Flash Memory SUMMIT

How Flash-Based Storage Performs on Real Applications

Session 102-C

Dennis Martin, President

Demartek®

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 → 💻

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 (DAS, NAS, SAN, 10/25/40/100 GbE, 16/32 GFC)
- We prefer to run real-world applications to test servers and storage solutions (databases, Hadoop, etc.)
- Demartek is an EPA-recognized test lab for *ENERGY STAR Data Center Storage* testing

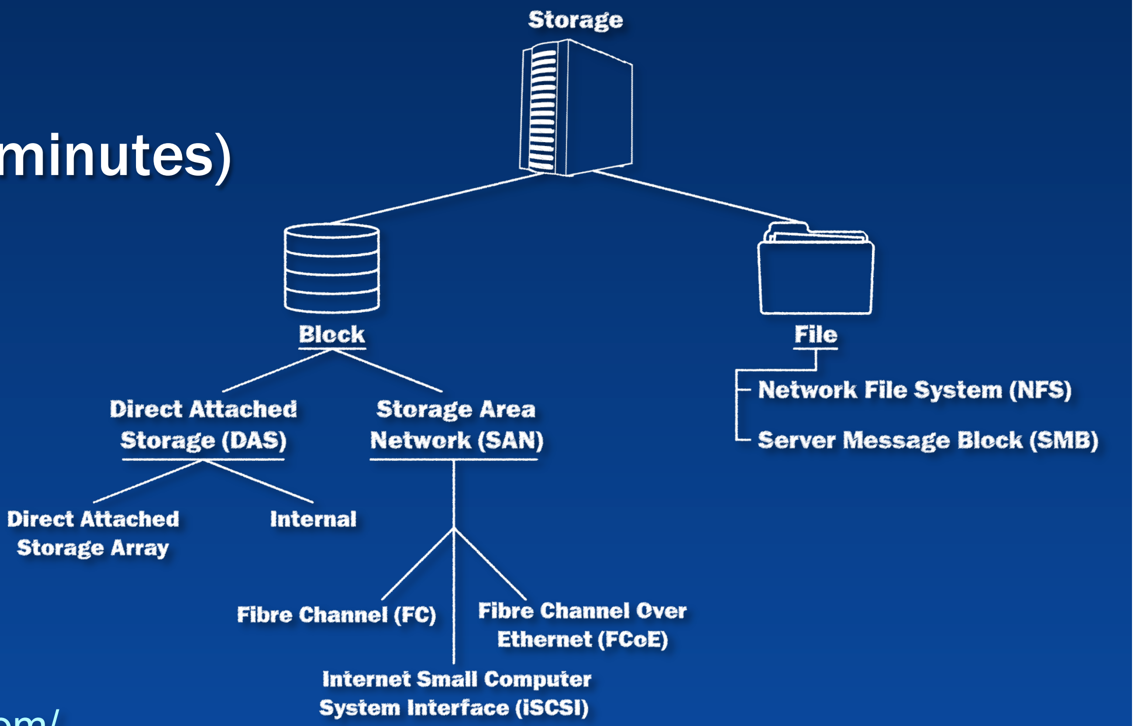
- Website: www.demartek.com/TestLab

SNIA Emerald™
RECOGNIZED TESTER

Demartek Tutorial Videos

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

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



Storage

Block

Direct Attached Storage (DAS)

Direct Attached Storage Array

Internal

Storage Area Network (SAN)

Fibre Channel (FC)

Fibre Channel Over Ethernet (FCoE)

Internet Small Computer System Interface (iSCSI)

File

Network File System (NFS)

Server Message Block (SMB)

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

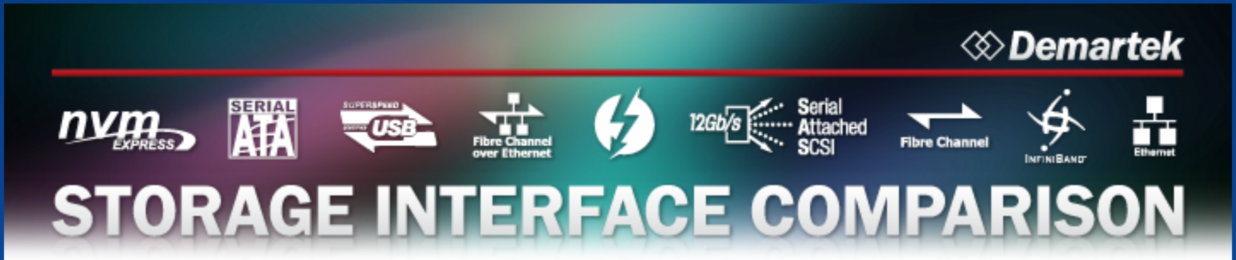
- Direct Attach Storage (DAS)
  - Storage controlled by a single server: inside the server or directly connected to the server ("server-side")
  - *Block* storage devices
- Network Attached Storage (NAS)
  - File server that sends/receives *files* from network clients
- Storage Area Network (SAN)
  -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

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 devices (SSDs, HDDs) and PCIe, NVMe

Storage Interface Comparison

- Demartek Storage Interface Comparison reference page
  - Search engine: *Storage Interface Comparison*
  - Includes new interfaces such as 25GbE, 32GFC, Thunderbolt 3


Demartek
nvm EXPRESS
SERIAL ATA
SUPERSPEED USB
Fibre Channel over Ethernet
12Gb/s Serial Attached SCSI
Fibre Channel
InfiniBand
Ethernet
STORAGE INTERFACE COMPARISON

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

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

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

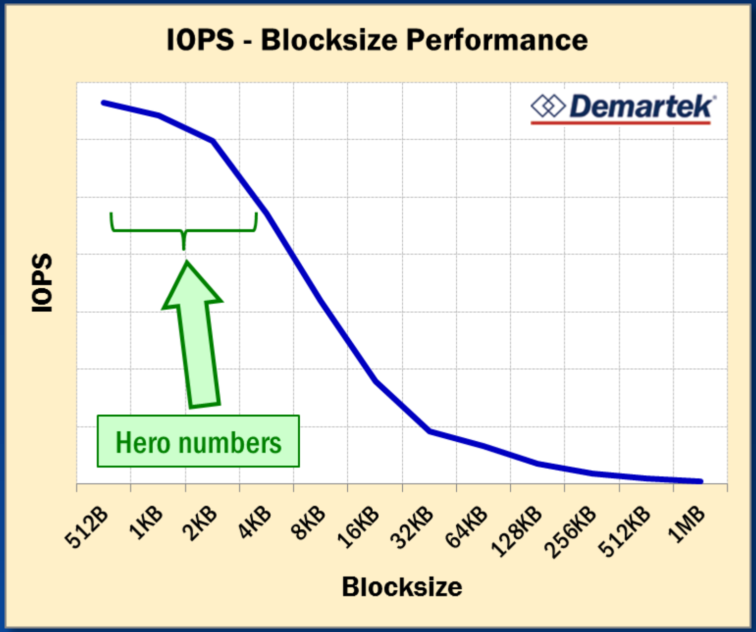
Real-world Workloads

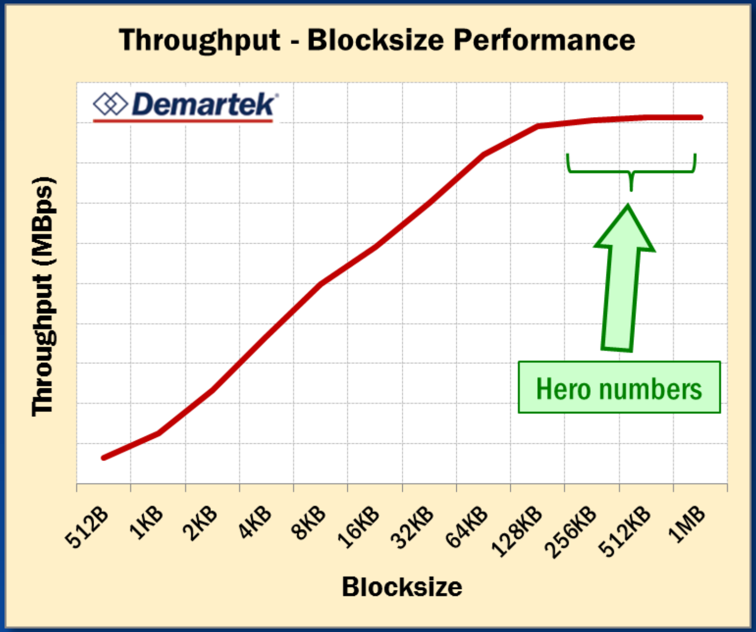
- Use variable levels of compute, memory and I/O resources as the work progresses
    - May use different and multiple I/O characteristics simultaneously for I/O requests (block sizes, queue depths, read/write mix and random/sequential mix)
- Many applications capture their own metrics such as database 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

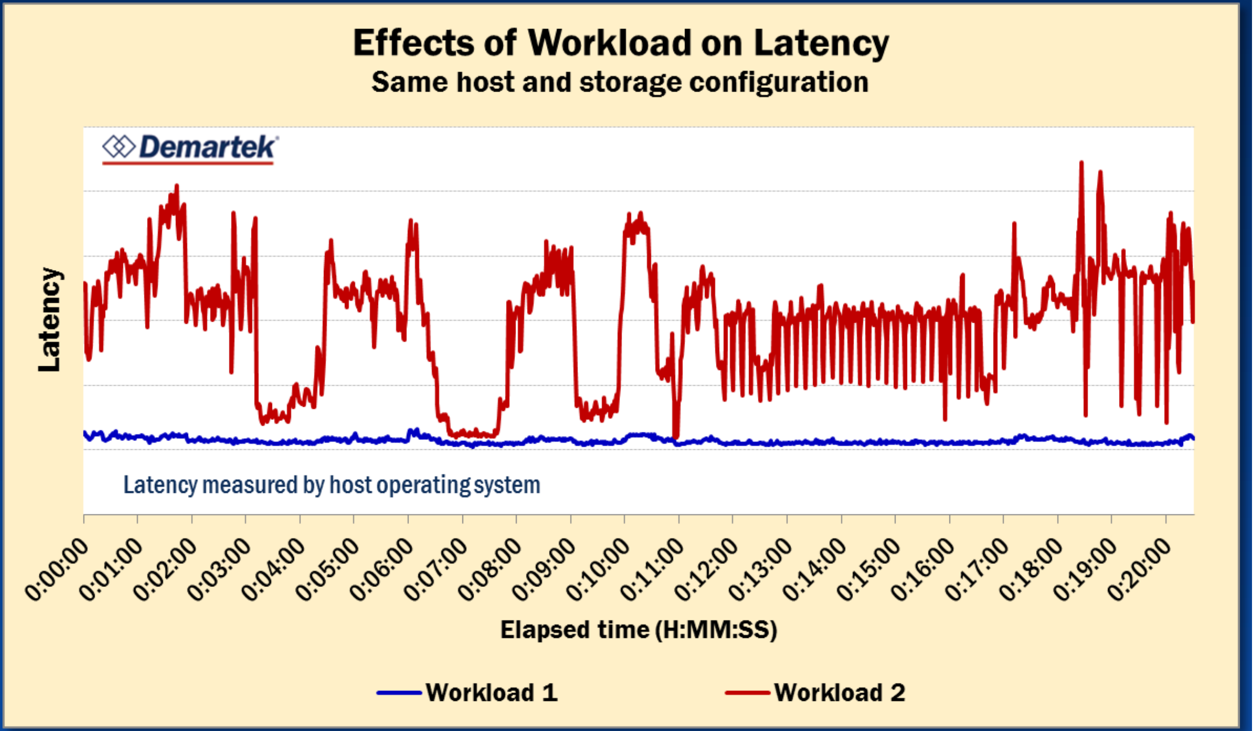
Generic Latency Results

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.

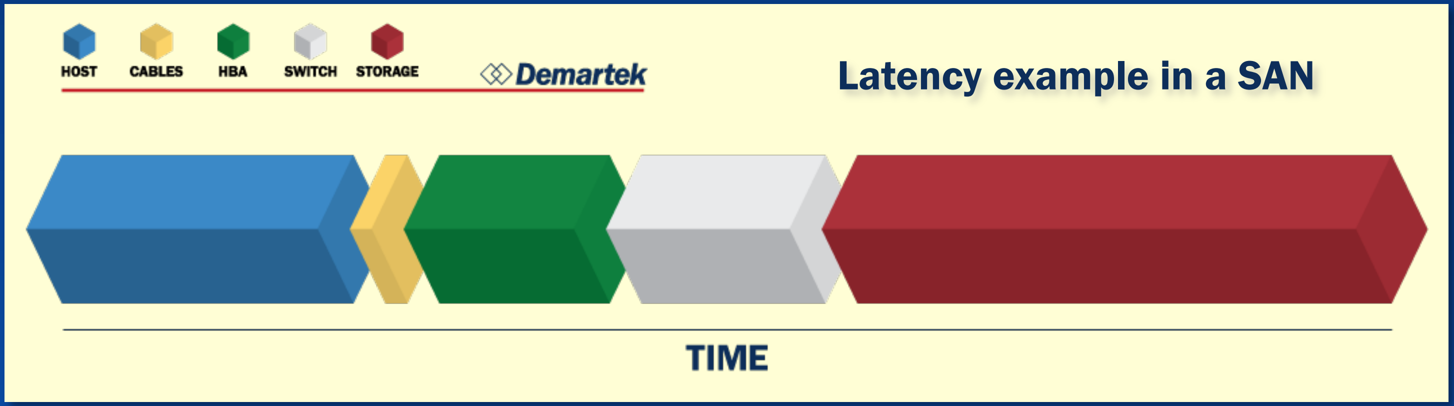


Demartek

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*

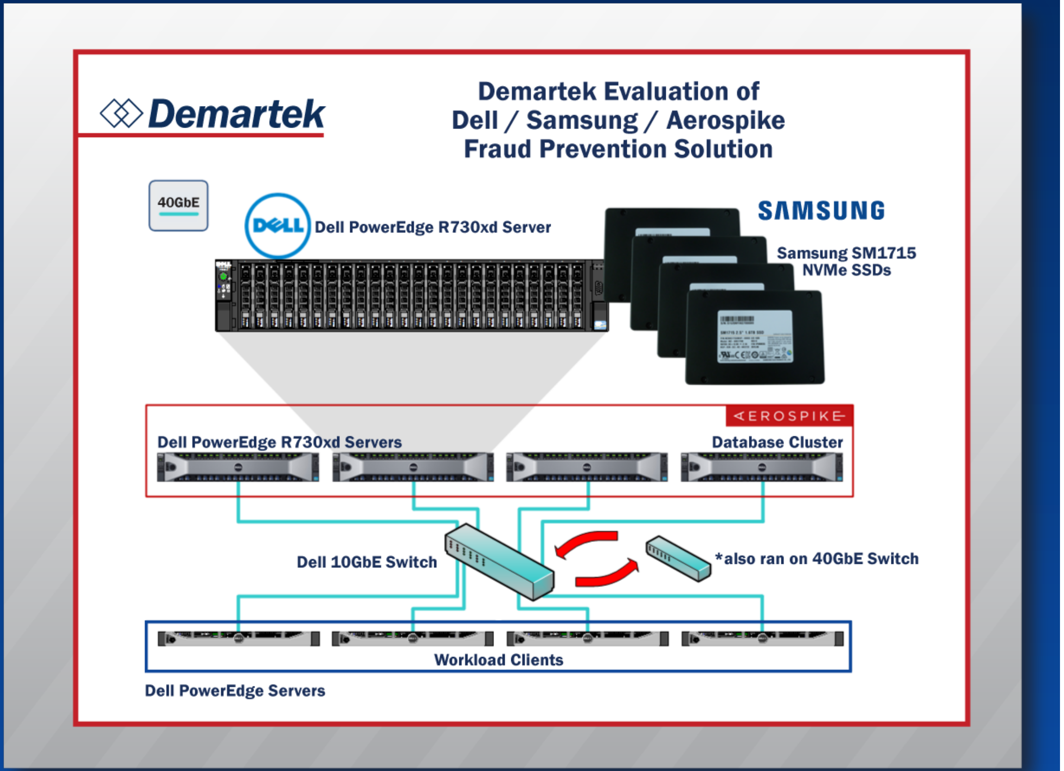


General Notes on These Tests

- SQL Server, Oracle database best practices:
  - Put database files and logs on different volumes
  - Different I/O patterns for database files and logs
- SQL Server and Oracle database will take as much machine as you make available (cores, memory, etc.)
  - Different results for 4-proc server with lots of memory vs. 1-proc server with small memory
- Earlier in 2016, we changed the format of our reports, so some of the graphs have a different style

NVMe & Credit Card Fraud Prevention

- Credit card fraud prevention
  - Retrieve data
  - Run fraud prevention analytics
  - Return a score in real-time
- Goals:
  - Meet customer SLA
  - High numbers of reads while maintaining good write rate
- NoSQL database stored on NVMe drives

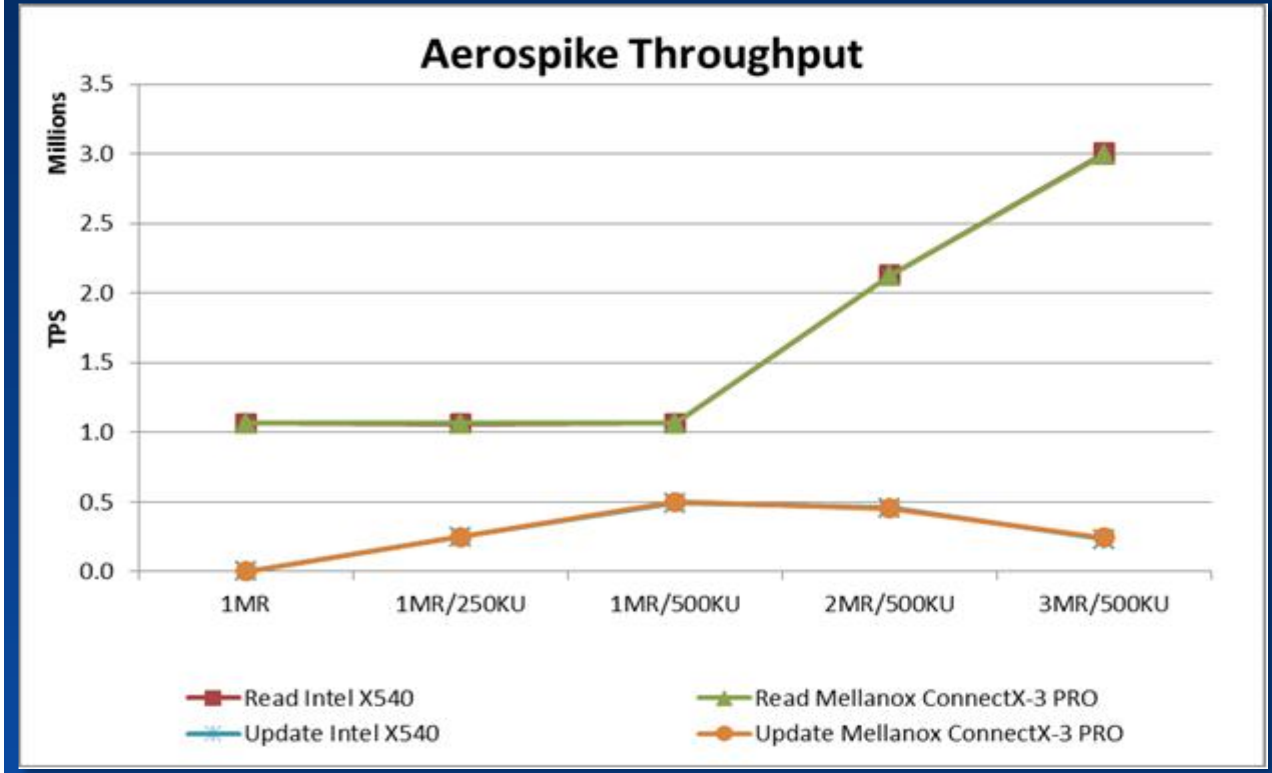


http://www.demartek.com/Demartek_Dell_Samsung_Aerospike_Fraud_Prevention_Evaluation_2015-12.html

NVMe & Credit Card Fraud Prevention

- Test Phase 1 – load 2 billion objects to database
- Test Phase 2 – run phase
    - Steady-state of 1 million database read operations per second
    - Add 250,000 database write/update operations per second
    - Increase write/updates to 500,000 per second
    - Increase reads to 2 million reads per second
    - Increase reads to 3 million reads per second

Transactions per second



Per Second Statistics

1MR = 1 million reads

2MR = 2 million reads

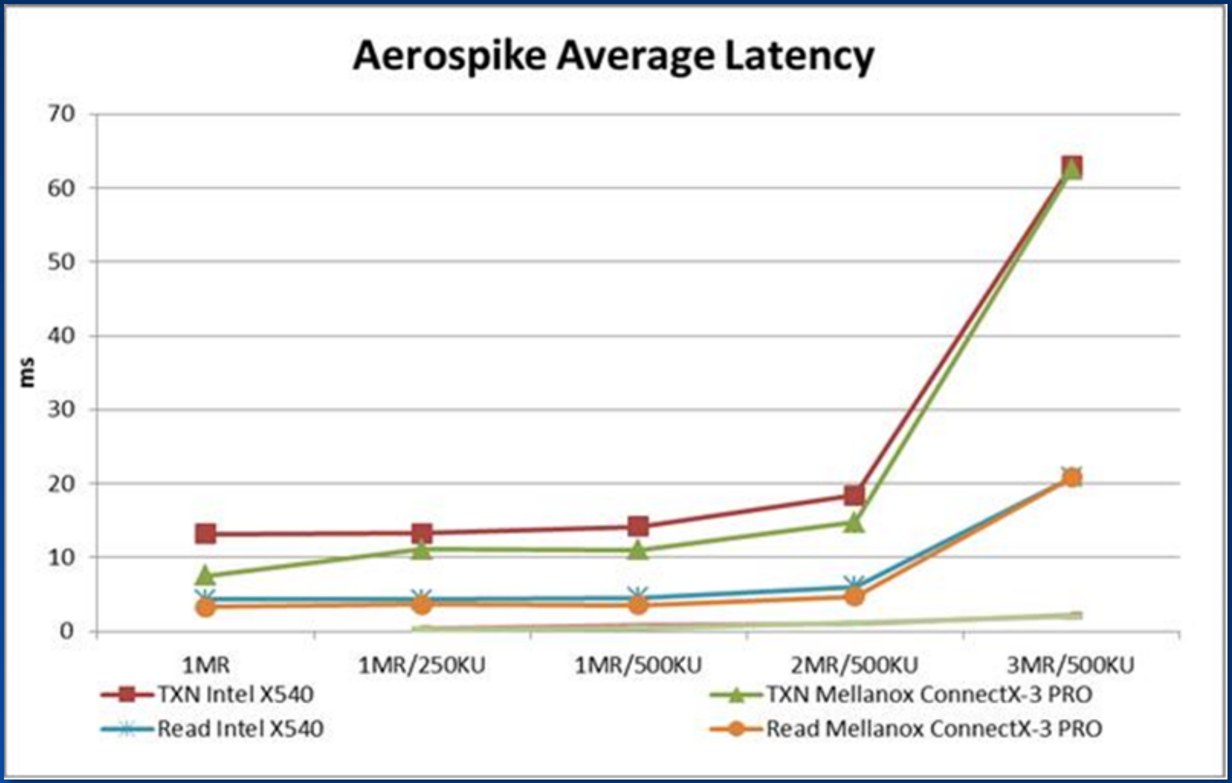
3MR = 3 million reads

250KU = 250,000 updates

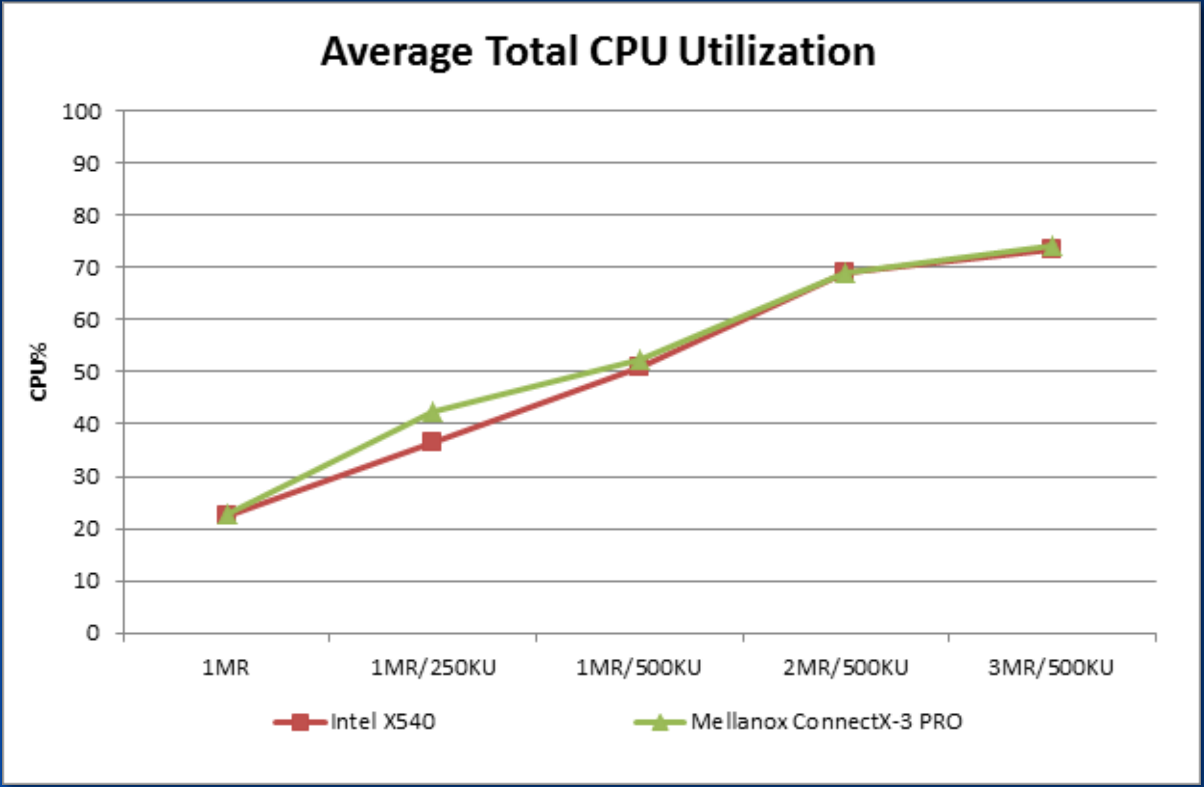
500KU = 500,000 updates

Application Latency

This is application latency, not storage device latency



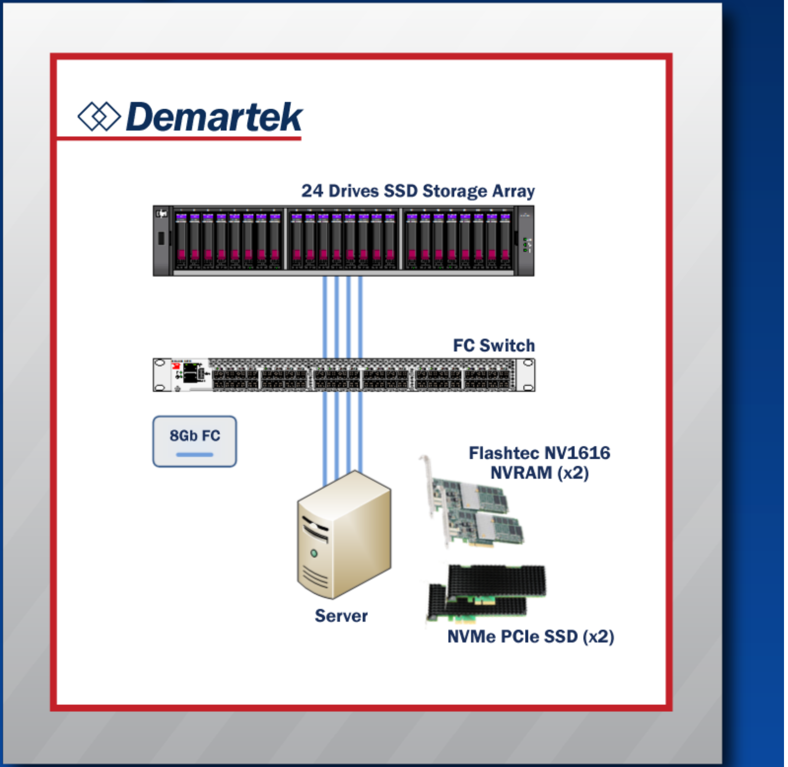
CPU Utilization



Demartek

NVRAM & Database Logs

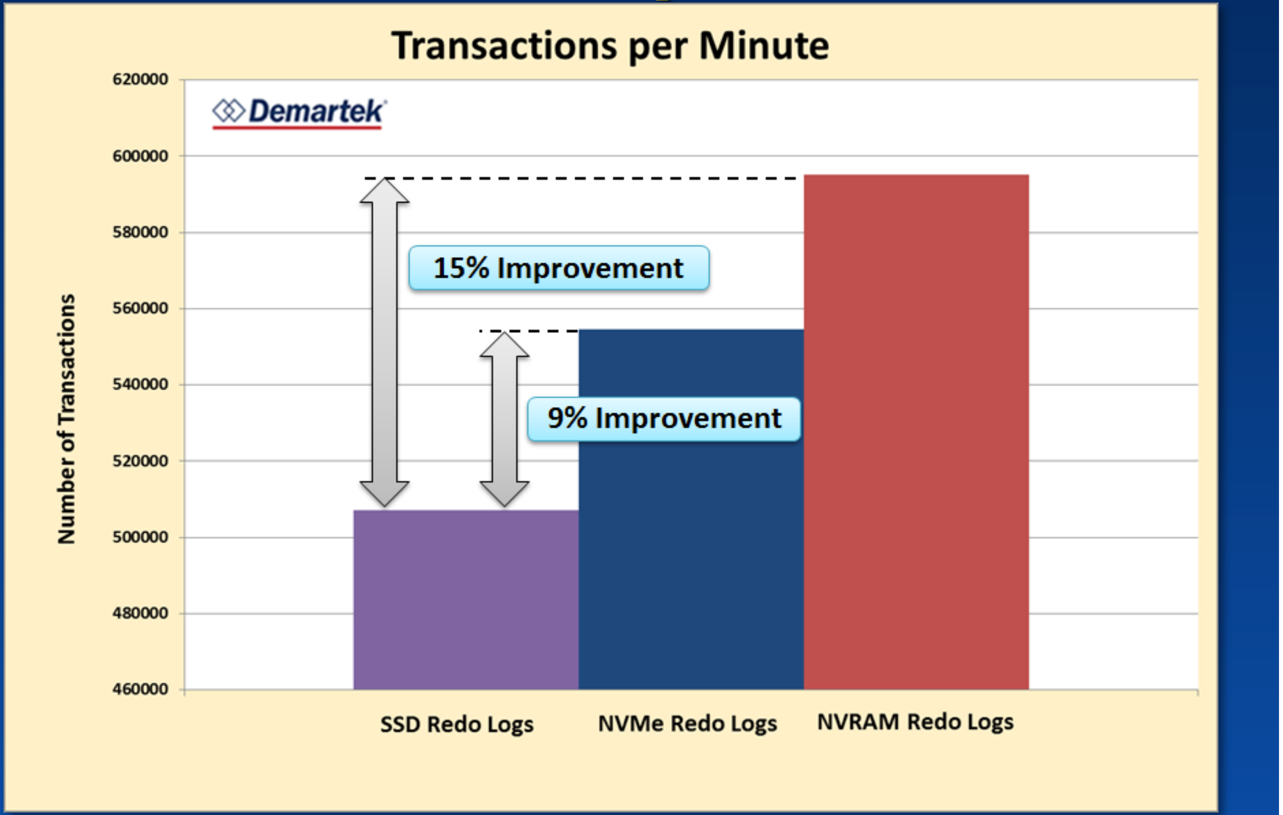
- Database logging
  - Database updates are logged to a journal or log file
  - Critical for recovery or rollback
  - Speed of log storage makes a difference
- Three types of log storage:
  - SSD storage array, NVMe drive, NVRAM
- Oracle database, OLTP workload
  - Log files are called “Redo Logs”



http://www.demartek.com/Demartek_Microsemi_Flashtec_NV1616_NVRAM_Database_Performance_2016-06.html

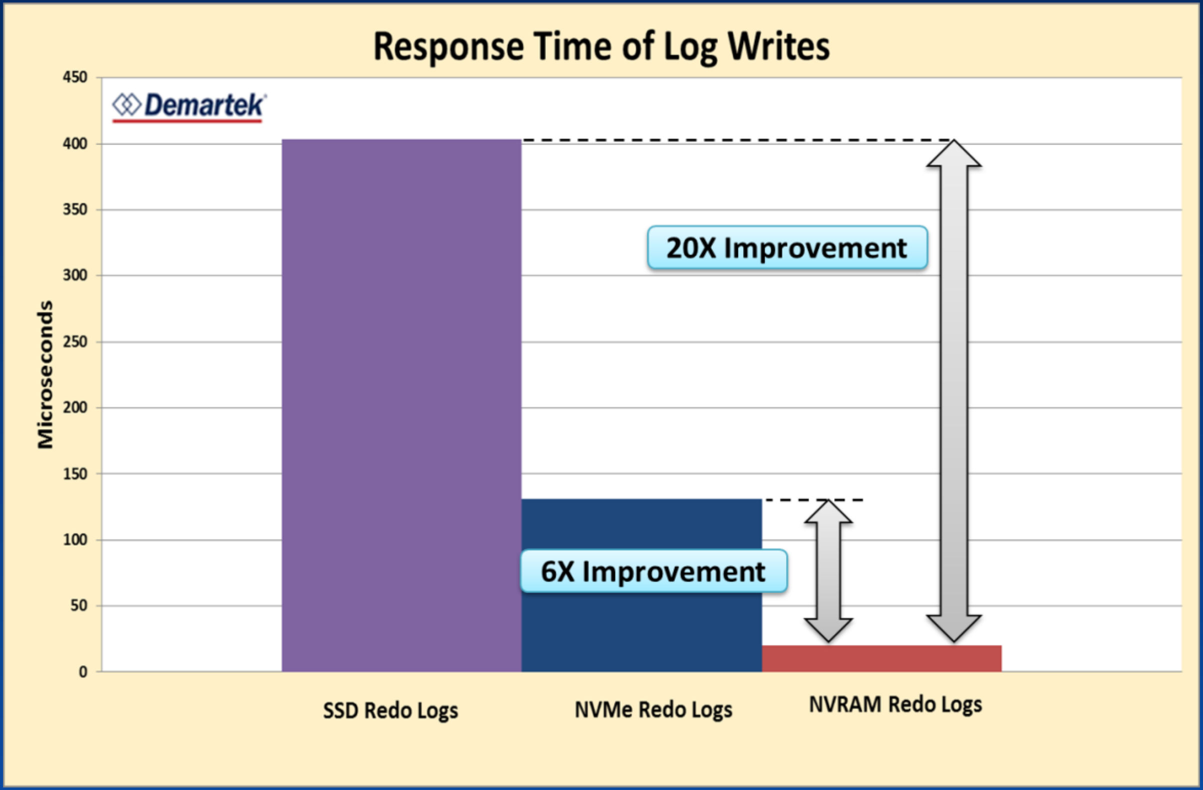
Database Transactions per Minute

Faster log writes improves overall database performance



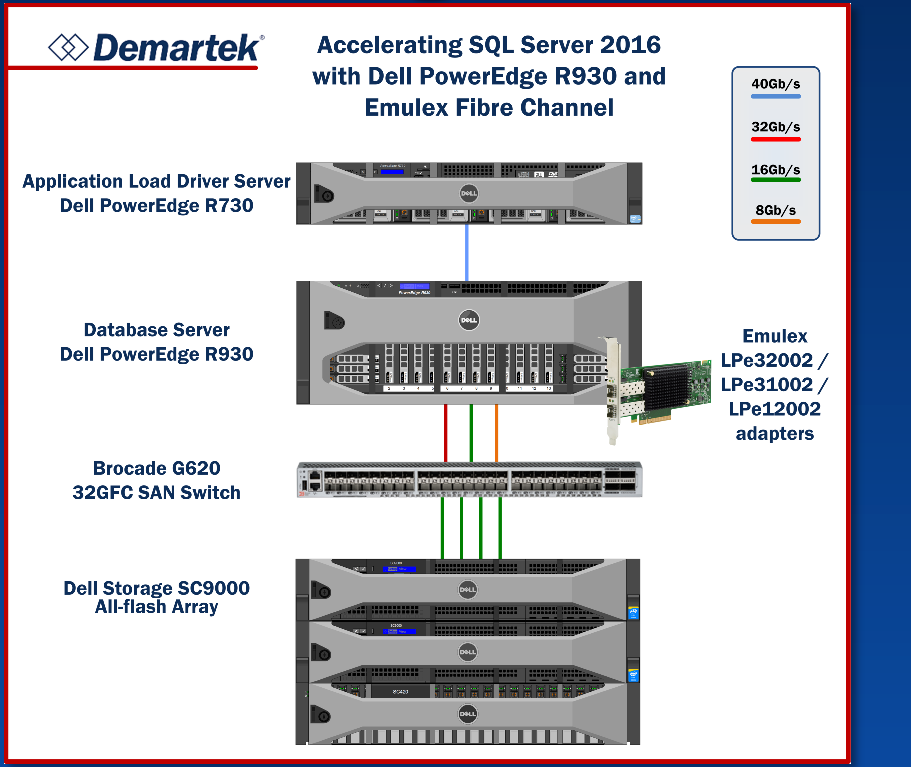
Demartek

Log Write Response Time (Latency)



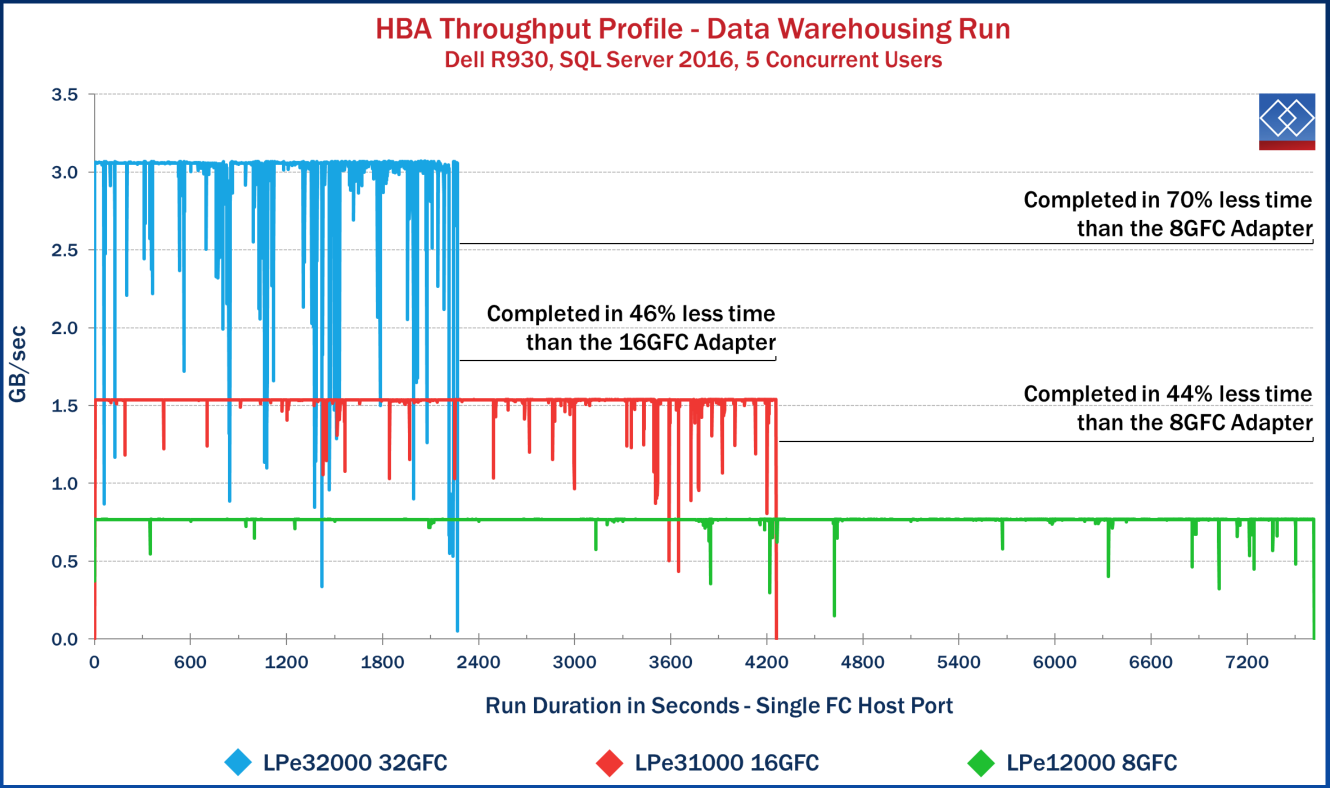
32GFC & Data Warehousing

- Data Warehousing
  - Decision Support
  - Complex analytics queries
  - Computes scores
  - Fixed set of work
  - Bandwidth-intensive workload
- Three generations of Fibre Channel technology:
  - 8GFC, 16GFC, 32GFC
- Microsoft SQL Server 2016


Demartek
Accelerating SQL Server 2016 with Dell PowerEdge R930 and Emulex Fibre Channel
40Gb/s
32Gb/s
16Gb/s
8Gb/s
Application Load Driver Server
Dell PowerEdge R730
Database Server
Dell PowerEdge R930
Emulex LPe32002 / LPe31002 / LPe12002 adapters
Brocade G620
32GFC SAN Switch
Dell Storage SC9000
All-flash Array

http://www.demartek.com/Demartek_Dell_R930_Emulex_32GFC_SQL_Server_2016_Evaluation_2016-06.html

Throughput



Completed in 70% less time than the 8GFC Adapter

Completed in 46% less time than the 16GFC Adapter

Completed in 44% less time than the 8GFC Adapter

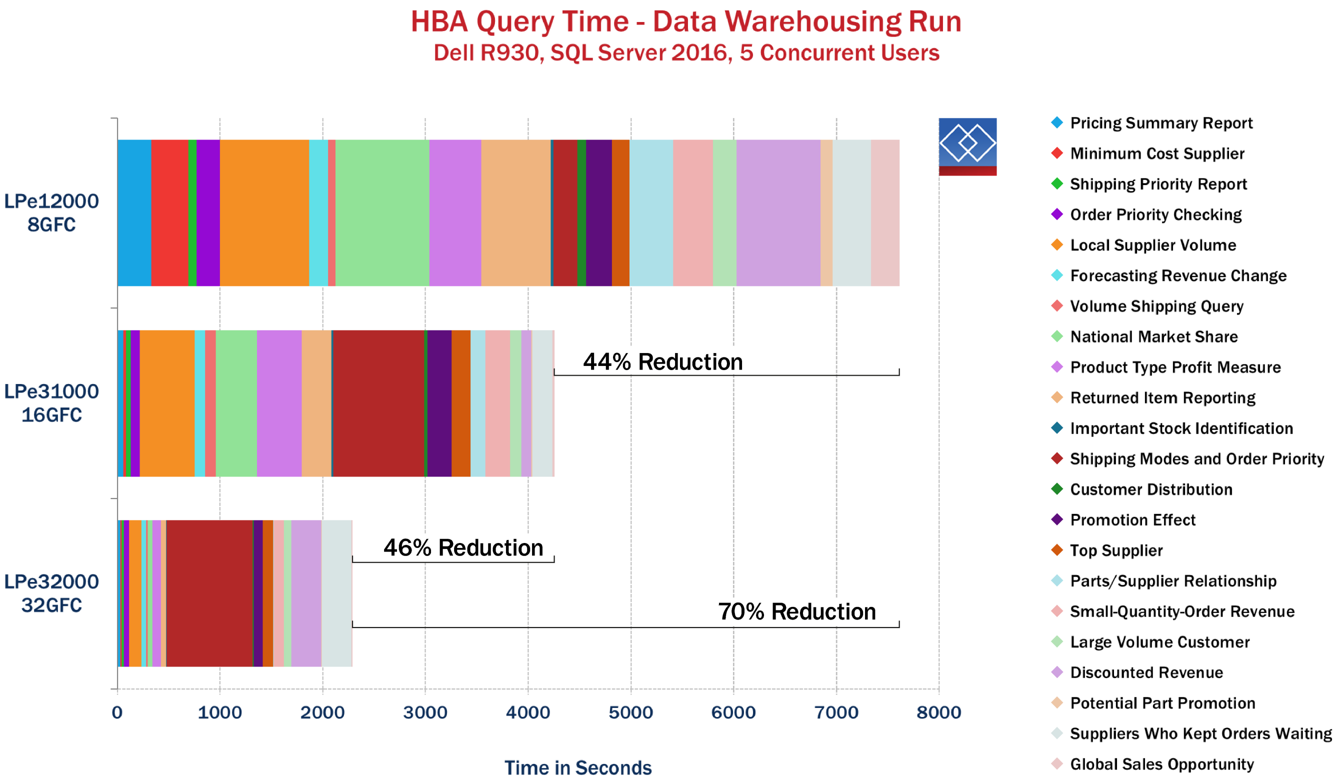
Application Latency

Time to Complete

8GFC: 127 minutes

16GFC: 71 minutes

32GFC: 38 minutes


HBA Query Time - Data Warehousing Run
Dell R930, SQL Server 2016, 5 Concurrent Users

100GbE RoCE and NVMe Storage

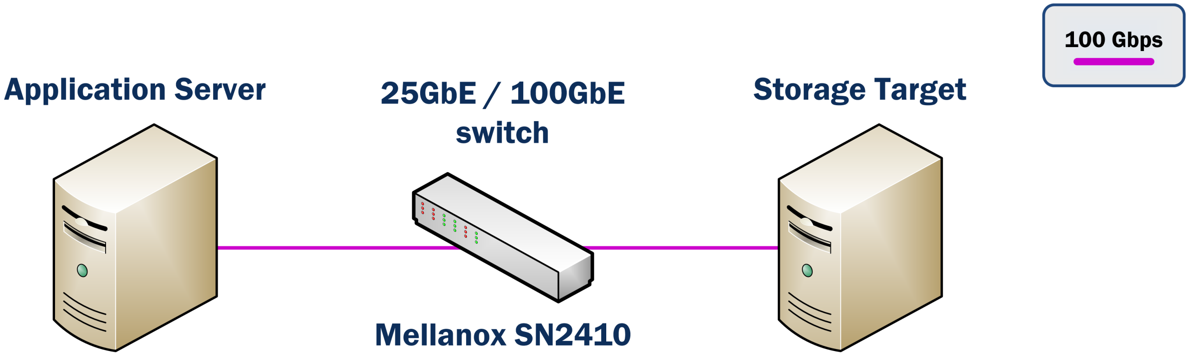
Work-in-Progress

Showing various deployments of RoCE equipment

If you make hardware that supports RoCE, contact me.

Demartek

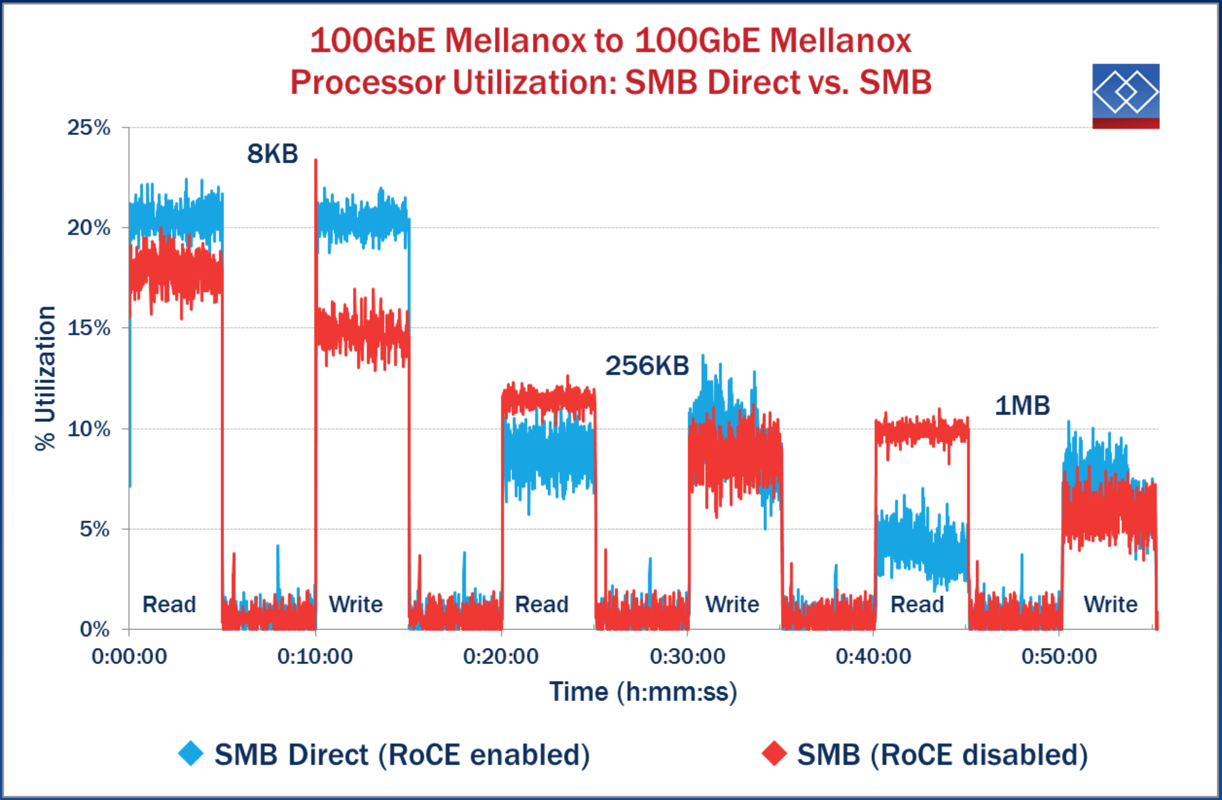
RoCE Deployment Guide Testing



2x Intel Xeon E5-2690 v2, 3.0GHz
(20 total cores)
4x Samsung SM-1715 NVMe SSD
1x Mellanox ConnectX-4 100GbE

4x Intel Xeon E5-4880 v2, 2.5GHz
(60 total cores)
4x Samsung SM-1715 NVMe SSD
1x Mellanox ConnectX-4 100GbE

RoCE CPU Utilization: File Share

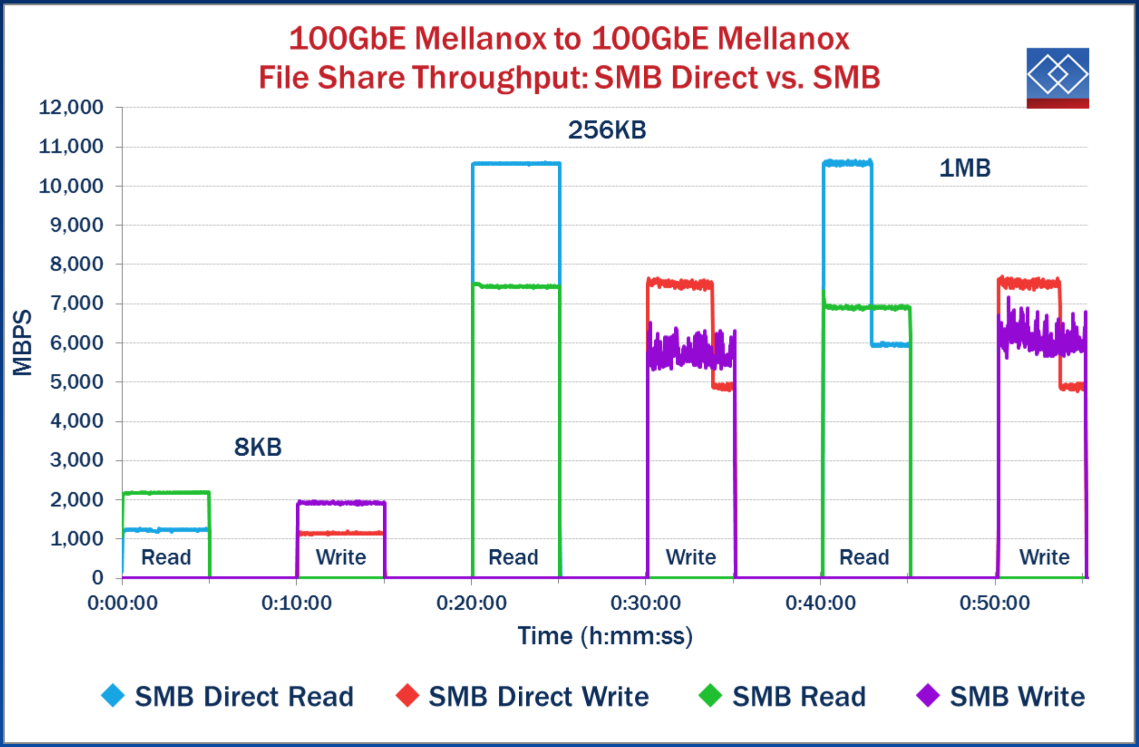


Demartek

RoCE Throughput: File Share

Windows SMB Direct

Large block size shows noticeable improvement in throughput, especially for file reads.

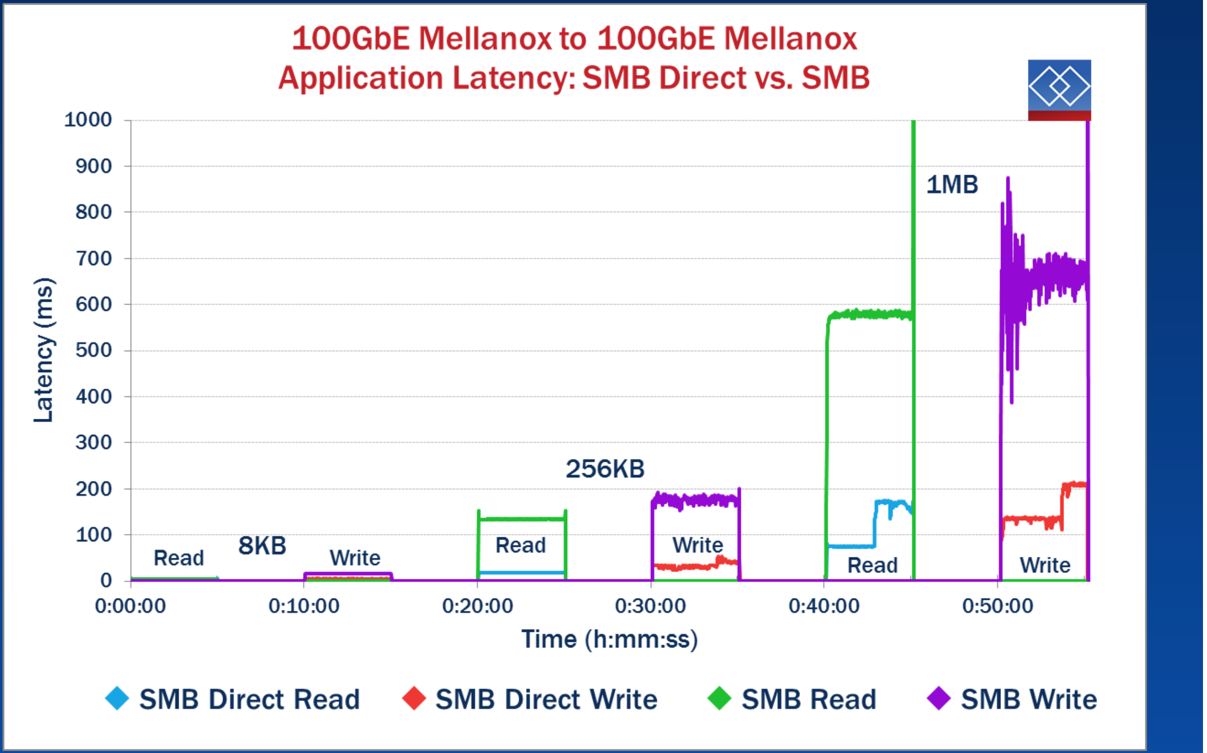


Demartek

RoCE Application Latency: File Share

Windows SMB Direct

Significant latency benefit for file workloads with SMB Direct.



Demartek

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

Demartek Free Resources

- Demartek SSD Zone – www.demartek.com/SS
- Demartek iSCSI Zone – www.demartek.com/iS
- Demartek FC Zone – www.demartek.com/FC
- Demartek SSD Deployment Guide
  www.demartek.com/Demartek_SSD_Deployment_Guide.html
- Demartek commentary: “Horses, Buggies and SSDs”
  www.demartek.com/Demartek_Horses_Buggies_SSDs_Commentary.html
- Demartek Video Library -
  http://www.demartek.com/Demartek_Video_Library.html

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

Thank You!

Demartek®

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

SUBSCRIBE TO THE DEMARTEK NEWSLETTER

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