SOLID STATE DRIVE APPLICATIONS IN STORAGE AND EMBEDDED SYSTEMS

Contributors

Sam Siewert, PhD
Atrato Incorporated

Dane Nelson
Intel Corporation

Index Words

Intel® X25-E SATA Solid-State Drive
Intel® X25-M SATA Solid-State Drive
SSDs
RAID

Abstract

Intel® X25-E and X25-M SATA Solid-State Drives have been designed to provide high performance and capacity density for use in applications which were limited by traditional hard disk drives (HDD), input/output (I/O) performance bottlenecks, or performance density (as defined by bandwidth and I/Os/sec per gigabyte, per Rack Unit (RU), per Watt required to power, and per thermal unit waste heat. Solid State Drives (SSDs) have also found a place to assist in capacity density, which is the total gigabytes/terabytes per RU, per Watt, and per thermal unit waste heat. Enterprise, Web 2.0, and digital media system designers are looking at SSDs to lower power requirements and increase performance and capacity density. First as a replacement for high end SAS or Fiber Channel drives, but longer term for hybrid SSD + Hard Disk Drive (HDD) designs that are extremely low power, high performance density, and are highly reliable. This article provides an overview of the fundamentals of Intel's Single Level Cell (SLC) and Multi Level Cell (MLC) NAND flash Solid State Drive technology and how it can be applied as a component for system designs for optimal scaling and service provision in emergent Web 2.0, digital media, high performance computing and embedded markets. A case study is provided that examines the application of SSDs in Atrato Inc.'s high performance storage arrays.

Introduction

Flash memory, especially NAND flash memory, has been steadily encroaching into new markets as the density has increased and the cost per gigabyte (GB) has decreased. First it was digital cameras, then cell phones, portable music players, removable digital storage, and now we are seeing the emergence of NAND based solid state drives (SSDs) in the consumer PC market. Some industry analysts predict that SSDs could be the single largest NAND market segment (in billions of GB shipped) by 2010.

The combined performance, reliability, and power of SSDs compared to traditional hard disk drives (HDD), explains the attraction of SSDs in the consumer marketplace.

Intel Corporation has launched a line of high performance NAND based solid state drives; the Intel® X25-M and X18-M Mainstream SATA Solid-State Drives utilizing MLC NAND, and Intel® X25-E SATA Solid-State Drive utilizing SLC NAND. The higher performance density and lower price point make them a practical choice for the storage and embedded markets.

> *"Some industry analysts predict that SSDs could be the single largest NAND market segment (in billions of GB shipped) by 2010."*

The purpose of this article is to examine the unique benefits of Intel® Solid State Drive (Intel® SSD) over traditional HDDs and competing SSDs, and to explore the benefits one could realize in using these new high performance SSDs in storage and embedded applications.

Solid State Drives versus Hard Disk Drives

Solid State Drives have no moving parts, unlike HDDs. The rotating media and servo-actuated read/write heads used to access HDD data are subject to mechanical failure and introduce seek and rotate latency. Capacity growth due to areal density advancement and low cost per gigabyte stored have been the main advantages of HDDs, but fast random access has always been a significant limitation.

> *"The rotating media and servo-actuated read/write heads used to access HDD data are subject to mechanical failure and introduce seek and rotate latency."*

Physical Differences

The main difference between an HDD and SSD is the physical media in which the data is stored. A HDD has platters that encode digital data with magnetically charged media. These magnetic platters spin at a high rate of speed (5,400, 7,200, or 15,000 revolutions per minute, or RPM) so that a servo-controlled read/write head can be positioned over the cylinders/tracks of sector data for data access. In an SSD the digital data is stored directly in silicon NAND flash memory devices. An SSD has no mechanical moving parts, which improves the durability in resisting physical shock or mechanical failure and increases performance density. An SSD then uses a controller to emulate a mechanical hard disk drive, making it a direct replacement for mechanical hard disk drives but with much faster data access due to the lack of the servo positioning latency in HDDs.

> *"An SSD then uses a controller to emulate a mechanical hard disk drive, making it a direct replacement for mechanical hard disk drives but with much faster data access."*

In addition to the memory, a solid state drive contains; an interface connector and controller, memory subsystem and controller, and a circuit board where all the electronics are housed.

Performance Differences

The main performance difference between HDDs and SSDs has to do with the limitation of the HDDs caused by the spinning mechanical platters. Two performance metrics that improve greatly are the random reads and writes per second, and the time it takes to enter and resume from a low power state.

Random read and write performance is measured in inputs/outputs per second, or IOPs. This is simply the number of reads or writes that can be completed in one second. A typical high performance 15-K RPM SAS hard drive can usually complete about 300 IOPs of random 4-kilobyte (KB) data. By comparison, the Intel X25-E SATA Solid-State Drive is able to process over 35,000 random 4-KB read IOPs, a difference of 117 times. The reason for this is that logical data locations on an HDD are directly mapped to ordered physical locations on the spinning physical disks. To access (read or write) that data, the disk must spin around to the correct location and the read/write head must move to the correct radius to access the data. Therefore, random data accesses require multiple ordered physical movements incurring significant mechanical access latency and significantly limiting performance.

> *"The main performance difference between HDDs and SSDs has to do with the limitation of the HDDs caused by the spinning mechanical platters."*

In a NAND SSD the data is stored in a virtual memory map on the NAND flash. Accessing any part of that is as simple as changing the address and executing the next read or write. Since most workloads are random in nature, especially as industries move toward multi-core compute engines, multithreaded operating systems, and virtual machines, random disk performance will only increase in importance.

The other main performance difference is the ability to resume operation from a low power state quickly. The lower power state for a HDD is accomplished by parking the read/write head off to the side and stopping the spinning platter. When the next read or write is requested the platter needs to be spun back up and the head has to be moved back in place, which can take on the order of seconds. In an SSD however, when it is not processing read or write requests, it can put itself into a low power state (through Device Initiated Power Management, or DIPM) and recover within a few milliseconds to service the next request. Hard drives take closer to seconds to do this, so they do not take full advantage of DIPM.

"In an SSD however, when it is not processing read or write requests, it can put itself into a low power state (through Device Initiated Power Management, or DIPM) and recover within a few milliseconds to service the next request."

Barriers to Adoption

With SSDs' higher performance, added reliability, and lower power, one may ask why they have not completely displaced HDDs. The main reason is cost, because SSDs cost many times more per gigabyte than mechanical hard drives today. There has been early adoption in markets that absolutely must have the performance, reliability, lower power, or resilience to shock and vibration, but mass consumer adoption will only happen when the cost of a comparably sized device approaches that of a Small Form Factor (SFF) hard disk drive.

"With SSDs' higher performance, added reliability, and lower power, one may ask why they have not completely displaced HDDs."

The second barrier is capacity, because SSDs typically have much smaller capacity than mechanical hard drives. As NAND flash densities increase, however, the capacity of SSDs will be large enough for most consumer, enterprise, and embedded marketplace needs.

Third, the limited number of write cycles per storage cell is a barrier to applications that mostly ingest data (50-percent writes, 50-percent reads) for later access. Flash memory, as it is erased and rewritten, will lose the capability to hold a charge after many program/erase cycles. This makes flash memory a consumable resource. However, with increased density, along with more comprehensive write wear-leveling algorithms, the longevity of solid state drives has improved.

Ideal Use Cases

Just because SSDs have an inherent advantage over HDDs in random read/write performance and in resumption from low power states doesn't mean that SSDs are better than HDDs in every case. Hard disk drives are excellent for storing massive amounts of data such as movies, music, and large amounts of digital content in general, due to their very low cost per gigabyte and continued improvements in areal density (gigabits/square-inch). Areal density improvements in HDD media technology have in fact followed or exceeded Moore's Law (capacity density doubling every 18 months or less), but access to that data has not improved at the same pace. The servo and rotational latency for access to data in HDDs has in fact been nearly the same for decades if you look at year-over-year improvements.

"Just because SSDs have an inherent advantage over HDDs in random read/write performance and in resumption from low power states doesn't mean that SSDs are better than HDDs in every case."

"Solid state drives, on the other hand, excel when the system requirements are more skewed toward performance, reliability, or power."

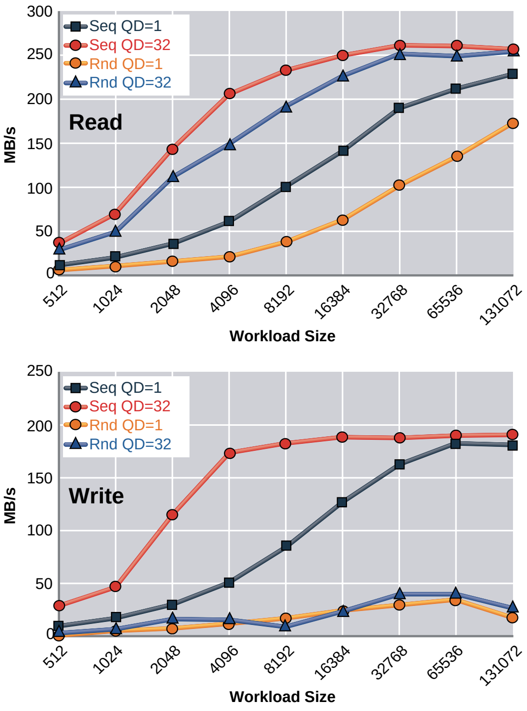

Figure 1: Performance for random/sequential read/write.Source: Intel Corporation, 2009

Solid state drives, on the other hand, excel when the system requirements are more skewed toward performance, reliability, or power. The performance and power metrics were described above, and the reliability of NAND storage in SSDs has been shown to be many times more reliable than mechanical hard disk drives especially in harsh environments.

Intel® Solid State Drive (Intel® SSD) Architecture and Design Considerations

The following sections describe the physical design and performance of Intel Solid State Drives.

Physical Design

Most solid state drives have a similar architecture; they all have a circuit board, interface controller, memory subsystem, and a bank of NAND flash memory. Intel SSD is no exception; it has 10 channels of NAND flash attached to the controller and it has a complex flash controller with advanced firmware, which allows it to achieve high random read write performance while at the same time managing the physical NAND to achieve the longest possible use of the drive.

Measuring SSD Performance

Traditional hard disk drive performance criteria apply directly to SSDs. The most common performance testing metrics are random and sequential sustained read/write bandwidth, random and sequential read/write IOPs, and power consumed in both active and idle states.

Sequential sustained read/write rates are mainly a reflection of the amount of parallel NAND channels that can be activated at once. Intel's 10 NAND channels allow for a very fast sequential throughput to the raw NAND, as is seen in the graph in Figure 1 showing sustained throughput versus data transfer size.

Random sustained read/write rates are mainly due to how well the controller and firmware can handle multiple outstanding requests. Newer SATA system architectures incorporate Native Command Queuing (NCQ), which allows multiple outstanding disk requests to be queued up at the same time. In random performance the Intel X25-M and X18-M Mainstream SATA Solid-State Drives, and Intel X25-E SATA Solid-State Drives provide read performance that is four times that of a typical 2.5" SFF HDD, twice that of a 3.5" enterprise HDD and for random IOPs they provide improvement by several orders of magnitude.

Sequential and random IOPs in SSDs are affected by the number of channels accessing the NAND memory, as well as the architecture of the controller and data management firmware running on that controller. In Figure 1 and Figure 2 you can see Intel's performance across various workload sizes.

Wear Leveling and Write Amplification

Since NAND flash wears out after a certain number of program and erase cycles, the challenge is to extract maximum use from all of the NAND cells. The SSD's controller firmware must make sure that the various program and erase cycles that come to the SSD from the host system are evenly distributed over all sectors of NAND memory providing even wear over the entire drive. If not designed correctly, a log file or page table can wear out one section of the NAND drive too quickly. Figure 3 shows how Intel handles these small writes and spreads the wear over the whole drive, which is shown by charting the program/erase cycle count of each NAND cell within the drive. As one can see, Intel's controller wears evenly across every cell in the drive by distributing the writes evenly.

The second main attribute that contributes to wear on the drive is called Write Amplification (WA), which is basically the amount of data written to the raw NAND divided by the amount of data written to the SSD by the host. This is an issue because NAND cells are only changeable in erase block sizes of at least 128 KB, so if you want to change 1 byte of data in the SSD you have to first erase the block that byte resides in and then update the entire block with that 1 byte modified. The problem arises that more program/erase cycles are being used up than the actual amount of data sent to the drive by the host. Without careful NAND data management, WA levels can range from 20—40x. This means more erases (20–40x) of the NAND are being done then required based on new data sent to the SSD. The ideal case would be a WA of 1.0, which means that exactly the same amount of data would be written to the NAND as would be written to the SSD by the host.

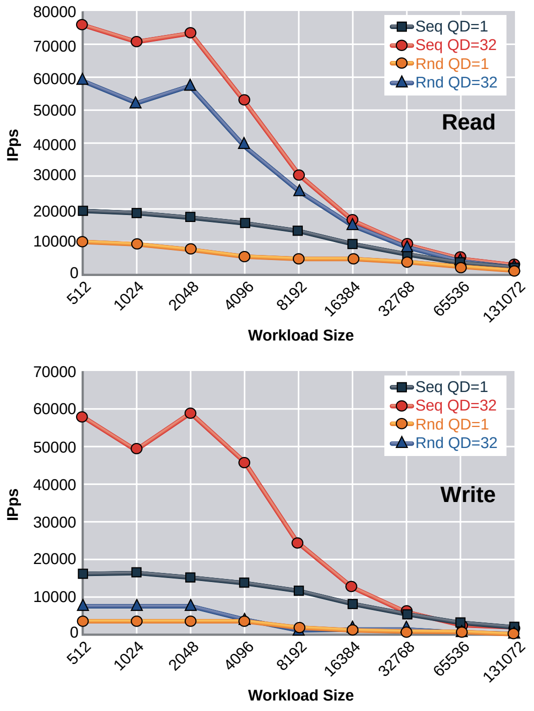

Figure 2: IOPs performance.
Source: Intel Corporation, 2009

Intel has taken a very close look at how to overcome this significant problem and has designed their controller accordingly. Intel's proprietary algorithms bring the WA of most compute applications very close to the ideal, and as one can see in the graph in Figure 4 for Microsoft Windows XP running MobileMark 2007 we measure a WA of less than 1.1.

Combining optimizations in both wear leveling and WA result in large increases to Intel SSD product longevity.

New Tier of Caching/Storage Subsystem

So far we have looked at a direct comparison between SSDs and HDDs without much examination of their application. There is the obvious direct-replacement market where HDDs are not meeting either the performance or reliability or power requirements of today's compute platforms. With high performance density SSDs the product designer has new options when designing embedded and scalable storage systems. The following sections examine how SSDs fit in today's storage and embedded products as well as how they could possibly be used in new ways to define tiered storage that enables new levels of access performance combined with scalability to many petabytes of capacity using both HDDs and SSDs.

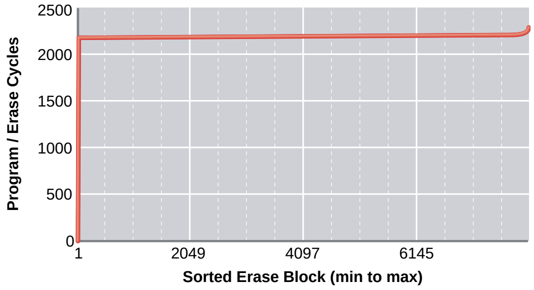

Figure 3: Erase cycle count showing Wear Leveling while running MobileMark 2007.
Source: Intel Corporation, 2009

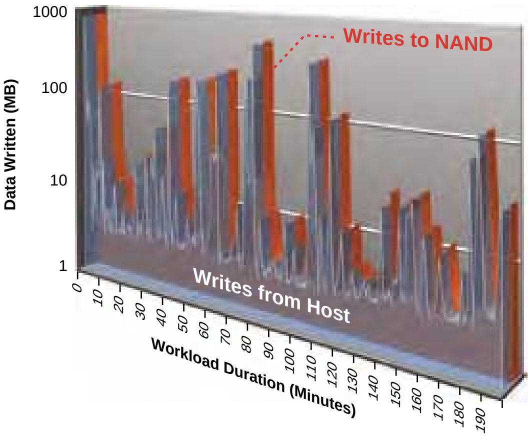

Figure 4: Microsoft* Windows XP Mobile workload writes. Source: Intel Corporation, 2009

SSD in Embedded Storage Applications

Emerging markets in Internet Protocol Television (IPTV), Video on Demand (VoD), the Digital Cinema Initiative (DCI), and Web 2.0 are bringing more on-demand high definition content to a broader base of users. This means that increased capacity and performance density is required from embedded devices as well as head-end, content distribution, and edge service systems. Storage in embedded mobile devices such as High Defintion (HD) digital cameras (1920 x 1080, 2048 x 1080, Digital Cinema 4096 x 2160, Red* Scarlet* 6000 x 4000 high resolution frame formats) and consumer devices are pushing embedded storage requirements to terabyte levels. Likewise, capacity requirements for head-end digital media services are reaching petabyte levels. For example, one hour of HD content un-encoded raw format for 1080p at 24 fps would require 489 gigabytes. A library of that content with 10,000 hours would require around 5 petabytes of formatted capacity. Most often video is encoded for delivery to consumer devices, with compression ratios that are 30 to 1 or more. Even with encoding, a library of 100,000 hours (similar to total content at Netflix*) encoded in typical high definition distribution/transport format requires 2 to 4 gigabytes per encoded hour on average, so at least 200,000 gigabytes or 200 terabytes total storage. Because of the multiplicity of transport encodings, content is stored in many formats, so capacity requirements are increased again. In this next section, we'll analyze how combinations of SSD and high capacity and performance density hard disk drives (HDDs) in tiered storage can help eliminate storage and I/O bottlenecks from both embedded and server systems and make the all-digital-content revolution a reality.

"Storage in embedded mobile devices such as High Defintion (HD) digital cameras (1920 x 1080, 2048 x 1080, Digital Cinema 4096 x 2160, Red Scarlet* 6000 x 4000 high resolution frame formats) and consumer devices are pushing embedded storage requirements to terabyte levels."*

Embedded Storage Growth

The increased capability of embedded storage and transport I/O in consumer devices has enabled the consumption of content at much higher bit rates. Progress in this embedded system segment has created demand for more high definition content from deeper content libraries. The emergence of affordable SSD for laptops over the next few years will help accelerate the demand for more on-demand high definition content. This means that the sources of content, starting with cameras, post-production, distribution, and finally delivery to consumers must all likewise upgrade to keep up.

"The emergence of affordable SSD for laptops over the next few years will help accelerate the demand for more on-demand high definition content."

General Architecture of Storage Applications

Today, most storage applications utilize HDDs, sometimes with redundant arrays of inexpensive disks (RAIDs) to scale capacity and performance. Embedded storage applications often make use of flash devices to store digital media, and small form factor HDDs to store larger content libraries. Content is often distributed on IEEE 1394 (such as FireWire*), USB 2.0 (Universal Serial Bus), or eSATA (external Serial Advanced Technology Attachment) external HDD when capacity is an issue, but this is less portable and often a significant I/O bottleneck. Media flash devices provide great I/O performance, but with very limited capacity (64 gigabytes is a typical high end device). For portable or semi-portable capacity and performance density, SSDs and SSD arrays will help change the landscape for portable storage architectures scaling to terabytes of capacity. As SSD cost continues down, the convenience, performance density, power, and durability of SSDs will likely drive mobile content storage completely to SSD. For system level content management with petabyte scale requirements, it is unlikely that SSD will replace HDDs for a very long time.

Today, most tiered storage moves content between flash media or SSD tiers and HDD tiers at a file level, with users actively managing how content is allocated between HDD and SSD tiers.

> "Today, most tiered storage moves content between flash media or SSD tiers and HDD tiers at a file level, with users actively managing how content is allocated between HDD and SSD tiers."

System Issues Today Using HDD

If we look at a 2K/4K format digital video camera typically used in cinema today, these cameras can produce 250 Megabits per second (Mb/sec) in JPEG 2000 (Joint Photographic Expert Group) format, which is about 25 MB/sec or 90 GB/hour. Today's 2.5" SFF mobile class HDDs can keep up with this data rate and have capacities up to 500 gigabytes, which provides reasonable capture support for a single camera. The drawbacks though are that one HDD can not support multiple cameras, they have lower MTBF (Mean Time Between Failure) when used in harsh environments (often the case in filming), and they are slower to download from the HDD to a backup RAID for post production. Some cameras support raw 2K/4K video capture, which is 53-MB per frame and at 30 frames/sec, 1.5-GB/sec data capture per stream. These types of emergent capture rates will require solid-state storage solutions.

> "Emergent capture rates will require solid-state storage solutions."

How SSDs Overcome These Issues

SSDs offer high-end digital 2K/4K/6K cameras the same advantages that smaller flash media provide consumers, but at capacities (160GB for Intel® X25-M SATA Solid-State Drive) that now make this a competitive option to HDD capture. This capacity offers approximately 2 hours of filming time and a capacity density that is competitive with SFF HDDs. The SSDs in this case would replace camera HDDs and offer lower power operation, equating to longer battery life, durability for filming in harsh environments, and high speed downloads to post-production RAID systems. The read rate of an Intel X25-E or X25-M SATA Solid-State Drive in sequential mode is at least four times that of typical SFF HDDs, so the download time will be far less. Even at raw 2K/4K rates of 1.5-GB/sec for uncompressed video ingest, it only requires 8 X25 SSDs to achieve full performance, however, at today's capacities (160 GB/SSD), the duration of ingest would only be 14 minutes (1.28 terabytes total SSD capacity for RAID0 mapping). One hundred percent ingest, rather than more typical 50 percent/50 percent write/read workloads is also a challenge for today's SSDs. Hybrid solutions with HDD backing SSD where SLC SSD is used as an ingest FIFO are perhaps a better approach and discussed in more detail in upcoming sections of this article.

> "The read rate of an Intel X25-E or X25-M SATA Solid-State Drive in sequential mode is at least four times that of typical SFF HDDs, so the download time will be far less."

Future Design Possibilities Exploiting SSD Advantages

The packaging of flash media into 2.5" and 1.8" SFF SAS/SATA (Serial Attached SCSI/Serial Advanced Technology Attachment) drives that are interchangeable with current SFF HDDs will help SSD adoption in the embedded segment of the digital media ecosystem. The SCSI (Small Computer System Interface) command set or ATA (Advanced Technology Attachment) command sets can both be transported to HDDs or SSDs over SAS with SATA tunneling protocols. This provides a high degree of interoperability with both embedded applications and larger scale RAID storage systems. As SSD cost per gigabyte is driven down and durability and maximum capacity per drive driven up by adoption of SSDs on the consumer side, the attractiveness of SSD replacement of HDDs for cameras will increase. Building hybrid arrays of SSD and HDD even for mobile field arrays provides a much better adoption path where cost/benefit tradeoffs can be made and systems right-sized. A

> "As SSD cost per gigabyte is driven down and durability and maximum capacity per drive driven up by adoption of SSDs on the consumer side, the attractiveness of SSD replacement of HDDs for cameras will increase."

key factor to success however is the development of software that can manage tiered SSD/HDD storage arrays for smaller mobile systems. This is even more important for post production, content delivery services, and the head-end side of the digital media ecosystem and will be covered in more detail in the following sections of this article.

> *"RAID storage system developers like Atrato Inc. have adopted SFF HDDs to increase performance density of HDD arrays."*

Storage Challenges

Since magnetic media storage density has kept pace with Moore's Law, both storage consumers and the storage industry have focused on cost per gigabyte and capacity density as the key metric. However, access to that stored data in general has not kept pace. Most often access performance is scaled through RAID systems that stripe data and protect it with mirroring or parity so that more HDD actuators can be used in parallel to speed up access. The upper bound for HDD random data access is in milliseconds, which has meant that the only way to scale access to storage is to scale the number of spindles data is striped over and to pack more spindles into less physical space. RAID storage system developers like Atrato Inc. have adopted SFF HDDs to increase performance density of HDD arrays. The Atrato V1000 SAID (Self-Maintaining Array of Identical Disks) has 160 SFF HDDs (spindles) packed into a 3RU (rack unit) array. This is presently the highest performance density of any HDD RAID solution available. At the same time, the emergence of SSDs in capacities that approach HDD (today on can get a 160-GB Intel X25-M SATA Solid-State Drive compared to 500-GB 2.5" SATA HDD) and cost per gigabyte that is only ten times that of HDD, has made tiered hybrid storage solutions for terabyte and petabyte scale storage very attractive. Rather than direct HDD replacement, tiered storage solutions add SSDs to enhance HDD access performance. The key is a hybrid design with RAID storage that is well matched to SSD tier-0 storage used to accelerate data access to larger HDD-backed multi-terabyte or petabyte stores. The fully virtualized RAID10 random access, no cache performance of the Atrato V1000 array is up to 2-GB/sec at large block sizes with IOPs up to 17K at small block size (measured with an HP DL580 G5 controller, where the rate limiting factor is the PCI Express* generation 1 and memory controller).

> *"Rather than direct HDD replacement, tiered storage solutions add SSDs to enhance HDD access performance."*

General Architecture of Storage Applications

Today most storage includes RAM-based I/O cache to accelerate writes on data ingest and to provide egress acceleration of reads through I/O cache read-ahead and hits to frequently accessed data. However, read cache often does little good for workloads that are more random and because the RAM cache sizes (even at 256 to 512 GB) are a very small fraction of capacity compared to petabyte back-end RAID storage (far less than one percent). Likewise, the cache miss penalty for missing RAM and going to an HDD backend is on the order of a 1000 to 1 or more (microsecond RAM cache access compared to millisecond HDD access). So, misses in RAM cache are likely and the penalty is huge, making RAM cache a wasted expenditure.

> *"As SSD cost per gigabyte is driven down and durability and maximum capacity per drive driven up by adoption of SSDs on the consumer side, the attractiveness of SSD replacement of HDDs for cameras will increase."*

Figure 5 shows access patterns to storage that range from fully predictable/sequential to full random unpredictable access. Both SSDs and the high spindle density solutions perform well for random access. The SSDs provide this with the best overall performance and capacity density compared even to the high density HDD arrays like the SAID if cost per gigabyte is not an issue. The most interesting aspect of both of these emergent storage technologies is that they provide a performance matched tier-0 and tier-1 for highly scalable storage. In summary, the SSDs are about ten times the cost per gigabyte, but ten times the capacity/performance density of the SAID and the SAID is ten times the capacity/performance density of traditional enterprise storage. This can further be combined with a 3.5" SATA lowest cost per gigabyte capacity tier-2 (archive) when very low cost infrequently accessed storage is needed.

In the following sections, we'll examine how to tier arrays with an SSD tier-0.

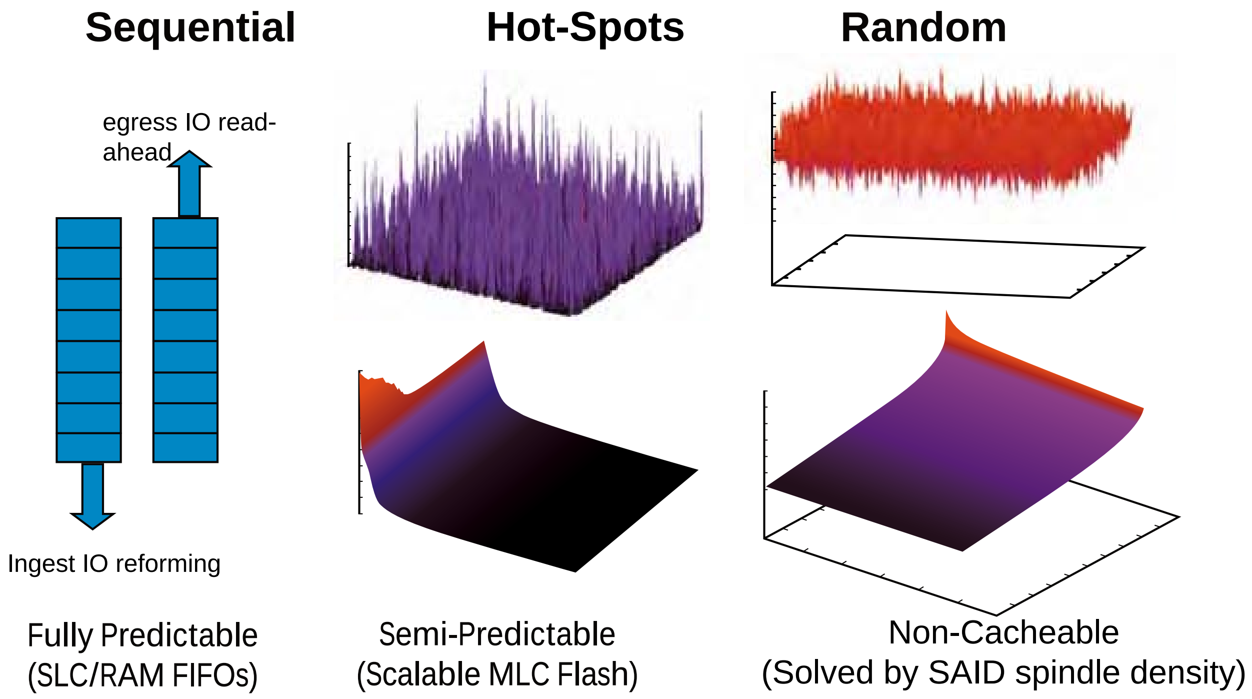

Figure 5: Performance range of access patterns observed by ApplicationSmart* Profiler. Source: Atrato, Inc., 2009

In Figure 5, totally random workloads are best served by storage devices with high degrees of concurrent access, which includes both SSD flash and devices like the Atrato SAID with a large number of concurrent HDD actuators. The biggest challenge arises for workloads that are totally random and access hundreds of terabytes to petabytes of storage. For this case, the SAID is the most cost-effective solution. For much smaller stores with totally random access (such as hundreds of gigabytes to terabytes), SSD provides the best solution. It is not possible to effectively cache data in a tier-0 for totally random workloads, so workloads like this simply require mapping data to an appropriate all SSD or highly concurrent HDD

"For totally predictable sequential workloads, FIFOs (First-In-First-Out queues) can be employed, with SLC SSDs used for an ingest FIFO and a RAM FIFOs used for block read-ahead."

array like the SAID based on capacity needed. The most common case however is in the middle, where data access is semi-predictable, and where SSD and HDD arrays like the SAID can be coordinated with intelligent block management so that access hot spots (LBA storage regions much more frequently accessed compared to others) can be migrated from the HDD tier-1 up to the SSD tier-0. Finally, for totally predictable sequential workloads, FIFOs (First-In-First-Out queues) can be employed, with SLC SSDs used for an ingest FIFO and a RAM FIFOs used for block read-ahead. The ingest FIFO allows applications to complete a single I/O in microseconds and RAID virtualization software is used to reform and complete I/O to an HDD tier-1 with threaded asynchronous I/O, keeping up with the low latency of SSD by employing parallel access to a large number of HDDs. The exact mechanisms Atrato has designed to provide optimal handling of this range of potential workloads is provided in more detail in upcoming sections after a quick review of how RAID partially addresses the HDD I/O bottleneck, so we can later examine how to combine SSDs with HDD RAID for an optimal hybrid solution.

"In general an HDD has a mean time between failure (MTBF) somewhere between 500,000 and 1 million hours."

Performance Bottlenecks that Exist Today

The most significant performance bottleneck in today's storage is the HDD itself, limited by seek actuation and rotational latency for any given access, which is worst case when accesses are random distributed small I/Os. Most disk drives can only deliver a few hundred random IOPs and at most around 100 MB/sec for sequential large block access. Aggregating a larger number of drives into a RAID helps so that all actuators can be concurrently delivering I/O or portions of larger block I/O. In general an HDD has a mean time between failure (MTBF) somewhere between 500,000 and 1 million hours, so in large populations (hundreds to thousands of drives) failures will occur on a monthly basis (two or more drives per hundred annually). Furthermore, environmental effects like overheating can accelerate failure rates and failure distributions are not uniform. So, RAID-0 has been enhanced to either stripe and mirror (RAID-10), mirror stripes (RAID-0+1), or add parity blocks every nth drive so data striped on one drive can be recovered from remaining data and parity blocks (RAID-50). Advanced double fault protection error correction code (ECC) schemes like RAID-6 can likewise be striped (RAID-60). So RAID provides some scaling and removes some of the single direct-attached drive bottleneck, but often requires users to buy more capacity than they need just to get better access performance, data loss protection, and reliability. For example, one may have 10 terabytes of data and need gigabyte bandwidth from it with small request sizes (32 K), which requires 32,768 IOPs to achieve 1 GB/sec. If each of the drives in the RAID array can deliver 100 IOPs, I need at least 320 drives! At 500 GB of capacity per drive that is 160 terabytes and I only need 10 terabytes. One common trick to help when more performance is needed from the same capacity is to "short-stroke" drives whereby only the outer diameter of each drive is used which often provides a 25-percent acceleration based on the areal density of the media.

"Virtualization of a collection of drives also requires RAID mapping and presentation of a virtual logical unit number (LUN) or logical disk to an operating system."

Virtualization of a collection of drives also requires RAID mapping and presentation of a virtual logical unit number (LUN) or logical disk to an operating system. This means that all I/O requested from the RAID controller must be re-formed in a RAM buffer and re-initiated to the disk array for the original request. The virtualization makes RAID simple to use and also can handle much of the error recovery protocol (ERP) required for reliable/resilient RAID, but comes at the cost of additional processing, store-and-forward buffering, and I/O channels between the RAID controller, the ultimate user of the RAID system (initiator), and the back-end array of drives. Applications not written with RAID in mind that either do not or cannot initiate multiple asynchronous I/Os often will not get full advantage of the concurrent disk operation offered by large scale RAID. Even with striping, if an application issues one I/O and awaits completion response before issuing the next, full RAID performance will not be realized. As shown in Figure 6, even if each I/O is large enough to stripe all the drives in a RAID set (unlikely for hundreds of drives in large scale RAID), the latency between I/O requests and lack of a queue (backlog) of multiple requests outstanding on the RAID controller will reduce performance.

"A much more ideal system would combine the capacity and performance scaling of RAID along with the performance density scaling of SSD in a hybrid array."

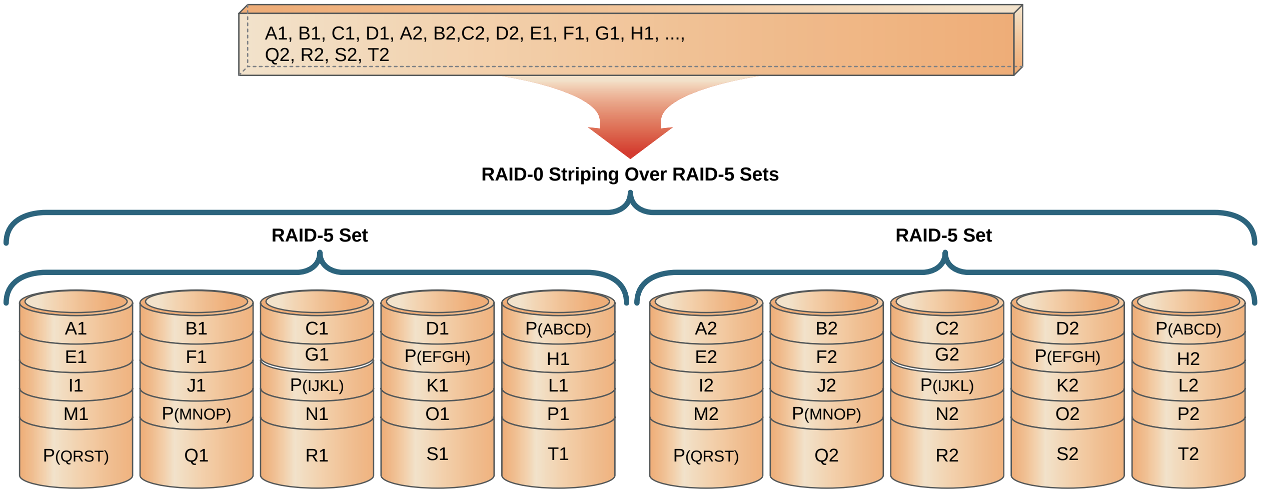

Figure 6: RAID set striping and striding example. Source: Atrato, Inc., 2009

A much more ideal system would combine the capacity and performance scaling of RAID along with the performance density scaling of SSD in a hybrid array so that users could configure a mixture of HDDs and SSDs in one virtualized storage pool. In order to speed up access with 10 terabytes of SSDs, one would have to combine 64 SSD drives into a virtualized array and stripe them with RAID-0. If they wanted data protection with RAID-10 it would increase the number of SSDs to 128. Even with lowering costs, this would be an expensive system compared to an HDD array or hybrid HDD+SSD array.

How SSDs Avoid These Bottlenecks

The bottleneck in embedded systems can be avoided by simply replacing today's HDDs with SSDs. The superior random read (and to a less extent write) provides a tenfold performance increase in general, albeit at ten times the cost per gigabyte. For small scale storage (gigabytes up to a few terabytes) this makes sense since one only pays for the performance increase needed and with no excess capacity. So, for embedded systems, the solution is simple drive replacement, but for larger capacity systems this does not make economic sense. What SSDs bring to larger scale systems is a tier that can be scaled to terabytes so that it can provide a 1-percent to 10-percent cache for 10 to 100 terabytes per RAID expansion unit (or SAID in the case of the Atrato Inc. system). Furthermore, the Intel X25-E and X25-M SATA

"The superior random read provides a tenfold performance increase in general, albeit at ten times the cost per gigabyte."

"An intelligent block-level managed solid-state tier-0 with HDD tier-1 can then accelerate ingest of data to a RAID back-end store, sequential read-out of data from the back-end store, and can serve as a viable cache for the back-end HDD store that is much lower cost than RAM cache."

Solid-State Drive SFF design allows them to be scaled along with the HDD arrays using common SFF drives and protocols. An intelligent block-level managed solid-state tier-0 with HDD tier-1 can then accelerate ingest of data to a RAID back-end store, sequential read-out of data from the back-end store, and can serve as a viable cache for the back-end HDD store that is much lower cost than RAM cache. In the following sections we will look at how SSDs are uniquely positioned to speed up HDD back-end stores geometrically with the addition of intelligent block management and an SSD tier-0.

Tiered Storage Using SSD and High Density HDD Arrays

The tiered approach described in the previous section can be managed at a file level or a block level. At the file level, intelligent users must partition databases and file systems and move data at the file container level based on access patterns for files to realize the speed-up made possible by tiers. Automated block level management using intelligent access pattern analysis software provides an increased level of precision in managing the allocation of data to the SSD tier0 and allows for SSD to be used as an access accelerator rather than a primary store. This overcomes the downside of the cost per gigabyte of SSDs for primary storage and makes optimal use of the performance density and low latency that SSDs have to offer.

Figures 7 through 9 show the potential for a coordinated SSD tier-0 with HDD tier-1 that is managed and virtualized by the Atrato Inc. virtualization engine. Figure 7 shows ingest acceleration through an SLC FIFO. Figure 8 shows sequential read-ahead acceleration through a RAM FIFO that can be combined with an MLC SSD semi-random read cache. The semi-random access SSD read cache has

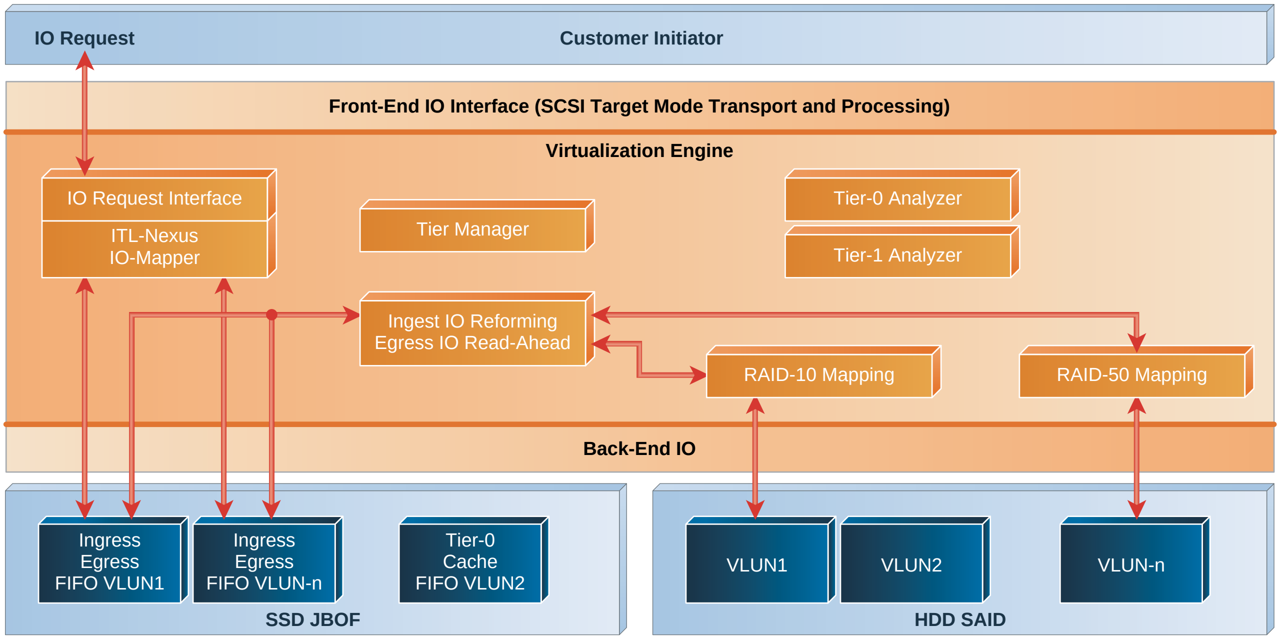

Figure 7: Ingest I/O reforming using SLC SSD and Egress read-ahead RAM cache. Source: Atrato, Inc., 2009

read hit/miss, write-through, and write-back-to-SSD operations. It can also be pre-charged with known high access content during content ingest. Any high access content not pre-charged will be loaded into SSD as this is determined by a TAM (Tier Access Monitor) composed of a Tier block manager and tier-0 and tier-1 access profile analyzers.

Ingest I/O acceleration provides a synergistic use of SSD performance density and low latency so that odd size single I/Os as shown in Figure 8 can be ingested quickly and then more optimally reformed into multiple I/Os for a RAID back-end HDD storage array.

Likewise, for semi-random access to large data stores, SSD provides a tier-0 block cache that is managed by the TAM profile analyzer and intelligent block manager so that the most frequently accessed LBA ranges (hot spots) are always replicated in the SSD tier-0. Figure 9 shows one of the many modes of the intelligent block manager where it replicates a frequently accessed block to the SSD tier-0 on a read I/O—the profile analyzer runs in the I/O path and constantly tracks the most often accessed blocks up to a scale that matches the size of the tier-0.

Overall, Figure 9 shows one mode of the intelligent block manager for write-back-to-SSD on a cache miss and HDD back-end read. The intelligent block manager also includes modes for write-through (during content ingest), read hits, and read misses. These tiered-storage and cache features along with access profiling have been combined into a software package by Atrato Inc. called ApplicationSmart* and overall forms a hybrid HDD and SSD storage operating environment.

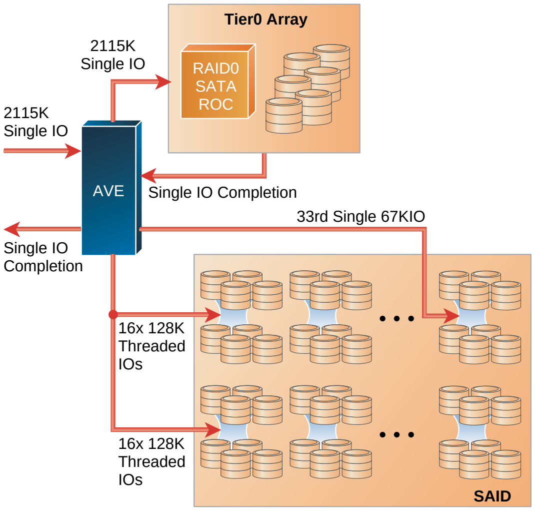

Figure 8: Ingest I/O reforming using SLC SSD and Egress read-ahead RAM cache. Source: Atrato, Inc., 2009

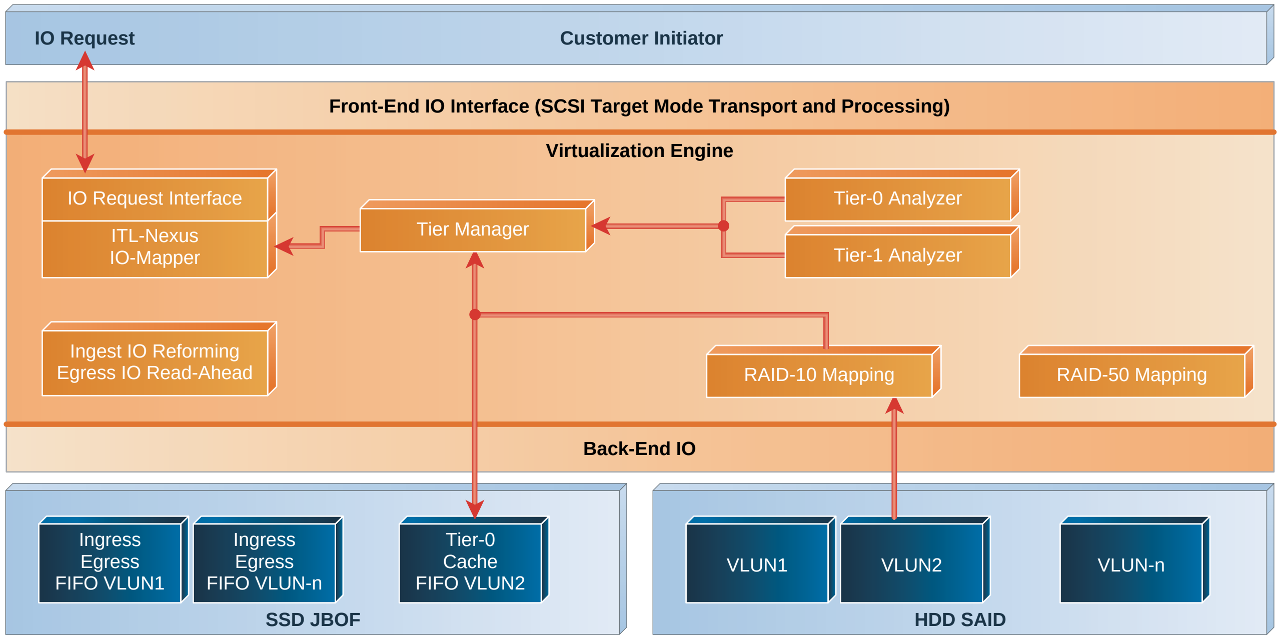

Figure 9: MLC SSD Tier-0 read cache opportunistic load of high access blocks on a read request. Source: Atrato, Inc., 2009

"The ability to scale to petabytes and maintain performance density comparable to SSD alone is the ultimate goal for digital media head-ends, content delivery systems, and edge servers."

This design for hybrid tiered storage with automatic block-level management of the SSD tier-0 ensures that users get maximum value out of the very high performance density SSDs and maximum application acceleration while at the same time being able to scale up to many petabytes of total content. Compared to file-level tiered storage with an SSD tier-0, the block-level tier management is a more optimal and precise use of the higher cost, but higher performance density SSDs.

SSD in Atrato Storage Application

For larger scale systems (tens to hundreds of terabytes up to many petabytes), SSDs are a great option for HDD access acceleration compared to RAM I/O cache due to scalability, persistence features, and cost per gigabyte compared to RAM. The ability to scale to petabytes and maintain performance density comparable to SSD alone is the ultimate goal for digital media head-ends, content delivery systems, and edge servers. As discussed previously, a tiered storage approach is much more efficient than simply adding additional HDDs in large arrays where more performance is needed even though the capacity is not.

Employing MLC Intel X25-M SATA Solid-State Drives as a read cache intelligently managed by the Atrato Inc. ApplicationSmart software and SLC Intel X25-E SATA Solid-State Drives for an ingest FIFO along with a RAM-based egress read-ahead FIFO, Atrato has shown the ability to double, triple, and quadruple performance from an existing V1000 RAID system without adding wasted capacity. Figure 10 shows a range of configurations for the Atrato V1000 with capacity ranging from 80 to 320 terabytes total capacity with SSD tier-0 1RU expansion units for access acceleration. This example was composed assuming the use of an Intel® Microarchitecture, codenamed Nehalem, the dual Intel® X58 Express chipset with off-the-shelf controller, which has at least 64 lanes of gen2 PCI-Express* and 8 total PCI-Express slots, 4 of which can be used for back-end SAID/SSD I/O and 4 of which can be used for front-end SAN or VOD transport I/O.

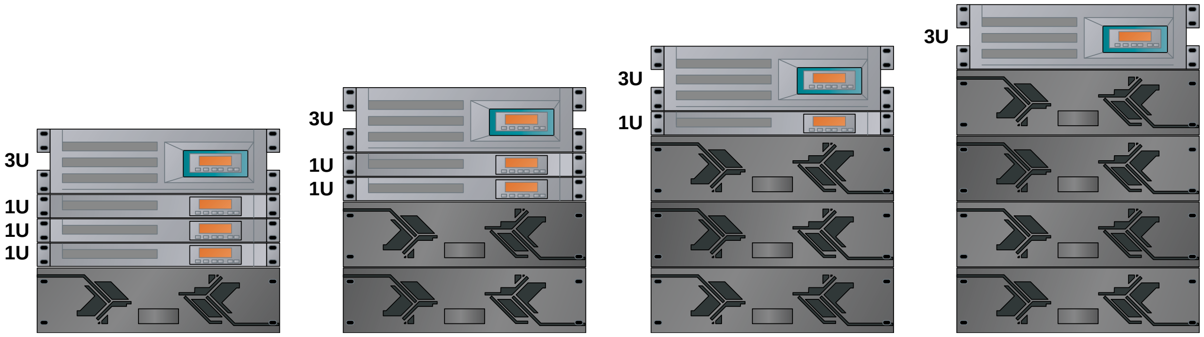

Figure 10: Scaling of SAIDs and SSD expansion units for access acceleration. Source: Atrato, Inc., 2009

There are 12 potential configurations that will allow customers to “dial in” the capacity and performance needed. Table 1 summarizes the configurations and the speed-up provided by SSD tier expansion units.

#SAID, #SSD Units	SSD Read Cache (TBs)	SSD ingest, egress (TBs)	BW (GBps)	IOPs	Capacity (TBs)	Cost, Capacity, Performance Normalized Score
4, 0	0	0	5.6	64000	320	2.4
3, 1	1.6	0.896	5.7	102000	240	2.4
2, 2	3.2	0.896	5.8	140000	160	2.3
3, 0	0	0	4.2	48000	240	2.2
2, 1	1.6	0.896	4.3	86000	160	2.1
1, 3	4.8	0.896	5.9	178000	80	2.1
2, 0	0	0	2.8	32000	160	2.0
1, 2	3.2	0.896	4.4	124000	80	1.9
1, 1	1.6	0.896	2.9	70000	80	1.8
1, 0	0	0	1.4	16000	80	1.7
0, 4	6.4	0.896	6	216000	6.4	1.2
0, 3	4.8	0.896	4.5	162000	4.8	1.0
0, 2	3.2	0.896	3	108000	3.2	0.7
0, 1	1.6	0.896	1.5	54000	1.6	0.2

Table 1: Cost, capacity, performance tradeoffs for SSD and HDD expansion units. Source: Atrato, Inc., 2009

Looking at a chart of the cost-capacity-performance (CCP) scores and total capacity, this would allow a customer to choose a hybrid configuration that has the best value and does not force them to purchase more storage capacity than they need (nor the power and space to host it). The CCP scores are composed of average cost per gigabyte, capacity density, and equally valued IOPs and bandwidth in performance, with equal weight given to each category so that a maximum possible score was 3.0. As can be seen in Figure 10 and in Table 1, if one needs between 100 and 200 terabytes total capacity, a 2 SAID + 2 SSD Expansion Unit configuration would be optimal. Furthermore, this would deliver performance that would exceed 4 SAIDs assuming that the access pattern is one that can cache 3.2 terabytes of the most frequently accessed blocks out of 160 terabytes (2-percent cache capacity).

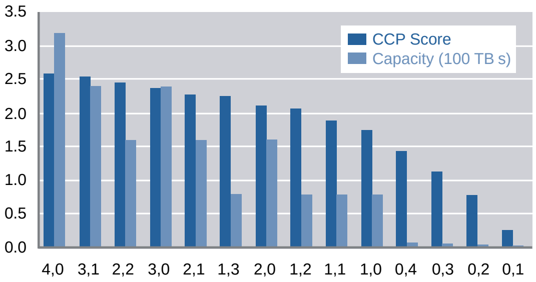

Figure 11: Cost, capacity, and performance score tradeoff. Source: Atrato, Inc., 2009

> *“Looking at a chart of the cost-capacity-performance (CCP) scores and total capacity, this would allow a customer to choose a hybrid configuration that has the best value.”*

> *"In general, storage I/O is such that I/Os can complete out of order and there are rarely data dependencies like there might be in a CPU cache."*

Computing the value of a read cache is tricky and requires a good estimation of the hit/miss ratio and the miss penalty. In general, storage I/O is such that I/Os can complete out of order and there are rarely data dependencies like there might be in a CPU cache. This means the penalty is fairly simple and not amplified as it might be when CPU cache causes a CPU pipeline to stall. A miss most often simply means an extra SAID back-end I/O and one less tier-0 I/O. The Atrato ApplicationSmart algorithm is capable of quickly characterizing access patterns, detecting when they change, and recognizing patterns seen in the past. The ApplicationSmart Tier-Analyzer simply monitors, analyzes, and provides a list of blocks to be promoted (most frequently accessed) from the back-end store and provides a list of blocks to be evicted from the tier-0 (least frequently accessed in cache). This allows the intelligent block manager to migrate blocks between tiers as they are accessed through the Atrato virtualization engine in the I/O path.

Figure 12 shows a test access pattern and Figure 13 shows the sorted test access pattern. As long as the most frequently accessed blocks fit into the tier-0, speed-up can be computed based on total percentage access to SSD and total percentage access to the back-end HDD storage. The equations for speed-up from SSD tier-0 replication of frequently accessed blocks are summarized here:

$$tier0\_LBA\_size = \left(\sum_{i=0}^{(sizeof\_sorted\_access\_counts-1)} evaluate\,(\,sorted\_access\_counts[i] > 0\,) \right) \times LBA\_set\_size$$

$$tier0\_hosted\_IOs = \sum_{i=0}^{(tier0\_LBA\_sets-1)} sorted\_access\_counts[i]$$

$$total\_sorted\_IOs = \sum_{i=0}^{(sizeof\_sorted\_access\_counts-1)} evaluate\,(\,sorted\_access\_counts[i]$$

$$tier0\_access\_fit = \frac{tier0\_hosted\_IOs}{total\_sorted\_IOs}$$

$$hit\_rate = tier0\_access\_fit \times tier0\_efficiency = 1.0 \times 0.6$$

$$speed\_up = \frac{T_{HDD\_only}}{T_{SSD\_hit} + T_{HDD\_miss}}$$

$$speed\_up = \frac{ave\_HDD\_latency}{(hit\_rate \times ave\_SSD\_latency) + ((1 - hit\_rate) \times ave\_HDD\_latency)}$$

$$speed\_up = \frac{1000\,\mu\text{sec}}{(0.6 \times 800\,\mu\text{sec}) + (0.4 \times 10000\,\mu\text{sec})}$$

In the last equation, if we assume average HDD latency is 10 milliseconds (10,000 microseconds) and SSD latency for a typical I/O (32 K) is 800 microseconds, then with a 60-percent hit rate in tier-0 and 40-percent access rate on misses to the HDD storage, the speed-up is 2.1 times. As seen in Figure 12, we can organize the semi-random access pattern using ApplicationSmart so that 4000 of the most frequently accessed regions out of 120,000 total (3.2 terabytes of SSD and 100 terabytes of HDD back-end storage) can be placed in the tier-0 for a speed-up of 3.8 with an 80-percent hit rate in tier-0.

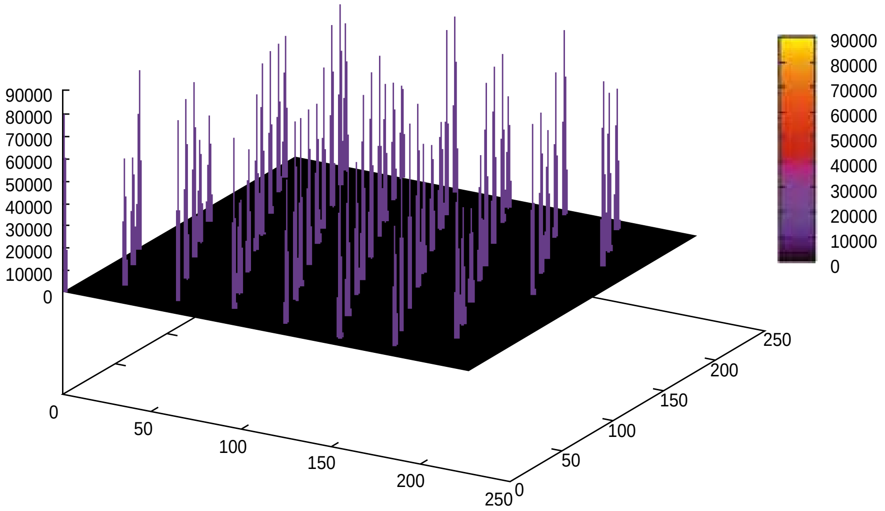

Figure 12: Predictable I/O access pattern seen by ApplicationSmart Profiler. Source: Atrato, Inc., 2009

Figure 13 shows the organized (sorted) LBA regions that would be replicated in tier-0 by the intelligent block manager. The graph on the left shows all nonzero I/O access regions (18 x 16 = 288 regions). The graph on the right shows those 288 regions sorted by access frequency. Simple inspection of these graphs shows us that if we replicated the 288 most frequently accessed regions, we could satisfy all I/O requests from the faster tier-0. Of course the pattern will not be exact over time and will require some dynamic recovery, so with a changing access pattern, even with active intelligent block management we might have an 80-percent hit rate. The intelligent block manager will evict the least accessed regions from the tier-0 and replace them with the new most frequently accessed regions over time. So the algorithm is adaptive and resilient to changing access patterns.

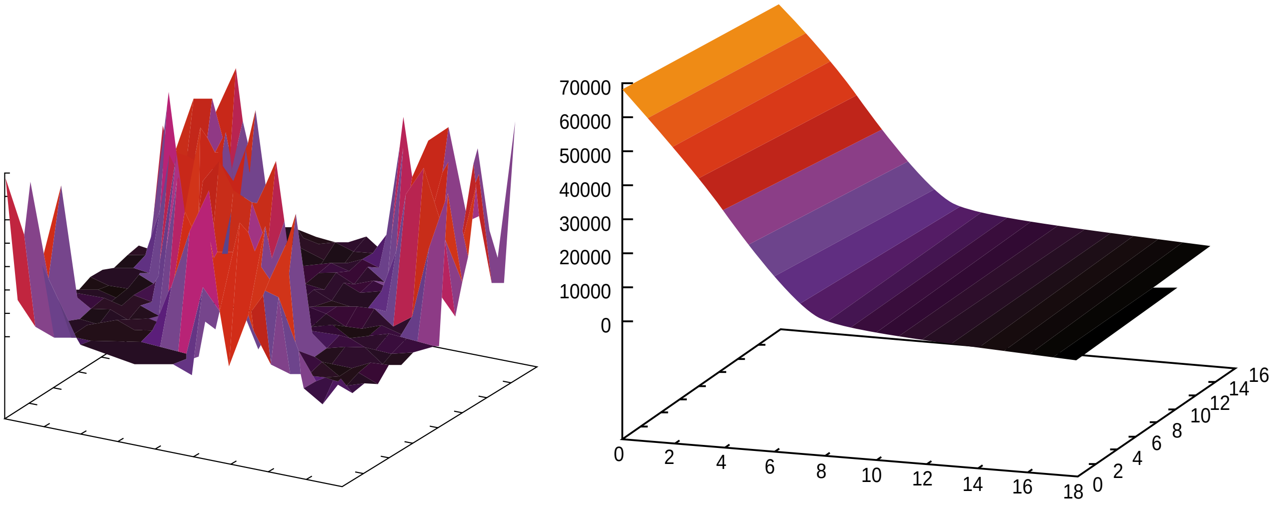

Figure 13: Sorted I/O access pattern to be replicated in SSD Tier-0. Source: Atrato, Inc. 2009

In general, the speed-up can be summarized as shown in Figure 14, where in the best case the speed-up is the relative performance advantage of SSD compared to HDD, and otherwise scaled by the hit/miss ratio in tier-0 based on how well the intelligent block manager can keep the most frequently accessed blocks in tier-0 over time and based on the tier-0 size.

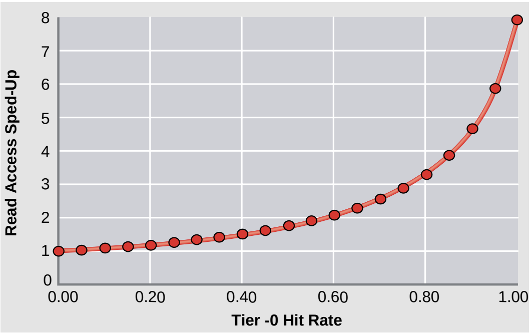

Figure 14: I/O access speed-up with hit rate for tier-0. Source: Atrato, Inc., 2009

It can clearly be seen that the payoff for intelligent block management is nonlinear and while a 60-percent hit rate results in a double speed-up, a more accurate 80-percent provides triple speed-up.

The ingest acceleration is much simpler in that it requires only an SLC SSD FIFO where I/Os can be ingested and reformed into more optimal well-striped RAID I/Os on the back-end. As described earlier, this simply allows applications that are not written to take full advantage of RAID concurrent I/Os to enjoy speed-up through the SLC FIFO and I/O reforming. The egress acceleration is an enhancement to the read cache that provides a RAM-based FIFO for read-ahead LBAs that can be burst into buffers when a block is accessed where follow-up sequential access in that same region is likely. These features bundled together as ApplicationSmart along with SSD hardware are used to accelerate access performance to the existing V1000 without adding more spindles.

Overview of Atrato Solution

The Atrato solution is overall an autonomic application-aware architecture that provides self-healing disk drive automation [9] and self-optimizing performance with ApplicationSmart profiling and intelligent block management between the solid-state and SAID-based storage tiers as described here and in an Atrato Inc. patent [1].

Related Research and Storage System Designs

The concept of application aware storage has existed for some time [2] and in fact several products have been built around these principles (Bycast StorageGRID, IBM Tivoli Storage Manager, Pillar Axiom). The ApplicationSmart profiler, Intelligent Block Manager and Ingest/Egress Accelerator features described in this article provide a self-optimizing block-level solution that recognizes how applications access information and determines where to best store and retrieve that data based on those observed access patterns. One of the most significant differences between the Atrato solution and others is the design of the ApplicationSmart algorithm for scaling to terabytes of tier-0 (solid-state storage) and petabytes of tier-1 (HDD storage) with only megabytes of required RAM meta-data to do so. Much of the application-aware research and system designs have been focused on distributed hierarchies [4] and information hierarchy models with user hint interfaces to gauge file-level relevance. Information lifecycle management (ILM) is closely related to application-aware storage and normally focuses on file-level access, age, and relevance [7] as does hierarchical storage management (HSM), which uses similar techniques, but with the goal to move files to tertiary storage (archive) [5][9][10]. In general, block-level management is more precise than file-level, although the block-level ApplicationSmart features can be combined with file-level HSM or ILM since it is focused on replicating highly accessed, highly relevant data to solid-state storage for lower latency (faster) more predictable access. Ingest RAM-based cache for block level read-ahead is used in most operating systems as well as block-storage devices. Ingest write buffering is employed in individual disk drives as well as virtualized storage controllers (with NVRAM or battery-backed RAM). Often these RAM I/O buffers will also provide block-level cache and employ LRU (Least Recently Used) and LFU (Least Frequently Used) algorithms. However, for a 35-TB formatted LUN, this would require 256 GB of RAM to track LRU or LFU for LBA cache sets of 1024 LBAs each or an approximation of LRU/LFU–these traditional algorithms simply do not scale well. Furthermore, as noted in [9] the traditional cache algorithms are not precise or adaptive in addition to requiring huge amounts of RAM for the LRU/LFU meta-data compared to ApplicationSmart.

> *"One of the most significant differences between the Atrato solution and others is the design of the ApplicationSmart algorithm for scaling to terabytes of tier-0 (solid-state storage) and petabytes of tier-1 (HDD storage) with only megabytes of required RAM meta-data to do so."*

Architecture

The Atrato solution for incorporating SSD into high capacity, high performance density solutions that can scale to petabytes includes five major features:

- Ability to profile I/O access patterns to petabytes of storage using megabytes of RAM with a multi-resolution feature-vector-analysis algorithm to detect pattern changes and recognize patterns seen in the past.
- Ability to create an SSD VLUN along with traditional HDD VLUNs with the same RAID features so that file-level tiers can be managed by applications.
- Ability to create hybrid VLUNs that are composed of HDD capacity and SSD cache with intelligent block management to move most frequently accessed blocks between the tiers.
- Ability to create hybrid VLUNs that are composed of HDD capacity and are allocated SLC SSD ingest FIFO capacity to accelerate writes that are not well-formed and/or are not asynchronously and concurrently initiated.
- Ability to create hybrid VLUNs that are composed of HDD capacity and allocated RAM egress FIFO capacity so that the back-end can burst sequential data for lower latency sequential read-out.

> *"Often these RAM I/O buffers will also provide block-level cache and employ LRU (Least Recently Used) and LFU (Least Frequently Used) algorithms. These traditional algorithms simply do not scale well."*

"In general, this algorithm easily profiles down to a single VoD 512-K block size using one millionth the RAM capacity for the HDD capacity it profiles."

With this architecture, the access pattern profiler feature allows users to determine how random their access is and how much an SSD tier along with RAM egress cache will accelerate access using the speed-up equations presented in the previous section. It does this by simply sorting access counts by region and by LBA cache-sets in a multi-level profiler in the I/O path. The I/O path analysis uses an LBA-address histogram with 64-bit counters to track number of I/O accesses in LBA address regions. The address regions are divided into coarse LBA bins (of tunable size) that divide total useable capacity into 256-MB regions (as an example). If, for example, the SSD capacity is 3 percent of the total capacity (for instance, 1 terabyte (TB) of SSD and 35 TB of HDD), then the SSDs would provide a cache that replicates 3 percent of the total LBAs contained in the HDD array. As enumerated below, this would require 34 MB of RAM-based 64-bit counters (in addition to the 2.24 MB course 256-MB region counters) to track access patterns for a useable capacity of 35 TB. In general, this algorithm easily profiles down to a single VoD 512-K block size using one millionth the RAM capacity for the HDD capacity it profiles. The hot spots within the highly accessed 256-MB regions become candidates for content replication in the faster access SSDs backed by the original copies on HDDs. This can be done with a fine-binned resolution of 1024 LBAs per SSD cache set (512 K) as shown in this example calculation of the space required for a detailed two-level profile.

- Useable capacity for a RAID-10 mapping with 12.5 percent spare regions
  - Example: (80 TB – 12.5 percent)/2 = 35 TB, 143360 256-MB regions, 512-K LBAs per region
- Total capacity required for histogram
  - 64-bit counter per region
  - Array of structures with {Counter, DetailPtr}
  - 2.24 MB for total capacity level 1 histogram
- Detail level 2 histogram capacity required
  - Top X%, Where X = (SSD_Capacity/Useable_Capacity) x 2 have detail pointers with 2x over-profiling
  - Example: 3 percent, 4300 detail regions, 8600 to 2x oversample
  - 1024 LBAs per cache set, or 512 K
  - Region_size/LBA_set_size = 256 MB/512 K = 512 64-bit detail counters per region
  - 4 K per detail histogram x 8600 = 34.4 MB

With the two-level (coarse region level and fine-binned) histogram, feature vector analysis mathematics is employed to determine when access patterns have changed significantly. This computation is done so that the SSD block cache is not re-loaded too frequently (cache thrashing). The proprietary mathematics for the ApplicationSmart feature-vector analysis is not presented here, but one should understand how access patterns change the computations and indicators.

"Feature vector analysis mathematics is employed to determine when access patterns have changed significantly."

When the coarse region level histogram changes (checked on a tunable periodic basis) as determined by ApplicationSmart ΔShape, a parameter that indicates the significance of access pattern change, then the fine-binned detail regions may be either re-mapped (to a new LBA address range) when there are significant changes in the coarse region level histogram to update detailed mapping, or when change is less significant this will simply trigger a shape change check on already existing detailed fine-binned histograms. The shape change computation reduces the frequency and amount of computation required to maintain access hot-spot mapping significantly. Only when access patterns change distribution and do so for sustained periods of time will re-computation of detailed mapping occur. The trigger for remapping is tunable through the ΔShape parameters along with thresholds for control of CPU use, to best fit the mapping to access pattern rates of change, and to minimize cache thrashing where blocks replicated to the SSD. The algorithm in ApplicationSmart is much more efficient and scalable than simply keeping 64-bit counters per LBA and allows it to scale to many petabytes of HDD primary storage and terabytes of tier-0 SSD storage in a hybrid system with modest RAM requirements.

"Only when access patterns change distribution and do so for sustained periods of time will re-computation of detailed mapping occur."

Performance

Performance speed-up using ApplicationSmart is estimated by profiling an access pattern and then determining how stable access patterns perform without addition of SSDs to the Atrato V1000. Addition of SLC for write ingest acceleration is always expected to speed-up writes to the maximum theoretical capability of the V1000 since it allows all writes to be as perfectly re-formed as possible with minimal response latency from the SLC ingest SSDs. Read acceleration is ideally expected to be equal to that of a SAID with each 10 SSD expansion unit added as long as sufficient cache-ability exists in the I/O access patterns. This can be measured and speed-up with SSD content replication cache computed (as shown earlier) while customers run real workloads. The ability to double performance using 8 SSDs and one SAID was shown compared to one SAID alone during early testing at Atrato Inc. Speed-ups that double, triple, and quadruple access performance are expected.

"Atrato Inc. has been working with Intel X25-M and Intel® X25-E Solid-State Drives since June of 2008."

SSD Testing at Atrato

Atrato Inc. has been working with Intel X25-M and Intel® X25-E Solid-State Drives since June of 2008 and has tested hybrid RAID sets, drive replacement in the SAID array, and finally decided upon a hybrid tiered storage design using application awareness with the first alpha version demonstrated in October 2008, a beta test program in progress this March, and release planned for the second quarter of 2009.

SSDs Make a Difference

Atrato Inc. has tested SSDs in numerous ways including hybrid RAID sets where an SSD is used as the parity drive in RAID-4, simple SSD VLUNs with user allocation of file system metadata to SSD and file system data to HDD in addition to the five features described in the previous sections. Experimentation showed that the most powerful uses of hybrid SSD and HDD are for ingest/egress FIFOs, read cache based on access profiles, and simple user specification of SSD VLUNs. The Atrato design for ApplicationSmart uses SSDs such that access performance improvement is considerable for ingest, for semi-random read access, and for sequential large

"Experimentation showed that the most powerful uses of hybrid SSD and HDD are for ingest/egress FIFOs, read cache based on access profiles, and simple user specification of SSD VLUNs."

block predictable access. In the case of totally random small transaction I/O that is not cache-able at all, the Atrato design recognizes this with the access profiler and offers users the option to create an SSD VLUN or simply add more SAIDs that provide random access scaling with parallel HDD actuators. Overall, SSDs are used where they make the most difference and users are able to understand exactly the value the SSDs provide in hybrid configurations (access speed-up).

"Atrato Inc. has found the Intel X25-E and Intel X25-M SATA Solid-State Drive integrate well with HDD arrays given the SATA interface."

Conclusions Made about Intel SSDs

Atrato Inc. has found the Intel X25-E and Intel X25-M SATA Solid-State Drive integrate well with HDD arrays given the SATA interface, which has scalability through SAS/SATA controllers and JBOF* (Just a Bunch of Flash*). The Intel SSDs offer additional advantages to Atrato including SMART data for durability and life expectancy monitoring, write ingest protection, and ability to add SSDs as an enhancing feature to the V1000 rather than just as a drive replacement option.

Atrato Inc. plans to offer ApplicationSmart with Intel X25-E and X25-M SATA Solid-State Drives as an upgrade to the V1000 that can be configured by customers according to optimal use of the SSD tier.

"The Intel X25-E SATA Solid-State Drives provide ingest acceleration at lower cost and with greater safety than RAM ingest FIFOs."

Future Atrato Solution Using SSDs

The combination of well managed hybrid SSD+HDD is synergistic and unlocks the extreme IOPs capability of SSD along with the performance and capacity density of the SAID enabled by intelligent block management.

Issues Overcome by Using SSDs

Slow write performance to the Atrato V1000 has been a major issue for applications not well-adapted to RAID and could be solved with a RAM ingest FIFO. However this presents the problem of lost data should a power failure occur before all pending writes can be committed to the backing-store prior to shutdown. The Intel X25-E SATA Solid-State Drives provide ingest acceleration at lower cost and with greater safety than RAM ingest FIFOs. Atrato needed a cost-effective cache solution for the V1000 that could scale to many terabytes and SSDs provide this option whereas RAM does not.

"For customers that need for example 80 terabytes total capacity, the savings with SSD is significant."

Performance Gained by Using Intel SSD

The performance density gains will vary by customer and their total capacity requirements. For customers that need for example 80 terabytes total capacity, the savings with SSD is significant since this means that 3 1RU expansion units can be purchased instead of 3 more 3RU SAIDs and another 240 terabytes of capacity that aren't really needed just to scale performance. This is the best solution for applications that have cache-able workloads, which can be verified with the Atrato ApplicationSmart access profiler.

Future Possibilities Opened Due to Intel SSDs

Future architectures for ApplicationSmart include scaling of SSD JBOFs with SAN attachment using Infiniband or 10G iSCSI such that the location of tier-0 storage and SAID storage can be distributed and scaled on a network in a general fashion giving customers even greater flexibility. The potential for direct integration of SSDs into SAIDs in units of 8 at a time or in a built-in expansion drawer is also being investigated. ApplicationSmart 1.0 is in beta testing now with a planned release for May 2009.

"The potential for direct integration of SSDs into SAIDs in units of 8 at a time or in a built-in expansion drawer is also being investigated."

Conclusion

Using Intel® Solid State Drive (Intel® SSD) for Hybrid Arrays

The Intel X25-E SATA Solid-State Drive provides a cost effective option for hybrid arrays with an SSD-based tier-0. As an example, Atrato has been able to integrate the Intel X25-E SATA Solid-State Drives in the V1000 tier-0 and with the overall virtualization software for the SAID so that performance can be doubled or even quadrupled.

A New Storage and Caching Subsystem

The use of RAM cache for storage I/O is hugely expensive and very difficult to scale given the cost as well as the complexity of scalable memory controllers like FB-DIMM or R-DIMM beyond terabyte scale. Solid state drives are a better match for HDDs, while being an order of magnitude faster for random IOPs and providing the right amount of additional performance for the additional cost, providing for easily justifiable expense to obtain comparable application speed-up.

SSDs for Multiple Embedded Storage Needs

The use of SSDs as drive replacements in embedded applications is inevitable and simple. On the small scale of embedded digital cameras and similar mobile storage devices, SSDs will meet a growing need for high performance, durable, low power direct-attach storage. For larger scale RAID systems, SSDs in hybrid configurations meet ingest, egress, and access cache needs far better than RAM and at much lower cost. Until SSD cost per gigabyte reaches better parity with HDD, which may never happen, hybrid HDD+SSD is here to stay, and many RAID vendors will adopt tiered SSD solutions given the cost/benefit advantage.

"Until SSD cost per gigabyte reaches better parity with HDD, which may never happen, hybrid HDD+SSD is here to stay, and many RAID vendors will adopt tiered SSD solutions given the cost/benefit advantage."

Acknowledgements

Nick Nielsen (Senior Software Engineer), Phillip Clark (Senior Software Engineer), Lars Boehenke (Software Engineer), Louis Morrison (Senior Electrical Design Engineer), and the entire Atrato, Inc. team who have all contributed to the ApplicationSmart software and integration of solid state disks with the V1000 intelligent RAID system.

References

[1] "Systems and Methods for Block-Level Management of Tiered Storage," US Patent Application # 12/364,271, February, 2009.

[2] "Application Awareness Makes Storage More Useful," Neal Leavitt, IEEE Computer Society, July 2008.

[3] "Flash memories: Successes and challenges," S.K. Lai, IBM Journal of Research and Development, Vol. 52, No. 4/5, July/September, 2008.

[4] "Galapagos: Model driven discovery of end-to-end application-storage relationships in distributed systems," K. Magoutis, M. Devarakonda, N. Joukov, N.G. Vogl, IBM Journal of Research and Development, Vol. 52, No. 4/5, July/September, 2008.

[5] "Hierarchical Storage Management in a Distributed VOD System," David W. Brubeck, Lawrence A. Rowe, IEEE MultiMedia, 1996.

[6] "Storage-class memory: The next storage system technology," R.F. Freitas, W.W. Wilcke, IBM Journal of Research and Development, Vol. 52, No. 4/5, July/September, 2008.

[7] "Information valuation for Information Lifecycle Management," Ying Chen, Proceedings of the Second International Conference on Autonomic Computing, September, 2005.

[8] "File classification in self-* storage systems," M. Mesnier, E. Thereska, G.R. Ganger, D. Ellard, Margo Seltzer, Proceedings of the First International Conference on Autonomic Computing, May, 2004.

[9] "Atrato Design for Three Year Zero Maintenance," Sam Siewert, Atrato Inc. White Paper, March 2008.

Author Biographies

Dr. Sam Siewert: Dr. Sam Siewert is the chief technology officer (CTO) of Atrato, Inc. and has worked as a systems and software architect in the aerospace, telecommunications, digital cable, and storage industries. He also teaches as an Adjunct Professor at the University of Colorado at Boulder in the Embedded Systems Certification Program, which he co-founded in 2000. His research interests include high-performance computing and storage systems, digital media, and embedded real-time systems.

Dane Nelson: Dane Nelson is a field applications engineer at Intel Corporation. He has worked in multiple field sales and support roles at Intel for over 9 years and is currently a key embedded products field technical support person for Intel's line of solid state drives.

Copyright