The Straight Truth: How Today’s Storage Performs on Real Workloads

TEST-101B-1

Dennis Martin, Principled Technologies
Agenda

- Principled Technologies: Who we are and what we do
- Interface vs. device speeds
- Synthetic vs. real-world workloads
- Performance results
- Industry trends & future directions

(If you’d like a copy of this presentation, I provide my contact information at the end)
Demartek → Principled Technologies

- I’ve spoken here at FMS for several years
- Many of you know me as the President and Founder of Demartek
- Principled Technologies (PT) acquired Demartek in September 2018
- Introducing PT…
Get the resources your project needs with top-of-the-line testing and creative facilities.

Modern, on-premises data centers

Dedicated client test beds

2,000-square-foot, fully equipped video production studio with green and white cycloramas

Take a 360° tour of PT

https://facts.pt/FBM-outcomes

Santa Clara, CA
August 2019

The Straight Truth: How Today’s Storage Performs on Real Workloads – Dennis Martin
The BenchmarkXPRT Development Community provides registered members with the opportunity to contribute to the process of creating and improving the XPRTs – benchmark apps that measure how well PCs, tablets, and smartphones handle everyday tasks. The XPRTs empower people all over the world to know how well their gear handles workloads like editing photos, playing movies, and browsing the Web.

http://benchmarkxprt.com
BenchmarkXPRT selector

http://facts.pt/the-xprt-selector
Interface vs. device speeds
# Interface vs. device speeds

<table>
<thead>
<tr>
<th>Application</th>
<th>Interface speeds</th>
<th>Device speeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bluetooth, Ethernet, Fibre Channel, InfiniBand, Mobile phones (3G / 4G / 5G), SAS, SATA, Thunderbolt, USB, Wi-Fi</td>
<td></td>
<td>HDDs, SSDs, NVMe, PCIe, consumer device storage cards (SD, CompactFlash, etc.), Memory (DRAM, NVDIMM, etc.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Throughput</th>
<th><strong>Bits</strong> per second</th>
<th><strong>Bytes</strong> per second</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviation</td>
<td>Mbps, Gbps, or Mb/s, Gb/s</td>
<td>MBps, GBps, or MB/s, GB/s</td>
</tr>
<tr>
<td>Case</td>
<td>Lowercase “b”</td>
<td>Uppercase “B”</td>
</tr>
</tbody>
</table>

*Converting from bits per second to bytes per second is not a simple divide by 8. You must know the encoding scheme for that interface.*
What about memory?

<table>
<thead>
<tr>
<th>Addressing</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Byte-addressable</strong></td>
</tr>
<tr>
<td></td>
<td>Applies to volatile and non-volatile types</td>
</tr>
<tr>
<td>Speed</td>
<td><strong>Clock rate, such as 2133 MHz</strong></td>
</tr>
<tr>
<td>Throughput</td>
<td><strong>Gigabytes per second (GBps or GB/s)</strong></td>
</tr>
</tbody>
</table>
Synthetic vs. real-world workloads
## Synthetic vs. real-world workloads

<table>
<thead>
<tr>
<th>Storage testing</th>
<th>Synthetic</th>
<th>Real-world</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O profile</td>
<td>Fixed, controlled, specific block size, read/write mix, length of time, random vs. sequential, etc.</td>
<td>Variable block sizes, read/write mix, length of time, random, sequential, etc. May use multiple different parameters simultaneously. May issue few or no I/O's for some periods of time.</td>
</tr>
<tr>
<td>CPU usage</td>
<td>Low to medium, but generally steady, focused on I/O tasks</td>
<td>Variable: low to high (performs other tasks in addition to I/O)</td>
</tr>
<tr>
<td>Memory usage</td>
<td>Generally fixed and small</td>
<td>Variable: small to large</td>
</tr>
<tr>
<td>Results</td>
<td>“Hero numbers”</td>
<td>Application-specific, such as database transactions per second, orders per minute, etc.</td>
</tr>
<tr>
<td>Who</td>
<td>Storage product vendors</td>
<td>End-users</td>
</tr>
</tbody>
</table>
Real-world workload types

<table>
<thead>
<tr>
<th>I/O pattern</th>
<th>Transactional</th>
<th>Streaming</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mostly random</td>
<td>Mostly sequential</td>
</tr>
<tr>
<td>Emphasis</td>
<td>I/O’s per second (IOPS)</td>
<td>Throughput (MBps or GBps)</td>
</tr>
<tr>
<td>Latency</td>
<td>Important</td>
<td>Not very important</td>
</tr>
</tbody>
</table>

*We observe different latency result patterns for different workload types.*
Blocksize performance relationships (synthetic benchmarks)

These generic performance curves apply to network and storage performance. For network performance, replace IOPS with Frames per Second (FPS).

Source: Demartek internal testing
Generic latency results (real-world workloads)

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 pink workload affected the latency of the green workload.

Source: Demartek internal testing
Real-world latency measurements

We measure end-to-end latency from the source to the target and back to the source, through all the hardware and software layers.

Single-server solution:
- Between processor and memory
- Between processor and storage

Hyper-converged solution:
- Between server and local storage
- Between server and remote storage in a different node

External storage solution:
- Between server and external storage system

Santa Clara, CA
August 2019
The Straight Truth: How Today’s Storage Performs on Real Workloads – Dennis Martin
Performance results
Two desktop computers:
- Intel Core i5 8600

Machine 1:
- 32GB RAM
- $2012

Machine 2:
- 16GB RAM
- 16GB Intel Optane mem.
- $1706

http://facts.pt/pqv8ua2
4-node VMware vSAN cluster

Each server:
- 2x Intel Xeon® Gold 6154 CPU @ 3.00GHz
- 384 GB RAM

Config 1
- 6x 2 TB NVMe SSDs (capacity)
- 2x 2 TB NVMe SSDs (cache)

Config 2
- 6x 2 TB NVMe SSDs (capacity)
- 2x 375 GB Optane SSDs (cache)

http://facts.pt/qtaj3ob
Upgrade to NVMe and Optane persistent memory

Storage config 1 (HDD)
- **Data**: 4-drive RAID10 volume
- **Logs**: 2-drive RAID1 volume

Storage config 2 (SATA SSD)
- **Data**: 4-drive RAID10 volume
- **Logs**: 2-drive RAID1 volume

Storage config 3 (NVMe + Optane)
- **Data**: 1 Intel Optane NVMe SSD
- **Logs**: Intel Optane persistent memory (Non-Interleaved App Direct Mode)

http://facts.pt/1mgym04

OLTP database workload

- **New server with NVMe (database) + Optane persistent memory (logs)**: 243,501 OPM
- **New server with SATA SSDs**: 206,770 OPM
- **4-year old server with HDD**: 8,616 OPM

http://facts.pt/1mgym04
SATA vs. SAS vs. NVMe storage

Storage configurations
- **NVMe**: 8x 960GB NVMe SSD
- **SAS**: 8x 960GB SAS SSD
- **SATA**: 8x 960GB Enterprise SATA SSD

Server configuration
- 2x AMD EPYC 7601, 32c/64t
- 128GB RAM

http://facts.pt/2h8emuf
Industry trends & future directions
## PCI Express (PCIe)

<table>
<thead>
<tr>
<th>Specification announced</th>
<th>Specification completed</th>
<th>Gigatransfers per second (GT/s)</th>
<th>x1 bandwidth (half-duplex)</th>
<th>x16 bandwidth (half-duplex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCIe 4</td>
<td></td>
<td>16</td>
<td>2 GB/s</td>
<td>32 GB/s</td>
</tr>
<tr>
<td>June 2017</td>
<td>May 2019</td>
<td>32</td>
<td>4 GB/s</td>
<td>64 GB/s</td>
</tr>
<tr>
<td>PCIe 6</td>
<td></td>
<td>64</td>
<td>8 GB/s</td>
<td>128 GB/s</td>
</tr>
</tbody>
</table>

Half-duplex speeds indicate the maximum rate for storage reads or writes in one direction only. For transmissions in both directions simultaneously (full-duplex), double the rate.

Source: PCI-SIG  [https://pcisig.com/](https://pcisig.com/)
## Ethernet, Fibre Channel & InfiniBand roadmaps

<table>
<thead>
<tr>
<th></th>
<th>Single-lane speeds (Gbps)</th>
<th>Four-lane speeds (Gbps)</th>
<th>Eight-lane speeds (Gbps)</th>
<th>Twelve-lane speeds (Gbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet</td>
<td>1, 2.5, 5, 10, 25, 50*</td>
<td>40, 100, 200*</td>
<td>100, 400*</td>
<td>-</td>
</tr>
<tr>
<td>Fibre Channel</td>
<td>8, 16, 32, 64*</td>
<td>128, 256*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>InfiniBand</td>
<td>8, 14, 25, 50*</td>
<td>32, 56, 100, 200*</td>
<td>-</td>
<td>96, 168, 300, 600*</td>
</tr>
</tbody>
</table>

*Speeds in purple generally require at least one PCIe 4.0 slot.

Links to public roadmaps listed on the “References” page later in this presentation.

---

Ethernet, Fibre Channel and InfiniBand are all dependent on two underlying technologies and will achieve new speed increments in similar time frames. The two technologies are: **PCI Express** and **Transceiver technology**.

---

* Speeds in purple generally require at least one PCIe 4.0 slot.

Links to public roadmaps listed on the “References” page later in this presentation.
“BEAVERTON, Ore.,—USA—July 23, 2019—NVM Express™, Inc. today announced the release of NVM Express™ (NVMe™) 1.4 Base Specification and that NVMe™ over Fabrics (NVMe-oF™) 1.1 specification has entered into final 45-day member review.”


- You will learn much more about **NVMe 1.4** and **NVMe-oF 1.1** at this conference in the NVMe track and in the expo hall.
- NVMe-oF 1.1 includes NVMe/TCP.
References

- Ethernet Alliance 2019 public roadmap:

- Fibre Channel public roadmap:
  https://fibrechannel.org/roadmap/

- InfiniBand public roadmap:
  https://www.infinibandta.org/infiniband-roadmap/

- PCI Express version 6 announcement:
  http://pcisig.com/pioneering-interconnect-industry-pci-sig%C2%AE-announces-upcoming-pcie%C2%AE-60-specification
Questions?

- I’m happy to answer questions after this session
  - I’ll be available during this conference

- Send me a message on LinkedIn
  - [https://www.linkedin.com/in/dennismartin](https://www.linkedin.com/in/dennismartin)
  - To ask questions about this presentation
  - To get a copy of this presentation
  (not putting my email here to avoid spam)
Principled Technologies posts public projects and materials to our website and announces them on a variety of social media outlets. Find us on any of the above.