OBJECTIVE ANALYSIS
Semiconductor Market Research

ENTERPRISE RELIABILITY, SOLID STATE SPEED

The SSD is gaining rapid acceptance in storage arrays, thanks to its extremely high performance. Unfortunately, the most widely adopted approach to using them – putting SSDs into a system designed around HDDs – ends up crippling the SSDs' performance while cheating the user of much that the SSDs have to offer. By the same token, many solutions to protect data and make it more reliable and highly available are also designed for slower media, not high performance SSDs. As a result, many enterprises are not only limiting their SSD performance, but also increasing their risks by purchasing SSD solutions that are not optimized for high performance, high availability, or data protection.

Newer systems abandon legacy HDD-based architectures to squeeze more speed out of flash. These systems use new topologies and software to better harness the SSDs' performance and reliability capabilities at a much more reasonable price. One such system is the Kaminario K2, which we will explore in some depth later in this white paper.

For the time being, let's look at the flash phenomenon to see how and why flash SSDs have suddenly burst onto the scene. We will then explore some of the knotty problems flash brings to enterprise storage design.

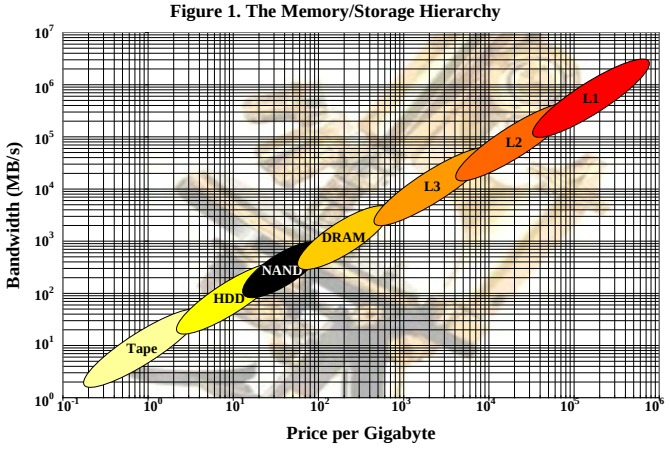
Why SSDs?

Three key trends are driving the enterprise to embrace the use of solid state storage:

1) Data requirements are mushrooming while performance needs are on a similarly steep growth curve
2) Costs cannot be allowed to balloon to match the two elements' growth rates
3) The data center manager must find a way to support data and performance growth while keeping within a relatively static fiscal budget.

Solid state storage allows the system's performance to grow while inexpensive capacity drives can be used for mass storage, thus minimizing the cost of capacity increases. The mix of these two – solid state storage for speed and capacity drives for mass storage – usually yields a significant performance improvement over more common SAN implementations using tiered HDDs in

Figure 1. The Memory/Storage Hierarchy



RAID configurations. In most cases, the performance increase of adding SSDs costs less than existing tiered HDD solutions.

The key reason SSDs can support such diametrically opposed goals is that they offer extraordinary speed for a price that can be justified through savings in other parts of the system.

As shown in Figure 1, the price per gigabyte for an SSD is high in comparison to that of an HDD, but some server applications use large numbers of HDDs at a fraction of their capacity to increase I/O bandwidth. In many cases, a single SSD can provide more speed than a number of HDDs at an adequate capacity for a competitive price.

SSD Reliability Concerns

NAND flash is a messy medium. One means by which chip architects pushed NAND flash costs below those of NOR flash (or any other memory technology, for that matter) was by compromising data integrity. In a move borrowed from the HDD industry, NAND flash stores data in a way that anticipates data corruption, then requires an external controller to scrub the data every time it is read from the device. As technology progresses and NAND prices are driven lower (through shrinking process geometries and the use of multi-bit cells), the level of data corruption grows, requiring ever-increasing levels of error correction.

Fortunately, such error correction coding (ECC) is well understood, and ECC is keeping pace with the degradation of NAND data integrity, offsetting increases in flash error rates.

However, another difficulty adds to this trouble. NAND flash has a wear-out mechanism that is unique to this technology. After a large number of erase/write cycles, bits start to lock up and can no longer be used. This adds to the number of bits that the ECC must correct. As an increasing number of bits become unusable, errors rise to approach the limits of the ECC engine's capabilities. At this point, that particular block must be removed from the pool of available memory.

SSD designers understand this and implement measures to reduce the number of erase/write cycles that the SSD performs. Most SSDs also maintain a set of reserve blocks in the background that can be added to the pool of available blocks when another block needs to be discarded.

Reading the above, some IT experts may start to worry about availability. If an SSD is composed of blocks that may be decommissioned and relies on spares to keep itself in good health, then at some point there will no longer be any spares, and the SSD will fail simply due to overuse. This is a very real problem that has received a lot of attention from SSD designers.

As SSD controller architecture matures, these designers have found ways to increase the number of writes the SSD can absorb, getting past the wear problem for the most part. Today's NAND SSDs are usually specified to be capable of withstanding as many as 30 complete overwrites a day during their warranty period (usually three to five years). Even then, once an SSD has run out of spare blocks and must be removed from operation, it is usually locked into a "Read-Only" mode; the entire contents of the SSD can be read and copied to a new device, but no writes will be accepted. Although this is still a problem, it is less painful than a total loss of the entire contents of the SSD.

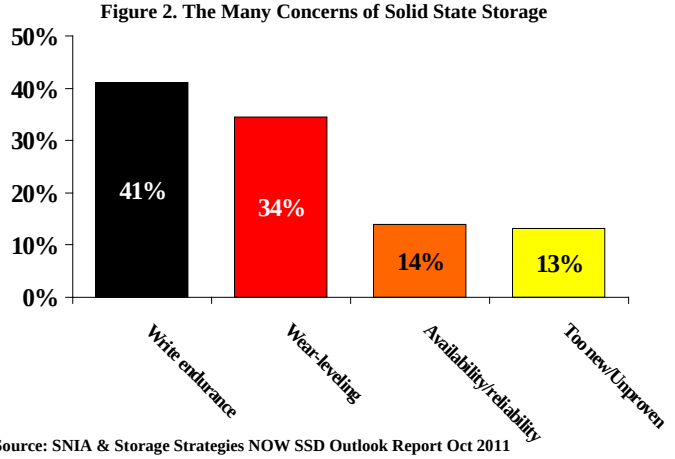
In addition to these wear issues, enterprises require the same standard availability and data protection techniques for their mission-critical data that are re-

quired of disk-based systems. These systems must also take into consideration the unique needs of SSDs in these environments. Solid state media is so fast and can produce so many IOPS that most standard high availability and data protection methods are insufficient and could impact data integrity and system speed.

None of this sits well with users who are concerned about the availability and integrity of their data. For this reason, some IT experts continue to be reluctant to adopt solid state storage.

Figure 2 shows some of the results of a survey performed by the Storage Networking Industry Association (SNIA) and Storage Strategies NOW (SSG-NOW) and detailed in an Outlook report SSG-NOW published in October 2011. SNIA and SSG-NOW surveyed 112 IT managers, asking what concerns they had about adopting SSD technology in their data centers. Endurance and wear were two of the greatest concerns, while availability, reliability, and concerns about the novelty of the approach were frequently cited as other important issues.

Figure 2. The Many Concerns of Solid State Storage



Source: SNIA & Storage Strategies NOW SSD Outlook Report Oct 2011

Reliable Solid State Storage

How can a storage system address these problems, especially with the stringent data protection and availability needs of the enterprise?

There are many well-established approaches to data protection. High availability is one that is commonly practiced – every part of the system is made redundant, and a failover mechanism redirects traffic away from any system element that has failed. Solid state systems require a high availability solution that is fast enough to failover and recover quickly without losing data or significantly impacting performance.

Another such approach is end-to-end data protection, in which data is protected not only through the ECC integrated into SSDs, but also with checksums and verification mechanisms. These verification mechanisms ensure that data will be read back from storage the way it was written in, or it will be written again. The system does not trust that data has been stored until a verification is issued and an acknowledgement is sent to the host.

To guarantee data availability, systems must ensure that data persists even in the event of a hardware failure. RAID (redundant array of inexpensive disks) is a topology that has found favor in storage arrays for its speed and data integrity. HDDs are arrayed in a topology that either stores multiple copies of data in different disks (mirroring), or uses parity to allow any single failed disk to be replaced and rebuilt without ceasing operation. Some RAID systems are striped – data is written to and read from several disks simultaneously to multiply I/O bandwidth.

RAID systems were designed for HDDs and have been optimized over decades of use for an HDD-based environment. As such, they are sometimes ill-suited for an SSD-based environment.

Here is a real-life example: Because of their wear mechanism, it is reasonable to assume that any two identical SSDs that were simultaneously installed into the same system and given similar workloads can be expected to wear out at about the same time. One IT manager we spoke with built a RAID system of SSDs, and after a period of time, one of the SSDs failed. While this manager was performing a rebuild, a second SSD failed, resulting in a system crash that could only be restored through the use of the backup tapes.

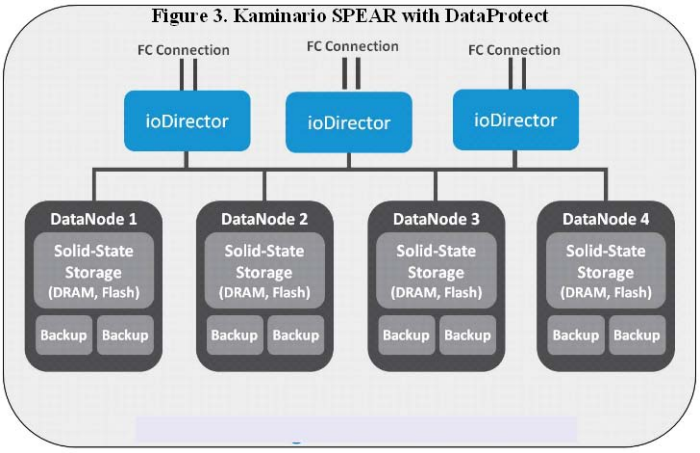


Figure 3. Kaminario SPEAR with DataProtect

RAID is configured for HDDs that fail infrequently and randomly. SSDs fail rarely as well, but fail predictably. Because of this, a standard RAID cannot be used with SSDs unless measures are implemented to prevent such scenarios from evolving.

If RAID needs to be optimized for SSDs, then it only makes sense that other data protection technologies such as snapshots and replication would follow suit.

Kaminario's Solution

Kaminario has created an architecture it calls SPEAR (Scale-Out Performance Storage Architecture) that has been devised to address these problems. The SPEAR SAN design is a clustered solid state storage array whose architecture has been optimized to redundantly store data while reducing the amount of costly high-speed storage in the system – all without impacting performance.

SPEAR also contains a rich storage software stack called DataProtect that provides automated high availability, nondisruptive operations, and extensive data protection. DataProtect reduces flash wear while increasing speed through highly advanced RAID 10, snapshots, and remote replication. DataProtect also manages the N+1 hot-swappable components, allowing them to be replaced for easy maintenance without rebooting while the system is still processing data. A call-home feature alerts system administrators when there is an issue, without disrupting operations.

Figure 3 illustrates a SPEAR system. This system is clustered to increase performance and to avoid any single point of failure.

The system consists of ioDirectors and DataNodes. The ioDirectors manage data onto and off of the media in the DataNodes, sending acknowledgements back to the system when written data has been verified.

Each DataNode consists of two types of storage: faster, costlier solid state storage in the primary storage tier, and less expensive backup storage in the secondary storage tier.

In Kaminario's fastest systems, the solid state storage is DRAM, but most systems will use flash SSDs. The backup storage can either be MLC flash SSDs or standard HDDs. The system is thus very flexible and can be configured to fit the datacenter's needs and budget.

DataProtect uses a storage approach that Kaminario calls RAID 10HD (Hybrid Distributed). RAID 10HD builds on standard RAID 10, using both striping

and mirroring – striping data across the fast solid state portion of multiple DataNodes (DRAM or flash), and mirroring to a less expensive backup storage medium (MLC flash or HDD).

Blocks A, B, C, and D have been added in Figure 4 to illustrate how data is striped across all the available DataNodes in the system (from two to 14). Each data block stored in the fast storage of any one DataNode is mirrored in the backup storage of a different DataNode.

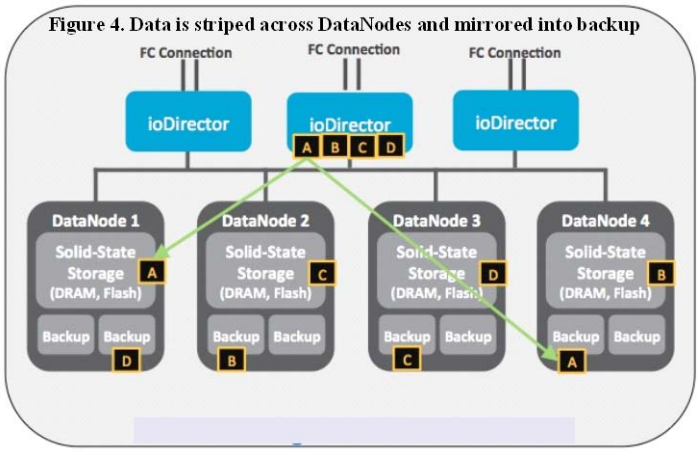


Figure 4. Data is striped across DataNodes and mirrored into backup

This means that each of the data blocks resides in two nodes – a copy in one DataNode's fast storage, and another in a different DataNode's backup.

Should any single DataNode fail, a spare DataNode will automatically begin replicating the contents of the failed DataNode. This occurs in the background as the data is read from or written to the other DataNodes.

To the system, the impact of this is a slowdown for the data that must be read from the backup storage. Kaminario showed us an example system in which a DataNode was suddenly pulled out of the live system. Performance was slowed as the system dynamically reconfigured itself, then it quickly resumed operations at full speed.

Kaminario has also focused attention on improving snapshot performance. Standard HDD-based snapshot approaches perform poorly on SSDs since these approaches were not designed to take advantage of faster media. The system quiesce time during snapshots can be reduced from seconds to milliseconds if the system uses snapshots that are purpose-built for SSDs. This not only harnesses the SSDs' speed to reduce snapshot turnaround time, but it also increases the volume of snapshots that can be taken over any period of time.

Kaminario's "K2 High Performance Snapshots" can be performed every second or less with a system quiesce time of 10 milliseconds. Today's standard SANs typically perform snapshots using a Copy-on-Write approach, in which a copy of the unchanged production data is written to the snapshot before the production volume is changed. Kaminario uses a Redirect-on-Write approach, in which a new storage block is assigned to the production volume and the old data is simply reassigned through a pointer to the snapshot. This significantly reduces write traffic to wear-sensitive SSDs while accelerating throughput, because it reduces the number of I/Os used to create the snapshots. This is enabled, in part, through Kaminario's "thin provisioning" approach to storage, in which blocks are provisioned to production storage rather than entire LUNs.

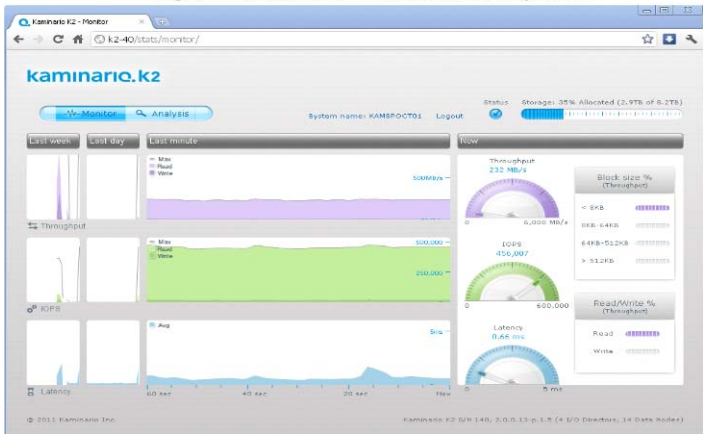
The system can store up to 8,000 snapshots, if necessary, before older snapshots need to be deleted.

Kaminario's remote replication process has been designed to use high performance solid state storage as both the source and the target to help system administrators meet their RPO and RTO goals. Remote replication can be performed as often as every 15 seconds, de-

pending on the size of the data set and the bandwidth of the communication channel to the replica system. Only the changed blocks are communicated to the remote system, accelerating replication and eliminating recovery time.

Real-time analysis of throughput, IOPS, and latency over time is key to maintaining the optimal performance of the K2. This is why Kaminario integrated an easy-to-use Performance Analysis GUI (Figure 5) in the K2 to analyze data traffic patterns and trends. System administrators can track performance parameters over time in a single view.

Figure 5. Kaminario K2 Performance Analysis GUI



Wrap-Up: A Good Solution

The storage community is just beginning to be offered systems that have been designed to take advantage of all of flash's speed. These systems perform significantly better than legacy HDD-based approaches that have been accelerated by replacing HDDs with flash. Although the simple addition of SSDs requires a far smaller engineering effort, more SSDs will be required to net the same performance, and all in all, that makes the system more expensive.

By teaming a smart, SSD-aware architecture with solid data protection, Kaminario has devised a very scalable, reliable, and cost-effective system.

Furthermore, the fact that HDD, flash, and DRAM can be mixed and matched to meet the system's needs gives the K2 great flexibility. Slower systems can be built using inexpensive SSDs for the fast storage and capacity HDDs for backup. The highest-speed systems will use DRAM for the fast storage and enterprise SSDs for backup. Between these two extremes lie myriad possible configurations to provide the right combination of speed and value.

The system assures high data availability and data protection while keeping costs reasonable through its unique architecture.

Jim Handy, February 2012