An Oracle White Paper
June 2010

Speeding Data Access with the Sun Flash Accelerator F20 PCIe Card

ORACLE

Introduction

With today's data-intensive applications, fast data access often translates into quick response times that yield higher productivity, faster time to market, positive user experiences, and greater revenues. For Web-based applications, high-performance computing applications, and data-driven applications—including online transaction processing applications, decision-support systems, and data warehousing applications—rapid data access is vital to achieving acceptable performance and maintaining mission-critical business processes.

In recent years, as more-powerful servers have become readily available, applications are tasked with processing increasingly large data volumes and solving more-complex analytical problems. At the same time, more than ever before, companies are deploying data-intensive application services to larger user groups. All of these factors can adversely impact performance, causing slow response times that hinder strategic business initiatives and frustrate end users.

CPU performance has generally kept pace with application demands over the years; however, there have been smaller gains in storage performance. Although rotational speed and capacity have improved with time, hard disk drive (HDD) technology has not significantly changed. To keep performance-hungry systems readily supplied with data, a new generation of storage solutions—incorporating enterprise-quality flash technology—is currently emerging. These new storage solutions help to eliminate common I/O bottlenecks and feed data more rapidly to today's multithreaded, multicore CPUs.

Oracle's Sun Flash Accelerator F20 PCIe Card—an innovative, low-profile PCI Express (PCIe) card that supports onboard, enterprise-quality, solid state–based storage—accelerates I/O performance. The Sun Flash Accelerator F20 PCIe Card delivers a tremendous performance boost to applications using flash storage technology—up to 100 K I/O operations per second (IOPS) for random 4 K reads, compared to about 330 IOPS for traditional disk drives—in a compact PCIe form factor. Thus, a single Sun Flash Accelerator F20 PCIe Card delivers about

the same number of IOPS as three hundred 15 K RPM disk drives, at the same time consuming a fraction of the power and space that those disk drives would require. Even in comparison to a solid-state drive (SSD), the card offers significantly greater capacity and better I/O performance.

Adding one or more cards to an Oracle rack mounted server turns virtually any Sun x86 or UltraSPARC processor–based system into a high-performance storage server in a modular fashion. To optimize application performance, solution architects can strategically deploy the card as a fast storage tier or data cache between the host processor and higher-latency disk storage. In addition to providing low-latency storage, the card can also function as a host bus adapter (HBA), enabling it to support a server's internal drive backplane and up to eight serial-attached SCSI (SAS) or serial ATA (SATA) devices.

This white paper describes the Sun Flash Accelerator F20 PCIe Card in detail, explaining its architectural features and highlighting typical deployment scenarios.

Accelerating Application Performance with Enterprise-Quality Sun FlashFire Technology

Based on Sun FlashFire technology, the Sun Flash Accelerator F20 PCIe Card packages 96 GB of enterprise-quality flash onto a low-profile PCIe card, making low-latency flash storage easily available to boost I/O performance while conserving energy consumption.

Evolving from early flash technology found commonly in MP3 players, cell phones, and digital cameras, enterprise-quality flash devices offer similar performance and power characteristics as their commercial counterparts, but with more-robust data integrity, reliability, availability, and serviceability features. Enterprise-quality solid-state components are now poised to change the way organizations deploy storage solutions. The Sun Flash Accelerator F20 PCIe Card uses enterprise-quality flash devices similar to Oracle's SSD offerings—with greater write performance and data protection and three times the capacity. This robust flash technology for the datacenter has undergone extensive quality assurance testing and component screening to optimize reliability, requiring a mean time between failures (MTBF) for flash components of more than 2 million hours.

The Sun Flash Accelerator F20 PCIe Card, as shown in Figure 1, combines four flash modules—known as Disk on Module, or DOM, units—each containing 24 GB of enterprise-quality SLC NAND[1] flash and 64 MB of dynamic random access memory (DRAM), for a total of 96 GB flash and 256 MB DRAM per PCIe card. Each card also incorporates a supercapacitor module that provides enough energy to flush DRAM contents to persistent flash storage in the event of a sudden power outage, which helps to enhance data integrity.

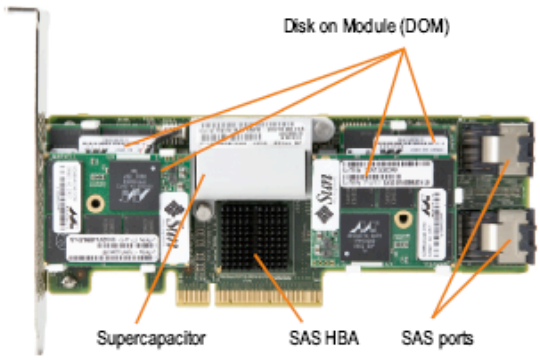

Figure 1. Sun Flash Accelerator F20 PCIe Card

[1]SLC NAND (single-level cell Not AND) flash devices store a single binary value in each memory cell.

The four DOMs are implemented via a piggyback connector to conserve space. The operating system (OS)—whether it is Oracle Solaris, Linux, or Microsoft Windows—treats the four DOMs as four separate high-performance 24 GB disk devices.

Sun Flash Accelerator F20 PCIe Card Highlights

The Sun Flash Accelerator F20 PCIe Card provides these benefits:

- **Low latency.** Flash technology can complete an I/O operation in microseconds, placing it between hard disk drives (HDDs) and DRAM in terms of latency. Because flash technology contains no moving parts, it avoids the long seek times and rotational latencies inherent with traditional HDD technology. As a result, data transfers to and from the onboard flash devices are significantly faster than what electromechanical disk drives can provide—a single Sun Flash Accelerator F20 PCIe Card can provide up to 100 K IOPS for read operations, compared to mere hundreds of IOPS for HDDs.
- **Enterprise-level reliability.** Reliability features help to increase availability and meet service-level agreement targets. The onboard flash components, which are subject to rigorous quality standards, are enterprise-quality SLC NAND devices managed by a flash memory controller. Each controller provides internal RAID, sophisticated wear leveling, error correction code (ECC), and bad block mapping (described in more detail in later sections) to provide the highest level of longevity and endurance. A supercapacitor unit flushes DRAM contents to flash storage if a power loss occurs. Even if a supercapacitor fails, the design maintains data integrity because it automatically enables write-through mode.
- **Simplified management.** The Sun Flash Accelerator F20 PCIe Card presents itself as an HBA to the server, with the four DOMs treated as four separate 24 GB disks. OS commands that manage disk drives apply equally to the DOM storage modules, so no special device drivers are required. In addition, firmware upgrades for the flash controller can be easily downloaded and applied as needed.
- **Flexible configurations.** The Sun Flash Accelerator F20 PCIe Card can be deployed in virtually any qualified Sun server that accepts a PCIe-based HBA. A variety of OSs have been qualified with the card:
  - Oracle Solaris 10 OS (Update 8)
  - Red Hat Enterprise Linux 5 (Update 3)
  - SUSE Linux 10 (SP2)
  - Windows 2003 (SP2)
  - Windows 2008 (SP1 or SP2)
- **Leading ecoresponsibility.** The solid-state DOMs operate at low power (approximately 2 watts for each 24 GB module), which is especially low in comparison to disk devices (typically around 12 watts each). The card itself consumes about 16.5 watts during normal operation.
- **Optimal storage value.** Based on the low-latency, enterprise-quality DOM storage modules, the Sun Flash Accelerator F20 PCIe Card offers low cost and power consumption given its ability to

accelerate I/O operations. Even in comparison to Sun flash SSDs, the card offers greater capacity and throughput.

The table provides a summary of performance specifications achieved with 32 outstanding I/O threads:

SUN FLASH ACCELERATOR F20 PCIE CARD PERFORMANCE

FEATURE	VALUE
Capacity per card	96 GB (4 x 24 GB)
Random 4 K read	100,110 IOPS
Maximum delivered random 4 K write	83,996 IOPS
Sequential read (1 M)	1,092 MB/sec
Maximum delivered sequential write (1 M)	501 MB/sec
Power consumption (normal running mode)	16.5 W

Although several other flash-based storage solutions exist, the Sun Flash Accelerator F20 PCIe Card provides the performance benefit of flash storage in a convenient and compact low-profile PCIe form factor. Occupying a single slot on the motherboard, the card's dense PCIe form factor is particularly beneficial for existing one rack unit (1U) servers or larger servers with a limited number of available disk slots. In addition, the PCIe form factor provides additional physical security because a PCIe card cannot be easily removed from a system, unlike a hot-pluggable disk drive. The card can function as an internal SAS/SATA HBA; therefore, it can also replace an existing HBA without the need to consume any additional slots, at the same time supplying low-latency, flash-based storage.

Sun Flash Accelerator F20 PCIe Card Architecture

The Sun Flash Accelerator F20 PCIe Card takes advantage of enterprise-quality NAND flash technology. Today, Oracle offers a range of innovative flash storage products, ranging from SSD drives for servers to the Sun Storage F5100 Flash Array. Specifically designed to accelerate database performance, the Sun Storage F5100 Flash Array houses up to 1.92 TB of flash storage in a compact 1U chassis. Building on experience in designing dense flash storage modules for these products, the design of the Sun Flash Accelerator F20 PCIe Card delivers a novel and high-performance storage architecture.

Sun Flash Accelerator F20 PCIe Card Components

As shown in Figure 2, the Sun Flash Accelerator F20 PCIe Card includes these major components, which are described in the subsequent subsections:

- Four onboard DOMs with enterprise-quality flash storage
- An eight-port 3 Gb/sec SAS host controller, which connects to the DOMs and the 12-port expander
- A 12-port SAS/SATA expander with two four-wide SAS connectors that can support up to eight internal disk drives
- An onboard supercapacitor module that helps to maintain data integrity

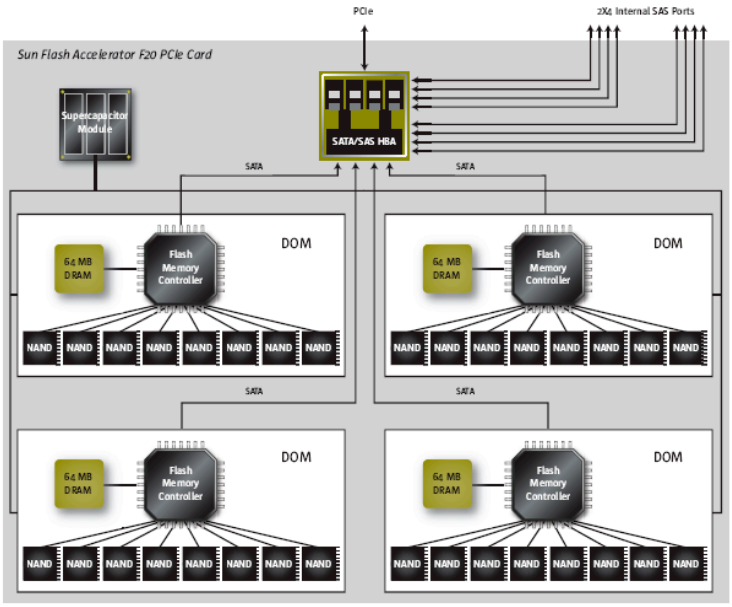

Figure 2. A logical block diagram of the Sun Flash Accelerator F20 PCIe Card

Disk on Module Design

Another innovation in storage technologies includes the design of the DOM. Each DOM combines SLC NAND flash components and a flash memory controller to provide an industry-standard SATA flash storage device in a compact, highly efficient form factor. The Sun Flash Accelerator F20 PCIe Card takes advantage of the DOM design to deliver high-performance storage in an optimal footprint.

Figure 3 shows a logical block diagram of a single DOM. Each DOM features:

- **SLC NAND flash.** Each DOM contains eight 4 GB SLC NAND components (four on the front side and four on the back), for a total of 32 GB, of which 24 GB is addressable for primary back-end storage. Excess NAND flash capacity is used internally to optimize both performance and longevity. Using available spare blocks, the controller can perform slower erase cycles independently in the background and map out faulty blocks so that they are not reused. The NAND devices are enterprise-quality components, which means they have an extended life span rating compared to commercial-grade flash components.

- **DRAM.** 64 MB of Double Data Rate 400 DRAM per DOM provides a local buffer cache to accelerate flash performance. In the event of a loss of power to the system, the content of DRAM is written automatically to the flash devices to maintain data integrity.
- **Flash memory controller.** Each DOM incorporates a Marvell flash memory controller—a SATA-2 controller that enables each DOM to communicate using standard SATA protocols. The controller manages NAND components and the DRAM buffer cache, and provides a communication interface to systems. To extend the life of NAND devices, the controller performs wear leveling and spontaneous 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, which 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.
- **Write-through functionality in firmware.** When needed, controller firmware can implement a specific mode of operation called *write-through*. This mode circumvents the caching in onboard DRAM, which accelerates flash performance during normal operation. Because write-through mode has a negative impact on I/O write performance, it is only activated when the energy backup circuitry (the supercapacitor module) has faulted. Write-through mode is also invoked during the initial power-on state until the supercapacitor becomes fully charged, after which write-through mode is automatically disabled.

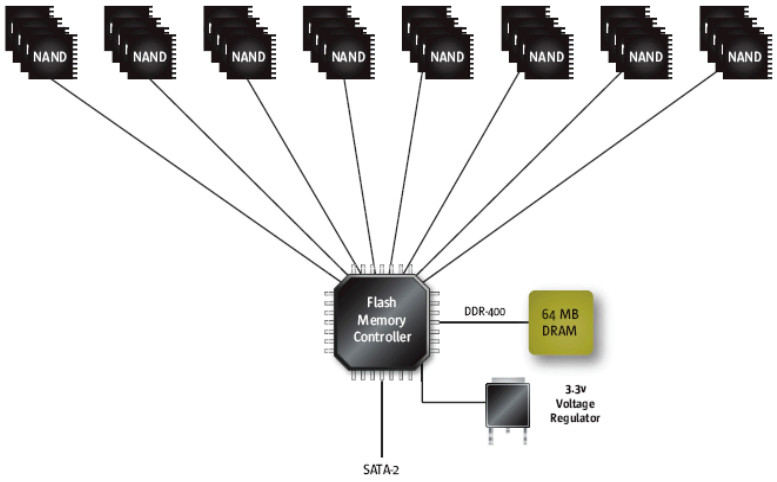

Figure 3. A logical block diagram of a single DOM

The DOM form factor is a semi-stackable design to fit required components within the defined physical envelope. Although DOMs do not interface with each other in a true stacking implementation, they overlap in the space consumed on the face of the PCIe card. The four DOMs are arranged in two pairs such that a taller connector positions one DOM over the top of another to conserve board real estate.

Enterprise-Quality NAND Flash for Reliability

These enterprise-grade SLC NAND devices offer specific reliability enhancements that exhibit greater endurance than commercially available flash components. In addition, extensive quality assurance testing and component screening further optimize NAND device reliability. SLC NAND devices are usually rated for 100,000 write-erase cycles. With advanced wear leveling, bad block mapping, and sparing, each DOM is rated for more than 2 million hours MTBF, which is greater than most disk drives.

Serial-Attached SCSI Controller

The SAS controller—an LSISAS1068E controller—is an eight-port controller that provides an eight-lane PCIe interface. With 3.0 Gb/sec SAS and SATA data transfer rates per port, the controller complies with the PCIe 1.0a specification and is compatible with SATA target devices. Four controller ports connect to the four DOMs, while the other four ports connect to the 12-port SAS expander, to support up to eight internal disk drives.

Twelve-Port Serial-Attached SCSI Expander

The Sun Flash Accelerator F20 PCIe Card contains a single 12-port LSISASx12 SAS expander. This expander supports SATA standards and is compliant with ANSI-defined SAS specifications. It provides connectivity to off-board, in-chassis storage (such as the eight-disk HDD backplane in the Sun SPARC Enterprise T5220 server) via two 3.0 Gb/sec four-lane internal SAS ports that use mini-SAS SFF 8087 connectors. For systems hosting 16 internal HDDs (such as the Sun Fire X4275 storage server), one of the four-lane SAS ports can connect to a LSISASx36 expander, which in turn, can connect to the 16 internal HDDs.

Supercapacitor Module

To increase reliability, the Sun Flash Accelerator F20 PCIe Card incorporates a capacitive energy backup system to help maintain data integrity during a power outage. The supercapacitor module resides in an easily detachable leaded cap pack, similar to a typical battery pack. The module provides sufficient power to flush data from DRAM to respective nonvolatile flash devices on each DOM.

Supercapacitors versus Batteries

If there is a sudden power outage, it is necessary to write out data in volatile DRAM to flash storage to maintain data integrity and achieve persistence. This requirement dictates the need for an energy backup solution such as batteries or supercapacitors. Batteries have a finite and lower functional life than supercapacitors, making supercapacitor technology better suited for this task.

Typically, batteries must be replaced every two to three years, depending on the type. Batteries also have issues of temperature sensitivity because both hot and cold affect the stored energy. In addition, batteries have higher internal resistance; if a large current load is needed for a short duration, batteries cannot provide it without compromising physical size and, hence, storage density. There is also an

issue of a battery's ability to deliver an instantaneous charge. Battery chemistry limits the immediate availability of energy, whereas a capacitor can instantly supply it.

Although elements of a supercapacitor are similar to batteries, supercapacitors do not suffer from wear through pure discharge (as in nonrechargeable batteries) or through charge/discharge cycles (as in rechargeable cells) as severely as batteries. Supercapacitors can also provide much-higher short duration current than an equivalent battery and permit usage in extended temperature ranges, enabling longer life expectancies. Generally, supercapacitors have a longer life expectancy compared to batteries, with estimated life spans of four to five years in a well-cooled chassis. Because high temperatures can have negative impact on life expectancy, it is best to locate the Sun Flash Accelerator F20 PCIe Card in PCIe slots that offer maximum airflow.[2]

Reliability, Availability, and Serviceability Features

Oracle's compute and storage products, including the innovative Sun Flash Accelerator F20 PCIe Card, are designed to preserve and protect mission-critical information assets. For this reason, their product architectures strive for

- **Reliability.** Furnish a high degree of data protection.
- **Availability.** Provide virtually continuous access.
- **Serviceability.** Incorporate components that help to resolve problems with minimal business impact.

Commonly referred to as RAS, the Sun Flash Accelerator F20 PCIe Card includes these RAS features:

- **Backup power to flush DRAM in the event of a power failure.** Integrated supercapacitors on the card provide sufficient power for each DOM to automatically flush DRAM, helping to maintain data consistency and availability.
- **Firmware support of write-through mode.** When a supercapacitor module fails, write-through mode is initiated, DRAM is avoided, and data integrity is no longer at risk. Depending on the application, this could result in a reduction in performance, especially if the I/O workload is write intensive, until the faulty supercapacitor module is replaced.
- **Easy serviceability.** Because of their simple modular design, DOM units and the supercapacitor module are easily replaced if a component failure occurs.

[2]Refer to the *Sun Flash Accelerator F20 PCIe Card Installation Guide* for card qualification and PCIe slot placement recommendations for particular servers.

- **Indicator LEDs.** Easily visible LEDs enable problems to be readily identified. The Sun Flash Accelerator F20 PCIe Card includes seven user-visible status indicators on the card bracket (see Figure 4):
  - **Four DOM status LEDs.** These LEDs remain off until each DOM is initialized, after which each is illuminated to a solid ON state.
  - **Energy backup system status LED.** This green LED illuminates steadily after the supercapacitor module has fully charged, indicating that it can supply power to flush DRAM to NAND storage in the event of an unplanned outage. This LED slowly blinks while the supercapacitor is charging.
  - **Energy backup system fault LED.** This amber LED illuminates to a steady ON state if there is a fault with the supercapacitor or its associated circuitry. The LED is off while the supercapacitor module is charging. (Note that both the energy backup system status and fault LEDs blink if a supercapacitor is not installed.)
  - **Card status power LED.** This LED illuminates to a steady ON state after all card components have successfully reached their preset working power levels.

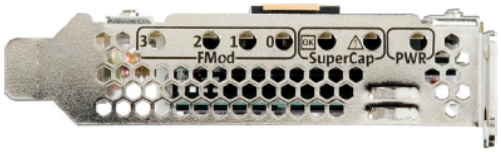

Figure 4. Status indicator LEDs

- **Reliability.** Designed and built for reliability, the DOM units are certified by the manufacturer to support 2 million hours MTBF. On each DOM, the controller enhances reliability in these ways:
  - It uses wear leveling to improve the life expectancy of the SLC NAND flash modules by minimizing writes to the same memory location.
  - It monitors and predicts media wear-out.
  - It corrects bad data as necessary with ECC.
  - It takes blocks out of service when their failure rate (detected after a failed write) becomes unacceptable.
  - It moves data to a known good location (and updates the corresponding mapping information in metadata).

  In addition, 8 GB (or 25 percent) of additional internal capacity is used to provide an extremely high level of write endurance and performance.
- **Data availability.** Mirroring technologies (such as storage pool mirroring) can duplicate data across DOMs using a volume manager. Failure of a DOM is equivalent to a catastrophic hard drive failure (although the card can still function as an HBA in this case). When mirroring across DOMs, the CPU sends writes to two or more DOMs simultaneously—it then has the ability to read from one or

more of these devices to retrieve data. In applications requiring fault tolerance, application or OS software must avoid a single point of failure and must replicate data onto additional Sun Flash Accelerator F20 PCIe Cards or other flash-based storage. Systems architects must take into account an application's need for data integrity or availability when configuring the Sun Flash Accelerator F20 PCIe Card into the storage infrastructure. Host-based data services can be used for replication, if needed.

To promote RAS, the Sun Flash Accelerator F20 PCIe Card is designed to function as a critical element in the overall storage infrastructure. The next section discusses common deployment scenarios and considerations for applying flash technology to accelerate I/O operations.

Deploying the Sun Flash Accelerator F20 PCIe Card

Flash Technology Considerations

Storage devices based on flash technology do not function like conventional disk drives. For example, unlike a conventional disk drive, data is not stored sequentially on the SLC NAND flash devices. Information that keeps track of the location of the data—the metadata—is also stored on flash storage. The metadata serves the additional purpose of tracking writes to individual storage elements.

With flash storage, there are critical operations that can affect data access time:

- **Block management.** Because access to an HDD address is by cylinder, track, and sector, a disk controller lays down data sequentially and translates the logical block address to disk geometry. A flash-based device, in contrast, places blocks anywhere in the NAND storage element, resulting in the need for tracking via the metadata to manage blocks within the device.
- **Maintaining tables of the number of writes.** In addition to mapping storage locations, the metadata tracks the number of writes to individual storage elements to perform wear leveling.
- **Wear leveling.** Wear leveling, which minimizes writes to the same memory location to extend device life, also requires data movement and metadata updates. If an I/O request is made during wear leveling or other housekeeping operations, the request must be delayed until the operation completes, which can increase latency.
- **Defragmentation of metadata.** Fragmentation of metadata can also affect performance and increase latency of I/O operations.

Along with the data itself, metadata must be protected to preserve data integrity. In the event of a power loss, if metadata is not written out to permanent storage, then access to the data is no longer possible, and it cannot be recovered. The amount of data stored in buffered volatile DRAM storage dictates the needed amount of independent energy storage to write out both data and metadata if there is an unexpected outage.

Aligning Data Transfers and Block Size

All flash memory has a native block size. Optimal performance is achieved when the size of the read/write data is an integer multiple of this block size and the data transferred is block aligned. Data transfers that are not block aligned and that do not use sizes that are a multiple of the block size can impact performance, especially for write operations. On the Sun Flash Accelerator F20 PCIe Card, the flash devices use a native 4 KB block size. Thus, the card delivers optimal performance for I/O operations when data is aligned on a 4 K boundary and when transfer sizes are a multiple of 4 KB.

There are a number of applications and environments in which the Sun Flash Accelerator F20 PCIe Card can be deployed successfully to accelerate data access. In most cases, the card is implemented as a layer of low-latency, fast storage between the host processor and higher-latency disk storage. The OS—whether it is Oracle Solaris, Windows, or Linux—accesses the DOMs of the four cards as four separate storage devices. When an administrator initializes DOM storage, configuring file systems or storage volumes using a native block size (or a multiple) is key to optimizing performance.

The next sections discuss three common deployment scenarios:

- General-purpose database acceleration
- Smart Flash Cache with Oracle Database 11g Release 2
- Hybrid storage pool technology in Oracle Solaris

Database Acceleration

Enterprises in every industry rely on fast access to business-critical information. With increasingly sophisticated applications that solve business problems, analyze data, and track customer relationships, database applications support more users and complex business processes than ever before. However, rapidly growing data volumes and compute-intensive applications are pushing databases to the limit, which can slow application response times.

Accelerating database performance is often critical to successful business initiatives. Because different database operations can stress underlying servers and storage systems, addressing both CPU and I/O bottlenecks is essential. To improve database performance, administrators today can take advantage of database partitioning, SQL tuning, query optimizations, and clever caching techniques. These strategies, along with powerful servers with large memories and multithreaded processor cores, are helping to alleviate CPU bottlenecks and speed processing throughput. Although striping I/O operations across multiple HDDs with faster interfaces can also help with the transactional and I/O demands of database environments, further performance improvement is now possible through flash-based storage. The Sun Flash Accelerator F20 PCIe Card provides an ideal solution to accelerate database application performance.

Database Deployment Considerations

The Sun Flash Accelerator F20 PCIe Card brings extremely low-latency random reads to database environments, which makes the card's DOM storage ideal for index and hot table placement. Optimizations such as mirrored disk drives or nonvolatile random access memory are still needed to

handle logging and data tables. Taking these considerations into account, organizations can better determine whether the card can help to accelerate database applications.

Assessing Opportunities for Performance Improvement

Because of the complexity in applications, system, and networking designs, it can sometimes be difficult to identify the precise cause of a performance issue. Determining whether an I/O bottleneck is limiting database response time is key to assessing whether the Sun Flash Accelerator F20 PCIe Card can help to improve application performance. Some methods to help determine whether the card can help to improve database performance include the following:

- Using I/O monitoring utilities that are part of the OS or management environment can help to determine if an I/O bottleneck truly exists, and whether the use of low-latency, flash-based storage can alleviate the bottleneck. Systems that run Oracle Solaris, for example, can use the `iostat(1M)` command. In addition, Oracle provides a utility that can assist with performance analysis—the Sun Flash Analyzer.[3] The Sun Flash Analyzer is an open systems, storage-centric Java application that thoroughly captures, summarizes, and analyzes storage workloads for Oracle Solaris, Windows, or Linux environments. Based on defined criteria, the utility enables quick identification of storage devices where I/O latency issues exist. By identifying database index logical unit numbers and storage devices, it is possible to measure whether I/O service or wait times are 10 milliseconds or higher. If so, it is likely that a Sun Flash Accelerator F20 PCIe Card can help to improve database performance. However, if total I/O service times are short (for example, approximately 1 millisecond), then it is likely that database indices are already optimized and cached.
- Databases often have internal reporting tools that can help identify system bottlenecks. For example, administrators can use the Oracle STATSPACK utility or its successor, Oracle Automatic Workload Repository, in existing Oracle Database Standard Edition and Oracle Database Enterprise Edition 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 Flash Accelerator F20 PCIe Card can often reduce latency in "db file sequential read" events, which can help to accelerate database performance. If "db file sequential read" is not a widely occurring event, the card is not likely to help improve database application performance.

Database Smart Flash Cache

Oracle Database 11g Release 2 introduces the Smart Flash Cache feature, which automatically places frequently accessed data in very fast flash storage, while most of the data is kept in cost effective disk storage. The software intelligently determines how and when to use the flash storage, and how best to incorporate flash into the database as part of a coordinated data caching strategy, all without the

[3]The Sun Flash Analyzer is available for download at no charge at sun.com/flash.

intervention of a database administrator. In addition, Oracle allows the user to provide directives at the database table, index and segment level to ensure that specific application data is kept in flash. Tables can be moved in and out of flash with a simple command, without the need to move the table to different tablespaces, files or LUNs. The Smart Flash Cache enabled on the Sun Flash Accelerator F20 PCIe Card, is a key component of Oracle Exadata V2 Storage Server, which delivers extreme performance and scalability for all database applications including OLTP, Data Warehousing, and consolidation of mixed database workloads.

Hybrid Storage Pool with Oracle Solaris ZFS

When designing storage solutions, system architects must often make trade-offs between functional requirements and cost. For this reason, datacenters frequently use a variety of devices to store, archive, and access information. One approach—creating a hybrid storage pool (HSP)—leverages the strengths of both rotational and flash-based storage technologies. Implemented via Oracle Solaris ZFS, an HSP takes the approach of placing data on the most appropriate storage media to optimize performance and balance costs.

On servers running Oracle Solaris, the Sun Flash Accelerator F20 PCIe Card is ideal for deploying HSPs on the ZFS file system. To minimize the impact of disk latencies and improve application performance, a storage administrator can create a new storage tier that holds frequently accessed data on the Sun Flash Accelerator F20 PCIe Card.

Reducing Read Latency

Systems often use memory to cache frequently accessed data for rapid access and improved performance. Once data is cached, future requests can be satisfied quickly by accessing the cached copy in system memory, rather than fetching it from disk. Policies determine which data is held in cached memory in an attempt to anticipate future needs. However, larger working sets that do not fit entirely into system memory cannot be effectively cached and must be retrieved from storage.

In many cases, flash storage can be used to implement an effective caching strategy. In ZFS, the adaptive replacement cache (ARC) automatically manages and balances cache content using the most frequently used and most recently used algorithms for storage and retrieval. ZFS also supports a second-level ARC (L2ARC) with smart caching and prefetching techniques. Applying flash technology in the Sun Flash Accelerator F20 PCIe Card to support the L2ARC cache can greatly boost read performance (see Figure 5).

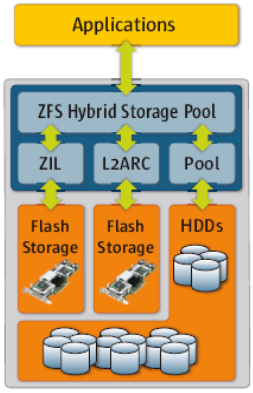

Figure 5. HSP striving to balance optimal performance against cost.

Implementing the L2ARC on flash storage can help to improve read performance by

- **Eliminating disk latency.** Both the ARC and L2ARC are used to satisfy read requests from clients. Using flash storage, read operations can be serviced in microseconds rather than milliseconds, which is typical for disk drive reads.
- **Speeding access to working sets.** Flash devices offer a faster way to access working sets that does not fit into available system memory.
- **Enhancing cache performance.** The L2ARC uses an evict-ahead policy. Cache entries are aggregated and predictively pushed out to flash devices to distribute overhead across large write operations and eliminate additional latency that can occur when an entry is evicted from cache.
- **Speeding system readiness by warming caches.** The L2ARC stores a directory of data blocks written to it. This practice helps to identify cache contents after a power or system failure, enabling the system to warm the cache proactively.

Reducing Write Latency

ZFS uses a log to record modifications to the file system. Shown also in Figure 5, the ZFS Intent Log (ZIL) enables applications that demand synchronous writes to a permanent storage medium to benefit from latency reductions and get work done while data is written asynchronously in the background. It can store small transactions to the file system in a dedicated flash storage pool before committing the transaction to disk. The ZIL stores enough information to replay the transaction, if needed. These records are freed once the data is committed to disk. The ZIL handles small and large writes differently:

- Small writes are included in the log record.
- Large writes are synchronized to disk, and the ZIL maintains a pointer to the synchronized data in the log record. As a result, the size of the ZIL tends to be small and is dictated by the number of IOPS from clients.

Several techniques are used to speed write throughput.

- ZFS manages the storage pool by aggregating high-bandwidth devices and low-latency devices separately. It dynamically determines whether a low-latency or high-bandwidth device should be used, depending on the amount of accumulated data in a transaction.
- Writes are acknowledged once the data is written to the ZIL. Multiple small transactions are aggregated, letting the system perform fewer commits to the hard disk drives in the storage pool and use fewer and larger I/O transactions to speed throughput.
- Placing the ZIL on a low-latency flash storage can help to improve server throughput by completing the write in microseconds instead of milliseconds as with writes to HDDs.

The DOMs on the Sun Flash Accelerator F20 PCIe Card can be separately dedicated to L2ARC and ZIL functions. The exact allocation of DOMs depends on the application and where the highest latency exists and needs to be addressed. For example, two DOMs can be dedicated for ZIL, and the remaining two DOMs can be used for the L2ARC. If more L2ARC is needed, another Sun Flash Accelerator F20 PCIe Card can be plugged into an available PCIe slot. It is advisable to use multiple DOMs for the L2ARC because it increases capacity and improves performance since concurrent I/O operations can occur simultaneously using multiple flash devices. To comply with 4 KB block-aligned data transfers (see "Flash Technology Considerations"), the storage administrator can use the `zfs(1M)` command "`set recordsize`" when creating the storage pool.

Conclusion

Today's complex business applications typically house massive data volumes and serve large numbers of users—a trend that drives performance requirements that are increasingly difficult to attain. To achieve fast response times for data-intensive applications, systems must be able to access data rapidly and transfer it quickly from storage to compute resources for processing. Many data-driven applications suffer from long latencies and slow response times due to I/O bottlenecks that limit throughput between storage and servers.

As flash technology moves into the enterprise, it holds promise for accelerating application performance, reducing bottlenecks, and helping lower datacenter energy consumption. In an innovative PCIe form factor, the Sun Flash Accelerator F20 PCIe Card incorporates low-latency, enterprise-quality flash technology with extremely low power consumption given the relative increase in IOPS. With a layer of flash-based storage between traditional disk media and host processors, today's powerful CPUs can experience less idle time waiting for I/O operations to complete, contributing to major improvements in application performance. For a variety of data-intensive workloads, Oracle's Sun Flash Accelerator F20 PCIe Card can dramatically eliminate I/O bottlenecks—helping organizations to enhance user productivity and expedite strategic applications.

Speeding Data Access with the Sun Flash Accelerator F20 PCIe Card
June 2010

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

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

SOFTWARE. HARDWARE. COMPLETE.