An Oracle White Paper
April 2010

An Economical Approach to Maximizing Data Availability with Oracle's Sun Storage 7000 Unified Storage Systems

ORACLE

Introduction

Storage is a trust business. Oracle's line of integrated storage appliances provide the reliability, availability, and serviceability (RAS) attributes necessary to preserve customer data and meet enterprise expectations. Some aspects of RAS for storage are well defined by IT organizations and the industry. Oracle's Sun Storage 7000 Unified Storage Systems meet these requirements by delivering features such as RAID-6, active-active clustering, integrity checking, redundancy, replication, and snapshot capabilities. Other aspects of the RAS design of Sun Storage 7000 Unified Storage Systems are new to the industry and offer exceptional value. In particular, the use of NAND Flash devices in conjunction with the Oracle Solaris Zettabyte File System (Oracle Solaris ZFS) defines a unique, highly reliable architecture that aims to provide the best possible price for performance ratio.

Reliability, Availability, and Serviceability Design Approach

In any area of system design — storage, networking, compute, or software — there exists a natural tension between the number of RAS features provided and the effect on availability. The cost in time and money to engineer and test the additional features can also be a factor. As system design moves from delivering four "nines" of uptime to five "nines" and beyond, the cost to engineer and test the next increment of reliability tends to increase dramatically — perhaps exponentially. These additional costs are often folded into the product price. Hence the DRAM module for a z/OS mainframe generally costs much more than the DRAM module found in a commodity x86 server — even if the capacity of the modules are identical.

Oracle's line of unified storage systems take advantage of commodity economics, allowing organizations to take advantage of the dramatically reducing price and increasing density of general-purpose compute cycles, exceptional raw capacity of Serial ATA (SATA) drives, and high performance of NAND Flash devices. Within these storage appliances, systematic design decisions control how and where availability is achieved. These efforts help preserve the economic benefits of industry-standard volume components while optimizing uptime levels.

In some cases, solving reliability issues at a higher level of the technology stack — through software innovations — can help minimize product costs. As an example, the Oracle Solaris ZFS software stack provides 19 "nines" of data reliability. By checksumming data and metadata and storing a unique checksum tree, Oracle Solaris ZFS protects data from the disk platter all the way to applications or over-the-wire data protocols. As a result, expensive, specialized NAND Flash devices or SATA disk spindles that provide data protection beyond current commodity attributes become unnecessary.

While considering the overall RAS design of Sun Storage 7000 Unified Storage Systems, keep Oracle's design philosophy firmly in mind. The objective is not to deliver the highest reliability at every level of the system architecture, rather to deliver the highest data availability at the lowest overall cost. This goal is achieved by solving each availability problem at the right layer of the storage system.

This technical white paper provides an overview of the RAS features of Sun Storage 7000 Unified Storage Systems. The content of this article assumes a basic understanding of Sun Storage 7000 Unified Storage Systems, common enterprise-class storage features, and Oracle Solaris ZFS. Specifically, the document discusses Oracle's advanced approach to managing and preventing fault events, provides a detailed review of the RAS features at every layer of Sun Storage 7000 Unified Storage Systems, and describes additional RAS features provided by Oracle's storage appliance management software, phone home support, and upgrade procedures.

Fault Management and Predictive Self Healing

Oracle's fault management architecture (FMA) defines a philosophy and set of common technology components that support self-healing features within Oracle products. Specific capabilities provided by FMA include first-fault telemetry capture, automated diagnosis of hardware failures, standardized fault messaging, and predictive and reactive fault isolation. The first set of FMA features were delivered as a part of Oracle Solaris 10 to enhance the fault management technology of Oracle's Sun servers with SPARC® processors. Today, a significantly larger set of capabilities are standard across Oracle's Sun platforms with SPARC, Intel®, and AMD Opteron™ processors. Sun Storage 7000 Unified Storage Systems utilize and build upon the standard FMA capabilities and design principles. An overview of FMA principles can be found at *http://opensolaris.org/os/community/fm/*.

Most traditional systems provide simple error reporting for hardware and software failures. Platforms often present error messages in the form of human readable —though often not easily decipherable — text strings intended for a system administrator or service engineer. Rather than helping to identify the underlying issue, error messages tend to provide a large stream of individual symptoms, such as "I have a fever, muscle ache, and sore throat", or in a computer system, "Disk 2 had a checksum error while reading block 12, a checksum error while writing block 14, and an I/O timeout while executing a SCSI write cache flush command".

Under the FMA model, the system can present the root-cause diagnosis — "I have the flu", or in a computer system, "Disk 2 is broken". Oracle's FMA model defines an architecture that:

- Captures error telemetry
- Stores the error event information in a structured form for post-mortem analysis
- Forwards the error event to a diagnosis engine
- Automates diagnosis of underlying problems, taking into account a large collection of simultaneously occurring errors that may have causal relationships
- Presents the diagnosis result to humans, rather than the underlying telemetry
- Through automated fault diagnosis, FMA can provide the highest accuracy, fastest time to diagnosis, and proper corrective action.

In an FMA system, a set of diagnosis rules for all detectable errors are defined in the design phase and delivered with the product. Each possible diagnosis result is assigned a unique code called the SUNW-MSG-ID. Designed to be used to classify problems that are occurring in the field, the SUNW-MSG-ID can be easily e-mailed or read over the phone to Oralce support. For example, the code ZFS-8000-D3 might correspond to the notion "a disk experienced one or more uncorrectable errors during an attempt to access the device label".

Each SUNW-MSG-ID corresponds to a human-readable message describing the problem, the software's automated response, a suggested corrective action, and links to a more detailed knowledge article on the *http://sun.com/msg* Web site. In addition, each individual problem diagnosis that is

completed on a running system is assigned a universally unique identifier (UUID) that can be used to uniquely identify that specific problem on that particular system. These UUIDs are used to recall the related telemetry, or as a key for a phone-home call. An FMA system answers the question "is anything broken on this system?" by reporting a list of (SUNW-MSG-ID, UUID) pairs, along with the human-readable diagnosis summaries corresponding to the SUNW-MSG-ID of each pair.

The FMA diagnosis and response capabilities standard in Sun Storage 7000 Unified Storage Systems are listed in Table 1:

TABLE 1. FMA DIAGNOSIS AND RESPONSE CAPABILITIES FOR SUN STORAGE 7000 UNIFIED STORAGE SYSTEMS

COMPONENT	ERROR DETECTION MECHANISMS	AUTOMATED RESPONSE
Software service	Process failure, core dump	Restart affected service and appropriate dependencies
CPU register file	Register file ECC errors	Offline CPU core
CPU L1, L2, and L3 cache	CPU cache ECC errors	Offline CPU core or all cores sharing an L3 cache
CPU HyperTransport (HT) or Quick Path Interconnect (QPI)	HT or QPI CRC errors	Fault diagnosis and reporting only
DRAM cells	DRAM ECC errors	Offline physical memory page containing bad cells
DIMMs and DIMM connectors	DRAM ECC errors	Fault diagnosis and reporting only
PCI Express Expansion cards	DMA errors, PIO Errors, Oracle Solaris ZFS checksum errors	Retire I/O device instances associated with a faulty card, thereby triggering network or I/O multipath failover
PCI Express lanes and connectors	DMA and PIO ECC errors, Oracle Solaris ZFS checksum errors	Retire I/O device instances associated with a faulty card, thereby triggering network or I/O multipath failover
Serial Attached SCSI (SAS) lanes and connectors	SAS I/O errors, CRC errors, Oracle Solaris ZFS checksum errors	Fault diagnosis and reporting only
SAS expanders	SAS I/O errors, Oracle Solaris ZFS checksum errors	Fault diagnosis and reporting only
Disk blocks	SAS I/O errors, RCC errors, SMART errors, Oracle Solaris ZFS checksum errors	Disk bad block remap, Oracle Solaris ZFS block self-healing
Disk devices	SAS I/O errors, RCC errors, SMART errors, Oracle Solaris ZFS checksum errors	Device offline, hot spare activation
Flash blocks	SAS I/O errors, RCC errors, SMART errors, Oracle Solaris ZFS checksum errors	Flash back block remap, Oracle Solaris ZFS block self-healing
Flash devices	SAS I/O errors, RCC errors, SMART errors, Oracle Solaris ZFS checksum errors	Device offline
Fans	i2c sensor errors	Fault diagnosis and reporting only
Power supplies	i2c sensor errors	Fault diagnosis and reporting only

At any given level of a storage system, the RAS design generally operates according to these rules:

- **A detection mechanism is implemented to discover that something is wrong.**
  As examples, a checksum is verified, an I/O succeeds or times out, or a cluster heartbeat is not received within a defined interval. Ideally, storage systems eliminate any possibility of silent data corruption — failures that are undetectable. To this end, the Oracle Solaris ZFS data architecture offers strong end-to-end checksums of data and metadata, providing 19 "nines" of data integrity.
- **The state associated with the detection mechanism is logged persistently.**
  To speed resolution, Sun Storage 7000 Unified Storage Systems represent the state as an FMA error report, store the report in the FMA error report log, and dispatch the error state for automated diagnosis. This automated diagnosis is provided for all key system components, including CPUs, DIMMs, Fan, PSU, Disk, and PCI Express cards.
- **Synchronous to the operation in question, an appropriate redundancy is applied — if one exists — to correct or bypass the failure.**
  Specific storage systems examples include utilizing extra ECC check bits to correct a flipped bit, retiring an I/O, directing an I/O to another disk in a mirrored pair or RAID group, restarting a daemon, or failing over a cluster.
- **Asynchronous to the operation in question, problem diagnosis is initiated**.
  Within Sun Storage 7000 Unified Storage Systems the appropriate FMA diagnosis engine examines the stream of error telemetry and attempts to identify and report the underlying fault in the system. A diagnosis may be a disk is broken, a bad cell exists in a DRAM, an L2 cache line is defective, or a power supply is dead.
- **Automated system responses are triggered to mitigate the impacts of the fault.**
  If appropriate, the FMA diagnosis engine result triggers an immediate automated response to prevent the fault from causing further systemic failure. Actions might include offlining a faulty disk, fencing off a DRAM page via the kernel's VM subsystem, or disabling the CPU core associated with a bad L2 cache line.
- **Diagnostic results are reported.**
  The FMA diagnosis result is presented through the software and hardware of the system to inform the customer and describe concisely the affected components and appropriate repair. Examples include the appearance of a pop-up in the Web browser interface, emission of an SNMP trap, transmission of a "phone home" message back to Oracle, update to the list of open problems, or illumination of a physical fault LED. Reporting actions include the following possibilities:
  - If Oracle's Sun Connection Readiness Kit (SCRK) or Auto Service Request (ASR) phone-home capability is enabled, and if the problem diagnosis matches a result that is associated with an immediate service action, a Radiance case is opened and response is initiated.
  - If the SNMP stack is enabled with a trap destination, a Fault Management Trap is dispatched to the pre-defined trap destination. The trap contains the SUNW-MSG-ID and UUID of the diagnosis.

- If a matching Alert rule is configured, an additional SNMP trap, syslog message, or e-mail message may be dispatched containing the SUNW-MSG-ID, UUID, and description of the problem.
- The Web browser user interface (UI) is updated for all users with a visual pop-up summarizing the alert, along with the Dashboard and Maintenance-Problems screens, described later in this document.

- **A service action is initiated.**
  At this phase, additional RAS features come into play such as the ability to locate the components to be repaired, electronic reporting of software or firmware versions, part numbers, or other attributes, and system support for on-line repair.

The remainder of this document first describes each of the major software and hardware components of Sun Storage 7000 Unified Storage Systems with respect to this RAS activity flow, and then discusses additional generic RAS capabilities.

Storage System RAS Design

In order to provide the highest possible data availability and integrity, Sun Storage 7000 Unified Storage Systems are designed to eliminate single points of failure. At every level of the system, strong error detection capabilities and appropriate data path redundancies are provided. Sun Storage 7000 Unified Storage Systems combine with Oracle's Sun Storage 7000 Management software to support the highest data path availability. Examples of these features include Oracle Solaris ZFS checksums, ditto blocks, RAID-Z, mirroring, hot spares, multipath I/O, Just a Bunch of Disks (JBOD) mirroring, active-active clustering, remote replication, network link aggregation control protocol (LACP) and IP Multipathing (IPMP).

This chapter walks through the major areas of a storage system from bottom to top, describing the RAS attributes and capabilities of Sun Storage 7000 Unified Storage Systems. When appropriate, storage redundancies are optional, allowing IT architects to make the necessary business trade-off between total system cost and availability.

Disk Block

The lowest level of the storage subsystem is the disk block, the unit of reading and writing to the disk drives. Sun Storage 7000 Unified Storage Systems use enterprise-grade SATA disk devices for raw capacity. SATA disk devices typically offer at least 10-bit ECC and 1.2 million hours mean time between failure (MTBF) rating.

Through the benefits of Oracle Solaris ZFS, Sun Storage 7000 Unified Storage Systems can provide even stronger data reliability guarantees beyond the benefits of any individual drive component. Oracle Solaris ZFS provides significantly stronger checksums for user data and metadata and stores these checksums in a Merkle tree — so as to detect phantom writes, mis-directed reads and writes, and other drive and HBA firmware errors. A checksum error detected by Oracle Solaris ZFS can identify data corruption that might have otherwise been silent to a disk, HBA, or other system path. In response to

checksum errors, Oracle Solaris ZFS automatically returns good data to the application or data protocol by using data elsewhere in the mirror or RAID stripe. In addition, Oracle Solaris ZFS automatically attempts to self-heal (rewrite) the bad data with the verified good data.

Another aspect of availability provided by individual drives is the use of extra capacity to reallocate disk blocks or sectors that encounter physical write errors. Although behavior varies across drives, in general all enterprise-class SATA drives today offer this functionality. A defect list or "g-list" is utilized to track bad blocks and a fixed number of extra blocks are reserved in advance. If a drive is unable to perform a successful write to a particular block, the block is added to the g-list and one of the previously reserved blocks is written instead. The logical address of the original block now maps to the spare block. Once the total capacity of the g-list is exhausted, the drive's Self-Monitoring, Analysis, and Reporting Technology (SMART) turns on the predicted failure indicator and the FMA storage diagnosis engine notes the drive as faulty, offlines the drive from use by Oracle Solaris ZFS, and triggers use of a hot-spare.

In addition to the drive's own defect list, Oracle Solaris ZFS provides added reliability for critical data blocks through the use of "ditto blocks". Oracle Solaris ZFS ditto blocks are extra copies of critical data blocks used in the event that reading from one copy returns an uncorrectable error from a given drive.

Any file system can be viewed as a tree of blocks. The inability to access a "higher" indirect block means that blocks it refers to also become inaccessible. As a result, the "value" of a given block increases with its height in the tree. Oracle Solaris ZFS currently supports two-way and three-way ditto blocks, and automatically allocates two-way ditto blocks for file system metadata, and three-way ditto blocks for metadata that is global to an entire Oracle Solaris ZFS Hybrid Storage Pool. Oracle Solaris ZFS also automatically allocates ditto blocks to be "far" from one another to increase their usefulness. In a typically RAID-Z configuration, Oracle Solaris ZFS allocates ditto blocks to be in separate virtual devices (vdevs), such that if an entire RAID stripe fails, the ditto copy of the particular metadata block likely still remains available in another RAID stripe.

Finally, Oracle Solaris ZFS also implements proactive background disk scrubbing at low priority. This procedure exercises disk blocks and detects latent checksum errors before they are encountered in the data path. Similar to errors detected by any Oracle Solaris ZFS data access, errors detected by the background scrubbing process can attempt to self-heal a bad block using good data located on the other side of the mirror or by using the RAID-Z parity to reconstruct the data.

Disk

The Oracle Solaris ZFS storage architecture provides the basic disk redundancies that are now standard in the enterprise storage market, including mirroring, RAID-Z (RAID-5), RAID-Z DP (RAID-6), and hot spares. During configuration of Sun Storage 7000 Unified Storage Systems, the management interface offers choices that facilitate the basic business trade-off between capacity, reliability, and performance. Based on this user input, the system optimizes the configuration of raw disk devices into an appropriate mirror or RAID layout. This management software also contains hardware profiles of Sun open storage devices from Oracle. As a result, the system takes specific hardware details into

account — such as the internal SAS controller wiring and the relationship of drive sets to fans and power supplies — to further optimize the layout of stripes and hot spares.

Devices grouped into Oracle Solaris ZFS mirrors and RAID groups operate as expected of enterprise-class storage:

- Writes occur to multiple devices
- Reads access multiple devices
- Devices with Oracle Solaris ZFS checksum errors prompt retrieval of good data from the other side of the mirror or reconstruction using a RAID parity encoding
- Good data is used to self-heal bad data in a mirror or RAID stripe
- Replacement devices are automatically resilvered upon insertion
- Failed devices are immediately replaced with a predefined hot spare

Since two or more disks can not be updated atomically, traditional RAID systems – regardless of the RAID parity algorithm – are susceptible to a write hole. For example, if a system crash or power outage occurs in between the update of data and parity in a RAID stripe, the stripe can be damaged. Enterprise RAID solutions often work around the write hole while maintaining performance by using hardware RAID cards with NVRAM on-board, or software RAID stacks that make direct use of NVRAM devices. Unfortunately, these devices add cost, eventually wear out, and can sometimes explode or leak.

In contrast, the Oracle Solaris ZFS RAID-Z algorithm can provide better device-level reliability by design. RAID-Z is a data and parity scheme that uses a dynamic stripe width – every block is composed of its own stripe – regardless of the block size. Therefore, every RAID-Z write is a full-stripe write in the Oracle Solaris ZFS copy-on-write transaction system. This approach eliminates both read-modify-write operations and the notion of a write hole. As a result, RAID-Z systems that are configured to prioritize the lowest cost per gigabyte over performance have no need of RAID cards. Configurations that require the highest possible performance can make use of Flash devices rather than NVRAM, thereby eliminating on-board battery modules as potential sources of system failure. For more details about Oracle Solaris ZFS and RAID-Z internals, please refer to *http://opensolaris.org/os/community/zfs/*.

To maximize the protection of critical system data, Sun Storage 7000 Unified Storage Systems use separate, mirrored boot devices — managed by Oracle Solaris ZFS — to store the system software, configuration metadata, and log files. This measure helps ensure that the system software itself is not stored on a single point of failure. In addition, each of the Oracle Solaris ZFS reliability attributes of mirroring also apply to these boot drives.

Flash Page and Blocks

Sun Storage 7000 Unified Storage Systems utilize Oracle Solaris ZFS Hybrid Storage Pools, a unique architecture that can deliver best price per performance. A Oracle Solaris ZFS Hybrid Storage Pool

incorporates DRAM cache, read-optimized Flash devices, write-optimized Flash devices, and commodity SATA devices for capacity, all transparently managed by Oracle Solaris ZFS. Taking advantage of the Oracle Solaris ZFS Hybrid Storage Pool architecture can help organizations gain the lowest:

- cost per Gigabyte
- cost per IOP
- cost per Megabyte/second
- W per IOP
- W per Megabyte/second

A detailed discussion of Oracle's Flash architecture is beyond the scope of this document. Additional details can be found in the article "Flash Storage Memory," Communications of the ACM, July 2008. For the purposes of discussing storage RAS design, a brief summary of the key attributes of Flash relevant to reliability of the overall storage system follows.

The Flash devices used in Sun Storage 7000 Unified Storage Systems are based on single-level cell (SLC) NAND Flash with a DRAM buffer for write performance. Within these systems, many Flash channels operate in parallel, and an energy storage module (ESM, or "super-capacitor") drains the write buffer into non-volatile Flash in the event of power loss.

Flash chips are organized by a controller into physical blocks, and represent the smallest units of the non-volatile write-erase cycles performed by the controller. These physical blocks are divided into pages, the unit of an actual read or write accepted by the device. In turn, pages can be presented as logical sectors or blocks to host software. A typical NAND Flash device with a SATA interface might offer 128K physical blocks, 2K pages (plus extra space used for ECC codes), and the usual 512 byte SATA sectors or logical blocks.

The common physical failure mode for Flash media is write-induced wear-out of a thin oxide layer between the two transistors that comprise each Flash cell. Once the layer wears out, the cell can fail to properly hold its value for the defined lifetime. Current 60 nanometer SLC NAND Flash guarantees on the order of 100,000 write-erase cycles per physical block before cells are likely to begin exhibiting symptoms of wear-out at a frequency higher than the specified bit-error rate.

Although the physical technology underlying Flash devices is quite different from that of rotating media, most of the reliability techniques applied in enterprise-grade Flash devices are quite similar. Specifically, Flash devices employed in Sun Storage 7000 Unified Storage Systems offer the following reliability measures:

- Utilize ECC or Reed-Solomon codes at the device level to protect data, on top of this mechanism Oracle adds significantly stronger Oracle Solaris ZFS data checksums
- Reserve excess capacity that can be used to remap bad physical pages on to a defect list, the exhaustion of this reserved space triggers a SMART threshold event
- Provide SMART-based monitoring of other internal components such as ESMs

- Implement firmware-managed wear-leveling that helps ensure that write/erase cycles are evenly distributed across the underlying Flash pages to maximize the overall lifetime of each page
- Support optional redundancy of Flash devices in the system itself, as described in the next section

The wear-leveling and bad-block remapping capabilities of the Flash devices, combined with the physical block lifetime, result in an overall rating of 2 million write-erase cycles for each Flash device. This rating can translate into an expected field lifetime of about five years of 24 x 7 x 365 enterprise storage use.

Flash Device and SSD

Appropriate redundancies are provided for Flash devices so that a single device is not a single point of failure. Sun Storage 7000 Unified Storage Systems making use of an Oracle Solaris ZFS Hybrid Storage Pool currently support two types of Flash devices:

- Read-optimized cache device that extends main memory with a secondary cache (Oracle Solaris ZFS Level 2 Adaptive Replacement Cache (L2ARC))
- Write-optimized log device that hosts the Oracle Solaris ZFS Intent Log (ZIL)

Read-optimized cache devices are used to implement a clean cache, storing the L2ARC behind the Oracle Solaris ZFS Level 1 Adaptive Replacement Cache (L1ARC) that resides in main memory (DRAM). The L2ARC is never the sole copy of any pending write. As a result, the data stored in the L2ARC remains clean and there is no need to provide additional redundancy at the level of read cache devices. If such a device fails — either through predictive diagnosis, a SMART threshold failure, or inability to execute I/O to the device — the Oracle Solaris ZFS file system transparently retrys the pending I/O query to the larger pool of SATA devices and returns good data back to applications. Therefore, a failed L2ARC device is handled identically to an L2ARC miss.

Write-optimized log devices are used to store the Oracle Solaris ZFS ZIL. As a result, synchronous writes from clients over NFS, CIFS, and other protocols can be committed to non-volatile storage and acknowledged as fast as possible. Subsequently, the newly written data is asynchronously migrated to a larger pool of slower SATA devices. Write-optimized log devices play the same role in system performance as that of NVRAM in a traditional storage system, but are arbitrarily scalable, significantly more power-efficient, and do not form a write-side bottleneck in a clustered storage system.

To permit scaling of ZIL capacity with redundancy, Sun Storage 7000 Unified Storage Systems can be configured to mirror ZIL devices in pairs. A mirrored ZIL device pair is handled identically to any Oracle Solaris ZFS file system mirrored pair:

- Writes occur to both sides
- Reads transparently access the faster of the two devices
- An Oracle Solaris ZFS file system checksum error on one side returns good data from the other side and self-heals the side with errors
- A replacement device is automatically resilvered upon insertion

If both sides of a mirrored pair of ZIL devices fail, Oracle Solaris ZFS can still continue to accept writes at lower performance. In this scenario, new ZIL entries are transparently written to free space in the larger pool of SATA devices. Therefore, even the pathological failure of every ZIL devices does not constitute system failure or data loss. Data loss can only occur if all ZIL devices fail and simultaneously an unrelated system outage occurs before ZIL records are written out to the SATA device pool (a window of about five seconds).

Projects that require the highest availability can configure ZIL Flash devices in pairs to be mirrored. Implementations that need to prioritize absolute lowest cost at the expense of some availability can configure a single ZIL device with knowledge of the simultaneous failure exposure. In each case, the integrity of committed data remains protected — Oracle Solaris ZFS is a transactional file system eliminating issues related to a partially-consistent non-volatile file system state.

SAS Expander

Sun Storage 7000 Unified Storage Systems use the industry-standard Serial Attached SCSI (SAS) protocol as a means of connecting storage controllers to disk and Flash devices. The devices associated with a particular storage system are grouped behind one or more SAS expanders. In this configuration, the SAS expanders are essentially switches that form connection vertices in a SAS topology. A typical SAS-1 expander offers 36 individual connections (phys). Some of these connections attach to devices. Other connections are grouped into ports (and then connectors) that can be used to link downstream to another expander or upstream to a SAS HBA in a storage controller.

Sun storage enclosures from Oracle such as Oracle's Sun Storage J4400 array provide redundant SAS expander paths. This architecture helps ensure that a single SAS expander is not a single point of failure for the system or solitary path to particular disks. In the Sun Storage J4400 array, two SAS expanders, each capable of accessing the 24 enclosed 3.5" disk devices, are provided on separate controller field replaceable units (FRUs) and connect to a passive backplane. If a particular SAS expander fails, the storage software transparently utilizes built-in Oracle Solaris Multipath I/O (MPXIO) to access the target device through another valid path that leads to a different SAS expander. In addition, the Sun Storage J4400 array supports independent replacement of SAS expanders by replacing the controller FRU.

Sun Storage 7000 Unified Storage Systems provide customer and field wiring diagrams that describe proper cabling of SAS enclosures. Utilizing these diagrams, organizations can create deployments such that no expander and expander connection is a single point of failure, and multipath I/O is always operational. Daisy-chain cabling such as that used by Sun Storage 7000 Unified Storage Systems is resilient to the failure of a SAS expander or connection failure along any one path to a particular chain of SAS enclosures. Additional two-way or three-way redundancy across entire JBODs and JBOD daisy-chains can also be configured, as described next.

SAS JBOD

To protect against the pathological failure of an entire JBOD enclosure — such as complete power loss or backplane failure — Sun Storage 7000 Management software permits the optional configuration

of "JBOD Mirroring". Two or three JBOD devices can be included in a mirror. Data is automatically replicated or laid out in a RAID configuration across JBOD devices. As a result, the complete failure of a single JBOD or JBOD chain, including the HBA, expanders, and connections, does not result in the inaccessibility of data. The JBOD Mirroring feature provides a straightforward trade-off between availability and the cost of effective capacity on the system. Selection of this approach is made as part of the storage appliance set-up procedure. Oracle's SAS JBOD devices also provide standard enterprise hardware redundancies such as hot-swap dual-redundant fans and power supplies. Please see specific product technical specifications for complete information on component redundancies.

SAS Host Bus Adapter

The SAS Host Bus Adapter (HBA) cards used by Sun Storage 7000 Unified Storage Systems typically offer two physical x4 SAS ports in a PCI Express low profile form factor. Redundancy can be created by utilizing Oracle Solaris MPXIO. In a system such as Oracle's Sun Storage 7410 system with two SAS HBAs and two JBOD daisy-chains, one port of each card is connected to the "A" expander path of one chain and the "B" expander path of the second chain. Therefore, MPXIO with the appropriate cabling also provides resilience against the failure of a SAS HBA. As shown in Figure 1, if a single HBA fails, the other HBA remains able to access all of the devices in either daisy chain.

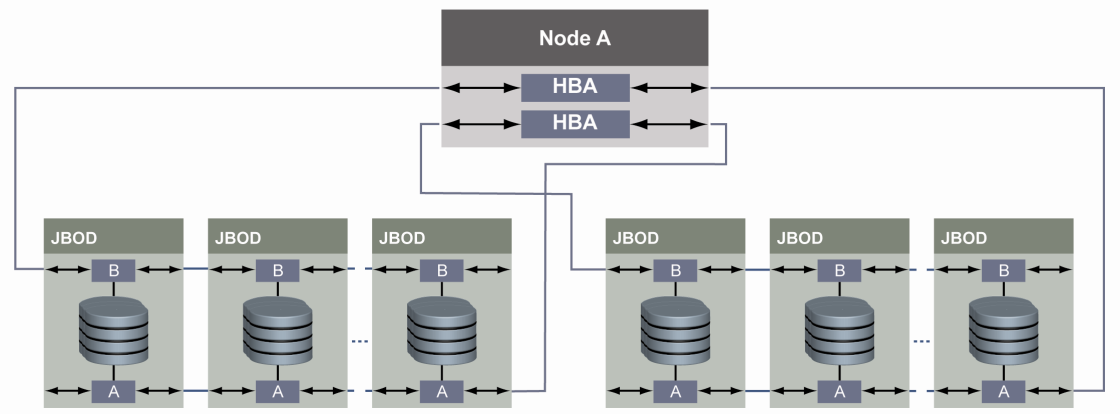

Figure 1. Redundant paths can be created by utilizing two SAS HBAs.

Storage Controller

Sun storage appliances from Oracle also provide the ability to configure redundant storage controllers — the head node including CPUs, memory, and the system software — in an active-active or active-passive cluster. The Sun Storage 7410 system supports two clustered nodes, that can be purchased together, or upgraded from a single controller to a clustered pair after the first controller has been deployed. Clustered controllers are connected together by means of a Sun cluster card from Oracle that provides three redundant communication links for exchanging heartbeats and configuration data. This method provides redundancy between controllers and not along the client data path. Oracle's storage clustering architecture is designed to be as simple as possible, to try to ensure the highest reliability of the cluster implementation.

A clustered storage system has multiple Oracle Solaris ZFS Hybrid Storage Pools. The Sun Storage 7410 system provides at most two. Each pool is assigned a default owner. If both cluster peers are active, the controller that services the pool becomes the owner.

In the event that both pools are assigned the same owner, the cluster is considered active-passive. In an active-passive scenario, the second controller sits idle waiting to take over upon a fault event. If each pool is assigned a different owner, the cluster is active-active. In the active-active scenario, each controller actively services one pool, and is able to take over for the other controller and its pool.

The take-over process that happens automatically upon loss of the heartbeat over all three communication links, or at the request of an administrator, triggers the following actions:

- Import of the Oracle Solaris ZFS Hybrid Storage Pool by the new controller
- Sharing of the related NFS, CIFS, iSCSI, HTTP/DAV, and FTP resources
- Activation of the failed set of IP addresses by means of a set of gratuitous ARP / IP DAD messages sent by the new controller

Exact take-over time varies based on the number of devices, characteristics of the network infrastructure, and associated client-side protocol retry intervals. However, typical take-over times for CIFS and NFS clients are in the 30 to 90 second range for most implementations.

Although each storage controller in a cluster is connected to the other by means of three redundant communication links, Oracle utilizes SAS Zoning to provide protection against a so-called cluster "split brain". A split-brain scenario refers to the generic notion of a cluster in which all explicit communication channels between cluster nodes fail, yet the cluster nodes remain operational. In this situation, each node concludes that the other node failed.

SAS Zoning is a feature of industry-standard SAS expanders whereby a particular SAS HBA initiator is granted permission to access a particular set of devices by means of a zone map. In a unified storage cluster, prior to importing an Oracle Solaris ZFS Hybrid Storage Pool on cluster take-over, the accessing controller reprograms the SAS zone map as a single expander transaction, granting itself access to the new devices and denying access to the other controller's initiators. Therefore, if all three cluster communication links fail but both controllers are still alive, only one can succeed in reprogramming the zone map and actually be able to successfully import the Oracle Solaris ZFS pool corresponding to those devices.

Enterprise storage implementations also replicate data between storage controllers as a means of providing resiliency against controller failure as well as larger-scale outages that affect the entire datacenter. Sun Storage 7000 Unified Storage Systems provide remote replication as part of the standard suite of software features for these type of disaster-recovery scenarios. Remote replication is asynchronous, transactionally-consistent, incremental copying of the state of one or more Oracle Solaris ZFS datasets from one storage appliance to another, over an IP network.

The storage appliance management software allows administrators to execute replication manually, scripted, according to a schedule (e.g. "every hour," "once per day at 5pm"), or as fast as the underlying link allows. Each replication transmission consists only of the transactional differences

between the sender's and receiver's copy of the dataset. Therefore once replication has occurred once, new transmissions are as efficient as the write workload can allow.

Any number of replication schedules can be configured on one appliance, and any Sun Storage 7000 Unified Storage Systems can replicate to and from any other appliance, regardless of the appliance type or the underlying storage configuration. In addition, transitive replication flows can be created, such as "A replicates to B, B replicates to C, C replicates to D, and so on." A replicated dataset can be made active and shared by the receiver (a "take-over") or by the sender (a "push-over"). Take-overs and push-overs of replicated datasets are performed by a human administrator or as part of a disaster recovery script.

Network Interface

When data is delivered from the Oracle Solaris ZFS layer to a network protocol such as NFS, CIFS, or iSCSI, Sun Storage 7000 Unified Storage Systems also make provisions for redundancy at the level of network interface cards, network wires and connectors, and network switches outside the system. The two primary redundancies that administrators can configure are use of the Link Aggregation Control Protocol (IEEE LACP), and IP Multipathing (IPMP). These two network redundancy mechanisms are complementary and may be configured simultaneously.

The LACP protocol permits multiple network physical datalinks (phys, or physical ports on the actual NIC) to be grouped together into a single aggregation, a logical datalink sharing a single logical Ethernet MAC address and able to utilize the bandwidth of the aggregated links. At the same time, an LACP aggregation can survive the failure of N-1 of its N aggregated ports, and still provide service, although at reduced performance. LACP must also be configured for the aggregated ports on the network switch, and thus does not span across switches.

The IPMP feature permits a group of IP addresses to be shared across a group of logical datalinks (these links may be individual physical ports or LACP aggregations of physical ports). IP connections are distributed across the datalinks in the group in round-robin fashion, and if an underlying datalink fails, the IP addresses that were assigned to this group member are immediately moved to another active group member. IPMP performs failure detection using both the link up/down detection capability of the underlying NIC, as well as an optional ICMP probe (ping) mechanism to a set of test addresses that can detect external switch/route failures. For this reason, LACP and IPMP are often useful to deploy in concert, with LACP providing increased bandwidth and N-1 redundancy of the connection to each network switch, and IPMP grouping together the LACP aggregations themselves, thus providing resiliency against the failure of a particular network switch or its path.

Generic RAS Features

By simplifying maintenance procedures, IT organizations can increase efficiency, avoid administrative error, and improve the uptime of storage systems. As described in this chapter, Sun Storage 7000 Unified Storage Systems offer enterprise-class features including intuitive management tools, phone home, lights out management, alerting, and rolling upgrade capabilities.

Hardware View

Sun Storage 7000 Unified Storage Systems provide a Maintenance-Hardware view to display the real-time status of the physical system and the various customer replaceable unit (CRU) and FRU elements. Images of each storage controller and connected JBOD storage devices are displayed. In addition, each physical chassis can be assigned a user-defined label that matches a sticker used in a datacenter or a rack location. Users can expand the view for a particular controller or JBOD to see front, rear, and top graphics of each physical chassis, as well as contained elements such as CPUs, DIMMs, fans, power supplies, disks, and PCI Express expansion cards.

The graphics within the Maintenance-Hardware view are interactive. Selecting a disk, DIMM, CPU, or other element in the picture highlights the description of that element. An FMA fault diagnosis turns the element's status icon and graphic to red. If an element is removed or added on the system — such as hot-plugging a disk — the graphics change in real-time to reflect the status of the physical system. Many Sun systems have blue or white indicator LEDs, including chassis and sub-components such as disks. Buttons are used within the management interface to represent these indicator LEDs. Selecting a button can enable and disable the indicator LED on the real component. Once activated, the graphical representation blinks along with the actual LED. Each enumerated element on the Maintenance-Hardware view includes an additional button that can be used to display the element part number, serial number, and other attributes.

Problems View

The Maintenance-Problems view displays the real-time status of the FMA diagnosis system. Each entry within this view indicates a fault that the diagnosis system believes to be present and not yet repaired. Faults are labeled with the SUNW-MSG-ID, UUID, and time of diagnosis. Administrators and service personnel can expand the entries in the problem list to view details such as the human-readable description of the problem, automated reaction if applicable, impact of the problem, and suggested corrective action. In addition, a Web interface offers a direct link to the most recent knowledge article on sun.com.

For each suspect component in the diagnosis, the relevant FRU is called out — using the same labels that appear in the Maintenance-Hardware view. At the time the diagnosis is made, the system checks if a fault LED is available on the physical system for that FRU. If available, the appropriate fault LED is illuminated and the maintenance-hardware view is updated with a red highlight for the FRU. For each new problem diagnosis, an alert is also posted to the alerts subsystem of the storage unit. Once a repair procedure is performed, removing the faulty FRU from the system, the problem is automatically closed and the entry is then removed from the problems list. It is important to note that some FRUs, notably PCI Express cards, do not have electronically readable serial numbers. As such, a manual "repair" button is provided to indicate that the repair of one of these components is complete.

Log View

The appliance management software keeps a set of persistent log files to provide a historical view of key events, including alerts, faults, system errors (includes legacy items not supported by FMA at this

time), and an audit trail. Each log file tab displays the most recent entries, provides the ability to browse through the entire set of log items, and supports sorting data by fields such as the entry time. Log file types include the following:

- The `alerts log` provides a history of all alerts emitted on the system.
- The `faults log` provides a history of every FMA problem diagnosis. Unlike the Maintenance-Problems view, the `faults log` includes entries for issues that have long since been repaired.
- The `audit trail` provides a history of administrative actions such as logins, the creation of shares, system reboots, and so forth.
- The `system log` provides access to underlying Solaris kernel error messages that have not yet been converted to use FMA. These messages are akin to the information in the traditional Solaris `messages (4)` file. The system log is not needed for common failure modes, and in general its content is unstable and requires developer-level knowledge.

Configuration and Services

The appliance software stack consists of the OpenSolaris kernel and a set of services. Oracle Solaris provides systems with an underlying Oracle Solaris Service Management Facility (Oracle Solaris SMF) to start, stop, and manage services. Please see `smf(5)` for an overview.

Oracle Solaris SMF starts services in parallel according to topological ordering of their dependencies, helping speed boot times. In the event of a service failure, Oracle Solaris SMF automatically restarts services based on specified dependencies. For example, if an ftp server core dumps due to a software defect, Oracle Solaris SMF automatically restarts the service. If a service is unable to come online, Oracle Solaris SMF places the service in maintenance mode. The deactivation of the service triggers a fault management diagnosis and the service is labeled as defective. This diagnosis result then drives the usual FMA actions within the storage system.
Actions include:

- visual alert notification
- user-defined alert triggers (e-mail, SNMP trap)
- update of the Maintenance-Problems display
- dispatch of a phone home call if Automated Service Response (ASR) is enabled

User-configurable services such as DNS are seen within the Configuration-Services view. Unlike a stock system running Oracle Solaris, the administrative interface on Sun Storage 7000 Unified Storage Systems provides a simplified view of configurable services. Services that are essentially implementation details of the operating system —such as the FMA service itself — are not presented. The appliance software also coalesces all services that implement a particular appliance facility. For example, the eight individually restartable services that comprise what one thinks of as "NFS" are presented as a single service composite.

The Configuration-Services view presents the administrator-visible service composites along with one of the following status descriptors:

- online
- offline (starting up, or waiting on a dependency)
- disabled (by an administrator)
- maintenance

The appliance interface provides administrators with the ability to request the restart of a component in the maintenance state. This feature can be helpful if the software defect was related to a particular configuration setting that has now been modified. Finally, one or more log files are kept for each service composite, such that legacy error messages (i.e. not FMA telemetry) can be quickly located for a given service. These logs are visible in the Configuration-Services view for each service. A service composite has one log file capturing the stdout/stderr output of each subcomponent and may also have other custom logs. For example, an HTTP error log is kept for the HTTP and WebDAV services and made available in the configuration-services view.

Configuration Backup

The appliance configuration can be backed up and exported to speed recovery in disaster recovery situations.

A configuration backup includes the following items:

- Metadata associated with the system as a whole, such as settings for NTP, NIS, LDAP, and other services
- Network device, datalink, and interface configuration
- User accounts, roles and privileges, preferences, and local user passwords
- Alerts and thresholds and their associated rules

However, a configuration backup does not include the following items:

- User data (shares and LUNs) — user data must be backed up separately, using NDMP backup software, snapshots, and/or remote replication
- Metadata directly associated with user data, such as snapshot schedules, user quotas, compression settings, and other attributes of shares and LUNs
- Analytics and logs — events can be redirected to external SNMP trap receivers or e-mail destinations using Alerts rules.
- System software — the system software is automatically backed up as part of the System Update capability

Alerts View

Oracle's appliance software provides a unified framework for publishing alerts from various software subsystems such as Oracle Solaris ZFS, FMA, and other key areas. Every fault diagnosis issued by FMA is automatically posted as an alert. However, alerts also include non-fault events such as the successful completion of a resilvering operation.

The Configuration-Alerts view can be used to create a system-wide rule list of alerts that can be dispatched to particular back-end mechanisms. The current set of supported back-end services includes SMTP (e-mail), SNMP (trap), and a set of actions to suspend and resume Sun Storage 7000 Unified Storage Systems DTrace Analytics worksheets and datasets. More information regarding DTrace Analytics worksheets and datasets can be found in the section titled "DTrace Analytics".

Using the Configuration-Alerts view, administrators can arrange arbitrary notification groupings. For example, the alias "admins@foo.org" can be set to receive an e-mail if a disk is removed or a cluster fails over. Regardless of the alert ruleset, FMA diagnosis events are sent to the phone home mechanism and appear on the Dashboard and Masthead — top part of the browser screen — of all active users. The set of alertable events are listed in Table 2.

TABLE 2. ALERT EVENTS

CATEGORY	EVENT	COMMENTS
Cluster	Cluster I/O link down	When a clustered system detects its cluster I/O link is offline
Cluster	Cluster I/O link failed	When a clustered system detects its cluster I/O link has failed
Cluster	Cluster I/O link up	When a clustered system detects its cluster I/O link is online
Cluster	Unexpected peer error occurred	When an unexpected error occurs on a cluster peer in a clustered system
Cluster	Communication to peer lost	When a clustered system loses communication with the cluster peer
Cluster	Incompatible software version on peer	When a clustered system detects that the software version on the peer is incompatible with the software on the local system
Cluster	Older software version on peer	When a clustered system detects that the software version on the peer is an older version than the local system
Cluster	Cluster peer panicked	When a clustered system detects the cluster peer has panicked
Cluster	Failed to set SP root password on cluster peer	When a clustered system fails to set the SP root password on the cluster peer
Cluster	Cluster rejoin failed on peer	When a clustered system detects its peer has gone down
Cluster	Cluster rejoin completed on peer	When a clustered system detects its peer has rejoined
Cluster	Cluster peer lost communication token	When a cluster communication token is lost on the local system
Cluster	Cluster rejoin failed	When a system fails to join the cluster
Cluster	Cluster rejoin completed	When a clustered system successfully joins the cluster
Cluster	Cluster takeover complete	When a clustered system takeover completes successfully
Cluster	Resources import failed during cluster takeover	When a resource fails to import during a cluster takeover
Cluster	Local cluster communication token lost	When a cluster communication token is lost on the local system

CATEGORY	EVENT	COMMENTS
Faults	Hardware fault diagnosed	When an FMA diagnosis of a hardware fault is made
Faults	Hardware defect detected	When an Oracle Solaris SMF service enters the maintenance state
Hardware	Appliance up	When the appliance boots up
Hardware	Appliance down	When the appliance is rebooted, reset, or power off is requested
Hardware	Chassis connected	When a new JBOD is cabled to the system
Hardware	Chassis removed	When an existing JBOD is disconnected from the system
Hardware	Component inserted	When a hot-plug component such as a fan is inserted
Hardware	Component removed	When a hot-plug component such as a fan is removed
Hardware	Disk inserted	When a disk is hot-plug inserted
Hardware	Disk removed	When a disk is hot-plug removed
Hardware	Service processor offline or unavailable	When the service processor is reset or physically removed
Hardware	Service processor online after outage	When the service processor comes online after an outage
Hardware	Failed to set SP root password on service processor	When the system fails to set the SP root password on the service processor
NDMP	Invalid NDMP restore	When an invalid NDMP restore is attempted
NDMP	Backup started	When an NDMP backup begins
NDMP	Backup finished	When an NDMP backup ends
NDMP	Restore started	When an NDMP restore begins
NDMP	Restore finished	When and NDMP restore ends
Phone home	Support bundle build failed	When a support bundle build fails
Phone home	Support bundle sent failed	When a support bundle fails to be sent
Phone home	Support bundle upload failed	When a support bundle fails to upload
Replication	Receive failed (cancelled)	When a particular data replication activity fails on the target
Replication	Receive failed (all others)	When a particular data replication activity fails on the target
Replication	Receive started	When a particular data replication activity begins on the target
Replication	Receive finished	When a particular data replication activity ends on the target
Replication	Send failed (cancelled)	When a particular data replication activity fails on the source
Replication	Send failed	
(all others)	When a particular data replication activity fails on the source	
Replication	Send started	When a particular data replication activity begins on the source
Replication	Send finished	When a particular data replication activity ends on the source
Services	Service failures	When a software service fails
Threshold	Threshold exceeded	When a watched statistic exceeds a defined threshold
Threshold	Threshold error	When a watched statistic is exceeded for a prolonged period
Threshold	Threshold normal	When a watched statistic returns to normal
Oracle Solaris ZFS	Resilver started	When an Oracle Solaris ZFS RAID or mirror resilvering operation begins

CATEGORY	EVENT	COMMENTS
Oracle Solaris ZFS	Resilver finished	When an Oracle Solaris ZFS RAID or mirror resilvering operation ends
Oracle Solaris ZFS	Scrub started	When an Oracle Solaris ZFS disk background scrubbing operation begins
Oracle Solaris ZFS	Scrub finished	When an Oracle Solaris ZFS disk background scrubbing operation ends
Oracle Solaris ZFS	Hot spare activated	When an Oracle Solaris ZFS hot spare is activated (e.g. after a disk fault occurs)

The Alerts view also permits system administrators to configure a global list of threshold alerts corresponding to system performance metrics. Associated thresholds can be used to trigger notifications using any of the backend services. The events generated by these alert rules correspond to those of category "threshold" in Table 2. The statistics that may appear on the watchlist are the same as those provided by DTrace Analytics, described below.

Each threshold alert rule defines a set of criteria including:

- A statistic to be monitored (e.g. "NFSv4 operations per second")
- A threshold value and direction (e.g. "above 100" or "below 1000")
- An optional time constraint (e.g. "only between 9am and 5pm on weekdays")

The "threshold exceeded" alert posts when the statistic first violates the constraint defined by the alert rule, or when the statistic continues to meet the watchlist criteria for a defined time period. For example, "repost every five minutes while NFSv4 operations per second remains above 100". Similarly, the "threshold normal" alert can be set to post if the statistic no longer meets the watchlist criteria for a defined time period. As in the example, "post normal if NFSv4 operations drops below 100 and stays there for one minute". Therefore, the watchlist provides full support for hysteresis by not posting a flurry of alerts when a statistic jitters up and down around the boundary defined by the administrator's watchlist rule. The "threshold error" alert fires when the statistic cannot be sampled.

An example of the e-mail message content generated by the SMTP alerts back-end for a threshold event is as follows:

```
SUNW-MSG-ID: AK-8000-TT, TYPE: Alert, VER: 1, SEVERITY: Minor
EVENT-TIME: Tue Jan 15 22:23:15 2008
PLATFORM: i86pc, CSN: 0000000000, HOSTNAME: catfish
SOURCE: svc:/appliance/kit/akd:default, REV: 1.0
EVENT-ID: 0c039cf6-29c9-6bf3-8588-f701edb0f1a4
DESC: cpu.utilization threshold of 1 is violated. Refer to
http://sun.com/msg/AK-8000-TT for more information.
AUTO-RESPONSE: None.
IMPACT: The impact depends on what statistic is being monitored.
REC-ACTION: The suggested action depends on what statistic is being
monitored.
CLASS: alert.ak.xmlrpc.threshold.violated
LIMIT: 1
SOURCE: svc:/appliance/kit/akd:default
STATNAME: cpu.utilization
THRESHOLDID: 4a45da64-9c12-4597-fbfe-a652eeb7f647
TIMESTAMP: 1200435795 (Tue Jan 15 22:23:15 2008)
UUID: 0c039cf6-29c9-6bf3-8588-f701edb0f1a4
```

The first section of the message is identical to the Oracle standard text format for all fault management events. The second section includes alert-specific details. In this example, this section includes the monitored statistic.

Alert rules, including threshold alerts, can also be used to suspend or resume DTrace Analytics worksheets and datasets. The use of thresholds to activate DTrace Analytics enables administrators to define sophisticated fine-grained instrumentation questions to be answered only when an aberrant condition is present on the system. For example, these features can be used in combination to define a request such as "When the number of CIFS Ops/second exceeds a given value for five minutes activate a DTrace Analytics query to record the CIFS client IP addresses and the CIFS file pathnames that are being accessed." DTrace Analytics and threshold alerts provide unprecedented 24 x 7 systems monitoring and ensure that the right data is gathered even if no human administrator is available.

Dashboard Status

A graphical real-time dashboard provides information on key system capacity and performance metrics. This view is the default view when logging into the appliance user interface. The dashboard displays the real-time status of the storage system. For Sun Storage 7000 Unified Storage Systems, the dashboard includes relevant metrics such as NFS, iSCSI, and CIFS operations per second, as well as memory and storage free and consumed space. The dashboard captures data over time, displaying metrics across the past week, day, hour, and in real-time, allowing administrators to quickly compare

the current performance of the system to the workload over any of these recent time periods. The dashboard also includes a scrolling list of recent alerts on the system, and provides a click-through to DTrace Analytics.

DTrace Analytics

Sun Storage 7000 Unified Storage Systems include unique real-time graphical DTrace Analytics, based on Oracles award-winning Oracle Solaris DTrace technology. DTrace Analytics provides the ability to monitor system performance statistics over time, observe the results as a set of real-time updating graphs in a Web browser, and most importantly "drill down" interactively on the data. Using DTrace Analytics, administrators can create new performance queries on-the-fly and sample the results with DTrace at low overhead to understand the live running system, safely, in production, as it processes the actual customer workload. DTrace Analytics can be used for performance analysis, latency debugging, capacity planning, and is intended for both customers and service personnel. A full treatment of DTrace Analytics is beyond the scope of this document. However, a complete list of the aspects of the system that can be monitored by DTrace Analytics and also used in Threshold Alert rules can be found in Table 3.

TABLE 3. ELEMENTS MONITORED BY DTRACE ANALYTICS

CATEGORY	ELEMENT	DRILLDOWN SUBCATEGORIES
CPU	CPUs	Percent utilization
CPU	Kernel spins	Type of synchronization primitive, CPU identifier, raw
CPU	Percent utilization	CPU mode, CPU identifier, application name, process identifier, user name, raw
Cache	ARC accesses	Hit/miss, file name, project, share, raw
Cache	ARC adaptive parameter	Raw
Cache	ARC size	Component, raw
Cache	ARC target size	Raw
Cache	DNLC accesses	Hit/miss, raw
Cache	DNLC entries	Raw
Cache	L2ARC accesses	Hit/miss, file name, project, share, raw
Cache	L2ARC I/O bytes	Type of operation, raw
Cache	L2ARC size	Raw
Data movement	NDMP bytes transferred to/from disk	Type of operation, raw
Data movement	NDMP bytes transferred to/from tape	Type of operation, raw
Data movement	NDMP file system operations	Type of operation, raw
Data movement	NDMP jobs	Type of operation, raw
Data movement	Shadow migration bytes	File name, project, share, raw
Data movement	Shadow migration operations	File name, project, share, latency, raw
Disk	Average number of I/O operations	State of operation, disk, raw
Disk	Disks	Percent utilization

CATEGORY	ELEMENT	DRILLDOWN SUBCATEGORIES
Disk	I/O bytes	Type of operation, disk, raw
Disk	I/O operations	Type of operation, disk, size, latency, offset, raw
Disk	Percent utilization	Disk, raw
Disk	ZFS I/O bytes	Type of operation, pool name, raw
Disk	ZFS I/O operations	Type of operation, pool name, raw
Memory	Dynamic memory usage	Application name, raw
Memory	Kernel memory	Kmem cache, raw
Memory	Kernel memory in use	Kmem cache, raw
Memory	Kernel memory lost to fragmentation	Kmem cache, raw
Network	Device bytes	Direction, device, raw
Network	Interface bytes	Direction, interface, raw
Network	IP bytes	Hostname, protocol, direction, raw
Network	IP packets	Hostname, protocol, direction, raw
Network	TCP bytes	Client, local service, direction, raw
Network	TCP packets	Client, local service, direction, raw
Protocol	CIFS operations	Type of operation, client, file name, share, project, latency, size, offset, raw
Protocol	FTP bytes	Type of operation, user name, file name, share, project, client, raw
Protocol	HTTP/WebDAV requests	Type of operation, response code, client, file name, user agent, size, latency, raw
Protocol	iSCSI bytes	Initiator, target, project, LUN, client, raw
Protocol	iSCSI operations	Initiator, target, project, LUN, type of operation, latency, offset, size, client, raw
Protocol	NFSv2 operations	Type of operation, client, file name, share, project, latency, size, offset, raw
Protocol	NFSv3 operations	Type of operation, client, file name, share, project, latency, size, offset, raw
Protocol	NFSv4 operations	Type of operation, client, file name, share, project, latency, size, offset, raw
Protocol	SFTP bytes	Type of operation, user name, file name, share, project, client, raw

Phone Home

Sun Storage 7000 Unified Storage Systems provide built-in support for product registration and phone home through the following services:

- Sun Connection Readiness Kit (SCRK) services
- Automated Service Response (ASR) services
- SupportFiles service deployed on Oracle's Web site

The HTTPS protocol is utilized to transmit product registration and phone home messages. As such, the storage system must be configured to allow outbound HTTPS connections or use an HTTPS proxy within the deployment environment. Phone home is enabled by default but may be disabled at the discretion of the customer.

The phone home subsystem sends the following messages:

- Registration messages are sent when the appliance is initially configured
- Heartbeat messages are sent every 24 hours while the appliance is functioning
- Fault messages are sent whenever FMA diagnoses a hardware or software problem
- Supportfiles bundles are sent at the request of an administrator or service personnel

In addition, Sun Storage 7000 Unified Storage Systems are compatible with an optional Sun Service Delivery Platform (SDP) from Oracle. Sun SDP is an appliance that can be connected in a datacenter to provide Oracle support engineers with remote access and fault aggregation for a variety of Oracle products. Sun Storage 7000 Unified Storage Systems export the standard Sun Fault Management SNMP MIB and trap mechanisms necessary to communicate with the SDP.

Please see the SCRK documentation for a detailed description of the remote registration process and the ASR documentation for information regarding heartbeat and fault messages. Supportfiles bundles are unique to Sun Storage 7000 Unified Storage Systems. However, these systems exploit Oracle's remote transfer portal at http://*supportfiles.sun.com*.

At the request of an administrator or service provider, or as the result of an FMA fault diagnosis, Sun Storage 7000 Unified Storage Systems can automatically construct a bundle of information to be transferred to back-line support. The storage system can also automatically upload that bundle to the supportfiles portal. The supportfiles bundle helps ensure that back-line support receives a full set of information relevant to the problem in question without having to rely on a series of manual steps. In addition, since the bundle capability is built into the software stack, enhancements can be delivered over time in concert with the software updates. As a result, new software features become serviceable at introduction.

The supportfiles bundle format is a compressed archive of relevant data files, and utilizes a naming convention that includes an FMA problem UUID. The elements that can be found in a supportfiles bundle are listed in Table 4.

TABLE 4. COMPONENTS OF THE SUPPORTFILES BUNDLE

ITEM	INCLUDED FOR	DETAILS
Audit log files	system failures, service failures	The system audit log includes recent administrative actions, such as modification of service properties, creation and deletion of file systems, and changes to network configuration. Also, all service actions performed at the underlying Solaris shell are audited, and these internal audit logs are included in the bundle. Please note these actions are only to be undertaken by qualified support personnel.
Disk and flash SMART data	FMA diagnosis	Following an FMA fault diagnosis that includes as a suspect any SATA disk or Flash device FRU, the system will attempt to capture relevant SMART data from the device, if possible, and include this in the supportfiles bundle for this problem. The data includes the SMART attribute thresholds, SMART error log, and other log pages of interest such as the vendor defect log pages, etc.

ITEM	INCLUDED FOR	DETAILS
Fault management events	FMA diagnosis	Includes XDR-encoded name-value lists of the complete list.suspect event, all embedded fault or defect events, and all of the error events referenced by this diagnosis (fmdump -e -U). In addition, the hardware chassis details (part number, serial number, firmware revision, etc.) are included for each suspect FRU.
Hardware chassis configuration	all bundles	All bundles include what is effectively an XML representation of all of the data corresponding the appliance Maintenance/Hardware view, i.e. a list of all hardware enclosures, their chassis serial numbers, and a summary of the enclosed components such as PCI Express cards, disks, SAS expander/controllers, Ethernet NICS, etc.
JavaScript error console	browser errors	Oracle's appliance browser interface includes a built-in JavaScript Error Console that captures JavaScript stack traces when uncaught exceptions are thrown or AJAX XML-RPC errors are detected. If the browser remains connected after such an error, the appliance software will transmit the contents of the error console back to the appliance for inclusion in the bundle.
JBOD SES logs and status	FMA diagnosis, system failures	Oracle's SAS storage enclosures, such as the Sun Storage J4400 array, include a SES management process that records the health status and a log of errors detected by the enclosure's sensors and the SAS expander. If the SAS expander and SAS HBA are still connected, these can be captured and included in a supportfiles bundle. (If the connection has failed, a serial console can be connected to the JBOD by a service engineer in order to retrieve this information.)
Kernel crash dump	system failures	Crash dumps are included in a bundle created following any time the system boots and discovers that it has previously panicked.
Process core file	service failures	Core files are included in a bundle created following any time a service enters the maintenance state and one or more core files corresponding to that service are found in the core file repository.
SAS HBA trace buffer	FMA diagnosis, system failures	Oracle SAS HBAs include a trace buffer of key SAS events and errors as recorded by the firmware on the HBA itself. Following an FMA diagnosis or system failure, a copy of the trace buffer is captured and included in the supportfiles bundle.
Service configuration	all bundles	An XML representation of customer-tunable service configuration (e.g. options for CIFS, NFS, and so forth) is included in all bundles. Properties that include encoded secure data, such as iSCSI passwords, are not included in the bundle for security reasons.
Service log files	service failures	Service log files (including both SMF logs and any service-specific binary log files) are included in a bundle created any time a service enters the maintenance state.
System event log (SEL)	system failures	A copy of the IPMI SEL is included by bundles created when the system boots following a hardware failure or software panic.
System service tag	all bundles	A copy of the appliance service tag XML is included in all bundles.
System software versions	all bundles	All relevant software version numbers are included in all bundles.

Service Tags

A Service Tag is built into Sun Storage 7000 Unified Storage Systems, simplifying access to key product metadata such as the appliance serial number, primary hardware chassis serial number, appliance type,

and software version number. The service tag can be retrieved by the Service Tag client program — downloadable from *http://sun.com* — and other service offerings developed by Oracle. The Service Tag client can also be used to register an existing Sun Storage 7000 Unified Storage System with the Sun Inventory Channel helping organizations add storage systems to their list of systems in the portal. An example Service Tag returned is as follows:

```
<service_tag>
       <instance_urn>urn:st:383f81e4-a59f-4b8f-e4fc-
bae031258410</instance_urn>
       <product_name>Sun Storage 7210</product_name>
       <product_version>2008.06.11.0.0,1-1.0</product_version>
       <product_urn>urn:uuid:afcee0df-edd1-11db-8c3c-
080020a9ed93</product_urn>
       <product_parent_urn>null</product_parent_urn>
       <product_parent></product_parent>
       <product_defined_inst_id></product_defined_inst_id>
       <timestamp>2008-07-08 16:26:00 GMT</timestamp>
       <container>global</container>
       <source>akd</source>
</service_tag>
```

Sun Storage 7000 Unified Storage Systems return an `instance_urn` corresponding to the Appliance Serial Number (ASN). The `product_urn` refers to the internal product registry identifier. Service Tag discovery and response is enabled on Sun Storage 7000 Unified Storage Systems by default. However, this feature can be disabled at the discretion of the administrator.

SNMP

Sun Storage 7000 Unified Storage Systems provide built-in support for heterogeneous systems management through SNMP, providing compatibility with a variety of enterprise management tools and open-source tools. The SNMP implementation supports both SNMPv2 and v3 and can be secured as necessary through the SNMPv3 security mechanisms. These storage systems provide a Sun Fault Management MIB and trap mechanism used to issue a trap for every FMA diagnosis of a problem on the system and to provide MIB browsing capability for all diagnosed faults. The list of MIBs supported by the Sun Storage 7000 Unified Storage System SNMP stack are included in Table 5.

TABLE 5. SUN STORAGE 7000 UNIFIED STORAGE SYSTEMS SNMP MIB SUPPORT

MIB NAME	REFERENCE	DESCRIPTION
SNMP Entity MIB	RFC 2737	SNMP entPhysical table
SNMP-FRAMEWORK-MIB	RFC 2271	SNMP framework attributes
SNMP-MPD-MIB	RFC 2572	SNMP message processing
SNMPv2-MIB	RFC 1907	System and SNMP groups
SUN-AK-MIB	Sun Storage 7x10 Documentation	Appliance identity and version strings

MIB NAME	REFERENCE	DESCRIPTION
SUN-FM-MIB	Sun Fault Management PRM	FMA diagnosis traps and fault browsing
SUN-STORAGE-MIB	Sun Storage 7x10 Unified Storage System Documentation	Oracle Solaris ZFS projects, shares, capacities

Lights Out Management

Sun systems from Oracle, including the new storage controllers that drive Sun Storage 7000 Unified Storage Systems, provide complete support for lights out management through an integrated service processor (SP). The SP is in effect a small, separate computer system inside of the system enclosure that provides:

- Remote power cycle capability
- Console redirection and console access over the network
- Fault management for hardware faults that prevent the operating system kernel from even starting
- Remote control through the industry-standard IPMI protocol

For Sun Storage 7000 Unified Storage Systems, the primary uses of the SP are remote power control through IPMI, remote console access, and the ability to access the System Event Log (SEL). These functions can even be utilized in the event that the system cannot even boot the operating system. Oracle's SP Lights Out Manager (LOM) is also compatible with industry-standard heterogeneous management tools such as IBM Tivoli Enterprise Console, Microsoft Operations Manager, HP OpenView Operations for UNIX, HP Systems Insight Manager, CA Unicenter Network and Systems Management, and BMC Patrol Enterprise Management.

Upgrade

Sun Storage 7000 Unified Storage Systems include a mechanism to efficiently update the software stack, on-disk formats, configuration properties, and component firmware (SAS HBAs, SATA disks, NAND Flash devices) to new versions. New versions of the software are delivered as a single downloadable image from *http://sun.com*. There are no patches of any kind for Sun Storage 7000 Unified Storage Systems. Instead, administrators use the built-in upgrade feature to upgrade from one version to the next over time — although one may of course skip ahead in the timeline. New software versions are classified as micro, minor, or major releases depending upon whether they contain only bug fixes or new features. Micro releases are designed to be able to be released within 24 to 48 hours if a critical fix is necessary. The upgrade itself is executed in several stages:

- The new software image is retrieved and then uploaded to the system using the Web browser interface or command-line interface. Several update images may be kept cached on the system.
- The upgrade process is initiated by an administrator, who unpacks the new image and prepares all of the new software that will be running on the system. During this stage, the system is still up and running the existing software, and actively servicing requests. Administrators can also specify upgrade options at this time.

- If the upgrade requires a system reboot to take effect, the reboot either takes place when manually initiated by the administrator, or at a pre-determined time specified when the upgrade operation was initiated. If the upgrade does not require a reboot, all administrative sessions will be reset and the new software is now active.
- If the upgrade requires a system reboot, then during the reboot process the remainder of the upgrade will execute. Administrators can select whether disk firmware upgrades take place all at once during system boot, or continue asynchronously in the background, stripe by stripe, after the system is again up and running.
- The system automatically saves previous images, facilitating rollback if an upgrade fails.
- In a clustered system such as Oracle's Sun Storage 7410 system, a rolling upgrade may be performed, wherein one cluster node asks the other to take over, is upgraded, then a fail-back occurs, and then the second upgrade.

The rollback procedure reverts all of the system software and all of the metadata settings of the system back to the state just prior to applying an update. This feature is implemented by taking a snapshot of various aspects of the system before the new update is applied, and rolling back this snapshot to implement the rollback. The implications of rollback are:

- Any appliance configuration changes are reverted and lost. For example, assume a system is running version V, and then is updated to V+1, and then the DNS server is changed. If a rollback is executed, then the DNS server setting modification is effectively undone and removed from the system permanently.
- Conversely, any changes made to user data are not reverted: if an update is completed from V to V+1, and clients then create directories or modify shares in any way, those changes still exist after the rollback.
- If the appliance is running version V, and has previous rollback targets V-1 and V-2, and the system is reverted all the way to version V (thereby "skipping" V-1), then not only are the system software settings and system software for V removed, but also for V-1. That is, after a rollback to V-2, it is as if updates V-1 and V never happened. However, the software upload images for V-1 and V are still saved on the system and can be reapplied (re-executing the update) after the rollback if desired.

If after applying an update, the system is back up and running, either the browser user interface (BUI) or the command line interface (CLI) can be utilized to initiate a rollback to one of two previously applied updates. If the system is not able to run at all after an update, Oracle's Sun Storage systems also provide a fail-safe rollback procedure. The fail-safe rollback is initiated by rebooting the system and, from the system console, selecting one of the fail-safe rollback menu entries from the boot menu. These entries are maintained automatically as updates are applied to the system. If a fail-safe entry is selected, the system will boot using the previous system software, and ask for human confirmation to complete the rollback.

Pulling It All Together

Today, businesses are digitized and maintain a large online presence. Protection of data assets is increasingly vital to business viability. While many enterprise storage management systems provide adequate data protection, most come with limited flexibility and at a high cost – especially for smaller deployments.

Sun Storage 7000 Unified Storage Systems offer enterprise-class storage features while leveraging open source software and industry standard components to lower costs and provide greater innovation. By taking advantage of Oracle's Sun Storage 7000 Unified Storage Systems organizations can realize the following benefits:

- Increase production uptime through a highly reliable architecture built with redundant components and extensive error detection and correction capabilities
- Speed problem resolution time with advance fault management, predictive self-healing, and real-time analytic capabilities
- Improve data integrity and protection with measures that extend beyond traditional storage system features
- Simplify management by providing a clear view of storage system health and alerts

For More Information

About the Author

Mike Shapiro is a Distinguished Engineer in Solaris Kernel Development. He holds a Masters of Science degree in computer science from Brown University. Mike's research and engineering interests are focused on technology to enhance the availability of computer and storage systems, including programming languages and debugging tools for developers, operating system technologies for handling and recovering from software and hardware faults and defects, and tools for administrators and users that improve the user experience.

Related Resources

- Oracle's Sun Storage 7000 Unified Storage Systems
  http://www.oracle.com/us/products/servers-storage/storage/index.htm
- Oracle Solaris ZFS
  http://sun.com/software/solaris/zfs.jsp
- Oracle Solaris DTrace
  http://sun.com/software/solaris/observability.jsp
- Oracle Solaris Predictive Self-Healing
  http://sun.com/software/solaris/availability.jsp

Oracle is committed to developing practices and products that help protect the environment

An Economical Approach to Maximizing Data Availability with Oracle's Sun Storage 7000 Unified Storage Systems
April 2010
Author: Mike Shapiro

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

SOFTWARE. HARDWARE. COMPLETE.