RAID System Administration Guide

Document Number 007-2113-002

Contributors

Written by Susan Ellis
Illustrated by Dan Young
Edited by Nancy Schweiger
Production by Gloria Ackley
Engineering contributions by Curtis Anderson, Jim Oliver

RAID System Administration Guide
Document Number 007-2113-002

Silicon Graphics, Inc.
Mountain View, California

Contents

Figures

Tables

About This Guide

The *RAID System Administration Guide* describes configuration, maintenance, and error recovery procedures for Silicon Graphics'® RAID (Redundant Array of Inexpensive Disks) product. The IRIX™ 5.0.1 version of the RAID device driver and the RAID administrative utility, *raid*(1M), are documented.

How to Use This Guide

Read "About This Guide" to learn what hardware and software is documented in this guide, what the prerequisites are for using RAID, about the documentation conventions in this guide, about product support for Silicon Graphics products, and where to go for more information about RAID.

Read Chapter 1, "Introduction to RAID," to learn RAID basics: what it is, the similarities and differences between RAID and non-RAID disks, the configuration options, the RAID LEDs, and RAID failure prediction and recovery features.

Turn to Chapter 2, "Formatting a RAID," to learn how to format a new RAID and reformat an existing RAID.

Chapter 3, "Routine Maintenance Tasks," covers a variety of tasks that system administrators occasionally may need to perform during normal operation of a RAID.

Chapter 4, "Recovering after a Disk Warning or Failure," describes how to replace a disk drive in a RAID unit after a predictive warning or a failure. It includes instructions for rebuilding a new disk drive with the same data as the replaced disk drive.

Chapter 5, “Error Messages,” contains lists of RAID error messages and suggested recovery procedures.

Appendix A, “Programming Hints,” contains tips for application developers on getting maximum I/O performance from a RAID.

Many readers will not need this entire guide. The chapters you need depend upon your role:

- System Administrator

  System administrators are the primary audience of this guide. The entire guide except for Appendix A explains configuration, routine maintenance, and error recovery procedures that are likely to be performed by system administrators.

- System Support Engineer

  System Support Engineers can find the procedure for formatting a RAID in Chapter 2.

- Application Developer

  Application developers should read Chapter 1 to learn about the RAID configuration options that affect the performance of their applications and Appendix A for information about optimizing I/O in applications that read and write data on RAIDs.

This guide assumes that you know how to become superuser, get disk controller and unit numbers from *hinv*(1M), and how to make a file system on a disk with *mkfs*(1M).

Hardware and Software Requirements

RAID is supported on workstations and servers that have fast and wide SCSI buses. This includes CHALLENGE™ L, CHALLENGE XL, POWER CHALLENGE L, POWER CHALLENGE XL, Onyx™ Deskside, and Onyx Rack.

This guide applies only to RAID units manufactured by Silicon Graphics. The commands and procedures documented in this guide cannot be used when disk drives, RAID controllers, or RAID unit enclosures purchased from vendors other than Silicon Graphics are used, even when those parts appear to be identical to the ones used in RAID units purchased from Silicon Graphics.

Systems with one or more RAIDs must have a non-RAID system disk.

RAID requires IRIX Release 5.0.1 or later. The subsystem *eoe1.sw.unix* contains all of the files needed for RAID.

Documentation Conventions

To distinguish between physical disks and what appears to IRIX to be a disk, this guide uses the following terms:

disk	one or more physical disk drives that appear to IRIX as a single disk
disk drive	a single disk hardware unit
RAID	the array of five disk drives that to IRIX appear as a single disk of type RAID
RAID unit	the RAID hardware including five disk drives, a controller, a backplane, and an enclosure

Three terms are used to describe the states of RAIDs and / or their disk drives:

operational	The RAID unit is operating normally.
failed	The RAID unit or a disk drive is not responding to commands, or it is responding but is not operating properly.
down	The RAID unit or a disk drive has failed or is not operating because it has been manually shut down with the *raid* **–d** command.

This guide uses these font conventions:

italics Italics are used for command and manual page names, file names, variables, and the names of *inst*(1M) products and subsystems.

`typewriter` Typewriter font is used for examples of command output that is displayed in windows on your monitor.

`bold typewriter`
Bold typewriter is used for commands and text that you are to type literally.

Product Support

Silicon Graphics provides a comprehensive product support and maintenance program for its products. For further information in the United States and Canada, contact the Technical Assistance Center at 1-800-800-4SGI. Elsewhere, contact your local service provider.

For More Information

For more information about RAID and disk management on IRIX, see the following sources:

- The University of California at Berkeley report on RAID:

  D. Patterson, G. Garth, R. Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)," University of California, Berkeley, Report No. UCB/CSD/87/391, December, 1987.

- The *IRIX Advanced Site and Server Administration Guide*

- IRIX manual pages on the *raid* administrative program and the RAID device driver:

  raid(1M), *usraid*(7)

- IRIX manual pages on various disk information and management tools:

  dvhtool(1M), *fx*(1M), *hinv*(1M), *mkfs*(1M), *mklv*(1M), *sar*(1)

Chapter 1

Introduction to RAID

Silicon Graphics' RAID (Redundant Array of Inexpensive Disks) product is a large capacity disk that provides protection against media failure. RAID stores parity information across a group of disk drives so that if a single disk drive in the group fails, users' data can be recovered.

This chapter briefly explains the characteristics of RAID and the features of the Silicon Graphics implementation of RAID. The sections in this chapter are:

- Section 1.1, "RAID Levels"
- Section 1.2, "RAID Stripe Depth"
- Section 1.3, "RAID and IRIX"
- Section 1.4, "RAID Configuration Options"
- Section 1.5, "RAID LEDs"
- Section 1.6, "RAID Failure Prediction"
- Section 1.7, "RAID Failure Recovery"

1.1 RAID Levels

RAID was first defined by Patterson, Garth, and Katz of the University of California, Berkeley, in their 1987 paper, "A Case for Redundant Arrays of Inexpensive Disks (RAID)" (see "For More Information" in "About This Guide"). Silicon Graphics' RAID product implements these RAID levels:

- RAID 3, striping with a single parity disk per group
- RAID 5, striping with parity spread over all disks

Silicon Graphics' RAID 3 is shown in Figure 1-1. It is a "4+1" implementation: data is spread across four disk drives and a fifth "check" disk drive stores parity information. A parity bit on the check disk drive is calculated by an exclusive-OR (XOR) of the corresponding bits on the four data disk drives: for each bit of data, if there is an even number of 1's on the data disk drives, the check disk drive contains a 0; if the number of 1's is odd, the check disk drive contains a 1.

Drive 0	Drive 1	Drive 2	Drive 3	Drive 4
. . .	. . .	. . .	. . .	. . .
P	52	56	60	64
P	51	55	59	63
P	50	54	58	62
P	49	53	57	61
P	36	40	44	48
P	35	39	43	47
P	34	38	42	46
P	33	37	41	45
P	20	24	28	32
P	19	23	27	31
P	18	22	26	30
P	17	21	25	29
P	4	8	12	16
P	3	7	11	15
P	2	6	10	14
P	1	5	9	13

Figure 1-1 Silicon Graphics' RAID 3 (Stripe Depth of 4)

If a data disk drive fails, the information on this disk drive can be reconstructed by recalculating the parity of the remaining good data disk drives and comparing that parity bit-by-bit to the parity bits on the check disk drive. For each bit, if the parities agree, the failed bit was a 0, otherwise it was a 1. If the check disk drive fails, its data can be reconstructed by recalculating parity from the data disk drives.

Silicon Graphics' RAID 5, shown in Figure 1-2, is similar to RAID 3, except that the parity information is distributed across all of the disk drives rather than just a single disk drive. This distribution of parity data gives better performance for multiple writes in parallel (in RAID 3, multiple writes must be sequential because all writes require a write to the check disk drive). Reconstruction of a failed RAID 5 disk drive is done in the same manner as for a RAID 3 disk drive.

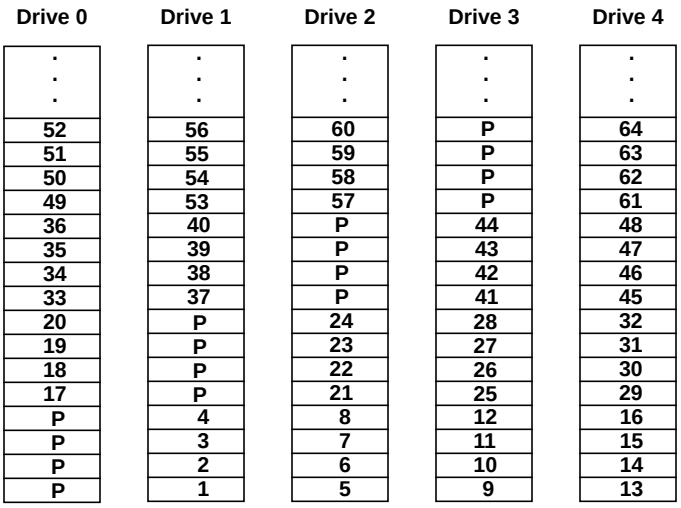

Figure 1-2 Silicon Graphics' RAID 5 (Stripe Depth of 4)

1.2 RAID Stripe Depth

Figure 1-1 and Figure 1-2 show the minimum "stripe depth" of 4. Each numbered slice of a disk drive is one 512-byte sector, and each disk drive in this example contains groups of four sequentially numbered sectors. Figure 1-3 and Figure 1-4 show a stripe depth of 8 for RAID 3 and RAID 5 respectively. The stripe size (the total number of sectors in each stripe) is the stripe depth times 4 (because each stripe uses four of the five disk drives for user data). The minimum stripe depth for any RAID is 4 and the maximum is 64.

The stripe width, depth, and size are analogous to the width, length, and area of a sheet of paper. For Silicon Graphics RAID, the width is always 4, and the depth is user setable. Since the total stripe size is always the depth times 4, this guide and the *raid*(1M) command use only the stripe depth rather than the more common "stripe size."

Drive 0	Drive 1	Drive 2	Drive 3	Drive 4
. . .	. . .	. . .	. . .	. . .
P	40	48	56	64
P	39	47	55	63
P	38	46	54	62
P	37	45	53	61
P	36	44	52	60
P	35	43	51	59
P	34	42	50	58
P	33	41	49	57
P	8	16	24	32
P	7	15	23	31
P	6	14	22	30
P	5	13	21	29
P	4	12	20	28
P	3	11	19	27
P	2	10	18	26
P	1	9	17	25

Figure 1-3 Silicon Graphics' RAID 3 with Stripe Depth of 8

Drive 0	Drive 1	Drive 2	Drive 3	Drive 4
. . .	. . .	. . .	. . .	. . .
40	P	48	56	64
39	P	47	55	63
38	P	46	54	62
37	P	45	53	61
36	P	44	52	60
35	P	43	51	59
34	P	42	50	58
33	P	41	49	57
P	8	16	24	32
P	7	15	23	31
P	6	14	22	30
P	5	13	21	29
P	4	12	20	28
P	3	11	19	27
P	2	10	18	26
P	1	9	17	25

Figure 1-4 Silicon Graphics' RAID 5 with Stripe Depth of 8

1.3 RAID and IRIX

Each RAID is viewed as a single 8 GB SCSI-2 compatible disk by the IRIX operating system. The RAID appears as a single disk at the PROM level also, but the PROMs cannot determine that it is a RAID. When a RAID is installed on a workstation or server, *hinv* from IRIX displays a single line of output for each RAID. An example is:

```
Disk drive: unit 5 on SCSI controller 0: RAID
```

A RAID is the same as a non-RAID disk in these ways:

- The default sector size is 512 bytes.
- The RAID has a volume header that is written just like any other volume header, and it uses the standard partition tables.
- You can use *mkfs*(1M) and *fx*(1M) (IRIX *fx* only, not standalone *fx*) as usual.
- Filesystems on a RAID are mounted in */etc/fstab*.
- Filesystems on a RAID can be exported for NFS mounting. (As with all NFS mounting, hard mounting a RAID filesystem is more reliable than soft mounting. It is particularly recommended for RAIDs.)
- Block and character device access is supported.
- The logical volume disk driver, *lv*(7M), supports RAIDs.
- The IRIS Volume Manager supports RAIDs.

A RAID is different from a non-RAID disk in these ways:

- A RAID must be formatted by the RAID administration program *raid*(1M).
- A RAID cannot be used as the system disk: it cannot contain the */* (root) and */usr* filesystems or the primary swap space.
- Third party disk drives cannot be used in a RAID.
- The RAID device driver, *usraid*(7M), is used rather than other device drivers, such as *dks*(7M) and *jag*(7M).
- The sector size cannot be changed.
- Standalone *fx* is not supported.
- *Add_disk*(1) cannot be used.

1.4 RAID Configuration Options

RAID configuration has two components:

- Configuring RAIDs on SCSI channels at hardware installation time
- Configuring a RAID at the time it is formatted

To perform this configuration, you must decide:

- The RAID level
- The stripe depth
- The number of RAIDs per SCSI channel

The next three sections explain how to choose the configuration options that will maximize your applications' I/O performance. You must choose the level of each RAID and the number of RAIDs per SCSI channel before RAID units can be installed in a system.

1.4.1 RAID Level and Stripe Depth

The number of users using a RAID, their applications, and the I/O characteristics of the applications that access the RAID all influence the choice of RAID 3 or RAID 5 and the stripe depth. For four combinations of RAID level and stripe depth, Table 1-1 shows the number of users, applications, I/O stream, and transfer size that is best suited for each combination.

The best stripe depth for maximum performance of a RAID depends upon the typical read/write size of the applications that access the RAID. Optimally, each read should access just one of the five disk drives in a RAID 5. For example, if the typical read is 2 KB, a good stripe depth would be 4: four sectors of 512 bytes each is 2 KB. The entire 2 KB can be read from a single disk drive.

RAID Level	Stripe Depth	Typical Number of Users	Typical Applications	Typical I/O Stream	Typical Transfer Size
3	small		not useful		
3	large	low	supercomputer applications, graphics	single stream of large requests (small number of requests per second and a massive amount of information in each request)	1 MB or more is optimal; the range is 64 KB and up
5	small	high	databases, NFS, file serving	multiple streams of small requests (a large number of requests for a small amount of information each time)	8 KB or less is optimal; the range is .5 KB to 64 KB
5	large	low	supercomputer applications, graphics	single stream of large requests (small number of requests per second and a massive amount of information in each request), mostly reads with few writes	1 MB or more is optimal; the range is 64 KB and up

Table 1-1 RAID 3 and RAID 5 Characteristics

A rule of thumb for choosing the stripe depth is to choose a stripe depth that accommodates the typical transfer size or is slightly larger. When the typical transfer size is unknown, a stripe depth of 32 is a good choice. Larger stripe depths are better for the controller because it is more efficient to read contiguous sectors from a single disk drive in a single read operation than many reads of small numbers of sectors from several disk drives.

The RAID level and stripe depth are specified when formatting a RAID (see Chapter 2, “Formatting a RAID”).

1.4.2 How Many RAID Units per SCSI Channel?

Because of packaging and cabling limitations, a maximum of 12 SCSI devices are supported on each SCSI channel (each RAID unit is counted as a single SCSI device). However, I/O performance can be affected by putting too many RAID units on a single SCSI channel.

You can use these rules of thumb when determining how many RAID units to attach to a channel:

- The rated capacity of a channel is 20 MB per second, but because of overhead, assume that the capacity is 18 MB per second.
- A RAID 3 at maximum throughput uses most of the capacity of a channel.
- A RAID 5 at maximum throughput uses about one quarter of the capacity of a channel.
- For RAID 5, the sizes of the transfers and the number per second determine how many transfers (and therefore RAIDs) can fill the capacity of the channel.
- A non-RAID disk at maximum throughput uses 3 MB to 4 MB per second.
- Putting slow devices such as tape drives and CD-ROM drives on a channel with one or more RAID units can adversely impact the performance of the RAIDs.
- In general, RAID units should be put on channels with other RAID units and non-RAID disk drives only.

RAID units are attached to channels at the time of hardware installation. Changing the configuration, for example, because you determine that there are too many RAID units formatted as RAID 5 on a single channel, requires a service call by your support provider.

The *hinv*(1M) command can be used to figure out how many RAID units are on each channel. Each controller listed in *hinv* output drives one channel. Counting the number of units listed for a controller tells you how many devices are on the channel for that controller. The *hinv* output for a controller with 3 RAID units and 5 other disk drives on its channel looks like this:

```
Integral SCSI controller 4: Version WD33C95A
Disk drive: unit 15 on SCSI controller 4: RAID
Disk drive: unit 14 on SCSI controller 4: RAID
Disk drive: unit 13 on SCSI controller 4: RAID
Disk drive: unit 5 on SCSI controller 4
Disk drive: unit 4 on SCSI controller 4
Disk drive: unit 3 on SCSI controller 4
Disk drive: unit 2 on SCSI controller 4
Disk drive: unit 1 on SCSI controller 4
```

1.5 RAID LEDs

Several sets of LEDs (Light Emitting Diodes) in the RAID unit display status information. This section contains diagrams that show the locations of these LEDs and tables that list their meanings.

1.5.1 RAID Controller LEDs

An array of eight LEDs on the front edge of the RAID controller (inside the RAID unit) displays status information for each of the five disk drives and for the RAID as a whole. These LEDs are shown in Figure 1-5 and their meanings are listed in Table 1-2. The table assumes that the RAID unit is powered on. The term “failed” means that the RAID unit or a disk drive is either not responding to commands or is responding but not operating properly. The term “down” means that the RAID unit or a disk drive has failed or it is not operating because it has been manually shut down with the *raid* **–d** command.

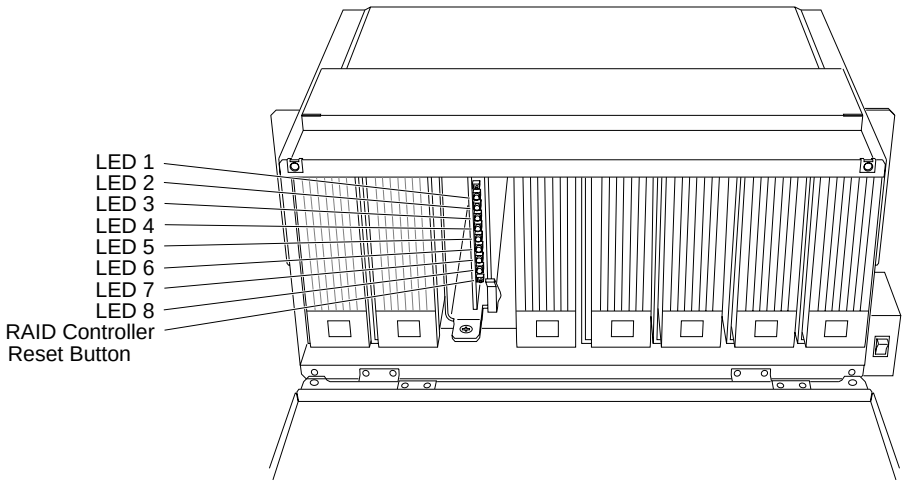

Figure 1-5 RAID Controller LEDs

LED	Lit Means	Off Means	Flashing Means
1	The RAID is processing commands.	The RAID is not processing commands.	The RAID is processing commands.
2	The RAID has failed.	The RAID has failed.	The RAID is operating normally (note that the on / off cycle can be as long as 10 seconds).
3	A maintenance command such as rebuilding a disk drive or checking parity is active.	The RAID is operating normally.	A maintenance command such as rebuilding a disk drive or checking parity is active.
4	Disk drive 0 is down.	Disk drive 0 is operational.	N / A
5	Disk drive 1 is down.	Disk drive 1 is operational.	N / A
6	Disk drive 2 is down.	Disk drive 2 is operational.	N / A
7	Disk drive 3 is down.	Disk drive 3 is operational.	N / A
8	Disk drive 4 is down.	Disk drive 4 is operational.	N / A

Table 1-2 RAID Controller LEDs

1.5.2 RAID Disk Drive LEDs

Figure 1-6 shows the fronts of two RAID disk drives. In the figure, they are shown rotated 90 degrees counterclockwise from their position when installed in the RAID unit. See also Figure 4-2.

An amber LED behind the bezel of the disk drive lights when the disk drive is down (failed or shut down with *raid* **–d**). When this LED is lit, the LED on the edge of the RAID controller that corresponds to this disk drive will be lit also (see Section 1.5.1).

The green LED on the front of each disk drive is lit when it is servicing a command.

Disk drives in the RAID unit are numbered 0 to 4, left to right. The number of each disk drive displayed on its front must match its drive number. The push buttons above and below the number are used to increase or decrease it.

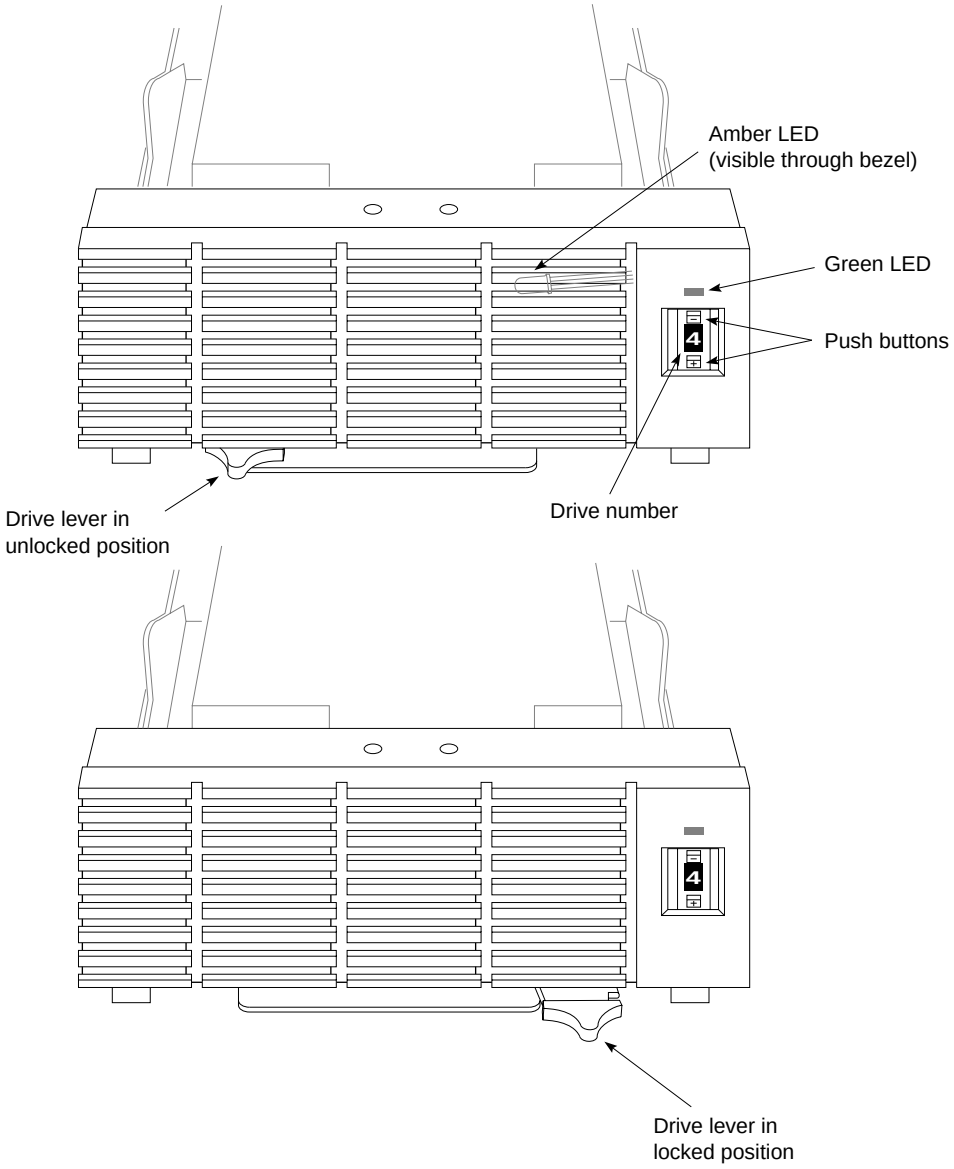

Figure 1-6 Disk Drive Front

1.6 RAID Failure Prediction

The 2.0 GB disk drives used in RAIDs have a sophisticated predictive failure analysis (PFA) feature. They use internal diagnostics and device information to attempt to give at least 24 hours' notice before they fail.

PFA information is written to the system console and to the file */var/adm/SYSLOG* when:

- A disk drive predicts that it will fail within 24 hours.
- The system is powered on and a disk drive has predicted its failure.
- The *raid* command is given with the **–L** option.

An example of a PFA message is:

```
disk n is predicting failure, replace as soon as possible
```

Chapter 5, "Error Messages," contains a complete list of PFA messages. Chapter 4, "Recovering after a Disk Warning or Failure," explains what to do if you get a PFA message.

1.7 RAID Failure Recovery

When a single disk drive in a RAID unit fails, the data on the failed disk drive can be reconstructed from the remaining disk drives, while IRIX and applications accessing the RAID continue to operate. However, if two disk drives fail or if a disk drive and the controller fail, the data cannot be reconstructed, and IRIX and applications can no longer access the RAID. Backup tapes should be made for data on a RAID just as on any other type of disk.

With RAID, you can "hot plug" a replacement disk drive: a disk drive can be removed from the RAID unit and a replacement inserted while IRIX is running and applications are accessing the RAID. See Chapter 4, "Recovering after a Disk Warning or Failure," for more information.

Caution: When a disk drive has failed, removing any disk drive other than the failed disk drive will **probably** result in the loss of **all** data on the RAID.

To reduce the risk of pulling the wrong disk drive when a disk has returned a PFA warning, you can use the **–d** option of the *raid* command to mark the disk down before pulling it. This will light the amber LED in the disk drive (see Figure 1-6) and the LED on the controller that corresponds to the disk drive (see Figure 1-5) to give a clear indication of which drive must be pulled.

Chapter 2

Formatting a RAID

When you format a RAID, you specify the addressing order of the sectors in the RAID. Formatting is required when a RAID is first installed on a workstation or server and when you want to change the RAID level, the stripe depth, or the number of sectors of each disk drive that are used. You may want to make changes because you are doing performance tuning or because applications with different I/O characteristics are now using the RAID.

Caution: All data on a RAID is lost when you reformat it. Be sure to make a backup of all files that are located on the RAID before reformatting it with the *raid*(1M) command.

Before formatting a RAID, be sure you know the answers to these questions:

- Which RAID level: RAID 3 or RAID 5? (See Section 1.4.1, "RAID Level and Stripe Depth.")
- What stripe depth? (See Section 1.4.1.)
- How many sectors of the RAID will be used? The default is the entire usable portion of each disk drive. Choose fewer sectors (leaving some portion of each disk drive unusable) when doing demonstrations in order to reduce the formatting time.

Follow these steps to format or reformat a RAID:

1. Get the controller and unit numbers of the RAID.
2. If you are reformatting the RAID, save all data on the RAID on tape.
3. Become superuser.
4. If you are reformatting the RAID, unmount any filesystems on the RAID.
5. Run *MAKEDEV*(1M) to create the RAID device files:

   ```
   cd /dev
   ./MAKEDEV rad
   ```

6. If you want to, you can check the health and connectivity of the RAID with *fx*(1M):

   fx "rad(*controller***,***unit***)"**

   fx performs a controller test and displays a menu. At the prompt, exit *fx*:

   fx> **exit**

7. Give this *raid* command to format or reformat the RAID:

   raid -*level* **-s** *depth* **-S** *sectors* **/dev/rdsk/rad***controller***d***unit***vh**

   where:

level	is the RAID level: 3 or 5
depth	is the stripe depth, and *depth* must be a multiple of 4 between 4 and 64
sectors	is the number of sectors per RAID to format (maximum 16777216)
controller	is the controller number (from *hinv*(1M))
unit	is the unit number (from *hinv*)

   –s *depth* and **–S** *sectors* are optional. If these options are not specified, the entire RAID is formatted and the stripe depth is 32. Formatting the entire RAID takes 15–20 minutes.

8. Invoke *fx* in expert mode and put a volume header on the RAID using these commands (shown with example output for *controller* = 4 and *unit* = 8):

```
fx -x "rad(controller,drive)"
fx version 5.0, Apr 19, 1993
...opening rad(4,8,)
...controller test...OK
fx: Warning - invalid label on disk, ignored
Can't get drive geometry, assuming 64 sectors/track
Can't get drive geometry, assuming 8 heads
Unable to get device geometry, assuming default
Scsi drive type == ULTRA U144 RAID SGI 1.0
warning: can't read sgilabel on disk
creating new sgilabel
...creating default bootinfo
Can't get drive geometry, assuming 64 sectors/track
Can't get drive geometry, assuming 8 heads
Unable to get device geometry, assuming default
...creating default partitions
...creating default volume directory

----- please choose one (? for help, .. to quit this
menu)-----
[exi]t [d]ebug/ [l]abel/ [a]uto
[b]adblock/ [exe]rcise/ [r]epartition/ [f]ormat
fx> label/create/all

...creating default bootinfo
Can't get drive geometry, assuming 64 sectors/track
Can't get drive geometry, assuming 8 heads
Unable to get device geometry, assuming default
...creating default partitions
...creating default sgiinfo
...creating default volume directory

----- please choose one (? for help, .. to quit this
menu)-----
[exi]t [d]ebug/ [l]abel/ [a]uto
[b]adblock/ [exe]rcise/ [r]epartition/ [f]ormat
fx> exit

label info has changed for disk rad(4,8,). write out
changes? (yes) yes
```

9. Confirm that the RAID is formatted properly:

   raid -p /dev/rdsk/rad*controller***d***unit***vh**

   The output should contain this information in a single line:

   /dev/rdsk/rad*controller*d*unit*vh: RAID*level*, sectors: *sectors*, stripe size: *depth*, OK

   See Section 5.1.1, “Messages from Format Operations,” if the message doesn’t end with “OK”.

10. To put one or more filesystems on the RAID, use *mkfs*(1M) to construct the filesystem(s), create the mount point(s), edit */etc/fstab*, and mount the filesystem as usual. For a complete description of the procedure, see Section 4.2, “Maintaining File Systems,” of the *IRIX Advanced Site and Server Administration Guide*.

11. If you reformatted the RAID, restore the data from tape.

Chapter 3

Routine Maintenance Tasks

RAIDs require routine monitoring and maintenance to ensure high reliability. This chapter describes a variety of routine administration tasks that are specific to RAID or particularly important for RAID:

- Section 3.1, “Making Tape Backups”
- Section 3.2, “Monitoring /var/adm/SYSLOG”
- Section 3.3, “Getting Configuration and Status Information”
- Section 3.4, “Checking the Integrity of Parity”
- Section 3.5, “Checking for Failed Disk Drives”
- Section 3.6, “Downloading New Firmware”

3.1 Making Tape Backups

Even though a RAID provides redundancy and disk drive failure prediction, it is still important to make regular tape backups of data stored on a RAID.

The *IRIX Advanced Site and Server Administration Guide* provides information about tape backup tools and strategies in Chapter 5, “Backing Up and Restoring Data.”

3.2 Monitoring */var/adm/SYSLOG*

Warnings and error information for RAIDs are written to the file */var/adm/SYSLOG* and the console by the *raid*(1M) command if the **–L** option is given and by the RAID device driver. (See *usraid*(7M) for more information about the RAID device driver.) */var/adm/SYSLOG* should be checked frequently, about every 12 hours, for messages that predict disk drive failure.

An easy way to monitor */var/adm/SYSLOG* automatically is to set up a *cron*(1M) job that checks for messages and mails the result to root or a system administrator. Follow these steps to set up a *cron* job:

1. Become superuser.
2. Write the current set of *cron* jobs to a temporary file:

   crontab -l > *tempfile*
3. Edit the temporary file and add a single line like this (shown on multiple lines here):

   ```
   0      4       *       *       *       egrep
   '/dev/*dsk/rad' /var/adm/SYSLOG |
   mail -s "RAID warnings" userid
   ```

   userid is the user ID that you want the mail sent to. At 4 a.m. every day, */var/adm/SYSLOG* is searched for a string included in all RAID messages and the output is mailed to *userid*.
4. Add another line just like the one in the previous step, but change the 4 to 16:

   ```
   0      16       *       *       *       egrep
   '/dev/*dsk/rad' /var/adm/SYSLOG |
   mail -s "RAID warnings" userid
   ```

   This line performs that same check as above, but at 4 p.m. every day.
5. Tell *cron* to use the new version of *tempfile*:

   crontab *tempfile*

Another way to monitor messages in */var/adm/SYSLOG* is to use *syslogd*(1M). See the *syslogd*(1M) manual page for details.

Section 5.1, “raid Error and Warning Messages,” and Section 5.2, “Messages from the RAID Device Driver,” list the messages that can appear in */var/adm/SYSLOG* and describe what to do about them.

3.3 Getting Configuration and Status Information

You can get configuration and status information about one or all RAIDs on a workstation or server by giving the **–p** option to the *raid* command. With this option, *raid* reports on disk drives that are down, RAIDs that have not been formatted, and internal configuration information in a RAID that is not consistent.

To get this information, give this command as superuser (omit the */dev/rdsk* argument to check all RAIDs):

/etc/raid -p /dev/rdsk/rad*controller***d***unit***vh**

controller is the controller number and *unit* is the unit number (both can be obtained from *hinv*(1M)) of a RAID. An example of the output from a correctly configured RAID is:

```
/dev/rdsk/rad0d3vh: RAID3, sectors: 16777216, stripe size: 64,
OK
```

Output is written to */var/adm/SYSLOG* and the console as well as being displayed in your shell window when you give the **–L** option on the *raid* command line.

Other messages that can follow the configuration information instead of `OK` are:

- `drive` *n* `is down`
- `not initialized`
- `must reformat`

3.4 Checking the Integrity of Parity

You can use the *raid* program to check the integrity of parity and correct errors on one RAID or all RAIDs. When checking integrity, all disk drives must be operational. Checking parity at least every few days is recommended.

To check parity on a RAID, give this command (omit the */dev/rdsk* argument to check all RAIDs) as superuser:

raid -i /dev/rdsk/rad*controller***d***unit***vh**

controller is the controller number and *unit* is the unit number (both can be obtained from *hinv*). *raid* silently repairs any parity problems it finds by regenerating the parity to match the data on the disk drives. This will take 15 to 20 minutes per RAID on a lightly loaded system.

To set up a *cron* job that checks parity every night at 1 a.m., follow the procedure in Section 3.2, except add this line instead (shown wrapped) to *tempfile*:

0 1 * * * raid -i -L |
mail -s "RAID warnings" *userid*

3.5 Checking for Failed Disk Drives

The *raid* command with the **–c** option can be used to check for failed disk drives:

- To check for disk drives marked down (failed disks and disks marked down with **–d**) on all RAIDs, give this command as superuser:

  raid -c

- To check for disk drives marked down on just one RAID, give this command as superuser:

  raid -c /dev/rdsk/rad*controller***d***unit***vh**

  controller is the controller number and *unit* is the unit number (both can be obtained from *hinv*).

This command takes a few seconds. Output is written to */var/adm/SYSLOG* and the console as well as being displayed in your shell window when you give the **–L** option, too.

When no disk drives are down, there is no output from the command. These messages are displayed if *raid* detects a down disk drive or other problem:

- `raid: warning: /dev/rdsk/rad`*controller*`d`*unit*`vh: drive` *n* `is down`
- `raid: error: /dev/rdsk/rad`*controller*`d`*unit*`vh: transfer of config info to/from disk 2 failed: I/O error`
  `raid: warning: /dev/rdsk/rad`*controller*`d`*unit*`vh: drive` *n* `is down`
- `raid: error: /dev/rdsk/rad`*controller*`d`*unit*`vh: controller has been replaced`
  `raid: error: /dev/rdsk/rad`*controller*`d`*unit*`vh: it doesn't match disk config info`
- `raid: error: /dev/rdsk/rad`*controller*`d`*unit*`vh: disk` *n* `has been replaced, but it wasn't marked down`
  `raid: error: /dev/rdsk/rad`*controller*`d`*unit*`vh: the RAID should be rebuilt`

Rebuilding the RAID with *raid* **–r** makes a down disk drive operational. For more information, see Section 4.1, "Replacing a Disk Drive."

Each time the system is rebooted, RAIDs are checked for failed disk drives automatically with this command:

```
raid -c -m -L
```

The **–m** option tells the *raid* command to check for disk drives that were replaced while the system was off. If it finds new disks, it marks them as failed so that any data on them is not used. See Section 5.1, "raid Error and Warning Messages," for more information.

3.6 Downloading New Firmware

RAID firmware can be downloaded to flash memory in the RAID controller. When IRIX software releases include new RAID firmware (see the *IRIX Release Notes*), you may need to download the new firmware into each RAID individually to update your system.

To update a particular RAID, first make sure that no applications are using the RAID, then give this command as superuser:

`raid -l` *firmware* **`/dev/rdsk/rad`***controller***`d`***unit***`vh`**

firmware is the pathname of the file that contains the new firmware, *controller* is the controller number, and *unit* is the unit number (both can be obtained from *hinv*). The downloading takes about a minute.

After downloading is complete, reset the RAID controller by follow this procedure:

1. Open the RAID unit as explained in steps 5 and 6 of Section 4.1, “Replacing a Disk Drive.”
2. Press the RAID controller Reset button shown in Figure 1-5.
3. Close the RAID unit as explained in steps 13 and 14 of Section 4.1.

Chapter 4

Recovering after a Disk Warning or Failure

This chapter contains several error recovery procedures:

- Section 4.1, “Replacing a Disk Drive,” describes the procedure for replacing a disk drive that has failed or has predicted its failure.
- Section 4.2, “Restoring Data from Tape after Two Failures,” explains the procedure to follow when two disk drives or a disk drive and the controller in a RAID have failed.
- Section 4.3, “Restarting a Hung RAID,” explains how to proceed if a RAID stops responding to commands.
- Section 4.4, “Resetting a RAID to Factory Defaults,” explains how to put a RAID into an uninitialized state.

4.1 Replacing a Disk Drive

It is important to replace a failed or about-to-fail disk drive as soon as possible to reduce the possibility of a second disk drive’s failing before the first is replaced. If a second disk drive fails, the data on the entire RAID must be recovered from tape backups.

Caution: Don’t remove the wrong disk drive! When a disk drive has failed, removing any disk drive other than the failed disk drive will **probably** result in the loss of **all** data on the RAID. Before removing any disk drive, confirm that you are removing the correct one.

Follow the procedure below to replace a disk drive in a RAID unit and rebuild the disk drive. The replacement disk drive must be the same type as the failed disk drive (contact your Silicon Graphics Sales Representative or distributor of Silicon Graphics equipment to purchase disk drives for RAID). During rebuilding, the contents of the non-failed disk drives are used to reconstruct the data that was stored on the failed disk drive.

1. Make a backup tape of the data on the RAID. This step is optional, but it is highly recommended.
2. Get the *controller* and *unit* of the disk drive you plan to replace from *hinv*(1M).
3. If the disk drive is predicted to fail, but has not actually failed yet, telling the controller to mark the disk drive as down is recommended. Give this command as superuser:

   raid -d *driveno* **/dev/rdsk/rad***controller***d***unit***vh**

   driveno is the drive number of the disk you are replacing. It was in any PFA messages. This command won't mark the about-to-fail disk drive as down if there is another disk drive in the RAID already marked down, or if the RAID controller and the disk drive contain inconsistent information. The output of this command explains if it found any disk drives already marked down, and if it was able to force the disk drive down as requested.

   If a disk drive is already marked down, it should be replaced before replacing a disk drive that has been predicted to fail.
4. Because you could lose the data on the RAID if you remove the wrong disk drive, you may want to reduce the chances of losing data. Choose and implement one of these options (least risky to most risky):

   - Shut the system down and power it off. Before powering off the system, mark the drive to be replaced with a Post-it™ note, since the LEDs that note its location won't be lit.
   - Quiesce the RAID. Stop any applications that are accessing the RAID and unmount the RAID filesystem(s).
   - Hot plug the new disk drive. No action is required.
5. Open the cabinet that contains the RAID unit so that you can see the front of the RAID unit. Figure 4-1 shows the front of one type of RAID unit enclosure; the other type of enclosure is very similar.

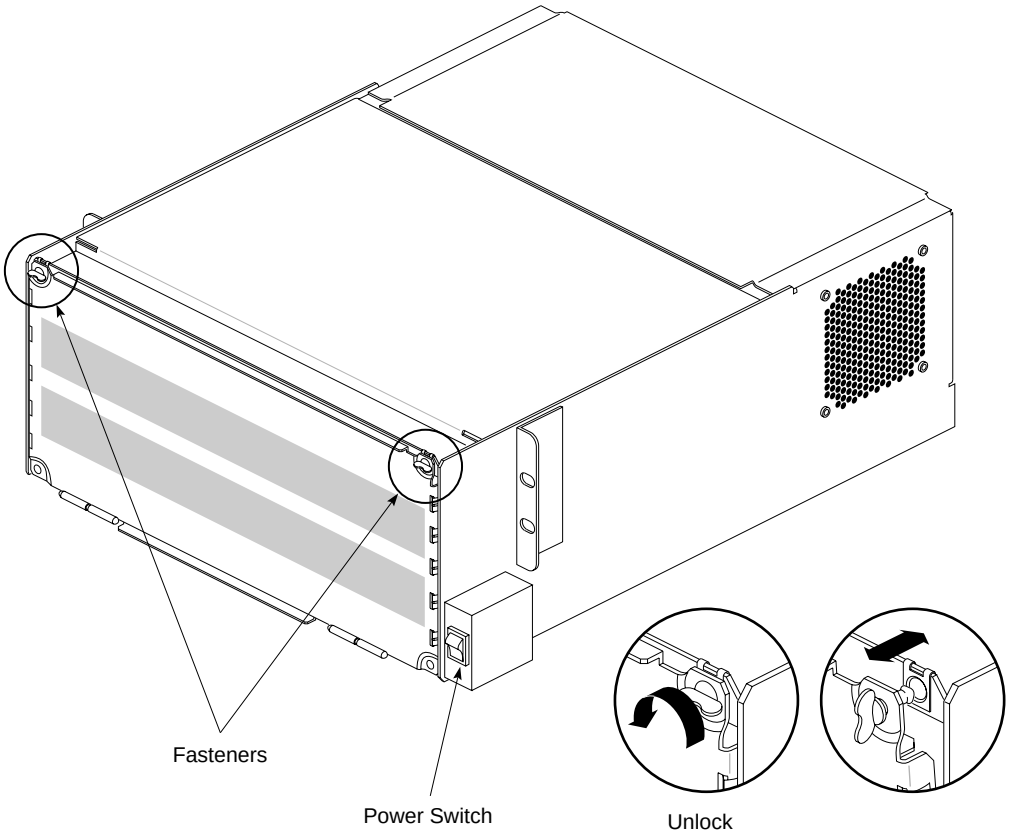

Figure 4-1 RAID Unit Front

6. Twist and pull the fasteners at the top of the unit front (see Figure 4-1) and lower the door into a fully open position.
7. Note the physical location of the disk drive you plan to replace. On a failed disk drive or on a disk drive that has been marked down, its amber LED is lit (see Section 1.5.2, “RAID Disk Drive LEDs”) and its LED on the edge of its controller is lit (see Section 1.5.1, “RAID Controller LEDs”). These LEDs are not lit on disk drives that have been predicted to fail but haven’t been marked down.
8. Push up on the drive lever on the lower left side of the disk drive you want to remove to release the disk drive module.
9. Slide the disk drive module out (see Figure 4-2).

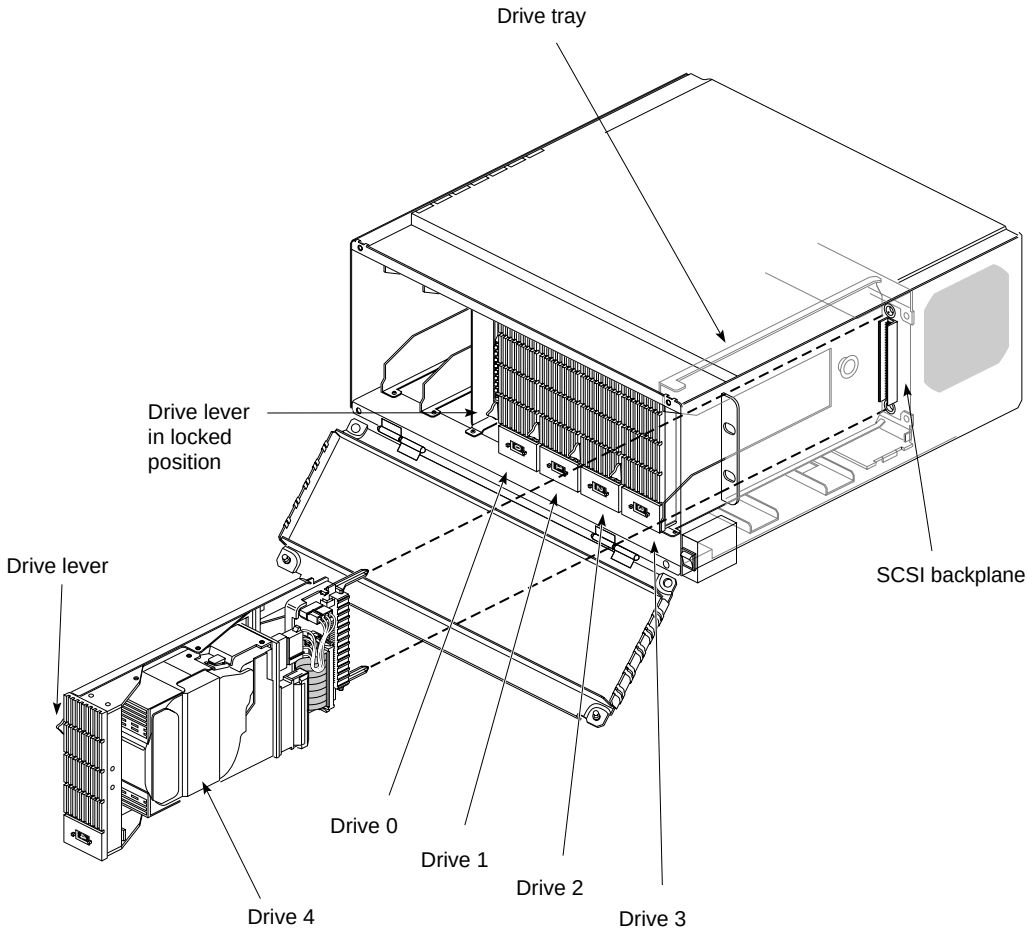

Figure 4-2 Removing or Installing a Disk Drive

10. Use the push buttons on the front of the replacement disk drive to set its drive number to the same number as the disk drive you just removed (see Figure 1-6).
11. Slide the replacement disk drive module into the opening in the RAID unit. As you slide in the module and push it to the back, its drive lever moves down.

12. Push down on the drive lever to lock it in position. Try pulling the module out to make sure that it is locked.
13. Raise the unit door and turn the fasteners to secure the front cover.

    Caution: The front cover must be in place during operation for proper air flow and to comply with FCC and other regulatory requirements.

14. Replace any panels of the cabinet that you removed earlier.
15. If you powered off the workstation or server, turn the power back on and wait one minute for the RAID disk drives to spin up.
16. If you unmounted the RAID filesystem(s), mount it again.
17. Rebuild the RAID by giving this command as superuser:

    `/etc/raid -r /dev/rdsk/rad`*controller***`d`***unit***`vh`**

    When the rebuilding of the RAID is complete (about 30 minutes on a lightly loaded system), the RAID is again fully functional.

4.2 Restoring Data from Tape after Two Failures

As explained in Section 1.7, “RAID Failure Recovery,” when two disk drives or a disk drive and the controller in a RAID fail, the data on the RAID becomes inaccessible. Examples of situations that result in two failures are:

- A disk drive fails, but the failure is not noticed or not repaired, and then a second disk drive fails.
- A disk drive fails and while attempting to hot plug a replacement for the failed disk, the wrong disk drive is accidently removed from the RAID.
- A disk drive fails, and then the controller fails.

If two disk drives have failed, follow this procedure:

1. Replace the failed disk drives by following steps 4 through 15 in Section 4.1.
2. Format the RAID as if it were a new RAID using the procedure in Chapter 2, “Formatting a RAID.”
3. Reload the files that were on the RAID from a backup tape.

A service call is required to replace a failed RAID controller.

4.3 Restarting a Hung RAID

If a RAID becomes hung, it doesn't respond to commands. This can occur as a result of an internal hardware fault in the controller. Two ways to attempt to restart it are:

- Reset the RAID controller.
- Power cycle the RAID unit.

Caution: When you perform either of these procedures, there is some risk of corruption of data on the RAID, just as there is when you press the Reset button on a workstation or server.

To reset the RAID controller, follow these steps:

1. Open the RAID unit as explained in steps 5 and 6 of Section 4.1.
2. Press the RAID controller Reset button shown in Figure 1-5.
3. Close the RAID unit as explained in steps 13 and 14 of Section 4.1.

Some RAID units have a power switch as shown in Figure 4-1. To power cycle this type of RAID unit, follow these steps:

1. Open the cabinet that contains the RAID unit so that you can see the front of the RAID unit.
2. Turn off power to the RAID unit (see Figure 4-1 for the location of the switch).
3. Wait 15 seconds.
4. Turn on the power.
5. Replace the front panel of the cabinet.

To power cycle RAID units that don't have their own power switches, you must power cycle the entire workstation or server.

4.4 Resetting a RAID to Factory Defaults

Under certain error conditions, for instance when configuration information in the RAID is internally inconsistent, you may need to make a last ditch effort to get the RAID into a known state. The *raid* **–z** option puts a RAID into an uninitialized state, just as it is when it is shipped from the factory.

Caution: Using the **–z** option of *raid* makes all data on a RAID inaccessible. Do not use it if you need any of the data on the RAID.

Give this command to restore the internal configuration information of a RAID to its factory default, uninitialized state:

`raid -z /dev/rdsk/rad`*controller***`d`***unit***`vh`**

controller is the controller number and *unit* is the unit number (both can be obtained from *hinv*). To use the RAID again, you must format it (see Chapter 2, “Formatting a RAID”).

Chapter 5

Error Messages

This chapter lists error messages that you might see on a system with one or more RAIDs installed and RAID LED error conditions that can be seen after opening the RAID unit. The causes and possible solutions of these errors are explained. This chapter contains these sections:

- Section 5.1, “raid Error and Warning Messages,” lists messages from the *raid*(1M) command. These messages appear as standard output and optionally on the console and in */var/adm/SYSLOG*.
- Section 5.2, “Messages from the RAID Device Driver,” lists messages from the RAID device driver (see *usraid*(7M)). These messages always appear on the console and in */var/adm/SYSLOG*.
- Section 5.3, “LED Error Conditions,” describes error conditions that are indicated by LEDs in the RAID unit.

5.1 *raid* Error and Warning Messages

Error messages from the *raid* command are always displayed in your shell window. When the *raid* **–L** option is given, messages are also written to the console and */var/adm/SYSLOG*.

5.1.1 Messages from Format Operations

`drive` *n* `is missing, formatting failed`

`drive` *n* `is not the same size as the other drives:` *m* `sectors`

`drive` *n* `is not a supported drive type: vendor:` *string1*`, product:` *string2*`, revision:` *string3*

All disk drives must be present, working correctly, and of the same make/model.

`/dev/rdsk/rad`*controller*`d`*unit*`vh: RAID`*level*`, sectors:` *sectors*`, stripe size:` *depth*`, drive` *n* `is down`

`/dev/rdsk/rad`*controller*`d`*unit*`vh: RAID`*level*`, sectors:` *sectors*`, stripe size:` *depth*`, not initialized`

`/dev/rdsk/rad`*controller*`d`*unit*`vh: RAID`*level*`, sectors:` *sectors*`, stripe size:` *depth*`, must reformat`

The RAID is not formatted properly. Try steps 7 through 9 of Chapter 2 again. If you get the error message again, there is a hardware problem. Contact your local service provider.

5.1.2 Messages from Integrity Check Operations

`cannot check integrity, drive` *n* `is down`

All disk drives must be up and functioning. There is no point in checking the parity integrity of the RAID when a disk drive is down.

5.1.3 Messages from Check Down Operations

`controller has been replaced, it doesn't match disk config info`

If the controller has been replaced, then any indication of down disk drives is not valid. The **–m** option wasn't specified, so no action is taken.

```
controller has been replaced, reprogramming to match disks
```

If the controller has been replaced, then any indication of down disk drives is not valid. The **–m** option was specified, so the controller configuration will be forced to match that of the disk drives in the RAID.

```
disk n has been replaced, but it wasn't marked down
```

This can happen if a disk drive is replaced while the system was down.

```
disk n is being forced down, the RAID needs to be rebuilt
```

The **–m** option was specified, so the disk drive is being marked down. It must be rebuilt using the **–r** option.

```
drive n is down
```

The disk drive has failed or been marked down. It it has failed, it should be replaced and rebuilt. If it was simply marked down, it can just be rebuilt.

5.1.4 Messages from Firmware Download Operations

```
can't download firmware to an uninitialized RAID, re-format
(-3|-5) the RAID, then try again
```

The RAID must be formatted and in good working order before new firmware can be downloaded.

5.1.5 Messages from Force Down Operations

```
cannot force a drive down on this RAID
```

There is some problem with the RAID. A disk drive can't be forced down at this time.

```
the controller has been replaced. use the '-c' and '-m'
arguments to reprogram the controller to match the config info
on the drives
```

If the controller has been replaced, then any indication of down disk drives is not valid. Using the **–c** and **–m** options forces the controller configuration to

match that of the disk drives in the RAID. After that is done, disk drives can be marked down using the **–d** option, if desired.

`disk` *n* `has been replaced, but it wasn't marked down`

This can happen if a disk drive is replaced while the system was down. The **–c** and **–m** options are used at system boot time to detect this situation and mark any such disk drives down. They must be rebuilt using the **–r** option.

`aborting: this would down a second disk in the array`

Every attempt is made to avoid marking a second disk drive down, but the **–f** option will forcibly mark any disk drive down. If a second disk drive is marked down, the RAID must be reformatted before it can be used for anything. All data will be lost.

5.1.6 Messages from Rebuild Operations

`cannot rebuild this RAID`

`all disks are OK in this RAID`

There is either a problem with the RAID and it cannot be rebuilt at this time, or there are no disk drives marked down.

`drive` *n* `is missing, rebuild failed`

`drive` *n* `is not the same size as the other drives:` *m* `sectors`

`drive` *n* `is not a supported drive type: vendor:` *string1*`, product:` *string2*`, revision:` *string3*

All disk drives must be present, working correctly, and of the same make / model.

5.1.7 Messages from All Operations

`more than one disk is marked down`

More than one disk drive has been marked down. This can only happen when an administrator forces down more than one disk drive. The only recovery is to reformat the RAID.

`drive` *n* `is missing`

`transfer of config info to/from disk` *n* `failed`

There is configuration information stored in reserved sectors on each disk drive in a RAID. That information could not be read from the indicated disk drive. Either the disk drive has totally failed or the disk drive is not plugged in. This message will probably be accompanied by messages from the kernel.

`drive` *n*`'s config info is unrecognizable`

`drive` *n*`'s config info contains illegal values`

`disk` *n* `config info doesn't match NVRAM info`

These are all indications that something is wrong with the disk drive in the given slot number.

`drive` *n* `should be in slot` *m*`, raid ignored`

The configuration information from the reserved sectors of each disk drive shows that the disk drives in the RAID have been moved around.

`controller and drives disagree on which drive failed,`
`controller says` *n* `failed, while drives say` *m* `failed`

The configuration information from the reserved sectors of each disk drive is inconsistent. It is unclear which situation is true. Depending upon which situation is true, all the data in the RAID could be lost. The controller's NVRAM claims that one disk drive failed, and the configuration information in the reserved sectors of that disk drive match all the other disk drives, but there is a second disk drive whose configuration information does not match. Probably a double failure has happened, and the RAID must be reformatted and reloaded from a backup tape.

`raid corrupted, must reformat`

The internal consistency of the RAID parity information has been compromised with no possibility of recovery. The RAID must be reformatted (wiping out all data on the RAID), and the contents reloaded from a backup tape.

`raid not initialized`

Formatting was not successfully completed. The RAID is not yet usable; it must be reformatted.

5.2 Messages from the RAID Device Driver

All messages from the RAID device driver begin with:

rad*controller*d*unit*s*partition*

RAID device driver messages are written to the console and */var/adm/SYSLOG*.

`disk` *n* `has failed, running in degraded mode`

The disk drive in slot *n* has failed; it should be replaced as soon as possible.

`disk` *n* `is predicting failure, replace as soon as possible`

The disk drive in slot *n* has returned a PFA warning. It will probably fail within 24 hours. It should be replaced as soon as possible. You should probably mark it down before replacing it, but that is optional.

`not ready, initializing command required`

The controller has not been configured, the configuration has been zeroed (with the *raid* **–z** option), or the controller configuration has been corrupted somehow. If the controller had previously been configured, try resetting the controller by pushing the controller Reset button (shown in Figure 1-5), otherwise a format operation is required.

`unrecovered hardware error, possibly a drive is not responding`

A disk drive is not responding, but the controller cannot tell if the disk drive is dead, simply not plugged in, or there is a fault in the bus connecting the controller to the disk drive. Check all the disk drives in the RAID unit.

`fixing parity integrity on stripe at LBA` *n*

An integrity check operation has detected that parity doesn't match the corresponding data for the stripe starting at the given block address. The parity is being updated to match the data automatically. If large numbers of these appear, or they recur often, contact your local service provider.

`recovered error`

A hard error was detected on one of the disk drives in the RAID, but the controller has reconstructed the data and taken recovery actions.

```
volume header not valid
```

The volume header on the RAID is not valid. *fx*(1M) should be used to write a new one.

```
retrying request
```

```
retries exhausted
```

A command has had an error and is being retried. After a certain number of retries, IRIX gives up and returns an error to the user.

```
fatal error on maintenance command
```

```
maintenance command LBA not changing, controller may need to
be RESET
```

```
unable to abort maintenance command, controller may need to be
RESET
```

A maintenance command (initialize disk drive, integrity check, or rebuild disk drive) has had an unexpected error. Retry the operation.

```
forcing a second drive down
```

The *raid* command was used to forcibly mark down a disk drive in a RAID where there was already one disk drive down. Data may have been lost.

```
FLASH EPROM programming error
```

```
FLASH EPROM sector protection error at page n
```

```
microcode download failed
```

```
new microcode has invalid checksum, ignored
```

```
new microcode has invalid signature, ignored
```

There was an error downloading firmware to the controller. Retry the operation.

```
auxiliary input 1 active
```

```
auxiliary input 2 active
```

There are auxiliary inputs on the controller for monitoring internal conditions. Contact your local service provider.

```
controller had internal parity error

internal target failure

unrecovered hardware error

unrecovered hardware error, 2nd fault while recovering

unrecovered hardware error, cannot find task ptr

unrecovered hardware error, recovery steps skipped

unrecovered hardware error, unknown cause

unrecovered hardware error, which stream is unknown

unspecified hardware error
```

The controller has reported an internal hardware fault. Use *fx* to run diagnostics, then contact your local service provider.

```
diagnostics failed on SCSI channel n: buffer test

diagnostics failed on SCSI chip channel n at ID m

diagnostics failed
```

Diagnostics were run on the controller and a hard fault was detected. Contact your local service provider.

```
corrupted definition page

corrupted partition page

corrupted units down info
```

The RAID configuration information (the RAID level, stripe depth, and failed disk drive information) returned by the controller is not in a recognizable form. Try resetting the RAID (Section 4.3, “Restarting a Hung RAID”). If that doesn’t work, try reformatting (see Chapter 2, “Formatting a RAID”). If that doesn’t work, use the **–z** option to zero out the controller configuration (see Section 4.4, “Resetting a RAID to Factory Defaults”). Contact your local service provider.

5.3 LED Error Conditions

RAID controller LEDs display these error conditions:

- LED 2 is not flashing.

  When LED 2 is not flashing (the flashing cycle can be very slow, up to 10 seconds), the RAID has failed, possibly due to an unrecoverable hardware error. Try both procedures in Section 4.3, “Restarting a Hung RAID,” to see if the RAID will start operating again. A service call by your support provider may be required to diagnose and repair the problem.

- Any of LEDs 4 to 8 are lit.

  Each lit LED indicates that a disk drive has failed. See Table 1-2 and Figure 1-5 to figure out which one. Chapter 4, “Recovering after a Disk Warning or Failure,” explains how to fix the problem.

When the amber LED in a RAID disk drive is on (see Figure 1-6 for the location of the amber LED), that disk drive is down. It could be down because it has failed or because it has been manually marked down with the *raid* **–d** command. See Chapter 4 for recovery procedures.

Appendix A

Programming Hints

This appendix contains suggestions for developers of applications that will be run on systems with RAIDs. These suggestions are designed to maximize performance of the RAIDs and the applications:

- Determine the typical size of reads and use this number to optimize the stripe depth. The *sar*(1M) command with the **–d** option reports the number of transfers to or from a block device and the number of bytes transferred.
- Hard mount filesystems on RAIDs that are NFS mounted rather than soft mounting them. (Hard mounting waits if the server is down, so data errors are returned after the connection times out. Soft mounting doesn't wait.)
- Expect different filesystem performance characteristics with RAID 3 than with a non-RAID disk. RAID 3 may not perform as you'd expect because of the many small things like inodes in a filesystem. For example, file creation and removal time may be very slow compared to data transfer times.
- Use raw or direct I/O instead of a filesystem to read more than 64 KB per disk access with RAID. To find out more about direct I/O, read about the F_DIOINFO command on the *fcntl*(2) manual page and the O_DIRECT flag on the *open*(2) manual page.
- Investigate using asynchronous I/O and direct I/O for large, data intensive applications if filesystem performance seems inadequate.
- When using a RAID as a raw disk, try to align I/O requests on stripe boundaries and make the sizes of requests some even multiple of the stripe depth.

Index

R

S

V

W

X