FUJITSU COMPUTERS
SIEMENS

Reliant UNIX *Online Documentation*

RAIDmaster

System Administrator's Guide

Edition February 1999

Copyright and Trademarks

1 Preface

The CS31, CS34, CS35 and CS37 RAIDmaster adapters are high-performance PCI RAID controllers. Under the Reliant UNIX operating system these controllers offer internal hardware caching and convenient creation of a RAID configuration tailored to the requirements of the individual environment. They also simplify the process of setting up the SCSI hardware and support redundant hot-swappable disk drives.

The functions are controlled via the user-friendly Storage Manager user interface.

1.1 Target groups

This manual is written for Reliant UNIX system administrators who are experienced in using and administering the operating system and have a sound knowledge of hardware.

1.2 Guide to the manual

This manual describes how to use the RAIDmaster adapters in Reliant UNIX systems:

* Chapter Preface
* Chapter Product overview
* Chapter Storage cabinets and devices
* Chapter PCI adapter installation
* Chapter The Storage Manager
* Chapter Theory of operation
* Chapter Troubleshooting guide
* Appendix: "Specifications" and "Assembly diagrams"
* Glossary

1.3 Notational conventions

The following notational conventions are used throughout this manual:

   Important information which you should note.

   Danger of data loss or damage to the device in the event of incorrect operation, or danger to health and safety.

typewritten text
Commands, options, variables, file names, user input and program output are represented by typewritten text within normal text.

italics
Variable parts of commands are written in *italics*, as are status indicators and specialist terms.

[Key symbols]and [Control elements of the user interface]
Keys are represented as they appear on the keyboard. If you explicitly have to enter uppercase letters, the shift key is specified, e.g. Shift + [A] for A. Keys that must be pressed simultaneously are linked by the '+' character (e.g. [Esc] + [A]) while keys which must be pressed one after the other are specified with nothing between them (e.g. [Esc] [Esc]).

Like key symbols, control elements of the user interface are shown as they appear on the screen, e.g. [View Log]Event Logs for the "View Log" button in the window.

Bear in mind that screen output on your system may not always be identical to the screen output shown in this manual.

2   Product overview

The CS31, CS34, CS35 and CS37 RAIDmaster adapters support:

- Reliant UNIX (5.42C or higher; for CS37: 5.44C or higher)
- SCSI-1, SCSI-2 and SCSI-3 with active termination for UltraSCSI (CS37: also for Ultra2 SCSI (LVD))
- 132MB/s PCI bus mastering
- Hardware scatter/gather function for top I/O performance
- Overlapped I/O and tagged command queuing for highest performance with SCSI-2/3 drives
- SCSI Storage Manager software
- RAID levels 0, 1 and 5
- Certified for Reliant UNIX (64-bit included)
- Up to 64MB of cache per adapter
- Key features include:Hot-swappable disk drives
- Hot spare capability (reserve disks)
- Event logging
- Optional e-mail notification
- Local configuration, diagnostics and I/O monitoring
- Optional: ECC protection for the cache memory
- Monitoring of ambient temperature and voltage
- Disk access statistics

2.1   The RAIDmaster adapters

2.1.1   The CS31 adapter

The CS31 PCI SCSI adapter is RAID-ready with integrated hardware caching.

The CS31's four SIMM sockets accept up to 64MB of cache, providing maximum performance and data integrity. RAIDmaster ECC technology detects and corrects data errors in the cache.

The CS31 model supports full 132MB/s bus mastering capabilities of the PCI bus, as well as SCSI-1 and SCSI-2 devices transferring data at up to 10MB/s across the SCSI bus (8 bit single-ended).

Environmental sensing capabilities are standard on the CS31 adapter. On-board temperature and voltage sensors for the server pass critical server management data to the Storage Manager for logging and for passing on to the system administrator.

The CS31 model is operated by a 20MHz 68030 microprocessor.

The CS31 model supports the addition of the CS01 dual SCSI bus expansion module, allowing the controller's single on-board SCSI bus to be expanded to three SCSI busses without requiring more than a single host slot. The three SCSI busses support up to 21 SCSI-1 and SCSI-2 devices.

 When storage cabinets are used, the number of available disks per SCSI channel is limited to the number of slots.

2.1.2   The CS34 adapter

The CS34 PCI SCSI adapter is RAID-ready with integrated hardware caching.

The CS34's four SIMM sockets accept up to 64MB of cache, providing maximum performance and data integrity. RAIDmaster ECC technology detects and corrects data errors in the cache.

The CS34 model supports full 132MB/s bus mastering capabilities of the PCI bus, as well as SCSI-1 and SCSI-2 devices transferring data at up to 20MB/s across the SCSI bus (16 bit differential).

Environmental sensing capabilities are standard on the CS34 adapter. On-board temperature and voltage sensors for the server pass critical server management data to the Storage Manager for logging and for passing on to the system administrator.

The CS31 model is operated by a 25MHz 68040 microprocessor.

The CS34 model supports the addition of the CS04 differential dual Wide SCSI bus expansion module, allowing the controller's single on-board differential SCSI bus to be expanded to three SCSI busses without requiring more than a single host slot. The three SCSI busses support up to 45 devices.

 When storage cabinets are used, the number of available disks per SCSI channel is limited to the number of slots.

2.1.3   The CS35 adapter

The CS35 PCI SCSI adapter is RAID-ready with integrated hardware caching.

The CS35's four SIMM sockets accept up to 64MB of cache, providing maximum performance and data integrity. RAIDmaster ECC technology detects and corrects data errors in the cache.

The CS35 model supports full 132MB/s bus mastering capabilities of the PCI bus, as well as SCSI-1, SCSI-2 and SCSI-3 devices transferring data at up to 40MB/s across the SCSI bus (16 bit single-ended).

Environmental sensing capabilities are standard on the CS35 adapter. On-board temperature and voltage sensors for the server pass critical server management data to the Storage Manager for logging and for passing on to the system administrator.

The CS35 model is operated by a 25MHz 68040 microprocessor.

The CS35 model supports the addition of the CS05 single-ended dual Wide SCSI bus expansion module, allowing the controller's on-board single-ended SCSI bus to be expanded to three SCSI busses without

requiring more than a single host slot. The three SCSI busses support up to 45 devices.

 When storage cabinets are used, the number of available disks per SCSI channel is limited to the number of slots.

2.1.4   The CS37 adapter

The CS37 PCI SCSI adapter is RAID-ready with integrated hardware caching.

The CS37's four SIMM sockets accept up to 64MB of cache, providing maximum performance and data integrity. RAIDmaster ECC technology detects and corrects data errors in the cache.

The CS37 model supports full 132MB/s bus mastering capabilities of the PCI bus, as well as SCSI-1, SCSI-2 and SCSI-3 devices transferring data at up to 80MB/s across the SCSI bus (16 bit LVD (as per Ultra2 standard)).

Environmental sensing capabilities are standard on the CS37 adapter. On-board temperature and voltage sensors for the server pass critical server management data to the Storage Manager for logging and for passing on to the system administrator.

The CS37 model is operated by a 66MHz i960 RD microprocessor and is fully compatible with the I20 (Intelligent Input/Output) specification.

The CS37 model supports the addition of the CS07 single-ended dual LVD SCSI bus expansion module, allowing the controller's on-board single-ended SCSI bus to be expanded to three SCSI busses without requiring more than a single host slot. The three SCSI busses support up to 45 devices, in both single-ended and LVD mode.

 When storage cabinets are used, the number of available disks per SCSI channel is limited to the number of slots.

The CS37 controller offers the following built-in performance features:

- **Predictive caching**: Disk read requests made by the host are analyzed to determine whether they are part of a disk access pattern. If a pattern is detected, the controller uses the pattern to predict which data the host is likely to request in the near future, then reads this data into the cache.

- **Intelligent hot spare drives**: If a drive fails, it is automatically replaced by a hot spare drive. RAID-1 and RAID-5 arrays are automatically rebuilt using the new drive. If multiple hot spares are available on a controller, an intelligent hot spare algorithm selects the best one based on capacity and bus location.

- **Logging environment parameters**: Standard on CS37 controllers. On-board temperature and voltage sensors pass critical hardware management data to the Storage Manager for logging and notification to the system administrator.

2.2   Expansion modules

RAIDmaster adapters provide an on-board connection for one SCSI bus. The expansion modules provide additional SCSI bus connections.

The expansion module CS01 provides two additional single-ended SCSI bus connections for the CS31 adapter.

The expansion module CS04 provides two additional differential Wide SCSI bus connections for the CS34 adapter.

The expansion module CS05 provides two additional single-ended Wide SCSI bus connections for the CS35 adapter.

The expansion module CS07 provides two additional single-ended LVD SCSI bus connections for the CS37 adapter.

2.3   RAIDmaster SCSI devices

RAIDmaster SCSI devices include hard drives in hot-pluggable carriers designed for insertion in any of the device bays of the released peripheral cabinets (see the Release Notes for your product).

SCSI hard drives

The SCSI hard drives, enclosed in hot-pluggable carriers, snap into any device bay in the peripheral cabinet.

A fault LED on each hard drive indicates a failed drive. Failed hard drives may be replaced without interruption of normal system operation by simply pulling out the drive carrier (hot plug function). When a replacement drive is inserted, the peripheral cabinet automatically spins-up the drive and starts the RAID rebuild process.

For immediate automatic replacement of failed drives, unused drives can be designated as "hot spares". Upon failure of any drive in a RAID-1 or RAID-5 array, the hot spare will automatically replace the failed drive and the missing data will be rebuilt onto the hot spare without user intervention.

2.4   Storage Manager

The Storage Manager is an easy-to-use application which performs on-line control, monitoring and analysis of RAIDmaster-equipped servers and workstations. Real-time I/O load analysis, on-line diagnostics, event logging and broadcasting, and device inventory management are available. The Storage Manager gives administrators unequalled flexibility and dramatically simplifies and reduces the cost of system administration.

The Storage Manager is part of Reliant UNIX for RM200/300/400. The Storage Manager's user interface runs in color text mode. The Storage Manager provides access to the RAIDmaster hardware and handles all management functions.

- Simplified installation

   The Storage Manager simplifies SCSI hardware installation. During hardware installation, the Storage Manager graphically guides you through device configuration. The Storage Manager even warns you of problems such as a failed hard disk.

- RAID configuration

   The Storage Manager makes RAID setup and configuration easy. All you need to do is answer a few questions and follow a simple procedure, then the Storage Manager constructs the RAID configuration that best meets your requirements.

- Real-time I/O analysis

   RAIDmaster is equipped to collect I/O statistics. The Storage Manager presents this data in real-time in an easy-to-understand format. By analyzing the relative number of disk reads and writes, the distribution of record sizes, and the number of cache hits and misses, system administrators can determine when they need to scale the I/O processing capabilities of their storage subsystem.

- On-line diagnostics

  To help analyze potential problem areas, the Storage Manager's built-in diagnostics may be executed on any combination of SCSI hard drives and disk arrays, allowing the hardware and media to be verified on-line, without interrupting normal system operation.

- Event logging

  RAIDmaster monitors and records the operating condition of your storage subsystem and maintains a complete log of key events. These event logs may be viewed through a scrolling window and filtered to show only events of interest. The Storage Manager's event logging capability shortens diagnosis time and allows the network administrator to take preventive action before users and data are affected.

- Event notification

  The Storage Manager allows you to designate who will be notified when a registered event occurs. Designated individuals are automatically notified of events through e-mail, allowing for timely response.

- Hardware inventory management

  The Storage Manager automatically discovers and inventories all SCSI devices attached to the RAIDmaster adapter in your system. Complete device information is at your fingertips including each device's type, make, model, operating status, transfer speed, capacity, and configured options. This information can be viewed on the screen or printed to create a complete record of your installation.

- Loading Firmware

  Firmware for the CS34/CS35/CS37 adapters is contained in flash ROM and may be upgraded by using Storage Manager.

3   Storage cabinets and devices

This chapter describes the released OLR (Online Replacement) cabinets (see Release Notes), SCSI devices and accessories, and gives instructions on configuring complete RAID subsystems with these products. OLR carriers are installed either in the peripheral cabinets or in the RM systems.

3.1   OLR SCSI devices

OLR SCSI devices include hard drives in hot-pluggable carriers designed for insertion in OLR cabinets. The hot-pluggable carrier connects the device's power and SCSI signals via a single DIN connector on the back of the housing. During removal of the carrier from the cabinet, the design ensures that DC power is present until the SCSI bus connection is broken. During insertion of the carrier, it ensures that power is applied before the SCSI bus connection is made, thereby minimizing signal glitches.

The SCSI-ID assignment to the individual slots is determined by the backplane coding of the OLR carriers. Additional termination is not required. In the case of peripheral cabinets, cables must be connected as described in the documentation provided with the cabinet.

Each OLR carrier has two LEDs that display the drive status. The green LED is the drive activity LED and is on whenever the drive is accessed. The red LED is the drive fault LED and indicates a failed drive. When a drive's fault LED is on solid (not flashing), the user should remove that drive from the storage cabinet and replace it with a good drive.

The red LED flashes to indicate several different types of operation:

- Red LEDs on all drives flash during power-up so that their functionality can be confirmed.
- Red LEDs on all drives in an array flash during the initial array-build operation.
- When a failed drive is replaced, the new drive's red LED flashes during the array-rebuild operation.
- A drive's red LED flashes during a low-level format operation.

The user should not remove a drive from the OLR Cabinet if either the green or red LED is flashing.

Failed hard drives may be replaced without interruption of normal system operation. Any failed drive may be removed by simply pulling out the drive carrier. Under Windows   NT, when a replacement drive is inserted, the RAIDmaster controller automatically spins-up the drive and starts the rebuild process.

Let the disk run out for at least 30   seconds before you remove the drive from the slide-in module.

For immediate automatic replacement of failed drives, extra unused drives can be designated as hot spares. Upon failure of any drive in a RAID-1 or RAID-5 array, the hot spare will automatically replace the failed drive and the missing data will be rebuilt onto the hot spare as a background activity without user intervention and without interruption of normal system operation.

OLR hard drives conform to the SCSI-2 specification and transfer data across the SCSI bus at fast-synchronous rates of up to 10MB/sec (8 bit), 20MB/sec (16 bit, wide) or 40MB/sec (Ultra2: LVD).

3.2   OLR subsystem configuration

OLR subsystems may be configured in a variety of ways, depending upon the storage capacity, performance, degree of fault-tolerance and types of SCSI devices desired (for detailed information on RAID systems see chapter Theory of operation).

To determine the proper configuration for your application, you should perform the following steps:

1. Determine RAID level(s) to be utilized. This should be based on the performance and capacity requirements of the system, and whether or not fault-tolerance is desired.
2. Determine the number and capacity of hard drives including hot spares. Depending on the RAID level, different numbers of drives are required to provide the same usable storage capacity.
3. Determine the type and number of RAIDmaster adapters and OLR cabinets to be used. Each adapter can

control up to seven SCSI devices on a single 8-bit SCSI bus or 15 devices on a 16-bit SCSI-Bus. Also consider the maximum number of OLR-devices per SCSI-bus in the storage cabinets.

Every RAIDmaster adapter can be upgraded to three SCSI busses per card with the SCSI bus expansion modules.

4. Select the amount of hardware cache memory (ECC SIMMs) to be installed. At least 4 MB is required for each adapter. The CS31, CS34 and CS35 adapters can be configured with up to 64 MB of cache, the CS37 with up to 64 MB. Adding more cache typically increases performance.

The following section describes the individual RAID levels including their advantages and disadvantages.

3.2.1 Determining the RAID level

Use the following descriptions as guidelines for determining which RAID level is best for your system. Multiple arrays of different levels may be configured on a single adapter and cabinet. Each array appears to your computer as one hard drive. The number of drives which may be included in an array depends on the RAID level.

RAID-0	**Data striping across multiple drives**
Benefits:	Maximized storage capacity
	Best performance
	Requires as few as two drives
Drawbacks:	No fault tolerance
Description:	A RAID-0 array is simply a group of hard drives which have been combined into a single logical storage unit using "striping" (See chapter Theory of operation). This logical storage unit looks like a single large disk drive to the computer, but the storage space is interleaved so that disk accesses will be evenly distributed across every drive in the group. A RAID-0 array must contain a minimum of two drives. The drives must all be connected to the same adapter.
Capacity:	Since no drive capacity is used for data redundancy, the capacity of a RAID-0 array will be equal to the combined capacity of all drives in the array.
Performance:	Adding more drives to a RAID-0 array increases performance in multiuser systems, since more user data requests can be serviced simultaneously.
Drive faults:	RAID-0 arrays are not fault-tolerant. If any drive in the array fails, the entire logical storage unit will fail.

RAID-1	**Mirrored data is written to pairs of drives**
Benefits:	Fastest fault-tolerant architecture
	Least performance degradation with failed drive
	Fastest rebuild
	Requires as few as two drives
Drawbacks:	Least efficient storage capacity
Description:	A RAID-1 array is a "mirrored" pair of drives which contain identical data. The RAID-1 array looks like a single drive to the computer and has the same capacity as a single drive,

but can perform two read operations at the same time. Both drives must be connected to the same adapter.

Capacity: Since a RAID-1 array is a "mirrored" pair of drives which contain identical data, the usable storage capacity of the array is half the capacity of the drives.

Performance: Both read and write operations are fast. Write operations are considerably faster than with RAID-5 arrays since much less overhead is required.

Drive faults: When a drive fails in a RAID-1 array, the array is said to be functioning in "degraded mode". Degraded mode in a RAID-1 array is slightly slower than normal operation, since all read operations must be rerouted to the mirrored drive. Write operations are not impacted. Array rebuilds occur quickly with minimal performance impact since the contents of the replaced drive are simply copied from its mirrored drive.A RAID-1 array only fails when both drives of the mirrored pair fail.

RAID-5   **Data striping with rotating parity**

Benefits: Fault-tolerant

Most efficient storage capacity

Drawbacks: Worst performance degradation with failed drive

Slowest rebuild

Requires at least three drives

Description: Each RAID-5 array consists of three or more drives which must be attached to the same adapter.

Capacity: Since the equivalent capacity of one drive in each RAID-5 array contains redundant data, the usable storage capacity of the array equals the combined capacity of all drives minus one drive.

Performance: Read operations are fast but write operations require much more overhead than in RAID-1 arrays and thus are slower. Adding more drives to a RAID-5 array group increases performance in multiuser systems since more user data requests can be serviced simultaneously.

Drive faults: When a drive fails in a RAID-5 array group, the array is said to be functioning in "degraded mode". Degraded mode and array rebuilds in a RAID-5 array are considerably slower than RAID-1 arrays, since missing data must be synthesized by XORing data read from the remaining drives in the array.Unless two drives within the same RAID-5 array fail, the RAID-5 will not fail.

3.2.2   Selecting RAIDmaster hard drives

Determine the number of hard drives you need to yield the total usable capacity required by your system. Refer to the preceding section to calculate the capacity depending on the RAID level selected. For this version, OLR hard drives are available in 2GB, 4GB and 9GB capacities and can be mounted in any device

bay in the OLR cabinet.

> If several SCSI buses are available on the adapter, you should plan redundant RAID arrays in such a way that associated drives are connected to different SCSI busses. If you use one disk per channel, you make sure that your system can still access the RAID array even if one SCSI bus fails or is blocked by a failed drive.

Since adding more drives to an array group typically increases performance, you may wish to configure your system using a larger number of smaller capacity drives. Although it is possible, it is not recommended that you mix drive capacities within an array, as this is a waste of storage space.

Optionally, RAID-1 and RAID-5 arrays can be protected by a hot spare drive, providing immediate, automatic array rebuild when a drive fails. Any hard drive model above may be configured as a hot spare as long as it is the same or larger capacity as the drives it protects. A single hot spare will protect all RAID-1 and RAID-5 arrays attached to the same adapter as the hot spare.

> If a SCSI Bus Expansion Module is used to configure an adapter for more than one SCSI bus, a single hot spare with an appropriate capacity attached to any of the adapter's SCSI busses will protect all arrays on all of the adapter's busses.
>
> To make sure that a hot spare is still available when a SCSI bus fails or is blocked by a failed drive, you should configure two hot spares on two different SCSI busses.

4 PCI adapter installation

Installing and configuring the RAIDmaster PCI adapter consists of the following steps detailed in this chapter:

1. One or more SIMMs, and if required an expansion module, are installed in the RAIDmaster adapter.
2. SCSI IDs, cables and termination are configured for SCSI devices in the RM300/RM400.
3. The adapter and SCSI devices are mounted in the RM300/RM400 or in the peripheral cabinets.
4. Storage Manager is run to verify proper hardware configuration and to configure disk arrays.

4.1 Installing SIMMs

RAIDmaster adapters support up to 64 MB of cache using four SIMM sockets. One, two, three or four SIMMs may be installed in the adapter (CS37: on the RAID expansion board). At least one SIMM is required for proper operation.

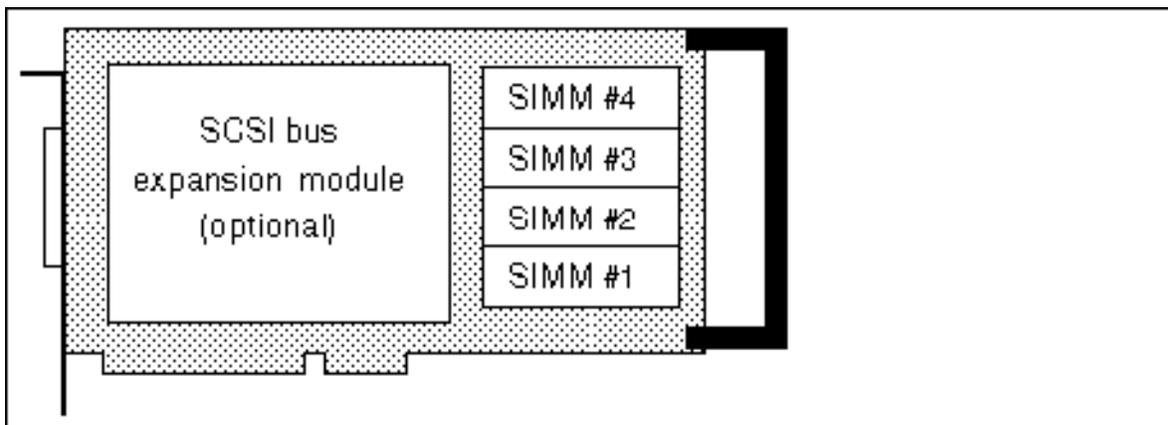SIMM sockets on PCI adapters reside on the main card (CS31, CS34 and CS35).

Figure 1: SIMM sockets on the CS31, CS34 and CS35 main card

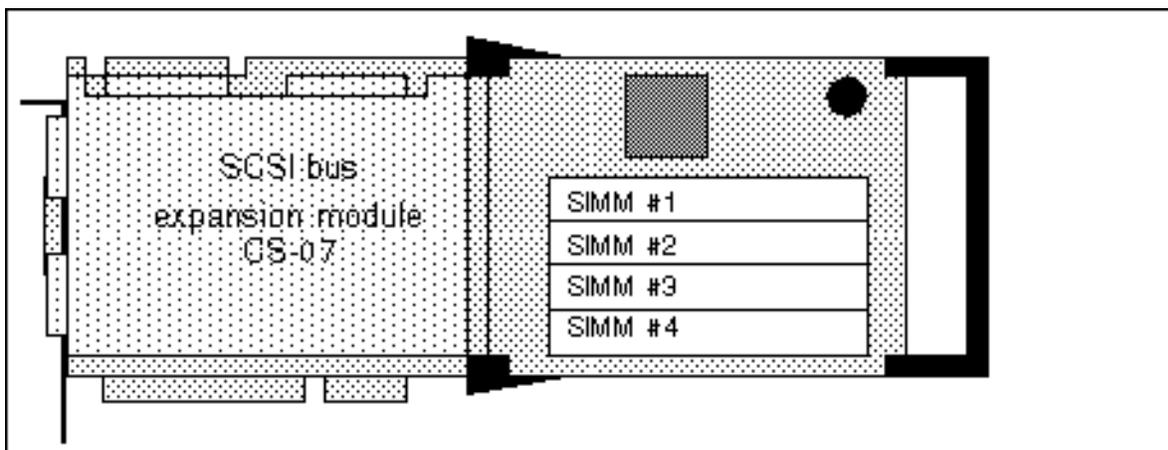The SIMM sockets on the CS37 PCI adapter reside on the RAID expansion module:

Figure 2: SIMM sockets on the RAID expansion module of the CS37 adapter

1. Install the SIMMs in the SIMM sockets. If different SIMM sizes are used, the highest-capacity module must reside on socket number 1.
2. To confirm that the SIMMs have been properly installed, the amount of cache memory present will be displayed by Storage Manager when the HBA icon is selected.
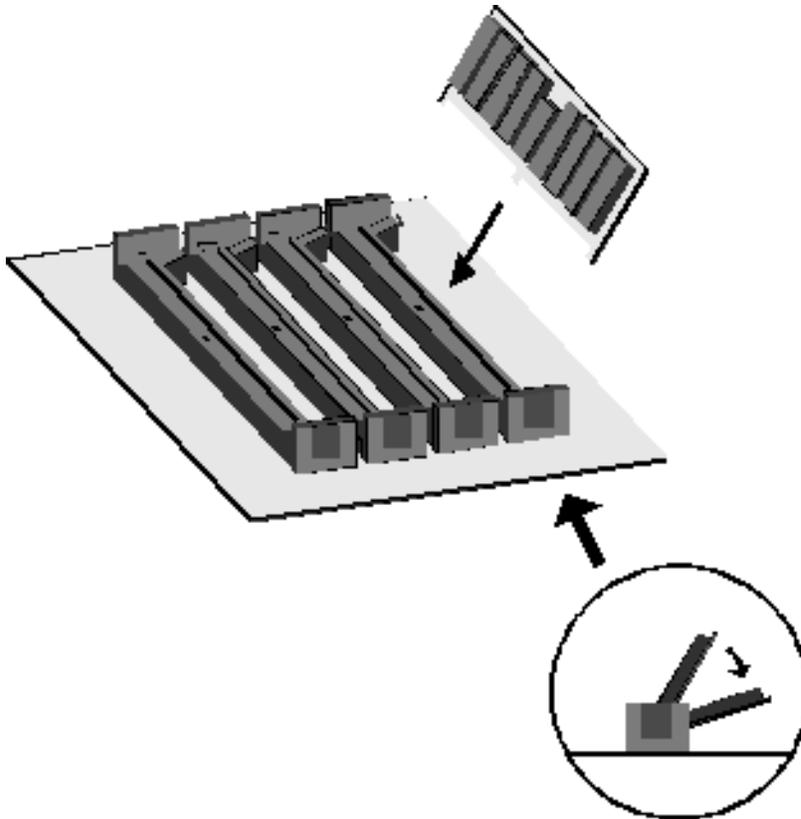
Figure 3: Installation of the SIMM modules

4.2   Installing the SCSI bus expansion module

Each RAIDmaster adapter can control up to seven (CS31) or 15   (CS34/CS35/CS37) SCSI devices on a single SCSI bus with one internal or external connector. Additionally, the adapters accept a SCSI bus expansion module. The bus expansion module adds two additional SCSI busses with internal connectors (CS37: plus 2 external connectors).

For C31, CS34 and CS35, the bus expansion module includes two adapter cables for the external SCSI connectors.

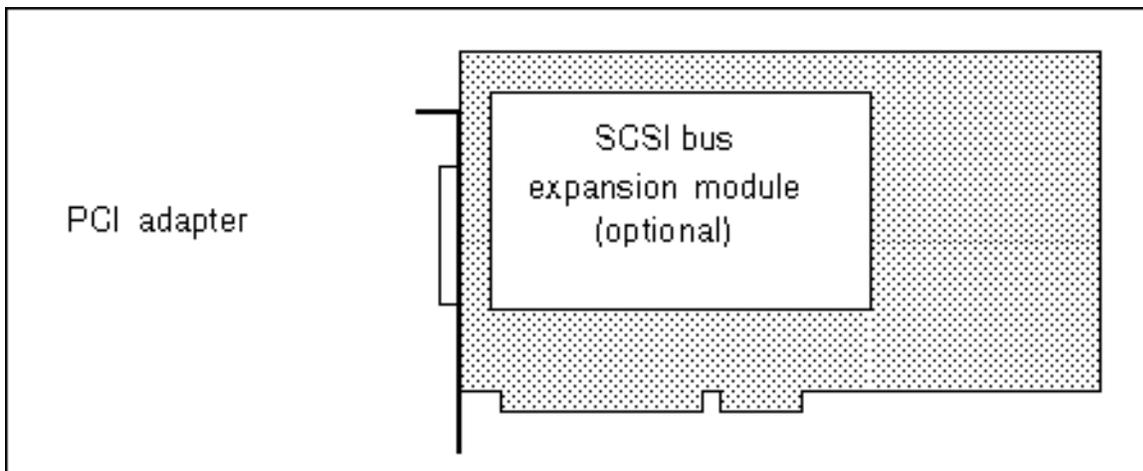In the case of CS37 controllers, only SIMM modules of the same size and type can be installed.

Figure 4: SCSI bus expansion module for CS31, CS34 and CS35 adapters

If multiple SCSI busses are required, a SCSI bus expansion module can reduce the number of RAIDmaster adapters required. Alternatively, more RAIDmaster adapters with fewer SCSI busses per adapter may be used in order to provide more I/O processing power.

1. If an expansion module is to be installed on the adapter, attach the module to the adapter. Use the nylon standoffs to secure the module.

2. If external SCSI busses are to be used with the expansion module, install the internal-to-external connector kit. Plug one end of the internal-to-external cable into the SCSI connector. The other end of the cable attaches to an external SCSI connector and mounting brackets. Attach this mounting bracket to any free card slot in the back of the PC.

 If the expansion module is properly installed, the corresponding icon will be displayed by the Storage Manager when the HBA icon is selected.

4.3   Hardware installation

Install the RAIDmaster adapter in any 32-bit PCI bus-mastering slot of the computer. Use the slot with the lowest slot number available.

Be sure to secure the adapter bracket to the host cabinet. The RAIDmaster adapter comes pre-configured to occupy SCSI-ID number 7 which should not be changed.

 Installing hard disk devices is described in chapter Storage cabinets and devices.

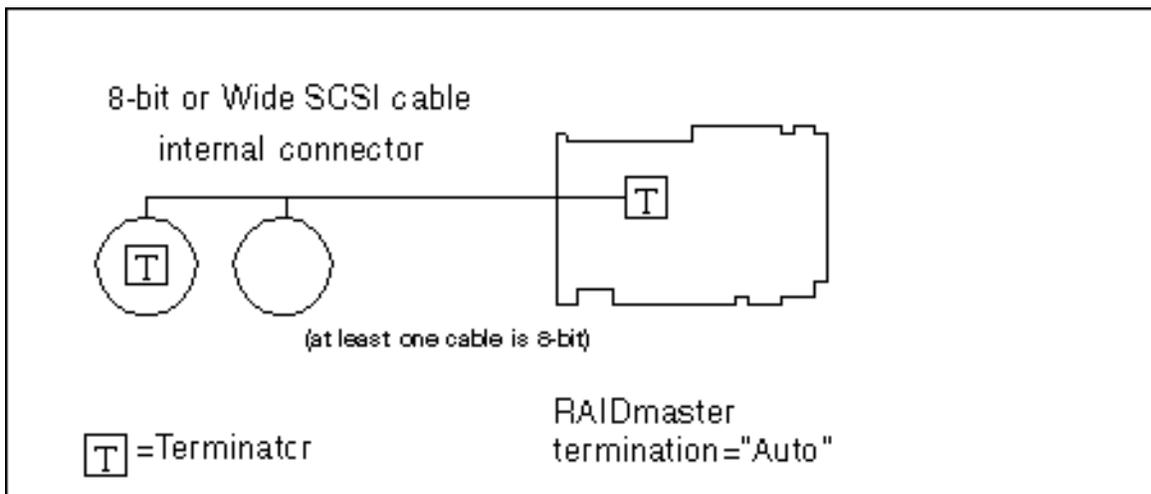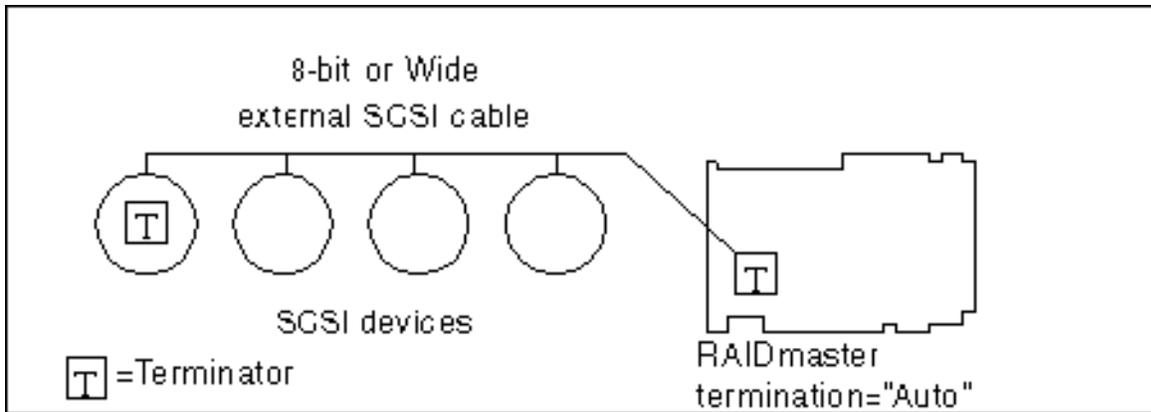## 4.4   Configuring SCSI termination

The devices at each physical end of a SCSI cable must be terminated. Depending on how you cable your system, you will either terminate two SCSI devices (such as hard drives), or the SCSI controller and one other SCSI device.

SCSI termination for RAIDmaster controllers is configured from the *Configure Host Bus Adapter* window accessible from the controller's *Information Window* in the Storage Manager. The controller has four termination settings:

Auto	The default setting. This can be used for all cabling conditions except where Wide cables are connected to both the internal and external SCSI connectors on the controller.
Disabled	Forces termination to OFF. Use this setting when Wide cables are connected to both the internal and external SCSI connectors on the controller.
Enabled	Forces termination to ON. The Auto setting is automatically enabled when you have only one cable connected to the controller.
High Only	Terminates only the signals that are present on Wide SCSI devices but not on Narrow SCSI devices. This allows Wide cables and Narrow cables to be simultaneously attached to the controller. The Auto setting automatically detects this condition and sets High Only.

The following illustrations show various SCSI cabling examples.

If only one SCSI cable is attached to the controller (to either the internal or external connector), the SCSI device at the opposite end of the cable must be terminated and controller termination set to Auto:

8-bit or Wide
external SCSI cable

SCSI devices

T = Terminator

RAIDmaster
termination="Auto"

8-bit or Wide SCSI cable
internal connector

(at least one cable is 8-bit)

RAIDmaster
termination="Auto"

T = Terminator

5 The Storage Manager

5.1 Introduction

The Storage Manager for RAIDmaster is part of the Reliant UNIX operating system and available in the SIdpt package on the system CD-ROM. The Storage Manager provides the following functions:

- Checks hardware configuration upon initial installation of the RAIDmaster adapter and after addition of new SCSI devices.
- Allows user to create, modify and delete disk arrays.
- Provides on-line functions for the SCSI storage subsystem such as event logging, diagnostics and I/O monitoring.
- The Logger allows remote notification of events via e-mail. It writes events to a file on the disk and reports events whose levels are greater than or equal to the current broadcast threshold to destinations specified by the user.

The online-help system of the Storage Manager explains how to use the Storage Manager. The most important functions of the Storage Manager are summarized in the Quickstart text, which is available in German and English.

The section Using the Storage Manager explains how to use this help system or the Quickstart text.

Another important section is Creating Reliant UNIX file systems on disk arrays which describes how to create Reliant UNIX ® file systems after configuring disk arrays.

5.2 Using the Storage Manager

The Storage Manager interface is very convenient, because semi-graphic representations are used for displaying the views, and interaction with the user takes place via forms, buttons, menus and masks.

You can start the Storage Manager with the dptmgr command.

If the message There appears to be another dptmgr loaded is displayed when the Storage Manager is started, a Storage Manager is already running or was aborted. Check whether another Storage Manager is still running. If this is not the case, you can restart the Storage Manager with the dptmgr -restart command.

For further information on the dptmgr command, refer to the Man Page.

The complete description of this user interface as well as a lot of background information about the storage system controlled by the RAIDmaster adapter is available as an on-line help system. The context-sensitive help provides you with important information in any situation. The Storage Manager also enables you to print any helptext from the on-line help into a file or on a printer of your system.

The following describes how to reach the on-line help system and the Quickstart text of the Storage Manager.

When the Storage Manager has been started and has checked the hardware configuration, the first screen to be displayed is always the Physical Configuration View. The presentation depends on your screen, i.e. if it is a monochrome or color screen. Therefore the descriptions describe the presentation on color screens and in parentheses the presentation on monochrome screens.

5.2.1 The on-line help system

If you are not familiar with the Storage Manager, you are advised to first use the on-line help system. At the top of your screen, you will see the menu bar with the menus File, Options, Help and Quickstart. One letter of each menu name is highlighted in pink (inverse). This letter is called "hotkey" and you can select a menu by typing the [Esc] key followed by the hotkey of the menu. For selecting the Help menu:

- Press the [ESC] key followed by [H].

As a result, you will see the pull-down menu of the on-line help system. If you are not familiar with the Storage Manager on Reliant UNIX ®, it is recommended that you find out more about this user interface. You can get

this information by selecting the menu item How to use it. The hotkey for this item is again [H], i.e. again:

✤ Press the [ESC] key followed by [H].

Then the Storage Manager will show you the first screen of How to use the Storage Manager. Here you get all the information necessary for scrolling helptexts and for navigating through the hyper link-based on-line help system. All other information can be found in the on-line help itself.

The table below briefly summarizes how you can control the Storage Manager via the keyboard (see section Notational conventions):

Storage Manager function	Key combination
Select pull-down menu or function button	[Esc] pink (inverse) letter
Exit pull-down menu	[Esc] [Esc]
Move to next icon	[Tab]
Move to previous icon	Shift + [Tab]
View icon information window	[Enter]
Scroll up or down	Ê or Ë
Scroll up or down one page	[PgUp] or [PgDn]
Move to top or bottom	[Home] or [End]
Select drive icon	spacebar

Table 1:

i   The TC10 console does not have an [ESC] key. It can be emulated using the key combination [CTRL] + [3].

i   Before you actually start working with the Storage Manager, you should read the README information, which is available under the menu item View README Information in the pull-down menu Help.

5.2.2   First steps with Quickstart

To learn the Storage Manager's standard functions, the Storage Manager for Reliant UNIX $^®$ provides the "Quickstart" function in addition to the help system. Under Quickstart the most important information is summarized in German and English on approximately eight print pages.

To call the Quickstart function, start the Storage Manager. The menu bar with the menus File, Options, Help and Quickstart is displayed. Here you select Quickstart by pressing [ESC] and then [Q] (the hotkey for the Quickstart menu).

In the following pull-down menu, you can select whether you want the Quickstart text to be displayed in German (menu item Schnellstart) or in English (menu item Quickstart). Use the Ê or Ë key to select the entry you want and press [Enter]. The Quickstart text is then displayed in the help window. Using the Ê and Ë keys you can scroll through the text.

You also have the option of writing the text to a file and print it. Press the [ESC] key and then [P] (hotkey for the Print function). The Print Settings window is displayed in which you can specify the name of the output file. Proceed as follows:

1. Press the [ESC] key and then [F] to select File/device.

2. Press the tabulator key to select the input field for the file name and enter an absolute path name.

3. Press the [ESC] key and then [P] to start writing the text to the file.

You can exit the Quickstart window with the key combination [ESC] [O].

i   You can also call the Quickstart text from within the on-line help system at any time using the key combination [ESC] [Q].

5.3   Adding and removing RAIDmaster adapters or SCSI devices

Adding or removing RAIDmaster $^{®}$ adapters and/or SCSI disks changes the hardware configuration. Two situations must be differentiated:

－   Existing disk arrays are not affected and no new arrays are to be configured: in this case, you only need to reboot the system. (Added hard disks can be used as SCSI disks, removed disks are no longer available.)

－   If disk arrays are affected by the modification or if new disk arrays are to be created, you must call the Storage Manager with the dptmgr command after rebooting the system in order to make these modifications.

Bear in mind that - as with every start of the Storage Manager - the modifications you make only come into effect when you reboot the system (init 0). The system must not be rebooted, however, before the build is completed, i.e. when the status flag B on the icons has disappeared. (For a detailed description of the status flags see the on-line help.)

5.4   Creating Reliant UNIX file systems on disk arrays

When you exit the Storage Manager and reboot your system, the configurations you have made by means of the Storage Manager become effective. Disk arrays now appear to the operating system as one hard disk.

First enter the autoconf -l command to find out which devices are available to the operating system.

If, for example, you have installed a RAIDmaster adapter with a SCSI bus on which a RAID-1 and a RAID-5 disk array are configured, this would be shown with the following entries:

```
...
ios1/scon60 DPT PM3224, FW07CH, ctrl=(0,0), 64MB, slot=PCI#2, logid=0
ios1/sport60 Host Adapter Port
ios1/sdisk600 DPT_RAID-5 -- partition not set
ios1/sdisk603 DPT_RAID-1 -- partition not set
...
```

As the remark partition not set is displayed, you must first initialize the array (which is considered as a disk by the operating system). For the RAID-5 array in the above example you can do this with the following command:

dksetup -U /dev/ios/sdisk600

If you then call autoconf -l, you will obtain the entry

ios1/sdisk600 DPT_RAID-5 -- DPT_RAID-5

Now you can create the partitions:

－   You can first call up the default settings of the partitions with the following command

dkpart -I $device\_name$s0

(in the above example: dkpart -I /dev/ios1/rsdisk600s0)

－   Then you can modify these partitions for the file systems. Bear in mind the maximum permissible file system sizes, e.g. 4GB for a ufs file system. To modify the partitions, enter the

dkpart -e $device\_name$s0

command and edit the entries for the partitions (see also "System Administrator's Reference Manual").

Now you must decide whether you want to use a disk array as a raw device, for example for a database system, or whether you want to create Reliant UNIX $^{®}$ file systems on this array.

If you want to use the disk array as a raw device, no further actions are necessary.

To create Reliant UNIX $^{®}$ file systems, proceed as follows:

－   Now create a file system with the newfs command in each partition you have configured.

If you configured, say, eight partitions in the RAID-5 array in the above example and want to create a ufs file system in partition 5, you would have to enter the following command

newfs -q /dev/ios1/sdisk600s5

5.5   Dual Level Arrays (as of CS37)

If you start the Storage Manager with the option *dptmgr -dual*, you can create dual level arrays, i.e. RAID-0, using RAID-1 or RAID-5 arrays in the Logical Configuration View of the Storage Manager. These combinations are often referred to as RAID-10, RAID-01 or RAID-50.

This option can only be used for creating a new RAID array; you cannot combine existing RAID-1 or RAID-5 arrays.

RAID10   After selecting RAID-1, the *dptmgr* allows you to select 2,4,6,8 .. drives. Two drives are used to built each RAID-1 array. If more than two drives are selected, a RAID-0 array is subsequently created using these RAID-1 arrays.

RAID50   To build a RAID-50 array, you must select RAID-5 and at least six drives. RAID-5 arrays are created from at least three drives, depending on the number of parity drives you specify. If you specify only one parity drive, one RAID-5 array is built using all the selected drives (as before). If you specify more than one parity drive, the number you specify determines the number of RAID-5 arrays used to built the RAID-0 array. If, for example, you specify nine drives, a RAID-0 array is built using three RAID-5 arrays, and so on.

5.6   Cache settings

To ensure data security even in the event of a power failure, the write cache mode on the RAIDmaster controller is set by default to write-through for all disks and logical drives. If you have an uninterruptible power supply (UPS), you can change the cache mode to write-back and profit from enhanced performance.

Any change made to the cache mode is made at your own risk. You should not switch to write-back mode if you do not have a UPS.

5.6.1   Modifying cache settings on the CS31, CS34, CS35

Before an array or hot spare is set up, you must use the Storage Manager (in the Device Configuration window) to set the cache mode to the same value for each physical drive to be integrated. This setting is then used for the array if the array is built immediately after the cache mode is set.

If you want to change the cache setting for existing arrays, you must use the tool dpttbl (see section Modifying the cache setting with the tool dpttbl) and then start and stop the Storage Manager once to activate the new configuration.

If the new cache setting is also to be activated for individual SCSI disks, you must do the following once you have completed the above-mentioned setting: set *dpt_force_cache* in */etc/conf/pack.d/dpt/space.c* to 0, build a new kernel and restart the system.

When using hot spare drives, it is highly recommended that you set all the arrays and hot spares (and their disks) on a RAIDmaster controller to the same value to avoid conflicts.

5.6.2   Modifying the cache setting on the CS37

The CS37 includes new cache algorithms in the firmware which makes it easier to modify the cache setting.

If you want to modify the cache setting for arrays and hot spares, you must specify the desired value using the Storage Manager (in the Configuration window) after the cache has been set up. The new value is effective

immediately.

If you want to change the cache setting for individual SCSI disks, you must set the value using the Storage Manager (in the Device Configuration window) and then set *dpti_force_cache* in */etc/conf/pack.d/dpti/space.c* to 0, build a new kernel and restart the system.

5.7 Modifying the cache setting with the tool dpttbl

5.7.1 Cache setting for CS31, CS34, CS35

The cache setting is stored in the reserved block (last block on the disk). The cache setting for an array is determined by the cache settings for the individual disks. Once an array has been built, you can no longer access the individual disks from dptmgr, and you can therefore no longer change the cache setting of an array.

If you want to change the cache configuration of an existing array, you have to use the tool dpttbl. dpttbl directly manipulates the reserved blocks of all the disks in the array. Once you have modified the cache configuration using dpttbl, you must start and stop the dptmgr once to activate the new setting. *dpttbl* displays a list of all the arrays and all the individual disks not in an array and asks which of the disks listed you want to modify.

Since the tool is very powerful, preconfigured options are available for displaying or modifying certain cache settings:

Options:	Meaning
-wl	Display the write cache setting
-wt	Set the write cache to write-through
-wb	Set the write cache to write-back
-rce	Activate read caching
-rcd	Deactivate read caching
-rcl	Display read caching
-wce	Activate write caching
-wcd	Deactivate write caching
-wcl	Display write caching
-p	Not only arrays but also individual disks are listed.

5.7.2 Cache setting for CS37

The CS37 includes new cache algorithms in the firmware which makes it easier to set the cache mode. The cache setting is specified from the *dptmgr* once an array has been built. The cache settings for arrays are no longer stored in the reserved block but in the RAID tables instead. Only the cache mode of a single SCSI disk is stored in the reserved block. You cannot use *dpttbl* to change the cache setting for an array or disk. It is also not possible to change the setting for read caching (always disabled for the CS37) or write caching.

6 Theory of operation

6.1 Disk caching

RAIDmaster adapters may be configured with 4 to 64 MB (CS37: 16 to 128 MB) of hardware cache. Adding more cache RAM to the RAIDmaster adapter will typically result in better RAID performance - especially in RAID-5 environments.

6.1.1 Software and hardware caching

The RAIDmaster cache is specifically designed to enhance performance of Reliant UNIX. Although software caches are almost always beneficial, in larger, more heavily loaded systems, hardware caching offers a significant advantage when combined with the software cache in the operating system.

The primary purpose of the operating systems's software cache is to eliminate disk reads. If the data requested by the application program can be serviced from the software cache, then a disk read will be eliminated. Software caches also improve disk write response time by immediately accepting data to be written from the application programs but postponing writes to disk, holding the data in cache until a time when the disk is idle. Software write caching provides performance benefits for lightly loaded systems with small numbers of users but experiences write performance degradation in larger multi-user environments with heavier loading. This is because frequent periods of system idle time are required to flush the software cache buffers to disk.
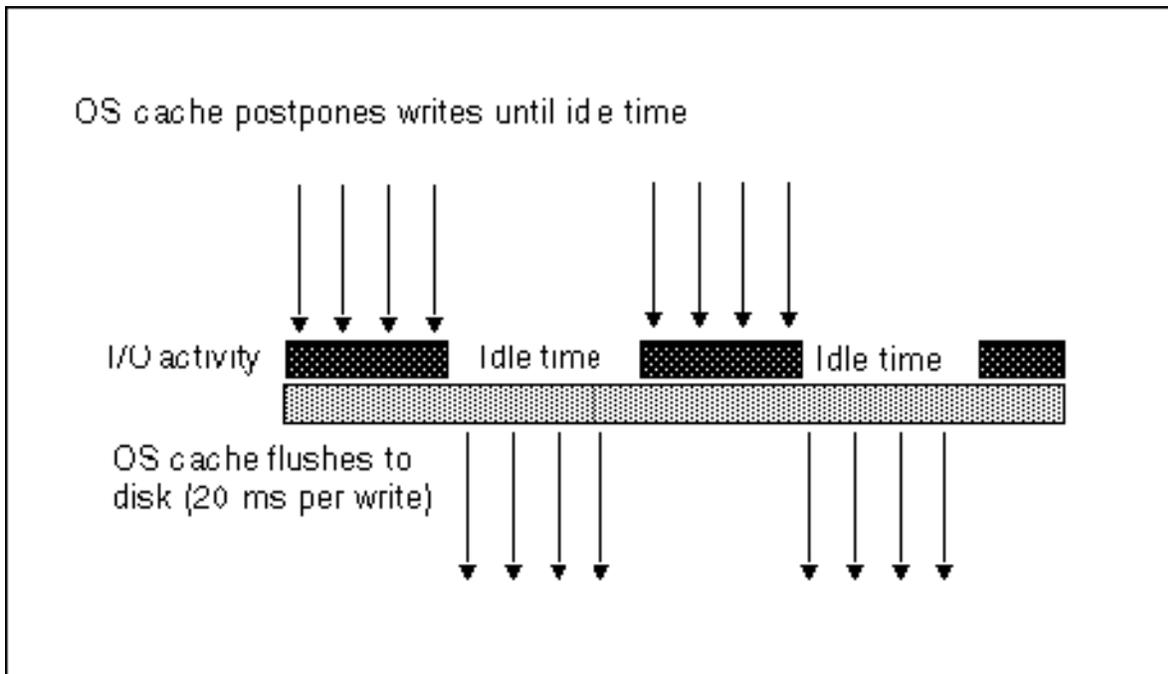

Figure 5: Caching postpones writes until idle time

Lightly loaded systems with a small number of users tend to have lots of idle time, but as more users are added, periods of system idle time become less frequent and shorter. This causes the software cache to fill up with dirty sectors waiting to be written to disk. These sectors take up space in the software cache which could be used for read caching - something that the software cache is much better at. Software cache flushing begins to interfere with other system activity as more users are added and the system has less idle time.
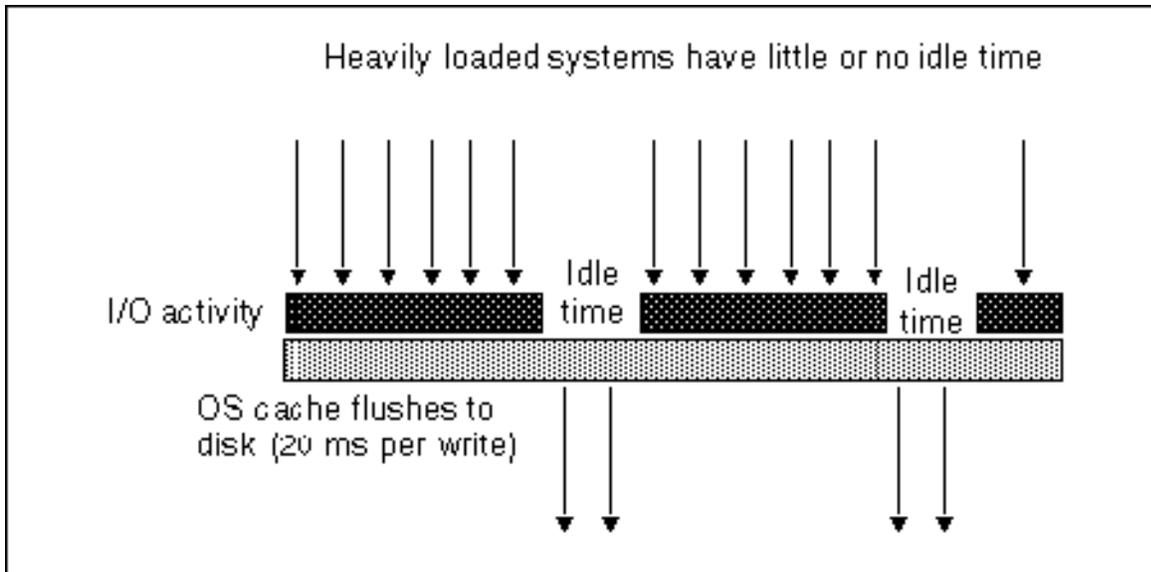
Figure 6: In heavily loaded systems, software cache flushing interferes with system activity

With RAIDmaster's hardware cache, the operating system's cache buffers flush much more quickly. The adapter cache receives the flushed data in a fraction of the time it would take without the cache. The RAIDmaster adapter then copies the data back to disk without interfering with other system activity. As an example, consider an operating system configured with one megabyte of software cache which contains only dirty data. With a 1-Kbyte record size and an average disk access time of 20ms, flushing the software cache would take 20 seconds. However, with the RAIDmaster adapter cache, the entire flush operation from the software cache would take approximately 1 second instead of 20 seconds. RAIDmaster would then flush its cache to disk using its separate I/O bus, concurrently with other system activity.
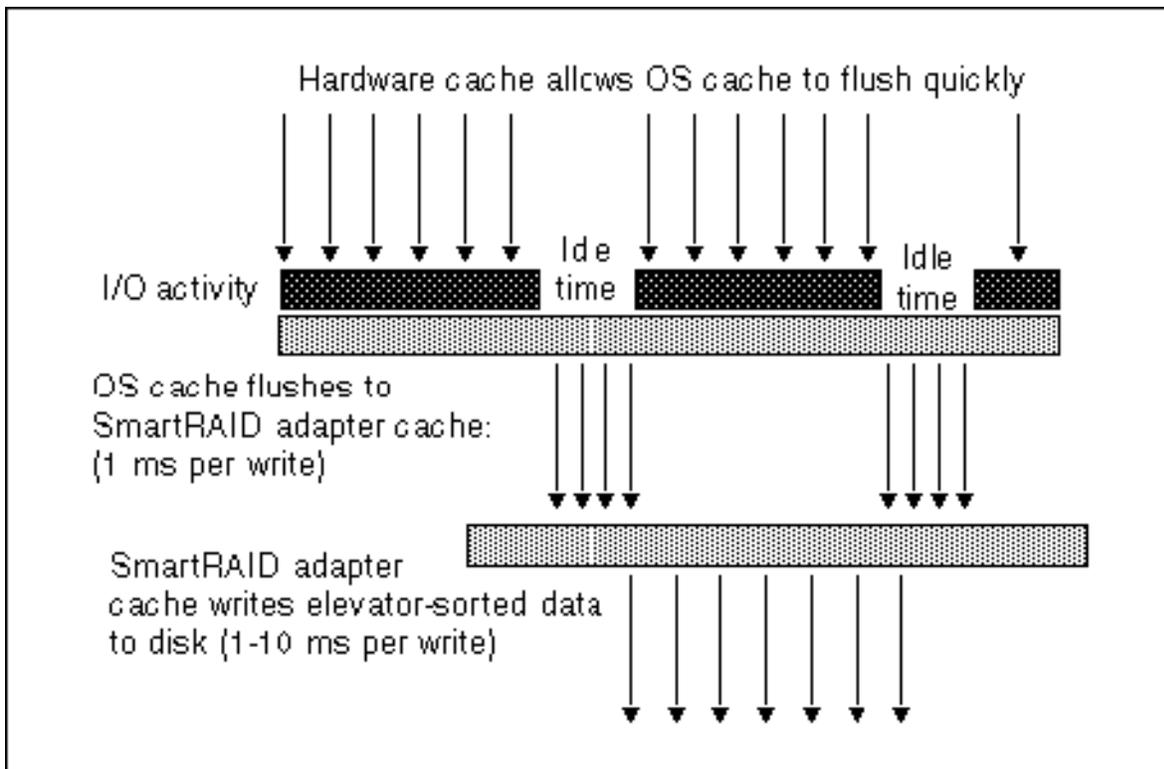

Figure 7: Hardware cache allows OS cache to flush quickly

6.1.2 Elevator sorting

Flush operations from the RAIDmaster hardware cache occur much faster than from the OS software cache, since all data in the adapter cache is kept in elevator-sorted order. Data is written to disk in order of increasing cylinder, head, and sector number, minimizing physical disk searches and rotational latency. This feature, called "elevator-sorted write-back", eliminates much of the head thrashing commonly associated with disk-intensive operating systems like UNIX. By elevator-sorting the sectors in cache, the average write access time to the drive may be reduced from the normal 20ms (seek plus rotation) time, to between 1 and 10ms.
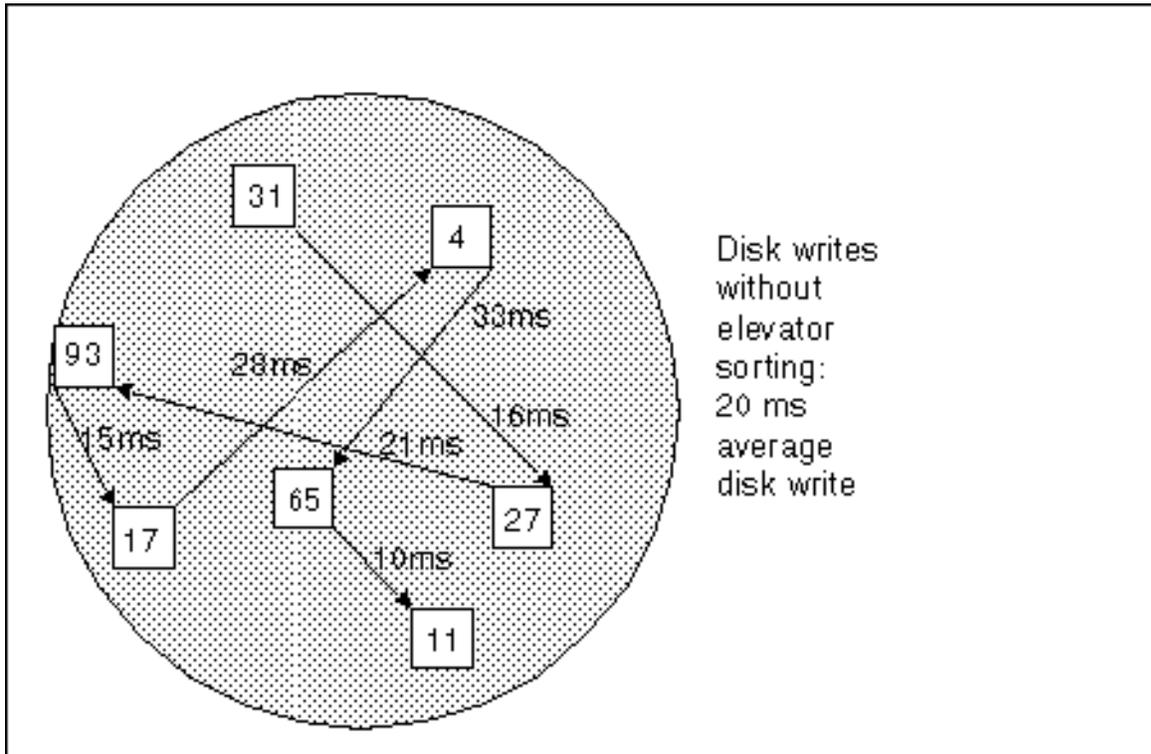

Figure 8: Disk writes without elevator sorting

A good analogy is an elevator in a large, multi-story building. Consider how much slower the elevator would be if it went to the requested floors in the same order the buttons were pressed, instead of in order of increasing floor number.

One additional benefit of elevator sorting is that multiple short disk write operations can often be combined within the adapter cache into one large write operation, thereby significantly reducing SCSI overhead.
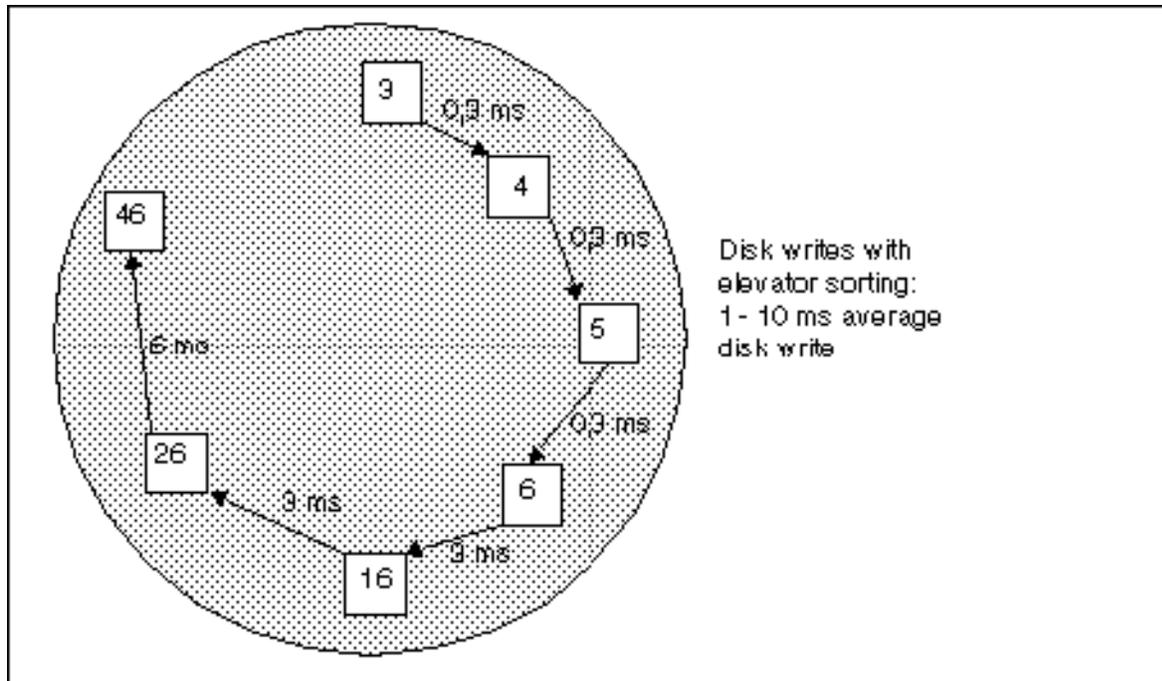
Figure 9: Disk writes with elevator sorting

6.1.3 Write-back and write-through algorithm

Write-back is a method of buffering data to be written to a slow device, such as a disk drive, by temporarily saving the data in a cache. The data can then be written at a later time when the device would otherwise be idle. In addition, the data can be processed through techniques such as elevator sorting, to increase write throughput. This type of caching generally provides better I/O performance.

In write-through mode, all data is written to disk for each write command before command complete status is returned to the computer (even though data may also be cached for later use). In write-back mode, writes to disk are deferred until after command completion.

Whenever RAIDmaster has not received an I/O request for 250ms it will flush dirty pages to disk. Since the data is written to disk in elevator-sorted order, several seconds of idle time will normally result in all dirty pages being copied to disk, even though it may have taken much longer for the computer to write the data to the adapter.

RAIDmaster also writes bursts of ordered sectors back to the disk whenever 80% or more of the cache becomes dirty. This flush operation will continue until the 80% limit is again reached, and the execution of further write commands from the host computer will be delayed. The maximum percentage dirty limit of 80% may be changed via the Storage Manager. Another suitable setting for this parameter is 50%.

6.1.4 Read-ahead algorithm (CS31, CS34, CS35)

Another way that average disk access time can be reduced is through "disk read-ahead caching". Since disk data is often grouped in clumps of contiguous sectors on the disk, it is usually beneficial to continue to read sequential sectors into cache following a sector that has been requested by the computer. This is sometimes referred to as the "locality of reference" principle. RAIDmaster's algorithms analyze the pattern of disk access, looking for cases where sequential I/O may be occurring. If it is determined that a read command from the computer is part of a pattern of sequential reads, the adapter will bring additional sequential sectors into cache so that future read commands will result in cache hits.
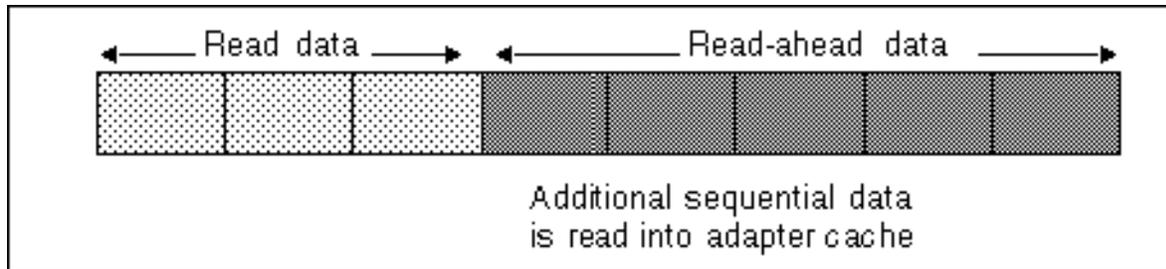
Figure 10: Automatic read-ahead

For disk read-ahead to be effective in a multi-user environment, it is necessary to cache a large number of read-ahead sectors from many different areas of the disk. This is because multi-user and networking operating systems contain highly fragmented files and I/O requests from many different users or tasks. Data which has been cached by a read-ahead operation may not be accessed until many commands in the future when the same user or task has another chance to access the disk. In addition, patterns of sequential I/O are harder to detect in large multi-user systems since they are typically interleaved with I/O from many other users. If operating system cache exists, disabling the read cache can enhance performance.
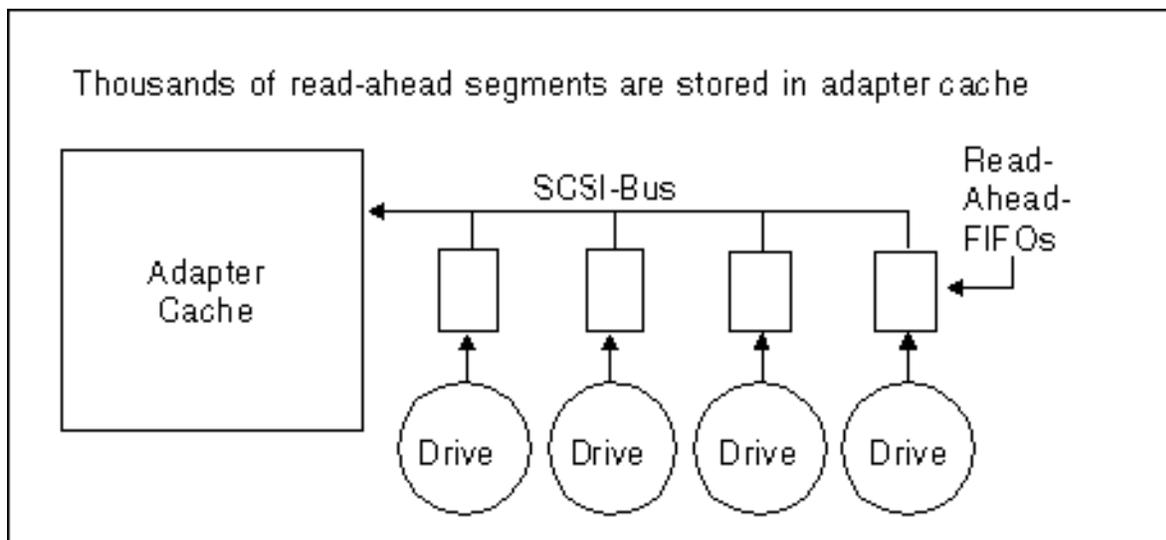

Figure 11: Caching read-ahead segments

This is why the SCSI drive's read-ahead FIFO, which can efficiently store only a limited number of data segments, may provide acceptable performance in small systems, but must be augmented by the much larger cache on the adapter in larger multi-user environments.

By default, RAIDmaster limits the amount of cache which can hold read-ahead data to 30%. This limit can be modified via the Storage Manager.

6.1.5 Predictive caching algorithm (as of CS37)

CS37's predictive caching dramatically increases disk read performance in computer systems with highly fragmented file structures. CS37 includes new pattern recognition algorithms, capable of predicting disk read requests from the host computer by recognizing patterns within seemingly random disk accesses. CS37's pattern recognition algorithms not only look for hidden sequential strings, but also analyze historical read/write patterns to determine when the host is re-accessing files that have been previously paged out of host cache.

Virtual cache

CS37 controllers use virtual cache to track the patterns of data access for the most recently accessed disk data without actually caching that data. The collected information can then be used for more intelligent predictive algorithms. Additionally, the use of virtual cache enables the controller to track how many hits would

have resulted from installing more controller or host cache. Virtual cache does not actually use RAM.

Read ahead

Since disk data is often grouped in clumps of contiguous sectors on the disk, it is sometimes beneficial to continue to read sequential sectors into cache following a sector that has been requested by the computer. The CS37 controller's algorithms analyze the pattern of disk access, looking for cases where sequential I/O may be occurring. If it is determined that a read command from the computer is part of a pattern of sequential reads, the controller will bring additional sequential sectors into cache so that future read commands can result in cache hits.

Pre-fetch

CS37's pre-fetch algorithm analyzes disk access patterns to detect if the host is reading data that it previously read. By using records stored in virtual cache, the caching algorithm determines the boundaries of the contiguous block of previously read data surrounding the current host read request, then reads this block into cache.

Predictive caching in a multi-user environment

In order for disk read-ahead to be effective in a multi-user environment, it is necessary to cache a large number of different areas of the disk. This is because multi-user and networking operating systems contain highly fragmented files and I/O requests from many different users or tasks. Data that has been cached by a read-ahead operation may not be accessed until many commands in the future when the same user or task has had another chance to access the disk. In addition, patterns of sequential I/O are harder to detect in large multi-user systems since they are typically interleaved with I/O from many other users. This is why a drive's read-ahead FIFO, which can efficiently store only a limited number of data segments, may provide acceptable performance in small systems, but must be augmented by the much larger cache on the CS37 controller in systems with more users.

6.1.6 Caching host reads

It is not desirable to use the RAIDmaster controller's hardware cache to duplicate the caching of the operating system. CS37 controllers with caching capability assume the presence of an operating system cache and cache only predictive (see section Predictive caching in a multi-user environment) and dirty data. Read cache misses are transferred directly from disk to host, bypassing the cache.

Older controllers (CS31, CS34, CS35) assumed that there was no operating system cache and cached all data read by the host.

6.1.7 Optimizing OS cache size

When configuring a server, it is important to take into account the effect of the operating system's cache on the disk subsystem. Although many disk read operations may be eliminated by the OS cache, all writes must get through to the disk. In effect, the OS cache acts as a read filter for the disk subsystem.

As the OS cache size is increased, more read hits are serviced from the OS cache resulting in fewer reads issued to the disk subsystem. However, the number of disk writes stays relatively constant. At some point, adding more cache to the OS will not result in significantly reduced disk I/O. At this point it is better to add more cache to the RAIDmaster to improve write throughput.

6.1.8 Optimizing Controller Cache Size

The cache is used primarily for predictive reads and elevator sorted write-backs. Increasing the size of this cache makes these algorithms more efficient. More cache RAM allows the controller to hold and elevator-sort more records, resulting in closer spacing between consecutive write-back segments. This results in lower average access time for disk write operations and thus higher disk throughput. The cache may be expanded incrementally as needed to maintain optimum system performance. More cache is typically required for systems with may active users than for single-user systems. The same holds true for systems with large data files as opposed to small files. If a system has a large number of users who all access the same data, less cache may be needed.

6.1.9   Tunable cache parameters (CS31, CS34, CS35)

The read and/or write caching functions can be disabled. In addition, there are various other caching parameters which can be changed. Both can be achieved via the Storage Manager.

Normally, RAIDmaster limits the largest read or write request that will be cached to 16 blocks (8 Kbytes) or 32 blocks (16 Kbytes, for CS37) respectively. This prevents long I/O operations such as file copies from displacing large amounts of data from cache. If the host issues a read or write request larger than 16 blocks or 32 blocks, respectively, in length, the RAIDmaster adapter will not cache that data. Enabling long I/O caching allows the adapter to cache all read and write data.

The maximum percentage of cache which is allowed to hold dirty data and the maximum percentage which can hold read-ahead data are limited to 80% and 30% respectively. These limits are adjustable via the Storage Manager.

6.2   P^3: Pipelined parallel processing (as of CS37)

Traditional RAID controller designs execute host command sequentially. Each command initiates a sequence of hardware activities that are processed one after the other until the command is complete. If any of these activities is delayed, hardware components must wait, compromising performance. DPT's P3 technology addresses this problem in five important areas:

- Real-Time Control System - P^3 technology processes a virtually unlimited number of host commands in parallel. CS37's Real-Time Control System (RTCS) breaks each command down into sub-tasks that are then routed to the hardware processors on the controller: the DMS processor, XOR processor, and SCSI and Fibre Channel processors.

- Hardware Request Schedulers (HRS) - Each hardware processor has an associated HRS that coordinates its queued requests, managing the queue so that all sub-tasks are processed efficiently. The disk HRS manages one request queue for each disk drive attached to the controller. The disk HRS maintains sorted disk queues, coalescing requests when possible, in order to achieve maximum throughput from each drive. This results in an increased number of read and write operations per spindle.

- Hardware RAID Processor - DTP's P^3 architecture is based on a new DPT-exclusive chip. This 0.4-micron, 120,000 gate standard-cell ASIC provides memory-to-memory and disk-to-memory data transfers that are fully concurrent with controller CPU execution. Depending on whether the PCI bus is 32-bit or 64-bit, data transfers occur at full 132 or 264 MB/sec data rates, respectively. Independent DMA or XOR processors execute scripts-based programs, enabling DMA and XOR tasks to be chained, and executed without controller CPU intervention. The hardware RAID processor also generates 128-bit parallel ECC for the controller cache memory. This powerful ECC is capable of correcting an error of up to 32 bits per 512-byte block of data.

- Intelligent caching - CS37 controllers have improved caching capabilities. Since most current operating systems are heavily cached, in default mode, the controller does not cache data read by the host. Instead, the controller cache is primarily used for predictive and write-back caching, and provides write data coalescing for efficient RAID-5 operation. For those environments where read caching on the controller will improve performance (particularly those environments without software disk caching), this option my be selected from Storage Manager.

  CS37 controllers also have improved predictive caching algorithms, a description of which may be found in the section Predictive caching algorithm (as of CS37).

- Virtual Cache - In addition to normal caching functions, CS37 controllers also maintain a large virtual cache used for pattern recognition and prediction. The total amount of cache space supported by a CS37 controller is fixed (128 MB for CS37). Some of this total cache space (max. 64 MB) is used when physical memory modules are installed on the controller to be used for standard cache. The space not used for standard cache is available for virtual caching. CS37 controllers support a total cache space of 128 MB with up to 64 MB of standard cache (depending on the number of memory modules installed). As an example, if 32 MB of cache memory were installed on a CS37 controller, then 128 MB - 32 MB = 96 MB of virtual cache would be available.

In addition to facilitating predictive caching, virtual cache also enables the controller to provide hit rate predictions for larger cache configurations without requiring the additional cache to be installed on the controller.

6.3   RAID

In 1987, Patterson, Gibson and Katz at the University of California Berkeley, published a paper entitled "A Case for Redundant Arrays of Inexpensive Disks (RAID)". This paper described various types of disk array, referred to by the acronym "RAID". The basic idea of RAID was to combine multiple small, inexpensive disk drives into an array of disk drives which yields performance exceeding that of a single large expensive drive (SLED). This array of drives appears to the computer as a single logical storage unit or drive.

The mean time between failure (MTBF) of the array will be equal to the MTBF of an individual drive, divided by the number of drives in the array. Because of this, the MTBF of an array of drives would be too low for many applications. However, disk arrays can be made significantly more fault-tolerant by redundantly storing information in various ways.

Five types of array architecture, RAID-1 through RAID-5, were defined by the Berkeley paper, each providing disk fault-tolerance and each offering different trade-offs in features and performance. In addition to these five redundant array architectures, it has become common to refer to a non-redundant array of disk drives as a RAID-0 array.

6.3.1   Disk striping

Fundamental to RAID is striping, a method of concatenating multiple drives into one logical storage unit (LSU). Striping involves partitioning each drive's storage space into stripes which may be as small as one sector (512 bytes) or as large as several megabytes. These stripes are then interleaved round-robin, so that the combined space is composed alternately of stripes from each drive. In effect, the storage space of the drives is shuffled like a deck of cards. The type of operating environment determines whether large or small stripes should be used.
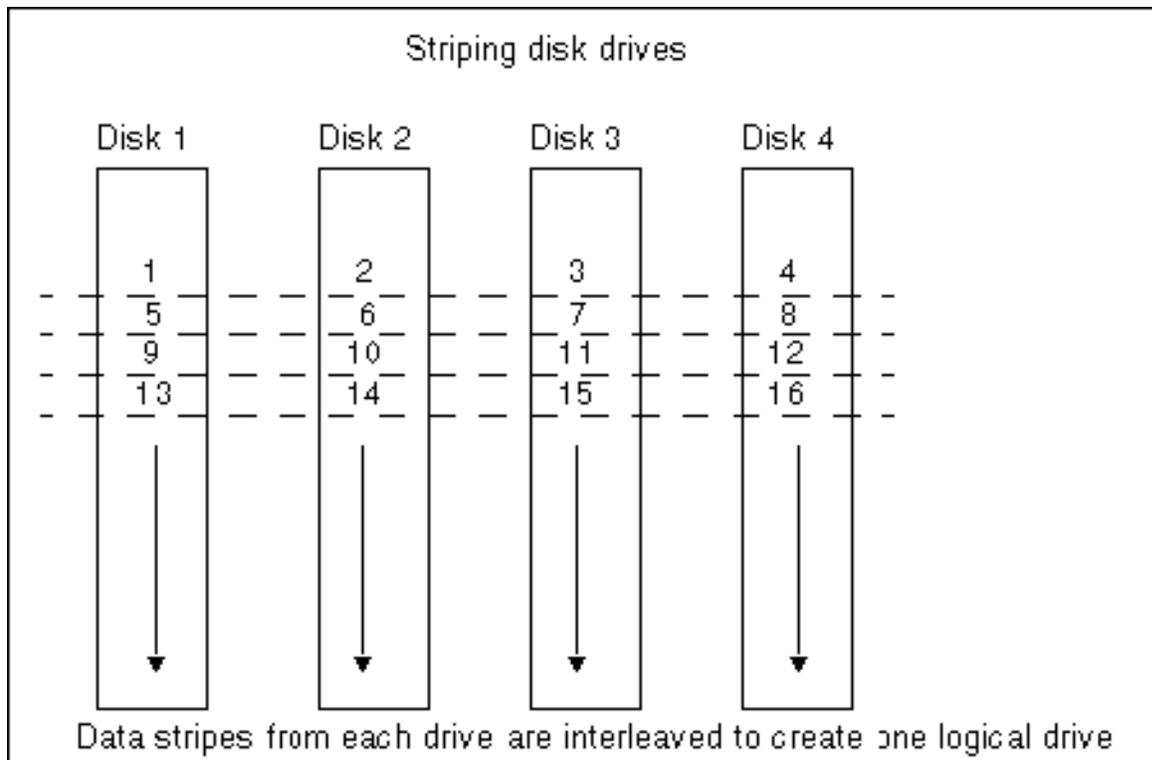

Figure 12: Disk striping

Most multi-user operating systems today, such as Reliant UNIX, support overlapped disk I/O operations across multiple drives. However, in order to maximize throughput for the disk subsystem, the I/O load must be balanced across all the drives so that each drive can be kept as busy as possible. In a multiple drive system without striping, the disk I/O load is never perfectly balanced. Some drives will contain data files which are frequently accessed and some drives will only rarely be accessed. By striping the drives in the array with stripes large enough for each record to fall entirely within one stripe, the records will be evenly distributed across all drives and the I/O load will be balanced. All drives in the array will thus be kept busy during heavy load situations. This allows each drive to work on a different I/O operation, and thus maximize the number of simultaneous I/O operations which can be performed by the array.

6.3.2   Definition of RAID levels

RAID-0 is typically defined as a non-redundant group of striped disk drives without parity. RAID-0 arrays are usually configured with large stripes, but may be sector-striped with synchronized spindle drives for single-user environments which access long sequential records. If one drive in a RAID-0 array crashes, the entire array crashes. However, RAID-0 arrays deliver the best performance and data storage efficiency of any array type.
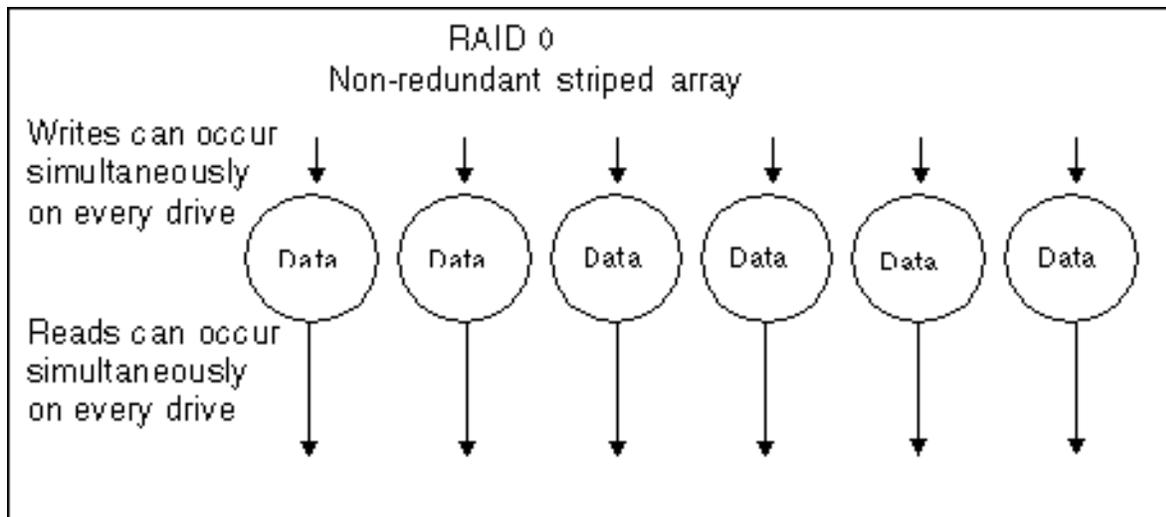

Figure 13: RAID-0 array

RAID-1, better known as "disk mirroring", is simply a pair of disk drives which store duplicate data, but appears to the computer as a single drive. Writes must go to both drives in a mirrored pair so that the information on the drives is kept identical. Each individual drive, however, can perform simultaneous read operations. Mirroring thus doubles the read performance of an individual drive and leaves the write performance unchanged. RAID-1 delivers the best performance of any redundant array type in multi-user environments.
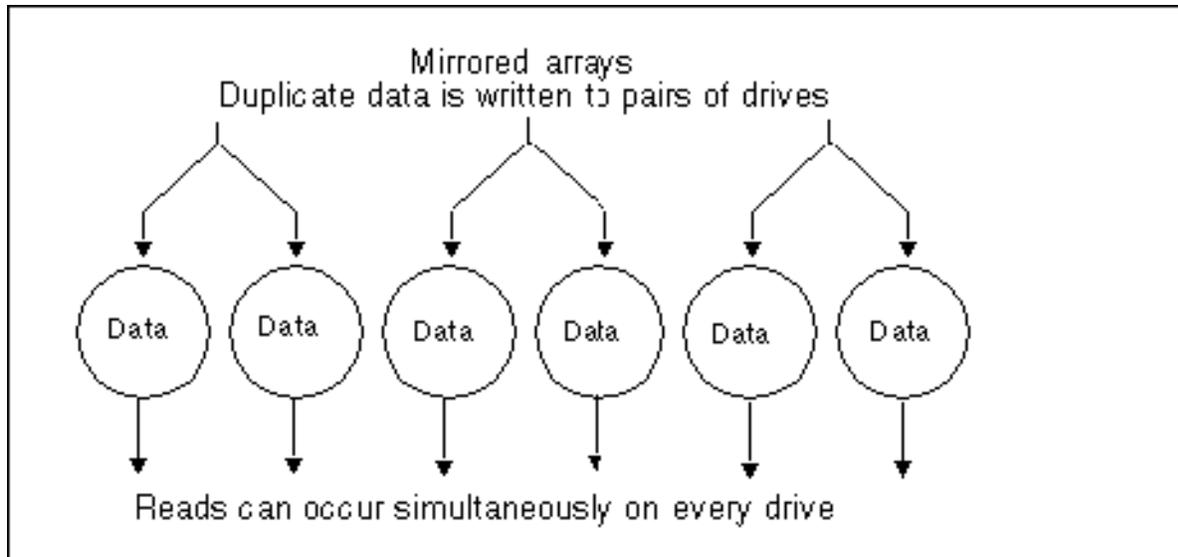
Figure 14: RAID-1 array

RAID-2 arrays sector-stripe data across groups of drives, with some drives relegated to storing ECC information. Since most disk drives today embed ECC information within each sector, RAID-2 offers no significant advantages over RAID-3 architecture and is not supported by RAIDmaster.
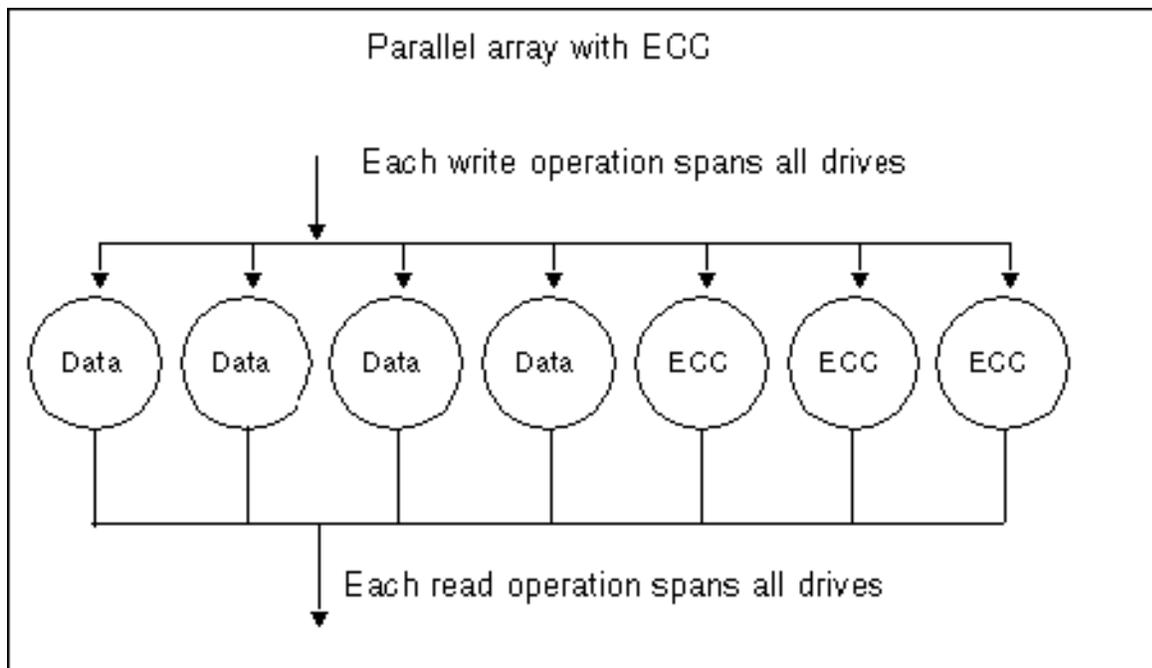

Figure 15: RAID-2 array

RAID-3, as with RAID-2, sector-stripes data across groups of drives, but one drive in the group is dedicated to storing parity information. RAID-3 relies on the embedded ECC in each sector for error detection. In the case of a hard drive failure, data recovery is accomplished by calculating the exclusive OR (XOR) of the information recorded on the remaining drives. Records typically span all drives, thereby optimizing the disk transfer rate. Since each I/O accesses all drives in the array, RAID-3 arrays cannot overlap I/O and thus deliver best performance in single-user, single-tasking environments with long records. Synchronized-spindle drives are required for RAID-3 arrays in order to avoid performance degradation with short records. Since RAID-5 arrays with small stripe sizes can yield similar performance to RAID-3 arrays, RAID-3 is not supported by
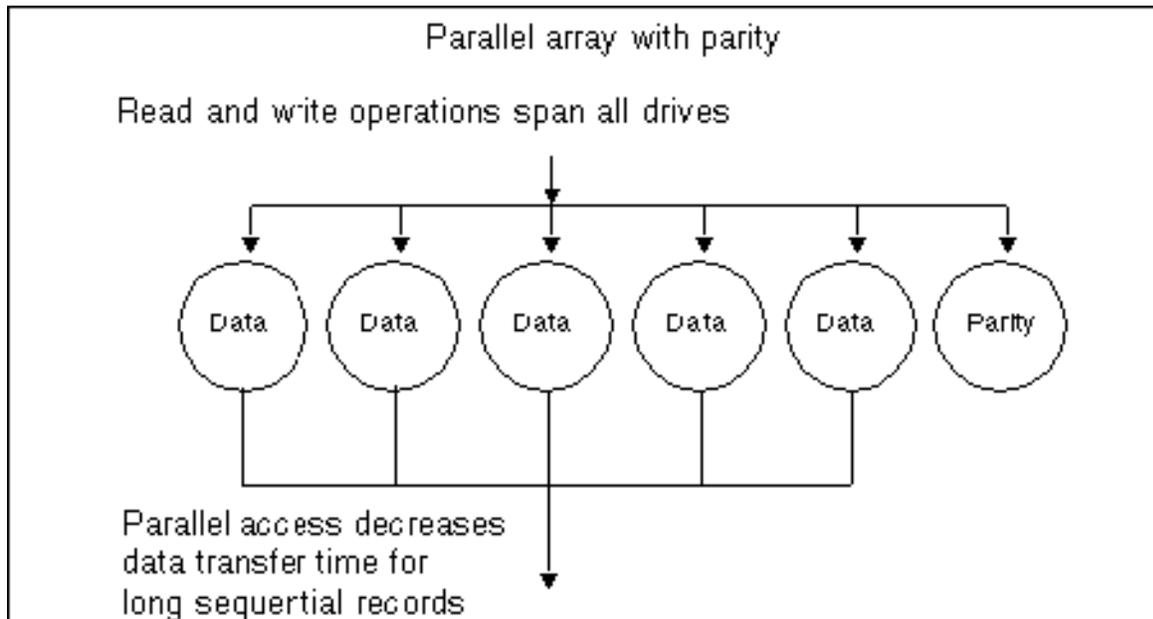
RAIDmaster.

Figure 16: RAID-3 array

RAID-4 is identical to RAID-3 except that large stripes are used so that records can be read from any individual drive in the array (except the parity drive), allowing read operations to be overlapped. However, since all write operations must update the parity drive, they cannot be overlapped. This architecture offers no significant advantages over RAID-5 and is therefore not supported by RAIDmaster.
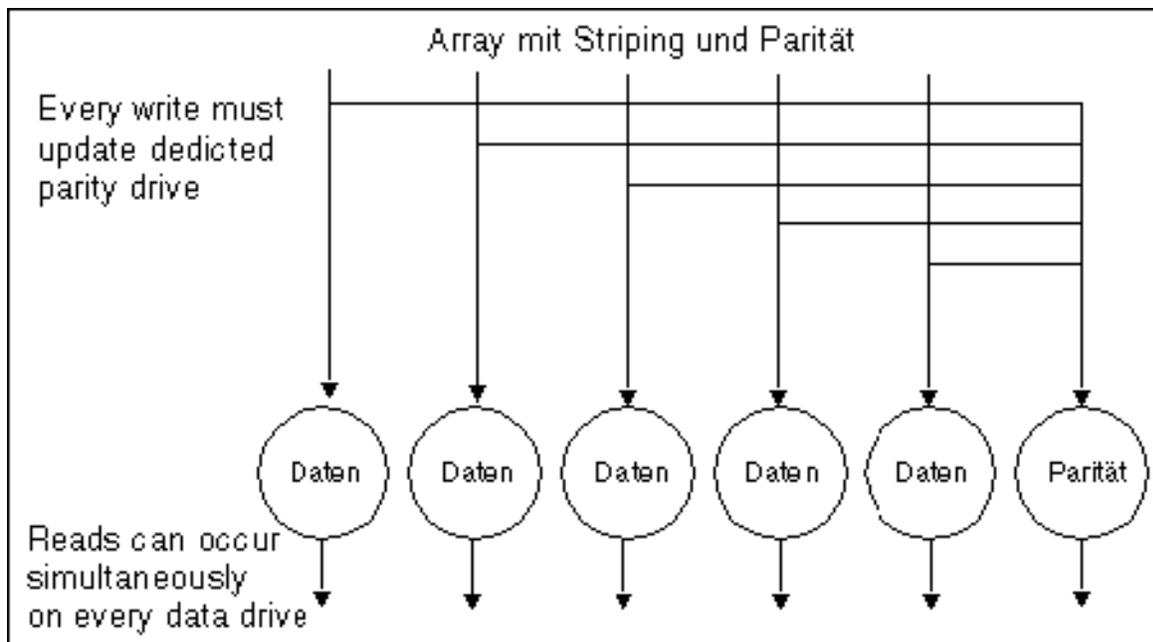
Figure 17: RAID-4 array

RAID-5, sometimes called a "rotating parity array", avoids the write bottleneck caused by the single dedicated parity drive of RAID-4. Like RAID-4, large stripes are used so that multiple I/O operations can be overlapped. However, unlike RAID-4, each drive takes turns storing parity information for a different series of stripes. Since there is no dedicated parity drive, all drives contain data and read operations can be overlapped on every

drive in the array. Write operations will typically access a single data drive, plus the parity drive for that record. Since, unlike RAID-4, different records store their parity on different drives, write operations can be overlapped.
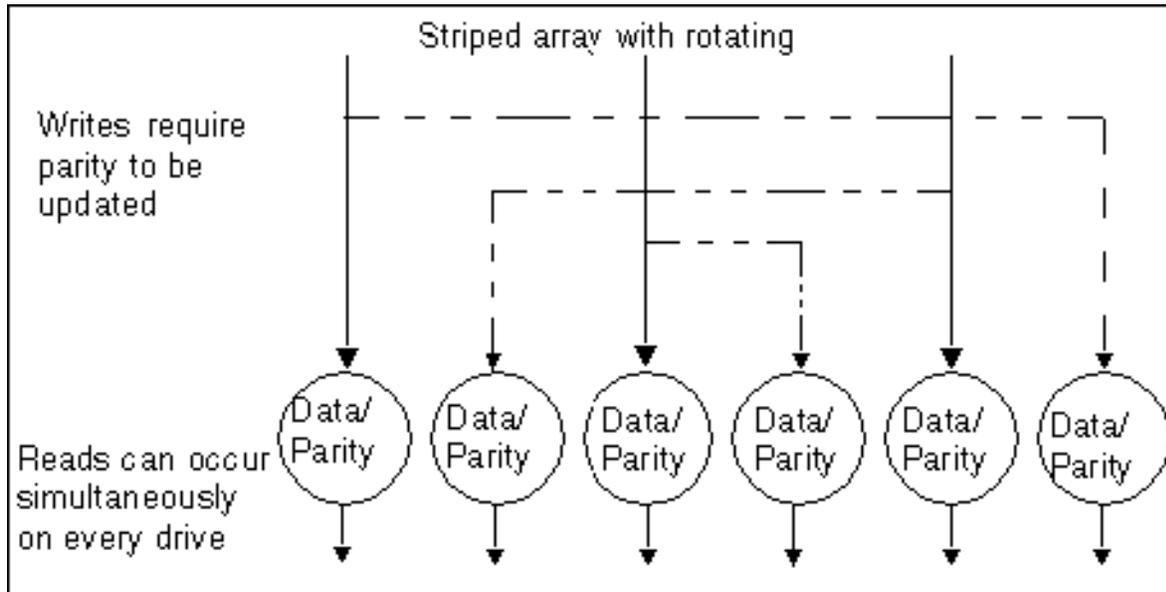

Figure 18: RAID-5 array

In summary:

RAID-0
  is the fastest and most efficient array type but offers no fault-tolerance.

RAID-1
  is the array of choice for performance-critical, fault-tolerant environments. In addition, RAID-1 is the only choice for fault-tolerance if no more than two drives are desired.

RAID-2
  is seldom used today, since ECC is embedded in almost all modern disk drives. RAID-2 is not supported by RAIDmaster.

RAID-3
  can be used in single-user environments which access long sequential records to speed up data transfer and provide fault-tolerance. However, RAID-3 does not allow multiple I/O operations to be overlapped and requires synchronized-spindle drives in order to avoid performance degradation with short records. RAID-3 is not supported by RAIDmaster.

RAID-4
  offers no advantages over RAID-5 and does not support multiple simultaneous write operations. RAID-4 is not supported by RAIDmaster.

RAID-5
  is the best choice for fault-tolerance in multi-user environments which are either not performance-sensitive, or which do few write operations. At least three drives are required for RAID-5 arrays.

6.3.3   Dual-level RAID (as of CS37)

In addition to the standard RAID levels, CS37 controllers can combine multiple hardware RAID arrays into a single array group. In such a dual-level RAID, the controller firmware stripes two or more hardware arrays of like RAID level into a RAID-0 single array.

Dual-level RAID achieves a balance between the increased data availability inherent in RAID-1 and RAID-5 and the increased read performance inherent in disk striping (RAID-0). These arrays are sometimes referred

to as RAID 0+1 or RAID 10 and RAID 0+5 or RAID 50 (see section Dual-level RAID (as of CS37)).

6.3.4   Creating data redundancy

RAID-5 offers improved storage efficiency over RAID-1 since parity information is stored, rather than a complete redundant copy of all data. The result is that three or more drives can be combined into a RAID-5 array, with the effective storage capacity of only one drive sacrificed to store the parity information. Therefore, RAID-5 arrays provide greater storage efficiency than RAID-1 arrays. However, this comes at the cost of a corresponding loss in performance.

When data is written to a RAID-5 array, the parity information must be updated. There are two ways to accomplish this. The first way is straightforward but very slow. The parity information is the XOR of the data on every drive in the array. Therefore, whenever one drive's data is changed, the other drives in the array which hold data are read and XORed to create the new parity. This requires accessing every drive in the array for each write operation.
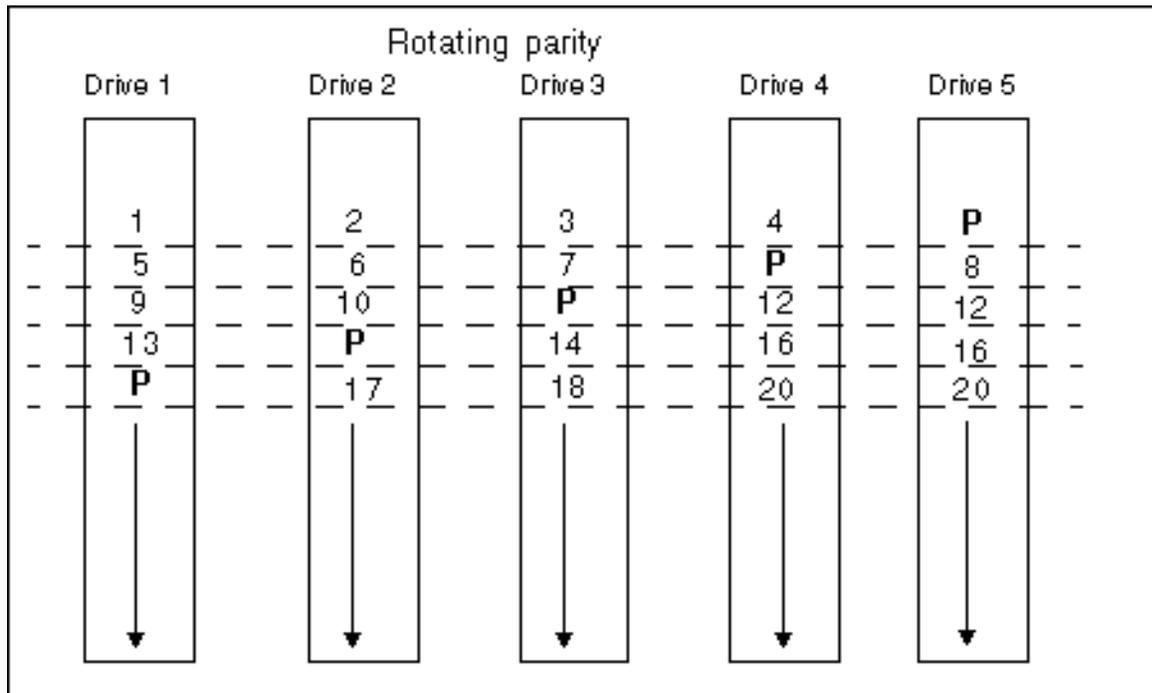

Figure 19: Rotating parity array

The second method of updating parity, which is usually more efficient, is to find out which data bits were changed by the write operation and then change the corresponding parity bits. This is accomplished by first reading the old data to be overwritten. This data is then XORed with the new data which is to be written. The result is a bit mask which has a one in the position of every bit which has changed. This bit mask is then XORed with the old parity information which is read from the parity drive. This results in the corresponding bits being changed in the parity information. The new updated parity is then written back to the parity drive. This results in only two reads, two writes and two XOR operations, rather than a read or write and XOR for every drive in the array.
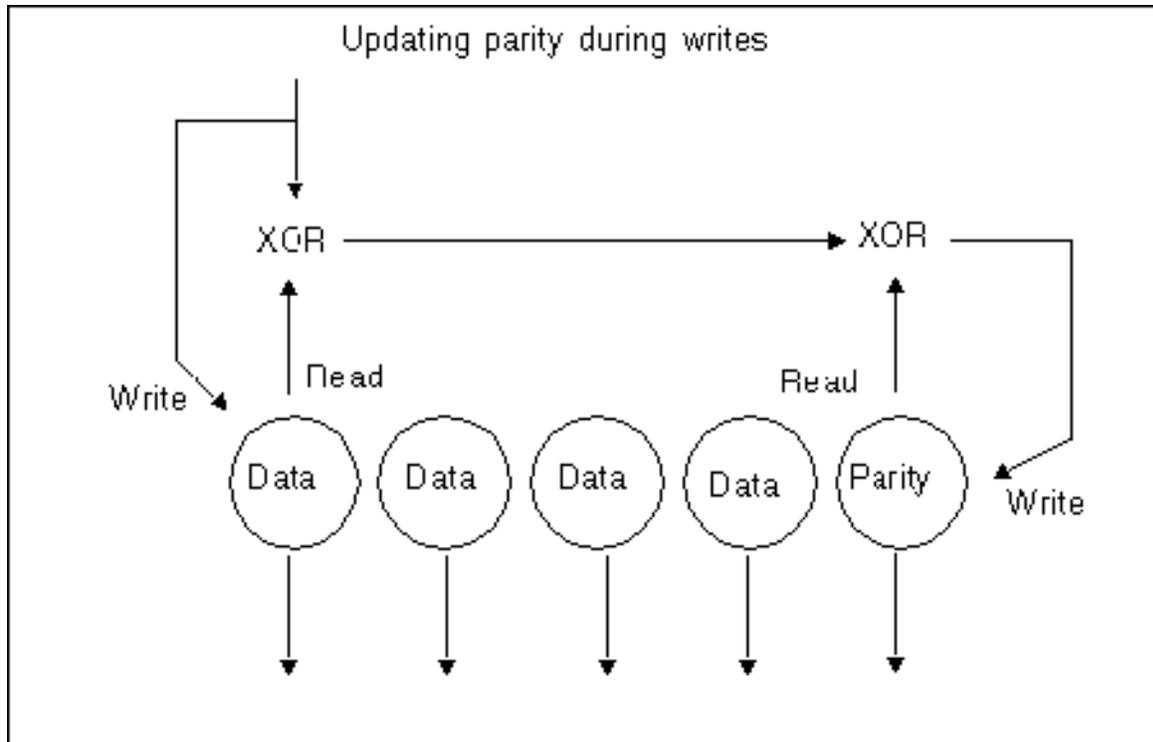
Figure 20: Regenerating parity information in a RAID-5 array

The cost of storing parity, rather than redundant data as in RAID-1, is the extra time taken during write operations to regenerate the parity information. This additional time results in a degradation of write performance for RAID-5 arrays over RAID-1 arrays by a factor of between 3:5 and 1:3. (i.e.   RAID-5 writes are only between 3/5 and 1/3 as fast as RAID-1 write operations.) Because of this, RAID-5 arrays are not recommended for applications in which performance is important. (The exception to this are applications which never write data.)

6.3.5   Recreating lost data

When a drive fails in a RAID-0 array, the entire array fails. In a RAID-1 array, a failed drive has little impact on performance, since data can be read from the duplicate mirrored drive. However, in a RAID-5 array, the data must be synthesized by reading and XORing the corresponding data stripes from the remaining drives in the array. This process is relatively slow and is thus called "degraded mode". The more drives contained in a RAID-5 array, the slower degraded mode operation becomes.
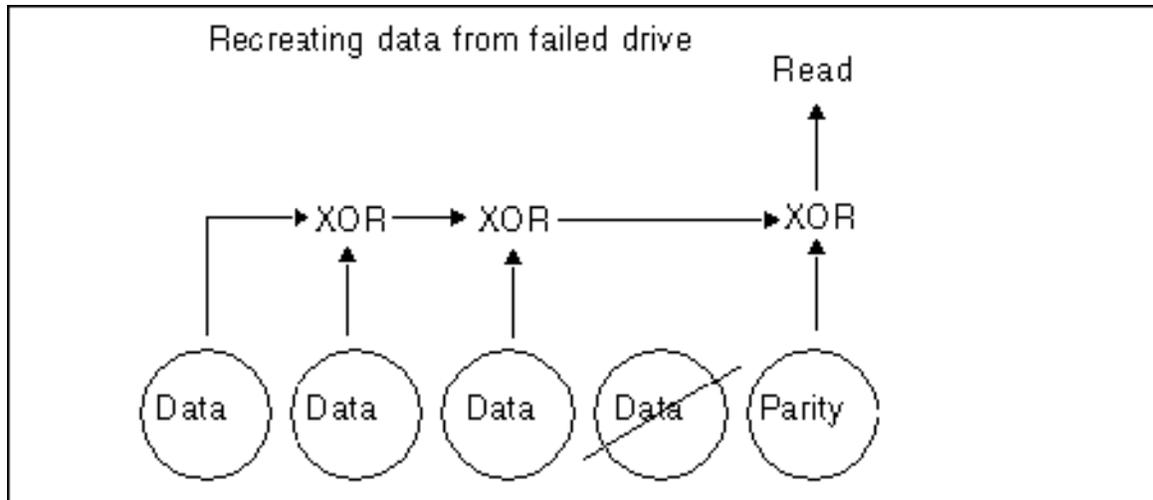
Figure 21: Recreating lost data in a RAID-5 array

6.3.6   Rebuilding a failed hard drive

When a failed drive is replaced in a RAID-1 or RAID-5 array, or if a Hot Spare is present, the RAIDmaster adapter will rebuild the lost data onto the new drive or Hot Spare. This rebuild operation occurs on-line while other normal reads and writes are being processed by the array. In the case of RAID-1 arrays, the rebuild will occur relatively quickly, since it simply involves copying data from the duplicate mirrored drive to the replacement drive. In the case of RAID-5 arrays, the new data to be written to the replacement drive must be synthesized by reading and XORing the corresponding data stripes from the remaining drives in the array. This will cause degraded performance to degrade even more. The more drives contained in a RAID-5 array, the longer the rebuild operation will take and the more pronounced the performance degradation will be during the rebuild process.

6.4   Command processing

RAIDmaster adapters process a command from the host computer as follows:

1. The CPU creates a command packet in system RAM and writes the address of this packet to a register in the RAIDmaster adapter. The CPU can then perform other tasks. All remaining operations will be performed by RAIDmaster. The command packet contains parameters, specifying the type of operation, the amount of data and the location on the peripheral device. The command packet may also contain pointers to a scatter/gather table and a status block. The scatter/gather table describes a list of data segments to be read or written to sequential locations on the peripheral device. These data segments may be scattered in many different areas of system RAM. The status packet is a place in system RAM for RAIDmaster to save status information concerning the successful or unsuccessful completion of a command.

2. RAIDmaster becomes the bus master (takes control of the system bus) and DMAs (direct memory accesses) the command packet and scatter/gather table into its local memory.

3. When RAIDmaster is ready to transfer the data, it again becomes the bus master and DMAs the data to the areas specified by the scatter/gather table.

4. When the command is complete, RAIDmaster becomes the bus master and DMAs the status packet into the proper RAM location.

5. RAIDmaster interrupts the CPU to tell the computer that the command has completed.

It supports overlapped I/O processing, command queuing and scatter/gather memory access.

Command overlapping

Peripheral controllers which cannot perform command overlapping can only execute one command from the host computer at a time. RAIDmaster can receive another command before the first command has finished executing. By allowing multiple outstanding commands, RAIDmaster adapters can optimize performance by overlapping commands for multiple SCSI devices, thereby keeping all devices as busy as possible.

Command queuing

The adapter is also able to queue multiple commands from the host computer and execute them out of order. Once a command has completed, RAIDmaster transfers a status packet into the host memory, containing status information plus information indicating which command was just completed. This allows commands to be executed in a different order than they were issued by the computer. RAIDmaster can thus queue commands and rearrange them so that they may be executed in a more efficient order for the particular peripheral configuration.

In addition, RAIDmaster supports tagged command queuing on the SCSI bus. This feature allows the adapter to transfer up to 64 SCSI commands to a single device if that device will accept queued commands. SCSI tagged command queuing may be disabled via the Storage Manager. However, this should not be necessary as the adapter will negotiate with each SCSI device to determine if it is capable of utilizing the command queuing feature.

Scatter/Gather

When using virtual memory addressing schemes, system memory may appear contiguous to the user but is actually fragmented into many widely scattered physical address locations. Because of this, it is often necessary when accessing a large amount of contiguous data from a peripheral device, to break up the transfer into many different locations in system memory. RAIDmaster adapters support scatter/gather, a method of providing multiple storage areas for data transfer in one command packet. This greatly increases performance.

I2O interface

The I2O (intelligent input/output) interface of the CS37 is optimized for universal I/O controller processing and allows all the I2O-compatible controllers in the system to be efficiently managed under one system driver. This OS-independent standard is protected by the I2O Special Interest Group (SIG).

6.5   SCSI interface

The Small Computer Systems Interface is an ANSI standard parallel interface designed to communicate with intelligent peripheral devices. RAIDmaster conforms to the SCSI specification, defined in ANSI document number X3.131-1986. In addition, RAIDmaster conforms to the ANSI SCSI-2 specification and CCS (common command set) documents.

6.5.1   SCSI device IDs

According to the SCSI specification, up to eight SCSI devices may be connected to a single 8-bit SCSI bus. All SCSI devices, including the RAIDmaster adapter, must be assigned a unique SCSI ID number. SCSI IDs, which are typically set via jumpers on peripheral devices, may be assigned any number from 0-7 (0-15 in the case of wide SCSI devices). The SCSI ID on RAIDmaster adapters is set by default to ID 7 and may be changed via the Storage Manager.

In 8-bit SCSI systems, the device with ID 7 has the highest priority and ID 0 has the lowest priority when arbitrating for use of the SCSI bus. In 16-bit SCSI systems, the sequence of priorities is 7 ... 0 and then 15 ... 8.

6.5.2   LUNs

Any SCSI device may contain up to eight sub-devices or logical units, which are assigned   logical unit numbers (LUN) 0-7. However, most SCSI devices other than bridge controllers consist of only one LUN and will be accessed as LUN 0 by RAIDmaster.

6.5.3   Commands and messages

All SCSI commands are supported by RAIDmaster. Using EATA or 120 protocol (the latter for CS37 only), RAIDmaster will pass on any SCSI command from the computer to the SCSI peripheral device.

RAIDmaster supports one-byte and extended messages, such as Abort, SCSI Reset, Initiator Detected Error, and Parity Error. SCSI message protocol is normally handled automatically by RAIDmaster without requiring intervention by the host computer. This allows much faster SCSI protocol handling and puts less load on the host CPU, freeing it for other tasks.

6.5.4   Disconnect/Reconnect

Full support for disconnection and reconnection by SCSI peripheral devices is provided by RAIDmaster adapters. The disconnect/reconnect message protocol is handled automatically by RAIDmaster without host intervention.

Disconnection allows a SCSI device and adapter to temporarily release control of the SCSI bus during periods of non-activity while a SCSI command is in process, so that another SCSI device can share the bus. However, disconnection and reconnection protocol adds unnecessary overhead when only one device is utilizing the SCSI bus. RAIDmaster adapters optimize performance by disabling disconnection during periods of single-device activity.

6.5.5   Transfer padding

In some unusual cases, a SCSI peripheral device may read or write more data than requested by the computer. If this occurs during a peripheral write operation, RAIDmaster will automatically transfer "pad-data" by adding additional bytes of FF_H to the end of the data to be written to the SCSI device. If this occurs during a peripheral read operation, RAIDmaster will automatically strip off additional bytes of data read from the SCSI device, thereby preventing host memory from being overwritten.

6.5.6   Synchronous/Asynchronous SCSI

RAIDmaster adapters are capable of handling mixed SCSI devices with different transfer speeds, as well as both synchronous and asynchronous protocol, on the same SCSI Bus. RAIDmaster automatically negotiates with each SCSI device during the power on sequence to determine the optimum transfer speed and method.

The preset maximum SCSI bus transfer rate of 10MHz synchronous may be changed via the Storage Manager. The default for CS37 may be higher (up to 40 MHz), depending on the device.

6.5.7 Fast/Ultra/Ultra2/LVD SCSI(2/3)

The original SCSI specification allowed transfer rates of up to 5 MBit/s. Standards supporting higher rates are identified by the word Fast. There are now a number of new standards which offer vast improvements in the handling of transfer frequencies, bus transfer rates and cable lengths:

Standard	MHz	MB/s	Devs. per bus	Cable length	Connectors
SCSI-1	5	5	8	6 m (differential: up to 25 m)	Internal: 50-pin, 2-rowExternal:50-pin, low-density Centronix-A
SCSI-2:					
Fast (Fast Narrow)	10	10	8	Up to four devices: 3 m; five or more devices:1.5 m (differential: up to 25 m)	Internal: 50-pin, 2-rowExternal:50-pin, low-density Centronix-A cable or 50-pin, high-density SCSI-2
Fast Wide (Wide)	10	20	16	Up to four devices: 3 m; five or more devices: 1.5 m (differential: up to 25 m)	Internal and external: 68-pin, high-density SCSI-2
SCSI-3:					Internal: – 68-pin, high-density SCSI-2 – 68-pin SCA and SCA/2 (for drive hot plugging) External: – 68-pin, high-density SCSI-2 – 68-pin, high-density Champ
Ultra (Fast20, Ultra Narrow)	20	20	8	1.5 m (LVD: 12 m)	
Ultra Wide (Fast Wide 20)	20	40	16	1.5 m (LVD: 12 m)	
Ultra2	40	40	8	12 m	
Ultra2 Wide	40	80	16	12m	
Ultra3	80	80	8	12 m	
Ultra3 Wide	80	160	16	12 m	

Table 2: SCSI standards

6.5.8 LVD

Non-differential SCSI busses are referred to as "single-ended". SCSI busses with 2 wires per signal line are known as "differential SCSI systems". LVD stands for "low voltage differential" and operates with a 3.3 V

potential difference between two signal lines. This form of mechanism allows for optimum equalization of magnetic and electrical noise that may affect the SCSI bus.

> If a single-ended device is connected to an LVD bus, the controller automatically switches back to single-ended operation.

> A **standard** differential controller is not capable of controlling single-ended devices.

6.5.9   SCSI termination

SCSI termination can be disabled via the Storage Manager for RAIDmaster PCI adapters. The default setting for adapter SCSI termination is "Enabled".

RAIDmaster adapters contain Alternative-2 active SCSI terminators for maximum reliability with fast SCSI devices. Active termination uses one 110   ohm resistor per signal, pulled up to a locally supplied voltage which is regulated at 2.85V.

Features of active termination include:

- Better immunity to fluctuations in TERMPWR
- Closer match to the characteristic impedance of the cable minimizes reflections
- Increased noise margin

Passive termination (called Alternative-1 in the SCSI-2 specification) is still utilized on many SCSI devices today. Passive termination employs one 220   ohm pull-up and one 330 ohm pull-down resistor per signal. For better signal integrity, especially in fast SCSI environments, SCSI devices with active termination should be used whenever possible. However, active and passive termination can be mixed on one SCSI bus, and that results in better reliability than using passive termination alone.

> Active termination is essential for ULTRA and ULTRA2 configurations (bus speed at least 20MHz).

6.6   RAIDmaster activity LEDs

A 10-segment LED display on RAIDmaster adapters provides a means to monitor adapter activity. There are several different adapter states which are indicated by the LED. To view the LED display, you must remove the side panel.

6.6.1   LED display during adapter idle

When no commands are in progress and all SCSI activity has ceased, the adapter enters the idle state. This is indicated by a rotating bit pattern in LED segments 1 through 8.

6.6.2   LED display for CS31, CS34, CS35

By viewing these LEDs, the user can determine information concerning cache hits and misses, disk reads and writes, and computer bus reads and writes. In addition several LEDs are provided for troubleshooting purposes. Refer to the chapter Troubleshooting guide for more information on interpreting these LED patterns.

Segment	Function
1	Busy
2	Computer bus transfer to adapter
3	Computer bus transfer from adapter
4	SCSI disconnect
5	SCSI active
6	Disk read
7	Disk write

8	Controller reset
9	Interrupt pending
10	DMA request pending

Table 3: LED blink codes for CS31, CS34 and CS35

LED display during power-up

During the power-up sequence, RAIDmaster's 10-segment LED display shows the amount of expansion cache present. The expansion cache size in megabytes is displayed as a binary pattern, with lit LEDs representing binary one. LED 1, the leftmost LED is the least significant binary bit. For example, if 4 MB of cache memory is installed, LED 3 will light. If 12 MB is installed, LEDs 3 and 4 will light.

All RAIDmaster adapters then enter a fifteen-second waiting period in order to give all SCSI peripherals time to power-up. During this period the LED alternately flashes segments 1, 2, 5, 6 and segments 3, 4, 7, 8.
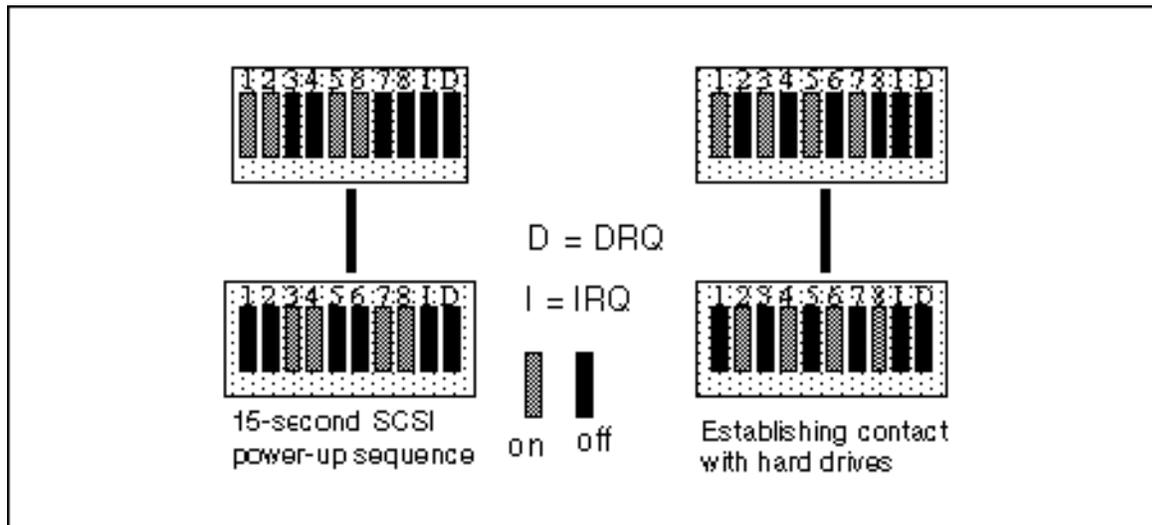

Figure 22: LED display at system startup (for CS31, CS34 and CS35)

After the 15-second waiting period, RAIDmaster attempts to establish contact with the SCSI hard drives. During this period, the LED alternately flashes segments 1, 3, 5, 7 and segments 2, 4, 6, 8. Depending upon the length of time the drives require for spin-up, this may take up to 5 seconds.

Once contact has been established with the hard drives or after the 20-second timeout period, RAIDmaster will enter the idle state and display a slower, rotating bit pattern which repeats approximately once per second.

6.6.3   CS37 controller LEDs

Nine activity LEDs on CS37 controllers provide a means to visually monitor controller activity. Several different controller states are indicated by the LED as outlined in the following sections. Two additional LEDs indicate the status of the cache memory on the controller.

LED display during power-up

During its power-up sequence, the controller passes through the following states in sequence, as indicated by the LEDs:

Memory mapped I/O   LEDs 6 and 7 and LEDs 5 and 8 flash
                                     alternately while the controller waits for
                                     the host computer to initialize the
                                     address translation unit (ATU) on the
                                     PCI bus.

Bridge	LEDs 5 and 7 and LEDs 6 and 8 flash alternately while the controller waits for the host computer to initialize its PCI-to-PCI bridge.
Fast-idle	The controller enters a fast idle pattern while waiting for the I20 initialization commands after the address translation unit (ATU) and PCI bridge have been enabled.

The ATU, the PCI bridge and the cache module (if installed) must be configured and enabled before the I20 initialization can be completed. If any one of these are not configured or become disabled during the fast idle phase, the pattern will return to the respective wait pattern for the affected component.

LED display during controller idle

When no commands are in progress and all bus activity has ceased, the controller enters the idle state. This is indicated by a rotating bit pattern in LEDs 1 through 8.

LED display during controller active

Once the power-up sequence has been completed, viewing these LEDs can help determine the operating status of the controller. These LED patterns are also useful for troubleshooting. Refer to the chapter on troubleshooting for more information on interpreting LED patterns.

LED	Function
1	Heartbeat. Indicates that controller interrupts are enabled and that the controller is alive. During controller activity, this LED flashes four times a second.
2	Indicates the occurrence of a non-maskable interrupt (NMI) to the I/O processor on the controller.
3	Indicates the controller's internal operating system is in its idle loop.
4	Indicates that the controller's internal operating system is processing an interrupt.
5	Reserved for future use.
6	Indicates that the cache controller is using DMA to perform a data transfer.
7	Indicates that the controller is generating parity information for a RAID-5 array (hardware XOR).
8	Indicates that there is a command on the SCSI or Fibre bus.
IRQ	Lights when the controller activates an interrupt on the host PCI bus.

Table 4: LED blink codes for CS37

The LED activity on the CS37 is not comparable with that on the CS31/35. Only the busy lamp (LED 1) indicates that the adapter is dealing with requests.
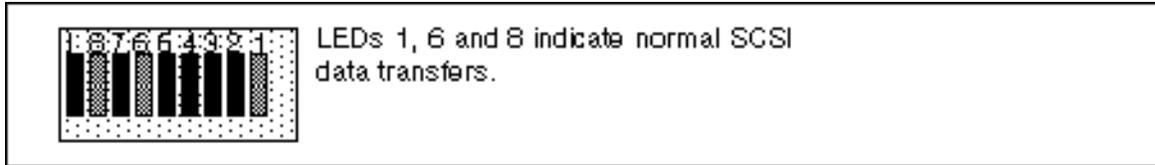
Figure 23: LED display for data transfer between cache and SCSI device

Cache status LEDs

Two LEDs on the CS37 controller indicate the status of the on-board cache RAM:

- The green ECC enabled LED, when lit, indicates that all installed SIMMs/DIMMs are DPT ECC SIMMs/DIMMs, and thus the controller cache and the internal data paths are ECC protected.

- The red ECC error LED, when lit, indicates that a correctable or non-correctable error has been detected in one of the ECC SIMMs/DIMMs. The LED will stay lit after the error has been corrected until the controller is powered-down. Cache failure information is recorded in the controller error log and can be viewed through Storage Manager.

6.7   Reset NVRAM

CS37 controllers retain their setup parameters even when powered off. These parameters are stored on the controller in a small area of non-volatile memory (NVRAM). There is a possibility that through improper configuration, the controller can be put into a state in which it hangs the system during boot. If this happens, the parameters in the NVRAM can be restored to their default settings by following the steps below:

1. Turn off power to the system.
2. Place a shorting jumper across pins 1 and 2 of P4 on the controller (see section CS37 adapter profile).
3. Power on the system and wait until the LEDs 3, 5, 7 and 8 on the controller begin flashing.
4. Turn off the power to the system and remove the jumper.

The controller can now be reconfigured using Storage Manager.

7 Troubleshooting guide

This troubleshooting guide provides answers to many commonly asked questions. If a situation occurs which is not covered in this chapter, or if the recommendations here do not correct the problem, contact Siemens customer support. We will be happy to assist you.

7.1 For CS31, CS34 and CS35 adapters

Problem:
> The RAIDmaster adapter is not detected on system startup.

Solution:
> This may indicate that several SCSI drives are configured with the same SCSI ID. Reconfigure the SCSI drives with unique SCSI IDs.

Problem:
> The controller produces whistling sounds.

Solution:
> Call the Storage Managers which switches off the whistling sound. As Storage Manager is not available when the system is booting, you must wait until Storage Manager can be accessed. This may take some time.

Problem:
> LEDs 4 and 8 flash once per second at power-up and the adapter does not respond.

Solution:
> This pattern indicates that the adapter has detected a problem with the SCSI cable termination.

Problem:
> LEDs 1, 2 and 8 flash once per second at power-up and the adapter does not respond.

Solution:
> This pattern indicates that there is no SIMM in socket #1 of the RAIDmaster adapter. Socket #1 must always contain the largest capacity SIMM.

Problem:
> LEDs 2 and 8 flash once per second at power-up and the adapter does not respond.

Solution:
> This pattern indicates that SIMM socket #1 of the RAIDmaster adapter does not contain the largest SIMM. Rearrange the SIMMs so that the largest capacity SIMM is in socket #1.

Problem:
> LEDs 1, 2, 3, 4 and 8 flash once per second at power-up and the adapter does not respond.

Solution:
> This pattern indicates that a combination of ECC and parity SIMMs have been installed on the RAIDmaster adapter. Remove either the parity or the ECC SIMMs.

Problem:
> LEDs 3 and 8 flash once per second at power-up and the adapter does not respond.

Solution:
> This pattern indicates that a combination of parity and ECC SIMMs have been installed on the RAIDmaster adapter and SIMM socket #1 does not contain a parity SIMM. Rearrange the SIMMs so that a parity SIMM is in socket #1, or do not mix parity and ECC SIMMs.

Problem:
> Various LEDs in the 1-4 range flash once per second and the adapter does not respond.

Solution:
> This pattern indicates an internal microprocessor trap in the adapter. Remove all SCSI devices, cables and option modules and retry. If the trap error disappears, reconnect everything, one device at a time, until the faulty device, cable or module is isolated. A microprocessor trap can also occur because of a

firmware EPROM error during normal operation. Try reseating the EPROMs in their sockets to ensure good contact.

Problem:

Some other combination of LEDs not listed above flashes once per second at power-up and the adapter does not respond.

Solution:

This indicates that a firmware trap has occurred. To determine the cause of the trap, run Storage Manager (from floppy if necessary) while the LEDs are still flashing. A description of the problem and possible solutions will be displayed upon entering the program.

Problem:

Although the SCSI devices can be accessed by the RAIDmaster adapter, the red fault LEDs on the devices in the OLR storage cabinet do not flash during boot-up and the RAIDmaster adapter does not detect drive swaps or cabinet failures.

Solution:

These symptoms indicate that the OLR storage cabinet status signals are not being properly received by the RAIDmaster adapter. One cause can be another SCSI device or non-Siemens cabinet connected to the external SCSI cable along with the OLR cabinet. Other devices will typically ground these signals.

7.2   For the CS37 adapter

Problem:

The RAIDmaster adapter is not detected on system startup.

Solution:

This may indicate that several SCSI drives are configured with the same SCSI ID. Reconfigure the SCSI drives with unique SCSI IDs.

Problem:

The controller produces whistling sounds.

Solution:

Call the Storage Managers which switches off the whistling sound. As Storage Manager is not available when the system is booting, you must wait until Storage Manager can be accessed. This may take some time.

Problem:

Some other combination of LEDs not listed above flashes once per second at power-up and the adapter does not respond.

Solution:

This indicates that a firmware trap has occurred. To determine the cause of the trap, run Storage Manager (from floppy if necessary) while the LEDs are still flashing. A description of the problem and possible solutions will be displayed upon entering the program.

Problem:

Although the SCSI devices can be accessed by the RAIDmaster adapter, the red fault LEDs on the devices in the OLR storage cabinet do not flash during boot-up and the RAIDmaster adapter does not detect drive swaps or cabinet failures.

Solution:

These symptoms indicate that the OLR storage cabinet status signals are not being properly received by the RAIDmaster adapter. One cause can be another SCSI device or non-Siemens cabinet connected to the external SCSI cable along with the OLR cabinet. Other devices will typically ground these signals.

Problem:

LEDs 1,6 and 8 are permanently on at full brightness and no SCSI activity is to be observed on the disks.

Solution:

There has been a SCSI bus malfunction. A SCSI reset may already be in progress, or you can use reinitdev to reinitiate the SCSI bus by hand (SCSI bus reset included).

Problem:

The controller does not respond and the IRQ LED (and any other LEDs) remains lit.

Solution:

The IRQ LED indicates that the controller IRQ assignment is pending. This usually indicates an IRQ conflict with another card. Ensure that each card is set to a unique IRQ.

Problem:

The controller does not respond and one of the following paterns of LEDs flash once per second at power-up:

Pattern	Meaning
1, 2, 5, 6, 7	None
1, 3, 5, 6, 7	High
2, 3, 5. 6. 7	Mismatch
1, 2, 3, 5, 6, 7	Invalid

Solution:

These patterns indicate that there is a problem with the memory modules on the controller.

None	Either no memory modules were detected on the controller, or there is no module in socket #1. Socket #1 must always have a 16MB or 64MB 60nx EDO memory module installed.
High	Too much memory has been detected on a CS37 controller. Rmove memory to bring the total to under 64MB.
Mismatch	Memory modules of mixed sizes have been detected or a SIMM slot was skipped. All installed modules must be the same size and must be filled sequentially from socket 1 to socket 4.
Invalid	A memory module of a size other than 16MB or 64MB has been detected. Install only valid sizes.

Problem:

The controller does not respond and various LEDs in the 1-4 range flash once per second.

Solution:

This pattern indicates an internal microprocessor trap in the controller. Remove all attached devices, cables and option modules and retry. If the trap error disappears, reconnect everything, one device at a time, until the faulty device, cable or module is isolated. If the error persists, contact Siemens Technical Support.

Problem:

After updating the CS37 controller firmware or BIOS and rebooting, the adapter does not respond.

Solution:

The firmware/BIOS update may have been unsuccessful. The controller is in a state in which it hangs the system during boot. If this happens, the parameters in the NVRAM can be restored to their default settings using the following procedure:

1. Turn of power to the sytem.
2. Place a shorting jumper across pins 1 and 2 of P4 on the controller (cf. illustration, see section CS37 adapter profile
3. Power on the system and wait until LEDs 3, 5, 7 and 8 on the controller begin flashing.
4. Turn off power to the system and remove the jumper. The controller can now be reconfigured using Storage Manager.

Problem:

After updating the CS37 controller firmware or BIOS and rebooting, LEDs 1 and 5 or 2 and 5 flash once per second.

Solution:

These patterns indicate that the adapter startup code detected a firmware checksum error or a flash error. Attempt the firmware update procedure again (see the "Firmware Update/Flash Button" section of the Storage Manager online help).

8 Appendix

8.1 Specifications

| Power requirements	
Voltage	5V ± 5%
Ripple and noise:	max. 50 mV peak-to-peak
Current	
CS31	2.20 A typical
CS34	1.25 A typical
CS35	1.15 A typical
CS37 + MR04	1.50 A typical
CS01	1.35 A typical
CS04	0.65 A typical
CS05	0.32 A typical
CS07	0.5 A typical

Table 5:

Environmental specifications	
Operating temperature:	0 to 50°C
Relative humidity (operating):	10% to 90% non-condensing
Maximum altitude for operation:	3,000 meters (10,000 feet)

Table 6:

8.2 Assembly diagrams

CS31 adapter profile
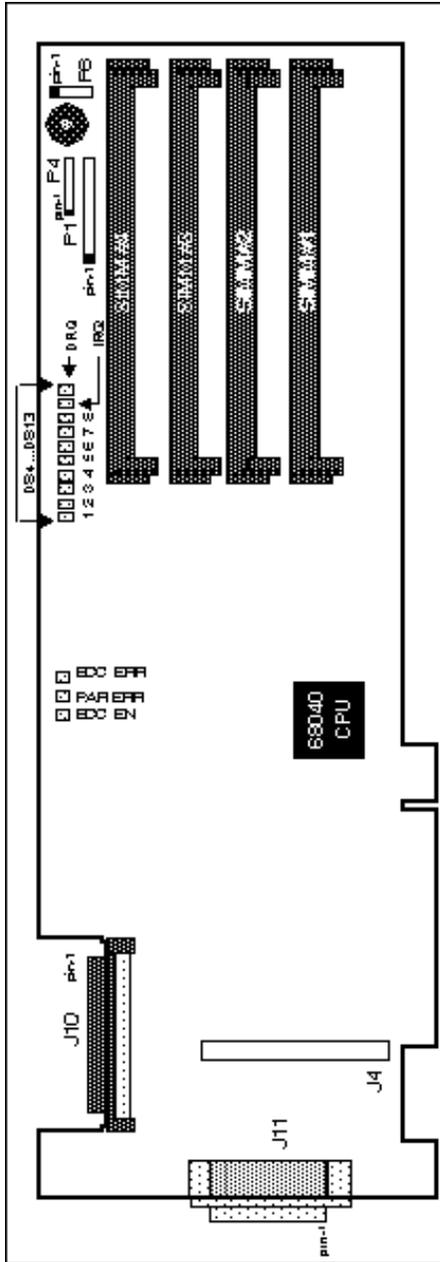
J4　Remote alarm connector
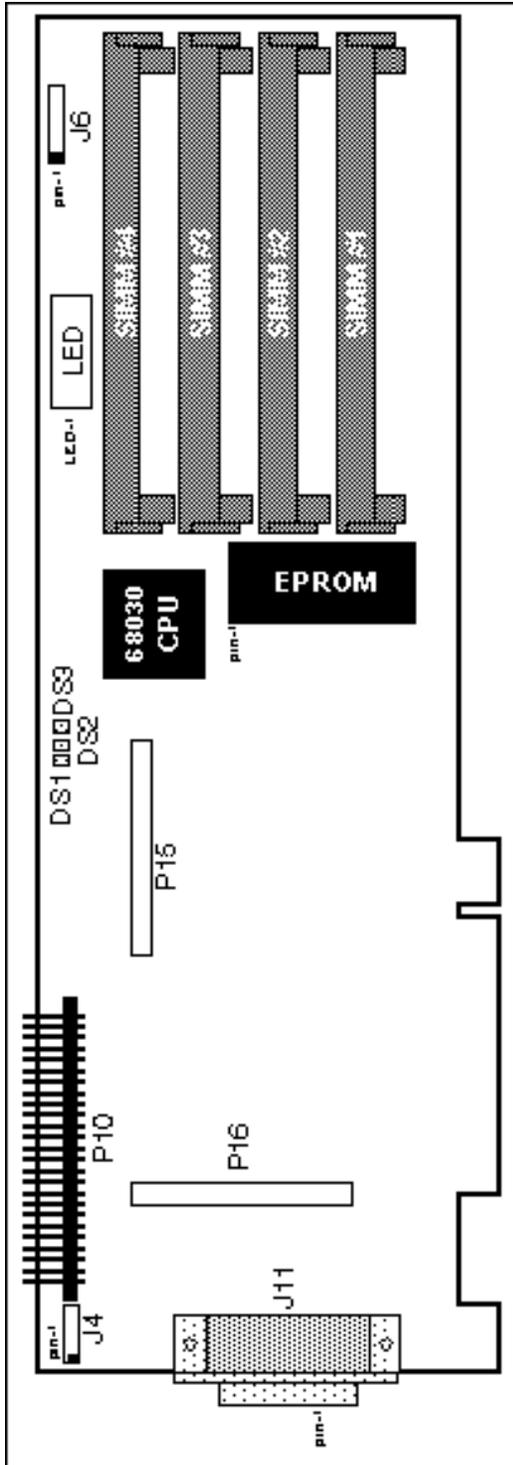
J6   Disk activity LED connector

J11   External 8-bit connector (bus 0)

P10   Internal 8-bit SCSI connector (bus 0)

P15, P16   CS01 connectors

DS1   ECC enabled LED (green)

DS2   Parity error LED (red)
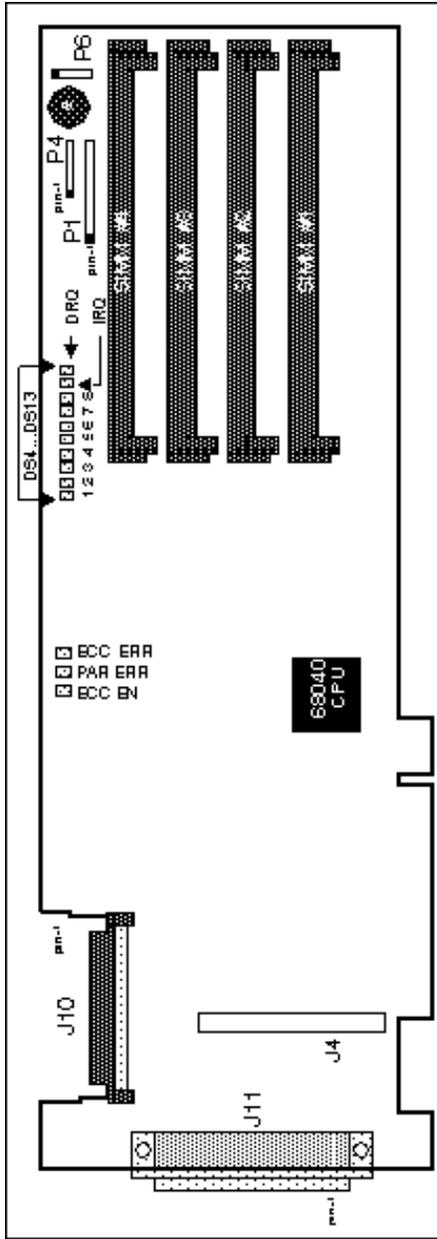
DS3   ECC error LED (red)

CS34 adapter profile

P1   Firmware flash clear jumper, pins 5

CS35 adapter profile

& 6

P4   NVRAM clear jumper, pins 1 & 2 / Remote alarm connector

P6   Disk activity LED connector

J4   CS04 connector

J10   Internal wide differential SCSI connector (bus 0)

J11   External wide differential SCSI connector (bus 0)

ECC ERR   ECC error LED

ECC EN   ECC enabled LED (green)

PAR ERR   Parity error LED (red)

DS4...DS13   Adapter activity LEDs

P1   Firmware flash clear jumper, pins 5

& 6

P4   NVRAM clear jumper, pins 1 & 2 / Remote alarm connector

P6   Disk activity LED connector

J4   CS05 connector

J10   Internal wide differential SCSI connector (bus 0)

J11   External wide differential SCSI connector (bus 0)

ECC ERR   ECC error LED

ECC EN   ECC enabled LED (green)

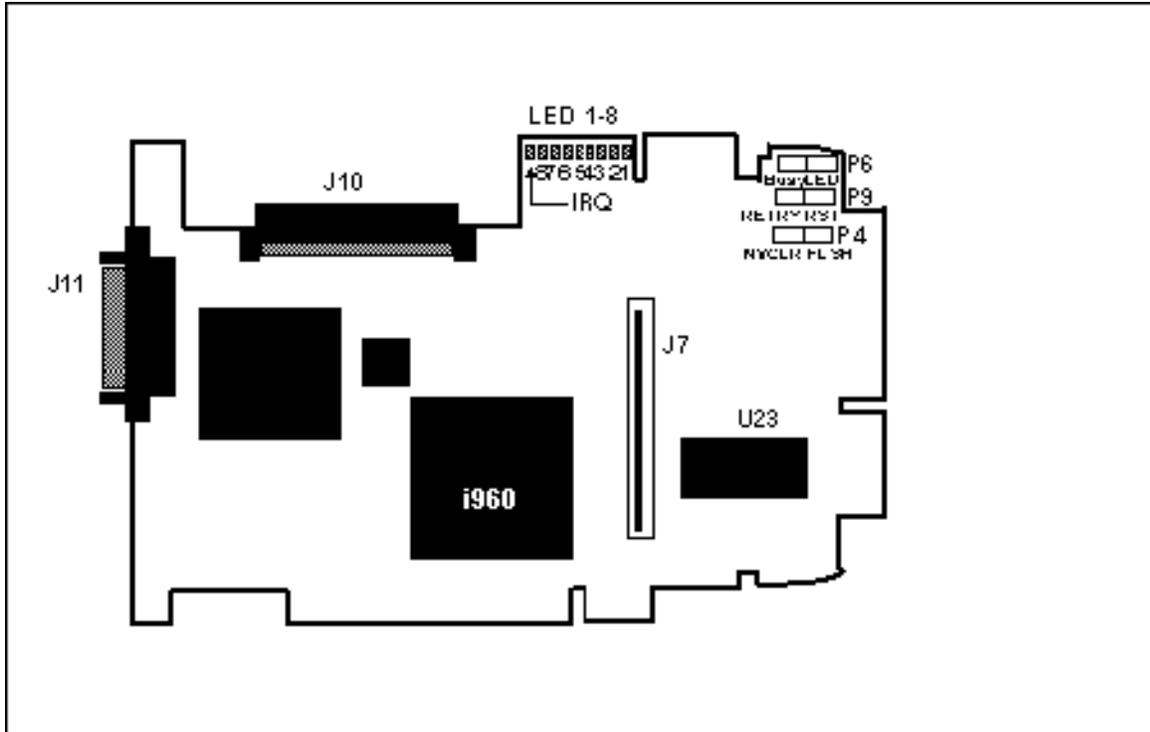PAR ERR   Parity error LED (red)

DS4...DS13   Adapter activity LEDs

CS37 adapter profile

P4   NVRAM clear (PIN 1-2) + FLASH enable (PIN 3-4)

P6   Busy LED connector

P9   Retry, (PIN 1-2), RESET (PIN 3-4)

LEDs 1-8 + IRQ
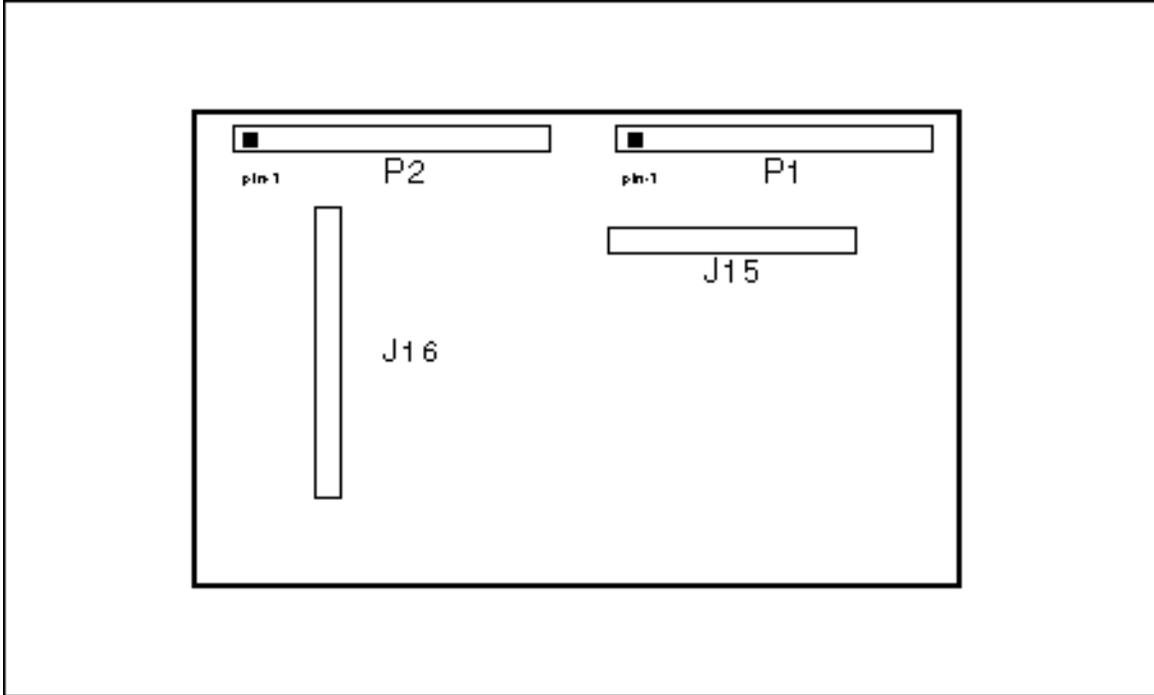   Adapter status, internal (see section LED display for CS31, CS34, CS35)

J7   CS06 RAID cache expansion connector
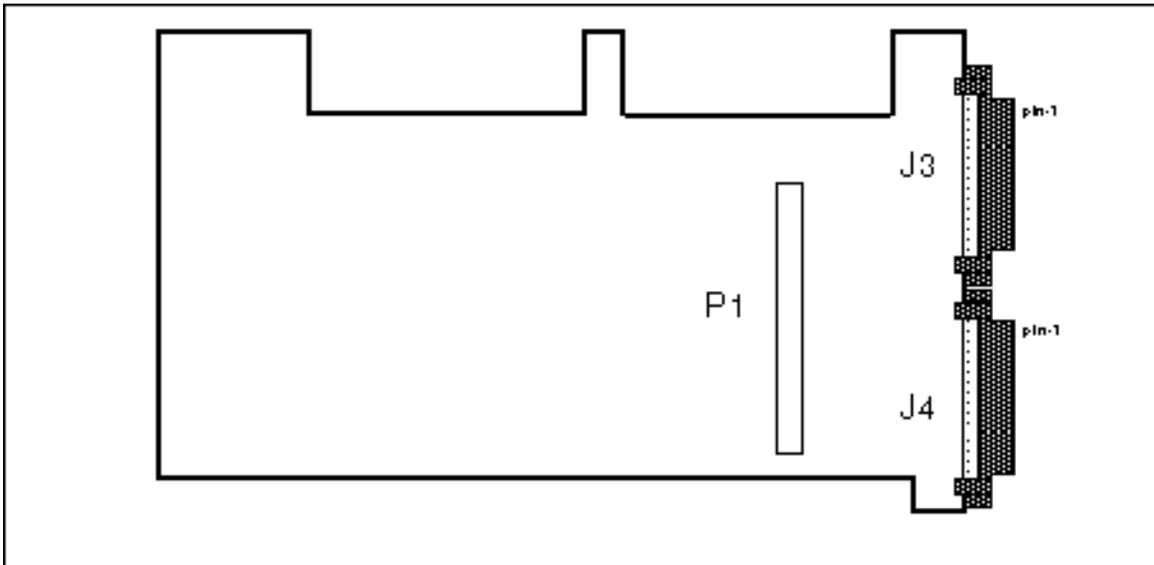
J10   Internal wide differential SCSI connector (bus 0)

J11   External wide differential SCSI connector (bus 0)

CS01 module profile

P1   Internal 8-bit SCSI connector (bus 1)

P2   Internal 8-bit SCSI connector (bus 2)

J15, J16
        Mating connectors for RAIDmaster adapter (reverse side)

CS04 module profile

J3   Differential wide SCSI connector (bus 1)

J4   Differential wide SCSI connector (bus 2)

P1   Mating connector for RAIDmaster adapter

CS05 module profile
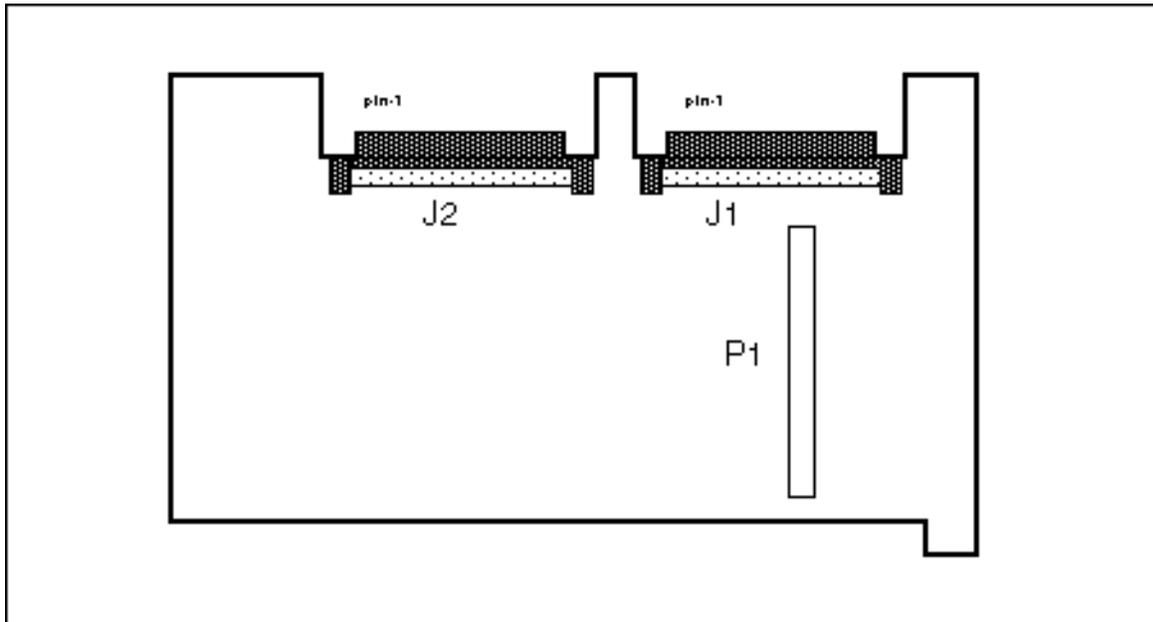
J1   Internal wide SCSI connector (bus 1)

J2   Internal wide SCSI connector (bus 2)

P1   Mating connector for RAIDmaster adapter

CS07 module profile
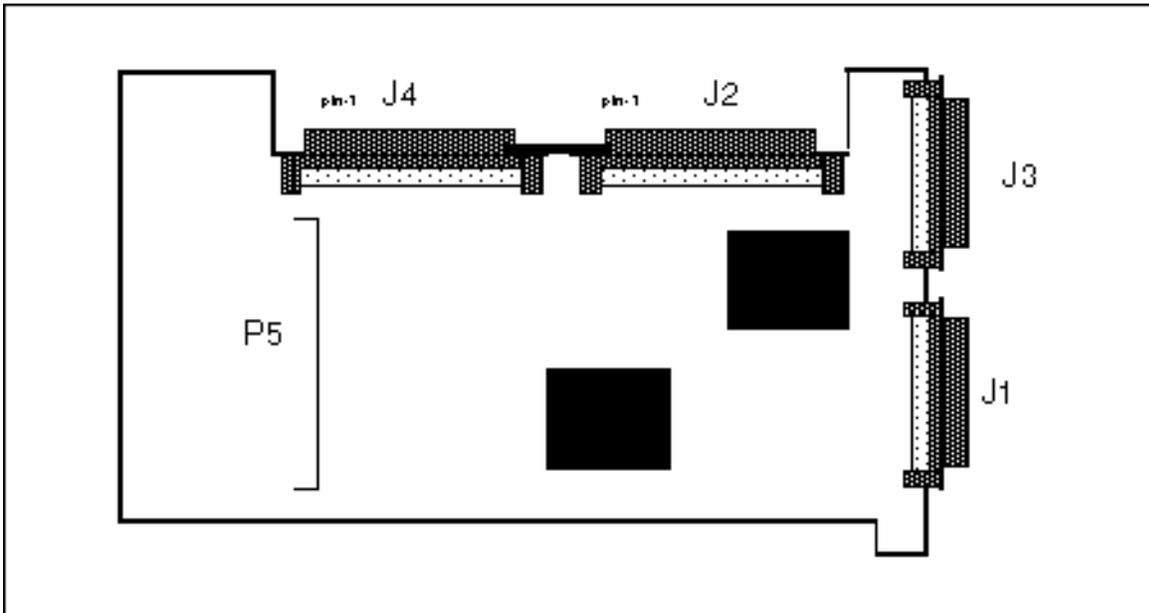
J1   Internal Ultra2/LVD Wide SCSI connector (bus 1)
J2   External Ultra2/LVD Wide SCSI connector (bus 1)
J3   Internal Ultra2/LVD Wide SCSI connector (bus 2)
J4   External Ultra2/LVD Wide SCSI connector (bus 2)
P5   Plug adapter for RAIDmaster controller

MR04 RAID cache expansion module
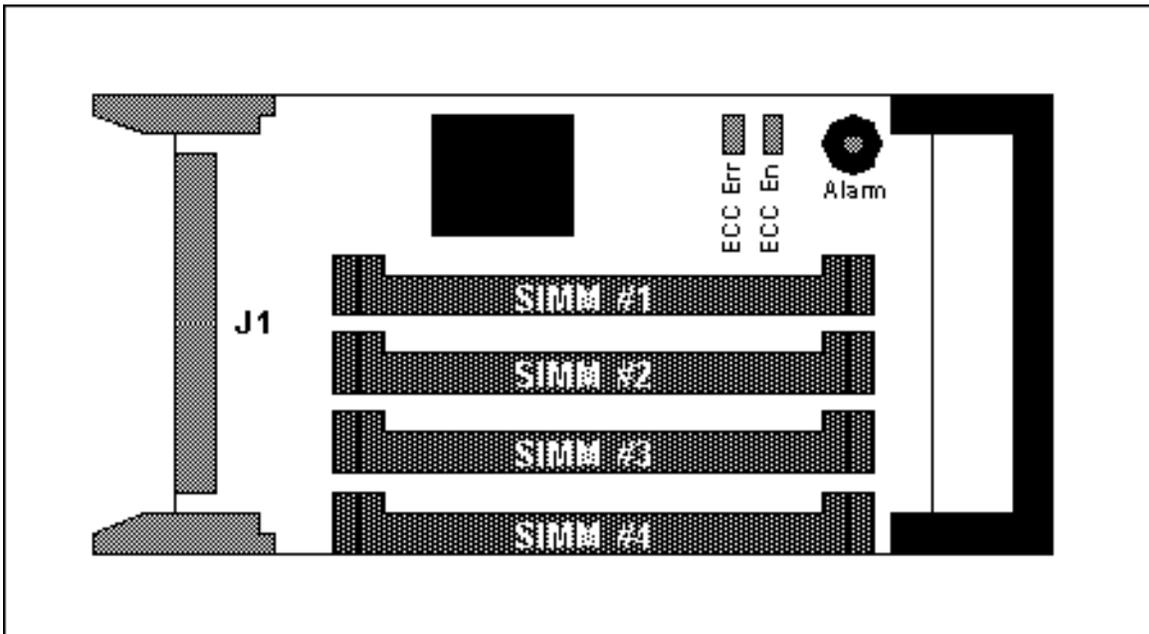
J1   RAIDmaster adapter connector
ECC En
      ECC enabled LED (green)

ECC Err
    ECC error LED (red)

Glossary

adapter
> SCSI Host Bus Adapter or HBA

ANSI
> American National Standards Institute.

array
> A group of disk drives which appear to the computer as a single LSU.

asynchronous
> Data transfer protocol which is not synchronized to a set timing interval. Asynchronous SCSI data transmitting devices must wait after each byte for acknowledgment from the receiving device. Either device can take as long as it wishes to send or acknowledge data. Asynchronous SCSI has no defined maximum transfer rate but is typically limited to 1.5 to 3 MHz.

bridge controller
> A SCSI device which appears as a single ID on the SCSI bus, but which bridges to multiple devices. Each of these devices is assigned a logical unit number or LUN. Bridge controllers are typically used to control remote RAID subsystems.

build
> The operation of initializing a redundant array by creating consistent redundant information. In the case of a RAID-1 array, one drive can be copied to the other. In the case of RAID-5 arrays, the parity information must be generated.

burst
> A term used to describe data words which are transmitted as a single group across a bus without interruption by another device.

bus mastering
> A method of data transfer which allows data to be moved between a peripheral controller and system memory without interaction with the host CPU or a third party DMA controller. This technique allows the peripheral controller to take control of the system bus and move data at up to 132MB/s.

cache
> A temporary fast storage area for data which would normally be accessed from a slower storage device. A cache management algorithm monitors the data access patterns and selects which data from the slower device is to be kept in the cache for quick access. Caches are normally transparent or hidden from the accessing device.

CD-ROM
> Compact Disk Read-Only Memory. A read-only storage device which retrieves up to 660 Mbytes of information from a removable laser disk similar to an audio compact disk.

command overlapping
> A feature which allows the simultaneous execution of more than one I/O command by a peripheral controller.

command packet
> An area of host computer memory which is used to store an I/O command and associated parameters for a peripheral controller which uses EATA protocol. Commands for DPT controllers are stored in command packets.

command queuing
> A feature which allows multiple I/O commands to be executed by a peripheral controller in a more efficient order.

degraded mode
>The mode of operation of a redundant array in which data is synthesized for read accesses to a failed drive.

device
>According to the SCSI specification, up to eight or 16 SCSI devices may be connected to a single SCSI bus. SCSI devices include peripherals such as disk and tape drives, optical devices, scanners, printers, and host adapters. Each SCSI device contains a SCSI ID number which can be set to 0-7 or 0-15 (Wide SCSI).

differential
>An electrical signal protocol which transmits information through a current loop rather than by changes in voltage, thereby reducing the susceptibility to electrical interference. Differential SCSI uses RS-485 transceivers to transfer 10MHz data at distances up to 25 meters (82 feet).

digital fault bus
>A proprietary Digital Equipment Corporation (DEC) bus standard that provides signals to host adapters for remote system management and device fault notification.

DIMM
>Dual In-line Memory Module - A standard way of packaging RAM on a small circuit board with a defined edge connector. See also SIMMs.

dirty page
>A cache page in which data has been written or modified but which has not yet been copied to the storage device. Once the data has been copied to disk, the page is said to be "clean".

DMA
>Direct Memory Access - A method of data transfer which allows data to be moved between a peripheral controller and system memory without interaction with the host CPU. The data is moved by the peripheral controller itself (bus mastering).

driver
>A software routine which receives I/O requests from higher levels within the operating system and converts these requests to the protocol required by a specific hardware device such as a SCSI adapter.

duty cycle
>The percentage of time in which an operation is active.

EATA
>Extended AT Attachment - The hardware-level high performance protocol used by the RAIDmaster adapter to interface to the host computer. EATA supports bus mastering, scatter/gather data transfers, and overlapped and queued commands. Valid for CS31, CS34 and CS35 only.

ECC
>Error Correcting Code - A method of generating redundant information which can be used to detect and correct errors in stored or transmitted data. ECC is used in RAIDmaster adapters to detect and correct errors caused by cache RAM defects.

EDO DRAM
>Extended Data Output DRAM is a type of RAM that improves memory read performance on faster systems. The user of EDO RAM is required on CS37 controllers.

elevator sorting
>A method of sorting records or cache pages by physical location on disk so that the information may be written to disk with less seek and rotational latency.

failed
>The mode of operation of a drive or array in which the drive or array, because of a malfunction, can no

longer be accessed.

Fast SCSI

The original SCSI specification defined synchronous data transmission speeds of up to 5MHz. By assuming transceivers which provide tighter timing margins, the SCSI-2 standard allows synchronous transfers of up to 10MHz. This provides a transfer speed of 10MB/s for an 8-bit bus, and 20MB/s for a 16-bit bus. Devices which utilize these faster timings are called fast SCSI devices.

Fast-20 SCSI

The SCSI-3 specification defined synchronous data transmission rates up to 20MHz. The 20Mhz transfer rate provides speeds up to 20MB/s on an 8-bit bus and 40MB/s on a 16-bit bus. Devices that support these timings are called Fast-20 SCSI devices.

flash ROM

A ROM on the controller containing firmware that can be reprogrammed via software without removing it from the board.

flush

Copying cached data which has not yet been written to disk, to the disk drive.

function button

An object used in a graphical user interface, on which the user can point and click with a mouse to execute a specific function.

GUI

Graphical User Interface. A software interface which interacts with the user through color graphics and a mouse.

HBA

Host Bus Adapter - Refers to RAIDmaster or other peripheral controllers which provide host computer access to the SCSI bus.

HDM

Hardware Device Module - In I2O terminology, the part of the split device driver that resides within the firmware of the device.

hit

A data access in which the requested data is found in the cache.

hit ratio

The ratio of cache hits to total disk accesses. A hit ratio of 100% means that all disk accesses can be serviced from the cache.

hot plug

The operation of adding or removing a device from a bus while transactions involving other devices are occurring over the bus.

hot spare

A spare disk drive which, upon failure of a member of a redundant disk array, will automatically be used to replace the failed disk drive.

hot swap

The operation of removing a failed disk drive which is a member of a redundant array and replacing it with a good drive while transactions involving other devices are occurring over the bus.

I2O

Intelligent I/O architecture - An I/O platform that provides a standardized software interface between I/O devices and operating systems.

icon

A graphical representation of an object or function used by a GUI.

icon information window

A window, displayed by Storage Manager upon selection of an icon, which contains information and functions pertaining to that device.

ID

A number from 0 to 7 (or 0 to 15 for 16-bit Wide SCSI) which SCSI devices use to address one another. SCSI IDs are typically selected by setting jumpers on the SCSI device.

IRQ

Interrupt Request Line - A hardware circuit on which a device can signal an interrupt to the CPU.

latency

The time required by a device to access stored data, excluding the data transfer time. Reducing disk latency results in more I/O operations per second being performed on a disk drive.

LRU

Least Recently Used - A cache management algorithm employed by RAIDmaster to determine the next cache page to delete and re-use once all pages have been filled with data. By deleting the page in cache which has gone the longest without an access, the LRU algorithm typically results in the most frequently accessed data being cache-resident.

LSU

Logical Storage Unit - A logical device on which the computer can store and retrieve information. An LSU may represent an individual disk drive or an array.

LUN

Logical Unit Number - Each SCSI device may contain up to eight sub-devices or logical units, assigned logical unit numbers 0-7. Typically, SCSI devices such as a disk or tape drives contain only LUN 0.

LVD

Low Voltage Differential - A SCSI interface operating with a 3.3 V potential difference between two signal lines. This allows for equalization of magnetic and electrical noise affecting the SCSI bus.

LVD SCSI

Low Voltage Differential SCSI - Provides the benefits of differential SCSI - added reliability and long cable lengths - without the need for external high voltage bus transceivers. LVD SCSI enables synchronous transfers of up to 40 MHz and is also know as Fast-40 or Ultra2 SCSI.

MIB

Management Information Base - A standardized format for object databases that describe network devices. SNMP uses this description when monitoring these devices.

mirroring

A common term for RAID-1. A method of creating disk-fault tolerance by redundantly storing information on pairs of drives.

miss

A data access in which the requested data is not found in cache. Cache misses result in the normal average disk latency of 20ms when accessing data.

MTBF

Mean Time Between Failure. The average time between expected failures of a device.

multi-level RAID

A method of combining multiple RAID-1 or RAID-5 arrays into a single array, providing increased performance in multi-user environments.

non-redundant array

An array with no fault tolerance (RAID-0). If one drive in a non-redundant array fails, the entire array will fail.

NVRAM

Non-volatile Random Access Memory - Memory that stores data even when system power is turned off.

RAIDmaster controllers store their setup information in NVRAM.

optimal

The mode of operation of a disk array in which no drive failures have occurred.

OLR

On-Line Replacement - The ability to swap a defective hard drive while the computer is in operation.

OS

Operating System - Software which manages the computer's resources and provides the operating environment for application programs.

page

The smallest region of cache which may be allocated for storing data. The RAIDmaster adapter can contain up to 128,000 pages, each storing one 512-byte sector from disk.

parity

A method of generating redundant information which can be used to detect errors in stored or transmitted data. Parity is used in standard RAM SIMMs and over the SCSI bus to detect data errors. Parity is used in RAID-5 disk arrays to reconstruct flawed or missing data sectors.

PCI

Peripheral Component Interconnect - An intelligent computer bus specification which supports 32-bit data paths, 132   Mbyte/sec data transfers from bus mastering devices and plug-and-play automatic configuration of peripheral cards.

predictive caching

CS37 controllers actively analyze your system's disk requests using virtual cache memory, improving system performance by predicting which data is likely to be needed in the near future, then reading that data into the physical cache.

pull-down menu

An item at the top of a GUI display which, when selected by clicking with a mouse, will cause a list of related functions to be displayed.

RAID

Redundant Array of Inexpensive Disks - A method of combining hard disks into one logical storage unit which offers disk-fault tolerance and can operate at higher throughput levels than a single hard disk.

rebuild

The operation of recreating data belonging to a failed member of a redundant disk array and writing that data onto a replacement disk drive.

redundant array

A fault tolerant array (RAID-1 or RAID-5).

SAF-TE

SCSI Accessed Fault-Tolerant Enclosure - A specification (co-developed by nStor Corporation, Inc. and Intel Corporation) which provides a standardized method for monitoring and reporting on the condition of disk drives, power supplies and cooling systems used in high availability LAN servers and storage subsystem.

The specification is independent of hardware I/O cabling, operating systems, server platforms, and RAI implementation because the enclosure itself is treated as a device on the SCSI bus.

scatter/gather
A feature which allows data to be transferred to or from multiple non-contiguous areas of host computer memory with a single I/O command.

SCSI
Small Computer Systems Interface - an ANSI standard parallel interface designed to communicate with intelligent peripheral devices. The SCSI definition, created by ANSI, is defined in document number X3.131-1986.

SES
SCSI Enclosure Services - A SCSI-3 command set that provides a method to manage and sense the state of the power supplies, cooling devices, displays, indicators, individual drives, and other non-SCSI components installed in an enclosure.

SIMM
Single In-line Memory Module - A standard way of packaging RAM on a small circuit board with a defined edge connector. CS37 controller cache can use 16MB or 64MB, 32 or 36-bit, single-sided, low-profile 60ns EDO SIMMs. DPT also offers 16MB and 64MB ECC SIMMS. See also DIMMS.

single-ended
An electrical signal protocol which transmits information through changes in voltage. Single-ended SCSI uses standard TTL signal-and-ground pairs to transmit information over the SCSI bus.

S.M.A.R.T.
Self-Monitoring, Analysis and Reporting Technology - A standard for disk drives that monitor their own health and report potential problems.

SNMP
Simple Network Management Protocol - An industry-wide standard protocol designed to allow users to remotely manage the components in a computer network.

status packet
An area of host computer memory which is used to store the completion status for an I/O command for a peripheral controller which uses the EATA protocol. RAIDmaster adapters return command completion status in status packets.

stripe
A contiguous region of disk space. Stripes may be as small as one sector or may be composed of many contiguous sectors.

striping
Also called RAID-0. A method of distributing data evenly across all drives in an array by concatenating interleaved stripes from each drive.

synchronous
Data transmission protocol which is synchronized to a defined time interval. Synchronous SCSI can transmit data faster than asynchronous SCSI because the transmitting device does not wait for acknowledgment of each byte from the receiving device. Instead, it continues to transmit data at the rate negotiated by both devices (up to 10MHz).

tagged command queuing
A feature of SCSI-2 protocol which allows SCSI commands to be executed out of order.

target
A SCSI device, such as a disk drive, that receives and executes commands from a SCSI initiator device such as a DPT RAID controller.

termination
A method of matching the transmission impedance of a electrical bus so as to eliminate signal

reflections from the physical ends of the bus.

TERMPWR

A signal on the SCSI bus which provides power for SCSI bus terminators on remote devices.

throughput

A term used to describe the amount of data which can be processed by a system in a given amount of time.

tool bar

A bar, near the top of the screen used by a GUI, which contains function buttons.

Ultra SCSI

The SCSI-3 specification defined synchronous data transmission rates of up to 20MHz. The 20MHz transfer rate provides speeds up to 20MB/s on an 8-bit bus and 40MB/s on a 16-bit Wide SCSI bus. Devices that support these timings are called Ultra SCSI devices. Ultra SCSI is also called Fast-20 SCSI.

Ultra2 SCSI

The SPI-2 specification defined synchronous data transmission rates of up to 40MHz. The 40MHz transfer rate provides speeds up to 40MB/s on an 8-bit bus and 80MB/s on a 16-bit Wide SCSI bus. Devices that support these timings are called Ultra2 SCSI devices. Ultra2 SCSI is also called Fast-40 SCSI.

Wide SCSI

A SCSI protocol and signal definition which provides a greater than 8-bit wide data path. Wide SCSI devices may support either 8 and 16-bit, or 8, 16 and 32-bit data transfers.

WORM

Write Once Read Multiple - An optical storage device similar to a CD-ROM which can write data one time only to any location on a removable laser disk. WORMs are primarily used as data archive devices because once written, the data can never be erased.

write-back

A method postponing data to be written to a slow device such as a disk drive, by temporarily saving the data in a cache. The data can then be written at a time when the device would otherwise be idle. In addition, the data can be processed through techniques such as elevator sorting, so that the write operations will go faster. This method should only be used with a BBU.

write-through

The opposite of write-back. In write-through mode, data must be written to the final destination before a write operation is completed. Default setting for CS31, CS34, CS35 and CS37.

XOR

Exclusive OR - A logical operation performed on two binary operands which yields a 0 in every bit position when the operands are both 1 or both 0, and a 1 in every bit position when the operands are dissimilar. Performing XOR operations on every bit in a word generates a parity bit for that word.