Business Machines Corporation
ICMP	Internet Control Message Protocol
IGMP	Internet Group Management Protocol
IP	Internet Protocol
IPL	Initial Program Load
ISA	Industry Standard Architecture
ITSO	International Technical Support Organization
JFS	Journaled File System
JTAG	Joint Test Action Group
LAN	Local Area Network
LCD	Liquid Crystal Display
LECS	LAN Emulation Configuration Server
LED	Light Emitting Diode
LLDB	Low Level Debugger
LP	Logical Partition
LR	Link Register
LES	LAN Emulation Server
LUM	Licence Use Management

LVCB Logical Volume Control Block

LVID Logical Volume Identifier

LVM Logical Volume Manager

MAC Medium Access Control

MAP Maintenance Analysis Procedure

MCA Micro Channel Architecture

MSR Machine Status Register

MSS Maximum Segment Size

MST Machine State Save Area

MTU Maximum Transfer Unit

NFS Network File System

NIS Network Information Service

NLS National Language System

NUA Network User Address

NVRAM Non Volatile Random Access Memory

OCS On-board Chip Sequencer

ODM Object Data Manager

OEM Original Equipment Manufacturer

PAP Password Authentication Protocol

PCI Peripheral Component Interface

PMR Problem Management Record

POST Power-On Self-Test

PP Physical Partition

PPP Point-to-Point Protocol

PTF Program Temporary Fix

PVC Permanent Virtual Circuit

PVID Physical Volume Identifier

QIC Quarter Inch Cartridge

RAID Redundant Array of Independent Disks

RIO Remote Input Output

RIP Routing Information Protocol

ROS Read Only Storage

RPC Remote Procedure Call

SCSI Small Computer System Interface

SIB System Interface Board

SIMM Single Inline Memory Module

SLIP Serial Link Internet Protocol

SMIT System Management Interface Tool

SMS System Management Services

SNA Systems Network Architecture

SPCN Serial Power Control Network

SRN Service Request Number

SRV Segment Register Value

SSA Serial Storage Architecture

SVC Switched Virtual Circuit

TCP Transmission Control Protocol

UDP	UNIX Datagram Protocol
UUCP	UNIX-to-UNIX Communication Protocol
VCI	Virtual Channel Indicator
VGDA	Volume Group Descriptor Area
VGID	Volume Group Identifier
VGSA	Volume Group Save Area
VMM	Virtual Memory Manager
VPD	Vital Product Data
VPI	Virtual Path Indicator
WAN	Wide Area Network
XDM	X Display Manager

Index

Symbols

Numerics

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

ITSO redbook evaluation

Problem Solving and Troubleshooting in AIX Version 4.3
SG24-5496-00

Your feedback is very important to help us maintain the quality of ITSO redbooks. **Please complete this questionnaire and return it using one of the following methods:**

- Use the online evaluation form found at http://www.redbooks.ibm.com/
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Which of the following best describes you?
_ Customer _ Business Partner _ Solution Developer _ IBM employee
_ None of the above

Please rate your overall satisfaction with this book using the scale:
(1 = very good, 2 = good, 3 = average, 4 = poor, 5 = very poor)

Overall Satisfaction _____

Please answer the following questions:

Was this redbook published in time for your needs? Yes___ No___

If no, please explain:

_____

_____

_____

_____

What other redbooks would you like to see published?

_____

_____

_____

Comments/Suggestions: (THANK YOU FOR YOUR FEEDBACK!)

_____

_____

_____

_____

SG24-5496-00
Printed in the U.S.A.

Problem Solving and Troubleshooting in AIX Version 4.3

SG24-5496-00