Benefits of Solid-State Storage

This Dell technical white paper describes the different types of solid-state storage and the benefits of each.

Jeff Armstrong

Gary Kotzur

Rahul Deshmukh

Contents

February 2012| Rev 1.0

Introduction

Traditional storage technology and storage interconnect technology has been evolving over the years, but hard disk drive performance has been restricted due to mechanical movement limitations and disk rotational speed, which has topped out at 15,000 rpm. This has led to a backend storage bottleneck for I/O-intensive and database-type applications.

Solid-state storage (SSS) promises to narrow the gap between server and storage performance. Solid-state storage is a non-volatile storage medium that uses integrated circuits rather than rotating magnetic media for storing information. Solid-state storage devices are available in a number of different form factors. Some of them are:

- **SSD**: solid-state drive
- **SSC**: solid-state card
- **SSM**: solid-state module (memory module)

Device interfaces include the traditional storage interfaces such as SATA, SAS and FC. More recently, SSS have been shipping in card form factors that plug directly into existing PCIe slots without the need for an intervening HBA providing connectivity to the system. Solid-state storage devices can also be made available to the server through traditional external storage networks, such as FC, SAS, or iSCSI.

PCIe-SSS

PCI Express solid state storage (PCIe-SSS) is a type of storage device that connects directly to the server's PCIe bus. Today's PCI-SSS devices are based on the standard PCIe card form factors. Devices typically implement an interface that is based on AHCI. They use a form of solid-state memory devices—NAND Flash—as the underlying storage media. NAND Flash is non-volatile, cost effective, and has low power consumption.

Two NAND types are in use today: single-level cell (SLC) and multi-level cell (MLC). SLC contains one bit per cell and is used for write-intensive applications where high write endurance and read/write performance is required. MLC typically uses two bits per cell and is used for applications that have high capacity requirements, minimal need for high write endurance, and where cost may be a factor.

Differences between SAS or SATA SSDs and PCIe-SSS

A SAS or SATA SSD is similar to a hard disk drive (HDD) in that it utilizes the same host interface, and therefore is accessed by the system through the same OS driver stack. While SAS/SATA solid state storage devices offer significantly better performance benefits over standard rotating-media HDDs, their performance can be limited by the latency and bottlenecks inherent in the driver stack. PCIe-SSS offers greater performance benefits since they are directly attached to the processor via the PCIe bus.

PCIe-SSS devices can be used as either storage caches or very fast storage tiers. As a storage cache, they can be used to cache such things as page files. As a storage tier, they can be used to hold frequently accessed data, often referred to as *hot data*. Close proximity and direct connection to the processor provides an access mechanism with much lower latency, making them especially suitable for these types of uses. New generations of PCIe will provide for even greater bandwidth than what is available today and will go even further in lowering the effective access latency.

Establishing a standard for PCIe-SSS

Form factor

Form factors for PCIe-SSS devices today are PCIe cards, which are standardized as a half-height/half-length (low-profile), full-height/half-length, and full-height/full-length. The ability to service the cards, especially in high-availability (HA) systems, is lacking. This was the motivation to form the SSD Form Factor Working Group[1]. Its charter is based on the promotion of PCIe SSDs by enabling serviceability, high-availability, ease of integration, interoperability, and scalability of SSS. This resulted in the group focusing on three areas: device form factor, connector, and hot-plug behavior.

The group has created two specifications: one detailing the connector and its pin-out, and another specifying a device form factor based on the 2.5" hard disk drive form factor. In addition to the form factor, this specification also defines critical device behaviors in support of the reliability, serviceability, and availability requirements that users demand. The connector specification defines a drive connector that supports backward compatibility with SAS and SATA and enables support for up to four lanes of PCIe. The SSD Form Factor Work Group, now more commonly known as the SSD FF WG, includes the founding companies Dell, EMC, Fujitsu, IBM, Intel, and other leading industry representatives. These companies are working toward specification approval and are driving toward making this the industry standard within the SFF standards body.

Choosing a solution

There are advantages to both types of solutions, depending on your needs and requirements. PCIe-SSD leverages the performance benefits of PCIe-SSS, where applications require ultra-high performance and benefit from the lowest latency and highest IOPs achievable, and support caching and primary storage usage models. The drive form-factor of the PCIe-SSD enables the serviceability that users expect, allowing use in high-availability and mission-critical applications as well as drive-scalability within a server or storage platform. If your solution needs have less-demanding latency requirements, then a SAS or SATA SSD can be used in a traditional host storage or external storage infrastructure utilizing software or hardware RAID to provide data protection.

Not all PCIe-SSS are architected the same, further complicating the decisions a user has to consider when making a purchase or planning for a system configuration. There are two primary device architectures that differ in the way they manage the NAND flash. Unfortunately, for the end user of NAND flash based devices—NAND flash—both SLC and MLC are consumable devices, in that both have a limited number of erase-write cycles available. Manufacturers have implemented numerous mechanisms for extending the write-life of the NAND flash. In fact, more than one mechanism will be implemented within a single device. All of these algorithms fall under a common categorization termed *flash management*. A device can have a fairly simple hardware implementation and have the flash management tasks implemented as a part of the OS device driver stack. Another approach is to have a more sophisticated device hardware implementation that supports the ability to have flash management tasks performed by the device itself. Flash management tasks are not trivial in terms of

---

[1] For more information on the SSD Form Factor Working Group, visit http://www.ssdformfactor.org/aboutus.html.

the amount of processor resources needed to perform the needed functions. Off-loading these tasks to the device itself makes for a much more scalable and higher-performing solution.

Advantages of Dell Express Flash PCIe-SSDs

There are a number of differences and advantages that Dell PowerEdge™ Express Flash PCIe-SSDs offer, compared to PCIe-SSC solutions. The PowerEdge Express Flash PCIe-SSD provides high-performance storage. It is built with enterprise-class SLC NAND flash designed for solutions requiring low latency, high input/output operations per second (IOPs), and enterprise-class storage reliability and serviceability. The device is PCIe Gen2-compliant and can be configured as storage cache or as a primary storage device in demanding enterprise environments, such as enterprise blade and rack servers, video-on-demand servers, web accelerators, and virtualization appliances. The advantages of Express Flash PCIe-SSDs are detailed in the following sections.

High performance

The Dell PowerEdge Express Flash PCIe-SSD enables outstanding IOP performance. It is designed to deliver sequential throughput on reads and writes of up to 1.8/1.2GBps. In addition, the Dell PowerEdge Express Flash PCIe-SSD does not burden the server processor and memory with NAND flash management overhead. Flash management is provided by the device controller running firmware on a high-performance ASIC that resides within the PCIe-SSD. This architecture provides superior IOPs and throughput performance without *borrowing* server resources.

Durability

The PowerEdge Express Flash PCIe-SSD's phenomenal performance extends beyond the hardware. We leverage specialized NAND management technologies to provide robust lifetime wear leveling and data protection. Furthermore, since NAND SSDs have a finite number of program and erase cycles, Dell warrants the PowerEdge Express Flash PCIe-SSD to a maximum amount of data written to the SSD in total bytes written. The SSD monitors these cycles, and Dell software management applications notify you when the warranty limits are reached.

Flexibility

You can apply Express Flash PCIe-SSD technology to a number of applications. Also, Express Flash PCIe-SSDs integrate with servers at the system bus and kernel level, creating a new flash storage cache/tier that can be used as a second-level database buffer cache for frequently read blocks or to store database objects, which requires low latency and high IO rates. Express Flash PCIe-SSDs can be mixed and matched for cache and primary storage capacity in the same system and can share these tasks with traditional HDDs as well.

Questions about PCIe-SSD

The following sections answer some potential questions about PCIe-SSD technology and Dell Express Flash drives.

Can I boot from PCIe-SSD?

Dell's initial release for PCIe-SSD does not include boot capability, but this feature is planned for future implementation.

How does RAID work with PCIe-SSD?

Standard hardware RAID is utilized for server storage or attached storage and provides RAID capabilities for standard storage peripheral interfaces (for example: SAS, SATA, FC, and iSCSI). These hardware controllers are not capable of providing RAID support for PCIe-SSDs, as these devices sit on the PCIe bus. PCIe-SSDs can utilize OS RAID, which is offered through a number of different operating systems, including:

- Microsoft Windows Server® 2008 R2 or later with Hyper-V virtualization (processor: x64/EM64T)
- Red Hat® Enterprise Linux® 6.1 64-bit or later (processor: x64/EM64T)
- SUSE® Linux Enterprise Server 11 SP2 64-bit or later (Processor: x64/EM64T)

Which applications can implement tiering with PCIe-SSDs?

There are a number of applications that can take advantage of Dell 2.5" Express Flash PCIe SSDs, including the following:

- Online transactional processing (OLTP)
- Online analytical processing (OLAP)
- Collaboration (Microsoft SharePoint®)
- Virtualization (VMware®, Hyper-V)

Dell Express Flash 2.5" PCIe-SSDs integrate with servers at the system bus and kernel level, creating a new flash memory tier which can be utilized to store database objects requiring high I/O rates (as well as low latency), or it can be used as a second level database buffer cache.

In OLTP and OLAP environments, you can expect significant improvements in the areas of IOPs, increased number of users, and query response times. Additionally, Express Flash 2.5" PCIe-SSDs are very flexible in how they can be deployed to enhance these database applications. Databases can be accommodated completely with the Express Flash drives, or frequently accessed components of the database can be placed on Express Flash drives.

Why do solid-state devices wear out?

To discuss why a flash-based SSD has a limited write-life, let's look at the mechanics of flash, how it varies across flash types, and what the techniques are used to improve its functionality.

A flash memory device is read and written in pages. A read is relatively straight forward—a read command with the address is issued and the respective data is returned. A write can only occur to those pages that are erased, therefore host write commands invoke flash erase cycles prior to writing to the flash. This write/erase cycling causes cell wear which imposes the limited write-life.

NAND flash devices can be either single-level cell (SLC) or multi-level cell (MLC). SLC only stores one bit of information and requires only two voltage levels to represent a 0 or 1. This is the simplest implementation of NAND and has the highest endurance, which is around 100,000 cycles. As future generations of flash move to smaller geometries, the endurance will be decreased (we are already seeing some that have 50,000 cycles). MLC usually indicates storing two bits of information and requires four voltage levels to represent 00, 01, 10 and 11. The cell wearing is similar between SLC and MLC, but since more voltage levels must be sensed, the endurance levels are significantly reduced. MLC is usually around 10,000 but newer generations are as low as 3,000–5,000 cycles.

A flash SSD is composed of a quantity of flash die to attain the high capacities. To improve the write-life of SSDs, several techniques are used, which can be applied to both NAND types. Host write accesses can occur to any location which can cause hot-spots, which causes premature wear in these locations. A technique called *wear-leveling* is used to prevent the hot-spots. Wear-leveling results in a nearly even distribution of write accesses across the total capacity of the SSD. Write-amplification is a measure of the ratio of the number of flash writes relative to the host write. As an example, if 2 flash writes are generated per host write, the write amplification is also 2. To reduce write-amplification, a technique called over-provisioning improves the garbage collection efficiency, thereby reducing write-amplification. Last, a technique typically applied to MLC uses lower-voltage levels during the write-cycle to ease the cell wear, thereby improving write endurance.

While it would be difficult to determine exactly how long an SSD would last, there are guidelines provided to help estimate. SSDs use a metric that was developed by JEDEC called TBW (terabytes written). While the actual write-life will be impacted by the work-load profile (for example: random or sequential, block size, or write-activity), the TBW provides an estimate, but your actual mileage will vary. To determine the expected life, one would take the TBW and divide by the expected average BW of the writes to the drive. Typically, aside from very demanding applications, SSDs should be expected to last more than three years.

Summary

Choosing a PCIe solid-state device like Dell PowerEdge Express Flash gives considerable performance advantages and is beneficial for latency-sensitive workloads. Express Flash is a standards-based server storage solution that can complement HDD storage or be used as a storage cache or an additional tier of storage, providing significantly higher IOPs and throughput performance. The Dell PowerEdge Express Flash PCIe-SSD 2.5” form factor, coupled with hot-plug capabilities, can provide you with unmatched serviceability and contribute to a new standard of availability.