Mark Moshayedi and Patrick Wilkison, STEC

For designers of enterprise systems, ensuring that hardware performance keeps pace with application demands is a mind-boggling exercise. The most troubling performance challenge is storage I/O. Spinning media, while exceptional in scaling areal density, will unfortunately never keep pace with I/O requirements. The most cost-effective way to break through these storage I/O

limitations is by incorporating high-performance SSDs (solid-state drives) into the systems.

While we often read in the press that SSDs will soon banish HDD (hard-disk drive) technology to the realm of tape storage, the fact is that SSD technology has only recently become ready for the enterprise. Not all SSDs are alike, and very few are appropriate for use as primary

ENTERPRISE SSDs

**Solid-state drives are finally ready for the enterprise.
But beware, not all SSDs are created alike.**

ENTERPRISE SSDs

storage devices in enterprise computing systems. Using flash storage in media players is fundamentally different from deploying the technology in 24/7 mission-critical operations.

With the advent of this new category of solid-state device, the potential for using SSDs in enterprise systems has become a reality, with profound implications for system performance. At the same time, leveraging the power of SSDs is difficult, and even identifying true enterprise-class SSDs is a major challenge. With these challenges in mind, we develop in this article a framework that can be used to assess SSD technology and determine its enterprise-readiness.

ARCHITECTING A PERFORMANCE SSD

The very first enterprise-class SSD was introduced in 2007. One of the key architectural dimensions of the product was that it combined the best attributes of two memory technologies: flash and DRAM. That dimension coupled with complex controller technology results in an entirely new class of SSDs for markets where performance is the key reason why customers would use SSDs instead of HDDs.

The primary applications that are now benefiting from this technology are those that are heavily dependent upon the drive I/O performance—for example, in enterprise storage and server applications where the I/O performance of the drives has a direct impact on the overall system performance and where cost is measured in performance and not just in capacity (cost/performance). In such applications, increased I/O is equal to increased revenue for the end user.

Two examples of such applications are:

- **Those that require a large number of the fastest HDDs (15,000-RPM FC/SAS drives).** To achieve fast I/O performance, these systems artificially limit the amount of

accessible drive space to only the outside orbit of the platters. This limits mechanical spindle movements. In addition, the data is striped across many drives to get to the desired IOPS (I/O operations per second) at the system level. One example of such an application has had as many as 9,000 drives in one system to achieve the required I/O performance. In this application an enterprise-class SSD has been able to replace the mechanical drives at the rate of 30 to one, reducing not only the number of drives from 9,000 to 300, but also the number of racks, enclosures, controllers, cables, switches, power supplies, and all other associated electronics. Switching to SSDs delivered not only a 50-percent upfront cost savings, but also a 90-percent cost savings associated with power and maintenance of the system over multiyear operations. Figure 1 illustrates the random I/O performance of an enterprise SSD that makes these cost savings possible.

- **Enterprise server applications where the overall performance of the system is limited by I/O performance of the drive subsystem.** In these applications, there is another category of enterprise SSD that offers much lower cost (relative to the highest-performance SSD) yet maintains overall performance characteristics of 10 times today's mechanical hard drives. With this approach, the additional cost associated with SSDs is justified by the overall performance-improvement gains at the server level and the power savings.

Another advantage of enterprise-class SSDs is eradicating latency. These SSDs provide access times in microseconds, rendering data-access times that are more like DRAM. With response times of this magnitude, SSDs behave more like main memory, yet possess all the “comforts” of stable disk storage in terms of persistence, communication protocols, form factor, etc. Enterprise SSDs provide native support for long-block data transfers (e.g., 520, 524, 528) that are obligatory for data-integrity and compatibility reasons in enterprise systems. In addition, because of the solid-state nature of SSDs, the mechanicals of a host system do not induce performance penalties as they do with HDDs.

A WORLD WITHOUT SPINNING PLATTERS

The key to achieving the right performance, capacity, and cost profile is to use NAND flash as the media in these drives. DRAM is too costly and obviously not a persistent storage medium (it loses data when power is removed). Future storage technologies claim to have an alternative means to address the needs of SSDs, but those technologies are not available today, have no clear path to scaling capacity-wise, and are too far away in terms of viability to be considered for SSD design for many years and possibly decades. NAND is the ultimate media choice, but it comes with its own series of design challenges.

NAND flash uses floating-gate technology and comes in two varieties: SLC (single-level cell, meaning the technology can store a single bit per memory cell) and MLC (multilevel cell, meaning the technology can store multiple bits per cell). SLC NAND flash costs twice as much as MLC flash; however, it is much more reliable and has much better endurance (write/erase cycles). It is also much faster (read/write speed) than MLC flash.

It is extremely important to note that not all NAND sources are the same. Although the theoretical benefits of SLC NAND and the specifications of all SLC devices suggest they are superior to MLC NAND, the reality is that not all SLC is as specified. An important take-away is that one must diligently test and screen the NAND in its fully packaged form, as well as within the SSD as part of the standard manufacturing test to ensure that the devices are appropriate for enterprise duty cycles.

In an overview of the common SLC NAND flash characteristics, the technology can be characterized as follows:

- Very fast sequential read (up to 40 MB per second per flash chip).
- Very fast random read performance (high number of IOPS).
- No rewrite capability; one must erase before a write.

- Erase is very slow and is done on a large block (128 KB today moving to 256 KB in the near future).
- Write is very slow and is done on a page level (usually 64 pages in a block, and current page size is 4 KB).
- Sequential write performance is fast (about 20 MB per second per flash chip).
- Random write is very slow.
- All NAND has a finite number of program/erase cycles, after which the physical blocks can no longer be programmed (SLC NAND can endure 100,000 program/erase cycles; MLC NAND can endure 5,000 program/erase cycles).
- All NAND has data-retention limitations, and the duration varies by the amount of utilization (SLC NAND has 10 years and MLC has five years of data retention; both

NAND is the ultimate
media choice, but it comes
with its own series of
design challenges.

SLC and MLC have declining data-retention characteristics when the cells are used, and the more intense the usage, the more precipitously the data-retention figure declines).

These divergent characteristics create quite the conundrum in terms of media management.

With these NAND flash characteristics, the first generation of low-end SSDs has the following performance characteristics:

ENTERPRISE SSDs

- Very fast sequential read and write performance.
- Very fast random read, which helps in fast application load and boot-up times.
- Very poor random write performance (less than one-tenth the speed of existing HDDs).
- Horrible performance under workloads with intermixed reads and writes.

Purely from a performance standpoint, the most pressing issue is achieving write speeds, particularly random write speeds, especially in workloads with both random reads and writes. Fundamentals of NAND programming introduce significant delays in the write performance of drives. At the heart of this phenomenon is the need to erase a block before writing. This process introduces latency and is the reason notebook SSDs are so slow in writes, particularly random writes and particularly when the host sends mixed reads and writes. Thus, the drive access patterns typified in the enterprise render these one-dimensional notebook SSDs worthless as a result of flaws in architecture and design.

THE PERFORMANCE AND RELIABILITY CHALLENGE

To achieve these ultimate drive-level performance characteristics, an entirely different product architecture is required from that used in notebook SSDs. An enterprise-class SSD is an optimized memory system with complex, tightly integrated hardware and software. At the heart of such a product is an elaborate chipset that performs all the vital communication protocols, as well as the critical media management. NAND is both a wonderful and a fickle medium. While the mechanical characteristics of NAND-based drives are obviously improved when compared with HDDs, NAND has unique reliability challenges not common to HDDs. It is the role of the SSD chipset and the coordinated manufacturing process that ultimately renders an SSD enterprise class.

SSD PERFORMANCE

An enterprise SSD implements high levels of parallel flash access and combines an optimal mix of two memory technologies—DRAM and NAND—to achieve the performance and reliability required by enterprise applications.

DRAM is extremely fast but requires power to maintain the data. DRAM devices are also used on disk drives, generally for cache to enhance the performance, but when used in disk drives they add to the risk of losing the data if power is lost. With enterprise-class SSDs, the DRAM is used for cache, but the drive adds power backup so when power is turned off, the device has enough power left to write the data from DRAM into flash (like the hibernate feature on laptops).

Enterprise SSDs use relatively large DRAM capacities within the drive (in the realm of a gigabyte). The use of DRAM helps the system overcome the biggest shortcomings of NAND, most notably random write performance. It allows the SSD to gather random writes and to use the good performance characteristics of NAND (relatively good sequential writes) to write the data at very high speeds. There is potential skepticism around the use of DRAM technology because all data written to flash would then become random data, but as explained earlier, NAND flash has very good random read characteristics.

By using DRAM in combination with power backup, the drive can effectively create a very fast nonvolatile memory device that can achieve access times and latencies that are 150 times faster than those of mechanical drives. This is one of the unique characteristics of SSDs that cannot be emulated by mechanical disk drives, no matter how many are used in parallel.

SSD RELIABILITY

Attaining the right levels of reliability for mission-critical applications requires that the drive maintain perfect data integrity, preventing any data loss or metadata corruption that would prevent the drive from rebooting following power removal. To achieve this, an enterprise-class SSD must have full data-path protection within the drive. Although this feature is currently used in enterprise-class HDDs, it is implemented in only one SSD. Figure 2 shows the salient architectural features of an enterprise SSD, including all the fully protected internal data flows.

Beyond the data-path protection, there are three pillars of media management that the drive must implement in a coordinated fashion:

Extensive full data-path error detection/correction.

An interesting phenomenon in SSDs is that the incidence of errors increases exponentially through utilization; this

translates into increasing effort from the drive to perform correction as the drive is exercised. Ultimately, all reads from the media will need to be corrected, which requires extensive and deep ECC (error correction checking) coverage. Performing such ECC well into the latter years of the drive's life without impacting system-level performance is a significant design challenge and is addressed only by enterprise-class SSDs.

Wear leveling. The SSD controller logic proactively deposits the writes into the optimal physical location in the NAND array such that no blocks are uneven or unduly worn. This is a delicate balance as the drive needs enough active wear leveling to ensure even drive wear, but the drive must not induce too much unnecessary data movement that could significantly over-exercise through excessive rewriting of data (see the section on write amplification later in this article).

Bad-block management. The process of managing bad blocks involves actively gauging the vitality of each independent block in the entire NAND array to ensure that bad blocks are removed from rotation and replaced with good blocks so that no data goes into corrupted blocks. An enterprise-class SSD implements bad-block management algorithms with multiple screens to determine the health of a block and to optimize the usable life of a

block, keeping it in rotation long enough to extract the maximum usable life without jeopardizing reliability or performance.

One area of concern with SSDs is the potential for drive corruption while the drive is moving data for house-keeping—the background operations required to manage the use and wear of the NAND blocks, as cited previously. The drive must constantly track and manage the physical utilization of the NAND array to ensure that the host gets maximum vitality from the drive. This process is run by both hardware and firmware, so the efficacy and reliability vary from vendor to vendor.

The two critical areas to assess are:

- The potential loss of metadata while background operations are taking place (which would render the drive incapable of booting following a power interruption and thus useless).
- Write amplification, which refers to exaggerated NAND programming resulting from brute-force techniques that fail to consider the impact of suboptimal NAND programming (all but one SSD in the market today induce inordinate amounts of write amplification, which dramatically reduces the vitality of the drive). Specifically, an erasure in NAND requires a complete 256-KB erasure. Even if the host writes 4 KB, most SSDs will do a full 256-KB erasure,

FIGURE 2

Enterprise SSD: A Look under the Hood

ENTERPRISE SSDs

which induces suboptimal wear of the NAND and can vary dramatically by SSD vendor. Some SSDs have write-amplification penalties, which render the life of the drive 75 percent less than true enterprise-class SSDs, which have considerably more efficient write-amplification techniques.

The final facet of this SSD housekeeping process is that most SSDs in the market that are not optimized for enterprise applications suffer from performance problems, because the background NAND management process ultimately becomes a foreground bottleneck. The notebook class of SSD needs relief (in the form of idle time) from the heavy enterprise duty cycles in order to do the housekeeping of the NAND. Enterprise applications need drives

to be poised for high performance 24/7 and thus cannot allow idle time.

Another important technique implemented in enterprise SSDs is the over-provisioning of NAND capacity, which is a vital means of achieving optimal performance and reliability. In the enterprise, there is no tolerance for varying performance in the drive. A drive cannot expect to have idle time available as a convenience to perform critical tasks. Having additional NAND within the drive will allow the drive to perform critical housekeeping tasks as background operations and then incorporate prepared blocks following background sanitization and preparation. When implemented properly, this technique significantly reduces write amplification and optimizes performance.

THE TECHNOLOGY APPLIED IN AN ENTERPRISE APPLICATION

While it is great to think through the sheer performance improvement of one drive technology versus another, let us now focus on the profound impact this has at the system level. Not only does SSD technology dramatically bolster system-level performance, but it also addresses one of the other most pressing issues in the data center: power reduction.

Enterprise-class SSDs present a compelling combination of performance and power savings that makes the technology a vital part of the storage technology spec-

trum. Figure 3 illustrates this power savings, comparing the power requirements to deliver 135,000 IOPS for an STEC enterprise SSD and a typical enterprise HDD.

It is important to note that SSDs have exceptional performance in small random transfers where the performance is optimized on 512-byte, 1-KB, 2-KB, 4-KB, and 8-KB random reads and writes. Once you have identified the appropriate role SSDs will play within the storage

To extract maximum system-level performance benefits, the key is to use SSD as a high-performance storage tier.

hierarchy, there are ways in which the system can tune access patterns to achieve maximum performance and reliability.

One key is to achieve the appropriate alignment of transfer sizes, which varies by product. Thus, you must know the SSD vendor intimately in order to implement optimal techniques to help achieve optimal performance and reliability.

In terms of compatibility, enterprise SSDs will work seamlessly within systems as drop-in replacements for HDDs. All of the aforementioned features are run entirely within the drive; SSDs are not dependent upon host-side file systems to perform all of the elaborate media management schemes. Various OEMs will develop unique ways to optimize system-level code to extract optimal performance, but there will be no requirement for the host to modify the manner in which the drives are addressed.

To extract maximum system-level performance benefits, the key is to use SSDs as a high-performance storage tier. As tiering of storage technologies proliferates (i.e., utilization of FC HDD as Tier 1, SATA HDD for Tier 2 and lower, and tape for archival), enterprise SSDs deliver an unprecedented performance profile; thus, they provide an entirely new tier of performance. Enterprise SSDs deliver performance more like main memory, so they can be used as a replacement for main memory, as implemented by Sun Microsystems with its ZFS acceleration design. SSDs can also be used to replace multiple high-performance HDDs, as EMC has implemented with its Symmetrix system. The convention is to refer to enterprise SSDs as Tier 0. Figure 4 shows a sample storage architecture, with an enterprise SSD in the Tier 0 position.

CONCLUSION

Enterprise storage and server OEMs are universally embracing SSDs to achieve the optimal balance of application demands, processor utilization, and cost. Clearly, SSD technology is emerging as the solution of choice for companies that need to improve the delivery of mission-critical applications while controlling costs and simplifying management. Not all SSDs are alike, however: to be truly enterprise class, a drive must be designed with the performance and reliability nuances of flash in mind. Drives that do not reflect these nuances will disappoint, as they will perform poorly and fail early—but those drives that are designed around flash will allow the technology to reach its full, disruptive potential. ◻

LOVE IT, HATE IT? LET US KNOW

feedback@acmqueue.com or www.acmqueue.com/forums

MARK MOSHAYEDI is president and CTO of STEC, where he has been for more than 16 years. Prior to that he worked in a variety of roles spanning engineering to sales at various companies including Texas Instruments, Sony, and Fujitsu. Throughout his career, he has specialized in storage and memory technologies. He earned his B.S. in electrical engineering from the University of California at Irvine and his M.B.A. from Pepperdine University.

PAT WILKISON is vice president of marketing and business development for STEC. He is responsible for STEC's products, spanning definition to introduction to management. He is also responsible for new market development. He earned his B.S. in systems engineering from West Point and his M.B.A. from the University of Southern California.

© 2008 ACM 1542-7730/08/0700 \$5.00