

# SUN™ STORAGE F5100 FLASH ARRAY STORAGE CONSIDERATIONS FOR ACCELERATING DATABASES

White Paper  
September 2009

## **Abstract**

Delivering new levels of performance with low price points and reduced needs for space and power consumption, the Sun Storage F5100 Flash Array provides breakthrough business value as a database accelerator. This white paper describes the array's innovative architectures and technologies, as well as storage deployment considerations, that bring tremendous business benefit to performance-hungry database applications.

## Contents

<b>Delivering Database Performance with Innovative Flash Storage</b> . . . . .	1
Sun™ Storage F5100 Flash Array—accelerating database performance . . . . .	2
Innovative storage system design. . . . .	3
Applications for the Sun Storage F5100 Flash Array . . . . .	4
Performance Considerations. . . . .	5
Software Considerations. . . . .	6
<b>Sun Flash Module Design</b> . . . . .	7
Module components . . . . .	7
Enterprise-quality NAND flash for reliability . . . . .	9
Low power consumption and fast I/O. . . . .	9
<b>Sun Storage F5100 Flash Array Architecture</b> . . . . .	10
Array architecture . . . . .	11
Sun Flash Modules . . . . .	12
LSI 36-port expanders . . . . .	12
Energy storage modules (ESM). . . . .	12
Enclosure features . . . . .	12
Front and rear perspectives . . . . .	13
Chassis design innovations . . . . .	13
Array management — Sun StorageTek CAM Software . . . . .	14
RAS features . . . . .	15
Flash technology considerations. . . . .	16
Supercapacitors . . . . .	16
<b>Deployment Considerations</b> . . . . .	18
Database acceleration deployment considerations . . . . .	18
Performance improvement assessment . . . . .	18
RAS requirements assessment . . . . .	19
System deployment assessment. . . . .	20
Configuration examples . . . . .	21
Simple availability configurations . . . . .	21
Performance tuning . . . . .	23
Index mirroring, migration, operation, and validation. . . . .	24
<b>Summary</b> . . . . .	26
For more information . . . . .	26
<b>Flash Basics</b> . . . . .	27

## Chapter 1

# Delivering Database Performance with Innovative Flash Storage

Enterprises in every industry—from manufacturing, telecommunications and financial services, to retail, insurance and healthcare—rely on fast access to business-critical information that is stored in databases in order to get the job done. With increasingly sophisticated applications being used to solve business problems, analyze data, and track customer relationships, enterprise databases support more users and complex business processes than ever before. However, these growing data volumes and compute-intensive applications are pushing databases to the limit, often slowing operations and increasing response times to users.

Finding ways to accelerate database performance is the key to success. Because the different types of operations performed by databases can stress underlying servers and storage systems, addressing both CPU and I/O bottlenecks is essential. Today, companies take advantage of database partitioning, SQL tuning, query optimization, and clever caching to improve database performance. These strategies, combined with powerful servers with more memory and processor cores using chip multithreading technology (CMT), are helping to alleviate CPU bottlenecks and speed processing throughput.

Yet I/O limitations remain. While spreading and striping I/O operations across multiple hard disk drives with faster interfaces has helped, the fact is hard disk drive technology has failed to keep pace with the transactional and I/O demands of database environments. Indeed, the CPU-to-storage bottleneck is hampering overall system performance—a trend that continues unabated as system performance outpaces disk throughput year over year. With rotational latencies and seek times impacting how fast a disk drive can locate data and begin transferring information to the host, CPUs often sit idle while waiting for the information needed to complete operations. This situation is aggravated by multicore, multsocket servers that have tremendous horsepower just waiting to be tapped.

Advances in storage technology are changing the game and creating new opportunities for storage deployment that can radically impact enterprise database performance. Once found only in MP3 players, cell phones, and digital cameras, Flash technology has moved beyond simple commodity use and found a place in the enterprise storage infrastructure. With breakthrough performance and power characteristics, as well as robust data integrity, reliability, availability and serviceability features, enterprise solid state devices incorporating Flash technology are poised to change the way organizations deploy database solutions.

## Sun™ Storage F5100 Flash Array—accelerating database performance

Continuing its long-standing tradition of innovating datacenter compute and storage solutions, Sun is introducing the revolutionary Sun Storage F5100 Flash Array—a simple, high-performance, eco-efficient solid state storage solution designed to help accelerate database operations. Incorporating enterprise-quality Flash memory modules that provide low latency at an affordable price point, the Sun Storage F5100 Flash Array delivers fast access to critical data and enhances the responsiveness of database applications, without requiring modifications to the application code.

To accelerate database performance, the array significantly reduces I/O bottlenecks by eliminating rotational and head seek delays found in traditional hard disk-based storage systems. Indeed, the Sun Storage F5100 Flash Array accelerates databases with over 1 million I/O operations per second (IOPS) performance and 100X less power and space than traditional disk-based solutions for the same performance level. Exceeding the IOPS performance of over 3,000 disk drives, and more than 4X the performance of other Flash technology-based systems, the Sun Storage F5100 Flash Array provides a cost-effective building block that can deliver breakthrough value for high-performance database applications and solutions.

Using low-latency solid state memory modules, the Sun Storage F5100 Flash Array delivers performance, capacity, and low power consumption in a compact enclosure. With almost 2 TB in a rack-mountable 1U chassis, it offers far greater capacity than a bank of solid state disk devices and at a range of affordable price points. A single array is designed as four separate SAS domains that can be attached to a number of servers. The Sun Storage F5100 Flash Array enables a variety of configurations, so storage architects can design flexible, cost-effective solutions that complement the existing storage infrastructure and meet performance, capacity, and availability goals. In this way the array helps to deliver substantial ROI for high-performance business-critical database applications. Table 1 provides a brief summary of array features.

Table 1. Sun Storage F5100 Flash Array features

Feature	Sun Storage F5100 Flash Array
Storage density	1.92 TB maximum per 1RU array
Performance	Random reads: 1.6 million IOPS (4 KB block size, 32 threads) Random writes: 1.2 million IOPS (4 KB block size, 32 threads) Sequential reads: 12.8 GB/sec (1 MB block size, 32 threads) Sequential writes: 9.7 GB/sec (1 MB block size, 32 threads) Read latency: 405 microseconds (4 KB block size, single thread) Write latency: 282 microseconds (4 KB block size, single thread)
Connectivity	Sixteen x4 mini-SAS ports (four x4 3Gb/s mini-SAS ports per domain)
Reliability	<ul style="list-style-type: none"> <li>• Redundant power supplies and fan modules</li> <li>• Supercapacitor-backed DRAM</li> <li>• StorageTek Common Array Manager for system management</li> </ul>
Power consumption	2.1 Watts per Sun Flash Module; 300 Watts per fully populated array

## Innovative storage system design

With a rack-mountable design that complements other Sun storage and server products, the Sun Storage F5100 Flash Array addresses fast-access database storage requirements at new economic levels. It offers:

- **Low latency.** Flash technology completes I/O operations in microseconds, which places it between hard disk drives and DRAM in terms of access times. Hard disk drives can complete an operation in milliseconds, while DRAM access time is in nanoseconds. Because flash technology contains no moving parts, it avoids the seek times and rotational latencies inherent with traditional hard disk drive technology. As a result, data transfers to and from the Sun Storage F5100 Flash Array are significantly faster than what electro-mechanical disk drives can provide — a single Sun Flash Module can provide thousands of I/O operations per second (IOPS) for write operations and tens of thousands of IOPS for read operations, compared to hundreds of IOPS for hard disk drives. For readers unfamiliar with the basics of flash technology, Appendix A provides a brief introduction.
- **Enterprise-level availability.** Reliability features help to increase availability and meet Service Level Agreement (SLA) targets. The Sun Storage F5100 Flash Array has a relatively small part count and is designed specifically for high reliability, availability, and serviceability (RAS). Sun Flash Modules incorporate features such as wear-leveling, ECC, and block mapping (described in more detail in Chapter 2), and are subject to rigorous quality standards to sustain enterprise-level reliability. An Energy Storage Module (ESM) contains supercapacitor units that help to flush Sun Flash Module metadata safely from DRAM to flash memory in the event of a sudden power loss. Redundant power supplies and fans also enhance reliability, helping to provide continuous operation.

- **Simplified management.** The Sun StorageTek™ Common Array Manager software provides a consistent interface to administer Sun storage products, including the Sun Storage F5100 Flash Array. The software supplies an easy-to-use interface to perform administrative tasks such as configuration, firmware upgrades, maintenance, and device monitoring.
- **Breakthrough value.** With the Sun Storage F5100 Flash Array, Sun delivers innovation that drives value. Based on high-speed, enterprise-quality flash memory modules, the array offers lower cost and lower power consumption per I/O operation, providing an enterprise-ready storage solution that delivers tremendous business value.
- **Flexible configurations and price points.** The Sun Storage F5100 Flash Array can be deployed in virtually any situation that accepts a SAS-attached storage appliance. It can be partially or fully populated, allowing storage architects to design the most cost-effective solution for the application at hand. Each array domain holds up to 20 flash memory modules that provide 480 GB of usable storage as a SAS domain. Configurations are available with either 20 (480 GB), 40 (960 GB) or 80 (1.92 TB) flash modules. To expand capacity and performance, additional flash modules can be added as upgrades to a single domain or across all four domains. When all four array domains are fully populated in the 1U rack, a single array provides a maximum capacity of 1.92 TB.
- **Leading eco-responsibility.** The Sun Storage F5100 Flash Array continues Sun's tradition of eco-responsibility by offering optimal performance as well as performance-per-watt. The solid state flash module operates at low power (approximately 2 watts for each module), which is especially low in comparison to disk devices (which are typically around 12 watts). The array features high-efficiency power supplies and an innovative chassis design that optimizes airflow and minimizes the need for cooling, which helps to lower energy-related costs and provide further thermal protection to the array's electronic components.
- **Ultra-dense Flash array packaging.** With a 1 RU chassis, the Sun Storage F5100 Flash Array holds up to 80 Sun Flash Modules in a minimum amount of space, making it easy to integrate into new and existing deployments. Sun's innovative ultra-dense I/O design provides 24 GB/sec throughput with 64 SAS1 lanes that can deliver 3 Gb/sec via 16 4 wide SAS1 ports, allowing organizations to get the performance they need where they need it.

## Applications for the Sun Storage F5100 Flash Array

Some organizations select flash array technology for its eco-friendly characteristics—flash arrays consume considerably less power and rack space than hard disk drives, especially when compared from an IOPS perspective. Sun's flash arrays are expected to be selected due to the inherent reliability and environmental characteristics of Sun Flash Modules. For example, Sun Flash Modules do not contain moving parts or

bearings that can wear out. There are no heads to crash or rotational vibration sensitivity, helping the modules to provide greater mechanical shock tolerance and higher mean time between failure (MTBF). However, flash arrays typically are selected for their performance characteristics. When applications depend on I/O throughput, Sun Storage F5100 Flash Arrays can help storage architects to deploy ultra-dense, high-performance, eco-friendly solid state storage in a variety of scenarios and eliminate many types of I/O bottlenecks—often without the need to redesign the existing storage infrastructure.

### Performance Considerations

The performance characteristics of the Sun Storage F5100 Flash Array are well-suited to the following I/O access types.

- Access patterns and transfer sizes—Access patterns (LBAs) that are on exact 4 KB boundaries, and read and write transfer sizes that are multiples of 4 KB, can take maximum advantage of the 4,096 byte page size used by Sun Flash Modules.
- Applications—Databases and related applications that perform most I/O in multiples of 4 KB are an ideal fit. Indeed, most databases perform I/O on 8 KB boundaries.
- Block access methods—Most generic device drivers, volume managers, and file systems assume a 512 byte block size for raw storage. Tools that issue I/O requests to arbitrary 512 byte boundaries or use odd transfer sizes cannot take maximum advantage of the I/O performance advantages of the array. However, a move to larger block sizes is underway. As more systems, interfaces, and best practices optimize for the 4,096 byte block size being implemented in many of today's storage and I/O systems, the greater the performance improvement for applications.

In terms of application access patterns and latency sensitivity, the ideal fit and optimal benefit can be found with smaller datasets, particularly those with very high I/O access densities (greater than 1 IOP/GB), and latency bound applications that do not benefit from caching.

Reducing the latencies associated with datasets that fit in the flash array by an order of magnitude or more often can significantly improve system performance. Databases are a prime example. A subset of the database that is 4 KB friendly can achieve a high access density of IOPS/GB. Indeed, use cases and reference architectures have shown double the performance when moving database indexes and hot tablespaces from high-performance traditional RAID systems to the Sun Storage F5100 Flash Array. For more information on these studies, see the technical white papers and resources listed in the reference section at the end of this document.

## Software Considerations

The Sun Storage F5100 Flash Array shares a design similar to four independent SAS JBOD (Just a Bunch of Disks) devices that can each attach to one to four host initiators. As such, the array can operate out of the box with industry-standard SAS storage HBA connections and the latest performance tuned MPT drivers on the Solaris OS. However, there are several key software considerations.

- Because the Sun Storage F5100 Flash Array can deliver extreme performance, it has been qualified with special performance enhancing SAS HBA firmware. This firmware trades off some connectivity to ensure the system can demand high I/O and take advantage of the array without creating bottlenecks.
- Several software tunables can be applied to ensure software aligns I/O requests on 4 KB boundaries and optimize performance. More information on these tunables can be found in *Chapter 4, Deployment Considerations*.
- Management and monitoring of the Sun Storage F5100 Flash Array is performed in-band with the latest version of the Sun StorageTek™ Common Array Management (CAM) software.

## Chapter 2

# Sun Flash Module Design

Sun has engineered a novel storage innovation — the Sun Flash Module or “FMod” — that combines SLC NAND flash components and a flash memory controller (FMC) to provide an industry-standard SATA device in a compact, highly efficient form-factor. The Sun Storage F5100 Flash Array takes advantage of Sun Flash Modules to deliver a high-capacity and high-performance storage array in an optimal footprint.

The array is divided into four separate array domains, each containing 20 Sun Flash Modules for a maximum of 80 modules per array. Peak performance metrics for an individual module are as follows:

- Random write: 15,300 IOPS
- Random read: 20,750 IOPS
- Sequential write: 118 MB/sec
- Sequential read: 265 MB/sec

### Module components

Figure 1 shows key components on the front and rear of each module, which uses a mini-DIMM JEDEC MO-258A form factor.

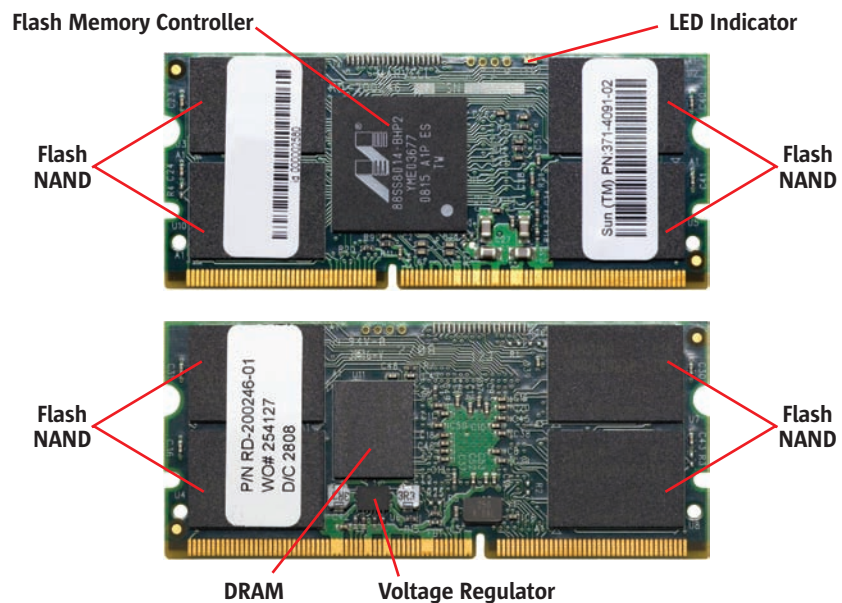


Figure 1. Front and Rear Views of a Sun Flash Module.

Module components include:

- **NAND SLC flash.** Each module contains eight 4GB SLC NAND components (four on the front side and four on the back), for a total of 32GB, of which 24GB is addressable for primary back-end storage. Excess capacity is used to optimize performance and longevity. Because of the availability of spare blocks, slower erase cycles can occur independently in the background and faulty blocks can be mapped out so they are not reused. For more details on NAND Flash, please see “Flash Basics” in Appendix A.
- **DRAM.** 64 MB of DDR-400 DRAM per module provides a local buffer cache to accelerate flash performance and maintain active data structures. In the event of a loss of power to the array, active data structures located in DRAM are written automatically to the flash devices to help ensure data integrity.
- **Flash memory controller.** Each Sun Flash Module incorporates a Marvell flash memory controller — a SATA-2 controller that allows each module to communicate using standard SATA protocols. The controller manages the NAND components and buffer cache, and provides a communication interface to systems. To extend the life of NAND devices, the controller performs “wear-leveling” and on-the-fly error correction. Wear leveling is a technique that decreases wear by minimizing writes to the same location. The controller is also responsible for tracking and mapping out faulty blocks. Faulty blocks are replaced with spare blocks that are mapped in when needed. In addition, the controller load-balances and interleaves data accesses to back-end NAND devices to accelerate I/O operations.
- **Voltage regulator.** A voltage regulator on each module down-regulates the 3.3V input voltage into 1.2V and 2.5V, which are used by the flash memory controller and DRAM.

Figure 2 depicts a logical block diagram of a single Sun Flash Module. Four 1 GB SLC NAND die are stacked together in a single package, forming each of the eight NAND flash devices. The NAND devices are enterprise-quality components, which means they have an extended lifespan rating compared to commercial grade flash components.

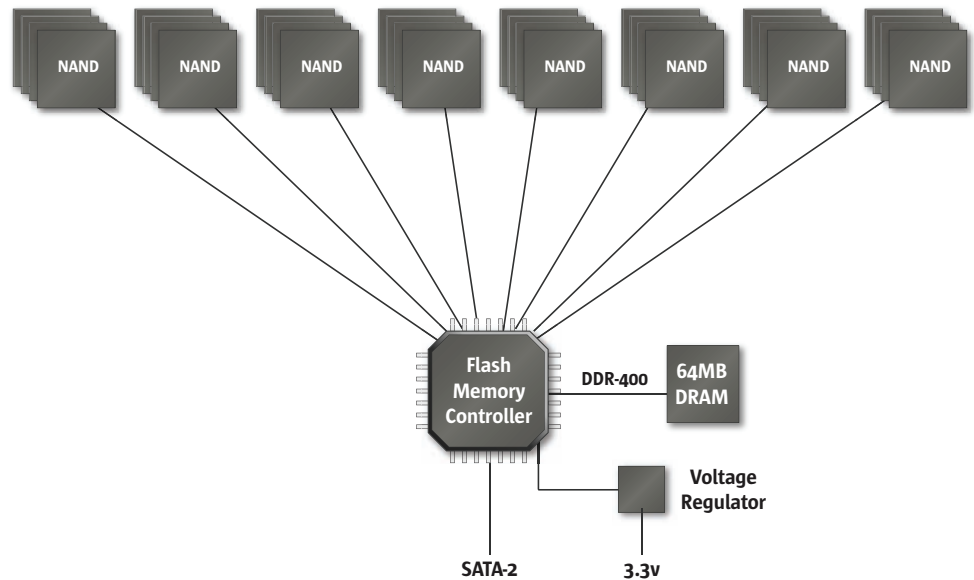


Figure 2. Logical block diagram of Sun Flash Module.

As shown in Figure 2, each NAND device communicates via a dedicated interface to the flash memory controller. The controller also interfaces to the 64MB DRAM and the voltage regulator, and provides a connection to the array's SAS expander, which communicates to the controller using standard SATA-2 commands.

### Enterprise-quality NAND flash for reliability

To develop enterprise-grade NAND SLC devices, Sun engineers worked closely with NAND manufacturers to make specific reliability enhancements. These design changes allow Sun's enterprise-quality SLC NAND devices to exhibit greater endurance. In addition, Sun performs extensive quality assurance testing and component screening to optimize NAND device reliability. SLC is usually rated for 100,000 write-erase cycles, whereas MLC Flash is usually rated between 5,000-10,000 write-erase cycles. NAND flash memory on Sun Flash Modules is certified for 2 million hours MTBF.

### Low power consumption and fast I/O

The cornerstone of the Sun Storage F5100 Flash Array's design is the Sun Flash Module, which consumes only 2.1w per module. Because of the module's low power demands, the array delivers virtually unmatched power and space efficiencies as well as low latency.

## Chapter 3

# Sun Storage F5100 Flash Array Architecture

The Sun Storage F5100 Flash Array is a dense, high-capacity, high-performance solid state storage array that can be attached via a SAS HBA to a server running a variety of operating systems, including OpenSolaris™ 2009.06, the Solaris™ 10 Operating System (Update 8), Microsoft Windows 2003 (SP2), Microsoft Windows 2008 (SP1 or SP2), Red Hat Enterprise Linux 4 (Updates 6 and 7), Red Hat Enterprise Linux 5 (Updates 2 and 3), SuSE Linux 9 (Update 4), and SuSE Linux 10 (SP1 and SP2). The array features a dense, compact design in a 1U chassis. Figure 3 shows the top cover removed and the locations of several key components.

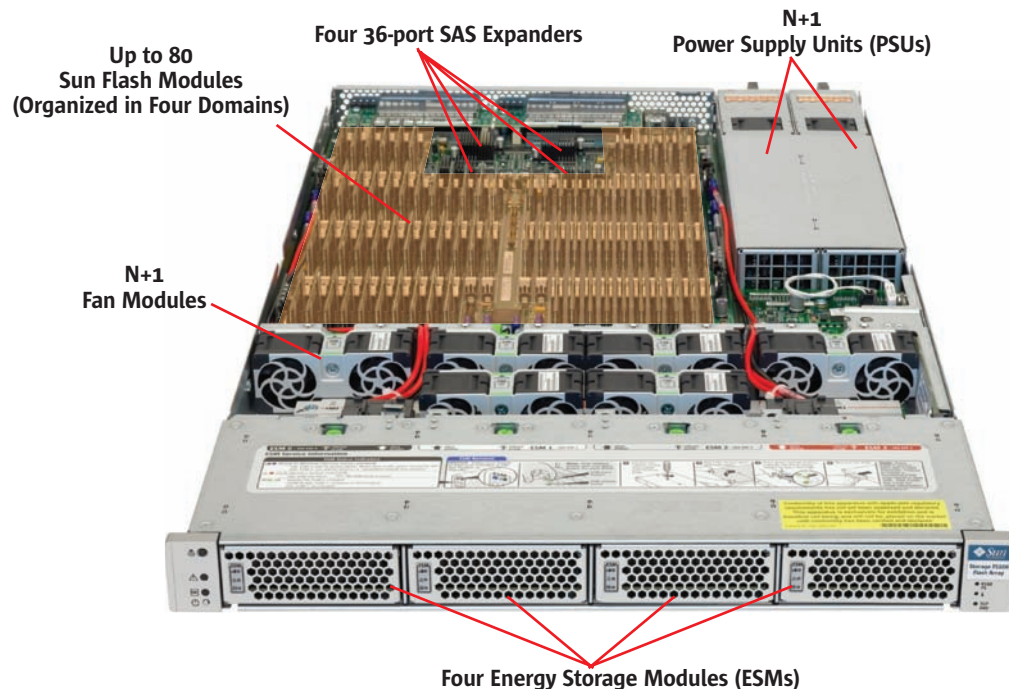


Figure 3. Major components.

The array is divided into four array domains that function as independent SAS domains. Each array domain contains:

- 20 Sun Flash Module slots. When the 20 slots in an array domain are populated, the domain provides up to 480GB of addressable storage capacity.
- Four x4 SAS ports for external host connectivity. Four ports per array domain enables tremendous performance and configuration flexibility (see Chapter 4 for example configurations).
- An LSI 36-port SAS expander. Each expander interfaces between 20 modules in the array domain and the corresponding four external host SAS connections.

- An energy storage module or ESM (one per array domain). This supercapacitor-based unit provides enough energy to flush data from module DRAM to persistent flash storage in each array domain.

## Array architecture

Figure 4 displays a block diagram for the Sun Storage F5100 Flash Array. Different colors are used to represent components in different array domains, including the ESM associated with each domain.

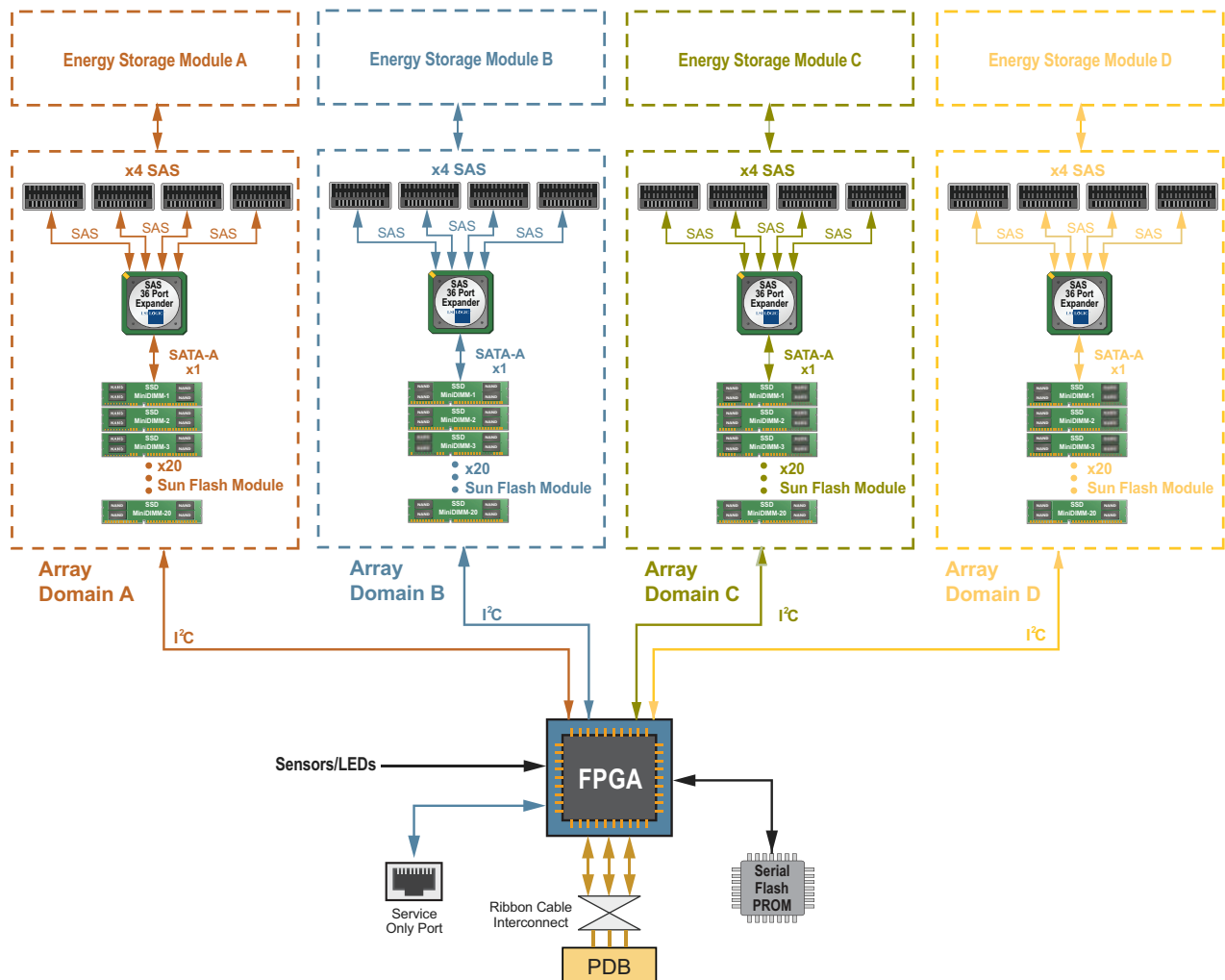


Figure 4. Sun Storage F5100 Flash Array simplified block diagram.

An FPGA (Field-Programmable Gate Array) communicates with all four domains via an inter-integrated circuit (I<sup>2</sup>C) bus and maintains array status information, including the status of Sun Flash Modules and the ESMs, as well as ambient and component temperatures. The FPGA reports status information or device failures to the Sun StorageTek Common Array Management software, which is used for in-band array management.

## Sun Flash Modules

Chapter 2 provides specifics on the design of the Sun Flash Modules. When populated, an array domain holds 20 modules. The modules are inserted into motherboard slots that feature color-coded latches corresponding to the array domain.

## LSI 36-port expanders

Four 36-port LSI SAS36 expanders reside on the array motherboard, one per array domain. The LSI SAS36 expanders support Serial ATA (SATA) standards and are compliant with ANSI-defined Serial Attached SCSI (SAS) specifications. Each expander provides 20 ports that interface to SATA links from the 20 Sun Flash Modules, along with 16 ports that connect to the four 4-lane SAS external host connectors for the array domain.

## Energy storage modules (ESM)

Controlled by the FPGA on the motherboard, the energy storage modules (ESMs) are energy storage devices used to provide backup power to the Sun Flash Modules. On each Sun Flash Module, data is cached on DRAM to enhance performance. In the event that the array loses AC power, the ESMs provide sufficient backup power to flush DRAM contents (both metadata and data) to NAND devices. In this way, the ESMs help to maintain data integrity in the event that a power loss suddenly occurs.

Designed by Sun, each ESM is a sealed unit containing 6 supercapacitors. The supercapacitor-based ESM design provides good reliability, especially compared to other technologies, such as battery-based solutions. Each ESM corresponds to a specific array domain and supplies 5 seconds of backup power to flush module DRAMs when array power is lost.

---

**Note** – Each ESM takes about 7 minutes to fully charge. During that time, flash modules in the array are accessible, but data corruption can occur if data is written to the array when the ESMs are not charged and a power loss occurs. Therefore it is recommended that administrators delay access to the array until the green OK LEDs on the ESMs stop blinking and are steady, indicating they are fully charged.

---

If an ESM experiences a fault condition, it automatically discharges so that the array can be powered down and the unit safely replaced. A Blue “OK to remove” LED on the front of the ESM module indicates that the unit is discharged and safe to remove.

## Enclosure features

The Sun Storage F5100 Flash Array enclosure occupies a single rack unit in a standard rack with rail depths of at least 28.125 inches (714 millimeters). Table 2 provides physical dimensions.

Table 2. Dimensions and weight of the Sun Storage F5100 Flash Array

Dimension	U.S.	International
Height	1.746 inches (1 RU)	44 millimeters
Width	16.75 inches	425 millimeters
Depth	28.125 inches	714 millimeters

## Front and rear perspectives

Figure 5 shows the front and rear panels of the Sun Storage F5100 Flash Array.

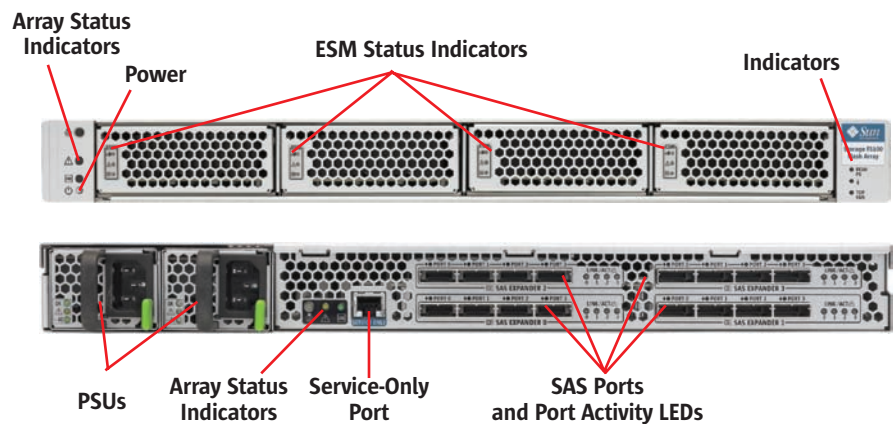


Figure 5. Front and rear panels.

External features and connections include:

- Front and rear array status indicator lights, reporting “locator” (white), “service required” (amber), and “activity status” (green)
- Front temperature, PSU, and fan status indicator lights
- Front ESM indicators, displaying ESM charging status, an ESM fault condition, and whether the ESM is fully discharged and can be safely removed
- Up to two PSUs (for N+1 redundancy) with integrated fans, each having a single, independent AC plug on the rear panel
- Rear PSU indicator lights, showing the status of each PSU
- Four banks of four x4 SAS ports (SFF 8088), with each bank connecting to a single array domain, along with LEDs representing I/O activity on each SAS connection
- A single service port (password-protected for access only by Sun Service)

## Chassis design innovations

Designed for hot-isle/cold-isle multi-racked deployments, Sun Storage F5100 Flash Arrays are optimized to conserve datacenter floor space, optimize airflow, and minimize power consumption. Features such as the honeycomb-shaped chassis ventilation holes help to provide the best compromise for strength, maximum airflow, and maximum electronic attenuation.

The Sun Storage F5100 Flash Array uses the same AC/DC 720w power supply units (PSUs) used in several Sun x64 and SPARC® servers. Configured for N+1 redundancy, the PSUs (each with a separate power cord) are highly efficient units. Variable speed fans (also leveraged from Sun server designs) operate at the lowest speeds possible to provide sufficient cooling while conserving power usage, prolonging fan life, and reducing acoustical noise. A green status light on a fan module indicates proper operation while an amber light indicates a fan fault. Fan modules are designed for redundancy — a backup fan enables continuity in the event of a single fan failure.

### Array management — Sun StorageTek CAM Software

An internal FPGA maintains array status information, including the status of Sun Flash Modules as well as ambient and component temperatures. The FPGA reports status information or device failures to the Sun StorageTek Common Array Management software, which is used for in-band array management via SCSI Enclosure Services (SES).

When deploying the Sun Storage F5100 Flash Array and other Sun storage products, the StorageTek CAM software provides administrators with a powerful, yet easy-to-use Java™ language-based graphical user interface for management. The CAM software enables online administration, a consistent interface across all operating systems, and the ability to monitor and manage one or all arrays from any location on the network. The software employs a wizard-driven, automated best practices model that takes the complexity out of configurations, saving time and reducing the chance of error. Administrative tasks such as asset discovery, configuration, re-configuration, expansion, and firmware maintenance can all be performed using the StorageTek CAM software.

Starting with version 6.4.1, the StorageTek CAM software supports the concept of zoning. Zoning allows a subset of storage devices to be allocated to specific SAS controllers within a single SAS subsystem. For example, when a single array domain in the Sun Storage F5100 Flash Array is configured using SAS zoning, different zones of flash device-based storage within a domain can be allocated to a particular HBA. This allows for up to 16 HBAs (4 HBAs per domain) to connect to a single array.

The StorageTek CAM software is available at no charge for a wide variety of host platforms, including Linux, Windows, and the Solaris Operating System. It can be downloaded from [www.sun.com/storagetek/management\\_software/resource\\_management/cam](http://www.sun.com/storagetek/management_software/resource_management/cam).

## RAS features

Corporate data and business information comprise critical business assets. Enterprise computing technologies strive to furnish a high degree of data protection (reliability), to provide virtually continuous access (availability), and to incorporate procedures and components that help to resolve problems with minimal business impact (serviceability). Commonly referred to as RAS, these capabilities are engineered into Sun's mission-critical computing and storage solutions.

The Sun Storage F5100 Flash Array is designed with these RAS features:

- *Backup power to flush DRAM in the event of a power failure.* Integrated supercapacitors in the ESMs provide 5 seconds of backup power, which is sufficient to automatically flush DRAMs on the Sun Flash Modules. This feature helps to protect data integrity — a key factor in maintaining data availability. When ESM failure is detected, FMods associated with that ESM change their behavior. Specifically, the existing contents of each FMod's DRAM buffer, if any, will be written to flash immediately and subsequent writes will not complete until the data has been stored to flash. Depending on the application, this may result in a dramatic reduction in performance. CAM management software monitors the status of the ESMs and notifies administrative and/or service personnel of any failures.
- *Redundant fan modules and power supplies.* Redundant fan modules and power supply units provide sufficient cooling and power to support continued operation, even if a PSU or fan module fails. When a downtime is scheduled, the faulty unit can be quickly and easily replaced.
- *Accessible components for improved serviceability.* When a scheduled downtime occurs, ESMs, power supply units, and fans can be replaced without completely removing the array from the rack.
- *Indicator LEDs on the front and back of the chassis enclosure.* Easily visible LEDs allow problems to be identified and isolated easily.
- **Reliability.** The Sun Flash Modules were designed and built for reliability. They are certified by the manufacturer for 2 million hours MTBF. Reliability benefits of the Sun Flash Module are enhanced by the controller which:
  - Uses wear leveling to improve the life expectancy of Sun Flash Modules by minimizing writes to the same location
  - Corrects bad data as necessary with ECC
  - Takes blocks out of service when their failure rate, detected after a failed write, becomes unacceptable
  - Moves data to a known good location (and updates corresponding mapping information)

## Flash technology considerations

Solid state storage devices based on flash technology do not function like conventional disk drives. Unlike a conventional drive, data is not stored sequentially on a flash-based device. Information that keeps track of the location of the data — the metadata — is also stored inside the flash storage device. The metadata also serves the additional purpose of tracking the number of writes to the individual storage elements.

Flash storage also needs to be managed and there are critical operations which can affect access time to the flash module:

- Since an HDD address is by cylinder, track, and sector, the data is laid down sequentially and an LBA (Logical Block Address) is easily translated to disk geometry. A flash memory-based device, on the other hand, can place a block anywhere in the NAND storage element, resulting in an additional level of tracking to manage every block within the flash-based device.
- Wear-leveling requires the movement of data and then updates to the metadata. If a request is made during wear-leveling or other housekeeping operations, the request must be delayed until the operation completes, which can impact latency.
- Maintaining tables of number of writes.
- Managing defragmentation of metadata areas.

Along with the data itself, the metadata must be protected to maintain data integrity. If there is insufficient time to write out the metadata to permanent storage, the data becomes corrupted and can not be recovered. The amount of data stored in buffered volatile storage dictates the need for independent energy storage to write out the data in the event of an unexpected power failure. Energy storage is typically implemented using batteries or supercapacitors.

Both the management tasks and the layout of the storage arrays have impacts on the performance of a solid state device versus a conventional hard drive. In addition, all flash memory has a native block size, and optimum performance is achieved when the size of the read/write data is an integer multiple of the block size and the data transferred is block-aligned. Data transfers that are not block-aligned and do not use sizes that are a multiple of the block size can impact performance, especially for write operations. Sun Flash Modules use a 4 KB block size.

## Supercapacitors

Flushing the data in the volatile buffer safely to flash storage in the event of sudden power loss requires an energy backup solution, such as batteries or supercapacitors.

Batteries have a finite and lower functional life than supercapacitors. Typical batteries have to be replaced every 2 to 3 years depending on the technology. Batteries also have issues of temperature since both hot and cold affect stored energy. In addition, batteries have higher internal resistance, so if a lot of current is needed for a short

duration, batteries cannot provide it without compromising flash storage sizes. With batteries there is also an issue of the availability of an instantaneous charge. Battery chemistry limits immediate availability of energy, whereas a capacitor can instantly supply it. Finally, there is also the problem of detecting when batteries must be replaced.

Supercapacitors have a much longer life. Although elements of a supercapacitor are similar to batteries, they do not suffer from wear out through pure discharge (non-rechargeable batteries) or charge/discharge cycles (rechargeable cells) as severely as batteries. Also supercapacitors can provide much higher short-duration current than an equivalent battery and allow for an extended temperature range usage, thus enabling a longer life expectancy. However, supercapacitors also have the following limitations:

- Some designs with supercapacitors use chemicals with shipping and disposal restrictions. The ESM in the Sun Storage F5100 Flash Array includes a label that describes the proper steps for device disposal.
- All supercapacitors have a finite life that is highly temperature-sensitive. The wear-out mechanism is a loss of capacitance. If the energy storage falls below what is needed to complete write-back operations, the data becomes at risk for corruption in the event of a power failure. The Sun StorageTek Common Array Manager (CAM) software monitors the status of the ESMs and reports failures as they happen, indicating the need for replacement.

## Chapter 4

# Deployment Considerations

A single Sun Storage F5100 Flash Array is designed to operate similar to four simple JBOD devices, and as such it offers tremendous configuration flexibility. This chapter includes information about the configuration rules for the Sun Storage F5100 Flash Array and its host bus adapters, as well as several sample configurations that highlight the balancing of functional requirements for availability and performance in database acceleration solutions.

### Database acceleration deployment considerations

The Sun Storage F5100 Flash Array brings very low latency random reads to database acceleration environments, making it ideal for index and hot table placement. Storage arrays with NVRAM are still needed to handle logging and data tables. Using these considerations as a checklist can help organizations assess the ability to successfully deploy Sun Storage F5100 Flash Arrays as database accelerators.

### Performance improvement assessment

Determining if a database deployment has an I/O bottleneck—and over how much storage—is key to assessing whether the inclusion of flash arrays can help improve system performance.

- Using I/O monitoring utilities that are part of the operating system or environment can help determine if an I/O bottleneck exists, and whether using a flash array can help the situation. In environments running the Solaris OS, the `iostat(1M)` command can help. Sun also provides downloadable utilities that can assist with this analysis, such as the Sun Flash Analyzer available at [sun.com/flash](http://sun.com/flash). By identifying database index LUNs and storage devices, it is possible to determine if service or wait times are 10 milliseconds or higher. If so, a flash array might help system performance. If these total service times are short, approximately 1 millisecond or less, then the indexes are being cached in the storage subsystem and are likely already optimized.
- Databases often have internal reporting tools that can help identify system bottlenecks. For example, administrators can use the Oracle STATSPACK or Oracle Automated Workload Repository (AWR) reporting tools in existing Oracle Standard Edition (SE) and Oracle Enterprise Edition (EE) database deployments. Go to “Top 5 Wait Events” and check to see if “db file sequential read” is one of the important wait events. The Sun Storage F5100 Flash Array reduces “db file sequential read” events to 1 to 3 ms for affected datasets, which can help dramatically accelerate database performance. If “db file sequential read” is not a widely occurring event, the Sun Storage F5100 Flash Array is not likely to help improve the performance of the deployment.

- In databases that are bottlenecked by storage I/O, a subset of the database typically exhibits the I/O bottleneck, such as deep trees in the indexes. A good rule of thumb is to place 25% of the database (indexes) on the Sun Storage F5100 Flash Array. Each array can support up to 960 GB of protected storage (1.92 TB raw). Use this ratio to determine the number of Sun Flash Modules per array, or number of arrays, needed.

### RAS requirements assessment

Sun Storage F5100 Flash Array database acceleration best practices use host-based mirroring for data protection. Deployments can be designed to avoid single points for failure, except for the compute node itself (single instance database.) Sun Storage F5100 Flash Array database accelerator best practices require some data, such as logs, to be placed on other NVRAM-based storage arrays, making the Sun Storage F5100 Flash Array an ideal add-on component for database acceleration deployments.

Potential RAS requirements include:

- If storage needs to be shared, look for alternate solutions. The SATA Sun Flash Modules cannot be shared by multiple initiators. However, a single flash array can be partitioned using SAS zoning for up to 16 different hosts.
- Some deployments use multiple paths to separate, redundant storage devices that are host mirrored. For those that require multipath I/O to database objects, look for alternate solutions. The interface on the SATA Sun Flash Modules do not support multipathing as of this writing.
- Sun Storage F5100 Flash Arrays use Solid State Sun Flash Modules that are inherently more reliable than mechanical rotating hard disk drives. Fans and power supplies are hot swappable. However, all other FRUs require a cold swap, including the Sun Flash Modules. If zero or near zero downtime for servicing is required, then a minimum of two Sun Storage F5100 Flash Arrays are needed, along with Sun Flash Module mirroring across separate Sun Storage F5100 Flash Arrays.
- For Oracle database deployments, use the Oracle Automatic Storage Management (ASM) software to protect database objects, such as indexes, on Sun Storage F5100 Flash Arrays. This is the preferred data protection strategy for Sun Storage F5100 Flash Arrays.
- Other host-based software mirroring, such as that provided by Solaris Volume Manager or Veritas Volume Manager (VxVM) software, can be used to mirror storage—with more complexity and risk, less built-in integration, and potentially higher costs. Documented practices are not provided for the combination of these tools and the Sun Storage F5100 Flash Array.
- Use other storage solutions if hardware data protection, such as RAID and NVRAM, is needed for all database objects.
- If a storage management simplification strategy is in use that dictates a single storage system for holding all parts of the database, deploy an alternate solution.

## System deployment assessment

In addition to performance and RAS considerations, check that the new or existing system deployment supports the addition of flash arrays for use as a database accelerator.

- At least two dual connector external eight lane PCIe SAS HBAs are needed to connect the Sun Storage F5100 Flash Array to servers in deployments. Two HBAs are qualified as of the writing of this document:
  - Sun StorageTek external SAS PCIe HBA (Sun part number SG-XPCIE8SAS-E-Z or SG-PCIE8SAS-E-Z)
  - Sun StorageTek external SAS HBA Express Module for blade servers (Sun part number SG-XPCIESAS-EB-Z or SG-PCIE8SAS-EB-Z)
- The SAS HBAs must have firmware level Phase 15 - MPTFW-01.27.03.00-IT. This limits connectivity to 20 Sun Flash Modules per HBA, but provides optimum performance results. For example, 40 or fewer Sun Flash Modules require at least two SAS HBAs. Similarly, 40 to 80 Sun Flash Modules housed in one or two Sun Storage F5100 Flash Arrays require at least four SAS HBAs, and 80 to 160 Sun Flash Modules in two Sun Storage F5100 Flash Arrays require at least eight SAS HBAs.
- To create a database accelerator solution with a partially populated (40 Sun Flash Modules) array and no single points of failure, four slots are required. Two slots are needed for Fibre Channel HBAs, and two slots are needed for SAS HBAs. If four slots are not available, an alternative solution is required.
- To create a fully mirrored solution using two Sun Storage F5100 Flash Arrays, ten 8-lane PCIe slots are required for the storage alone. For most servers, an I/O expander chassis is likely required. If sufficient PCIe slots are not available on the database server, or if the HBAs are not qualified for the planned database server, look at alternate solutions.
- Use the Solaris 10 OS Update 8 to take advantage of a number of device driver enhancements. Patches are available for earlier Solaris 10 OS updates.
- Each Sun Storage F5100 Flash Array requires only 1 RU of rack space for each Flash array in the deployment. Each array has a nameplate power of 720 watts per power supply to establish its power distribution design. Use the power calculators at [sun.com/powercalculators](http://sun.com/powercalculators) to estimate actual power and cooling consumption for a given deployment. Power consumption for the first five to 10 minutes after power up includes the energy needed to recharge the energy storage modules over and above the steady state chassis power consumption.
- The SAS interconnect cables are limited to three meters. Verify that the rack, rack wiring, and inter-rack wiring if needed, can support interconnects between the server with SAS HBAs and the flash array(s.)

For maximum SAS PCIe HBA performance, a new version of the HBA firmware (firmware level Phase 15 - MPTFW-01.27.03.00-IT) must be downloaded. This version increases maximum performance from ~50K IOPS to over 100K IOPS per HBA. To download this firmware, please go to <http://www.lsi.com/support/sun>.

Suggested alternate storage solutions include the Sun StorEdge 9900 series, Sun Storage 7000 Unified Storage Systems, and Sun StorageTek 2000 and 6000 array series.

## Configuration examples

This section presents three database acceleration configuration examples for the Sun Storage F5100 Flash Array, and discusses how each configuration meets different implementation goals for performance, capacity, and availability.

- *Availability considerations.* To meet data availability targets defined in Service Level Agreements (SLAs), array data must be mirrored through the connected host — there is no mirroring in the array itself.
- *Performance goals.* Array performance can be influenced by a number of factors beyond the configuration itself, including application-related issues as well as the application-generated I/O mix, and the processing power of the host connected to the array. Configuration decisions that impact performance include the number and type of hosts and the number of independent HBA channels connected to the array. Key factors in maximizing throughput are whether applications produce threaded I/O requests, host compute capabilities, and the number of host HBA cards connected to the array.

The examples below discuss how each given configuration meets design goals for capacity, availability, and performance.

### Simple availability configurations

Figure 6 shows the simplest recommended configuration with two HBA cards installed in the database server. Since each HBA features two SAS channels, this configuration supports connections to all four array domains, and up to 960 GB of raw storage (40 Sun Flash Modules). Host mirroring across HBAs and domains allows data in two array domains to be mirrored to two other domains, reducing the usable storage capacity to 480 GB.

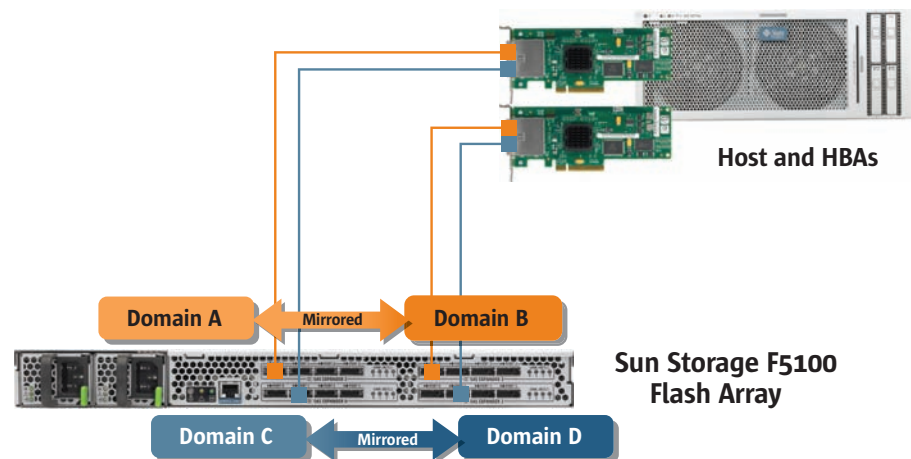


Figure 6. Single host with two HBAs connected to four array domains (two domains per HBA).

For maximum capacity with a single Sun Storage F5100 Flash Array, four HBA cards must be used in the database server, as shown in Figure 7, to connect to all four array domains and to access 1.92 TB of storage. As before, host mirroring across SAS domains can be used to increase data availability, which reduces usable capacity to 960 GB.

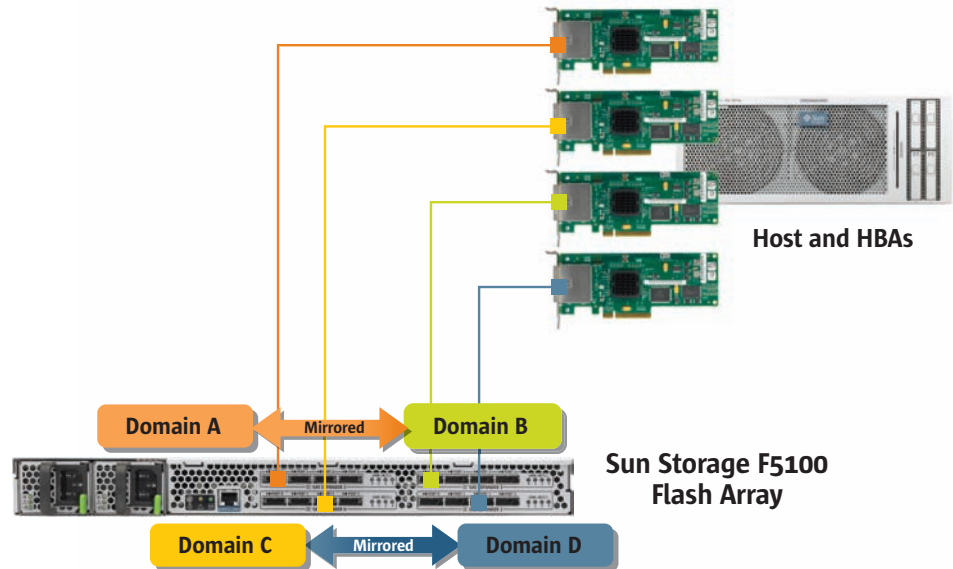


Figure 7. Single host with four HBAs connected to four array domains (one domain per HBA).

Figure 8 depicts a high-performance, high-capacity configuration that reduces downtime when replacing FRUs in one of the mirrored chassis. In this configuration, eight SAS HBA cards and their associated 8 lane PCIe server slots are needed to connect to two fully populated Sun Storage F5100 Flash Arrays and provide 3.84 TB of raw storage. Each domain is mirrored to its counterpart domain in the mirror chassis, reducing usable capacity to 1.92 TB.

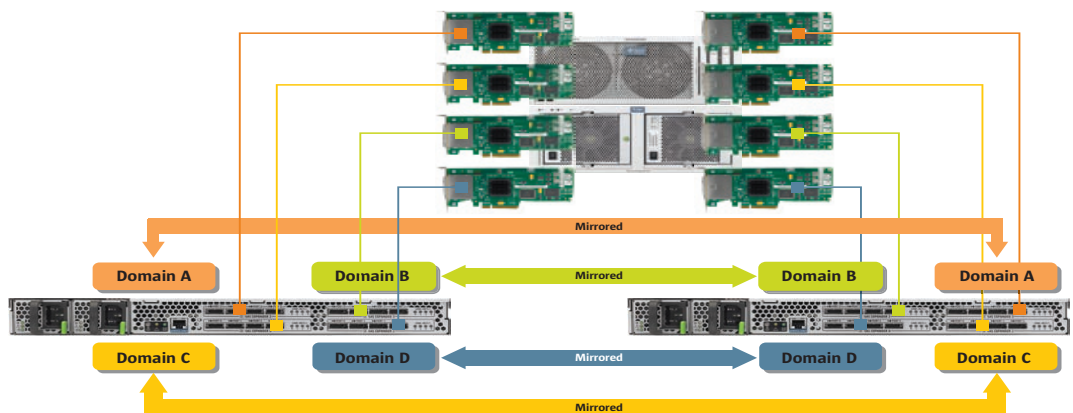


Figure 8. Single host connected to four SAS zones.

## Performance tuning

When first brought online Sun Storage F5100 Flash Arrays are seen as collections of disk drives. Each Sun Flash Module appears as a SAS attached SATA-based solid state drive. Several tuning best practices can help ensure the flash array speeds indexes and delivers maximum performance.

- Be sure the SAS HBA driver is up to date before bringing the storage online.
- Ensure that the firmware in each HBA is qualified for the Sun Storage F5100 Flash Array. The latest firmware can be found at <http://www.lsi.com/support/sun>.
- Adjust block size and alignment parameters that can impact performance. Since Sun Flash Modules use a 4 KB block size, optimal performance occurs using a 4 KB (or a multiple of 4 KB) block size. All partitions on Sun Flash Modules are aligned to start on 4 KB aligned boundaries. This is achieved by using the Solaris OS `format(1M)` command on SPARC® platform, or the `fdisk(1M)` command on x64 platforms, as described in the “Sun Storage F5100 Flash Array Product Notes”. No tuning is required for SPARC platforms. Simply ensure that standard slices are used to format volumes. On x64 systems running the Solaris OS, block 0 contains the partition table. As a result, formatting starts at block 1, which is not 4 KB aligned.
- Sun Services can help configure and design effective storage solutions using the Sun Storage F5100 Flash Array. Experienced Sun consultants can help to analyze and optimize performance and fine-tune configurations for demanding applications.
- This document recommends using software mirroring to protect raw flash array drives. For Oracle databases, Oracle Automatic Storage Management (ASM) can be used to mirror the systems, yet let the Oracle database manage the raw storage. It is recommended that the Solaris ZFS file system be used to protect key availability portions of the database, such as the archive log and flash recovery areas. It is not recommended to place indexes and hot tablespaces in Solaris ZFS file systems.

## Index mirroring, migration, operation, and validation

While a full discussion is beyond the scope of this paper, it is important to describe the final high-level steps for implementing the database acceleration practices defined in this document. Exact determinations and recommendations usually depend more on business needs than the technology. Typically, these steps are handled by the database manager. However, system administrators or storage administrators can be responsible for some of these steps, particularly setting up mirroring and data protection schemes for the array volumes.

- Establish data protection practices—The use of software mirroring is recommended. Third-party volume management tools, such as Symantec's Volume Manager (VxVM), or operating system utilities such as the Solaris Volume Manager (SV), can be used. For Sun Storage F5100 Flash Arrays, it is recommended

that the database provide any mirroring needed. For example, using Oracle ASM with Oracle databases simplifies management and delivers optimal performance due to the level of integration inherent with database-managed storage. See *Accelerating Databases with the Sun Storage F5100 Flash Array* for practical examples of balancing the performance and capacity of cost-constrained OLTP and VLDB applications.

- Migrate indexes to the flash array—There are many ways to establish the indexes on the flash array. Working with the various administrative groups that support the database deployment can help determine the best course of action.
  - Add mirrors to existing indexes and resilver to the flash array. Break the mirrors to the old index LUNs. In some environments, these tasks can be completed while the database is online.
  - Perform tablespace migration. Some databases support online tablespace migration (including indexes).
  - Rebuild indexes. Some databases support online index rebuilding to new locations, such as mirrored flash arrays.
  - Use offline techniques if business needs dictate. For example, take the database offline, and backup and restore the indexes to the mirrored flash arrays.
- Operation and validation—Once the database is online, it is important to perform an initial assessment of the service or wait times on the indexes. Service times are expected in the range of 1 to 3 milliseconds, down from the typical 10 to 15 milliseconds. If the index service times remain high, review the installation for discrepancies. If index latencies are reduced, a major database I/O bottleneck has been removed. Large improvements in total transaction throughput (~2x) and reduced total transaction times are anticipated if the system was principally experiencing I/O read latency bottlenecks. Some systems have other bottlenecks, and easing the I/O bottleneck can help make these issues easier to see, identify, and resolve.

## Chapter 5 Summary

Businesses face a substantial number of challenges in responding to storage performance requirements for data-hungry applications. Although some high-performance storage solutions exist today, economic realities make most of these solutions unattainable. Since many high-performance storage solutions incorporate proprietary controllers and devices, they usually impose high operational costs, burdening the datacenter with high demands for power and cooling, footprint, and administrative complexity.

Sun is well-known for its innovative engineering and its ability to deliver value in storage and compute solutions. With the Sun Storage F5100 Flash Array, Sun delivers breakthrough database storage performance at new economic levels. Based on enterprise-grade flash modules designed by Sun, the array raises the bar for I/O performance, as well as performance per watt and performance versus cost. Designed to behave similarly to a JBOD device, the array is a simple building block that integrates easily into existing architectures when higher performance and database acceleration is needed, without the need to rearchitect the existing storage infrastructure.

### For more information

Once the array is announced, the following resources will be available on *sun.com*.

Description	URL
Sun Storage F5100 Flash Array	<a href="http://sun.com/storage">sun.com/storage</a>
Sun Storage F5100 Flash Array Performance Information	<a href="http://blogs.sun.com/BestPerf/entry/1_4_million_iops_in">blogs.sun.com/BestPerf/entry/1_4_million_iops_in</a>
Flash Performance Tuning	<a href="http://wikis.sun.com/display/Performance/Home#Home-Flash">wikis.sun.com/display/Performance/Home#Home-Flash</a>
Sun Open Storage Initiative	<a href="http://sun.com/openstorage">sun.com/openstorage</a>
Sun Power Calculators	<a href="http://sun.com/powercalculators">sun.com/powercalculators</a>
Sun StorageTek Common Array Management Software	<a href="http://sun.com/storagetek/management_software/resource_management/cam">sun.com/storagetek/management_software/resource_management/cam</a>
Solaris ZFS	<a href="http://sun.com/solaris/zfs.jsp">sun.com/solaris/zfs.jsp</a>
Sun Services	<a href="http://sun.com/services">sun.com/services</a>

Other references include:

- Leventhal, Adam. “Flash Storage Memory”, Communications of the ACM, July 2008.
- Wright, Jeffrey. *Accelerating Databases with the Sun Storage F5100 Flash Array*, SunWIN #567882.

## Appendix A

### Flash Basics

Nearly everyone is familiar with some sort of commercially available flash device, from memory cards used in MP3 players, cell phones, and digital cameras to store music, photographs, and other digital information, to removable USB drives used to backup and transport data from one machine to another. Technological advancements are moving NAND flash technology past simple commodity use and making it a reasonable storage alternative for the enterprise. Robust data integrity, reliability, availability and serviceability features, combined with breakthrough performance and power characteristics, have made it possible to create a new class of solid state storage device.

Originally developed by Toshiba in the 1980s, flash memory is low-cost, non-volatile computer memory that can be electrically erased and reprogrammed. Data accesses to flash memory complete in microseconds, in comparison to milliseconds for rotational media and nanoseconds for DRAM memory.

Two different types of flash memory gates — NOR or NAND — provide the basis for flash devices and dictate how erase and read operations are performed. With dedicated address and data lines and a fully memory-mapped random access interface, NOR flash supports random access to any location. Combined with long erase and write times and large voltage requirements, NOR flash is well-suited to code that needs to be updated infrequently.

In contrast, NAND flash provides block access to data. With a smaller chip area per cell, NAND flash supports greater storage densities, provides greater endurance due to smaller current requirements, and costs less per unit of storage. Hence NAND flash is the more commonly implemented variety. Figure A-1 shows the basic design of a NAND gate.

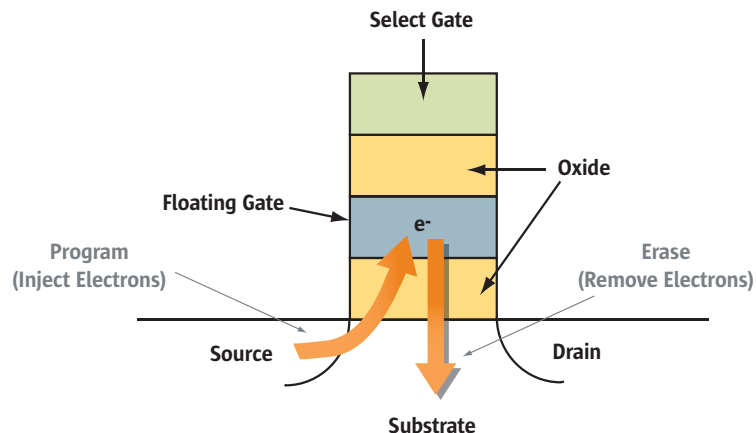


Figure A-1. NAND Gate.

There are two types of NAND flash devices: single-level cell (SLC) and multilevel cell (MLC). SLC stores a single binary value in each memory cell, while MLC supports four or eight values per cell and may contain two or three storage bits (Figure A-2). Because of its longer lifespan and better performance, SLC NAND flash devices prevail in enterprise applications.

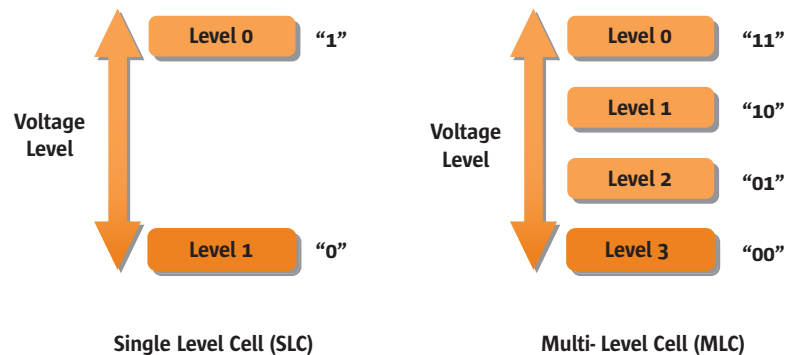


Figure A-2. SLC and MLC NAND.

Storage devices perform two basic types of I/O operations: reads and writes. Flash technology greatly outperforms rotational media for read operations because there is no mechanical process involved — read times are not delayed by “seek time” as in the case of rotational media. As such, latencies are generally reported in tenths of milliseconds. Flash devices also do not need spin-up time on startup, which can potentially speed up boot time.

Manufacturers rate the lifetime of a flash memory device in write/erase cycles because every write operation must be preceded by an erase operation, which impacts the device's lifespan. A read access on a flash device does not result in a write/erase operation, which impacts the longevity of flash memory. NAND flash memory on Sun Flash Modules is designed with 2 million hours MTBF.

For more information on characteristics of flash memory devices, see the article “Flash Storage Memory” in the July 2008 issue of Communications of the ACM.





**Sun Microsystems, Inc.** 4150 Network Circle, Santa Clara, CA 95054 USA **Phone** 1-650-960-1300 or 1-800-555-9SUN (9786) **Web** [sun.com](http://sun.com)



© 2009 Sun Microsystems, Inc. All rights reserved. © 2009 Sun Microsystems, Inc. Sun, Sun Microsystems, the Sun logo, Java, OpenSolaris, Solaris, StorageTek, and ZFS are trademarks or registered trademarks of Sun Microsystems, Inc. or its subsidiaries in the U.S. and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. and other countries. Products bearing SPARC trademarks are based upon architecture developed by Sun Microsystems, Inc. Information subject to change without notice. Printed in USA SunWIN #567883 09/09