SSD flash drives enter the enterprise

By Ray Lucchesi

Solid-state disk (SSD) drive manufacturers have solved most of the problems associated with flash memory technology, paving the way for use in enterprise-level storage systems

Until recently, the idea of using solid-state disk (SSD) flash drives in an enterprise storage subsystem would have been deemed ludicrous. However, recent trends in NAND technology have made SSDs more viable in the enterprise storage market. The technology has become economical enough to favorably compare to traditional disk drives—at least in a price-performance context. In addition, many seemingly insurmountable shortcomings have been resolved.

Many disk array vendors, including EMC and Hitachi Data Systems (HDS), are adding SSDs to their disk systems. This technology could radically change the IT storage marketplace and seems ripe to do so.

Why SSD drives?

Flash-based SSD devices have a number of characteristics that are advantageous in enterprise storage applications, but perhaps most important is random read speed. In the past, applications that read data randomly often resorted to drive short stroking to gain significant performance advantages. In fact, with striping and short stroking of 10 high-end 15,000rpm hard disk drives performance gains of more than 16× are attainable, resulting in more than 3,000 random reads per second.

SSD flash drives can improve random read performance even more significantly. For example, one SSD drive can attain anywhere from 5,000 to 20,000 random reads per second; 10 SSD devices could easily handle 50,000 to 200,000 random reads per second. SSD drives are able to achieve their superior random read performance because they have almost no seek time and absolutely no rotational time reading NAND data.

Similar intrinsic technology advantages also afford SSDs a slight advantage during sequential read activity. However, the performance advantage is not as significant—e.g., about 250MBps for SSDs versus about 160MBps for 15,000rpm hard drives. Another advantage of SSD flash drives is their efficient power consumption. Comparing an active enterprise-class hard disk to an active SSD drive, the SSD drive uses only one-half to one-third the power needed by a typical 15,000rpm disk drive. Thus, replacing 10 hard drives with one SSD drive could result in considerable power, cooling, and space savings.

Who benefits from SSDs?

Obviously, enterprises needing superior random read performance will benefit most from SSD flash drives. Many vendors have specialized performance monitoring software and/or services to determine this requirement. Woody Hutsell, Texas Memory Systems' executive vice president, says that most customers looking at SSD technology usually have a good intuitive sense of their needs and can quantitatively support this intuition in a straightforward manner.

Data-warehousing applications with numerous random read operations are a natural fit for SSD flash (sometimes referred to as SSDf) devices. For example, data analytics analyzing customer purchase records, stock transaction history, and/or options data primarily do random reads. Many other applications can also benefit from SSD drives.

NAND workarounds

"Flash devices are asymmetric media," says David Flynn, chief technology officer at FusionIO. Simply stated, SSD flash drives can read blazingly fast but write excruciatingly slow. Writing NAND data first requires a NAND block erasure then a NAND page-programming (write) pass, both of which take a relatively long time. For example, one SSD drive has specifications of 130 random writes per second and 18,000 random reads per second, enabling the drive only to maintain a random write rate less than 1% of its read rate. In contrast, hard disk drives can typically maintain write rates close to 90% of their read rates.

As such, SSD write throughput and random write IOPS rate are major issues. To provide faster write throughput, most SSD vendors add parallelism in the write data path so that multiple NAND dies are written in parallel. However, data path parallelism doesn't help the random write rate. For this issue, many SSD vendors also provide a DRAM cache in the drive itself, which caches random write data and later de-stages this data sequentially to the NAND memory. But a write cache is no panacea. The cache must be large enough to smooth out random write activity, must support high data integrity by being ECC- or parity-protected, and must be able to be flushed to NAND for power failures.

Another problem with NAND is wear out. NAND technology can support only so many erase-program (write) passes before it fails—a condition typically referred to as the NAND write endurance problem.

NAND comes in two technologies—single-level cell (SLC) and multi-level cell (MLC). SLC typically supports up to 100,000 erase-program operations before failure, while MLC NAND supports only a fraction of SLC write cycles. As such, enterprise SSD drive vendors use SLC NAND and support wear leveling. With wear leveling, an SSD flash drive supports an onboard virtual memory/log structured file scheme and SSD block data is written to a physical NAND location having only minimal wear. In this way, write activity can be spread across all the NAND data cells in the drive, and thereby mitigate NAND's write-endurance problem.

Samsung and Sun recently announced SLC NAND that supports 500,000 write cycles before cell failure. These new SLC parts were discovered by selectively sampling chips to find ones that support a higher write endurance. It's noteworthy that these results were accomplished without any NAND technology changes and provide hope for future NAND technology offerings.

In combination with wear leveling, many drive vendors also reduce NAND's write-endurance issues by over-provisioning their SSD flash drives (i.e., supplying more NAND flash capacity in the SSD drive than its rated capacity). Obviously, write endurance is less of a critical issue with significant over-provisioning. With SSD over-provisioning, additional sparsely populated NAND blocks can be identified and re-written elsewhere in an efficient manner. Such background "garbage collection" benefits write performance by providing a fresh, empty NAND block to support write activity.

However, SSD wear leveling and write parallelism, while improving write-endurance and performance issues, cause yet another challenge: For every data byte written to an SSD flash drive more than one data byte is written to NAND flash. Intel calls this effect write amplification and wear leveling efficiency. For some SSD flash devices, these two factors can more than triple the data written to the NAND flash on the SSD drive. Of course, more over-provisioning can help mask this issue.

Even though wear leveling may result in writing more data than necessary, it can mitigate another troublesome, but lesser known, issue of NAND storage. Specifically, NAND memory can be negatively impacted by read and program (write) disturbs arising from over accessing a particular NAND location. This overuse of NAND locations causes bits within the NAND block to erroneously change values. Wear leveling, by redirecting SSD writes to lesser-used NAND locations, thus reduces the potential for program or write disturbs.

Another technology offered by many SSD vendors to combat read-and-write disturbs is sophisticated multi-bit error correction codes (ECC) such as 6-bit correction/8-bit detection ECC. Combined with error recovery procedures that refresh or move data from deteriorating NAND blocks, the read-and-write disturb issues of NAND storage appear controllable.

Finally, another issue surrounding SSD flash drives is cost. On the street, enterprise SSD drives cost anywhere from 10× to 30× more than comparable enterprise hard drives on a $/gigabyte basis. However, this cost differential can be justified given the elimination of 10 to 30 short-stroked hard drives.

"Most IT shops will use SSD devices for transient data that needs to be read very quickly," says Claus Mikkelsen, chief scientist at HDS. "For example, stock analysts who took five hours to make buy/sell decisions can now make these decisions in minutes with SSD devices."

But why only transient data? The answer requires a more discriminating look at NAND technology. NAND uses high voltage (20V) to perform program/write operations and occasionally this high voltage shorts out a NAND die, resulting in a catastrophic write short. One NAND die is about 8Gb, or 1GB, of data and losing this much data is similar to losing a disk drive.

One drive vender determined that some manufacturers' NAND chips had a catastrophic write short rate of 4% defects per million (DPM) dies per year. At that rate, a NAND die has an MTBF of approximately 219,000 hours. However, each 128GB SSD flash drive typically has 200 8Gb NAND dies and thus is expected to experience a die failure approximately every 45 days. Even at a 2% DPM failure rate, SSD drives using these NAND chips are expected to experience a chip failure every 90 days. In contrast, an enterprise-class hard drive has an MTBF rating of approximately 2 million hours, or 0.4% DPM, resulting in an expected failure only after 220 years of operation. As such, SSD flash drives have only a fraction of the reliability of hard drives.

RAID technology at either the SSD drive or die level could provide protection in handling catastrophic write shorts. However, this subsystem technology was not meant to accommodate such high DPM rates. RAID 5 works best with block—not die—failures. In addition, any RAID-5 array would require at least three active SSD drives, as opposed to one, to provide the storage necessary to write parity blocks and support block rebuild operations for adequate protection against die failures.
Moreover, with any introduction of a new performance-enhancing technology, the architecture of the entire storage subsystem should be reviewed. Traditionally, the storage performance bottleneck has always been the hard drive. With SSD flash devices, the performance bottleneck has moved upstream. How well system, subsystem, and server vendors adjust to this new reality will dictate SSD performance in the enterprise. For example, placing an SSD flash drive behind a subsystem's expensive DRAM cache to boost read performance makes little sense as any performance gain from DRAM cache is not significantly greater than that from the SSD drive alone. Also, drive interfaces were not meant for SSD's fast read rate and can often limit SSD sequential throughput.

Finally, while SLC is the SSD NAND technology of choice, primarily because of its write endurance, MLC technology dominates the NAND market. MLC NAND is found in many consumer devices, such as music players, digital cameras, cell phones, and USB thumb drives. As such, MLC technology represents 90%-95% of the market for NAND flash, but only has one-tenth the write endurance and performance of SLC. Given MLC's manufacturing dominance, it's unclear how long SLC NAND can be produced at economical prices. Vendors such as Samsung and Intel-Micron (a partnership) have stated that they are committed to providing SLC NAND technology for the foreseeable future at economical price points, but market forces may dictate otherwise.

Conclusion

The near future of storage subsystem environments may be further complicated by the recent introduction of SSDs into enterprise storage systems. The blazing random read rates of this technology may prove to be overwhelmingly attractive to many storage

customers. In fact, SSDs could replace the typical short-stroked disk drives prevalent today, especially with the critical cost differential gap narrowing significantly.

Also making SSD flash drives more attractive to enterprise-class users are the many advances made to resolve, or at least mask, current SLC NAND flash issues through sophisticated and advanced drive controller technology. For example, parallelism and DRAM cache combat write throughput and random write performance issues, and wear leveling and ECC alleviate issues of write endurance and flash disturbs. In addition, vendors have over-provisioned their SSDs to further decrease SSD shortcomings.

NAND reliability problems, on the other hand, have not been adequately addressed. A 4% DPM rate, or even a 2% rate, is a glaring problem. As lithography advances shrink NAND cell geometry even further, this problem may become even more pronounced.

Ultimately, whether SLC and MLC NAND technology will both be available, or just MLC, will be decided by the market. On the other hand, there are other storage technologies emerging from research labs vying for the enterprise storage market. The battle between SSD flash and traditional hard drives may have just begun, but the war for enterprise storage dominance will eventually expand beyond this lone battlefield.

About the author

Ray Lucchesi *is president of Silverton Consulting, a Storage, Strategy & Systems consulting services company, based in the USA offering products and services to the data storage community.*

mailto:info@silvertonconsulting.com

http://www.silvertonconsulting.com