SSDs in the Mainstream

Quick Note

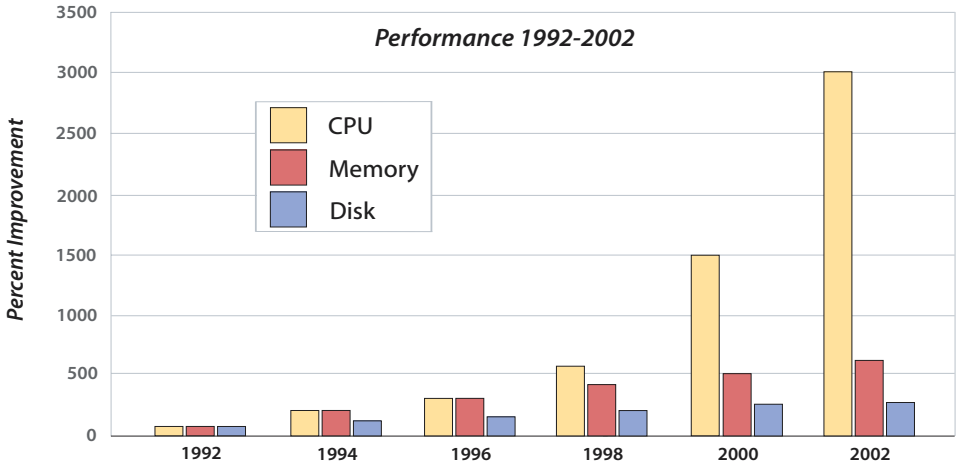
David Freund
21 January 2004

The largest single solid-state disk (SSD) sale ever was just announced by Texas Memory Systems. Dynamic Solutions International, an OEM for Texas Memory Systems, recently delivered a whopping 2.5-terabyte SSD configuration. But it will not, contrary to what one might expect, be used for some new, specialized, technical application. Instead, it's being installed to accelerate access to an existing storage farm that has grown to more than 300 terabytes—and that serves such pedestrian commercial applications as relational database software.

SSDs are extremely fast. They are able to access data in microseconds instead of the several milliseconds typical of hard-disk drives (HDDs)—a speed differential of at least two or three orders of magnitude. Despite their enormous performance advantage, SSDs have nonetheless been limited to niche applications for two major reasons. First, their cost-per-MB is *much* greater than traditional rotating magnetic media. Second, they require that IT administrators manually place specific files on them, and monitor their performance to ensure that the most-frequently accessed files are on the fastest, expensive hardware. Three trends in the storage industry could change that, however, and help SSDs become far more practical and attractive in mainstream datacenters than they have been since their appearance in the late 1970s.

The Storage Performance Gap

During the last decade, microprocessor performance has largely kept pace with Moore's Law, which (roughly stated) predicts that performance will double every 18 months. That has done a lot to boost system performance. Unfortunately, other system components have not kept pace, slowing the pace of overall system-performance improvements.



Sources: IBM Research, Microsoft Research

Similarly, disk capacity has been growing at a tremendous pace, about 60% per year, making onboard storage of data much cheaper. But the speed of data access has grown more slowly, holding back performance gains even further. The average HDD data-transfer rate, usually measured in megabytes per second (MB/s), has grown at about 40% per year. But the critical metric for most commercial applications is the number of I/O operations per second (IOPS), which has grown at a measly 16% per year.[1]

Why the huge discrepancy? Because today's disks are based on design principles created 50 years ago when the first disk drives were built: spinning platters of magnetic material with record/playback heads positioned over the portions of the magnetic media being accessed. Each head is attached to an "arm," the arms are connected together, and the whole assembly is moved back and forth by a special electro-mechanical device called an "actuator" to get the heads to the required positions for reading and writing data. The basic design has worked well, but is still conceptually tied to the days when music was scratched from vinyl disks by diamond needles. That kind of retro approach works in fashion, but not so well in high-level technology implementation.

1. As a result, the number of IOPS per GB of capacity has been *declining* over the years—a decidedly unhelpful performance trend.

That longevity is partly because of the pace with which disk-drive manufacturers have been able to improve areal density (the number of gigabits per square inch of magnetic media) to boost capacity—while continuing to drive down costs. As a result, high-capacity drives, now into the hundreds of gigabytes, have become common even in desktop PCs. And there have been economic incentives aplenty for this capacity growth, driven by an explosive demand for more and more data in both business and home contexts—whether for rich media like audio or video, or large databases for Internet-mediated transactions like those at Amazon.com or eBay.

Unfortunately, there have not been similar economic incentives for drive manufacturers to develop faster electro-mechanical components to store and access the data on this ever-more-dense magnetic medium. And even if there were incentives, the mechanical nature of the device makes innovations at rates comparable to Moore's Law extremely unlikely. So today's drives continue to have a single actuator for data access, and those actuators have become the most critical resource—and bottleneck—for physical-disk performance.

One would think that attacking this problem would be an obvious next move for disk manufacturers, but that simply hasn't been the case. The reasons are primarily economic. Whenever companies have attempted to bring even modest improvements to market—by adding more actuators to each drive,

for example—the additional cost killed the effort. It's been easier and cheaper to get more actuators by simply purchasing more drives.

In the past, several small companies have attempted—and failed—to sell drives using multiple actuators.[2] But it takes tremendous investment—and determination—to drive such a "new" design into the marketplace, since it would be competing with the volume economics available in commodity single-actuator designs. Performance increases for single-actuator drives have more than kept ahead of the "good enough" response-time requirement of most customers, reducing the demand for alternatives.

For these economic reasons—in addition to perhaps more-fundamental matters of the underlying performance limitations of mechanical devices compared to electronic ones—the gap in performance between disk drives and DRAM continues to grow.

Minimizing the Performance Gap

That performance gap has a significant impact on overall system performance. Boosting the speed on only one part of a system can leave bottlenecks that occur in other parts, limiting the benefit to the entire system.[3] As microprocessors continue to outstrip the data access speed of other components, they end up wasting more time waiting for data to be read from memory or disk. Some tactics can minimize the differential, however, to bring the IOPS rate closer to the level needed to keep up with the CPU.

Spread Out the I/O. One option is to use multiple disks and access them in parallel—the approach used in RAID storage arrays. To achieve an increase in IOPS, one must increase the number of disk actuators, the electro-mechanical assemblies that move each disk's magnetic heads over its spinning media. Rather than adding actuators to each disk, it's possible to add them by using more disk drives. Since disk capacity has been growing much more quickly than access speed, this means a lot more capacity must be purchased to obtain the number of drives necessary to achieve a balanced system.[4] The result is higher actual throughput, though at lower performance per GB of capacity. Even with tricks such as "short-stroking"—changing firmware to limit how far a disk's actuator can travel from the spinning platters' outer edge—the trend is not encouraging. Overabundant storage is wasteful if the performance boost can't be rigorously cost-justified. And short-stroking ensures that some of each disk's capacity is wasted, no matter what the benefit. Nevertheless, customers have essentially three types of drives to choose from:[5] cheap, large, and slow; expensive, small, and fast; and not-so-expensive, not-so-big, and not-so-fast.

Avoid the I/O. Another option, which database-software makers have aggressively pursued, is to perform as much work as possible in the server's memory so as to avoid disk I/O whenever possible. There are few applications that can use this tactic, however. First, it requires significant development effort to tune an application to avoid I/O. Second, databases are growing so large that the amount of main memory that must be used to keep the "I/O-avoidance" levels acceptable has also increased, often to prohibitive levels. The DBMS vendors have spent decades on the algorithms they use to avoid "disk hits" to gain every possible ounce of performance. It's doubtful that most ISVs would replicate that work, or that end-user customers could afford to do so.

2. See, for example, http://www.storagereview.com/guide2000/ref/hdd/op/actMultiple.html
3. This concept is generally referred to as Amdahl's Law.
4. This also means that, for high-performance systems, comparisons based on price per MB, which are favored by many storage vendors, are often not terribly useful.
5. This stratification is not based on interconnect types such as Fibre Channel, Serial ATA, SCSI, etc. In fact, it's common within the product lines for each of these interconnects.

Cache the I/O. This option has been common for many years in storage products.[6] Array controllers handle I/O operations between a server and a disk-array LUN.[7] The controller maintains a local copy of recent transfers in its local memory—in effect, caching a LUN's data on behalf of the requesting server(s). The largest array-cache sizes at this writing are in the top end of the array-performance (and price) spectrum, at 128 GB. Along with other design tweaks meant to increase I/O performance, market leaders EMC, Hitachi, and IBM are continually racing to increase their cache sizes as quickly as memory density and pricing allow. The latest crop of network- or storage-fabric-resident storage-virtualization appliances like IBM's SAN Volume Controller (SVC) and Sun's PSX-1000 N1 Data Services Platform also include sizeable built-in caches to accelerate I/O operations.[8]

An Alternate Route

SSDs can be used as another form of cache, but have historically been too expensive to build much of a market. Currently, SSD costs average more than 30 times the price per MB of the same storage capacity on a high-end, caching disk array such as the Symmetrix DMX from EMC or the Lightning 9900 from Hitachi. The high price is such a significant barrier that only customers with extreme performance needs have been able to justify the comparatively high purchase and ongoing operational costs.

6. The growing gap in CPU and DRAM performance is a major concern, too, and is dealt with through the use of multiple levels of faster memory to cache main-memory contents.
7. A disk-array controller presents "logical unit numbers" (LUNs) to a number of connected compute servers.
8. Others that are based on a pure Fibre Channel switch design avoid caching, preferring instead to forward the Fibre Channel message to its proper destination as quickly as possible. See Illuminata report "Storage Virtualization Wars Part I: The Philosophical Debates" (July 2003) and see Illuminata report "Storage Virtualization Wars Part II: Vendors Speak Volumes" (July 2003) for a more detailed discussion of competing virtualization designs.

The sale by Dynamic Solutions International, however, provides an example of SSDs being used as just such a cache, despite the difference in price. The company further hopes that this deal, along with others that it and Texas Memory Systems have been winning, will serve as a harbinger of better days for SSD. The keys to such marketplace success: three trends—the adoption of tiered storage classes, storage automation, and storage virtualization—in addition to drops in SSD prices that someday might make them comparable to HDDs—could make SSDs not only attractive, but practical.

Tiered Storage. Of the three, the increasing popularity of tiered online storage probably raises the greatest potential for SSD. In traditional hierarchical storage schemes, only high-value data is stored on disk; the rest goes into optical jukeboxes, or tape—relatively cheap alternatives that come with a high performance penalty.

Now, however, many users are interested in storage devices that offer different "classes" of service (performance, level of data protection, etc.) and correspondingly different per-unit prices. Storage vendors are responding. Serial-ATA drives are appearing in datacenter-class storage arrays, and are being sold to store infrequently-accessed data. The more-expensive mid-range and high-end arrays with SCSI and Fibre Channel drives and large caches would then be reserved for moderately and highly frequently-accessed data, respectively. As this type of "tiered" storage-class model becomes the norm, the acceptance of an even higher-performing class—the fast and pricey SSD—would be easier than in the past.

Storage Automation. Newer storage-management products that automatically migrate data from one storage device to another based on user policies could also make SSDs more practical, by moving data automatically between SSDs and traditional disk-storage arrays, mitigating the need for the manual tuning that's been associated with—and is often the fly in the ointment of—SSDs. However, great care and planning remains a necessity to make

sure applications can handle such data movement, or that some form of volume-management software is used to hide the details of which devices are storing what data files.

Storage Virtualization. Storage virtualization products can provide that level of abstraction, by emulating disk drives so that servers never know they're linking to a heterogeneous storage tier, rather than a single set of disks. By using logical disk volumes to hide the physical details of what's being stored where, this type of product can lend application transparency to the use of SSDs. Many of these virtualization appliances (FalconStor's is a perfect example) can create a *de facto* storage tier by identifying heavily used regions of a disk, and caching in memory the data on it, thereby increasing performance. Combining automated data migration with virtualization would extend the transparency (somewhat) to the IT administrators. Without such virtualization, IT administrators will not be able to effectively manage their increasing storage capacities, and their downside risks (e.g. cost of operations, risk of downtime incidents) will increase.

SSD/HDD Price Convergence. SSD prices have been coming down, as recent Texas Memory Systems product launches illustrate. Its 16-Gbyte RamSan-320, introduced last July, has a list price of $36,000, or just over $2,000 per GB. That's a 64% drop per GB from the $5,600-plus/GB it was charging for the 8-GB, $45,000, RamSan-220, launched just four months earlier. There have also been a few reports that per-GB SSD prices are coming down at a faster rate than those of HDDs. But the numbers we've seen in these reports don't appear to be congruent with long-term DRAM and disk-drive price trends we've seen. It would seem that SSD prices are dropping faster than the price of the DRAM chips of which SSDs are largely made. Unless other factors are involved, this is not a sustainable trend; if it were, the price of an SSD would eventually be lower than the cost of the DRAM inside it. If, on the other hand, SSDs don't rebound to their old price levels, the smaller difference in price between SSDs and HDDs could also help increase the attractiveness of SSDs to a broader audience.

The combination of these three trends enables an SSD to be used as a top tier storage device, with only the most-frequently accessed (and most important) data kept on it *automatically* through storage-management software. As some of the SSD-held data becomes less "hot," it's moved automatically to a lower-tiered device and replaced with data from elsewhere that has grown hotter. No less critical, all of these machinations would be hidden from the data owners through virtualization. As far as the applications (or their users) are concerned, the data has never moved—which avoids costly (and sometimes risky) application maintenance and downtime. It's this combination that will make the use of expensive, high-performance SSD storage arrays attractive to mainstream business customers.

Conclusion

The recent sale of 2.5 Terabytes of SSD storage by Dynamic Solutions International is a significant milestone—especially because it's being used to accelerate workloads seen in many commercial IT shops. Does this sale by itself mean SSDs are going mainstream? No, not by itself. But if recent trends toward tiered storage, automated data placement, and virtualization hold steady—and they give every indication of doing so—then solid-state disk will become dramatically more attractive to a much broader range of customers and for a much larger set of applications. At some point, it may even break out of its current niche to become a standard way that high-performance storage is delivered.