




# How Flash-Based Storage Performs on Real Applications

## Session 104-C

Dennis Martin, President



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

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

# About Demartek

- ◆ Industry Analysis and ISO 17025 accredited test lab
- ◆ Lab includes enterprise servers, networking & storage (DAS, NAS, SAN, 10GbE, 40GbE, 16GFC)
- ◆ 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](http://www.demartek.com/TestLab)



# 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

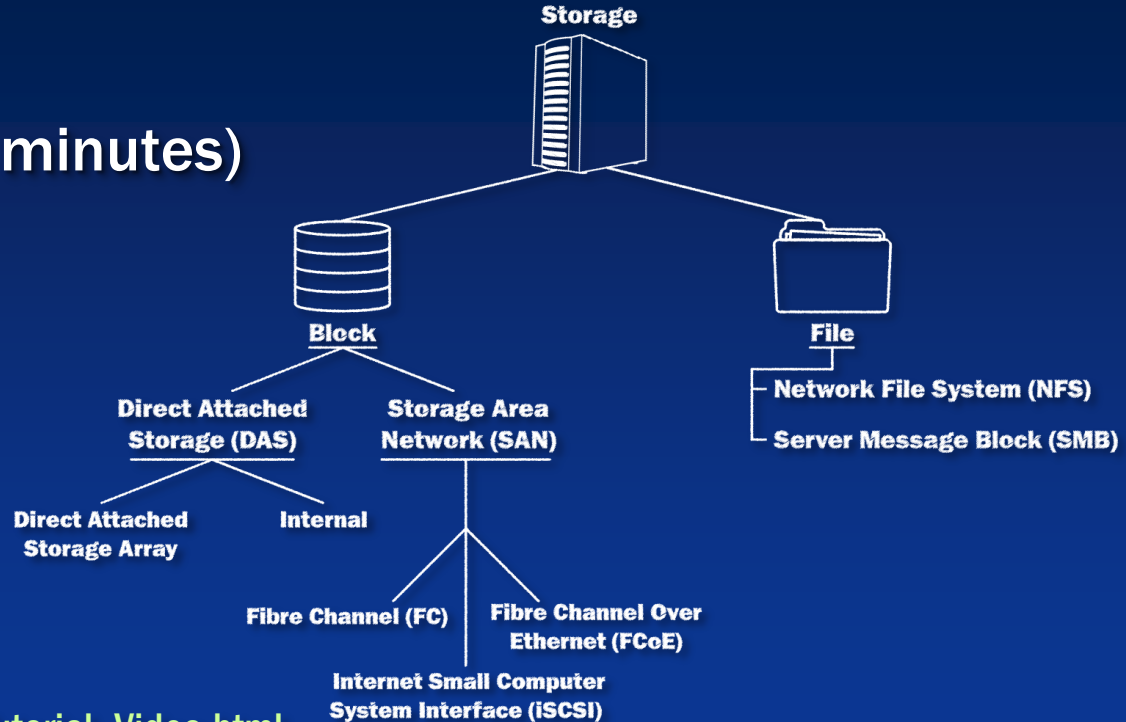
# Enterprise Storage Architectures

## ► 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

# Demartek Tutorial Videos

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



[http://www.demartek.com/Demartek\\_Tutorial\\_Video.html](http://www.demartek.com/Demartek_Tutorial_Video.html)

# 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



[http://www.demartek.com/Demartek\\_Interface\\_Comparison.html](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 ( $\mu\text{s}$ )
  - 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

# 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



# 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](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

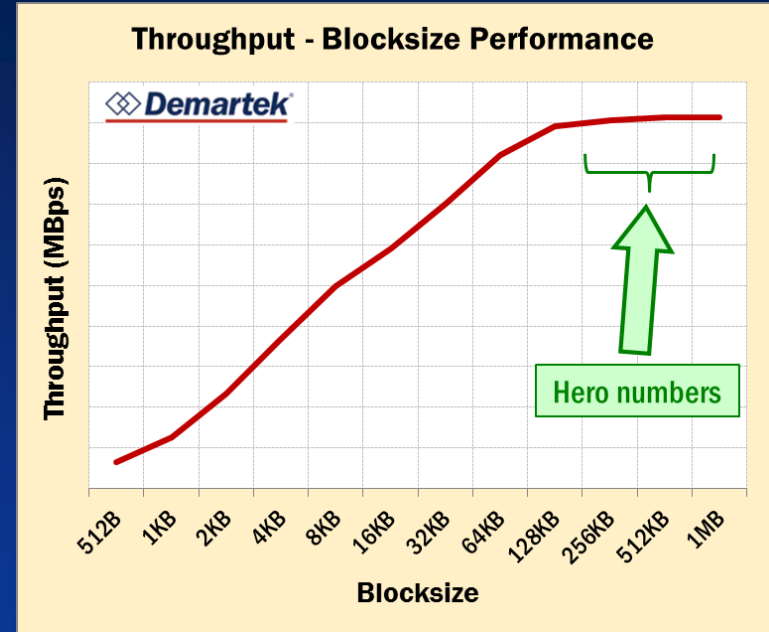
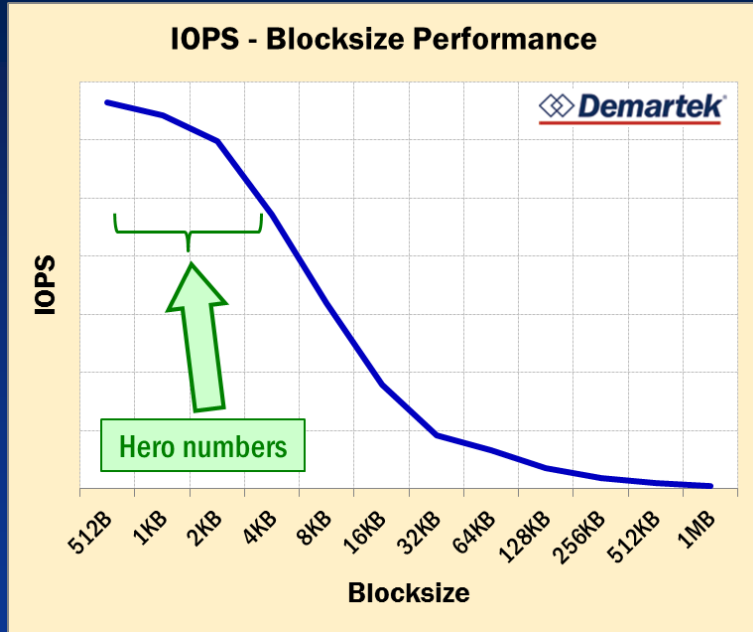
# Real-world Workload Types

- ◆ 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*



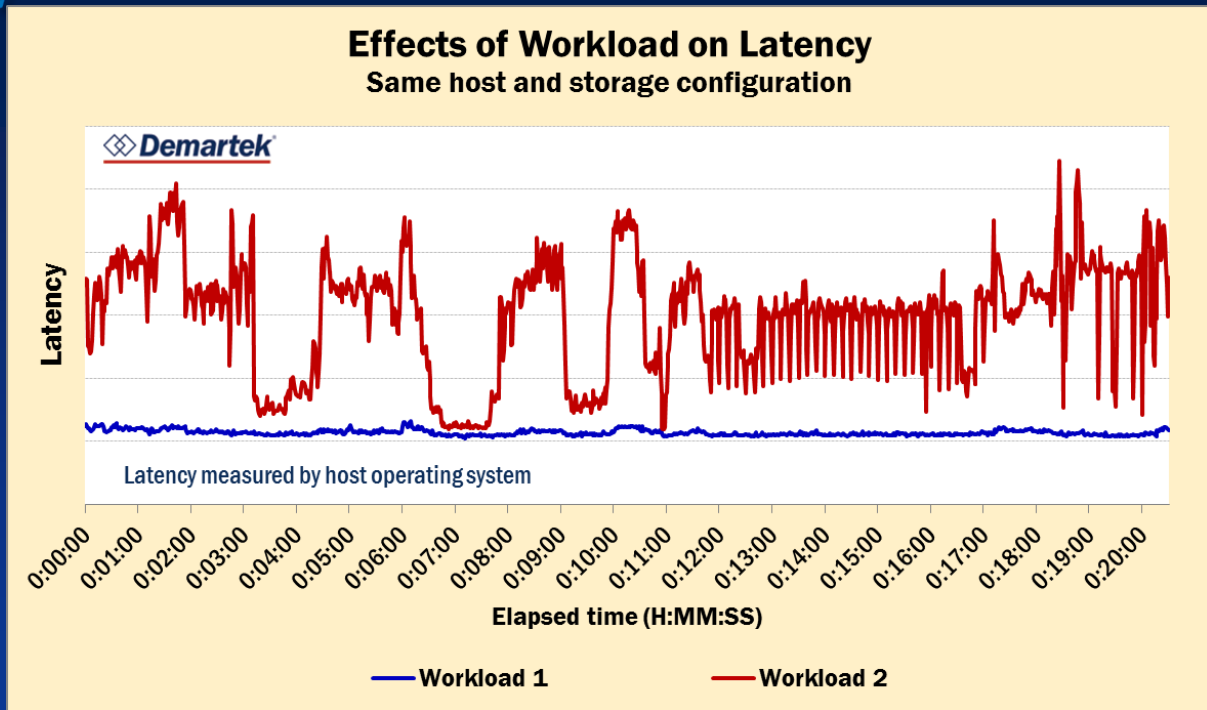
# Performance Results

# Generic IOPS and Throughput Results



These performance curves generally apply to network and storage performance

# Generic Latency Results

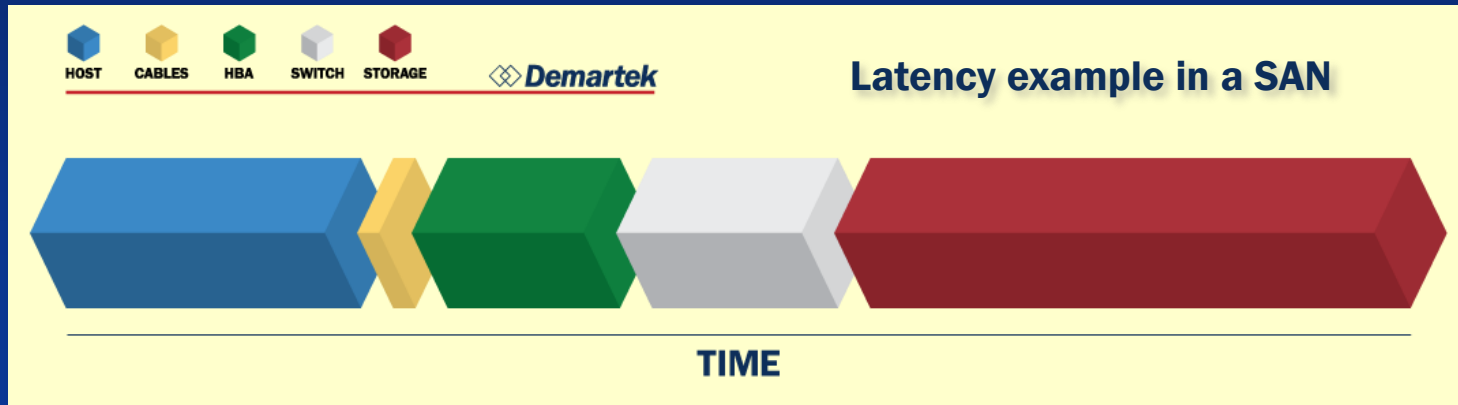


The nature of each workload has a large impact on latency

# 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*





# Two Different Systems

► Which has better performance?



Mario Kart is a trademark of Nintendo

## Purpose of Tests (Need Answers)

- ◆ Flash allows more work to be completed in less time
  - What is the effect on host CPU utilization (physical and virtual)?
  - What is the effect of server memory on storage performance?
- ◆ How do caching and tiering compare?
- ◆ What are some real-world workload block sizes?
- ◆ Can flash storage handle multiple workloads?
- ◆ Different workload settings were used for each set of tests
- ◆ Most of the measurements were taken at the *host server*

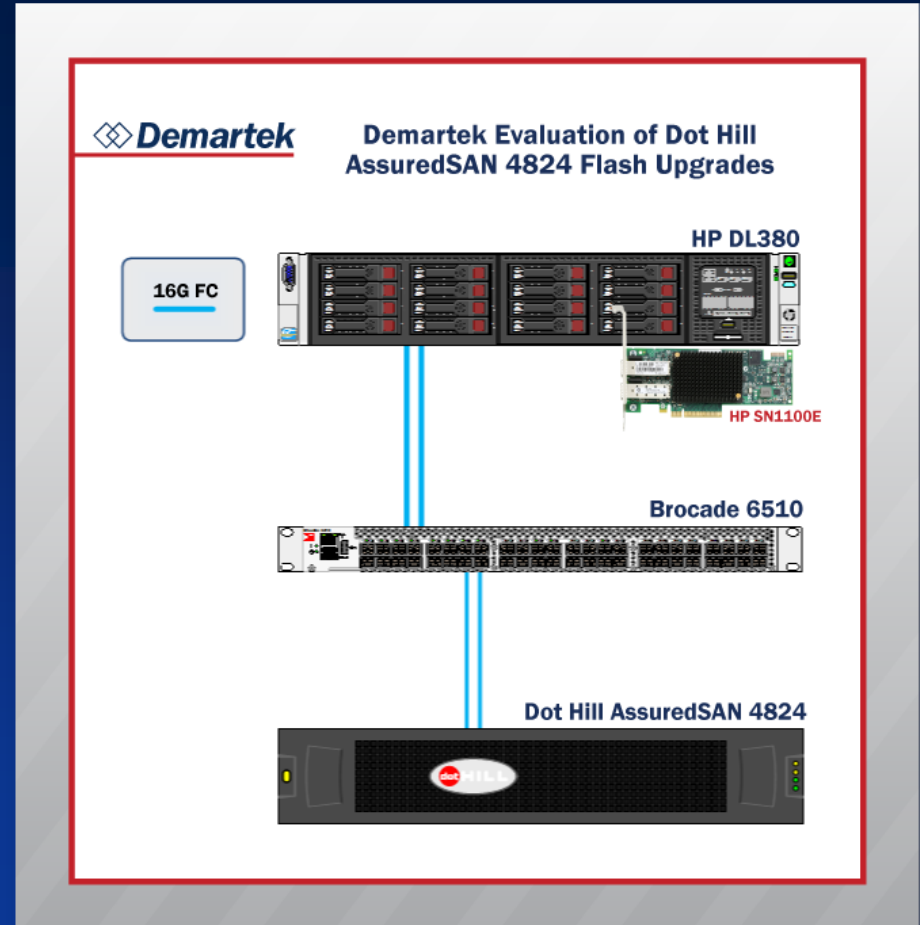
# General Notes on These Tests

- ◆ **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

# Flash Products Tested

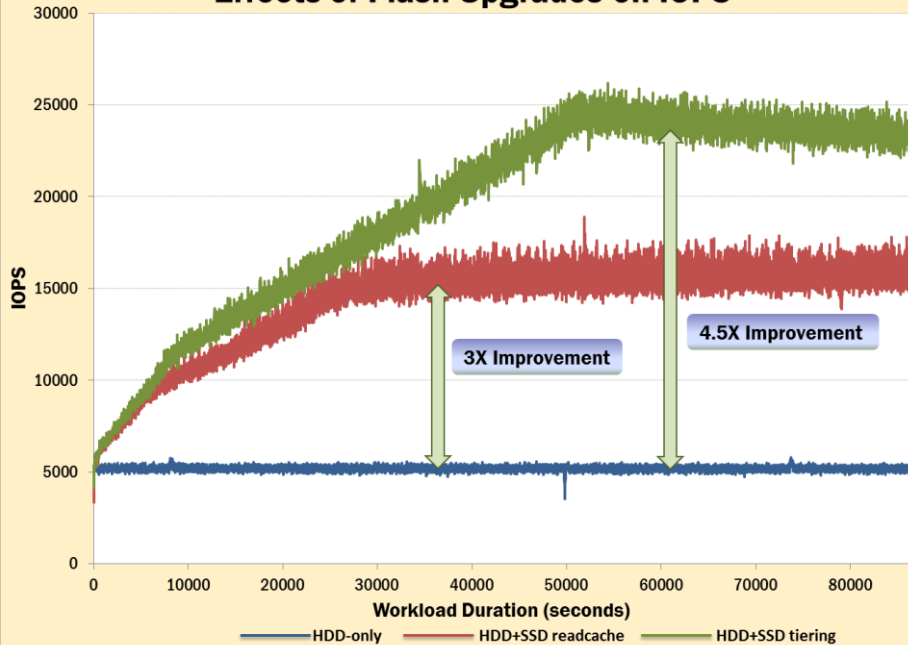
- ◆ Dot Hill hybrid storage array (caching and tiering)
- ◆ PMC-Sierra NVMe PCIe card vs. SATA SSDs
- ◆ Samsung SM-1715 NVMe PCIe cards (four)
- ◆ Kaminario K2 all-flash array (two mixed workloads)
- ◆ Violin Memory 7300 all-flash array (four mixed workloads)

- ◆ **Dot Hill AssuredSAN 4824**
  - 20x 900GB 10K RPM HDDs
  - 4x 400GB SSDs
- ◆ **Three configurations**
  - HDD only: No SSDs used
  - Two SSDs used as a cache
  - Four SSDs used as a tier
- ◆ **Microsoft SQL Server OLTP**

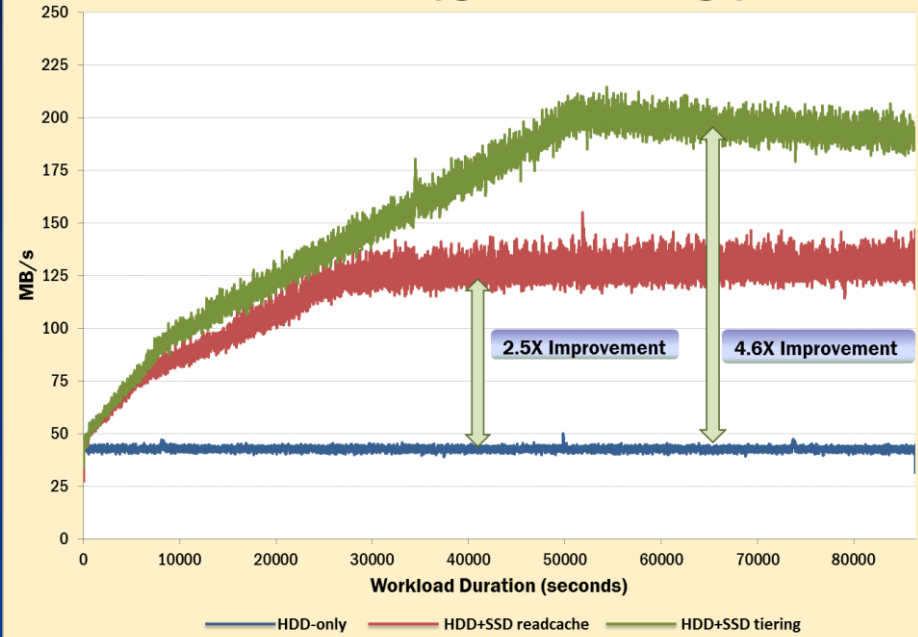


# IOPS & Throughput

### Effects of Flash Upgrades on IOPS



### Effects of Flash Upgrades on Throughput



# Latency and Transactions per Second

### Effects of Flash Upgrades on Latency



### Database Transactions per Second



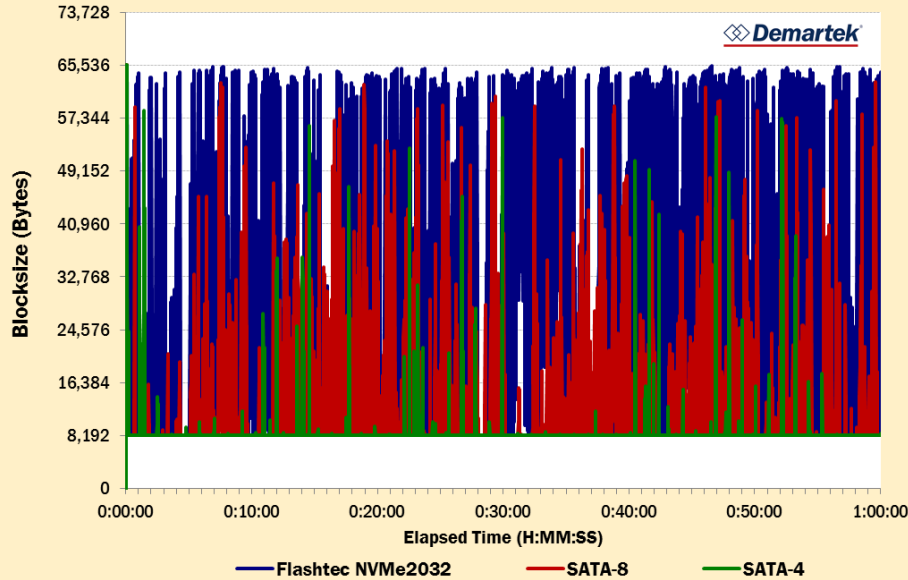
# NVMe SSD vs. SATA SSD (Inside Server)

- ◆ 1x PMC Flashtec NVMe2032 board
- ◆ 8x SanDisk Extreme Pro SSD (among the best SATA SSDs)
- ◆ Single processor, 8 GB RAM
- ◆ Microsoft SQL Server OLTP workload
- ◆ Three configurations:
  - NVMe board configured into four logical volumes
  - 8x SATA SSDs managed by Windows Storage Spaces, four volumes spread across all eight devices
  - 4x SATA SSDs as four individual devices – one volume per device

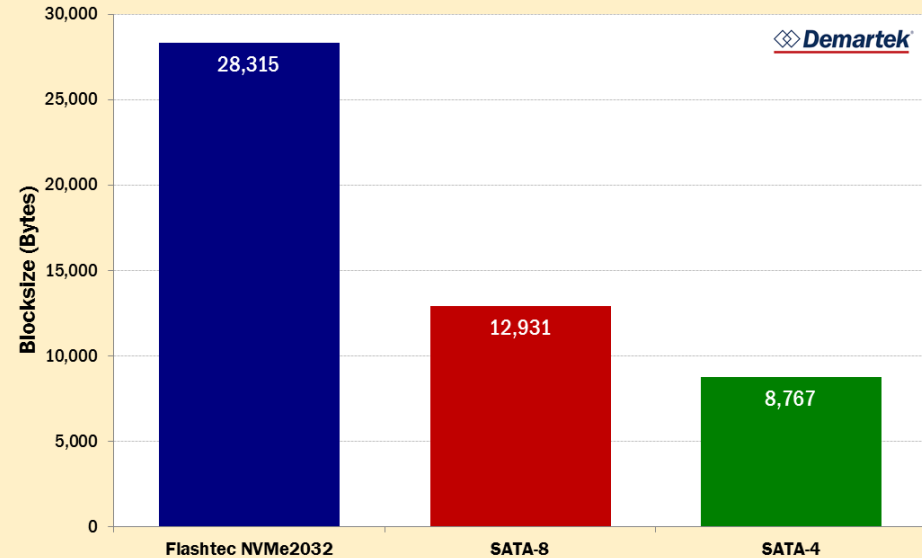


# Workload Block Sizes

Read Blocksize - OLTP Workload



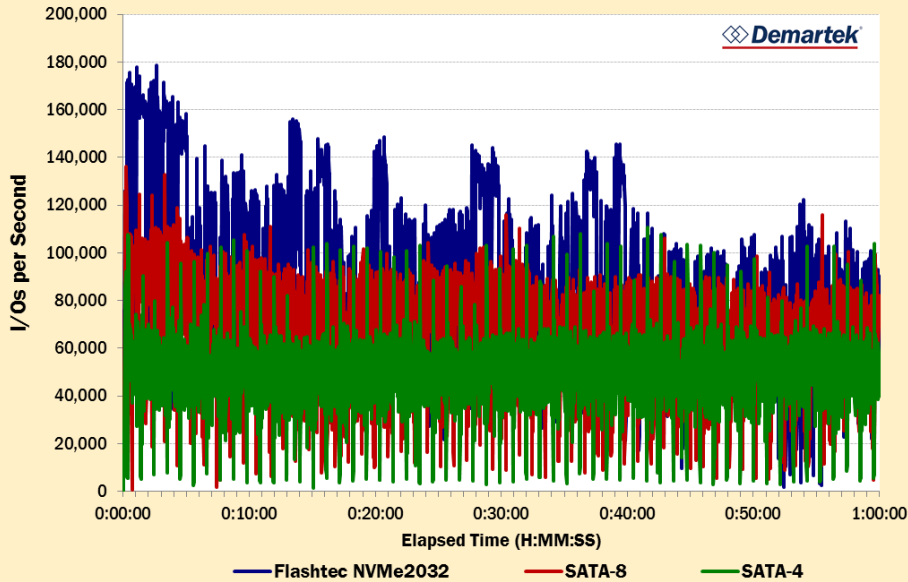
Average Read Blocksize - OLTP Workload  
60 minute run, mixed block sizes



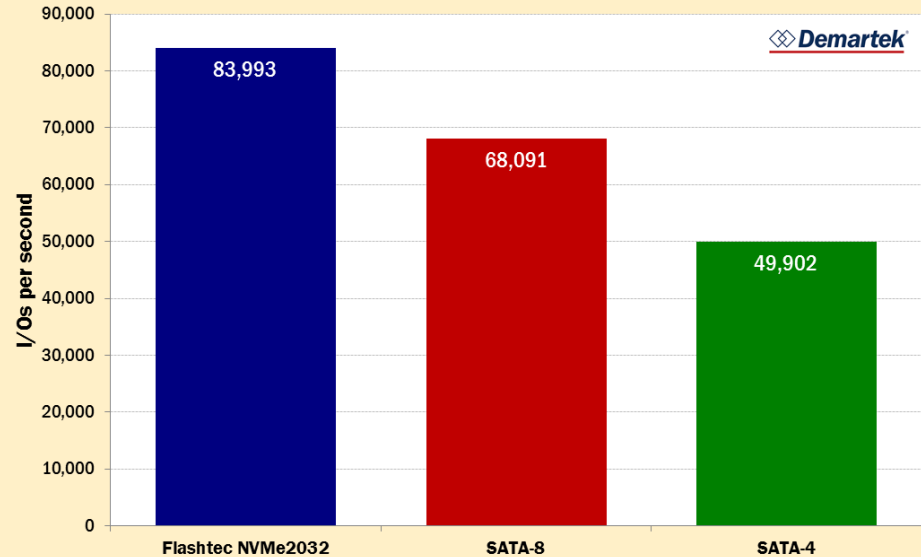
Full report: [http://www.demartek.com/Demartek\\_PMC-Sierra\\_Flashtec\\_NVMe2032\\_Evaluation\\_2015-09.html](http://www.demartek.com/Demartek_PMC-Sierra_Flashtec_NVMe2032_Evaluation_2015-09.html)

# NVMe IOPS

**Read IOPS Comparison - OLTP Workload**



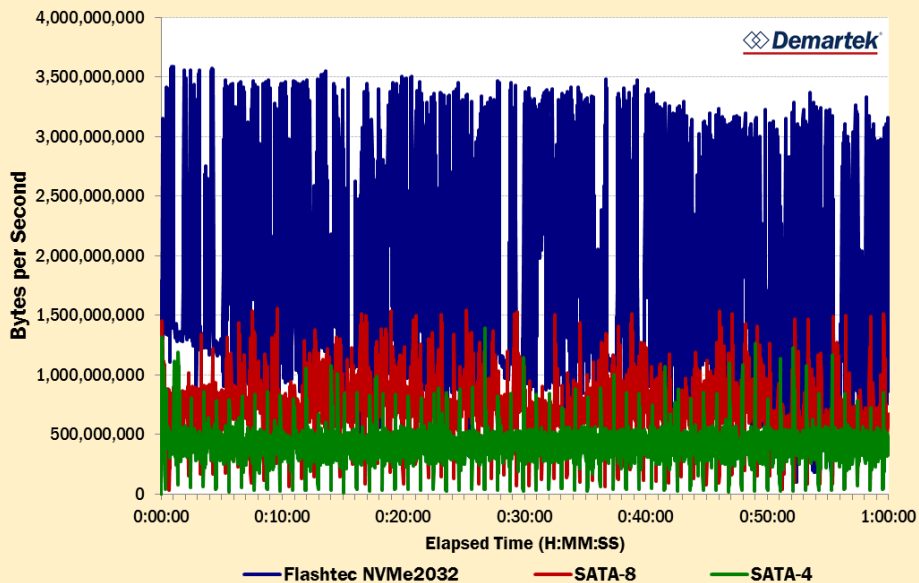
**Average Read IOPS - OLTP Workload**  
60 minute run, mixed block sizes



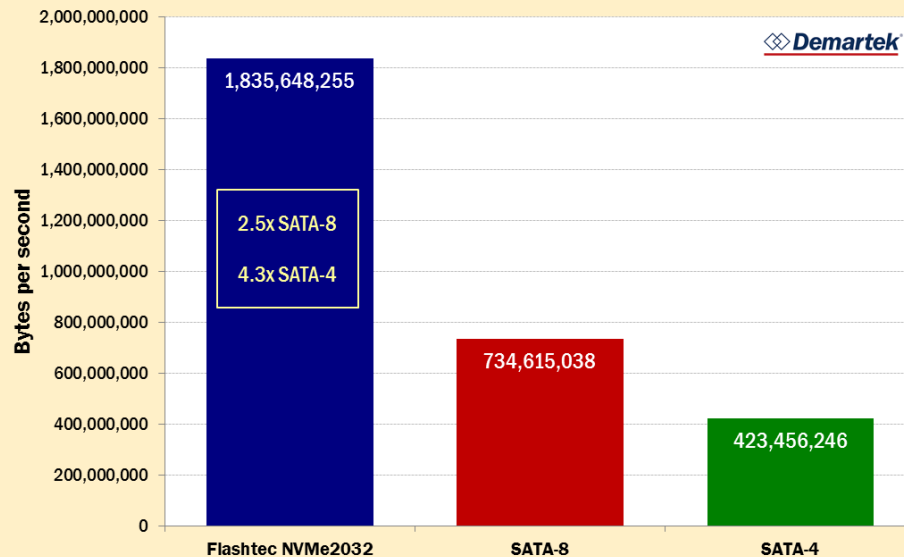
Full report: [http://www.demartek.com/Demartek\\_PMC-Sierra\\_Flashtec\\_NVMe2032\\_Evaluation\\_2015-09.html](http://www.demartek.com/Demartek_PMC-Sierra_Flashtec_NVMe2032_Evaluation_2015-09.html)

# NVMe Throughput

Read Throughput Comparison - OLTP Workload



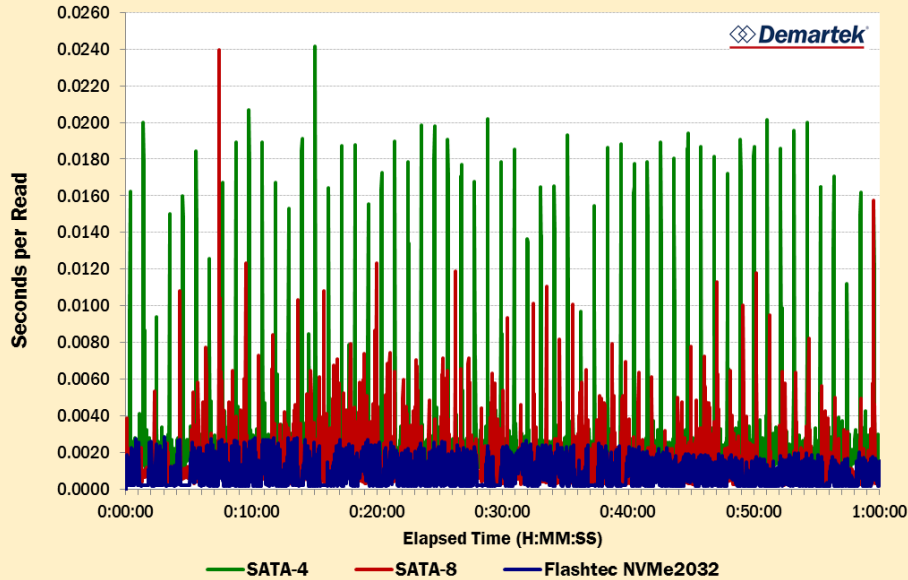
Average Read Throughput - OLTP Workload  
60 minute run, mixed block sizes



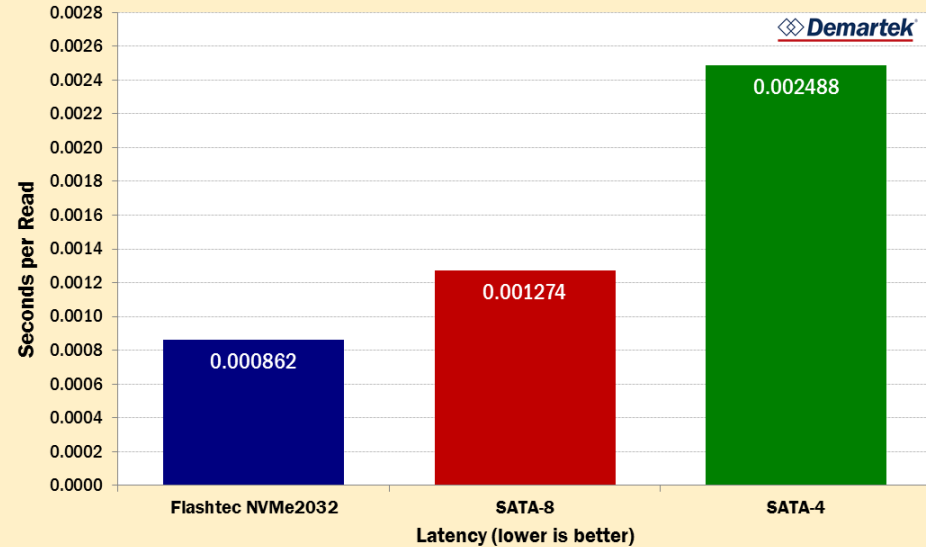
Full report: [http://www.demartek.com/Demartek\\_PMC-Sierra\\_Flashtec\\_NVMe2032\\_Evaluation\\_2015-09.html](http://www.demartek.com/Demartek_PMC-Sierra_Flashtec_NVMe2032_Evaluation_2015-09.html)

# NVMe Latency

Average Read Latency - OLTP Workload



Average Read Latency - OLTP Workload  
60 minute run, mixed block sizes

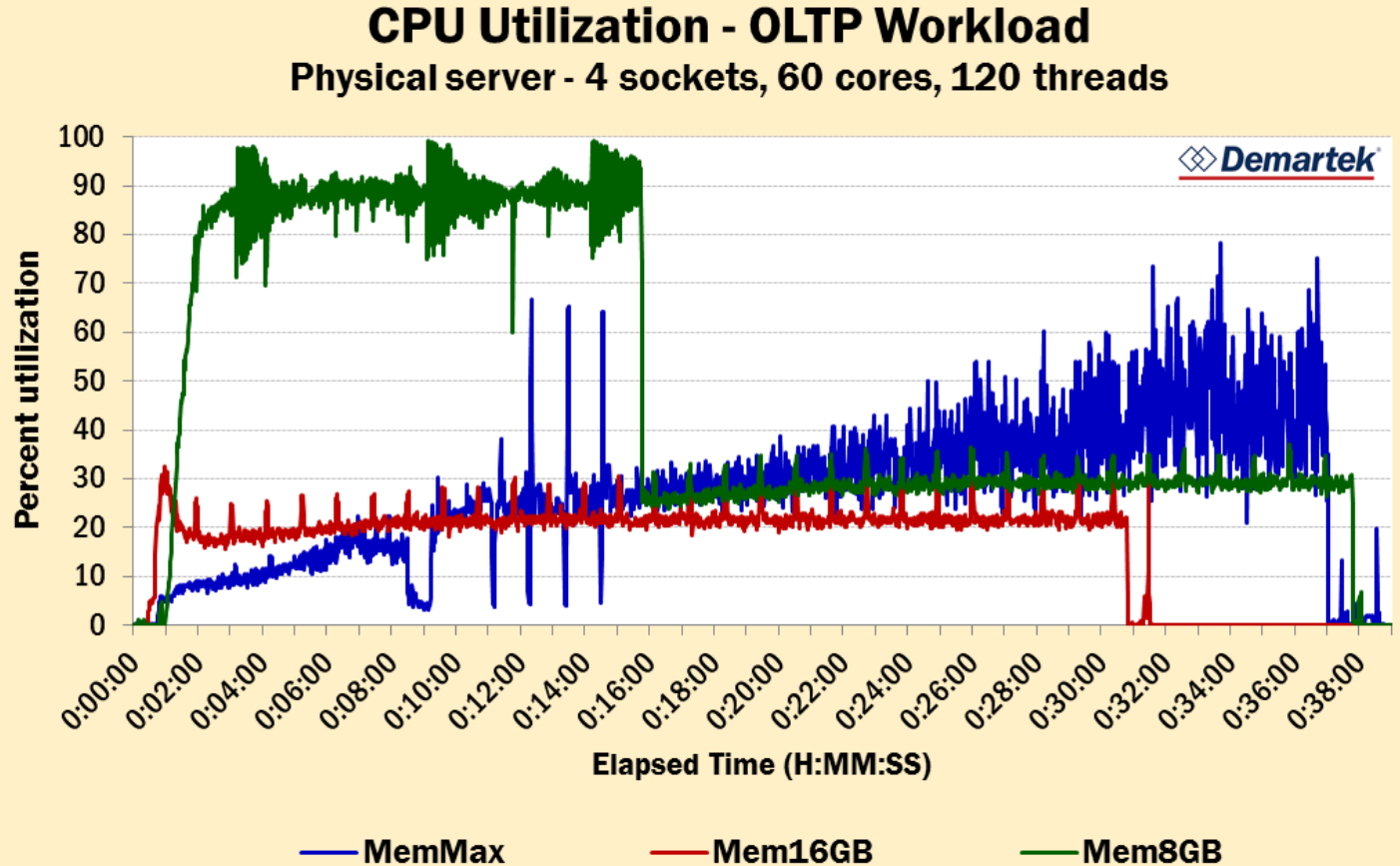


Full report: [http://www.demartek.com/Demartek\\_PMC-Sierra\\_Flashtec\\_NVMe2032\\_Evaluation\\_2015-09.html](http://www.demartek.com/Demartek_PMC-Sierra_Flashtec_NVMe2032_Evaluation_2015-09.html)

# Multiple NVMe cards

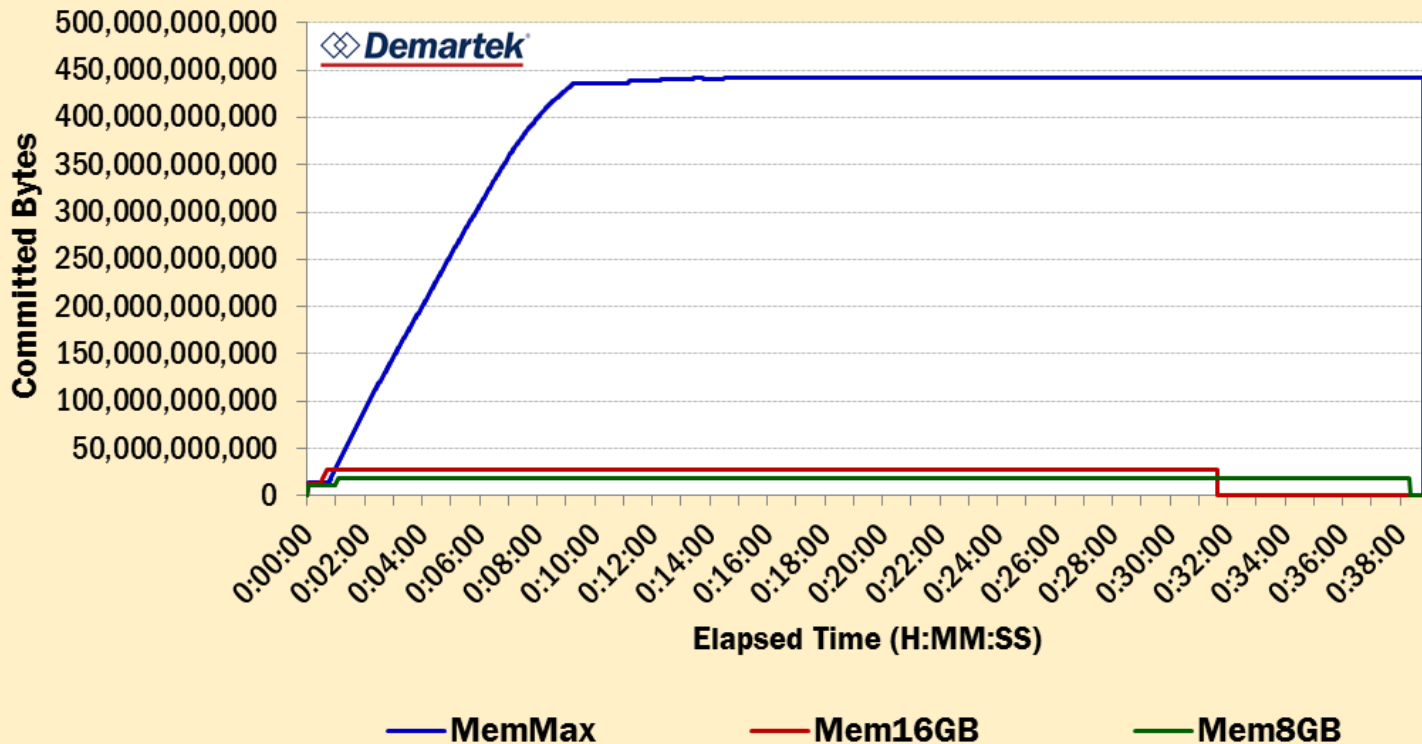
- ◆ **Four Samsung SM-1715 NVMe PCI cards**
  - In-box Windows NVMe drivers
  - 4 LUNs/volumes: allocated one to each NVMe card
- ◆ **Dell PowerEdge R920**
  - 4x Intel Xeon E7-4880 v2, 2.5 GHz, 60 cores, 120 threads
  - 416 GB RAM
- ◆ **Three SQL Server memory settings: max, 16GB & 8GB**
- ◆ **SQL Server OLTP workload**

Limiting RAM  
allocated to  
SQL Server  
affects CPU  
utilization.



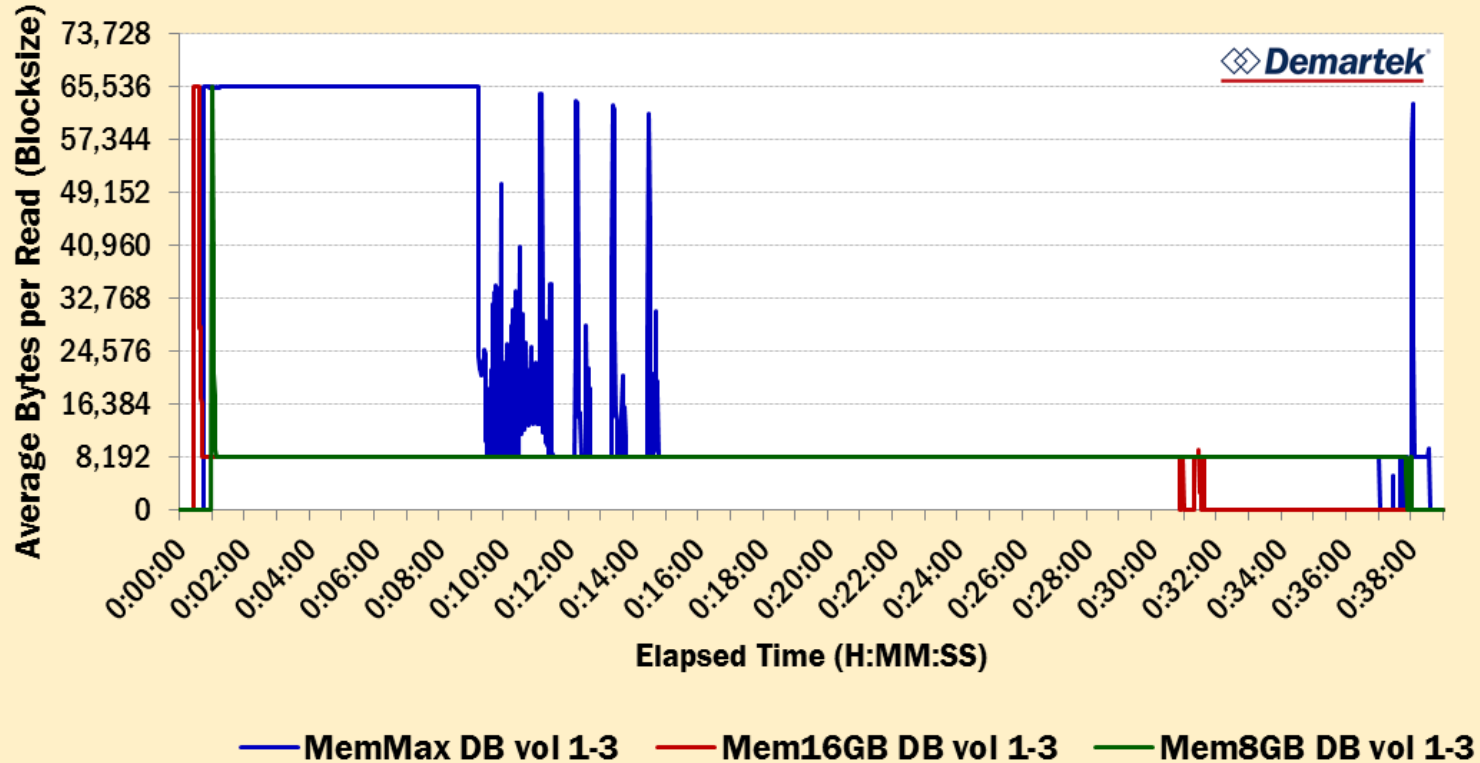
Database applications specifically use RAM to avoid performing I/O.

## Memory Usage OLTP Workload



Bigger RAM buffers means larger block sizes for I/O.

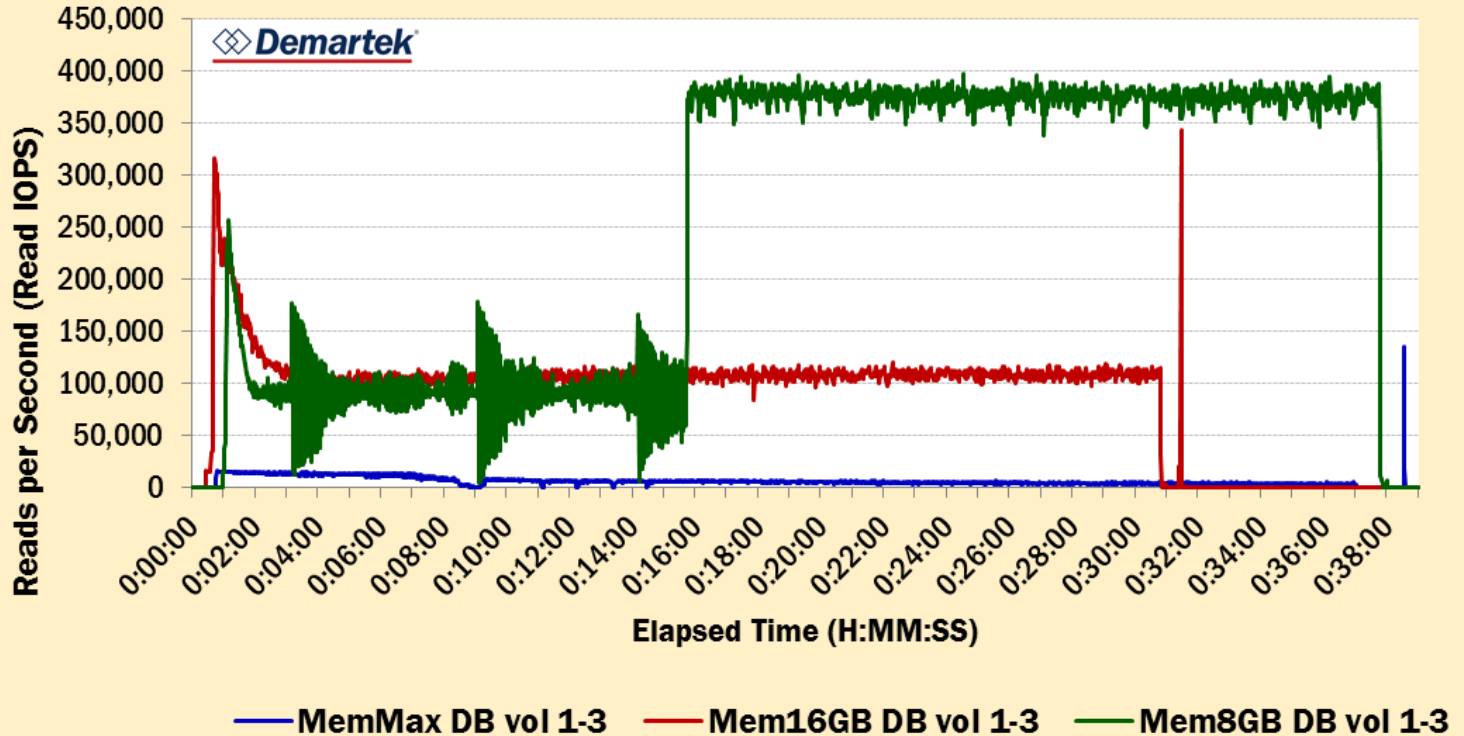
## Database Read Blocksize OLTP Workload





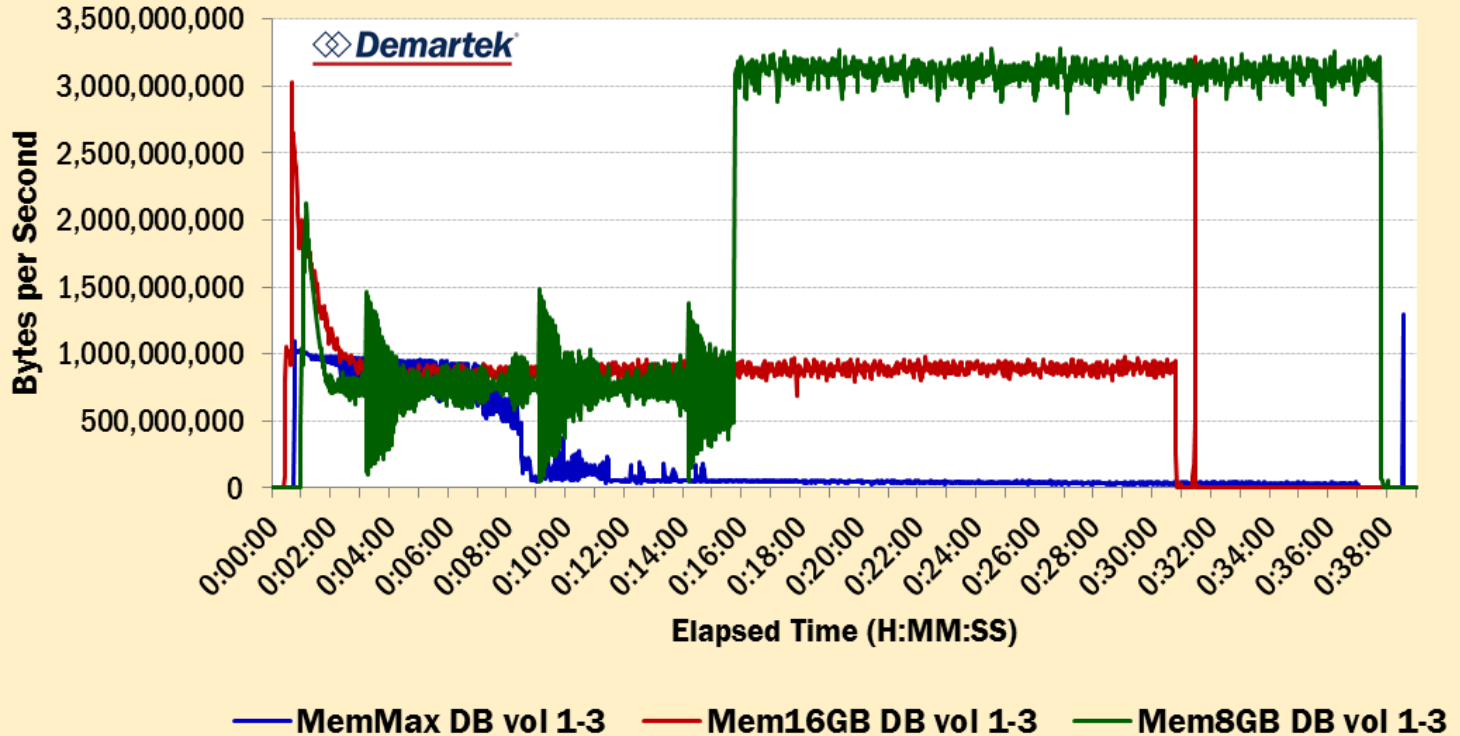
Large  
memory  
means fewer  
I/O  
operations

## Database Read IOPS OLTP Workload



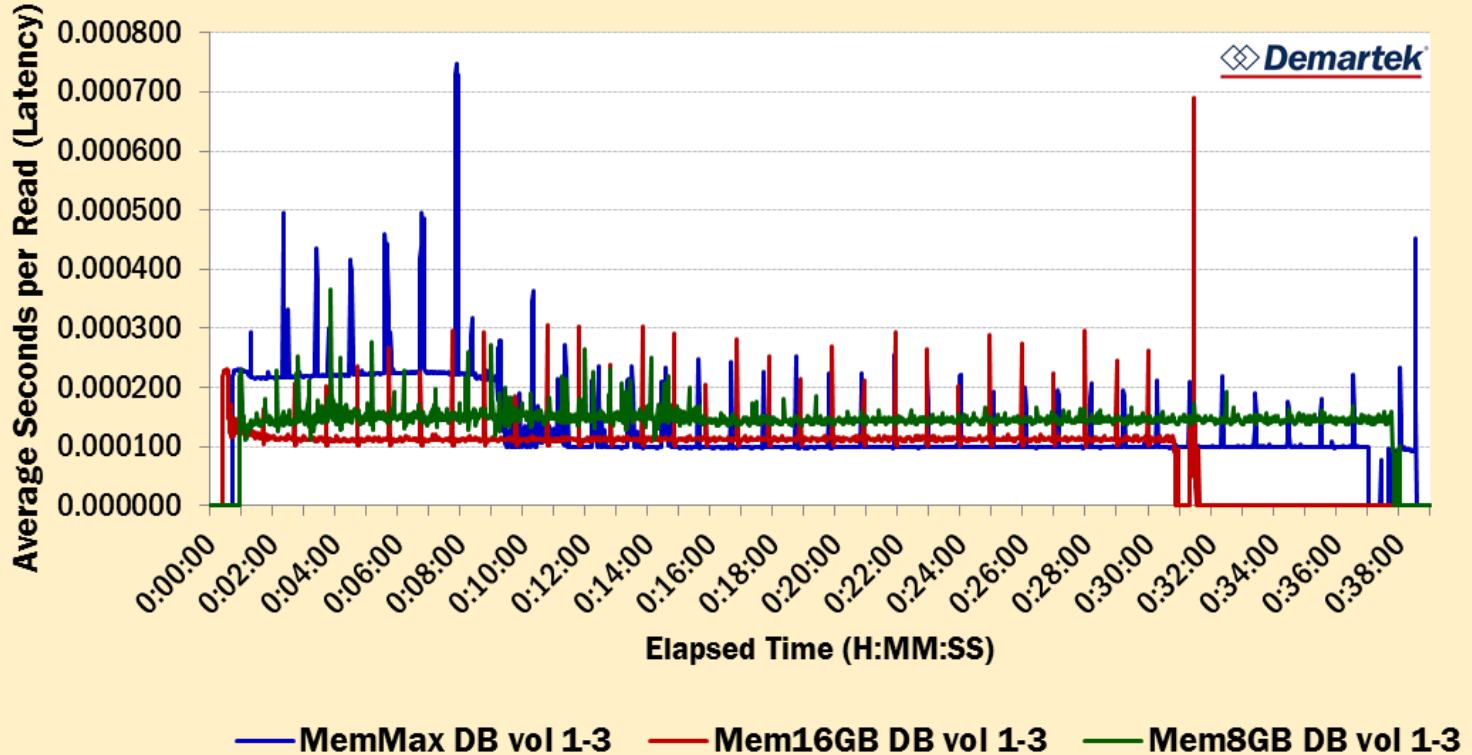
Smaller  
memory  
makes the  
storage work  
harder

## Database Read Throughput OLTP Workload



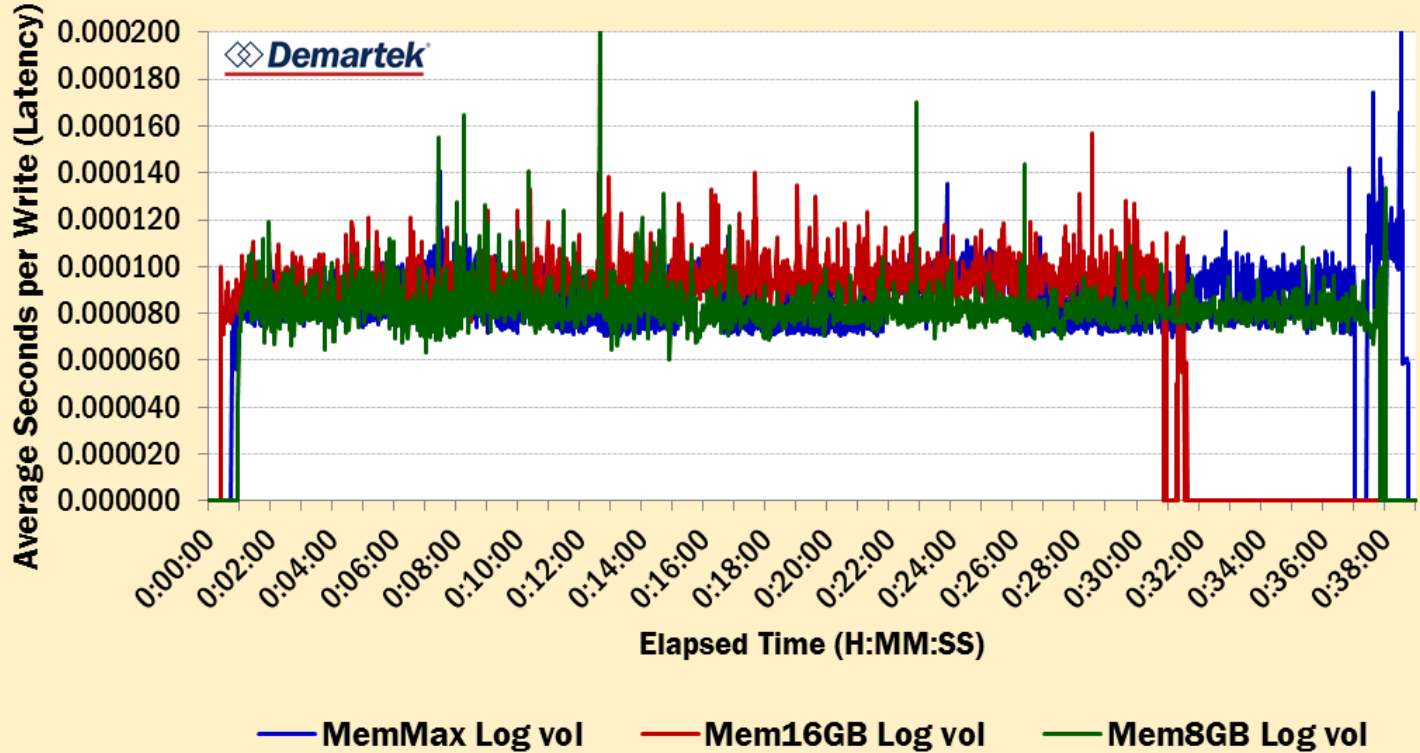
Read latencies approaching 100  $\mu$ s for the Samsung SM-1715 NVMe cards

## Average Database Read Latency OLTP Workload



Write latencies approximately 80  $\mu$ s for the Samsung SM-1715 NVMe cards

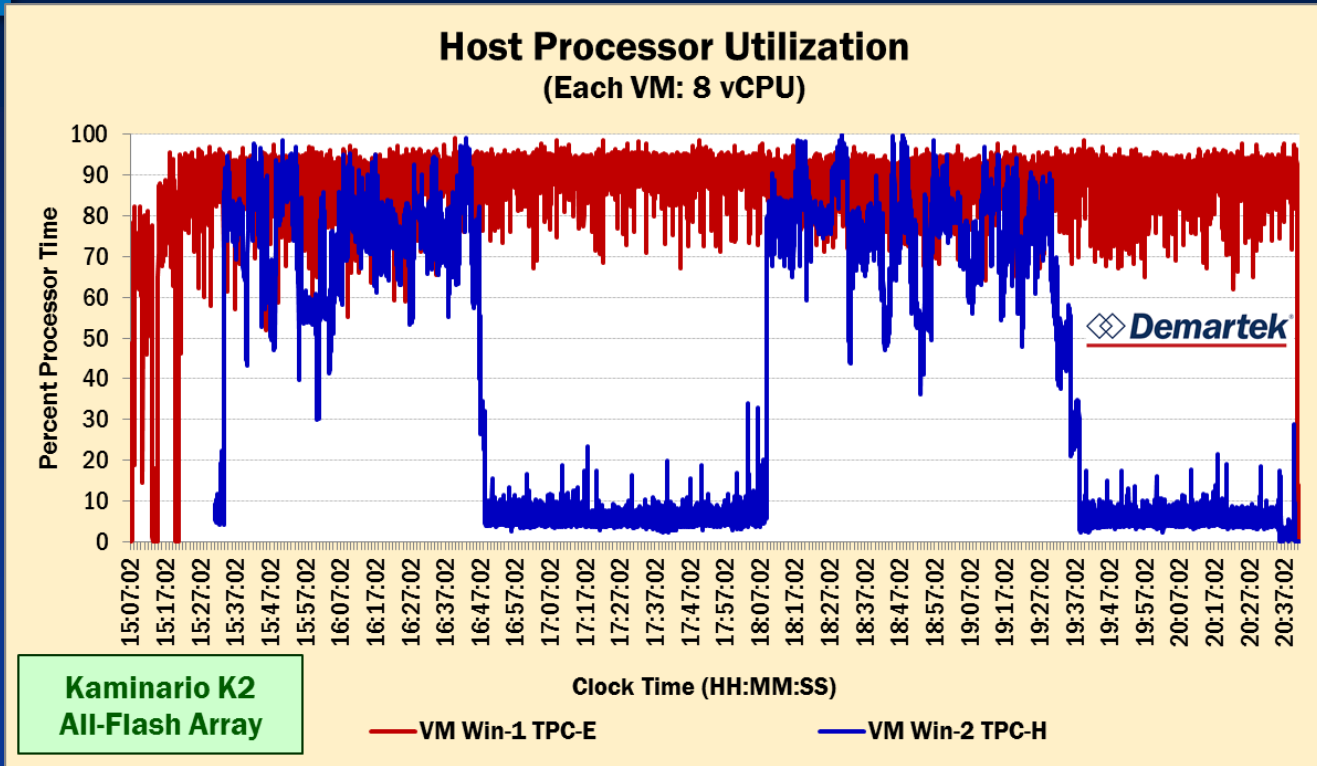
## Average Log Write Latency OLTP Workload



# Two SQL Server Workloads

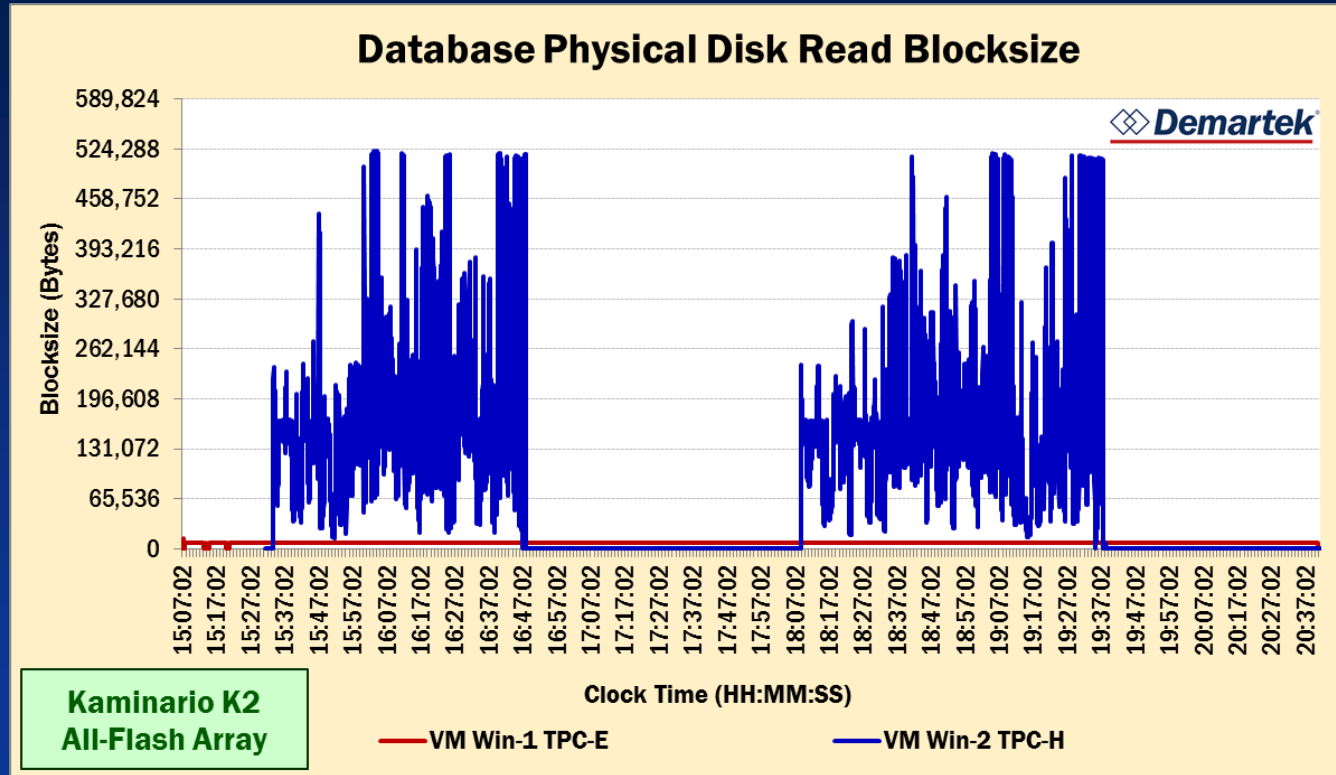
- ◆ Kaminario K2 all-flash array
- ◆ Two workloads from one VMware server simultaneously
  - VM Win-1 (8 vCPU) SQL Server OLTP (TPC-E)
  - VM Win-2 (8 vCPU) SQL Server Data Warehousing (TPC-H)
- ◆ Incremented workloads until host CPUs > 90% utilization
- ◆ 8GFC SAN configuration
- ◆ Four 1TB LUNs - Two for each VM (Database and Log)

# Host CPU Utilization (VMware VMs)



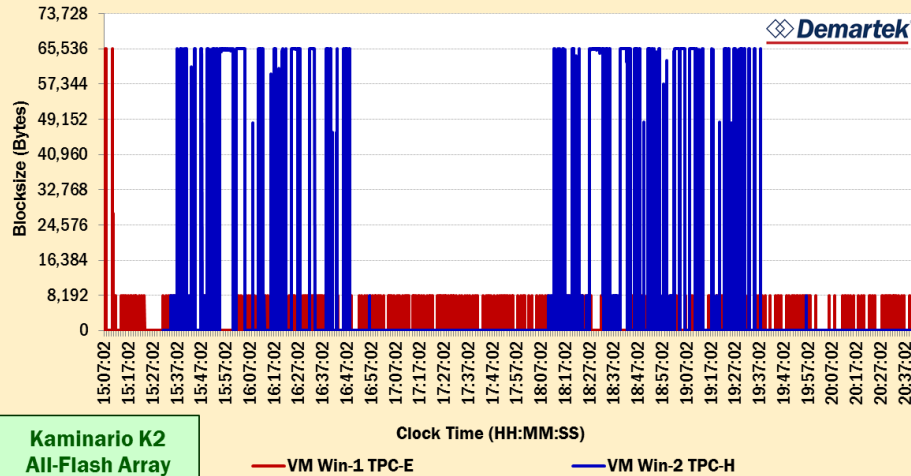
# Database Read Block Size

OLTP workload  
(red line)  
are mostly 8K  
reads while data  
warehousing reads  
are very large  
blocks (blue line)



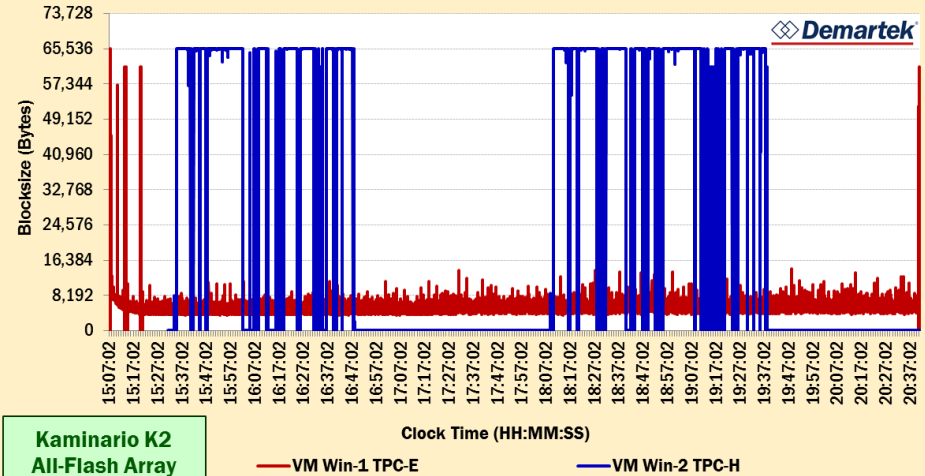
# Log Volume Block Sizes

Log Physical Disk Read Blocksize



Mixture of 8K and 64K Reads

Log Physical Disk Write Blocksize

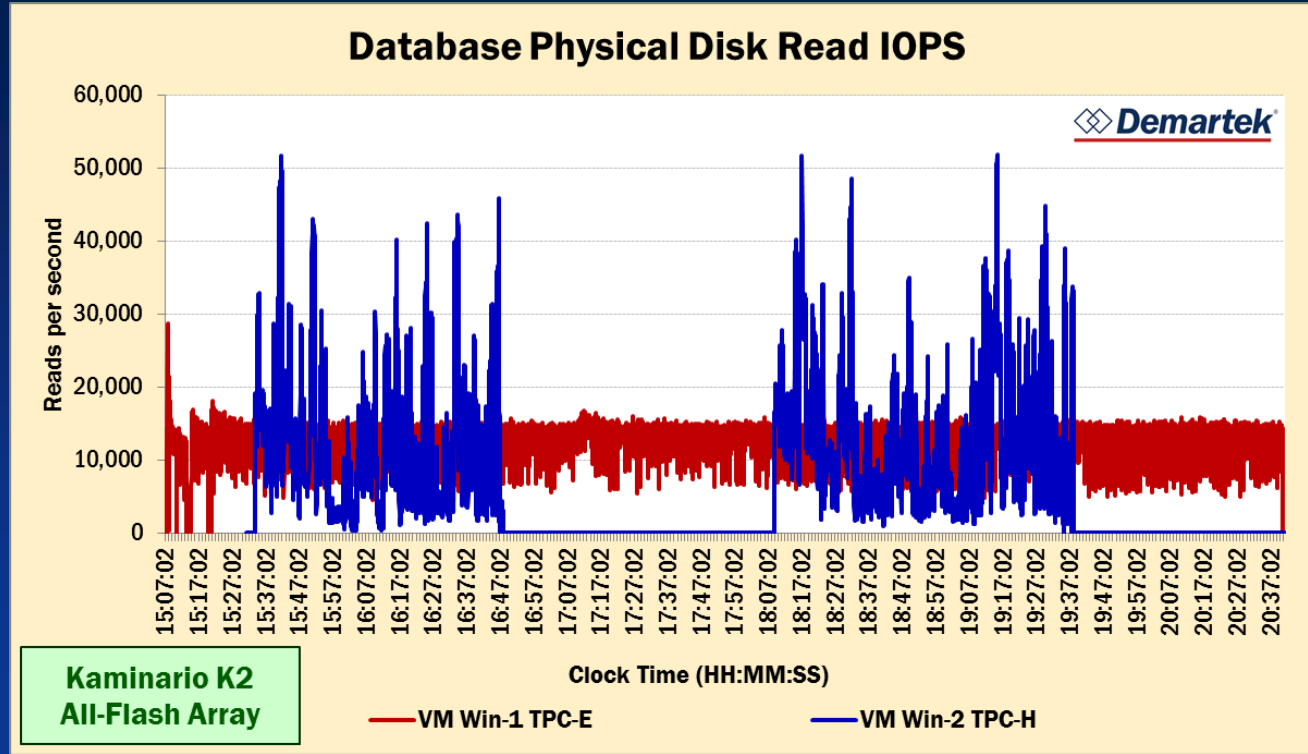


Log writes are small, variable block  
“sequentialish” writes for OLTP workload  
but 64K for data warehousing workload



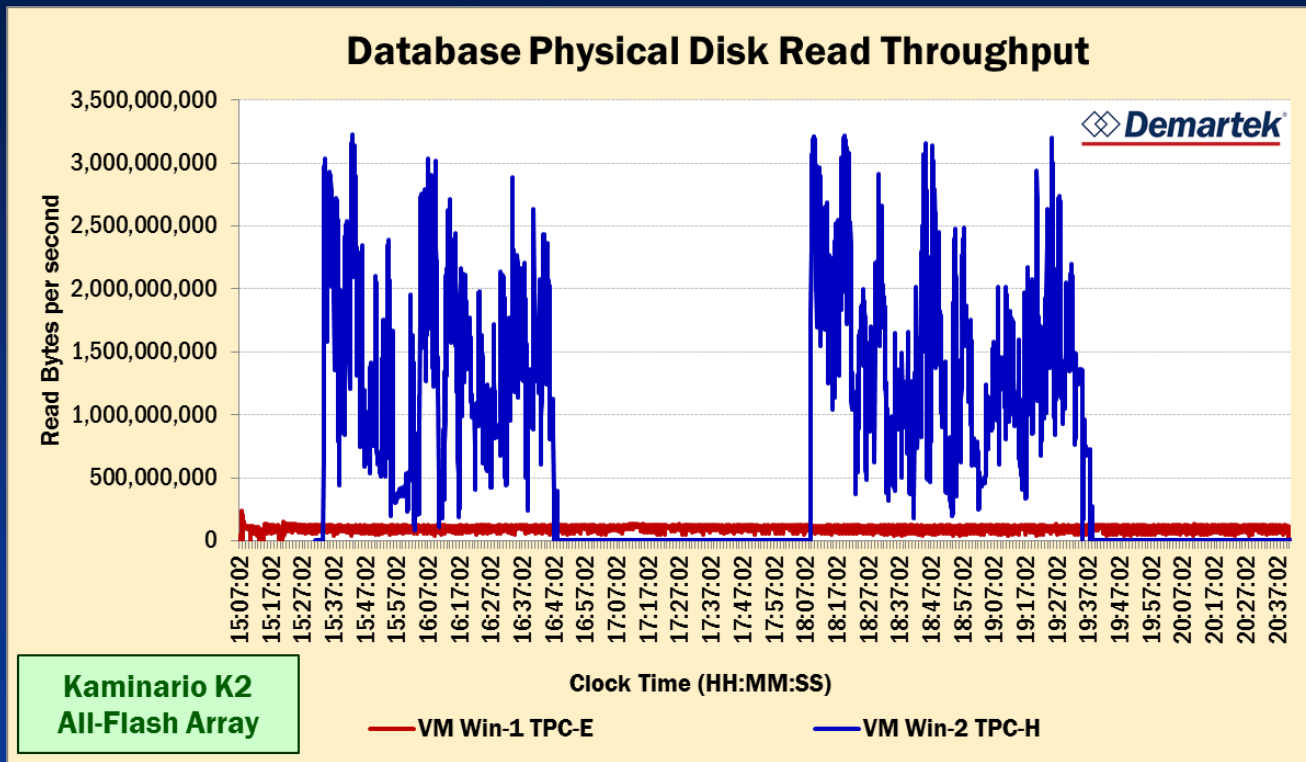
# Database IOPS

IOPS limited by host CPUs (VMs) running at 90%+ utilization



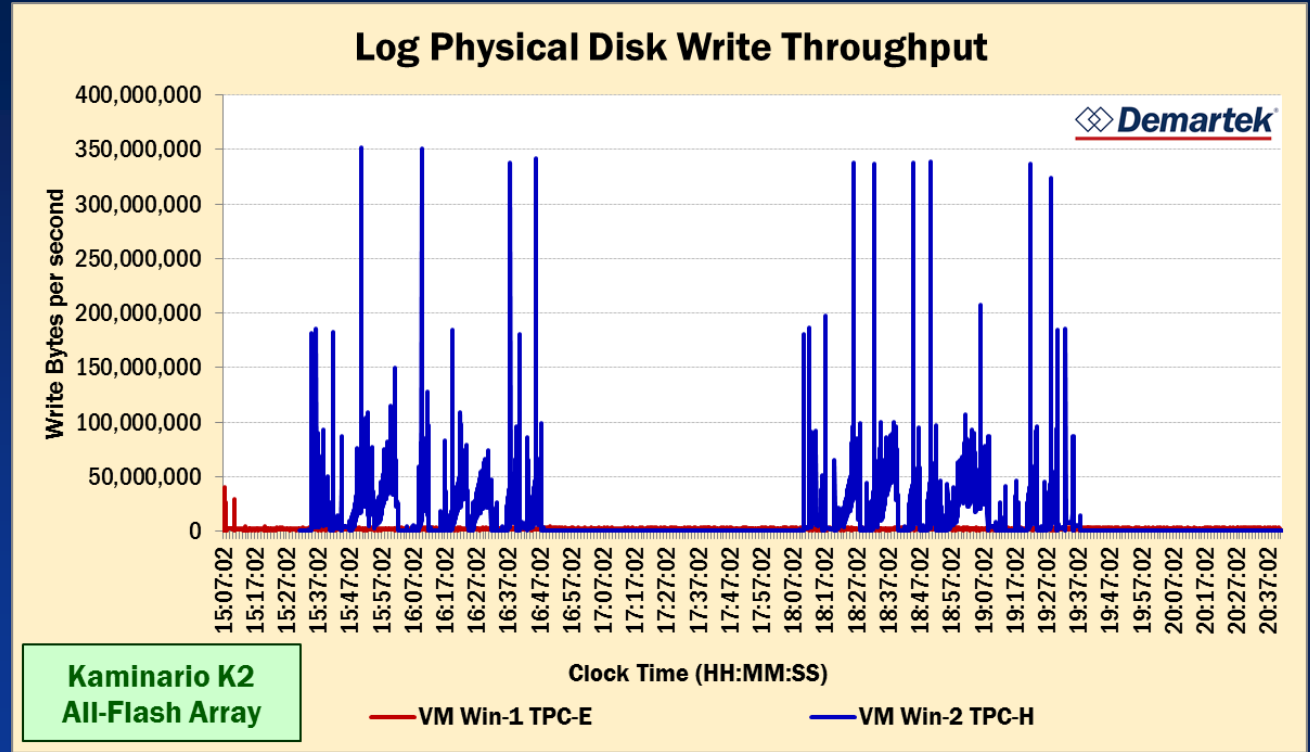
# Database Throughput

Database throughput limited by host CPUs (VMs) running at 90%+ utilization



# Log Write Throughput

Log write activity  
proportional to  
database activity

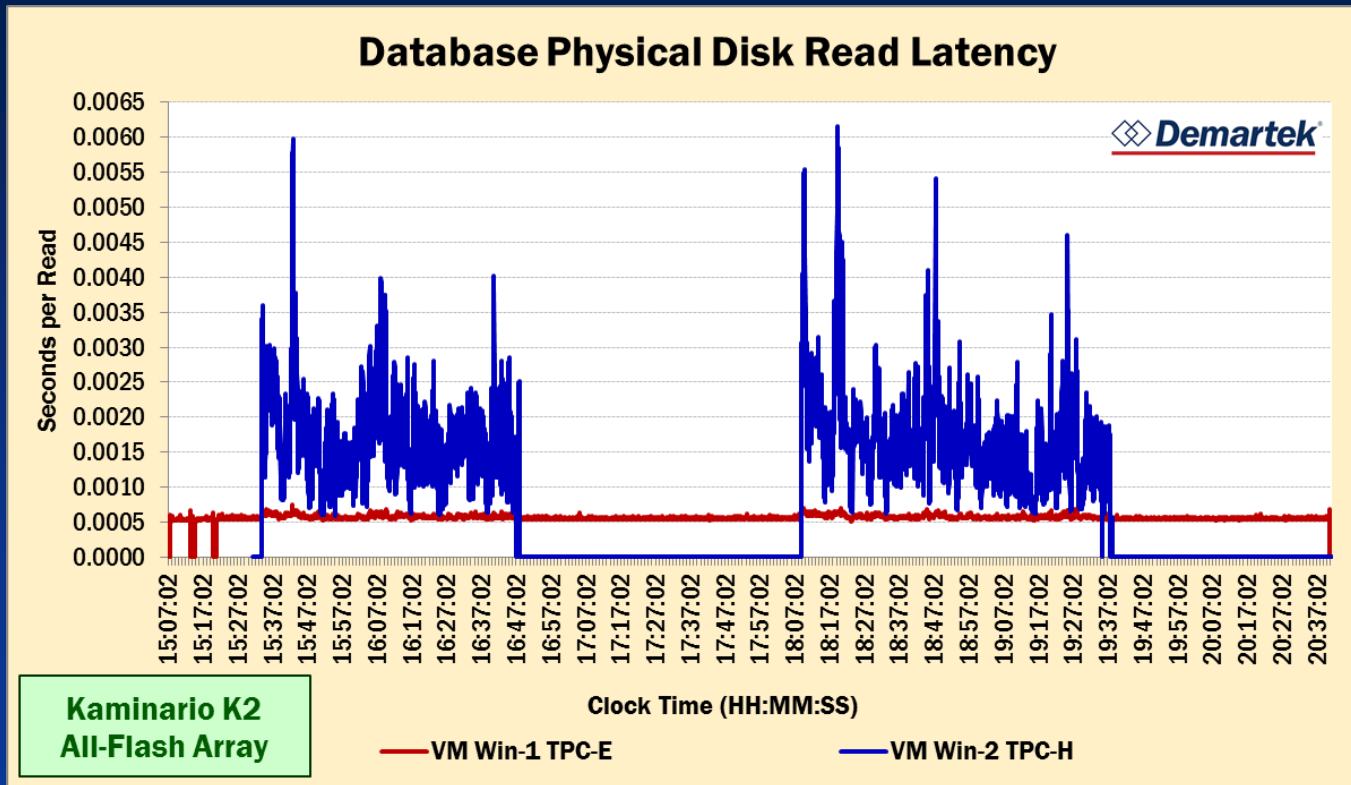


# Database Latency

This data warehousing workload (blue line) is brutal on latency, even for all-flash arrays.

We see similar latency spikes on all the all-flash arrays that we have tested.

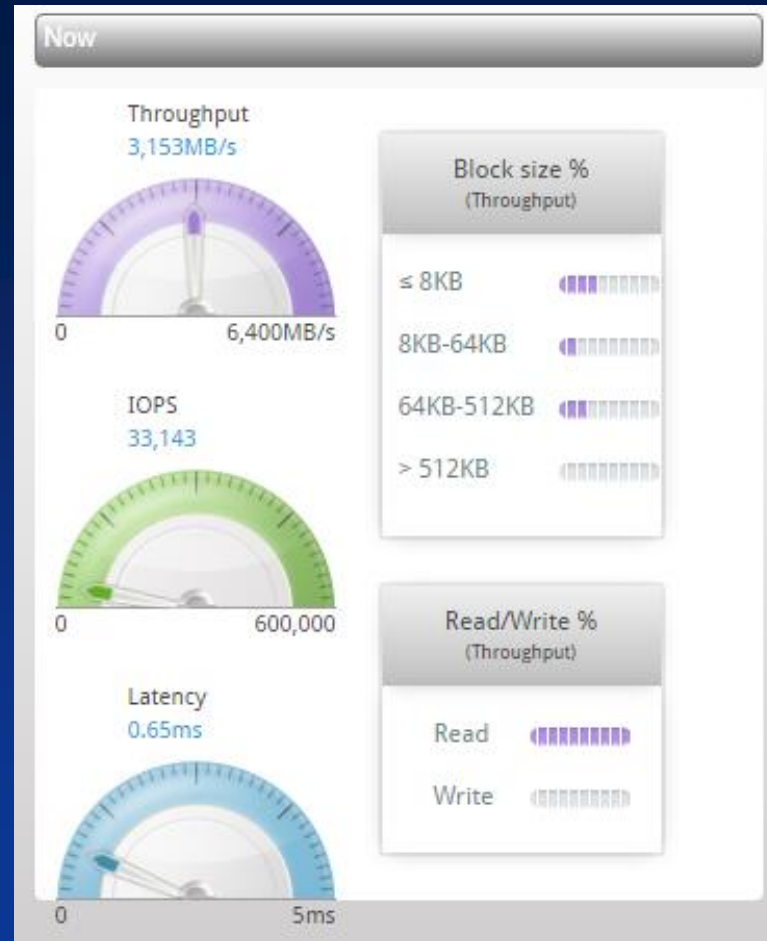
Not all of this latency is due to the storage array.



The location of the measurement matters.

In this screen shot from the Kaminario system, the overall latency is 0.65ms but the latency measured at the host is higher.

Remember the “*Latency Example in the SAN*” on an earlier slide.



# Four Workload Tests

- ◆ Four Workloads on the same all-flash array simultaneously
  1. Web server (four instances)
  2. SQL Server OLTP workload
  3. Exchange Server Jetstress
  4. Microsoft SharePoint
- ◆ Workloads were added incrementally every 15 minutes
- ◆ Full report:

[http://www.demartek.com/Demartek\\_Violin\\_Memory\\_7300\\_FSP\\_Multiple\\_Workloads\\_Evaluation\\_2015-09.html](http://www.demartek.com/Demartek_Violin_Memory_7300_FSP_Multiple_Workloads_Evaluation_2015-09.html)

# Configuration

- ◆ **Two physical servers running Microsoft Windows Hyper-V**
  - Both: 2x Intel Xeon E5-2690 v2, 3.0 GHz, 20 cores, 40 threads
  - Both: 256 GB RAM
  - VMs spread across both servers
- ◆ **16GFC Infrastructure**
- ◆ **Violin Memory 7300 All-flash Array**
  - 35 LUNs/Volumes configured for the four workloads

## Block Sizes

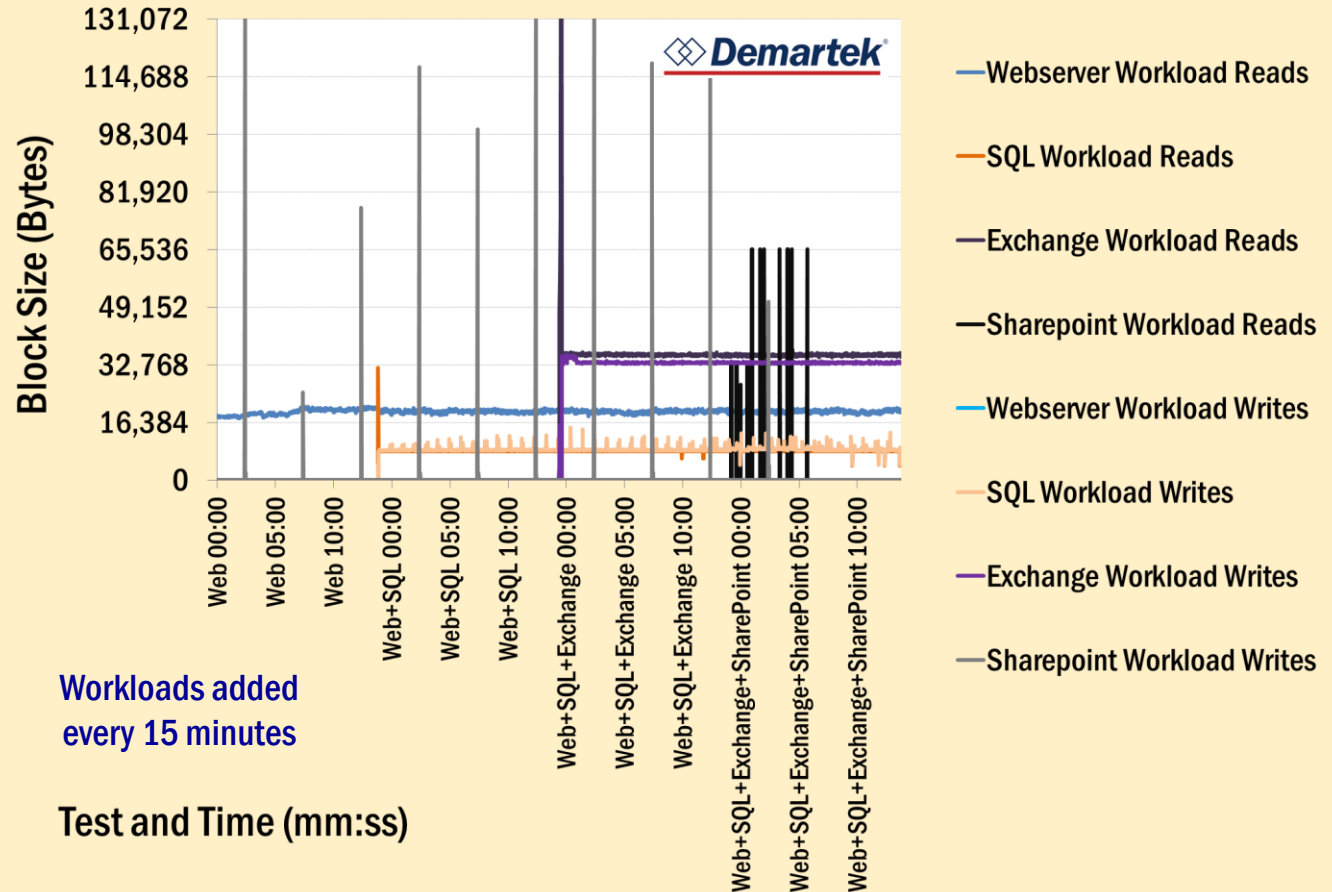
Web ~ 20K

SQL ~ 8K

Exchange ~ 32K

SharePoint ~  
small, 32K, 64K

## Read and Write Block Sizes for 4 Workload Test



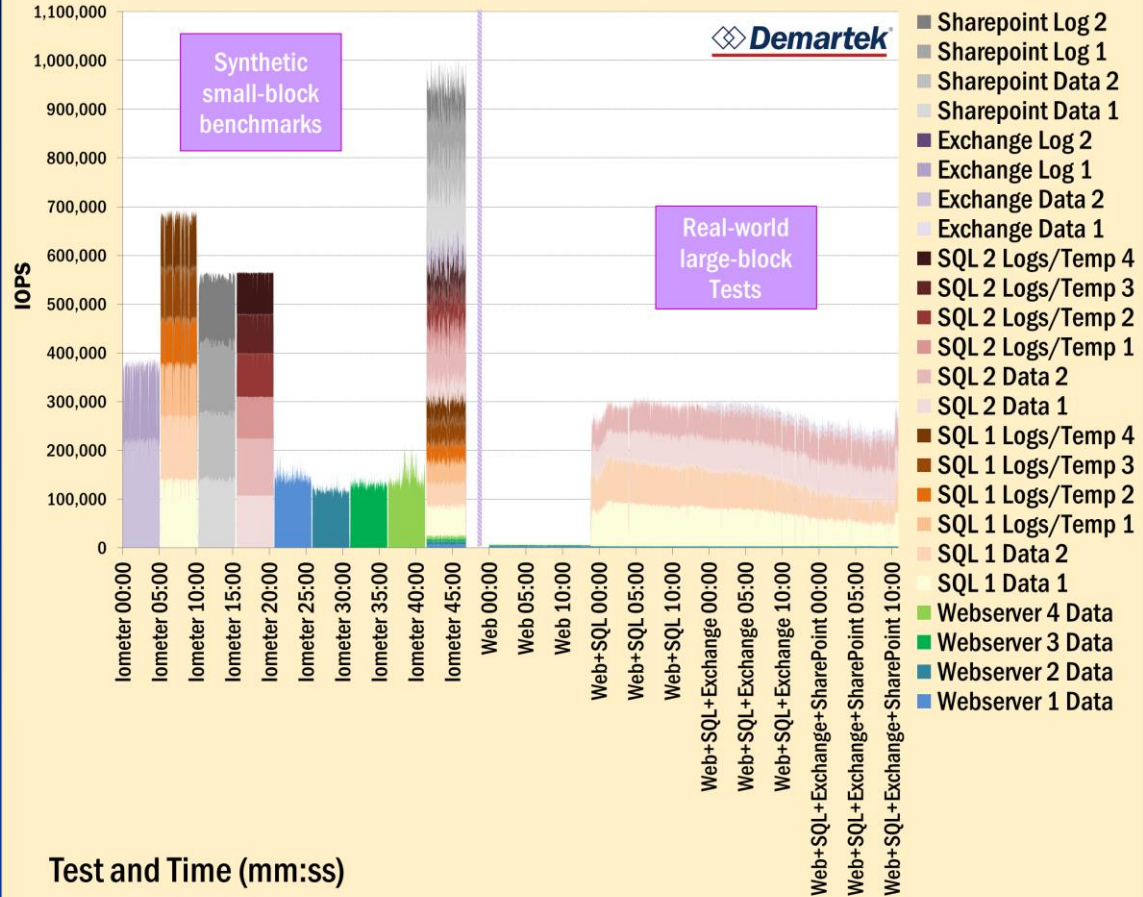




As a sanity check, we first ran a synthetic benchmark on each of the workload LUNs to see what the storage system in our configuration could do.

We achieved almost 1M IOPS.

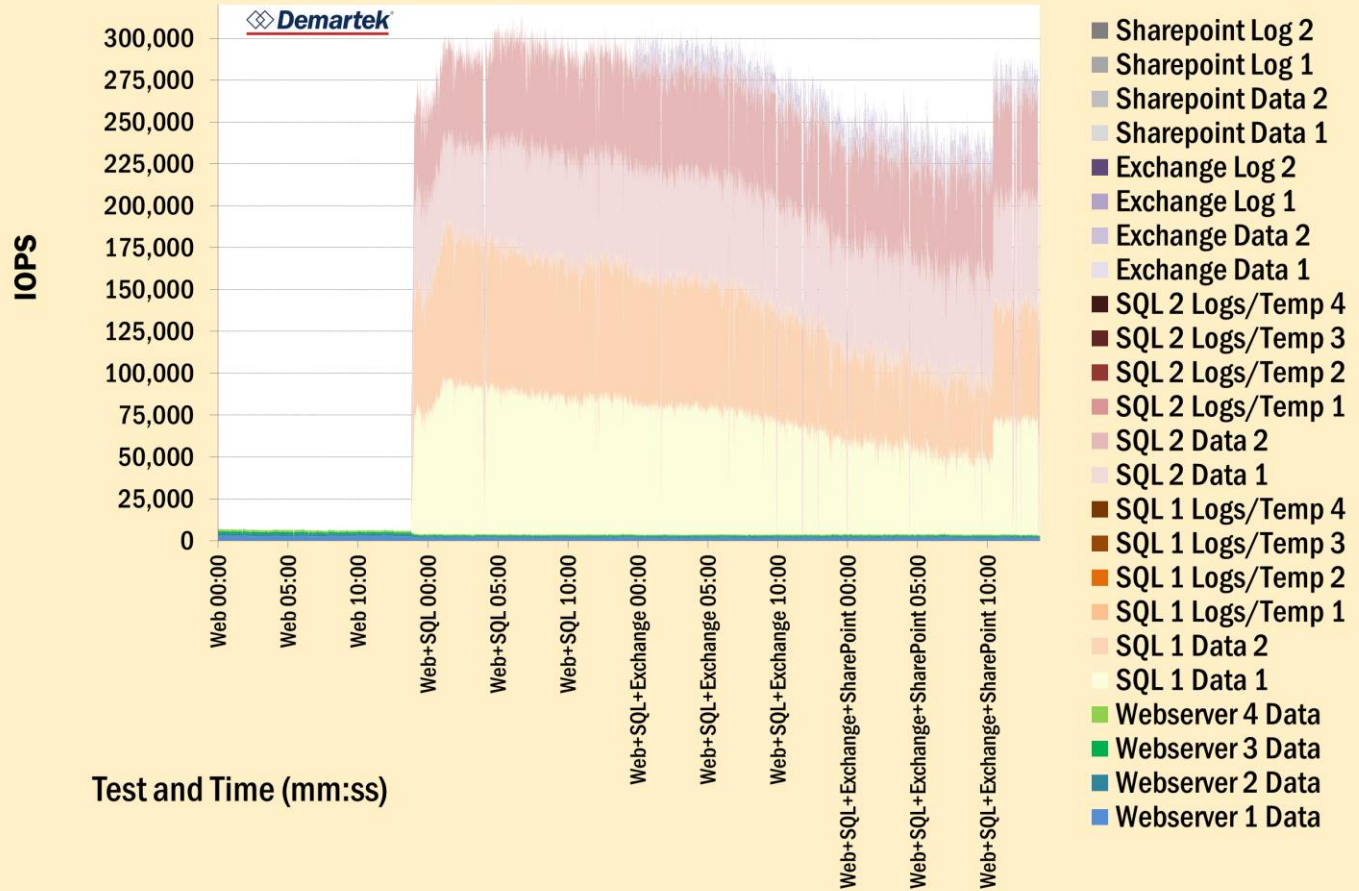
Violin 7300 FSP IOPS for Iometer and 4 Workload Test



Test and Time (mm:ss)

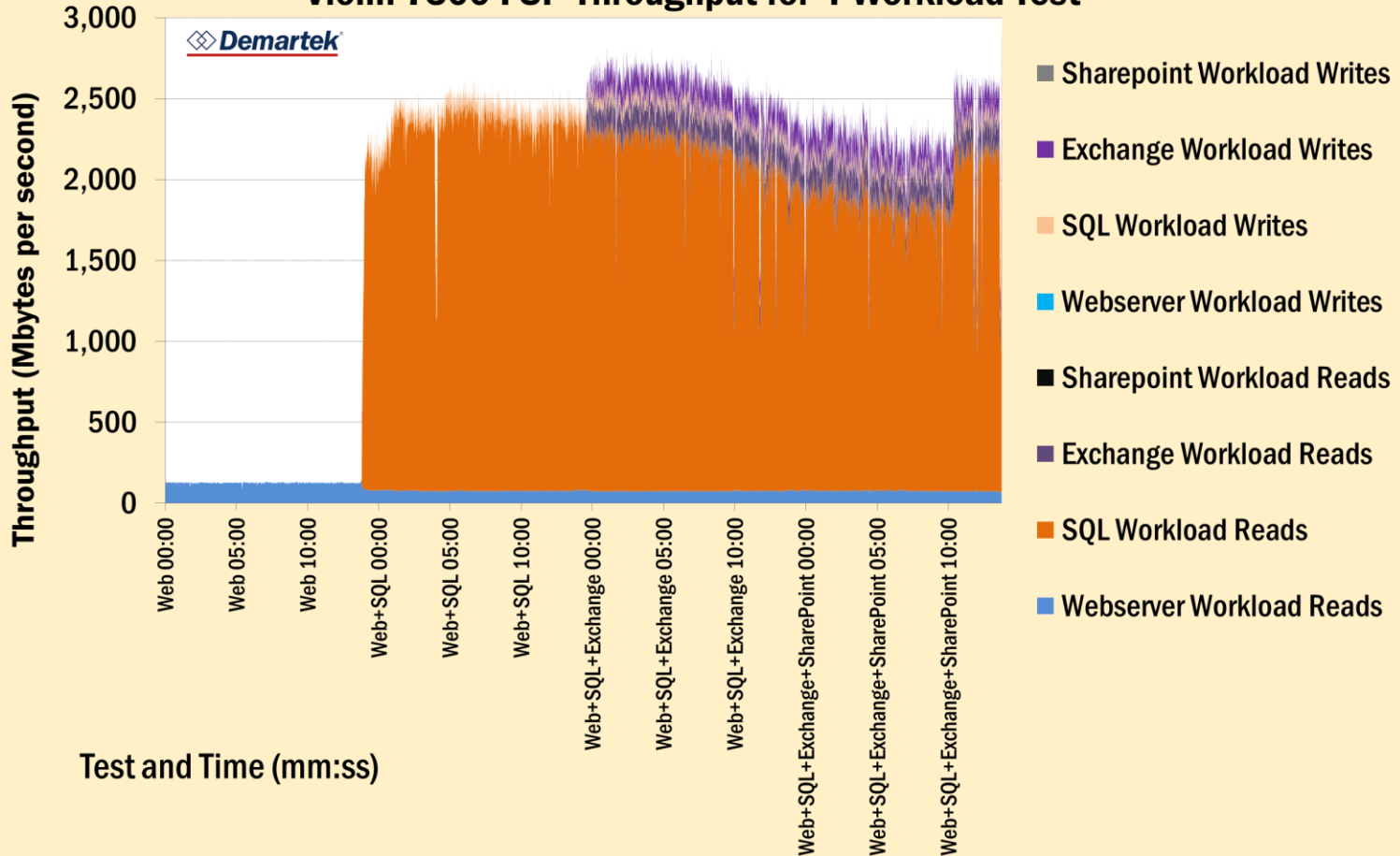
The SQL Server workload was responsible for the majority of the IOPS. It also used the smallest block size, on average, of all the workloads.

### Violin 7300 FSP IOPS for 4 Workload Test



The SQL Server workload was responsible for the majority of the throughput.

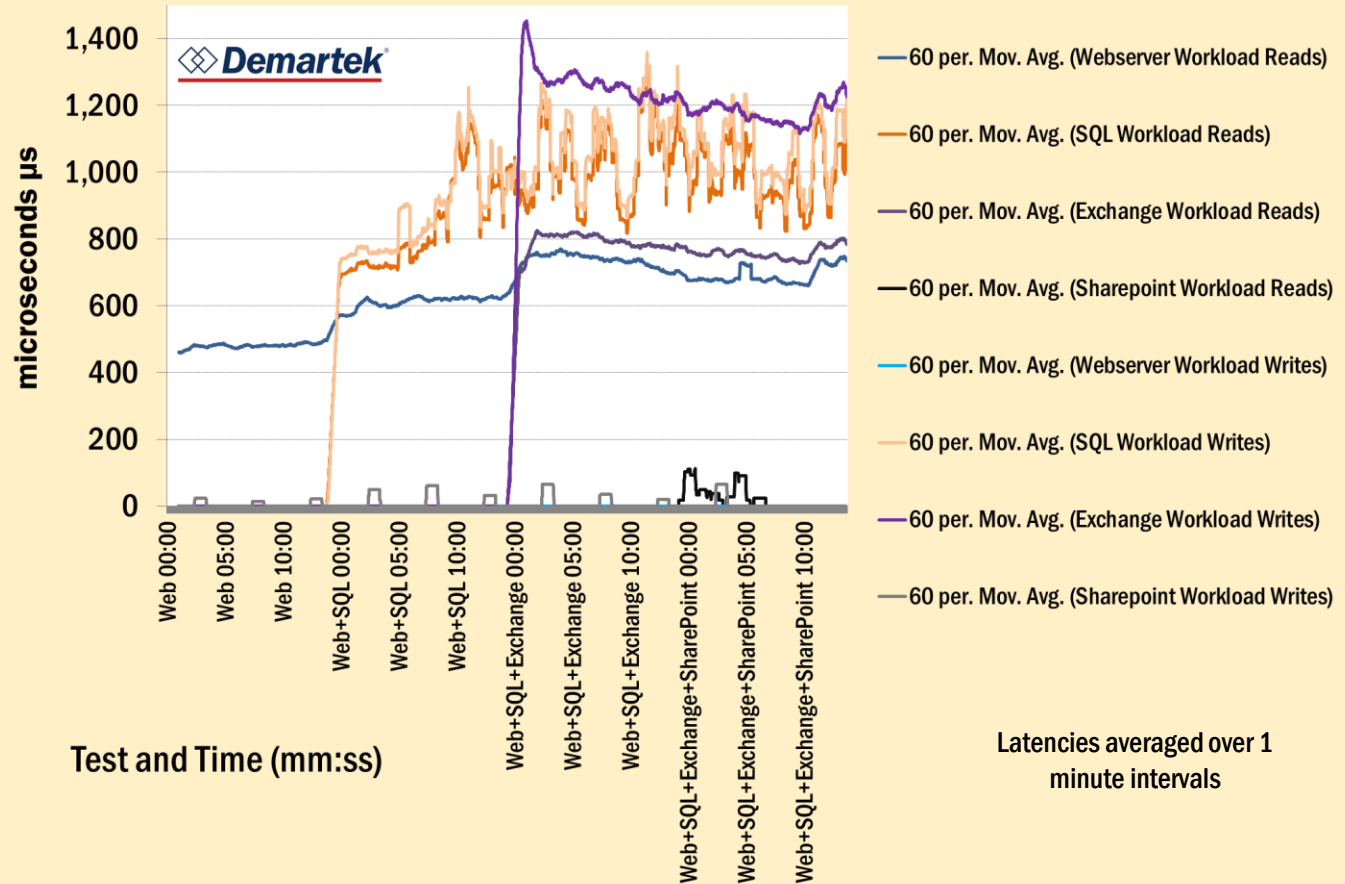
## Violin 7300 FSP Throughput for 4 Workload Test



Single workload latencies were below 1ms, but slightly higher when multiple workloads of different types were added.

Latencies measured at the host server.

## Violin 7300 FSP Averaged Latencies for 4 Workload Test



# Conclusions

- ◆ Real-world workloads can be “messy” compared to synthetic workloads
  - Variable I/O characteristics and multiple factors influencing performance
- ◆ It is not enough to say that an all-flash array can deliver sub-millisecond latencies for a single workload
- ◆ Look for more Demartek multiple workload test results

# Demartek Free Resources

- ◆ Demartek SSD Zone – [www.demartek.com/SSD](http://www.demartek.com/SSD)
- ◆ Demartek iSCSI Zone – [www.demartek.com/iSCSI](http://www.demartek.com/iSCSI)
- ◆ Demartek FC Zone – [www.demartek.com/FC](http://www.demartek.com/FC)
- ◆ Demartek SSD Deployment Guide  
[www.demartek.com/Demartek\\_SSD\\_Deployment\\_Guide.html](http://www.demartek.com/Demartek_SSD_Deployment_Guide.html)
- ◆ Demartek commentary: “Horses, Buggies and SSDs”  
[www.demartek.com/Demartek\\_Horses\\_Buggies\\_SSDs\\_Commentary.html](http://www.demartek.com/Demartek_Horses_Buggies_SSDs_Commentary.html)
- ◆ Demartek Video Library - [http://www.demartek.com/Demartek\\_Video\\_Library.html](http://www.demartek.com/Demartek_Video_Library.html)

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

# Thank You!



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](http://www.demartek.com/newsletter)