Sun StorEdge™ T3 Array

A Technical White Paper

Copyright Information

Printed in the United States of America
901 San Antonio Rd., Palo Alto, California 94303 U.S.A.

Please
Recycle

Contents

Sun StorEdge™ T3 Array

Introduction

To accommodate customer storage needs in a volatile, information-driven marketplace, Sun introduces the Sun StorEdge™ T3 array with modular architecture that has been designed for three-dimensional scalability of capacity, bandwidth, and transaction rate. Fundamental to the architecture are the principles of simplicity, reliability, availability, diagnoseability, serviceability, manageability, and performance.

Figure 1. The Sun StorEdge T3 array.

Designed with flexibility in mind, the Sun StorEdge T3 array answers your storage capacity needs. Multiple Sun StorEdge T3 array units can be configured to serve large and growing storage demands and to achieve optimal performance in throughput—including both bandwidth and transactions. This storage system offers redundancy to strengthen reliability, availability, and serviceability within the workgroup or the data center environment. The Sun StorEdge T3 array includes array management software that allows a large number of arrays to be centrally administered. This capability allows an organization to grow its storage capacity without increasing administration cost and complexity.

Architecture Fundamentals

The Sun StorEdge T3 array's architecture begins with a basic *controller unit*. The standalone controller unit is the smallest possible array configuration. The architecture integrates disks, data cache, hardware RAID[1], power, cooling, uninterrupted power supply (UPS), diagnostic capabilities, and administration into a versatile, standalone component. The controller unit includes external connections to a data host (or hub or switch), and to a management network (see figure 2).

1 RAID (redundant array of independent disks) is widely used in the storage industry as a means to maintain data integrity even when a disk drive fails. This document assumes the reader has a general acquaintance with RAID fundamentals.

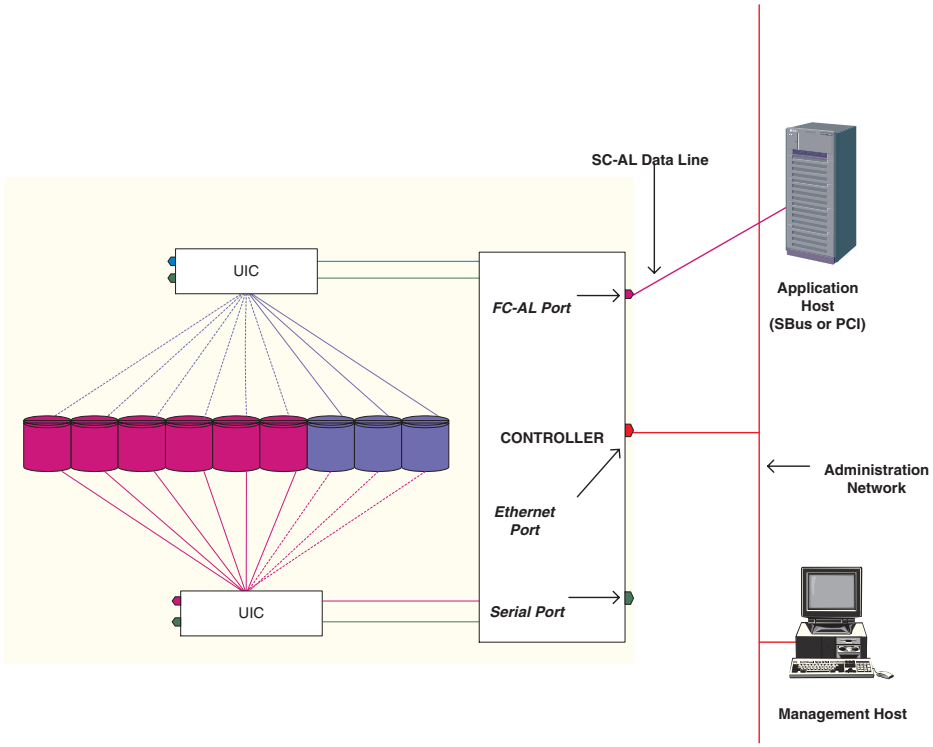

Figure 2. Logical view of the Sun StorEdge T3 array's controller unit, with separate data path connected to an application host and administration path connected to a management host.

The user can double storage capacity and spindle count[2] with an *expansion unit*. This is the same enclosure as a controller unit, less the RAID controller module (see figure 3). The expansion unit is connected to a controller unit in daisy-chain fashion with a pair of redundant unit interconnect cables.

The controller and expansion units are the basic building blocks used to create meaningful application configurations. Because the expansion unit lacks any external host connections, it can only be used in conjunction with a controller unit.

2 Transaction processing performance is often dependent on the number of disk drives, or spindles, available rather than on the total storage capacity available.

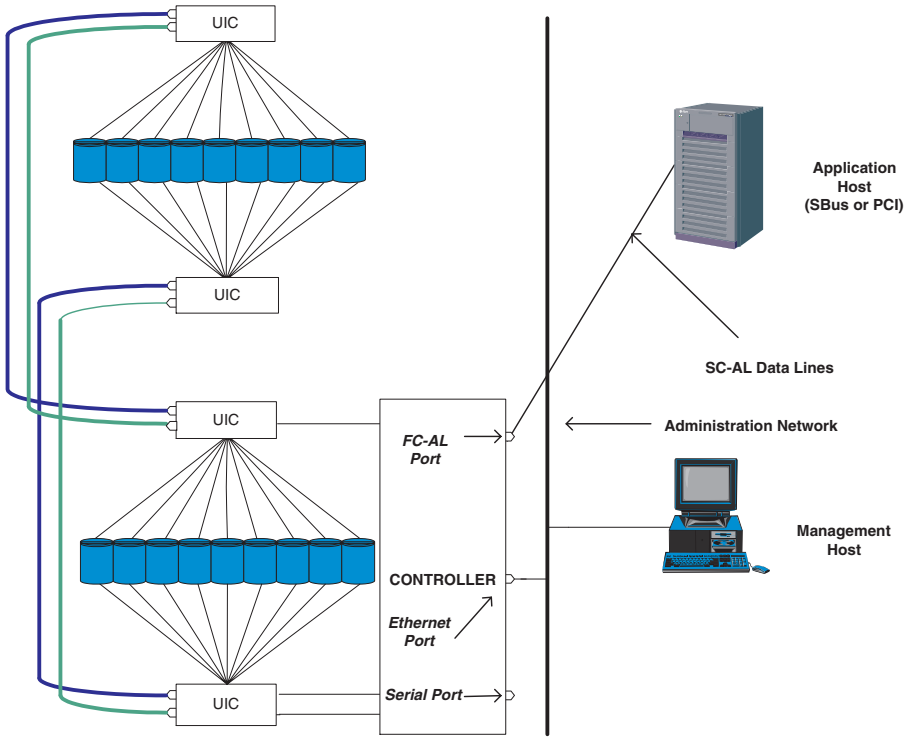

Figure 3. Logical view of the Sun StorEdge T3 controller unit plus expansion unit.

Partner Group Reliability

Two controller units may be paired in a *partner group* to create a configuration with redundant controllers and redundant data and management paths, allowing for cache mirroring, controller failover, and path failover capability (see figure 4). The partner group is thus the minimum storage configuration for enterprise environments that call for high availability. As with standalone controller units, partner groups may be configured with expansion units to double capacity and/or spindle count (see figure 5).

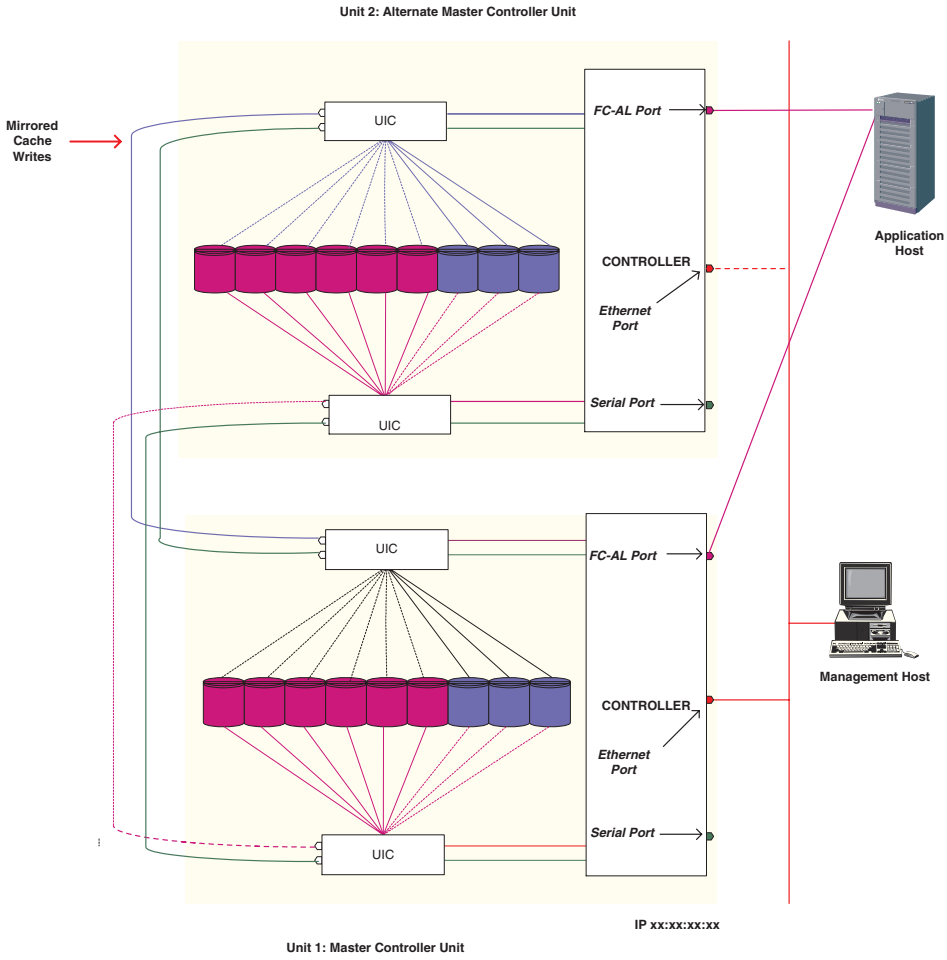

Figure 4. Logical diagram of a partner group with separate data and administration paths, with data paths connected to an application host and administration paths connected to a management host.

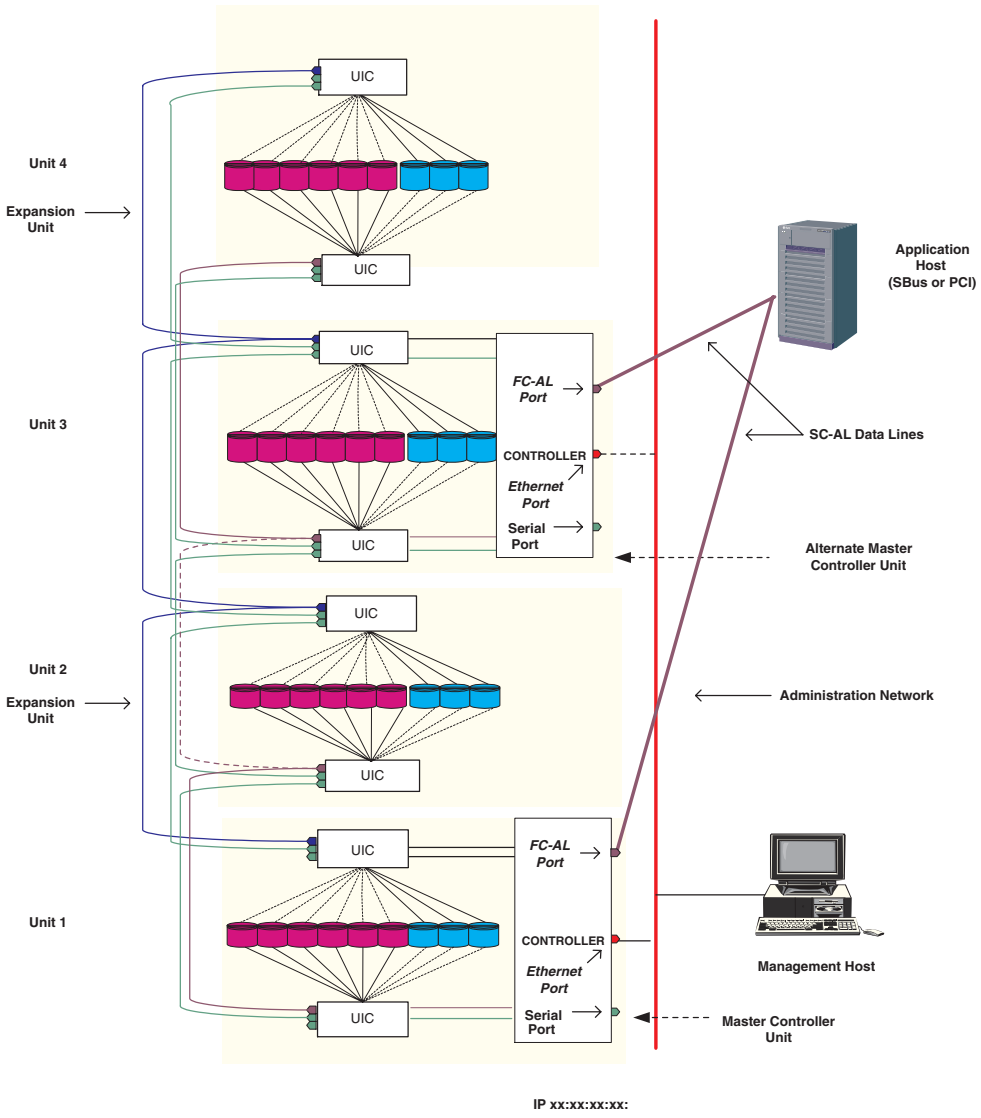

Figure 5. Sun StorEdge T3 array partner group with expansion units.

Separation of Data and Management Paths

The Sun StorEdge T3 array controller module includes three external interfaces. A Fibre Channel Arbitrated Loop (FC-AL) port transports data to the application host. An Ethernet port handles administrative (configuration, monitoring) communication with the management host. And, finally, an RS232 serial port is used for advanced service procedures, such as boot diagnostics.

Only application data travels across the FC-AL channel, and only administrative information moves across the network channel. This separation of responsibilities has several advantages. It enables greater reliability, because diagnostic reporting is preserved even when the host channel is down. It provides greater performance, because administrative traffic does not interfere with application I/O. It also provides greater security: a junior system administrator may be granted access to monitor and service the unit without access to the application server or even application data on the Sun StorEdge T3 array.

In addition, the separate administrative path enables greater efficiency and productivity in the data center by allowing for centralization of administration. A site may have multiple, heterogeneous, geographically distributed application servers with local Sun StorEdge T3 arrays. All Sun StorEdge T3 arrays can be connected via Ethernet and TCP/IP to a single management server, which provides centralized administration with a single user interface.

3-D Scalability

The Sun StorEdge T3 array's architectural design provides for 3-D scalability. Controller and/or controller/expansion unit configurations can be added to meet requirements in capacity, bandwidth, and transaction rate as business requirements grow. This scalability and flexibility protects the original investment, and it allows the customer to "pay as you grow." Sun StorEdge T3 arrays simply can be added to the existing storage infrastructure as business requirements change. In addition, the centralized administration capability of the Sun StorEdge T3 array answers increasing storage needs without adding management complexity.

Capacity Increments

The Sun StorEdge T3 array's architecture allows storage to grow with an application (see figure 6). Add partner group configurations as needed; add hubs to connect multiple partner groups to one pair of host adaptors or multiple hosts; add switches to create a storage area network (SAN). Capacity ranges from a minimum configuration of 163 gigabytes (GB) in a single unit (using 18-GB drives) up to 2.6 terabytes (TB) in a rack of two four-unit partner groups (using 36 GB disks).

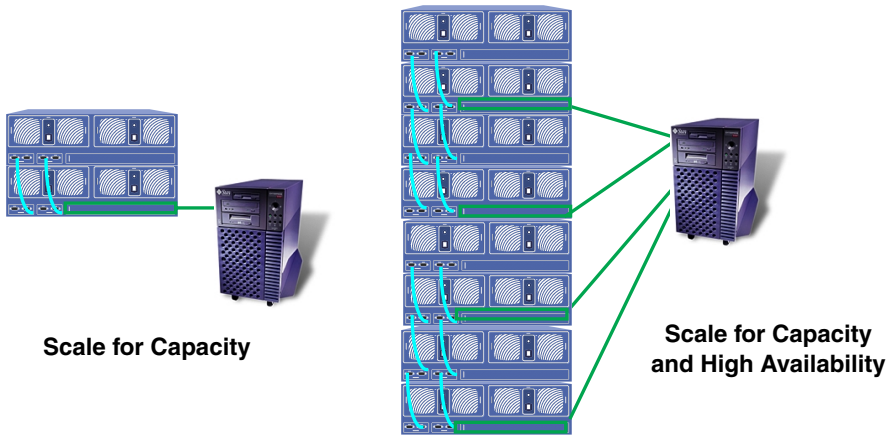

Figure 6. Scale for capacity with an expansion unit, or create a rack configuration of two partner groups, each with two controller units and two expansion units.

Throughput Increments

Applications with high bandwidth requirements, such as image processing, data mining, decision support (DSS) and high performance computing (HPC), typically consume multiple host channels to transport as many MB/sec of data as possible. Random I/O environments such as online transaction processing (OLTP) require storage with low latency and the ability to process as many I/O operations per second (IOPS) as possible. Capable of serving both environments, the Sun StorEdge T3 array's architecture allows customers to

add controllers to a configuration (up to one per disk tray), thus scaling the back-end processing power as needed to meet the application I/O requirements (see figure 7).

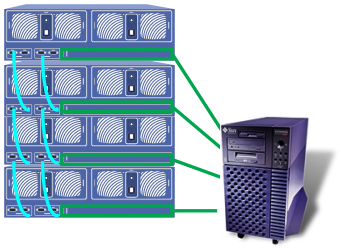

Figure 7. Scale bandwidth using four controller units and four host connections.

Spindle Increments (Scaling IOPS)

OLTP applications are often limited not by back-end processing power but by the number of spindles rotating to accept or deliver data. Because the Sun StorEdge T3 array's flexibility allows customers to order configurations, including expansion units (and thus including more spindles), the customer may adjust the disk-to-controller ratio higher to keep an application from becoming spindle-bound. The Sun StorEdge T3 array also can support a wide range of disk sizes, which allows the customer even greater flexibility to adjust capacity along with spindle count (see figure 8).

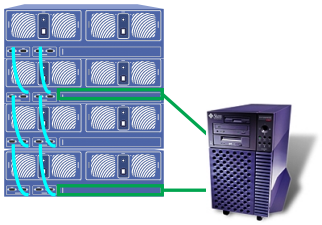

Figure 8. Customers can scale IOPS to serve a traditional OLTP environment using a partner group with two expansion units along with two controller units, which doubles the number of spindles per host channel.

Hardware Architecture

Reliability and Serviceability via FRUs

In the Sun StorEdge T3 array, all active components are designed to be N+1 redundant, including disks, power supply, fans, and UPS. On the back end, loops, loop switching, diagnostics, and administration channels are also redundant. When configured in a partner group, even controllers, host channels, and external administration channels are redundant.

Active components are consolidated into four types of FRUs: disk drive, power/cooling unit (PCU), unit interconnect card (UIC), and controller. All FRUs are hot swappable to prevent servicing downtime and minimize mean time to repair (MTTR).

FRU Replacement

To adhere to the Sun StorEdge T3 array's promise of ease of use, all FRUs in this Sun™ array are designed for replacement with the same short series of steps:

- Use the provided tool (or a coin or other thin, rigid object) to depress a latch. The latch will pop out to reveal a tab with a large hole.
- Insert a finger into the tab and pull out the FRU.
- Insert the replacement component, making sure it is firmly seated, and press the pull-out tab into position.

Larger, heavier FRUs include two tabs. Each type of FRU is a unique size to help ensure the FRU is inserted in the proper location. Moreover, the chassis and all FRUs include guides to promote correct insertion of each FRU.

FRU Identification

All FRUs include a FRU ID EEPROM. Each FRU ID includes read-only information, such as part number, revision level, date of manufacture, and serial number. Some FRUs also contain writable information, such as firmware version and battery warranty status.

FRU IDs improve serviceability by enabling hardware and firmware version checking through software to determine whether these are consistent and at current, supported revision levels. FRU IDs also enable configuration information to be transmitted to service personnel as part of monitoring alerts.

FRU Descriptions

Disks

Nine dual-ported fibre channel disks reside within a single Sun StorEdge T3 array (see figure 9). The disk drives are concealed by a removable front bezel that provides electromagnetic interference (EMI) shielding. Each disk FRU consists of a custom enclosure that holds either a half-height (1.6-inch) or low profile (1-inch) disk drive (refer to the Sun StorEdge T3 array data sheet for current drive capacities). Individual disk drives are not visible to the application host; rather they are configured into one or two RAID 5, RAID 1, or RAID 0 logical volumes. The ninth disk may optionally be configured as a standby drive (also known as a hot spare drive).

Each drive has a private region of 200 MB reserved for system use. All remaining capacity is available for use by the application host. On the master and alternate master controller units, the system area is used on all drives as a nine-way mirror, containing a copy of the operating system, file system, and firmware. Multiple versions of firmware may be saved, allowing the flexibility to back out or revert to an earlier version if necessary. On the master controller unit, the system area also includes configuration information, system log, and other assorted files for internal use.

The system area of expansion units is empty, as it is reserved for future use. If an expansion unit should be upgraded to a controller unit at some future time, the system area then may be populated without affecting existing data.

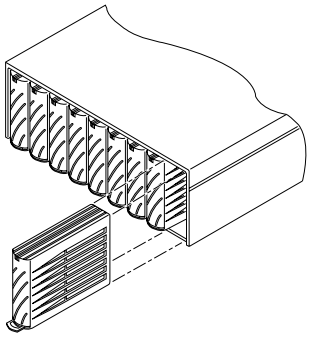

Figure 9. Front profile of the Sun StorEdge T3 array, depicting the array's disk drives.

Power Cooling Unit (PCU)

Each Sun StorEdge T3 array's tray includes two redundant power and cooling units (see figure 10). Each has an external power connection, allowing for connection to two independent power grids. There is one internal 325-watt auto-switching power supply per PCU. In case of external power failure or PCU failure, one power supply is sufficient to power the Sun StorEdge T3 array indefinitely.

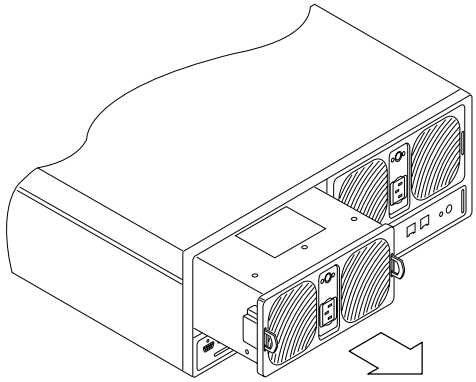

Figure 10. Sun StorEdge T3 array's rear view, highlighting the PCU.

Each PCU includes two cooling fans. The fans are N+1 redundant, which means that three of the four fans are sufficient to cool the Sun StorEdge T3 array. Each fan draws its power from the Sun StorEdge T3 array's midplane. Therefore, even if the internal power supply in a PCU should fail, the fans in

that PCU will continue to operate. In the situation where a PCU must be replaced, the remaining two fans provide sufficient cooling for at least 30 minutes. Since MTTR for a PCU is less than one minute, this provides ample margin for FRU replacement. In the case of a power supply failure, the FRU should be left in place until its replacement is available and ready to be swapped in, so that the FRU's fans can continue to provide cooling.

Each PCU includes a battery that functions as a UPS. If all external power to the Sun StorEdge T3 array should fail, the single battery is sufficient to keep the entire Sun StorEdge T3 array running long enough for write data to be flushed from cache to disk, and an orderly shutdown of the array to be accomplished. Two batteries are provided for redundancy, one per PCU. The batteries are periodically cycled to help ensure proper functioning if and when needed.

Controller Card

The controller card provides cache, RAID management, administration, diagnostics, and external interfaces (see figure 11). Controller units include one controller FRU. Two controller units are paired in a partner group for cache mirroring and controller redundancy. Expansion units, which are used to add disks to a controller unit, have no controller FRU.

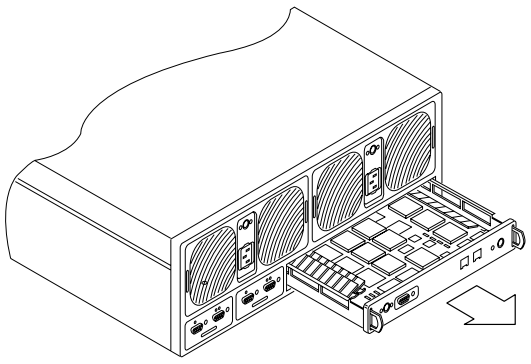

Figure 11. Sun StorEdge T3 array's rear view, highlighting the controller card.

Unit Interconnect Card (UIC)

The unit interconnect card includes back-end loop-switching hardware, diagnostic state registers, and proprietary external ports used to link multiple Sun StorEdge T3 arrays in a linear, array-to-array or "daisy-chain" fashion (see figure 12). There are two redundant unit interconnect cards per Sun StorEdge T3 array.

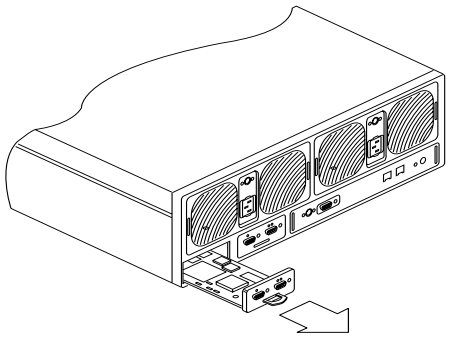

Figure 12. Sun StorEdge T3 array's rear view, highlighting the UIC.

Sun StorEdge T3 Array Chassis

All FRUs connect to a midplane within the Sun StorEdge T3 array. The midplane is integral to the Sun StorEdge T3 array's chassis, and it has no active components other than an FRU ID EEPROM and temperature sensors. The midplane/chassis can be replaced, although it should be a rare occurrence in which substitution is required. The midplane/chassis has a mean time between failures (MTBF) of 7.8 million hours, making replacement due to externally inflicted damage far more likely than internal midplane failure. Chassis replacement can be accomplished by disconnecting all external cables, removing all FRUs, replacing the chassis, and then replacing the FRUs and reconnecting the cables.

Cabling

The external ports on the Sun StorEdge T3 array's controller use standard cables with standard connectors:

- The FC-AL port is a standard DB9 copper connector. Each controller is shipped with a media interface adapter (MIA), which converts the port from copper to short-wave Fibre.

- The Ethernet port is a standard RJ45 10BaseT connector providing 10 megabit per second (Mbps) connectivity.
- The serial port is a standard RS232 connection via RJ11. Note that use of the serial port is restricted to advanced diagnostics by trained service personnel.
- The UICs use a standard DB9 connector, but with non-standard signalling. This allows the cable to carry both fibre channel (for application data) and serial (for monitoring and configuration) protocols to pass through a single connector and a single proprietary cable. The result is simplified, daisy-chain cabling between units, and higher reliability due to fewer cables and cable connections.

Controller Architecture

The controller is both the data processing and administrative ”brain” of the Sun StorEdge T3 array. It provides all the Sun StorEdge T3 array’s external interfaces and controls all back-end activities, whether they be related to data management or administration (see figure 13). The controller’s data host interface is a Qlogic 2100 fibre channel (FC-AL) interface ASIC. It connects to a 64-bit, 33-MHz PCI bus, which functions as the backbone of the Sun StorEdge T3 array. Also residing on the PCI bus is 256 MB of SDRAM cache, with a proprietary inline FPGA XOR engine that has 2 MB of VRAM. Two more Qlogic 2100s provide the interfaces to two back-end FC-AL loops. Finally, there is a bridge chip on the backbone, providing a transition to a 32-bit, 33-MHz PCI administration bus to the controller CPU.

The controller has an administration bus, which connects FLASH PROM/FRU ID, external 10BaseT Ethernet port, and serial line interface to an external RS232 serial port. The external ports use standard connectors: DB9 for FC-AL, RJ45 for Ethernet, and RJ11 for the serial port. Connector pinout specifications are provided in the product documentation. The administration bus also connects to two internal serial lines to the unit interconnect cards.

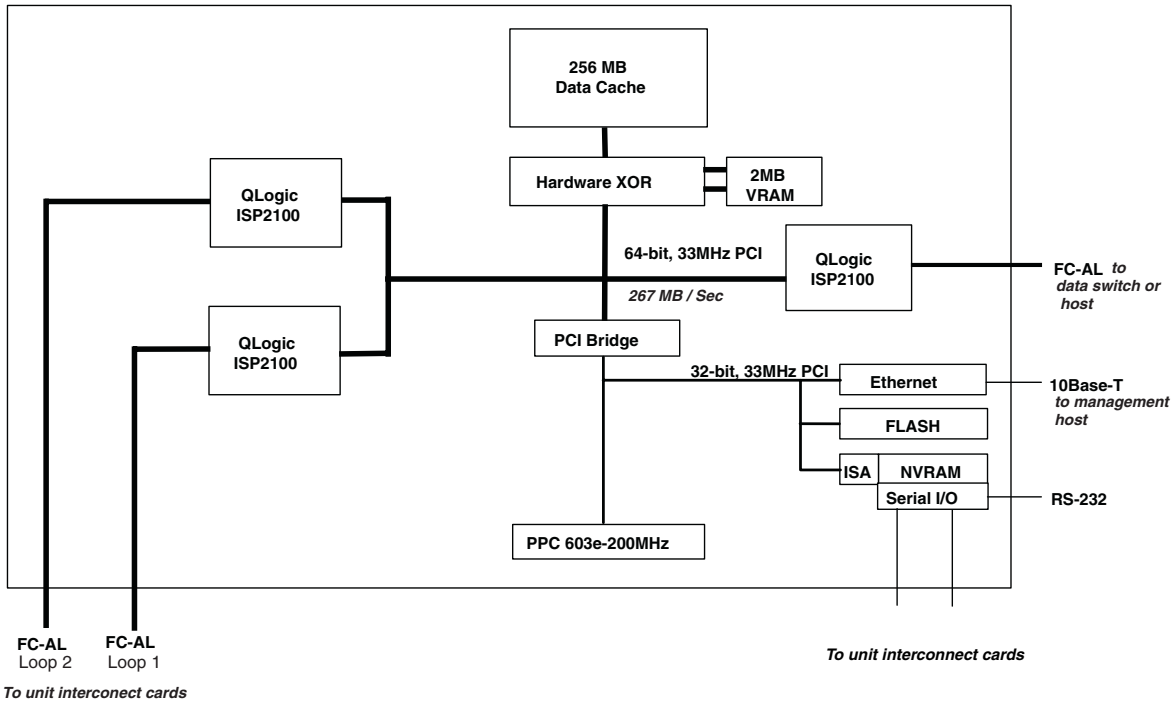

Figure 13. Controller architecture.

The controller CPU is a PowerPC 603E running at 200 MHz. Note that the CPU is not on the system bus; its involvement with application data is limited to managing the data, not manipulating it. Although the CPU controls DMA transfer of data between host interface FC-AL ASIC and cache, and between cache and back-end FC-AL ASICs, data never travels through the CPU itself. Even XOR parity calculations are performed not by the CPU but rather by the inline XOR engine, as data moves in and out of cache. This efficient data flow is a key factor in achieving superior RAID 5 performance in the Sun StorEdge T3 array.

Because the Ethernet and serial lines are also isolated from the controller backbone, it is not possible to transfer data through the external Ethernet or serial ports. They are available exclusively for administration, just as the external FC-AL loop is available exclusively for application I/O.

Unit Interconnect Card Architecture

Unit Interconnect Cards

The unit interconnect card (UIC) has three inter-related functions: join units in a daisy chain, perform back-end loop switching, and complete diagnostics. There are two UICs per Sun StorEdge T3 array, one for each back-end loop.

Considering its multi-functional duties, the UIC architecture is deceptively simple (see figure 14). Loop resiliency circuitry (LRC) provides internal fibre channel switching and bypass capability, connecting all back-end fibre channel components. This includes all nine drives, one of the back-end Qlogic 2100 fibre channel interfaces on the controller board, and one pair of ports on the Unit Interconnect Card, used to join units together in a daisy chain. The LRC also provides fibre channel clock regeneration, crucial to preventing signal jitter.

The remainder of the UIC consists of an 8051 CPU connected to FLASH PROM FRU ID, control/sense registers, and a dual UART. The CPU is connected to the LRC through the registers, which collect and maintain component status and diagnostic states. The administration serial line from the controller CPU connects through the midplane to the UIC CPU, and then to the dual UART, which extends one serial line to each of the two back-end external ports.

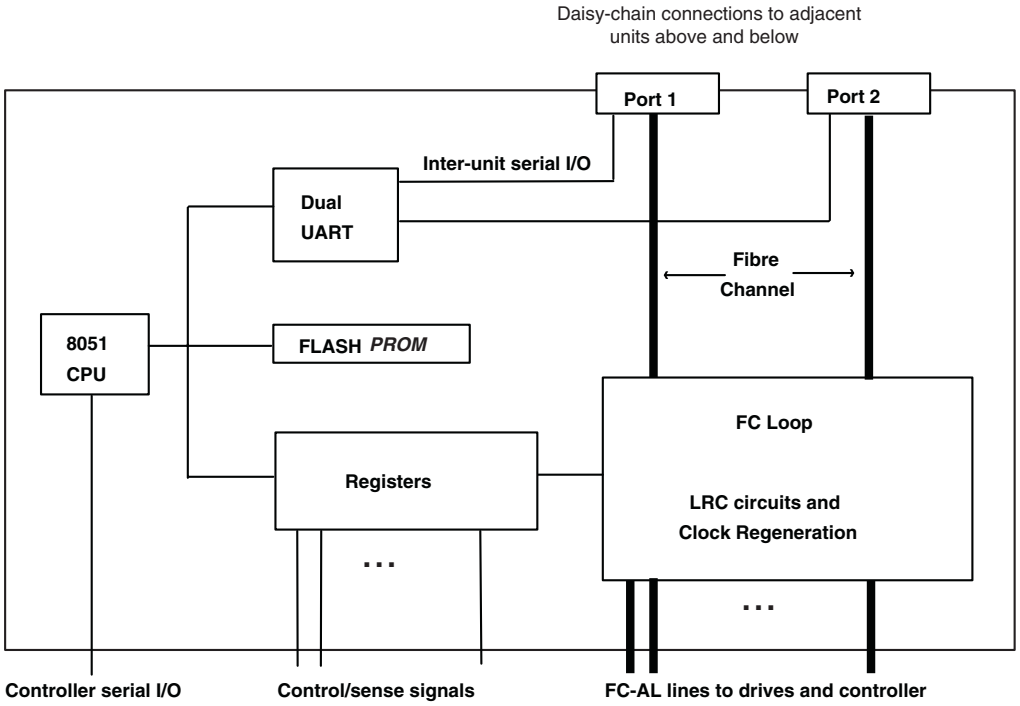

Figure 14. Unit Interconnect Card Architecture.

The LRC switching capability gives the Sun StorEdge T3 array unique back-end reliability, availability, and serviceability (RAS) capabilities. In normal operation, each UIC selectively enables a subset of its components for purposes of load balancing, cache mirroring, and redundancy. First, consider normal operation (see figure 15).

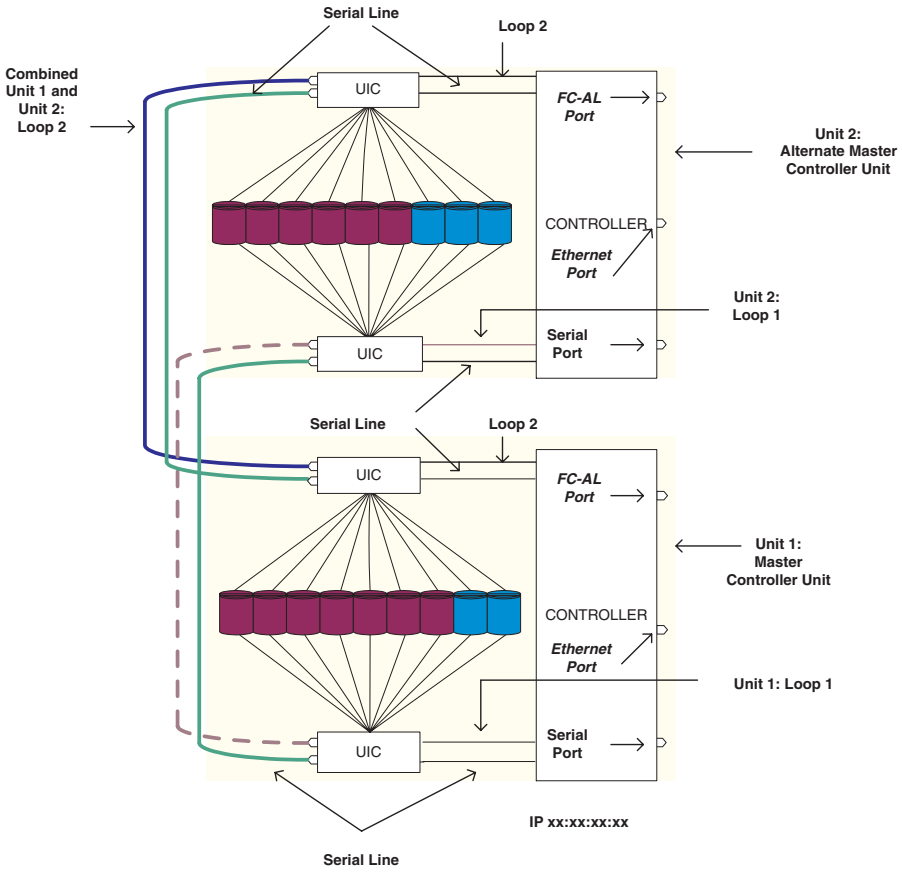

Figure 15. Sun StorEdge T3 array partner group configuration. Loop components are enabled with each UIC in normal operation.

The Sun StorEdge T3 array's Unit 1, Loop 1 UIC enables the Loop 1 back-end controller ASIC and drives 4 through 9. Any I/O requests to drives 4 through 9 on Unit 1 travel through this loop. Note that the Loop 1 back-end daisy-chain ports are not enabled on the loop. Therefore, Loop 1 in Unit 1 and Loop 1 in Unit 2 are maintained as two independent loops.

The Loop 2 UIC in Unit 1 enables the Loop 2 back-end controller ASIC, drives 1 through 3, and the Loop 2 back-end daisy chain ports. The same applies to Unit 2. Because the daisy chain ports are enabled, there is just one Loop 2 for the entire partner group. So a single Loop 2 spans both units: from the Unit 1 Loop 2 ASIC to its Loop 2 UIC (including drives 1-3), to the Unit 1 Loop 2 back-end daisy-chain ports, to the Unit 2 Loop 2 back-end daisy-chain ports, to its Loop 2 UIC (including drives 1-3), to the Unit 2 Loop 2 controller ASIC.

This back-end configuration of three loops for the partner group provides static load balancing in normal operation. Two of the loops, one in each unit, each carry the load for six drives. The third loop, which spans both units, also carries the load for six drives (three in each unit), plus any mirrored cache writes.

Administration

The LRC switching function also provides unique and powerful diagnostic capabilities. In case of failure on a loop, a diagnostic routine can be run in which the UIC systematically switches components in and out of the loop until the offending component has been identified. Then, the UIC can switch off, or bypass, that component until it is replaced. By fencing off the failed component from the loop, the loop can be restored to active use.

Unit Interconnect Cable

In addition to the fibre channel port used to link the data path between units, the UIC includes a serial port to link the administrative path between units. The fibre channel and serial ports are combined into a single non-standard physical connector. A single non-standard cable, called the unit interconnect cable, combines the fibre channel and serial lines that link the Sun StorEdge T3 arrays.

Data Flow

The application data path is used to process I/O between the application host and disks exclusively. No configuration or monitoring is performed over the data path to the host, other than normal SCSI inquiry requests. No configuration or monitoring is handled over the internal data paths, other than to store/retrieve configuration and monitoring data to/from the reserved system area on the disk drives.

Data movement is by DMA to and from the Qlogic 2100 ASICs. All data goes through the cache and inline XOR engine. Because all XOR operations are completed as data moves in and out of cache, there is virtually no performance penalty for calculating RAID 5 parity on the Sun StorEdge T3 array. When write-behind mode is enabled, host writes are acknowledged when they reach cache, and are later destaged to disk. When cache mirroring is enabled in a partner group configuration, host writes are acknowledged only after they both reach cache and are copied to the partner controller's cache. Write data is later destaged to disk according to cache destage rules, based on idle time, utilization, and error conditions.

Within the Sun StorEdge T3 array, all data travels through cache. The system uses static load balancing to spread I/O across the two back-end loops. Data destined for drives 4-9 is sent through Loop 1, while cache mirroring data, plus data destined for drives 1-3, is sent through Loop 2. Should one loop become disabled, the surviving loop handles the full back-end load.

Standard Data Flow

A simplified version of the Sun StorEdge T3 array's read and write data flow patterns are as follows:

Read Data

- I/O request from host to front-end FC-AL ASIC
- I/O request from FC-AL ASIC to CPU
- If not a cache hit, request data from disk(s), transfer data from disk(s) to back-end ASIC(s); DMA data from back-end ASICs to cache
- DMA data from cache to front-end FC-AL ASIC
- Data from front-end ASIC to host

Write Data (write-behind and cache mirroring enabled)

- I/O request from host to front-end FC-AL ASIC
- I/O request from FC-AL ASIC to CPU
- I/O received by front-end ASIC from host
- DMA from front-end ASIC to cache

- DMA from front-end ASIC to back-end Loop 2 ASIC; DMA from back-end ASIC to back-end Loop 2 ASIC on partner unit; DMA from ASIC to cache in partner unit
- Once data has reached both caches, return ACK to host
- Destage data from cache to disk (see cache section for details)

In the case where write-behind cache is not enabled, data will be written from cache to disk(s) before acknowledgment is sent to the host.

Maintaining Data Availability

Redundant Controller Units

To achieve high availability using the Sun StorEdge T3 array, two controller units (plus any associated expansion units) are configured in a partner group (see figure 4). Although united in a partner group, each controller unit processes data to disk independently. However, write data that is placed in cache to be destaged at a later time is mirrored to the partner controller unit cache prior to returning an ACK through the host interface. The mirroring is accomplished over one of the two back-end FC-AL loops. Under normal operation, Loop 1 in each controller unit remains independent of the partner's Loop 1. Loop 2, which provides the cache mirroring path, is a continuous loop between the two partners.

Redundant Host Data Paths

Each controller has its own data path to the application host (or hub or switch). In normal operation, each path to a given controller carries only data for the volumes in the same StorEdge T3 array as that controller i.e., the path serves as the *active* or *primary* path for those volumes. However, each controller and path is also capable of carrying data intended for its partner controller and path, i.e., the path also serves as the *secondary* or *passive* path for its partner if necessary.

Failover Strategies

The failover scheme for the Sun StorEdge T3 array distinguishes between two types of failures: path failures and controller failures. The failover approaches for both failure types have much in common, but they also exhibit some distinct differences.

Path failure

Path failure occurs when I/O to a unit is interrupted for any reason except controller failure. The failure could be in a cable, media interface adapter (MIA), host adapter, or even a non-I/O root cause such as removal of an application host system I/O board. Regardless of the cause of the interruption, I/O requests targeted at a LUN will eventually time out. The I/Os are then redirected to the alternate path for that LUN, namely the path to the other Sun StorEdge T3 array in the partner group. The redirection is managed on the application host by the alternate pathing software appropriate to that host. (On Solaris™ platforms, the user can choose between Solaris Operating Environment AP (alternate pathing) and VERITAS DMP (dynamic multipathing). On other platforms, the user can choose between VERITAS DMP and the native failover driver available with the Sun StorEdge T3 array.)

When a controller receives an I/O request targeted at a LUN that belongs to its partner controller, it verifies that its partner controller is healthy, and then takes over control of the LUN. This procedure is called a LUN failover. The back-end connection between Loop 1 of Unit 1 and Loop 1 of Unit 2 is healed, so that there is now a single Loop 1 for the entire partner group, as well a single Loop 2. (Recall that there is already a single Loop 2 in normal operation, which provides a path for cache mirroring). I/O targeted at the partner's LUN is directed across the back end (Loop 1 or Loop 2, as appropriate) and into the proper drives.

In a path failure scenario, both controllers remain healthy. So, if write-behind cache was enabled, it remains enabled. Writes go into the controller with the live path, are written into the local cache, copied via Loop 2 into the partner's cache, acknowledged to the application host, and, in due course, destaged to disk across the back-end channels (see figure 16).

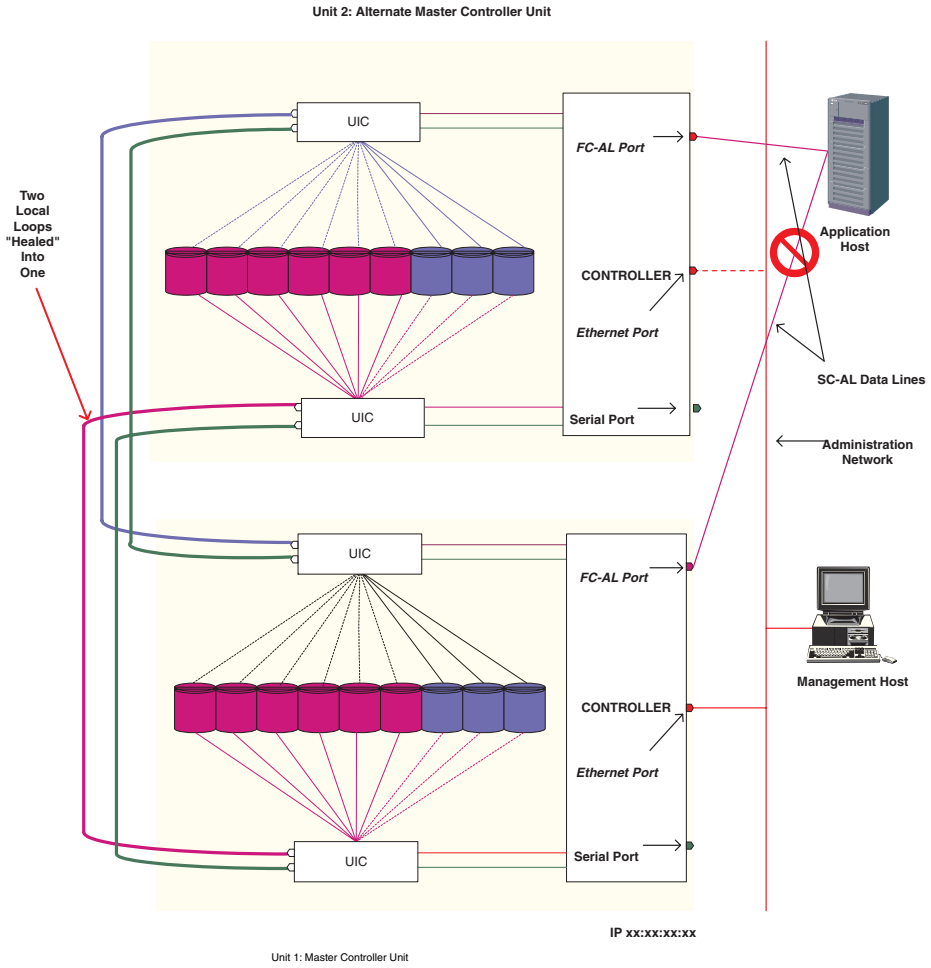

Figure 16. Illustration of path failure to Unit 1.

Even when a failure occurs on the data path to the master controller, administrative procedures continue unchanged. The controller and administrative path remain healthy, so the master controller continues administration even though it may have temporarily ceased performing data I/O operations.

Also note that no special communication, or “heartbeat” is needed between the application host and partner group to initiate failover or failback. The occurrence of I/O down the alternate path (except for SCSI inquiry or

read/write of block 0 of any LUN) automatically triggers failover. Likewise, resumption of I/O down the primary path (except for SCSI inquiry or read/write of block 0 of any LUN) automatically triggers failback. It is the responsibility of the alternate pathing software to "ping" the primary path periodically (by sending a SCSI inquiry or read/write of block 0 of any LUN), to see if it has been restored.

Controller Failure

To the application host, a controller failure appears identical to a path failure, and the response and recovery procedures are also identical. I/O requests down one channel time out. The host-based alternate pathing software re-routes I/O down the failover channel. The software periodically pings the primary channel, and when it gets a response, re-routes I/O back to the primary channel. The only difference is that the time it takes to effect a controller failover will be slightly longer than that needed to effect a LUN failover.

On the Sun StorEdge T3 array, the path failover resulting from controller failure causes LUN failover, as with any other path failover. But there the similarity ends. Loss of communication heartbeat informs the surviving controller that its partner controller has failed. The surviving controller takes significant additional recovery actions on both the data and administration fronts.

Write-behind cache is disabled, as well as cache mirroring. Back-end Loop 1 on both partners are joined into a single loop, just like Loop 2 (see figure 17). Any uncommitted write data in the surviving cache is flushed to disk, including mirrored uncommitted write data destined for LUNs of the failed controller.

If the alternate master controller has failed, then no administrative changes are needed. If the master controller has failed, then the alternate master controller must take over the role of the master controller. The alternate master controller takes on the MAC and IP addresses of the master controller, as well as the host name, activates its dormant Ethernet connection, and resumes IP activity on its administration path. As far as the network and any management consoles are concerned, nothing about the network topology has changed; the alternate master controller, for all intents and appearances, is now the master controller (see figure 17).

Because the MAC and IP addresses for the partner group have not changed, there is no need to change routing tables and maps.

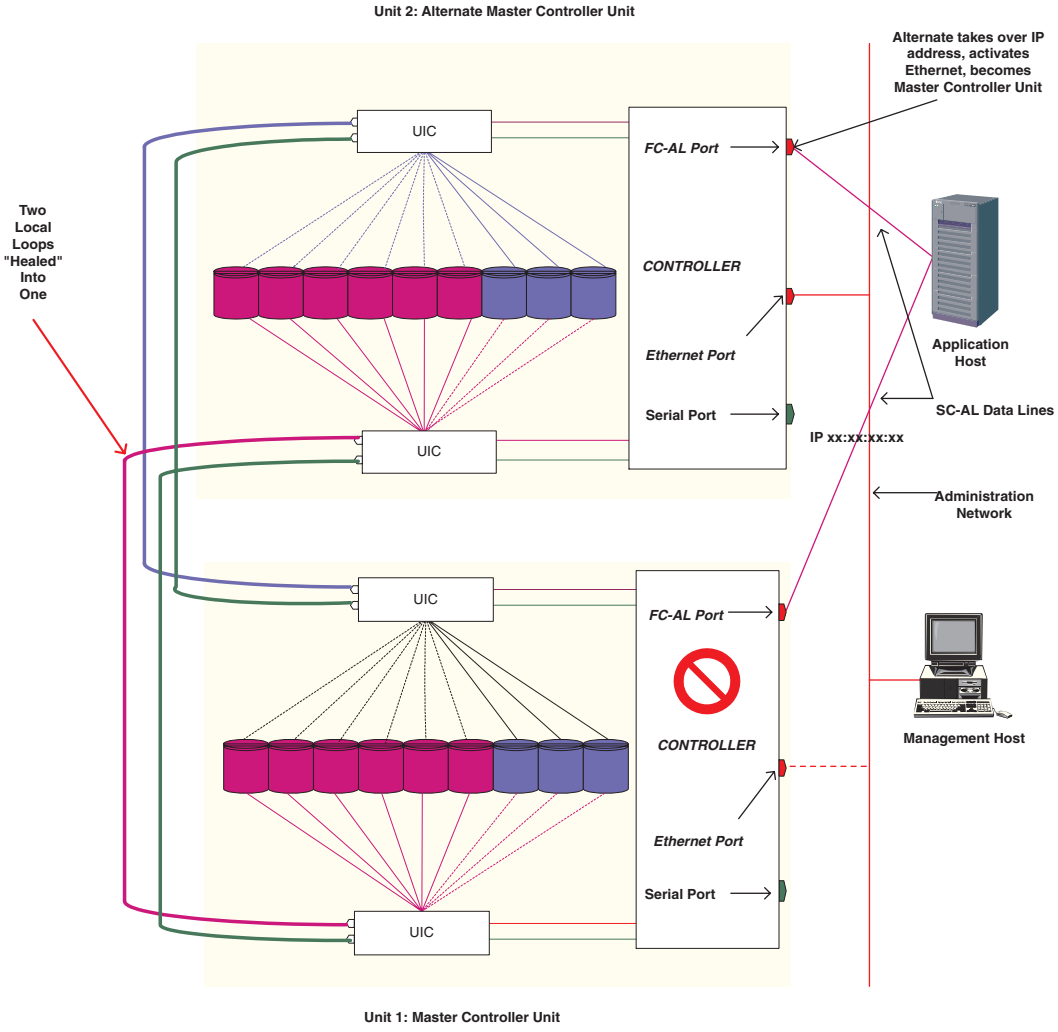

Figure 17. Sun StorEdge T3 array partner group with one failed controller. The alternate master controller takes over as master controller, and two loops are “healed” into one.

When a failed Sun StorEdge T3 array controller is replaced, insertion of the replacement controller is automatically detected, the controller is booted, and the unit’s heartbeat is restored. LUN failback is achieved the same manner as non-controller path failover: when the host pings the primary path and

receives a response, it re-routes I/O back to the primary path. However, if the controller failure resulted in administrative controller failover, controller replacement does not cause administrative failback. The former alternate master controller continues to act as the new master controller, rather than suffer the overhead of the administrative failback. The former alternate master controller will continue to act as master controller until a system reset or power cycle, or failure of that controller. In any of these cases, the controller in Unit 1 is restored as master controller.

Note that even when the alternate master controller takes over as master controller, the populated system area continues to be a nine-way mirror on Unit 1. This means the system boots from a firmware image on the drives of Unit 1, and the syslog continues to be written on Unit 1 drives.

Data Cache

To understand how the Sun StorEdge T3 array processes data through the cache, it is helpful to be familiar with both common cache terminology and some terms unique to the T3 array. Before reading this section, the reader may wish to refer to the cache glossary found at the end of this document.

The primary purpose of the data cache in the Sun StorEdge T3 array is to provide a low latency buffer for write data, allowing writes to be quickly acknowledged to the application host. The cache is especially crucial to RAID 5 write performance, because it can coalesce several partial-stripe writes into a single read/modify/write operation. A secondary benefit of the cache is to buffer read data, allowing for low latency on repeated reads of the same data.

Adaptive cache is a key feature of the Sun StorEdge T3 array. The algorithms used for allocating, coalescing, and flushing data are automatically and dynamically adjusted based on I/O patterns. This limits the amount of cache configuration needed to be performed by the user, thus greatly simplifying administration, improving ease of use, and enabling optimal cache behavior for current I/O patterns.

Each Sun StorEdge T3 array controller includes 256 MB SDRAM data cache. Cache organization and behavior are tightly coupled with LUN stripe width and Sun StorEdge T3 array block size (the amount of data in the stripe that goes on each disk). The Sun StorEdge T3 block size is a system configuration parameter set by the user to be 16 KB, 32 KB, or 64 KB. The cache buffer size equals the block size. Therefore, the block size configuration parameter defines

both the size of the cache buffers and the unit of data written to each disk in a RAID stripe. Because the cache size is fixed at 256 MB, the number of cache buffers varies. There are 16384, 8192, or 4096 cache buffers, depending on block size of 16 KB, 32 KB, or 64 KB, respectively (see figure 18).

Each cache block is composed of eight segments. This means that segment size is 2 KB, 4 KB, or 8 KB, for block size of 16 KB, 32 KB, or 64 KB, respectively. Segmentation of the cache block is crucial to performance of the adaptive cache, because the segment size defines the Sun StorEdge T3 array's atomic I/O to disk. This means that not only is it possible to optimize for partial stripe reads and writes, but it is also possible to optimize for partial block reads and writes.

Note that the host I/O size is not necessarily the same as the Sun StorEdge T3 array block size. As noted below, there are cases where optimal performance will be achieved when the segment size matches the host I/O size or where the stripe size matches the host I/O size.

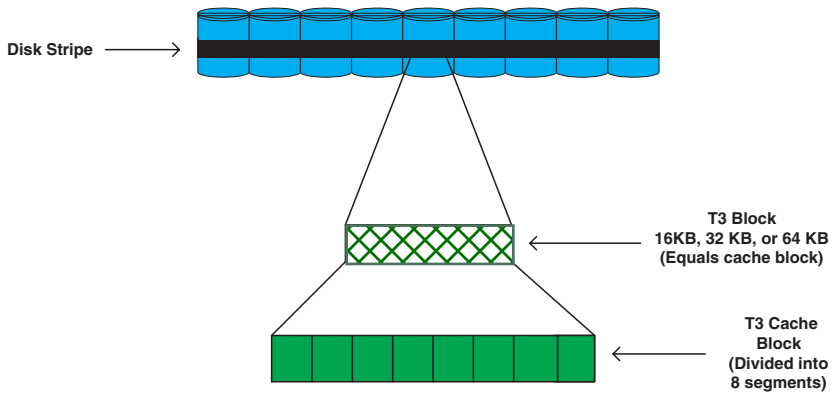

Figure 18. The Sun StorEdge T3 array cache block includes eight segments—each 2 KB, 4 KB, or 8 KB, depending on block size.

Read Cache Versus Write Cache

All data travels through cache, and all data in the cache is read cache. Therefore, all data—whether read or written into cache, whether dirty or clean—is available for subsequent host read requests as a cache hit. Dirty write

data is limited to 80 percent of cache, to keep some cache space available for subsequent writes and for read requests that are cache misses. Writes are cached only when write-behind mode is enabled. Dirty write data will be flushed to disk under the following conditions:

- Demand flushing. When the dirty write data 80 percent threshold has been reached, the CPU will cause as many as 20 stripes of dirty data to be flushed to disk. The least recently used stripes are chosen to be flushed.
- Idle time flushing. If no host requests are received for a full second, one stripe is flushed. After 10 milliseconds (ms) with no host requests, two stripes are flushed. After another 10 ms, four stripes are flushed, and so on. This continues up to a maximum of 128 stripes per flush, until either the cache is emptied of dirty write data or a host command is received. The stripes chosen for each flush are those least recently used.
- LUN flushing. If a volume is unmounted by the user, any dirty cache associated with that volume is flushed. All host commands are queued while the LUN data is flushed.
- Controller flushing. All dirty data in the cache is flushed when the system is shut down, when the user manually forces a "sync" operation, and when there is a controller failure or power failure.

Adaptive Cache Optimizations

Following are adaptive cache behaviors for different I/O patterns.

Small-Block Random Writes (OLTP)

Minimum write size from host into the cache is a segment (1/8 of a Sun StorEdge T3 array's block). (See figure 19.) If a host write is smaller than one segment, then the entire segment must be read from disk and modified by the write. The block is held in cache as long as possible to allow subsequent random writes of additional segments in the same block to occur. When the entire block is filled, it can be written as a single atomic write to disk, thus consolidating eight host writes into a single disk write. Even when less than an entire block must be written, if the segments are contiguous, they can be written as a single atomic write to disk, without having to read the remainder of the block from disk into cache. If, for some reason, one or more noncontiguous segments in a block must be written (for example, because the 80 percent write threshold was reached), then a read/modify/write sequence of the entire block must be performed.

- Note that while the algorithm of segment write from host with block writes to disk is especially crucial to RAID 5 partial stripe write performance, it benefits RAID 1 write performance as well.

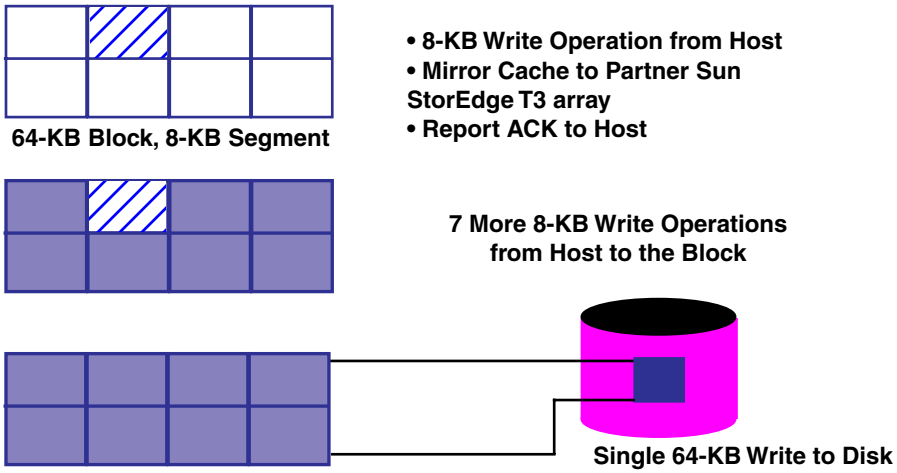

Figure 19. Sun StorEdge T3 array's Adaptive Cache, depicting a small block random write which is often used in OLTP.

Large-Block Sequential Writes

The ideal host I/O size for large-block sequential writes is equal to the Sun StorEdge T3 array's block size, or an integer multiple of Sun StorEdge T3 array's block size. This allows full block atomic writes to be performed. If RAID 5 is being used, this further allows for parity to be calculated in atomic units of segments.

Another feature of Sun StorEdge T3 array's adaptive cache is that even with write-behind mode in effect, large-block sequential writes are treated as write-though data. There is little advantage to holding large-block sequential writes in cache; they are unlikely to be read again soon as cache hits. Furthermore, large sequential I/O tends to saturate cache, eventually resulting in the same effect as write-through mode, but meanwhile monopolizing the cache at the expense of other random I/O that may be occurring. So when the Sun StorEdge T3 array detects large-block sequential I/O, it writes to disk before sending an acknowledgment to the host, thus freeing up the same block for the next sequential write, and keeping the remainder of the cache available for random writes.

Small-Block Sequential Read

The cache read-ahead parameter is configurable. If more than two host I/O blocks (not Sun StorEdge T3 array blocks) are read consecutively, then the entire array block that holds those I/O blocks will be read into cache. The default setting is “on,” which indicates that read-ahead mode is enabled. A setting of “off” will disable the read-ahead mode. The read-ahead parameter may be viewed and set from any administrative CLI or GUI tool that has write access to the Sun StorEdge T3 array. Small-block sequential read examples follow:

- Example 1: Sun StorEdge T3 array block size of 64 KB, host I/O block size of 8 KB, read-ahead enabled. Two consecutive 8-KB blocks are read by the host, causing two 8-KB array segments of a 64-KB block to be read into cache. Because read-ahead is enabled, the Sun StorEdge T3 array will read the remainder of its 64-KB block (i.e., six more 8-KB segments) into cache.
- Example 2: Sun StorEdge T3 array block size of 64 KB, host I/O block size of 2 KB, read-ahead enabled. Two consecutive 2-KB blocks are read by the host, causing one 8-KB T3 segment, four host I/O blocks, to be read into cache. Because read-ahead is enabled, the array will read the remainder of its 64-KB block (i.e., seven more 8-KB segments, 28 more host I/O blocks) into cache. Thus, at total of 32 host I/O blocks will be read into cache: the two requested blocks plus 30 more.
- Example 3: Sun StorEdge T3 array block size of 64 KB, host I/O block size of 64 KB, read-ahead enabled. Two consecutive 64-KB blocks are read by the host, causing two entire 64-KB blocks to be read into cache. Even though read-ahead is enabled, the array will not read any additional data into cache, because there is no remaining portion of a Sun StorEdge T3 array block to read.

From the third example above, users can see that if the host I/O block size multiplied by two is as large or larger than the T3 array block size, then the read-ahead parameter has no effect, and read ahead will never occur, even if the read-ahead parameter is enabled.

Administration

Administration Path

On the Sun StorEdge T3 array, the administration path provides connectivity from the controller card in the master controller unit (and alternate master controller unit) to all FRUs throughout the entire partner group. It also connects the master controller unit (and alternate master controller unit) to the external Ethernet and serial ports. It is over this path that configuration, diagnostics, and monitoring takes place.

The Sun StorEdge T3 array has two internal serial lines connecting the CPU with all non-disk FRUs. These lines are used as redundant internal administration paths, communicating configuration, control, monitoring, and diagnostic information. The serial lines extend through the daisy-chain unit interconnect cables to all units in a partner group, creating a single administrative domain.

The external Ethernet administration path is used to exchange configuration and monitoring information between the Sun StorEdge T3 array's CPU and the management host(s).

The CPU has no access to the application data, and no application data is available via the administration path. This separation of data and administration paths provides security by shielding application data from those individuals performing everyday service and administration. The path disunion also means that a path for communicating with the Sun StorEdge T3 array remains available even if the data path or application host has failed.

In a configuration of more than one Sun StorEdge T3 array, one controller unit is always designated as the master controller unit, and its partner is the alternate master controller unit. Although both controllers are physically connected to the external management network, only the master controller's connection is active. All administration and all external communication, on behalf of all units in the configuration, is conducted via the master controller. Only in case of master controller failure does the alternate master controller take over administration.

Configuration Overview

In keeping with the goal of a simple, easy to use product, the Sun StorEdge T3 array requires the user to contend with only a few system and logical volume configuration options. The Sun StorEdge T3 array includes basic volume management, which operates with the goal of presenting one or two "perfect disks" (i.e., logical volumes that never fail) to the application host. The nine disks in each unit (controller or expansion unit) can be configured as one or two SCSI logical units (LUNs), with the option of a hot spare disk. Each LUN may be independently assigned a RAID level (5, 1, or 0). Each tray comes preset with a default volume configuration. More complex configurations, such as n-way mirroring, plaiding (stripes of stripes), or slicing into small partitions are accomplished using a host-based volume manager.

System configuration parameters include the block size of the logical volumes, cache usage, and the run time priority (relative to application I/O) of various diagnostic and maintenance routines, including volume reconstruction, parity checking, and diagnostics. System configuration parameters settings are global to all Sun StorEdge T3 array trays in a partner group. All system configuration parameters are preset to commonly used default values.

When operating partner group configurations, host data path failover is supported via host-based alternate pathing software. For the Solaris™ Operating Environment, either VERITAS DMP (dynamic multipathing) or Solaris operating environment AP (alternate pathing) may be used. For other major operating environments, users can choose between VERITAS DMP or the custom drivers for the Sun StorEdge T3 array. Consult the Sun StorEdge T3 array documentation for current availability.

Data Volume Configuration

In each Sun StorEdge T3 array, the nine disk drives are configured into one or two logical volumes or LUNs, which are the atomic units presented to the application host. In other words, the application host does not see the individual disk drives.

Sun StorEdge T3 array configuration tools employ the following configuration rules:

- LUNs must consist of a contiguous sequence of whole disks.
- A disk may not be partitioned into different LUNs.

- LUNs may not span physical trays.
- There is a maximum of two LUNs per tray.
- The minimum size for RAID 1 LUNs is two disks, and maximum size is nine disks.
- The minimum size for RAID 5 LUNs is three disks and maximum size is nine disks.
- If a hot spare is used, it is always drive number 9, it must be used with all LUNs in a given tray, and it must be declared when the first LUN on a tray is created.

These rules may appear restrictive, but they provide for great simplicity of configuration. The user has to make only these decisions:

- Will there be a standby (hot spare) drive?
- How many LUNs, one or two? If two, how many disks for each LUN?
- What RAID level is required for each LUN?

Recommended Configurations per Tray

In general, a configuration of a single LUN per tray is recommended over dual LUNs per tray, because dual LUNs create additional administrative overhead without adding value.[3] Dual LUN configurations are recommended only in the case where a small RAID 1 boot volume or log volume is desired.

The following is a list of recommended configurations by required RAID level:

- RAID 5: single LUN, either 9 disk (8+1) without hot spare or 8 disk (7+1) with hot spare. Sun StorEdge T3 array hardware and firmware have been optimized for RAID 5. In most cases, RAID 5 will outperform RAID 1. If read/write ratio is 1:1 or higher, use RAID 5.

3 If small 1-GB or 2-GB LUNs are desired, 14 partitions (two LUNs x 7 Solaris partitions) still do not create enough volumes to utilize all disk capacity. In this case, use of host-based volume management software (such as VERITAS Volume Manager) will be needed to create the required quantity of subvolumes, regardless of whether there are one or two native volumes on the Sun StorEdge T3 array.

- RAID 1: single LUN, either nine disk without hot spare or eight disk with hot spare. RAID 1 is recommended if the customer is seeking highest reliability possible, or if the application read/write ratio is less than 1:1 (i.e., more than 50 percent writes). Note that on the Sun StorEdge T3 array, a RAID 1 volume may consist of an odd number of disks. This is because the array uses a diagonal RAID 1 scheme (see figure 20). A block on one drive is mirrored on the adjacent drive. In other words, with nine-disk RAID 1, disk 1 data is mirrored on disk 2, disk 2 data is mirrored on disk 3, and disk 9 data is mirrored on disk 1.

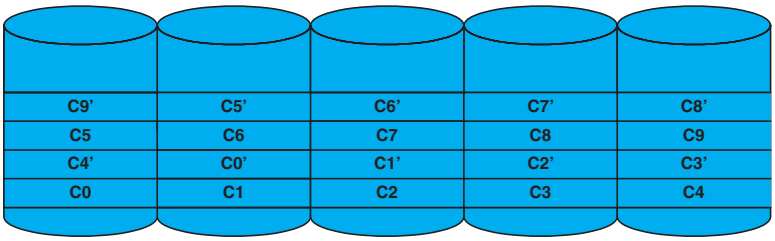

Figure 20. Disk configurations for the Sun StorEdge T3 Architecture. With RAID 1, users gain the reliability benefits of Raid 1+0, the performance benefits of striping across all disks in the group, and the ability to include an odd number of disks in the disk group.

- RAID 0: single LUN of 9 disks. Use of RAID 0 is advised only in conjunction with an external form of data protection, such as using host-based mirroring across two Sun StorEdge T3 arrays, or the rare case of using host-based RAID 5 stripes across multiple trays.

Monitoring and Maintenance

Every customer's application environment is unique, as is their administration of their application environment. Sun's philosophy of administration for the Sun StorEdge T3 array is based on flexibility. The T3 array's architecture makes no attempt to dictate one approach, tool, or application programming interface (API) for administration. Instead, the Sun StorEdge T3 array's architecture offers users a variety of APIs and tools that can be used, at each user's discretion, on their own or in conjunction with an existing independent administrative environment.

Supported APIs include a system log (syslog), SNMP traps, and a Jiro™ technology Managed Object (MO) platform for the Sun StorEdge T3 array, based on http tokens. Some APIs, such as syslog and SNMP, enable read-only activity, which is suitable for monitoring. Other APIs, such as native firmware commands and http tokens, support both read and write activity, and thus are suitable for both monitoring and configuration.

User tools include the Sun StorEdge T3 array's native CLI (command line interface), which is documented in the Sun StorEdge T3 array Administrator's Guide, as well as other CLIs and GUIs (graphical user interfaces), which vary by platform. (See reference documents for each platform for more information.)

Supported user interfaces include the status LEDs on the physical unit, a native CLI, the Component Manager GUI (see figure 21), Sun Remote Service (SRS), and STORtools (CLI or GUI). These tools are typically operated from a management host, which communicates with the Sun StorEdge T3 array via its Ethernet port. An exception is the StorTools™ utility, which operates from the application host to diagnose problems in the data path from the host to the Sun StorEdge T3 array.

The user also can write scripts that log into the Sun StorEdge T3 array via telnet and use the native CLI as an API. Also, the user is free to employ an independent administration application by leveraging one of the APIs mentioned above. The syslog is stored locally on the Sun StorEdge T3 array, and may be exported to any management host that is accessible via the management network. Each administration API and tool has varying functional capability for the Sun StorEdge T3 array. Customers should consult with their sales support team and the tool documentation to determine which API or tool (or combination) is appropriate for their environment.

Serviceability

One of the beneficial side effects of using a hardware RAID product is separation of host and back-end channels, which eases problem identification. The Sun StorEdge T3 array's architecture expands on this concept by using sophisticated diagnostic firmware combined with loop-switching capability. This allows for innovative use of back-end loops, both in normal operation and in case of hardware failures. Both back-end loops are utilized for I/O and cache mirroring, which enables superior throughput. Hardware problems are automatically isolated to the appropriate field replaceable unit (FRU). In the case of back-end loop problems, which are typically difficult to diagnose, the

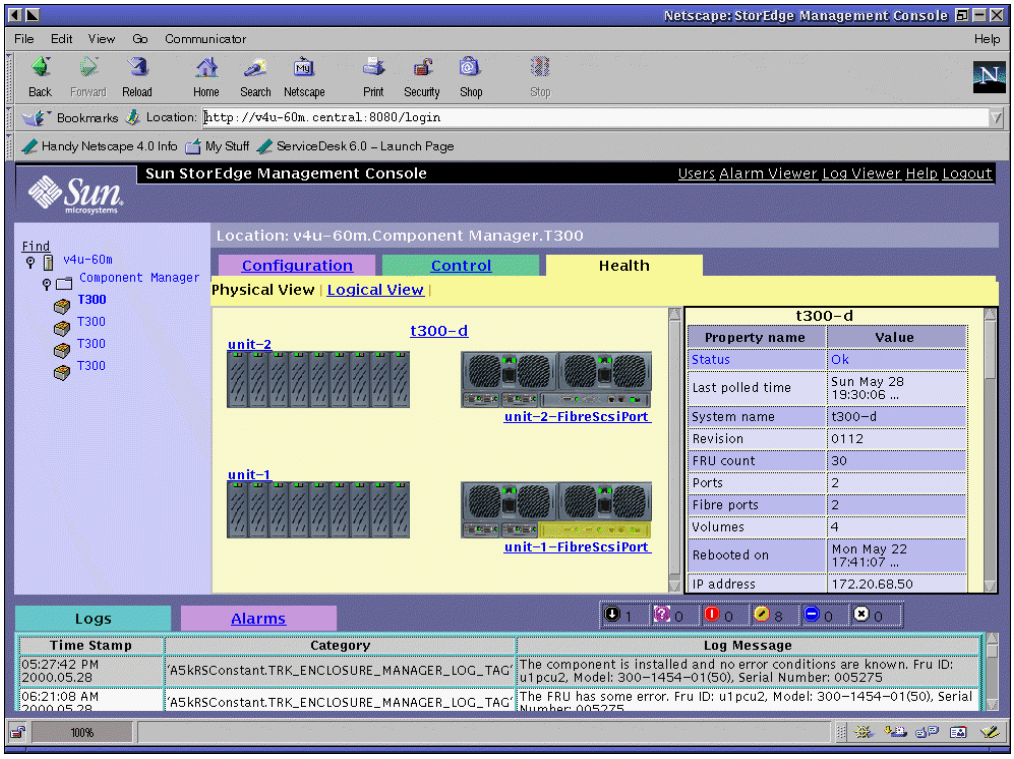

Figure 21. Snapshot of Component Manager depicting Sun StorEdge T3 array's physical view health screen.

problem can be isolated to a specific FRU, and furthermore that FRU can be bypassed in the loop. This allows the loop to continue I/O operations and provide optimal performance even in degraded conditions.

Diagnostics

The level of diagnoseability offered within the Sun StorEdge T3 array is unprecedented. All FRUs are constantly monitored. FRU failures can be detected without intervention from the user or the application host, and the failure reported via the various managment APIs. And because the T3 array has n+1 redundancy, once the Sun StorEdge T3 array has identified the failed FRU, the FRU can be replaced without loss of data availability. Furthermore,

any state that is automatically changed upon a failure (such as disabling write-behind cache upon a controller failure) is automatically restored upon replacement of the failed FRU.

The array's loop-switching capability combined with loop diagnostic software allows even back-end loop failures, traditionally difficult to diagnose, to be easily isolated, and the offending FRU fenced out.

Hardware Upgrades and Repairs

FRU Replacement

Although the FRUs in Sun StorEdge T3 arrays are designed for high reliability, FRU redundancy means that if a component does fail, users need not experience an application interruption.

To minimize MTTR, the Sun StorEdge T3 array's FRUs are hot swappable. Upgrades and repairs are performed simply by replacing the component. (See page 10, "FRU Replacement" for replacement procedures.)

FRU failures are automatically detected by the master controller unit and reported via the LEDs on the failed FRU, and via the various APIs on the Sun StorEdge T3 array. Replacement considerations for specific FRUs are provided below.

- PCUs are redundant and may be replaced without any impact on the Sun StorEdge T3 array's operation.
- Controller failure/removal/replacement will, of course, cause a controller failover. Data continues to be available to the application host.
- Unit Interconnect Card failure/removal/replacement will cause back-end loop failover. Data continues to be available to the application host.
- Disk failure/removal/replacement invokes logical volume reconstruction. If a disk fails, data availability to the host is maintained, due to the redundancy provided by use of RAID 1 or 5. If RAID 1 is used, data is available from the mirror copy on the adjacent drive. If RAID 5 is used, data is reconstructed from parity as data moves through the XOR engine on its way to cache. Upon disk failure, the proper monitoring notifications are given. If a standby (hot spare) drive is available, the system will begin reconstructing data onto the standby drive. Once the failed drive has been

replaced, the data on the standby drive is copied onto the replacement drive and the hot spare is returned to standby status. If no standby drive is available, the system continues to run in degraded mode until the failed drive is replaced, at which time reconstruction takes place directly onto the replacement drive.

The time required for the RAID reconstruction is variable, based on disk size, system availability, and the relative priority of reconstruction versus application I/O. If priority is given to application I/O, disk reconstruction can take several hours. If priority is given to reconstruction, duration can be less than one hour for an 18-GB drive, at the expense of throttling application I/O to almost nothing. Users may reference the Sun StorEdge T3 array product documentation for details regarding disk reconstruction priority configuration.

Midplane Replacement

In the rare circumstance where a midplane must be replaced, all FRUs are removed, the entire chassis is swapped out, and the FRUs are restored. Note that the system takes its MAC address from the FRU ID of the master controller unit midplane. Therefore, if that midplane must be replaced, the system MAC address will change, and any MAC-to-IP address mapping in the administrative network must be updated accordingly.

Software Upgrades and Repairs

All firmware (on the controller, controller FLASH PROM, UIC, and UIC FLASH PROM) can be easily upgraded. First, transfer new firmware from the management host to the Sun StorEdge T3 array system area via ftp. Several different versions of firmware can be stored in the system area, so old firmware may be retained, allowing the flexibility to back out or revert to an earlier version, if necessary. Second, instruct master controller unit to use the new firmware.

The user must perform a system reboot to invoke new firmware in the case of controller firmware upgrade only.

Summary

In a marketplace that requires businesses to remain agile and adaptable to growth and change, storage that can aid and abet a dynamic information infrastructure is invaluable. Sun StorEdge T3 array is designed with extremely high RAS standards to bring dramatic benefits to customers' data centers.

To achieve heightened reliability and availability, the Sun StorEdge T3 array components were designed to be N+1 redundant and field replaceable. Its four FRUs are hot swappable, decreasing downtime, minimizing MTTR, and contributing to ease of use and serviceability. A component does not interrupt applications on the Sun StorEdge T3 array.

By assembling Sun StorEdge T3 arrays in a partner group, customers not only can initiate cache mirroring and controller failover, but also begin to scale for capacity, bandwidth, and/or IOPS. The modular nature of the Sun StorEdge T3 array allows customers to purchase for capacity or speed as needed.

Thanks to the Sun StorEdge T3 array's low administrative requirements and centralized managment capabilities, customers can reduce administrative effort and cost and allot their time to more intensive tasks. RAS standards significantly add to customer cost savings by supporting continuous data availability and unfailing mission-critical application performance.

Glossary of Cache Terms

Block	An overly used term. Often used to describe the amount of data sent or received by the host per I/O operation. Also used to describe the size of an atomic read/write operation to/from a disk. In the context of the Sun StorEdge T3 array, represents the size of each cache buffer, and also the disk interleave factor (also known as stripe unit, chunk, interlace factor). Sun StorEdge T3 array block size can be 16 KB, 32 KB, or 64 KB.
Cache hit	A read or write request for data that is already in cache. Therefore, a request can be serviced without needing to go to disk.
Clean data	Any read data or write data that has been committed to disk. In other words, a copy of data that is safely on disk.
Dirty data	Write data that is in cache and has been acknowledged to the application host, but which has not yet been committed to disk.
Read-ahead	Sequential data that has been read from disk into cache without having actually been requested by the application host, in anticipation that it will be requested by the host. When the request occurs, it can be serviced as a low latency cache hit, thus improving host application performance.
Stripe size	Total amount of data in a disk stripe; i.e. block size multiplied by number of data disks in the stripe.
Stripe width	Total number of disks in a disk stripe.
Segment	Another overly used term; in the context of the Sun StorEdge T3 array, 1/8 of a cache buffer. In the Sun StorEdge T3 array, a segment is the smallest size of I/O possible between cache and disk. Segment size is 2 KB, 4 KB, or 8 KB, depending on block size.
Write-behind mode	A data write is acknowledged to the application host as soon as it is in (mirrored) cache, without having yet been committed to disk, in order to reduce write latency. Also known as write-back or fast-write mode.
Write-through mode	A data write is acknowledged only when data is fully committed to disk.