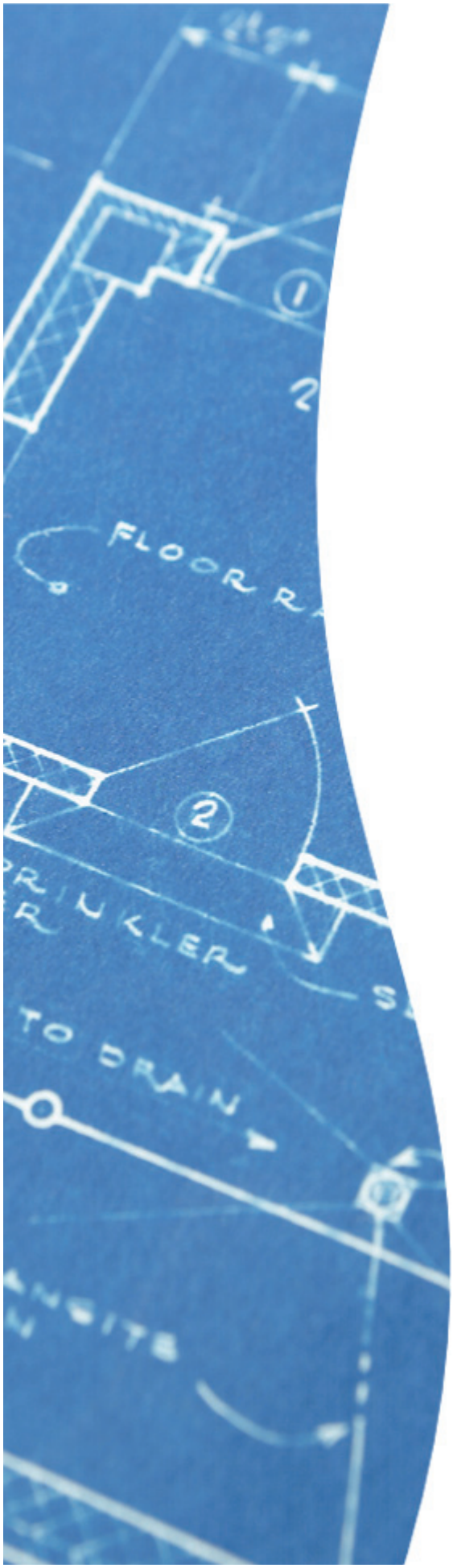

DEPLOYING HYBRID STORAGE POOLS

With Sun Flash Technology and the Solaris ZFS™ File System

Roger Bitar, Sun Microsystems

Sun BluePrints™ Online

Part No 820-5881-10
Revision 1.0, 10/31/08

Table of Contents

Storage Infrastructure Challenges

Businesses increasingly rely on datacenters to provide access to services, applications, and data. As demand rises and applications gain complexity, datacenter infrastructure must provide massive capacity and fast access to information in order to keep pace with business priorities. Today companies can add storage capacity easily by augmenting infrastructure with additional hard disk drives and arrays. Unfortunately, the devices that provide the highest capacity fail to provide the performance needed to keep systems supplied with the data for processing.

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. Consider that a quad-core server with a maximum memory configuration can generate hundreds of thousands of I/O operations per second (IOPS) — yet the entire complex of disk drives available to the system can only perform thousands of IOPS combined. The disparity in IOPS between the system and disks caused by rotational and seek latencies forces CPU cycles to be lost waiting for I/O to complete, impacting system throughput and application performance (Figure 1).

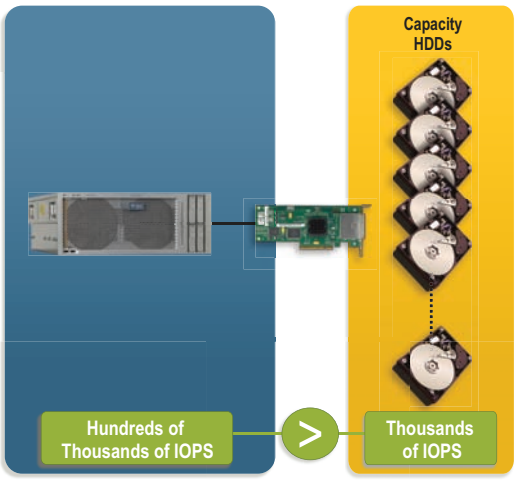

Figure 1. Hard disk drives leave systems waiting for data

To compensate, IT managers typically add more external devices and DRAM to help speed throughput. More DRAM lets systems store larger working sets in memory to avoid disk latency, and adding disk spindles can help increase throughput by letting I/O operations be performed in parallel. However, the result is an expensive infrastructure that is difficult to manage.

This Sun BluePrints™ article provides an overview of Flash technology, and discusses its introduction into a new tier of storage infrastructure. It explains how companies can utilize Flash technology and the Solaris ZFS™ file system to take advantage of the high performance of enterprise solid state drives (SSDs) and the low cost of high-capacity hard disks to create hybrid storage pool solutions that help balance system performance and cost.

Flash Technology Moves to the Enterprise

Originally developed by Toshiba in the 1980s, Flash memory is low-cost, non-volatile computer memory that can be electrically erased and reprogrammed. Two different types of memory — NOR and 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.

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 storage device.

Enterprise Solid State Devices

Enterprise solid state devices based on Flash technology consist of three main components: NAND Flash, DRAM, and a controller (Figure 2).

- *NAND Flash*
  NAND Flash is used for primary back-end storage, and requires blocks to be erased prior to writing data. While NAND Flash provides fast read access times, it takes 1.5 milliseconds to erase a block. Maintaining a range of spare blocks that are available for use helps to alleviate erase time bottlenecks.
- *DRAM*
  DRAM provides a local buffer to accelerate Flash write performance and maintain active data structures. A supercapacitor is used to protect data and permit it to be flushed to the media in the event of power loss.
- *Controller*
  A controller manages the back-end storage and buffer cache, and provides a communication interface to systems. To extend the life of the device, the controller works to minimize writes to the same location to decrease wear, and tracks and maps bad blocks so they cannot be reused. While mapping bad blocks out of the available address space can impact performance over time, doing so

results in gradual device failure rather than sudden failure. In addition, information is load balanced and interleaved to speed performance, and ECC is supported to provide data integrity.

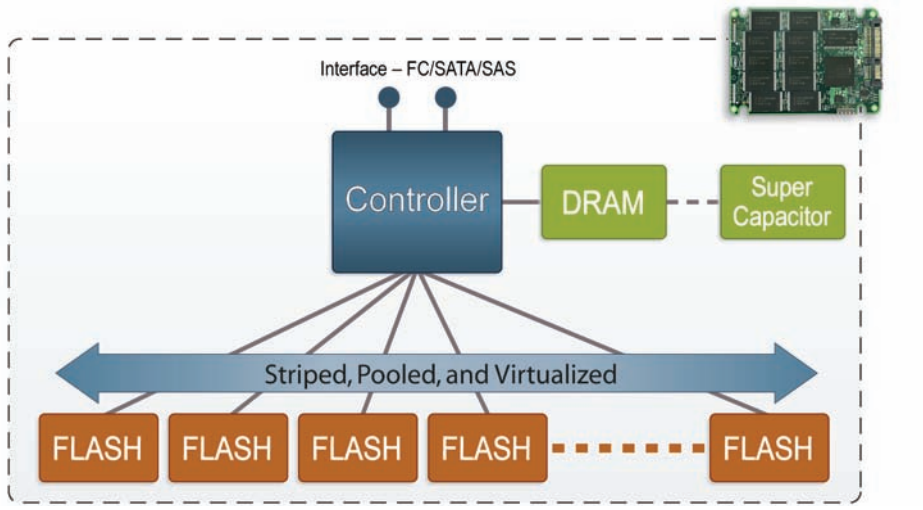

Figure 2. Enterprise solid state drive high-level architecture

Several advancements in Flash technology characteristics are making it possible to utilize SSDs in the enterprise datacenter.

- *Performance*
  Flash technology completes operations in microseconds, placing it between hard disk drives (milliseconds) and DRAM (nanoseconds) for access time. Because Flash technology contains no moving parts, it avoids the seek times and rotational latencies associated with traditional hard disk drives. As a result, data transfer throughput to and from solid state storage media is faster than electro-mechanical disk drives can provide — with enterprise SSDs providing tens of thousands of IOPS compared to hundreds of IOPS for hard disk drives.

- *Low power consumption*
  Hard disk drives draw significant amounts of power to operate the motor and spin the media. In contrast, the use of efficient Flash integrated circuits and a lack of motors and other mechanical parts result in enterprise SSDs consuming a fraction of the power of conventional hard disk drives. In fact, enterprise SSDs use only 5 percent of the power used by hard disk drives when idle, and as little as 15 percent when performing operations. As a result, enterprise SSDs produce less heat in the system chassis.

- *Cost*
  While Flash devices are more expensive per gigabyte than a comparable disk drive, Flash memory costs are dropping significantly year over year. In addition, as electricity costs rise and Flash memory costs decrease, the relative cost per available gigabyte and cost per IOPS of Flash memory improves. For example, hard

disk drives cost approximately $1.25/IOPS, compared to only $0.02/IOPS for enterprise SSDs. Since hard disk drives must be powered on to be available, the low power consumption of enterprise SSDs makes them a smart choice for datacenters looking to reduce operating costs. While a gigabyte of mechanical disk costs less than a gigabyte of Flash memory, the fact that Flash memory outperforms hard disk storage by at least an order of magnitude in reading and writing data makes the cost per gigabyte of Flash devices exceptionally low.

- *Reliability*
  While enterprise SSDs provide similar functionality to traditional hard drives, they offer improved reliability features. Both hard disk drives and enterprise SSDs support bad block management, wear leveling, and error correction codes (ECC) to foster data integrity. However, unlike hard drives that use a motor to spin magnetic media and a read/write head that must move to perform operations, enterprise SSDs contain no moving parts — data is stored on integrated circuits that can withstand significant shock and vibration. In fact, enterprise SSDs operate in a wider thermal operating range and wider operational vibration range than hard disk drives to deliver significantly higher Mean Time Between Failure (MTBF) (2.0 million hours versus 1.2 million hours).

The Hybrid Storage Pool — A New Way to Manage Storage and Speed Application Throughput

As companies look for ways to correct the imbalance between system processing needs and storage system throughput capabilities, finding an approach that optimizes both low cost per GB and cost per IOPS is essential. Doing so requires making trade-offs between cost and performance. Today organizations make that decision based on how quickly data is needed or available budget. As a result, datacenters use a variety of device types to store, archive, and access information (Figure 3).

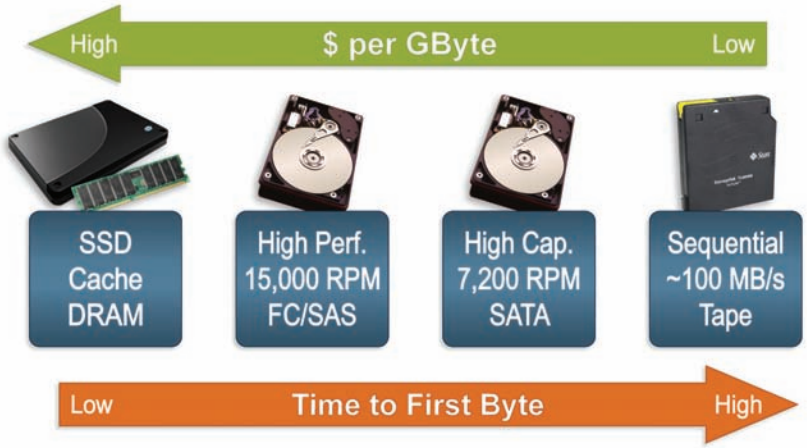

Figure 3. Enterprise devices present cost and performance trade-offs

Unfortunately, such solutions often keep applications waiting for data, slowing overall performance. Over the last decade CPU and Flash technology benefitted from Moore's Law, delivering performance improvements. Yet storage systems have not made similar strides (Figure 4). Hard disk drives still use the similar mechanical designs based on a moving arm and rotating disk platter, and rotational speeds have begun to stagnate. As a result, new drives tend to provide increased capacity at existing rotational speeds.

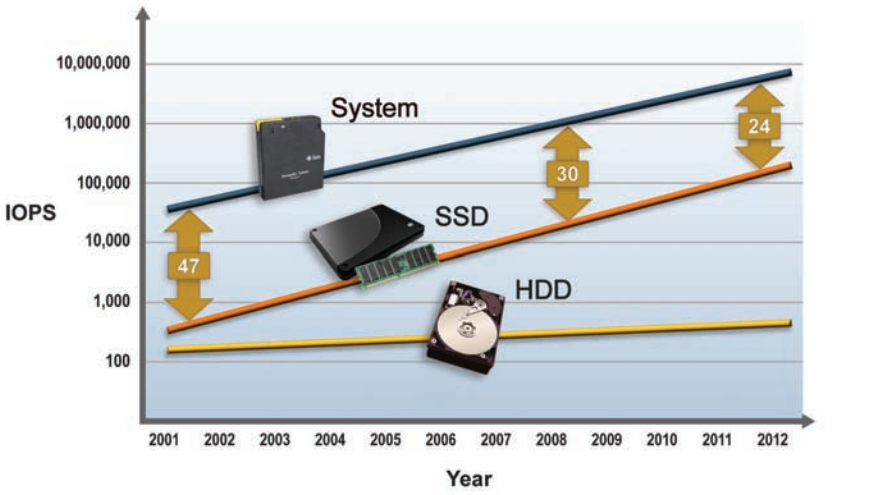

Figure 4. CPU, Flash, and hard disk drive technology advancement trends

With hard disk drive performance affected by seek, rotation, and transfer times, the latency created when transferring data to and from drives results in I/O bottlenecks. Drives simply cannot transfer data as fast as systems make requests. Processes are put to sleep waiting for information, and CPUs switch to other tasks to continue performing work. Furthermore, the strategy of adding expensive DRAM to systems in an effort to help alleviate the slowdown breaks down as data sets continue to double every two years. The result — the high-powered processors in the system are constantly waiting for data on which to operate. With applications spending significant portions of time waiting for storage requests to complete, application performance remains sluggish.

Today applications running on current multicore, multisocket servers are increasingly held back by slow storage systems. Technological advancements are changing the way storage devices can be used to rebalance systems and storage and create optimized solutions. While hard disk drives provide the capacity needed to handle large amounts of I/O, they are slow to perform. On the other hand, enterprise SSDs provide required IOPS, yet cannot provide the capacity needed at competitive price points.

Replacing all hard disk drives in a system with enterprise SSDs is not economical for most datacenter storage infrastructures. The right approach combines the strengths of both technologies. Enterprise SSDs can be placed in a new storage tier to assist hard disk drives by holding frequently accessed data to minimize the impact of disk latencies

and improve application performance. By utilizing enterprise SSDs to handle CPU I/O, and hard disk drives to store massive data sets, a hybrid storage pool gives organizations significant performance gains without sacrificing capacity.

How the Hybrid Storage Pool Works

The hybrid storage pool places data on the appropriate storage to maximize performance and get costs under control. Using this new approach, a server accesses data stored on a combination of enterprise SSDs and hard disk drives. Communication routes are established over multiple host adapters to parallelize I/O and speed throughput. In comparison to the previous example, a quad-core server with a maximum memory configuration connected to enterprise SSDs and hard disk drive storage via multiple host adapters is balanced (Figure 5). Both the system and combined storage pool can generate hundreds of thousands of IOPS.

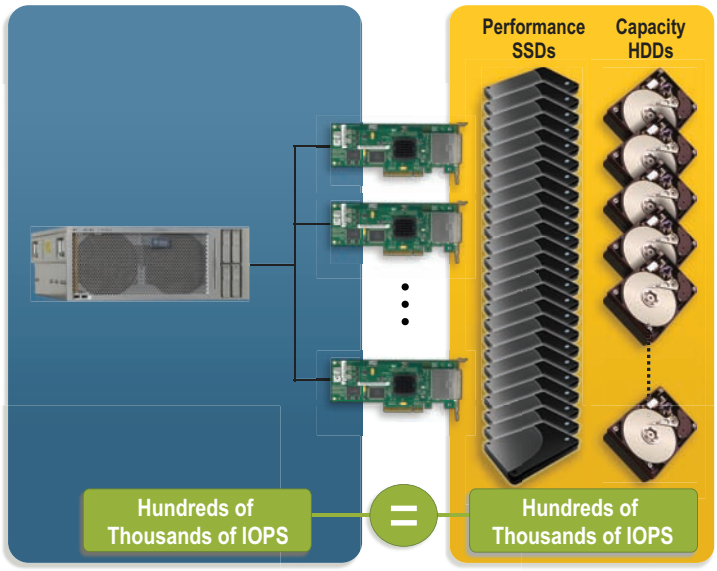

Figure 5. The hybrid storage pool provides a tiered architecture that lets data be placed on the right kind of device to balance system performance

Organizations can take advantage of the hybrid storage pool by combining enterprise SSDs and the Solaris ZFS file system.

Enterprise Solid State Devices

Enterprise solid state devices available today offer 16 GB to 128 GB of capacity and additional buffer cache in 2.5 inch or 3.5 inch form factors. Very thin and supporting a SATA-II interface, these devices deliver up to 250 MB/sec sequential read, up to 150 MB/sec sequential write, up to 30,000 random read IOPS and 7,000 random write IOPS. Continuing the industry trend to help reduce datacenter energy requirements, many enterprise SSDs utilize only 2.0 Watts when operating and provide the reliability features and robustness needed in enterprise environments.

Solaris ZFS™ File System

Solaris ZFS is an enterprise-class, general-purpose file system that provides virtually unlimited file system scalability and increased data integrity to large-scale solutions. Providing up to 21 billion YottaBytes[3] of capacity, this 128-bit, open source file system integrates traditional file system functionality with built-in volume management techniques. By automatically allocating space from pooled storage when needed, Solaris ZFS simplifies storage management and gives organizations the flexibility to optimize data for performance.

Key capabilities of Solaris ZFS related to the hybrid storage pool include:

- *Virtual storage pools* — Unlike traditional file systems that require a separate volume manager, Solaris ZFS introduces the integration of volume management functions. Breaking free of the typical one-to-one mapping between the file system and its associated volumes, Solaris ZFS introduces the storage pool model. Solaris ZFS decouples the file system from physical storage in the same way that virtual memory abstracts the address space from physical memory, allowing for more efficient use of storage devices. Space is shared dynamically between multiple file systems from a single storage pool, and is parceled out of the pool as file systems request it. Physical storage can be added to storage pools dynamically, without interrupting services. When capacity is no longer required by one file system in the pool, it becomes available to other file systems.
- *Data integrity* — Solaris ZFS uses several techniques to keep on-disk data self-consistent and eliminate silent data corruption, such as copy-on-write and end-to-end checksumming. Data is written to a new block on the media before changing the pointers to the data and committing the write. Because the file system is always consistent, time-consuming recovery procedures like fsck are not required if the system is shut down in an unclean manner. In addition, data is read and checked constantly to help ensure correctness, and any errors detected in a mirrored pool are automatically repaired to protect against costly and time-consuming data loss and previously undetectable silent data corruption. Corrections are made possible by a RAID-Z implementation that uses parity, striping, and atomic operations to aid the reconstruction of corrupted data.
- *High performance* — Solaris ZFS simplifies the code paths from the application to the hardware, delivering sustained throughput at near platter speeds. Block allocation algorithms accelerate write operations, and consolidate many small random writes into a single, more efficient sequential operation. Indeed, an I/O scheduler bundles disk I/O to optimize arm movement and sector allocation to speed throughput. In addition, an intelligent prefetch performs read ahead for sequential data streaming, and can adapt its read behavior on the fly for more complex access patterns. Furthermore, data is striped automatically across all available storage devices to balance I/O and maximize throughput. Solaris ZFS

3. 1 YottaByte is equal to 1 trillion terabytes, or 10^{24} bytes.

immediately begins to allocate blocks from devices as soon as they are added to the storage pool, increasing effective bandwidth as each device is added to the system.

- *Simplified administration* — Solaris ZFS automates many administrative tasks to speed performance and eliminate common errors. Creating file systems is fast and easy. There is no need to configure, or reconfigure, underlying storage devices or volumes — these tasks are handled automatically when devices are added to a storage pool. In addition, administrators can guarantee a minimum capacity for file systems, or set quotas to limit maximum sizes.

Solaris ZFS provides a seamless and easy way to administer hybrid storage pools, taking advantage of the performance of enterprise-class SSDs and inexpensive hard disk drive capacity. Unlike traditional volume managers that simply use SSDs in a RAID stripe, Solaris ZFS integrates the volume manager with the file system and can use enterprise SSDs more effectively. For example, Solaris ZFS can use SSDs intelligently as a cache for both application and file system metadata, placing latency-critical data structures appropriately on Flash media and using algorithms to optimize data placement. In addition, Solaris ZFS provides acceleration of both read and write operations, and lets administrators configure the system to match workload demands (Figure 7). These concepts are explored in the sections that follow.

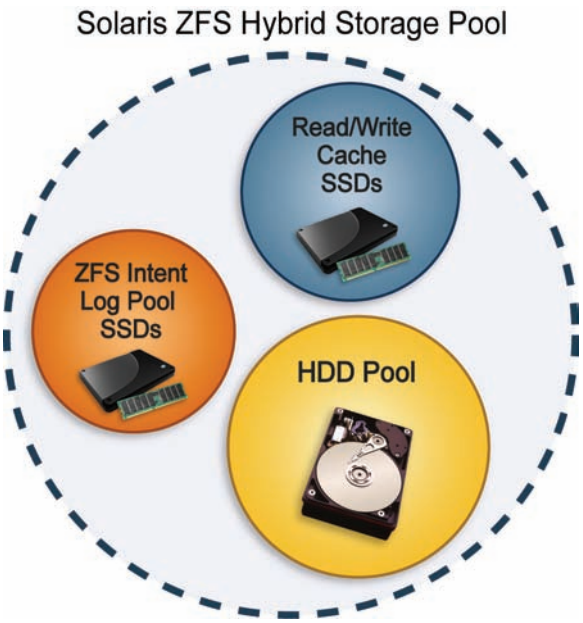

Figure 6. Solaris ZFS automates storage management and helps balance system performance with the ZFS Intent Log, a read cache, and a high capacity storage pool

Reduced Read Latency

Systems use memory to cache frequently accessed data for rapid access and improved performance. Once data is stored in the cache, future requests can be satisfied quickly by accessing the cached copy rather than fetching it from disk. Policies determine which data is held in the cache in an attempt to anticipate future needs. However, large working sets that cannot fit into memory can cause the cache to be ineffective.

Flash storage can be used to enhance caching operations in systems. Solaris ZFS combines main memory and enterprise SSDs into a large read cache and uses an Adaptive Replacement Cache (ARC) for its cache replacement algorithm. The ARC manages and balances the cache content using most frequently used (MFU) and most recently used (MRU) algorithms for storing data to, and retrieving data from, memory. A second-level ARC (L2ARC[4]) with smart caching and pre-fetching techniques lets Solaris ZFS use enterprise SSDs as a second-level cache to further speed read performance. Defective Flash blocks are treated as a cache miss rather than data loss, with information retrieved from hard disk to satisfy the request. The checksums built into Solaris ZFS are used to catch cache inconsistencies.

Using Solaris ZFS and an L2ARC stored on Flash devices helps:

- *Eliminate disk latency*
  Both the ARC and L2ARC are used to satisfy read requests from clients, and aim to avoid blocking a read request due to disk latency. Read operations can be serviced by the combined caches rather than disk drives. As a result, applications block for no more than the duration of Flash latency (< 100 us) rather than the latency of disk drives (up to 4 ms).
- *Speed access to working sets*
  Flash devices offer a faster way to access working sets that do not fit into available memory. While Flash devices are more expensive than fast hard disk drives per unit of storage, caching a very large working set on Flash devices costs less than storing all the data on fast disks when the performance advantages of Flash technology are taken into account.
- *Enhance cache performance*
  The L2ARC uses an evict-ahead policy. Cache entries are aggregated and predictively pushed out to Flash devices in order to distribute overhead across large write operations and eliminate additional latency that could arise when an entry is evicted from the cache. A ring buffer is used to manage the L2ARC replacement policy. When the end of the cache is reached, entries are stored at the beginning of the cache to avoid potential fragmentation. While it is possible for entries to be overwritten in the L2ARC prematurely, the most frequently accessed data still resides in DRAM-based cache.

4. The L2ARC is currently available only in the OpenSolaris™ operating system.

- *Speed system readiness by warming caches*
  The L2ARC stores a directory of data blocks written to the L2ARC. This practice helps to identify cache contents after a power or system failure and warm the cache. Instantly warming the cache reduces the time needed to restore production systems after planned or unplanned outages or other system resets.
- *Reduce the volatility of cache content*
  Since the L2ARC writes to flash devices slowly, and data on the system can be modified very quickly, it is possible for the contents of the L2ARC to be different than the data stored on disk. During normal operation dirty and stale entries are marked and ignored. After a system reset, stale data can be read from cache devices. However, metadata kept on the device and the checksums built into Solaris ZFS are used to identify this condition and seamlessly recover by reading the correct data from disk.

Reduced Write Latency

Solaris ZFS uses a log to record modifications to the file system. The Solaris ZFS Intent Log (ZIL) allows applications that demand synchronous writes to a permanent storage medium to benefit from apparent 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 enterprise SSD pool before committing the transaction to disk. The ZIL stores enough information to replay the transaction, if needed. These records are freed after 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 pointers 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.

- Solaris 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 I/O throughput. The file system writes uniformly to each byte in the intent log SSD to help alleviate flash wear out.

- Placing the ZIL on a low-latency enterprise SSD can help improve server throughput. For example, internal testing revealed ZIL latency in the 80 us to 100 us range. In this configuration, Solaris ZFS wrote the ZIL to the enterprise SSD in 8 KB chunks, with each write completing in 80 us — far faster than the milliseconds needed to access a hard disk drive.

Applications That Can Benefit

A variety of applications can take advantage of the hybrid storage pool approach. For many applications, adding enterprise Flash devices to a Solaris ZFS pool to create a hybrid storage pool provides transparent access to devices, and helps improve performance and reduce costs. Other classes of applications can benefit from being placed on a file system created from a pool that consists largely of enterprise Flash devices. In this scenario, applications must be configured to be aware of the Flash devices. Doing so extends the hybrid storage pool concept to applications where I/O latency is a critical factor to success. For example:

- Database applications can store log journals and indexes on enterprise SSDs.
- Distributed file systems, such as parallel NFS, as well as high-performance computing (HPC) applications, can place frequently accessed metadata on enterprise SSDs.
- Web 2.0 applications that use a memcached distributed memory system can also benefit from the use of enterprise SSDs.

How Flash Technology Can Help

As Flash technology moves into the enterprise, it holds promise for accelerating application performance and reducing datacenter energy consumption. By combining high-performance enterprise SSDs with high-capacity hard disk drives into a hybrid storage pool that is automatically managed by Solaris ZFS, IT organizations can rebalance systems, eliminate I/O bottlenecks, and improve the end user experience. As a result, IT organizations can take advantage of unique device characteristics and deploy systems that address specific application and workload problems.

About the Author

Roger Bitar is a staff engineer and technical marketing manager at Sun, and is responsible for developing technical product materials that provide compelling and differentiated technical evidence of Sun solutions. Over the last nine years, Roger has worked on improving ISV application performance and scalability on Sun platforms.

Acknowledgements

The author would like to recognize the following individuals for their contributions to this article:

- Adam Leventhal, Sun Microsystems Laboratories
- Denis Villfort, Systems Marketing

References

Hybrid Storage Pools: The L2ARC Blog
`http://blogs.sun.com/ahl/entry/hybrid_storage_pools_the_l2arc`

Narasimhan, Om. “Optimizing Systems to Use Flash Memory as a Hard Drive Replacement,” *Sun BluePrints Online*, June 2008. To access this article online, go to
`http://wikis.sun.com/download/attachments/17957083/820-4689.pdf`

Related Resources

Leventhal, Adam. “Flash Storage Memory,” *Communications of the ACM*, July 2008.

Megiddo, Nimrod and Dharmendra S. Modha. “ARC: A Self-Tuning, Low Overhead Replacement Cache,” *Proceedings of FAST ‘03: 2nd USENIX Conference on File and Storage Technologies*, March 2003.

Ordering Sun Documents

The SunDocs℠ program provides more than 250 manuals from Sun Microsystems, Inc. If you live in the United States, Canada, Europe, or Japan, you can purchase documentation sets or individual manuals through this program.

Accessing Sun Documentation Online

The `docs.sun.com` Web site enables you to access Sun technical documentation online. You can browse the `docs.sun.com` archive or search for a specific book title or subject. The URL is
`http://docs.sun.com/`

To reference Sun BluePrints Online articles, visit the Sun BluePrints Online Web site at:
`http://www.sun.com/blueprints/online.html`

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

10/08