Flash Memory Summit

The Performance Story: An Independent Evaluation of Flash Storage

Dennis Martin, President

Demartek®

Flash Memory Summit

Agenda

- About Demartek
- Enterprise Datacenter Environments
- Storage Performance Metrics
- Synthetic vs. Real-world workloads
- Performance Results – Various Flash Solutions

(new since last year's Flash Memory Summit presentation)

Some of the images in this presentation are clickable links to web pages or videos ➔ 💻

Flash Memory Summit

About Demartek

Click to view this one minute video
(available in 720p and 1080p)

http://www.demartek.com/Demartek_Video_Library.html

Flash Memory Summit

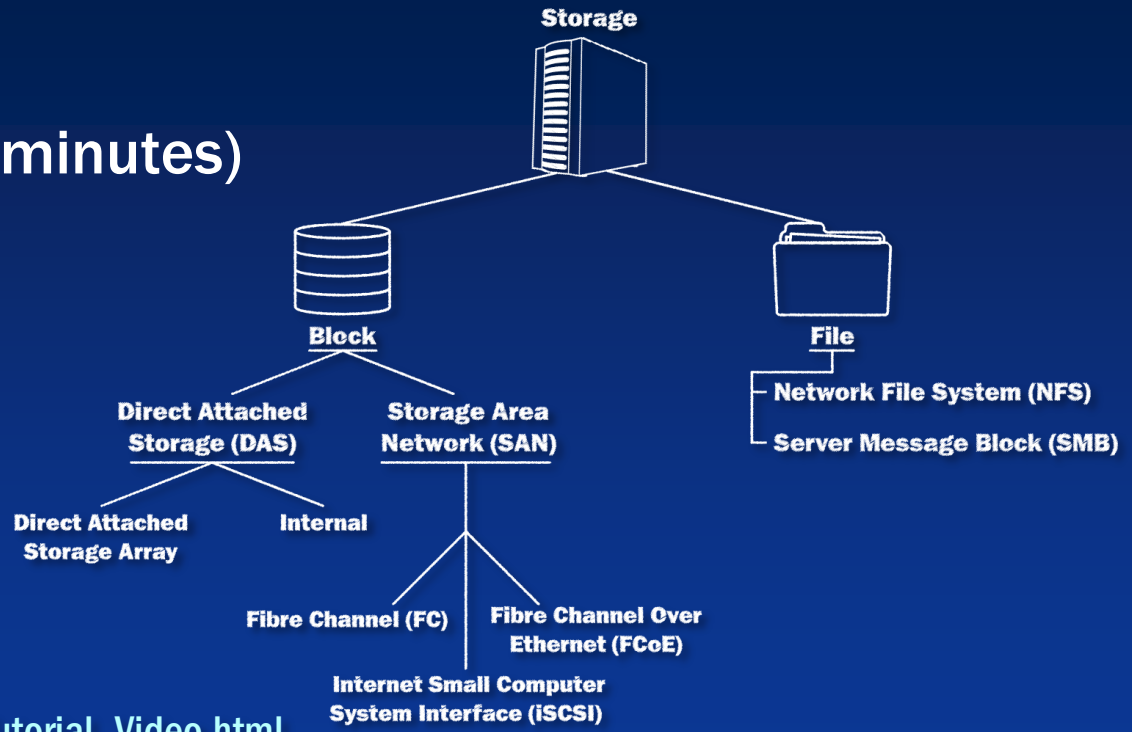
About Demartek

- Industry Analysis and ISO 17025 accredited test lab
- Lab includes enterprise servers, networking & storage
  - 6/12 Gb SAS, 10/25/40/100 GbE, 16/32 GFC, NVMe over Fabrics
- We prefer to run real-world applications to test servers, storage and HCI solutions (databases, Hadoop, etc.)
- Demartek is an EPA-recognized test lab for *ENERGY STAR Data Center Storage* testing
- Website: www.demartek.com/TestLab

SNIA Emerald™ RECOGNIZED TESTER

Demartek Tutorial Videos

- Short videos (3 – 4 minutes)
- Storage Basics



http://www.demartek.com/Demartek_Tutorial_Video.html

Enterprise Datacenter Environments

- Typically support a large number of users and are responsible for many business applications
- Often have specialists for applications, operating environments, networking and storage systems
- Have a large amount of equipment including servers, networking and storage gear
- Multiple types and generations within each category
- Reliability, Availability and Serviceability (RAS)
- Complex systems working together

Enterprise Storage Architectures

▸ Flash Can Be Deployed In Any of These

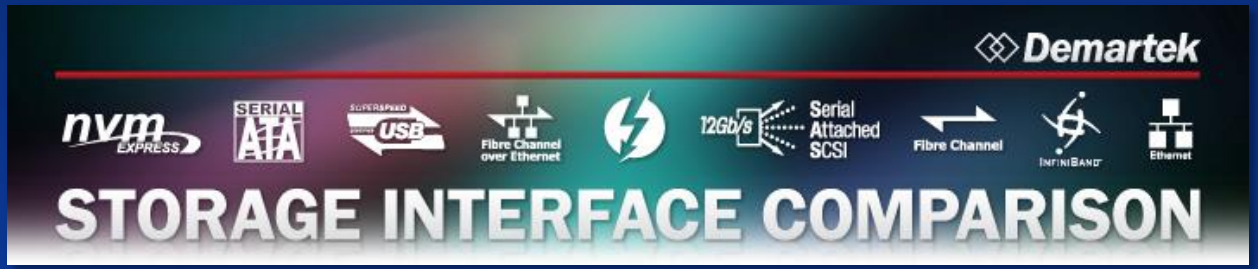
- Direct Attach Storage (DAS)
  - Storage controlled by a single server: inside the server or directly connected to the server ("server-side")
  - ***Block*** storage devices
- Network Attached Storage (NAS)
  - File server that sends/receives ***files*** from network clients
- Storage Area Network (SAN)
  - Delivers shared ***block*** storage over a storage network fabric

Interface vs. Storage Device Speeds

- *Interface* speeds are generally measured in bits per second, such as megabits per second (Mbps) or gigabits per second (Gbps).
  - Lowercase “b”
  - Applies to Ethernet, Fibre Channel, SAS, SATA, etc.
- *Storage device* and system speeds are generally measured in bytes per second, such as megabytes per second (MBps) or gigabytes per second (GBps).
  - Uppercase “B”
  - Applies to storage devices (SSDs, HDDs) and PCIe, NVMe

Storage Interface Comparison

- Demartek Storage Interface Comparison reference page
  - Search engine: *Storage Interface Comparison*
  - July 2017 updates for PCIe 4.0, PCIe 5.0, SFP28, QSFP28, OM5
  - More roadmap and other updates planned for August 2017 edition



http://www.demartek.com/Demartek_Interface_Comparison.html

Storage Performance Metrics

Storage Performance Metrics

▸ IOPS & Throughput

- IOPS
  - Number of Input/Output (I/O) requests per second
- Throughput
  - Measure of bytes transferred per second (MBps or GBps)
  - Sometimes also referred to as “Bandwidth”
- Read and Write metrics are often reported separately

Storage Performance Metrics

▸ Latency

- Latency
  - Response time or round-trip time, generally measured in milliseconds (ms) or microseconds (µs)
  - Sometimes measured as seconds per transfer
  - Time is the numerator, therefore lower latency is faster
- Latency is becoming an increasingly important metric for many real-world applications
- Flash storage provides much lower latency than hard disk or tape technologies, frequently < 1 ms (*workload dependent*)

I/O Request Characteristics

▸ Block size

- Block size is the size of each individual I/O request
  - Minimum block size for flash devices is 4096 bytes (4KB)
  - Minimum block size for HDDs is 512 bytes
    - Newer HDDs have native 4KB sector size ("Advanced Format")
  - Maximum block size can be multiple megabytes
- Block sizes are frequently powers of 2
  - Common: 512B, 1KB, 2KB, 4KB, 8KB, 16KB, 32KB, 64KB, 128KB, 256KB, 512KB, 1MB, 2MB, 4MB

I/O Request Characteristics

▸ Queue Depth

- Queue Depth is the number of outstanding I/O requests awaiting completion
  - Applications can issue multiple I/O requests at the same time to the same or different storage devices
- Queue Depths can get temporarily large if
  - The storage device is overwhelmed with requests
  - There is a bottleneck between the host CPU and the storage device
- Some interfaces have a single I/O queue, others have multiple

I/O Request Characteristics

▸ Access Patterns: Random vs. Sequential

- Access patterns refers to the pattern of specific locations or addresses (logical block addresses) on a storage device for which I/O requests are made
  - Random – addresses are in no apparent order (from the storage device viewpoint)
  - Sequential – addresses start at one location and access several immediately adjacent addresses in ascending order or sequence
- For HDDs, there is a significant performance difference between random and sequential I/O

I/O Request Characteristics

▸ Read/Write Mix

- The read/write mix refers to the percentage of I/O requests that are read vs. write
  - Flash storage devices are relatively more sensitive to the read/write mix than HDDs due to the physics of NAND flash writes
  - The read/write mix percentage varies over time and with different workloads

I/O Request Characteristics

▸ Full Duplex and Half Duplex

- Full Duplex
  - Traffic flows in both directions at the same time (between server and storage), for example: reading and writing simultaneously
  - Total speed is the sum of the speeds in each direction
- Half Duplex
  - Traffic flows in only one direction at a time between server and storage, for example: reading or writing separately
  - Total speed is the speed in one direction only

Synthetic vs. Real-world Workloads

Synthetic Workloads

▸ Purpose

- Synthetic workload generators allow precise control of I/O requests with respect to:
  - Read/write mix, block size, random vs. sequential & queue depth
- These tools are used to generate the *"hero numbers"*
  - 4KB 100% random read, 4KB 100% random write, etc.
  - 256KB 100% sequential read, 256KB 100% sequential write, etc.
- Manufacturers advertise the hero numbers to show the top-end performance in the corner cases
  - Demartek also sometimes runs these tests

Synthetic Workloads

▸ Examples

- Several synthetic I/O workload tools:
  - Diskspd, fio, IOmeter, IOzone, SQLIO, Vdbench, others
- Some of these tools have compression, data de-duplication and other data pattern options
- Demartek has a reference page showing the data patterns written by some of these tools
  - http://www.demartek.com/Demartek_Benchmark_Output_File_Formats.html

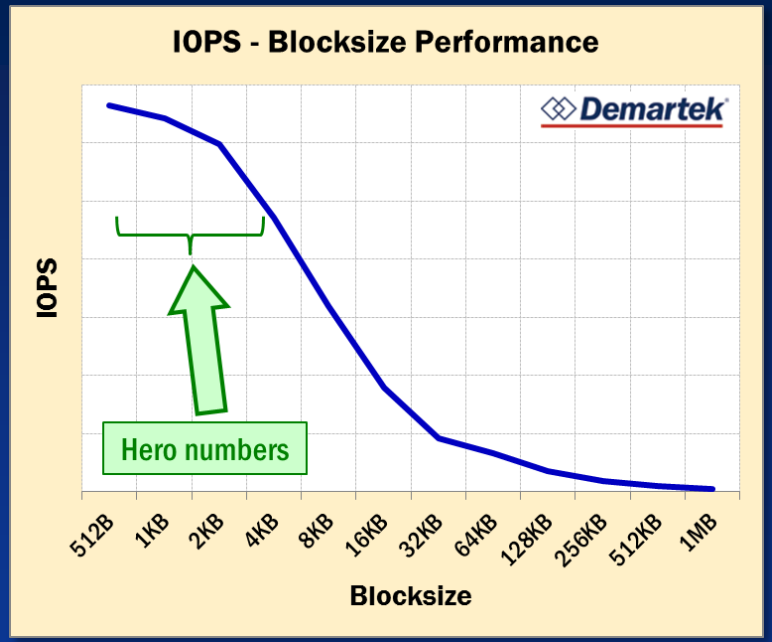
Real-world Workloads

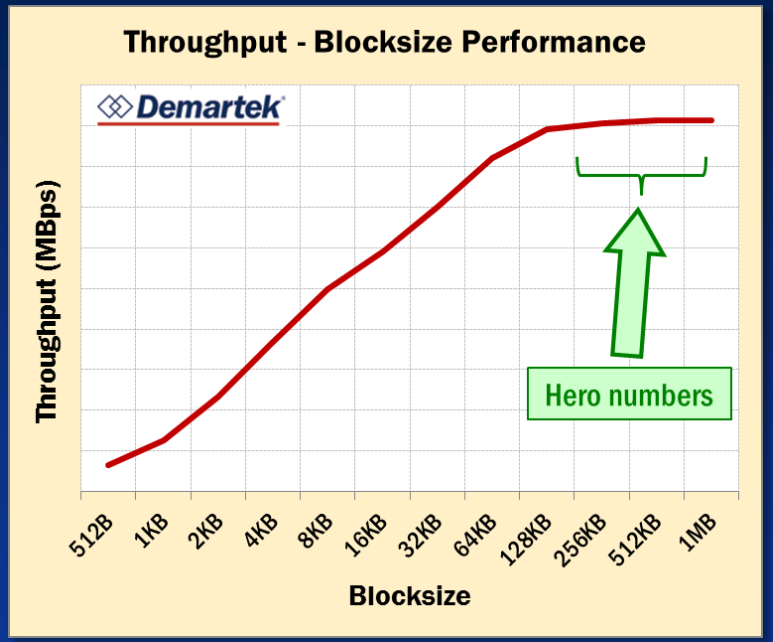
- Use variable levels of compute, memory and I/O resources as the work progresses
  - May use different and multiple I/O characteristics simultaneously for I/O requests (block sizes, queue depths, read/write mix and random/sequential mix)
- Many applications capture their own metrics such as database transactions per second, etc.
- Operating systems can track physical and logical I/O metrics
- *End-user customers have these applications*

Real-world Workload Types

- Transactional (mostly random)
  - Generally smaller block sizes (4KB, 8KB, 16KB, etc.)
  - Emphasis on the number of I/O's per second (IOPS)
- Streaming (mostly sequential)
  - Generally larger block sizes (64KB, 256KB, 1MB, etc.)
  - Emphasis on throughput (bandwidth) measured in Megabytes per second (MBps)
- *Latency is affected differently by different workload types*

Generic IOPS and Throughput Results





These performance curves generally apply to network and storage performance

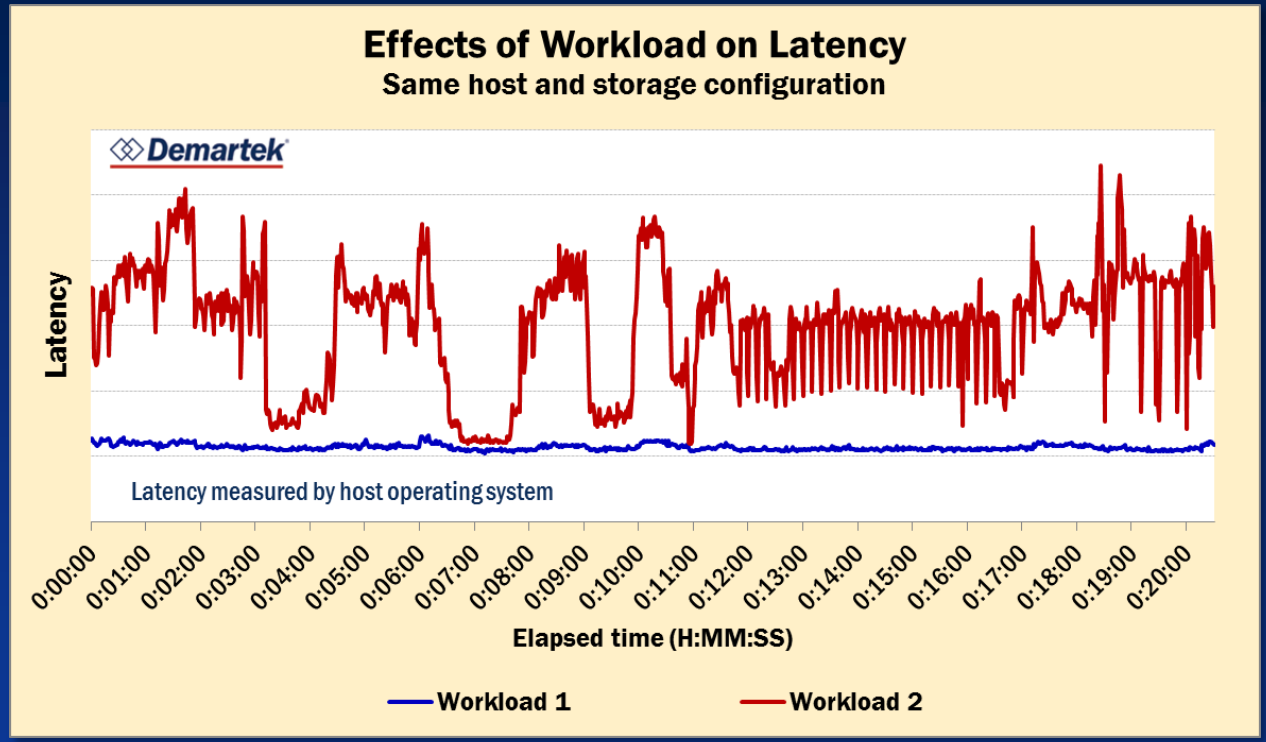
Generic Latency Results

One all-flash array.

Two different workloads running simultaneously.

The nature of each workload has a large impact on latency.

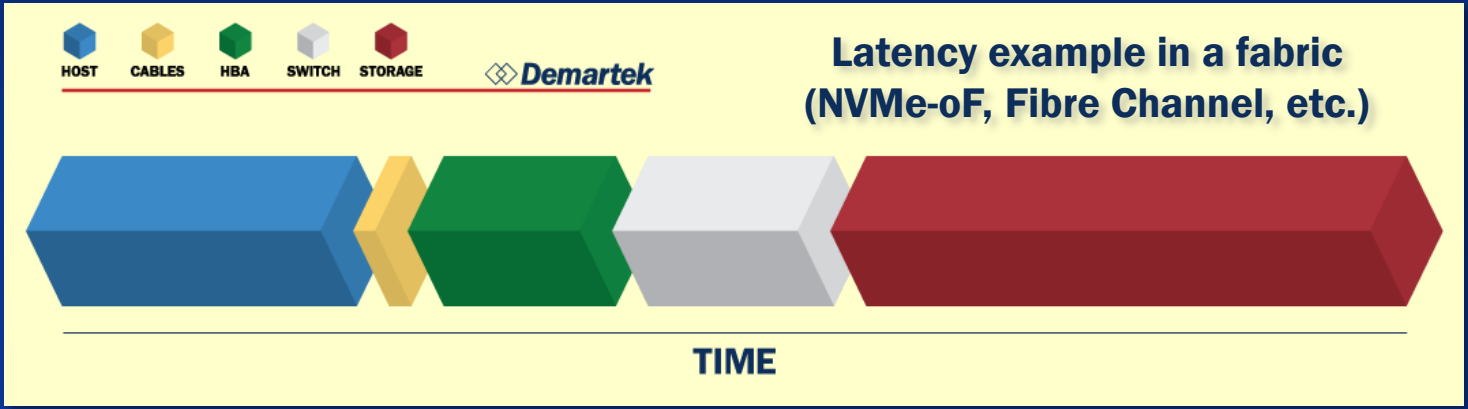
At 06:00 & 10:00 the red workload affected the latency of the blue workload.



Storage Performance Measurement

▸ Multiple Layers

- There are many places to measure storage performance, including software layers and hardware layers
  - Multiple layers in the host server, storage device and in between
  - *The storage hardware is not the only source of latency*



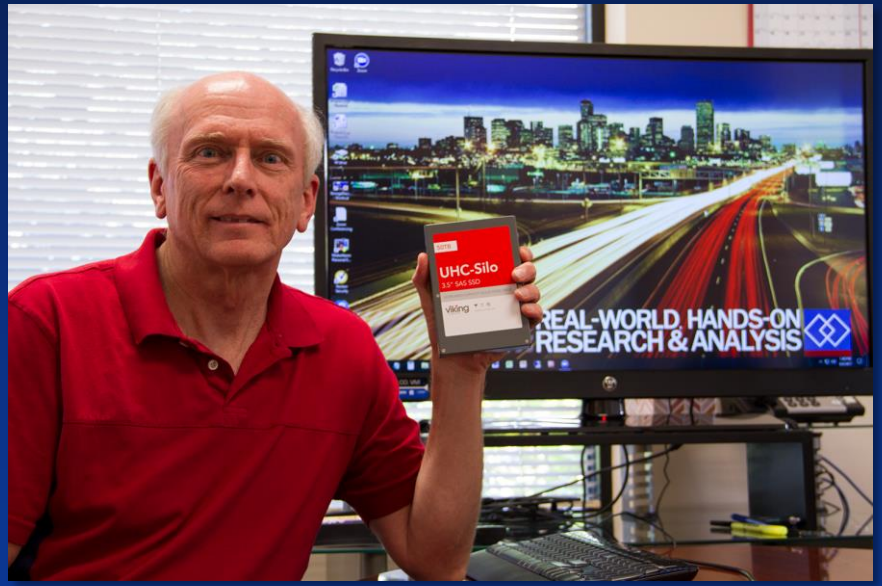
Demartek – Independent Test Lab

- We are not a product manufacturer
- We work with most product manufacturers
- We use almost every interface, device type, etc.
- We run system-level tests with real operating systems and applications – just like end-users
- We test current and new technologies

Flash Memory Summit

ExaDrive® – 50 TB 12Gb/s SAS SSD

- Publicly announced on August 7, 2017
- We tested this drive in our lab in Colorado
- Largest capacity single drive that we have tested to date
- 3.5-inch drive (LFF)



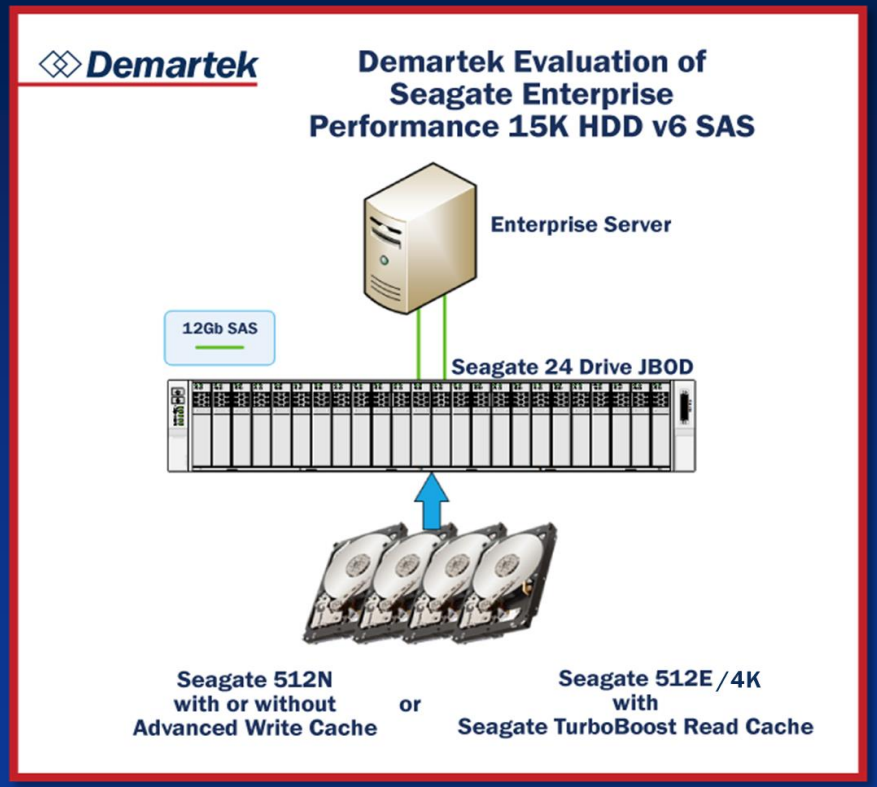
Flash Memory Summit

ExaDrive® – 50 TB 12Gb/s SAS SSD

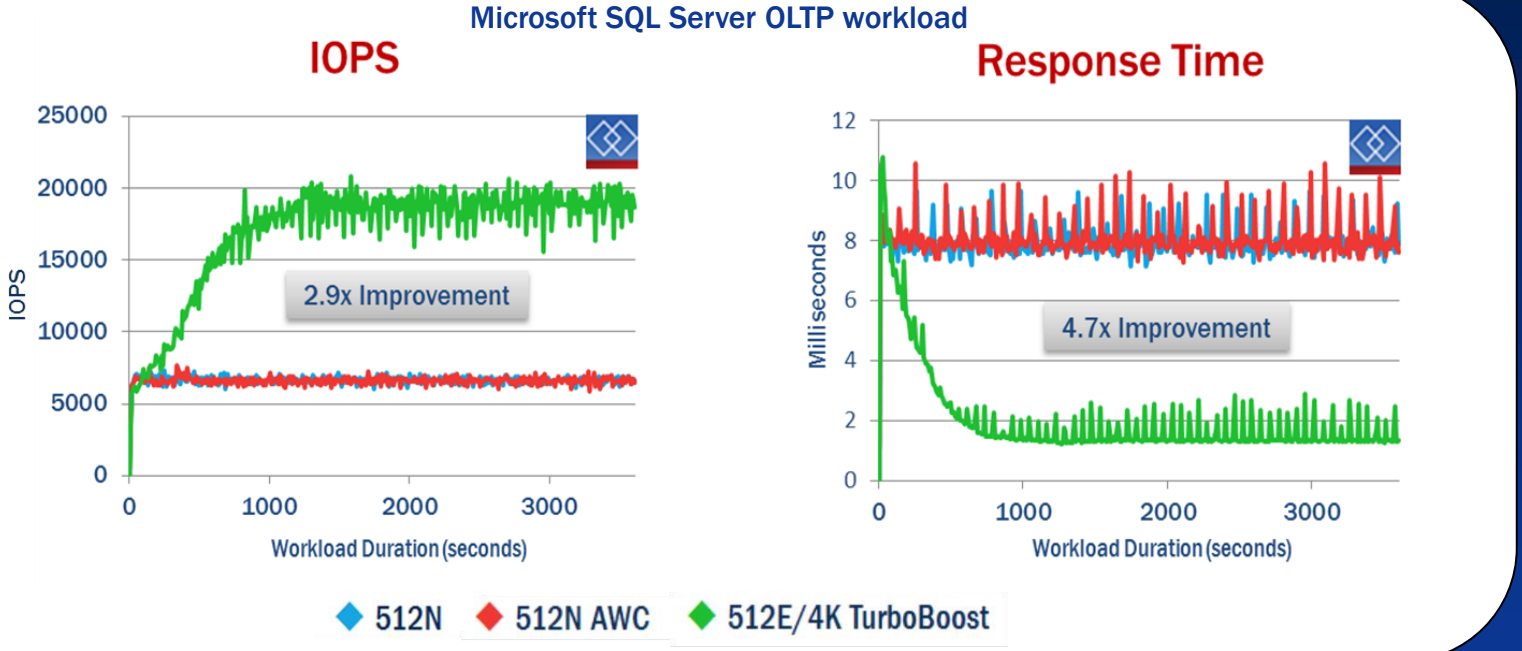
- Full report pending
- 50 TB raw capacity (without de-dupe or compression)
- Inserted into our 60-drive 12Gb/s SAS JBOD
- Recognized immediately by host
- Idle power: 7 – 8 watts
- Average active power consumption: 14.5 – 16.5 watts
- Flash as an archive device?

Seagate TurboBoost HDD

- 15K RPM HDD with NAND flash buffer
- 900 GB capacity
- Advanced format drive (4K with 512B emulated)
- Mixed workloads



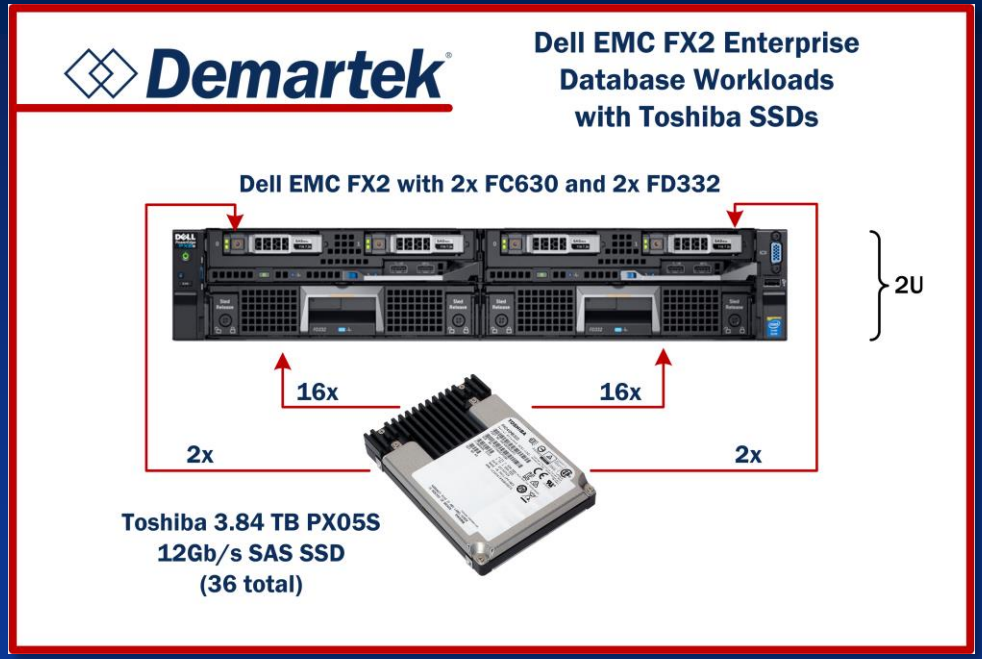
TurboBoost HDD Performance Results



http://www.demartek.com/Demartek_Seagate_TurboBoost_Cache_15K_HDD_Evaluation_2017-04.html

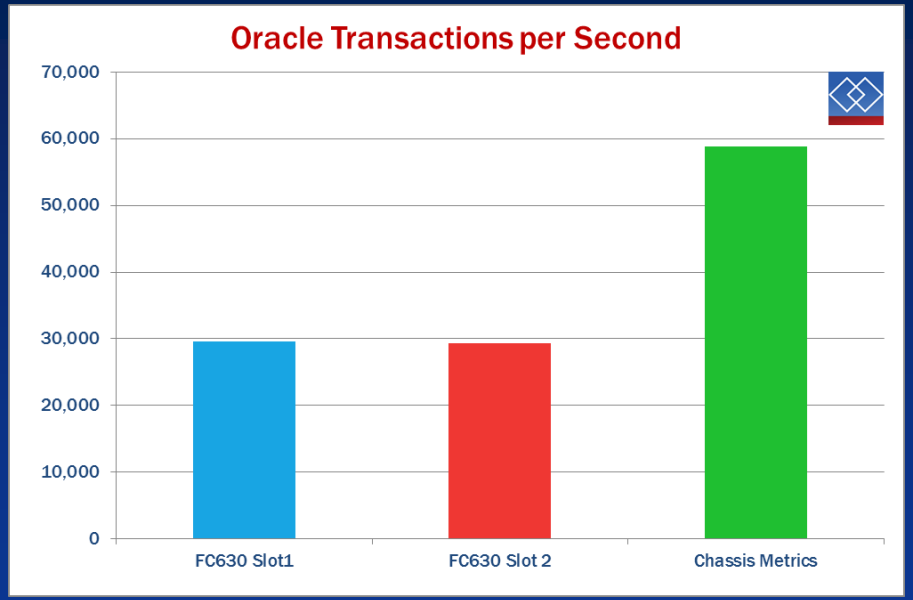
Oracle Database on Dense Platform

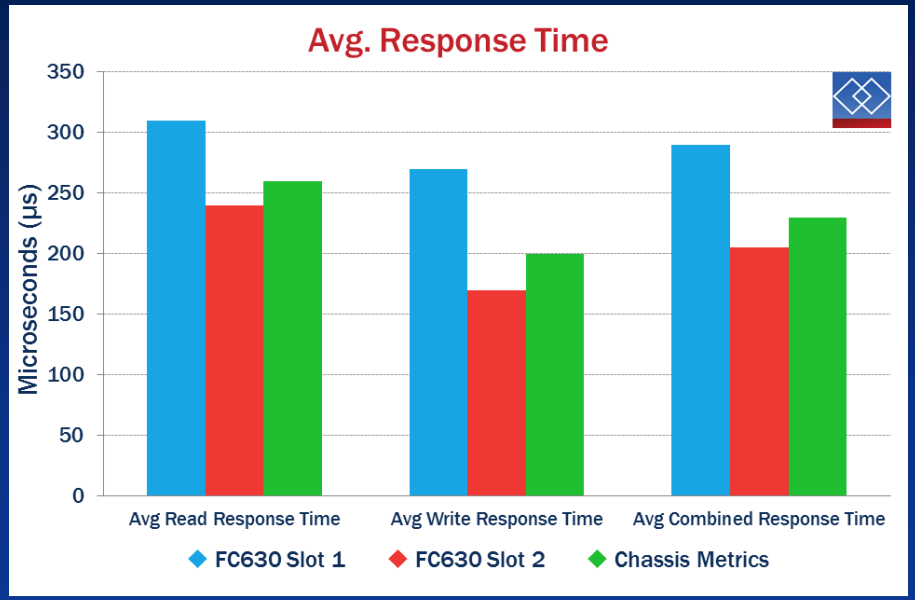
- Dense compute and storage configuration
- 36x Toshiba 3.84 TB 12Gb/s SAS SSDs
- 2.5-inch (SFF)
- 2 instances of Oracle
- 2 sets of storage



Flash Memory Summit

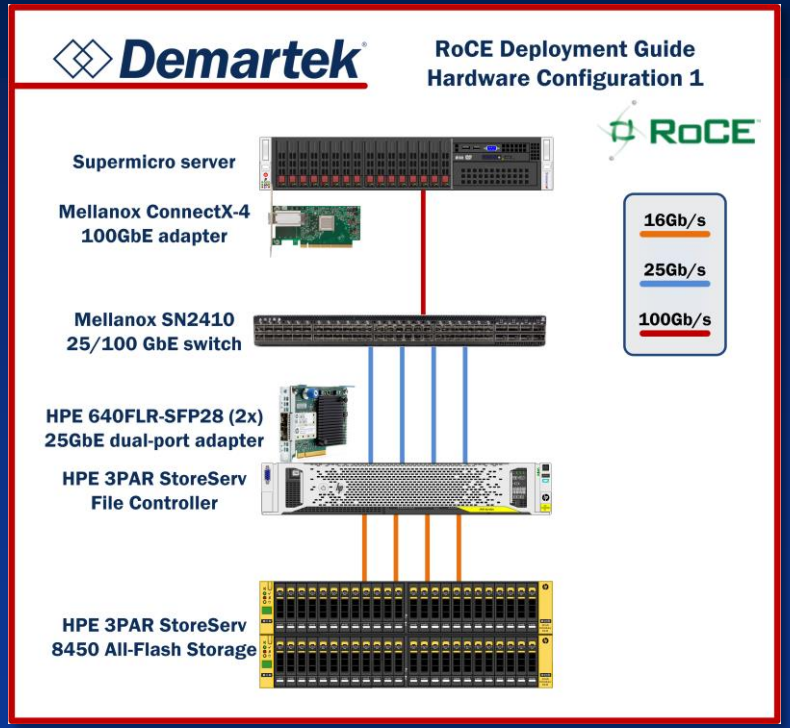
Oracle Performance Results

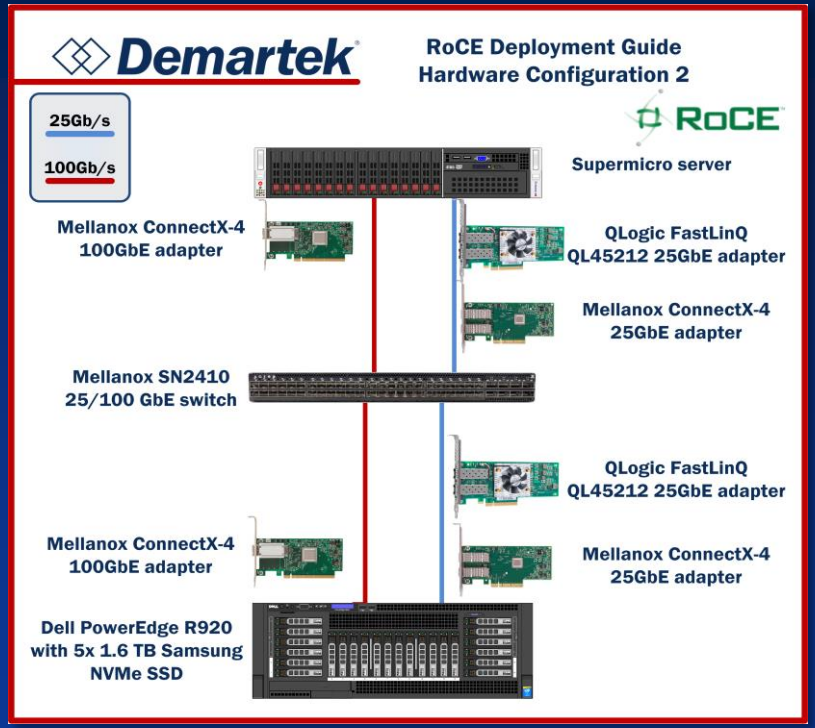




Full report pending – available soon at www.demartek.com/news

RoCE Deployment Guide





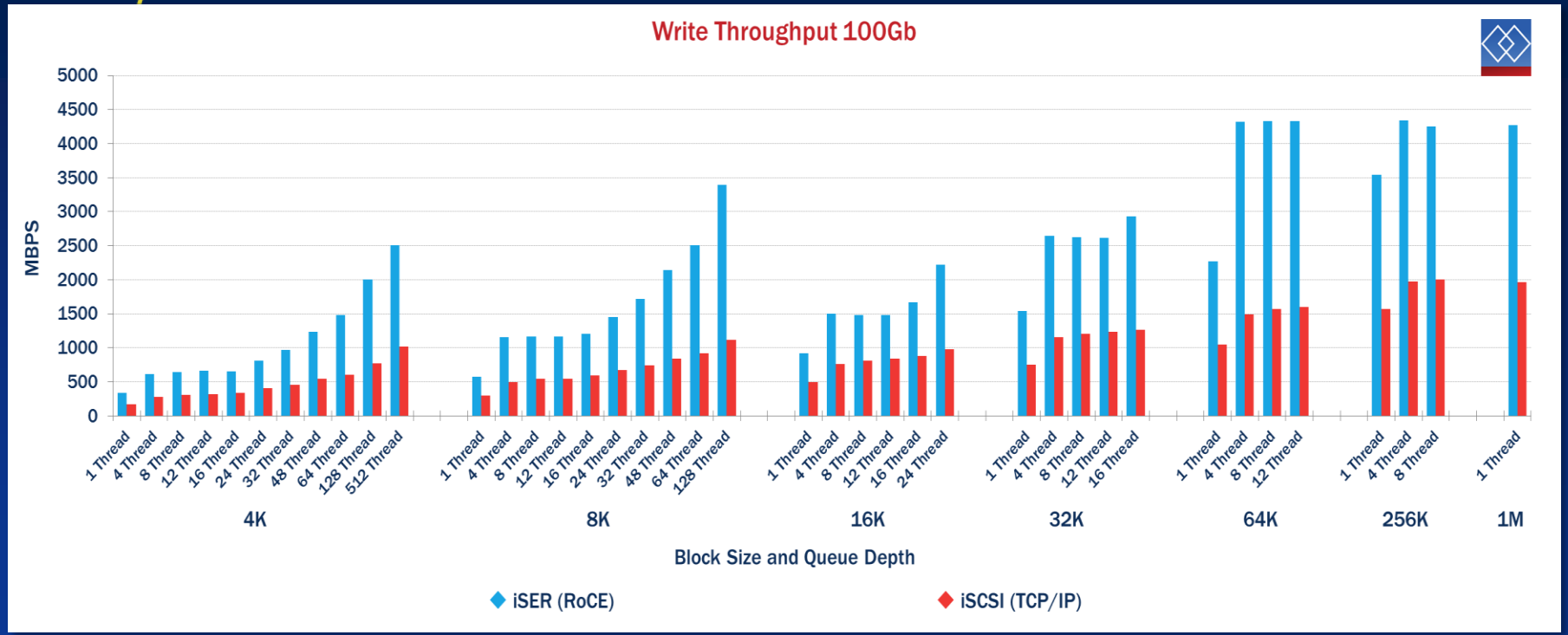
RoCE Deployment Guide

- Windows performance tests
  - Storage Spaces Direct
- Linux performance tests
  - iSER vs. iSCSI
- 25GbE
- 100GbE

Linux iSER Throughput – 100 Gbps

http://www.demartek.com/Demartek_RoCE_Deployment_Guide.html

Linux iSER Throughput – 100 Gbps

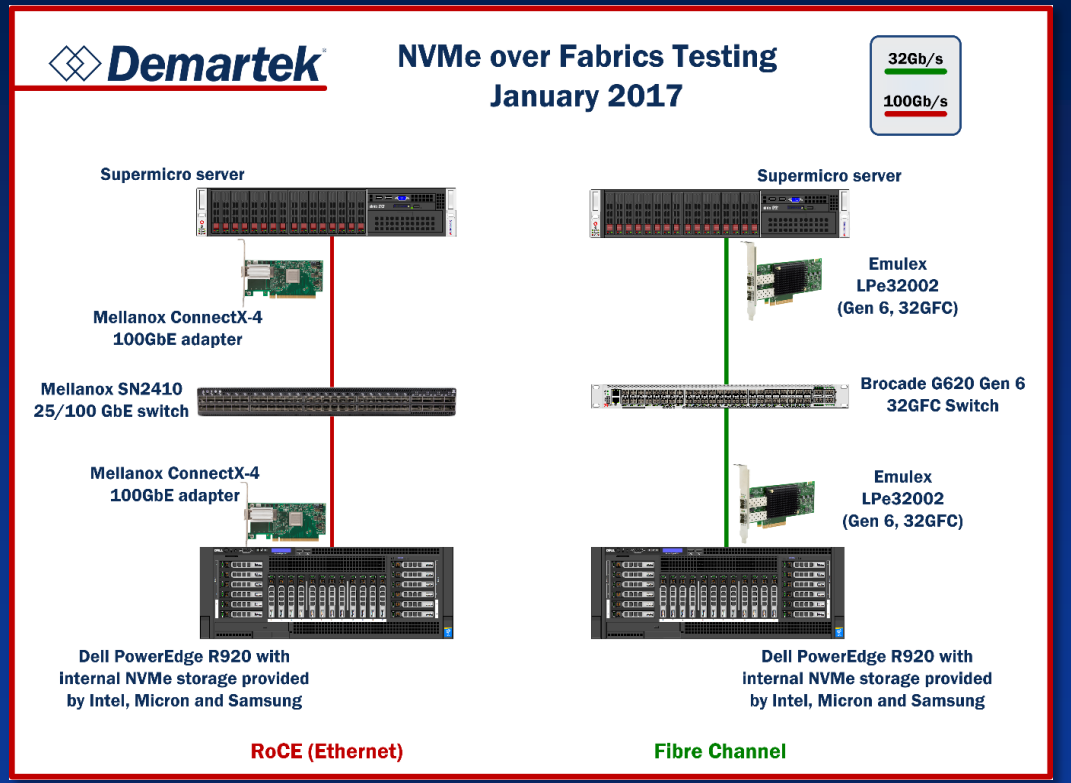


http://www.demartek.com/Demartek_RoCE_Deployment_Guide.html

Flash Memory Summit

NVMe over Fabrics (NVMe-oF)

- We ran NVMe-oF with
  - RDMA
  - Fibre Channel
- Mixture of NVMe drives
- Full report is pending



NVMe-oF Observations

- Latency is workload dependent
- RDMA and Fibre Channel each have different advantages and disadvantages as a storage fabric

Demartek NVMe-oF Rules of Thumb

To achieve maximum throughput in a storage target (without oversubscription):

- At least one 25Gb or faster network port for each NVMe drive (PCIe 3.0 x4) for large-block sequential I/O
- Dual-port 25GbE or 32GFC adapters need PCIe 3.0 x8
- For every two NVMe drives and network ports 16 lanes of PCIe 3.0 are needed (FC has more headroom)
- Prospects are better with PCIe 4.0

http://www.demartek.com/Demartek_NVMe_over_Fabrics_Rules_of_Thumb_2017-08.html

Flash Memory Summit

Conclusions

- Real-world workloads can be “messy” compared to synthetic workloads
  - Variable I/O characteristics and multiple factors influencing performance
- New flash technologies are yielding very interesting results
- Look for more Demartek workload test results with various forms of flash (NVDIMM, persistent memory, etc.)

Demartek Free Resources

- Demartek SSD Zone – www.demartek.com/SSD
- Demartek iSCSI Zone – www.demartek.com/iSCSI
- Demartek FC Zone – www.demartek.com/FC
- Demartek commentary: "Horses, Buggies and SSDs" www.demartek.com/Demartek_Horses_Buggies_SSDs_Commentary.html
- Demartek Video Library - www.demartek.com/Demartek_Video_Library.html
- Demartek News – www.demartek.com/news
- This presentation - http://www.demartek.com/Demartek_Presenting_FlashMemorySummit_2017-08.html

Performance reports, Deployment Guides and commentary available for free download.

Flash Memory Summit

Thank You!

Demartek®

Demartek public projects and materials are announced on a variety of social media outlets. Follow us on any of the above.

Sign-up for the Demartek monthly newsletter, *Demartek Lab Notes*.
www.demartek.com/newsletter