Benefits of PCI Express Flash

This technical white paper describes different flash technologies and the benefits of PCIeSSD 2.5" drives.

Saptadeep Chanda,

Dell | Enterprise Solutions Group

Contents

Tables

Figures

July 2014 | Version 1.0

Applications are data starved

Virtualization and the explosion of IOPs (input/output per second)-intensive applications for SAP, Oracle®, Microsoft® SQL Server®, big data analytics, virtual desktop infrastructure (VDI), and many others drive near real-time analysis and an appetite for performance that has to be fed by flash-based solid state devices (SSDs). Subsequently, the industry is seeing a growing number of organizations deploying flash-based solid-state storage solutions as these drives provide faster access to data, resulting in an increase in performance and lower latency, and because demand for these devices is changing the economics of flash storage. Flash performance at the cost of disks is a real consideration and organizations no longer have the cost impediment consideration for implementing flash-based storage to improve the performance of their applications.

These SSDs come in three different interfaces —SAS, SATA and PCI Express® (PCIe) SSD. While SATA and SAS SSDs, which can provide 8x more IOPs than hard disk drives (HDDs), have provided significant improvements in I/O performance and latency over traditional rotating media storage, their performance is limited by the interface and the latency inherent with the older SCSI-based stack. PCIeSSDs which can provide 10.5x more IOPs than 16 HDDs are directly connected to the PCI bus, which provides a direct connection to the CPU and memory complex. This tightly coupled connection enables the PCIeSSD with a significant performance and latency advantage over SAS and SATA SSDs. Performance testing has shown the PCIeSSDs to have considerably more IOPs capabilities and significantly higher number of database transactions per second over traditional SAS and SATA SSDs.

Dell, a front runner in leading industry-wide standards and in driving technology innovations, was a founding member of the SSD Form Factor Working Group (SSD FF WG), driving the specification and standards for the 2.5-inch PCIeSSD form factor and also of the non-volatile memory express (NVMe) consortium helping drive a standardized high-performance host controller interface for PCIeSSDs designed for current and next-generation non-volatile memory. Dell Inc. first launched Express Flash PCI SSDs in May 2012 on its 12th-generation PowerEdge servers. Initial products were based on single-level cell (SLC) non-volatile NAND storage (Appendix A). These PCIeSSD devices came in a 2.5-inch form factor and were front-loadable, hot-swappable, ultra-low latency devices. Furthermore, Dell PowerEdge server R920 is the first in the industry to ship with NVMe PCIeSSD based on multi-level cell (MLC) technologies. [1]Dell helped drive the standards for the non-volatile memory express specification which standardizes the PCIeSSD protocol and will drive industry adoption. By providing standards-based PCIeSSD technology and enabling high availability, serviceability, ease of integration and enhanced scalability, the end customer will have a wider choice of options, can limit expensive customization and can plan for the future more efficiently.

Storage form factor choices

Below is a quick overview of the different types of storage devices and form factors that we will talk about in this paper:

1. **Hard disk drives (HDDs)** —rotating media storage with lower costs and high capacity, but high latency and low performance. HDDs can be 10K and 15K SAS drives, which are also known as Tier-1, or 5.4K and 7.2K rpm SATA drives, which are also known as Tier-2.

1 http://www.samsung.com/global/business/semiconductor/minisite/Greenmemory/news-event/press-release/detail?contentsSeq=13301&contsClassCd=N

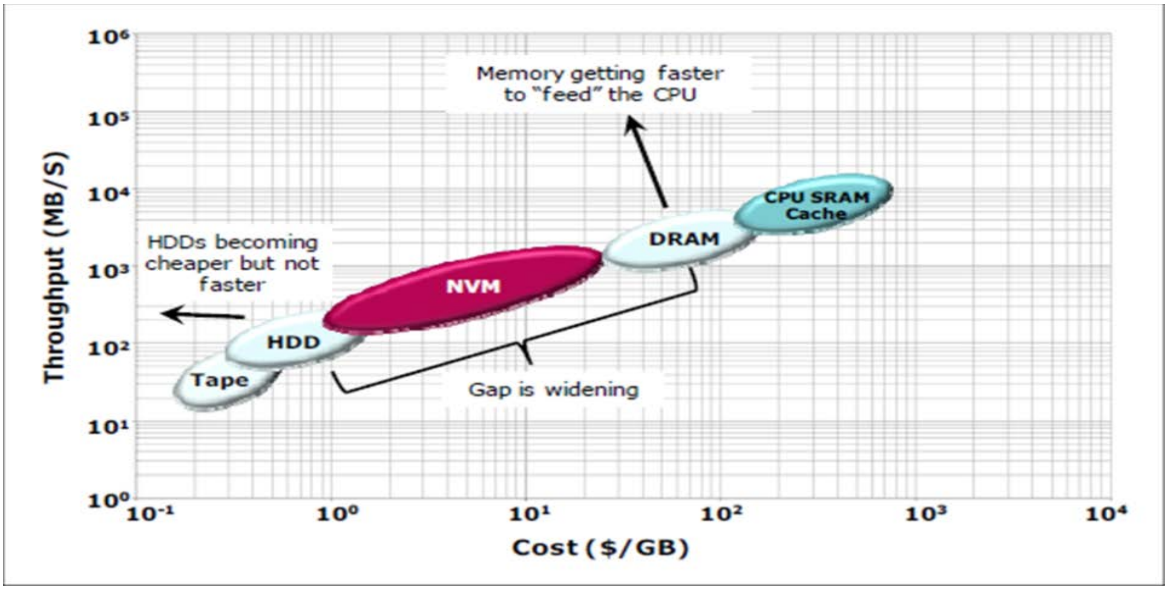
2. **Solid state drives (SDDs)** — unlike rotating media storage, these use integrated circuit assemblies as memory to store data persistently. SSDs are designed around enterprise application I/O (input/output) requirements with the primary attribute focus being on random I/O performance and reliability. SSDs could be a SAS, SATA or PCIeSSD.
3. **SATA SSD** — these devices, unlike SAS SSDs (or HDDs), use the serial ATA (SATA) interface for connecting to the server. These can reach up to 6Gbps.
4. **SAS SSD (or HDDs)** — these are SAS-based SSDs (or HDDs), meaning that the devices use a serial-attached SCSI interface for connecting to the server. The bandwidth of SAS SSDs can reach up to 12Gbps.
5. **2.5-inch PCIeSSD** — these are 2.5-inch SSD devices that run on PCI lanes and faster than the SAS or SATA interfaces, delivering up to 32Gbps.
6. **PCIe solid state card (SSC)** — These are solid state card form factors and the interface protocol is not an industry standard but proprietary. These form of PCIe storage devices have higher capacity but lack the RAS features provided by 2.5" PCIeSSD form factor.

Why PCIeSSD flash storage?

Performance and low latency are paramount to companies running Oracle, SQL, virtual desktop infrastructure (VDI), and other online transaction processing (OLTP)-related database applications, which often demand high IOPs capabilities from the network-based storage subsystem. These workloads need to deliver data faster to applications, and due to the physical constraints of the rotational storage, latency is a reality when dealing with these demanding database applications. PCIeSSD storage can dramatically increase data throughput depending on the nature of the workload (Read/Write along with sequential vs. random data reads and writes.)

Figure 1 Need for NVM (NVMe 2013) illustrates the widening gap in price/performance between DRAM and HDDs on the server. This means that while capacity is increasing and cost is decreasing for an HDD, performance is not keeping pace. While bandwidth and speed of memory is increasing, the cost of DRAM memory remains high. Slow rotational storage devices are not able to keep up with the continually improving processor (CPUs), and memory (DRAM) is extremely costly to provide the capacity needed to deliver the requisite performance. Flash storage (NAND storage devices) and especially PCIeSSDs fill in this gap beautifully, creating an "I/O memory tier" in the system. These devices are much cheaper than memory (DRAM) and are able to deliver significant improvement in performance over traditional HDDs or SAS/SATA SSDs.

Figure 1 Need for NVM (NVMe 2013)



PCIeSSDs are best suited for workloads that need tremendous amounts of IOPs as well as those that need high-bandwidth performance. In industry tests, PCIeSSDs have provided more number of IOPs over traditional SAS/SATA SSDs and also outperformed other forms of SSDs, providing the best cost/IOPs to the user (Appendix B).

The architecture for PCI SSDs utilizes dedicated PCIe lanes. The following four factors strongly contribute to lowering latency and enhancing application performance.

- PCI lanes have no host bus adapter (HBA) overhead resulting in very low latency
- PCI lanes have scalable link speed (0.25/0.5/1 GBps)
- PCI lanes have scalable port width (x1 to x16)
- PCI lanes are full duplex, can handle multiple Queues requests, Out Of Order processing

Overall, PCIeSSDs benefit from not only bypassing legacy interfaces such as SAS and SATA and altogether eliminating latency when retrieving data via a PCI data path, but also by their ability to directly attach to the chipset or CPU, thereby eliminating the need to use an external HBA. This configuration saves 7–10 watts of power and overhead costs of maintaining an HBA. Hence, to get more performance for applications running on the server, this approach places the flash drives directly on the high-speed PCI bus which enables the server to leverage the PCIeSSD flash storage as an extension of the server memory cache.

Table 1 Comparison of different form factor (Akber Kazmi, 2013)shows while PCIeSSDs 3.0, the one that is used for the PowerEdge R920 and other 13th-generation PowerEdge servers, can go up to 4GBps, SAS can only go up to 1.5GBps and SATAe* 1GBps.

Table 1 Comparison of different form factor (Akber Kazmi, 2013)

	12G PowerEdge Servers (PCIe 2.0)	PowerEdge R920/PowerEdge 13G servers (PCIe 3.0)
PCIe (x4)*	20Gbps	32Gbps
SAS	6Gbps	12Gbps
SATA	6Gbps	N/A
SATAe**	N/A	8Gbps

* R920 and 13G servers are coming with PCIe 3.0 x1 (number of lanes)
**SATA express that is compatible with both SATA and PCIe.

As we can see from Table 1 Comparison of different form factor (Akber Kazmi, 2013)above, the data bandwidth of the PCIeSSDs has almost doubled from the previous generation. These performance gains were made through various architecture and protocol management improvements. There are two usage categories where PCIeSSDs have proven benefits including:

1. **Caching:** Data could be cached at the compute layer using caching software to identify data that is in high demand and hold it in cache reducing fetch times, latency and increasing application performance. Dell Fluid Cache for SAN is an example of caching solution offered by Dell that can improve application performance.
2. **Application acceleration/performance:** Flash drives can be used to accelerate log file writes, where traditional drives can take a significant amount of time due to their physical constraints.

PCIeSSDs are optimized to provide fast cache at an affordable price, and could be specifically used for most of the challenges faced by customer that want to accelerate their applications while keeping other factors such as cost and power management in check (Appendix B).

What usage cases and workloads benefit from PCIeSSDs?

PCIeSSD delivers the right amount of application acceleration with SSDs that can be used for storage caching or as a fast storage tier. These drives can easily fit in almost any server environment with minimal additional overhead (front-loadable, hot-swappable disks), and provide a lower latency of data. Customers big on OLTP and OLAP will find these benefits extremely useful for their businesses. Among many others, this will benefit customers in the following three ways: Reducing average response time, increasing transactions per second, and increasing number of concurrent users. PCIeSSDs are best suited for applications that require random read and writes requiring high performance and optimal cost per IOPs (Appendix B). However, if the application demands a greater number of sequential reads and writes, caching makes little sense, hence, PCIeSSDs are not the most optimum solution.

Broadly, the three generic use cases of PCIeSSDs are:

1. **All:** All data on PCIeSSD —This could be used for the application that requires faster response times and exceptional IOPs. This is a less costly option compared to storing all data on DRAM.

2. **Hybrid:** Partial data on PCIeSSD — This is used for databases and applications that require quick access to the temp files or log files to accelerate database applications. Hot files, indexes, metadata can all be placed in very fast PCIeSSD to speed up the response times and help accelerate database transactions.
3. **Caching:** SAN/DAS data on PCIeSSD —This is used as a cache medium where the most frequently accessed data is kept right next to the server to accelerate IOPs and average response times per transactions. Fluid Cache for SAN, which is Dell's own branded solution, uses this use case to accelerate OLTP and OLAP based transactions.

Within the generic uses cases described above, a few of the specific scenarios where Dell PCI express flash storage could be used are:

1. Need I/O performance, low latency and availability of placing flash close to application.
2. Need to resolve bottlenecks caused by the difference in the performances of storage and Central Processor
3. Applications and workloads are local.
4. One to one relationship between server and application.
5. Need to add high speed cache to an HDD tier.
6. Need accelerated performance to mission critical applications, while maintaining data integrity.
7. Examples of workloads: OS or hypervisor boot; hot data caching; database acceleration; performance acceleration etc.

Table 2 Drive recommendations for applications and optimal I/O profiles below summarizes select applications along with their I/O profiles along with a recommended drive type for specific workloads and applications. As we can see, SSD drives are ideal for workloads that are highly random and the payload sizes are 4 to 8KB, and all these applications could potentially use PCI bus lanes to accelerate or cache the data for faster access and higher performance.

Table 2 Drive recommendations for applications and optimal I/O profiles

Application	Payload size (in Bytes)	Read/Write percentage mix	Random/Sequential percentage mix	Recommended drive type
Web file server	4K, 8KB	95/5	75/25	SSD
Database, Online Transaction Processing (OLTP)	8KB	70/30	100/0	SSD
Email – Microsoft Exchange	4KB	70/30	100/0	SSD
Decision Support System (DSS)	1MB	100/0	100/0	SSD
File server	8KB	90/10	75/25	SSD
Microsoft SQL Server logging	64KB	0/100	0/100	HDD/SSD*
Web server logging	64KB	0/100	0/100	HDD/SSD*
Media streaming	64KB	95/5	0/100	HDD/SSD*

*SSD's are an option depending on the performance requirements of the database.

PCIeSSD with NVMe interface protects data

PCIeSSD flash storage devices are designed with non-volatile NAND memory, which means the data in the memory of the drive is persistent and not lost in the event of an abrupt power outage. However, adoption of these devices has been inhibited by different proprietary implementations and unique drivers. Without standard drivers, and a consistent feature set, Computer OEMs and customers were challenged by the need to validate custom products from each SSD vendor. To remove inconsistencies and define an industry-standard interface for non-volatile memory (NVM), a consortium of 94+ companies introduced a standard specification for NVM, called NVM Express (NVMe). NVMe delivers the full potential of non-volatile memory and addresses the needs of enterprise, data center and client systems with standards-based protocols and interface for PCIeSSDs. This standard was introduced to maximize performance while significantly reducing the latency of IO. At launch, NVMe showcased more than a 50% improvement Table 3 Latency Comparison (Cobb & Huffman, 2013)of latency over SCSI/SAS devices and this was achieved by an optimization of the stack to process commands.

Table 3 Latency Comparison (Cobb & Huffman, 2013)

	Time per 2000 CPU cycles	Number of cycles
SCSI/SAS	6.0 µs	19500
NVMe	2.8 µs	9100

Benefits of the NVMe interface include:

1. Performance across multiple cores to quickly access critical data
2. An optimized register interface and command set that uses the minimum number of CPU clocks per I/O for higher performance and lower power
3. Scalability with headroom for current and future NVM performance
4. End-to-end data protection capabilities and support for standard security protocols, such as Trusted Computing Group
5. Lower power consumption resulting in a lower total cost of ownership and carbon footprint

NVMe standardization drives simplification

Dell drives standards, standards drives adoption, adoption drives volume, and volume drives cost. Following are the three most important strategic benefits of NVMe standardization efforts.

Standard drivers —Eliminates the need for OEM to qualify a driver for each SSD vendor and enables broad adoption across a wide range of industry-standard and proprietary operating systems.

Consistent feature set — All SSD implementations are required to have baseline features and optional features are implemented in a consistent manner.

Industry ecosystem — Standardized development tools such as analyzers, emulators and test platforms.

Conclusion: Benefits of PCIeSSD

1. Dell PowerEdge Express Flash NVMe PCIeSSDs have twice the performance of previous generation PCIeSSD devices
2. Number of IOPs has almost doubled from that from previous generations (Appendix C)
3. Front-access, hot-plug 2.5-inch PCIeSSD devices with the ability to "hot-add" additional devices and avoid the need to insert cards into PCIe slots that require taking the server offline
4. Device-based flash management, reducing the overhead costs (I.e., costs associated with HBA)
5. Supports a wide range of applications including OLTP, OLAP, collaborative environments, and virtualization where there is random access to data versus sequential for reads and writes
6. Easy fit in almost any server environment with minimal additional cost that could be associated with bringing the server down when replacing PCIe card form factors
7. Deliver the right amount of application acceleration with SSDs that can be used for storage caching and caching software, including Fluid Cache for SAN, Fusion IO, and direct-attach storage (DAS). A second usage includes as a fast storage tier used in conjunction with software-defined storage that includes Microsoft Windows® Storage Spaces, OpenStack™ Ceph, VMware® and others
8. Lower latency of data

Appendix A – Technologies used in the PCIeSSD

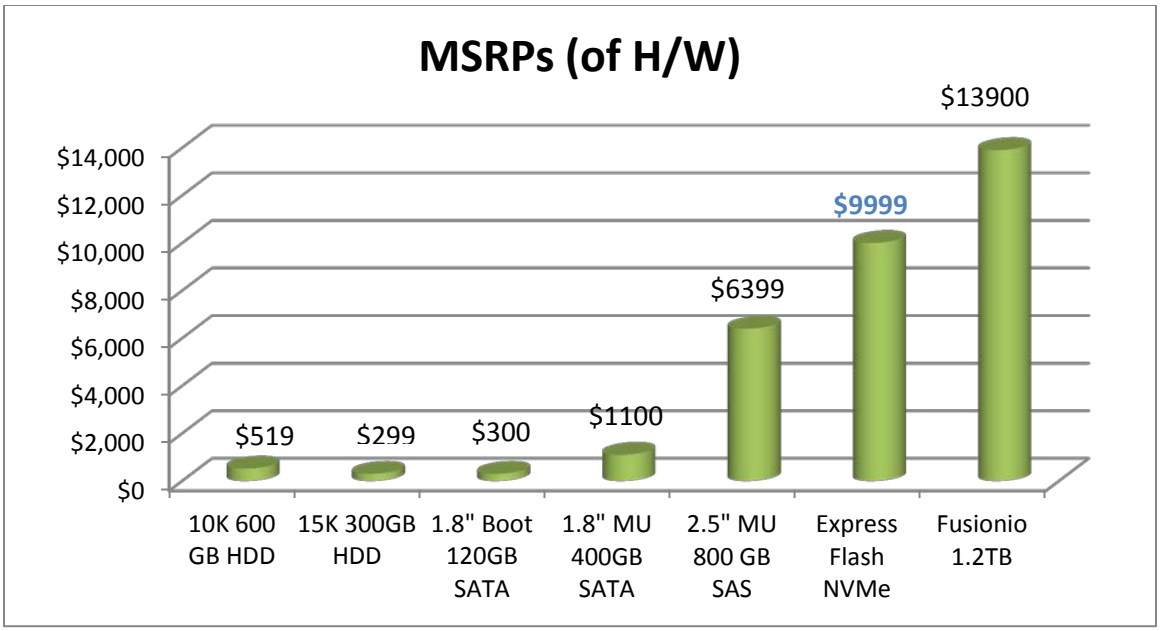
Flash storage fundamentally uses a different medium, such as silicon or electronic NAND gates with very different performance, cost and data-retention characteristics that may impact the economics and operations of application workloads in the data center. Today, there are three common types of enterprise flash drives in the market based on write endurance. Flash drive selection for each of the application workloads may impact the data center economics and operations.

- **Read intensive swim lanes**: This use case is generally for applications or workloads with heavy read IO profiles. Drives filling this requirement typically have Write Endurance limits of 1-to-3 drive fills per day. They are also the least expensive of the enterprise class flash drives. Dell recommends this for 90/10read/write ratio.
- **Write intensive swim lanes:** This use case is generally for applications or workloads with heavy write IO profiles. Write Endurance of these flash drives is typically 20-to-30 drive fills per day. The increased write endurance comes at a cost; making this the most expensive and most reliable enterprise flash drive. Dell will not ship any SLC based drives on its products. However, the demand is shifting towards more of read intensive and mixed use cases rather than write intensive use cases.
- **Mixed Use swim lanes**: This use case drives attempt to strike a balance between write and read intensive drives both from a Write Endurance and cost perspective. Typical Write Endurance for these drives is somewhere between 5-to-10 drive fills per day

Appendix B - Performances

The maximum standard retail price is the lowest for HDD and highest for Fusion IO 1.2 TB PCIe card form factor. Figure 2 MSRP of Hardware shows that prices go up as we move from traditional rotating media, SATA, SSD to a PCIeSSD 2.5-inch form factor followed by the card form factor.

Figure 2 MSRP of Hardware



Furthermore, performances go up to almost proportional with price, but not at the same rate. There is a huge spike in performance from 2.5-inch SAS to a 2.5-inch PCIeSSD with NVME interface and it dips down for the PCIe card form factor. Figure 3 IOPs further illustrates that the performances of PCIeSSDs and PCIe card form factor differ by a huge margin.

Figure 3 IOPs

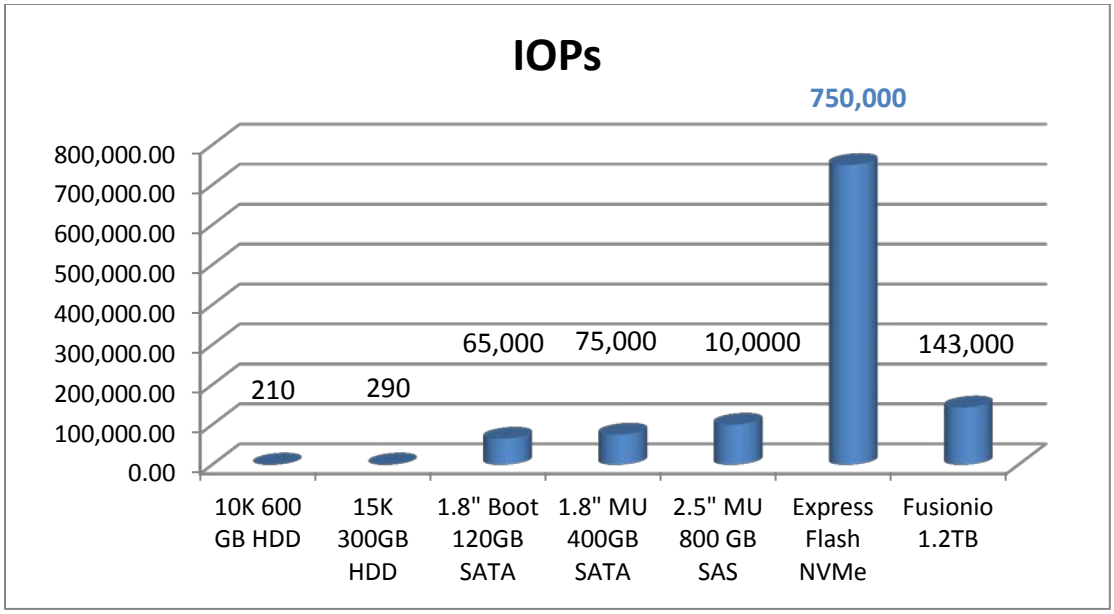
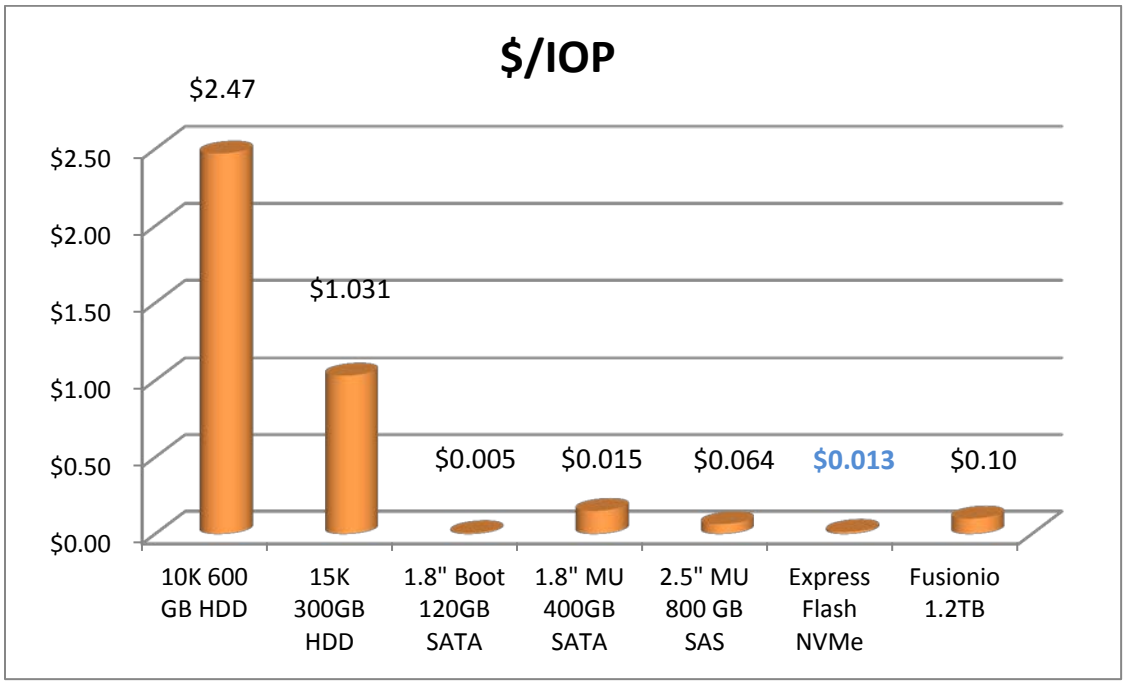


Figure 4 $/IOP illustrates that the best price per IOP is obtained through PCIeSSD with NVMe interface (1.8" 120GB Boot is primarily used for booting purposes and cannot be used as a caching media). These three graphs empahsize the fact that even though the cost might be a little higher but the best value proposition is provided by the 2.5" PCIeSSD form factors.

Figure 4 $/IOP



Appendix C – Flash Gen 2 X 4 & Gen 3 X 4 – IOP – 4K RR/RW

SNIA.org Benchmark		Dell-2.5" 350GB SLC	Dell Express flash NVMe 1600GB MLC	Dell Express Flash NVMe 1600GB MLC
		PCIe Gen2 x4 12G Servers	PCIe Gen2 x4 12G Servers	**PCIe Gen3 x4 R920 & 13G Servers**
Performance	Seq. Read – 1M QD = 16	1.38 GB/s	1.7 GB/s	3.0 GB/s
	Seq. Write – 1M QD = 16	1200 MB/s	1350 MB/s	1350 MB/s
	RR - 4KB QD =256	**420,000 IOPs**	**420,000 IOPs**	**750,000 IOPs**
	RW - 4KB QD = 256	210,000 IOPs	110,000 IOPs	110,000 IOPs
	Endurance – Total Bytes Written (TBW)	350GB = 25PB	400GB = 5PB 800GB = 10PB 1600GB = 20PB	400GB = 5PB 800GB = 10PB 1600GB = 20PB
	Warranty	TBW or 5 years, whichever comes 1st	TBW or 5 years, whichever comes 1st	TBW or 5 years, whichever comes 1st
	EOL Data Retention	3 months	3 months	3 months
	Avg. Latency RR - 4KB QD = 1	70us Micro-seconds	90us	90us
	Avg. Latency RW - 4KB QD = 1	290us	40us	40us