The Real Story on Flash Storage Performance
Session TEST-101B-1
9:45 a.m. - 10:50 a.m. PDT,
Tuesday, August 7, 2018
Ballroom G
This Presentation

https://www.demartek.com/FMS2018/
Agenda

- About Demartek
- Synthetic vs. Real-world workloads
- Performance Results – Various Flash Solutions
  (new since last year’s Flash Memory Summit presentation)
- Industry Trends & Future Directions

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

Click to view this one minute video

https://www.demartek.com/Demartek_Video_Library.html
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, NVMe, NVMe over Fabrics
- We prefer to run real-world applications to test servers, storage and HCI solutions (databases, VMware, IoT, etc.)
- Demartek is an EPA-recognized test lab for ENERGY STAR Data Center Storage testing
- Website: https://www.demartek.com/TestLab/
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
Synthetic vs. Real-world Workloads
Synthetic Workloads

- 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
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*
Performance Results
Adding NAND Flash to HDDs

- 24x Seagate TurboBoost HDDs with flash cache in each drive
- Multiple synthetic & real-world workloads

https://www.demartek.com/SeagateEnhancedCache/
Adding NAND Flash to HDDs

Microsoft SQL Server OLTP workload

IOPS

2.9x Improvement

Response Time

4.7x Improvement
Adding NAND Flash to HDDs

VMware ESXi Bootstorm: Fixed amount of work, 60 Win10 VMs

Santa Clara, CA
August 2018

The Performance Story: An Independent Evaluation of Flash Storage – by Demartek
12 NVMe Drives in Cloud Server

- HPE AMD EPYC cloud server cluster
- 100 GbE network
- Excelero NVMesh
- Yahoo Cloud Serving Benchmark (YCSB)

https://www.demartek.com/HPE-Cloudline-CL3150-Benchmark/
Yahoo Cloud Serving Benchmark (YCSB)

- Common cloud datacenter workloads
  - **Workload A**: Update heavy (50% read, 50% write)
  - **Workload B**: Read mostly (95% read, 5% write)
  - **Workload C**: Read only (100% read)
  - **Workload D**: Read latest (new records inserted and then read)
  - **Workload E**: Short ranges (ranges of reads, such as email threads)
  - **Workload F**: Read-modify-write

- Uses NoSQL database (MongoDB, Cassandra, etc.)
Yahoo Cloud Serving Benchmark (YCSB)

- Common cloud datacenter workloads
  - **Workload A**: Update heavy (50% read, 50% write)
  - **Workload B**: Read mostly (95% read, 5% write)
  - **Workload C**: Read only (100% read)
  - **Workload D**: Read latest (new records inserted and then read)
  - **Workload E**: Short ranges (ranges of reads, such as email threads)
  - **Workload F**: Read-modify-write

- Uses NoSQL database ([MongoDB](https://www.mongodb.com), Cassandra, etc.)
Cloud compute / storage nodes

- Each server was configured identically
  - One node was designated the compute node
  - Two nodes were designated the storage nodes (where the application database resided)

- All the data had to traverse the network

- In the event of a compute node failure, it can be replaced without moving any data
YCSB Database Record Counts

- 700,000 records (700K)
- 200,000,000 records (200M)
- 500,000,000 records (500M)

- Fixed amount of work to be processed
With 12 NVMe drives in each server, we found that the bottleneck was the 100GbE network.

See my NVMe over Fabrics Rules of Thumb later in this presentation.
Workload F (the longest of the three we chose)

<table>
<thead>
<tr>
<th>Workload F</th>
<th>Milliseconds</th>
<th>Seconds</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>700K records</td>
<td>43120</td>
<td>43</td>
<td>0.7</td>
</tr>
<tr>
<td>200M records</td>
<td>2777729</td>
<td>2778</td>
<td>46.3</td>
</tr>
<tr>
<td>500M records</td>
<td>5230121</td>
<td>5230</td>
<td>87.2</td>
</tr>
</tbody>
</table>
Results: Workload A

YCSB Workload A Latency
50%/50% Read/Write Mix

99th Percentile Latency (µs): 440
95th Percentile Latency (µs): 330
Min Latency (µs): 200
Average Latency (µs): 145.18

Latency (µs)

Total Records: 500,000,000

Santa Clara, CA
August 2018
Results: Workload B

YCSB Workload B Latency
95%/5% Read/Write Mix

- 99th Percentile Latency (μs): Update 355, Read 460
- 95th Percentile Latency (μs): Update 310, Read 211
- Min Latency (μs): Update 87, Read 46
- Average Latency (μs): Update 144.03, Read 211.43

Total Records: 500,000,000

Santa Clara, CA
August 2018

The Performance Story: An Independent Evaluation of Flash Storage – by Demartek
Results: Workload F

YCSB Workload F Latency
Read, modify and update existing records

- 99th Percentile Latency (μs):
  - Update: 453 μs
  - R/M/W: 353 μs
  - Read: 713 μs

- 95th Percentile Latency (μs):
  - Update: 293 μs
  - R/M/W: 478 μs
  - Read: 206 μs

- Min Latency (μs):
  - Update: 82 μs, R/M/W: 46 μs, Read: 132 μs

- Average Latency (μs):
  - Update: 204.36 μs, R/M/W: 351.28 μs, Read: 144.13 μs

Total Records: 500,000,000
The Performance Story: An Independent Evaluation of Flash Storage – by Demartek

NVMe over Fabrics (FC-NVMe)

- Comparison of FC-SCSI to FC-NVMe
- Same hardware, different protocol

Santa Clara, CA
August 2018

https://www.demartek.com/ModernSAN/
Results: Random Read 4KB

Random Read 4KB Latency vs. IOPS

- At least 34% lower latency
- 53% higher IOPS at 450 μs
- 54% higher IOPS at 2300 μs

Note: all measurements taken on a single-node A700s. Standard implementations are dual-node.
Results: Random Read 4KB (zoom-in)

Random Read 4KB Latency vs. IOPS (zoom in)

Note: all measurements taken on a single-node A700s. Standard implementations are dual-node.

At least 34% lower latency
Results: Oracle 80-20 8KB

Oracle 80-20 8KB
Latency vs. IOPS

Note: all measurements taken on a single-node A700s. Standard implementations are dual-node.

- 47% higher IOPS at 1085 µs
- 58% higher IOPS at 375 µs

At least 18% lower latency
NVDIMM comments

- Faster technology can have some interesting effects.
- We installed some NVDIMMs in a server running Microsoft SQL Server. Because of the speed of the NVDIMMs, we had to adjust the SQL Server recovery interval setting. The default setting was slowing things down.

Industry Trends & Future Directions
NVMe over Fabrics Rules of Thumb

https://www.demartek.com/NVMeoF-Rules/

The Performance Story: An Independent Evaluation of Flash Storage – by Demartek
Balanced Configuration without oversubscribing the PCIe lanes

- Dual-port 25GbE or 32GFC adapter (x8 lanes)
- Processor complex
- 8 PCIe 3.0 lanes
- NVMe drive (x4 lanes)
- NVMe drive (x4 lanes)

https://www.demartek.com/NVMeoF-rules/
Demartek 25GbE Deployment Tips

PRACTICAL TIPS FOR DEPLOYING 25GbE TECHNOLOGY…

BECAUSE THERE ARE SOME THINGS YOU NEED TO KNOW THAT MIGHT NOT BE OBVIOUS.

LEARN MORE >

https://www.demartek.com/25GbE-Tips/
Demartek RoCE Deployment Guide


https://www.demartek.com/RoCE/
Demartek Storage Interface Comparison reference page
- Search engine: Storage Interface Comparison
- Recent updates for PCIe 5.0, U.3, Fibre Channel, FC-NVMe & SATA

https://www.demartek.com/Storage-Interface-Comparison/
U.2 and U.3 backplanes

U.2 – SFF8639

U.3 – SFF-TA-1001
Rev. 1.0 was ratified in November 2017 and Rev. 1.1 was ratified in May 2018.

https://www.demartek.com/Storage-Interface-Comparison/
Roadmaps

- PCIe 4.0 – 1.0 spec. published October 2017
- PCIe 5.0 – revision 0.7 published May 2018
  - Target of Q1 2019 for spec. complete
- NVMe and NVMe over Fabrics (NVMe-oF) – next revision in 2019
- Ethernet & Fibre Channel – some of the same technology will drive single-lane 50GbE and 64GFC.

https://www.demartek.com/Storage-Interface-Comparison/
Demartek Free Resources

- Demartek FC Zone – www.demartek.com/FC/
- Demartek iSCSI Zone – www.demartek.com/iSCSI/
- Demartek NVMe Zone – www.demartek.com/NVMe/
- Demartek SSD Zone – www.demartek.com/SSD/
- 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
https://www.demartek.com/FMS2018/
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