Enterprise Flash Storage
Annual Update

Flash, It’s not just for tier 0 anymore
Or
Flash is the new black

Howard Marks
Technologist Extraordinary
and Plenipotentiary
Your not so Humble Speaker

- 30+ years of consulting & writing for trade press
- Occasional blogger at TechTarget
- Recently Chief Scientist DeepStorage, LLC.
  - Independent test lab and analyst firm
- Technologist Extraordinary and Plenipotentiary
  - VAST Data
    - I promise to keep the sales down, I’m new to it anyway

Howard@VASTdata.com  @DeepStorageNet
Agenda

- Review 2017-2018 events, predictions
  - Flash is just normal
  - The shift from SSD to NVMe
  - NVMe over fabrics the new lingua franca
    - Is Tier 0 sustainable
  - 3D Xpoint and Storage Class Memory
- A look at a few illustrative examples
A Decade+ of Enterprise Flash

2007
- Rackmount SSDs
- Texas Memory
- Violin Memory
- Fast but niche

2010
- SSDs in DISK arrays
- High cost
- Endurance fears
- Hybrids emerge

2014
- Flash goes commercial
- All Flash Arrays
- Costs = high performance HDD

2017
- Flash is mainstream
- Full data services & data reduction
- Cost effective for primary storage

2020
- Democratizing flash
- Data intensive applications
- 3D Xpoint starts small/fast cycle again
The Tipping Point Tipped

- **2017**
  - Enterprise SSD 25X capacity HDD $/GB

- **2019**
  - 1 TB SSD < $100
  - Enterprise SSD 10-12¢/GB (3.5X)
  - WD exits 10 & 15K RPM HDDs
  - SK Hynix announces 128 layer 128 Tb chip

<table>
<thead>
<tr>
<th>SSD Type</th>
<th>GB</th>
<th>Cost</th>
<th>$/GB</th>
<th>HDD Multiple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel P3700</td>
<td>1600</td>
<td>$405</td>
<td>0.25</td>
<td>8.44</td>
</tr>
<tr>
<td>Intel P3520</td>
<td>2000</td>
<td>$535</td>
<td>0.27</td>
<td>8.92</td>
</tr>
<tr>
<td>16TB HDD</td>
<td>16000</td>
<td>$480</td>
<td>0.03</td>
<td>8.92</td>
</tr>
<tr>
<td>Micron 5120 ION</td>
<td>7680</td>
<td>$800</td>
<td>0.10</td>
<td>3.47</td>
</tr>
</tbody>
</table>

Source: NewEgg.com 8/1/2019
The Tipping Point Tips, Part Deux

- AFA market share passes hybrids

Source: IDC Worldwide Quarterly Enterprise Storage Systems Tracker
## All Flash Player Joins the Big Boys

<table>
<thead>
<tr>
<th>Company</th>
<th>1Q19 Revenue</th>
<th>1Q19 Market Share</th>
<th>1Q18 Revenue</th>
<th>1Q18 Market Share</th>
<th>1Q19/1Q18 Revenue Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Dell Technologiesa</td>
<td>$2,355.9</td>
<td>34.4%</td>
<td>$2,219.6</td>
<td>34.0%</td>
<td>6.1%</td>
</tr>
<tr>
<td>2. NetApp</td>
<td>$894.9</td>
<td>13.0%</td>
<td>$890.1</td>
<td>13.6%</td>
<td>0.5%</td>
</tr>
<tr>
<td>3. HPE/New H3C Groupb</td>
<td>$745.4</td>
<td>10.9%</td>
<td>$652.2</td>
<td>10.0%</td>
<td>14.3%</td>
</tr>
<tr>
<td>4. Hitachi</td>
<td>$452.7</td>
<td>6.6%</td>
<td>$457.9</td>
<td>7.0%</td>
<td>-1.1%</td>
</tr>
<tr>
<td>T5. IBM*</td>
<td>$320.0</td>
<td>4.7%</td>
<td>$364.1</td>
<td>5.6%</td>
<td>-12.1%</td>
</tr>
<tr>
<td>T5. Pure Storage*</td>
<td>$289.5</td>
<td>4.2%</td>
<td>$236.4</td>
<td>3.6%</td>
<td>22.4%</td>
</tr>
<tr>
<td>Rest of Market</td>
<td>$1,800.3</td>
<td>26.2%</td>
<td>$1,709.0</td>
<td>26.2%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Total</td>
<td>$6,858.6</td>
<td>100.0%</td>
<td>$6,529.3</td>
<td>100.0%</td>
<td>5.0%</td>
</tr>
</tbody>
</table>

The Toshiba Memory Soap Opera

- 2006 Toshiba buys Westinghouse
- 3/2017 Westinghouse chapter 11
  (AP1000 Reactors $9B loss)
- 9/2109 Toshiba sells memory unit $18B
  - WAIT – Western Digital/SanDisk sue
    - Lawyers make money, waste time
- Sale Closes 6/2018
  - 9/2018 Fab 6 opens at Yokkaichi
- June 16 Power failure at Yokkaichi
  - 6+ EB NAND production lost
- Also June CEO Yasuo Naruke, goes on sick leave
Toshiba becomes Kioxia

- IPO planned for 9/2019
  - Before both power failure and CEO illness
- Rebrand effective October 1, 2019.

- **Kioxia:**
  - Kioku meaning “memory”
    - Japanese
  - Axia meaning “value”
    - Greek

- **Pronunciation:**
  kee-ox-ee-uh
The Party’s Over, again

- 2008-2015 SSD $/GB -30%/yr
- 2016-2018 maybe 30% total
- Last year I said “Expect 30+% CAGR”
- I thought:
  - Supply is easing
    - 96 layer+ QLC
    - Process improvements
    - New fabs
- Fabs cut back starts
- Next 2 quarters flat, back to 25-30% CAGR
Enterprise SSD Evolution

- Data center NVMe ≈ SAS/SATA volume
- SSDs and HDDs now both ≈ 16 TB
- Greater Differentiation
  - Performance and cost vary 5X or more
  - SLC returns as SCM
- New form factors remain proprietary
  - M.2 didn’t work in data centers
  - Samsung NGSFF
  - Intel Ruler
SSD Differentiation

- **Storage Class Memory SSDs**
  - More on this later

- **Dual-port enterprise**
  - DRAM/Supercap

- **Single port enterprise**
  - NVMe and SATA for HCI, HPC, Etc.

- **Low cost single port**
  - Hyperscaler’s tail
QLC SSD Endurance by Workload

QLC SSD Drive Writes by Workload

- 1855
- 1842
- 1823
- 1784
- 1712
- 378
- 182
- 91

- 128KB Sequential
- 90% 128K Seq, 10%
- 80% 128K Seq, 20%
- 70% 128K Seq, 30%
- 50% 128K Seq, 50%
- 16KB Ransom
- 8KB Random
- 4 KB Random
Open Channel SSDs

The disaggregation of flash storage

Today - Monolithic model
- Host
- Hardware managed (SSD)
  - SSD Drive
    - Address mapping
    - Garbage collection
    - Wear leveling
    - Manage bad blocks
    - Manage media
    - Power failure

Denali model
- Host
  - Software defined (Direct)
    - Address mapping
    - Garbage collection
    - Wear leveling
  - SSD Media drive
    - Manage bad blocks
    - Manage media
    - Power failure

  - Software defined (Offloaded)
    - SoC or FPGA
      - Address mapping
      - Garbage collection
      - Wear leveling
    - Accelerators
    - SSD Media drive
      - Manage bad blocks
      - Manage media
      - Power failure
PCIe Advances

- **PCIe 4.0**
  - Doubles bandwidth/lane to 2GBps
  - Driven by 100Gbps Ethernet & NVMe
  - Power systems shipping now
  - x86 Next server chipset release

- **PCIe 5.0 close on its heals**
  - .7 version issued May 2018
  - Adoption planned Q1 2019
  - 400Gbps Ethernet ≅ x16 slot
  - Servers and such 2020?

<table>
<thead>
<tr>
<th></th>
<th>Spec Date</th>
<th>Raw Bandwidth per lane</th>
<th>x8 Gbps</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCIe 1</td>
<td>2003</td>
<td>2.5GT/s</td>
<td>16</td>
</tr>
<tr>
<td>PCIe 2</td>
<td>2007</td>
<td>5.0GT/s</td>
<td>32</td>
</tr>
<tr>
<td>PCIe 3</td>
<td>2010</td>
<td>8.0GT/s</td>
<td>64</td>
</tr>
<tr>
<td>PCIe 4</td>
<td>2011</td>
<td>16GT/s</td>
<td>128</td>
</tr>
</tbody>
</table>
NVMe Over Fabrics (NVMe-oF)

- Extends/encapsulates NVMe semantics over
  - Ethernet with RMDA
  - Fibre Channel
  - Infiniband (no products yet announced)
  - TCP
- Adds name spaces and discovery
- 10-50μsec protocol and network overhead
NVMe-oF Models

- **JBOF**
  - Just Fabric-SSD bridges
  - HA optional
- **JBOF+**
  - Adds slice/dice and RAID
  - Also manage in client models
- **NVMe-oF Array**
  - All the abstractions and services of SCSI over Fibre Channel
  - Lower latency of NVMe-oF
NVMeOF Pioneers Shakeout

- Mangstor
  - Reborn as EXTEN
  - Software NVMe-oF JBOF+
- Apeiron – 40Gbps Ethernet switch in JBOF
- E8 – Dual controller array – basic services
  - Acquired by AWS
- Excellero – Low CPU SDS, RDMA
NVMe-oF Use Cases

- Intra-storage system SAS replacement
- HPC/skunkworks/Rackscale
  - RDMA to JBOFs
- Hyperscale
  - TCP to expand to data center scale
- Enterprise
  - Primarily arrays
  - NVMe runs over Fibre Channel for these customers
NVMe Over Fibre Channel

- Fibre Channel
  - Zero copy vs RDMA
  - Flow and congestion control
- Gen5 (16) and Gen6 (32Gbps) Fibre Channel
- One fabric for SCSI and NVMe
- Keeps storage network in storage domain
- The safe move in enterprise
**NVMe over TCP**

- Encapsulates NVMe verbs in TCP
- Relies on TCP low control
- NIC offload optional
- No switch config requirements
- Nominal latency addition
- Supporters:
  - SolarFlare
  - Cavium
  - Toshiba
NVMe JBOFs Emerge

- **Today’s JBOFs are x86 servers**
  - Eg: Toshiba KumoScale
  - High flexibility
  - High cost

- **NVMeoF ASICs**
  - Vastly reduce costs
  - Sampling from
    - SolarFlare Xilinx
    - Kazan Networks
    - Attala Systems
    - Mellanox

---

[Diagram showing NVMeoF architecture with Host Server, Ethernet Network, Attala Data Node, NVMe SSDs, and FPGAs connected through PCIe Switch with NVMe Driver and OFED stack/RNIC driver.]
**NVMe™ Feature Roadmap**

### NVMe™ Base
- **NVMe™ 1.2 – Nov ’14**
  - Namespace Management
  - Controller Memory Buffer
  - Host Memory Buffer
  - Live Firmware Update

### NVMe ofabric
- **NVMe-oF™ 1.0 May’16**
  - Transport and protocol
  - RDMA binding

### NVMe™ MI
- **NVMe-MI™ 1.0 Nov’15**
  - Out-of-band management
  - Device discovery
  - Health & temp monitoring
  - Firmware Update

### NVMe™ 1.3
- Sanitize
- Streams
- Virtualization

### NVMe™ 1.4*
- IO Determinism
- Persistent memory Region
- Multipathing

### NVMe-oF™-1.1*
- Enhanced Discovery
- TCP Transport Binding

---

* Released NVMe™ specification

Planned release

* Subject to change
Storage Class Memory

- A controversial term
  - As well defined as Software Defined

- For me:
  - Inherently persistent
  - Latency between DRAM and NAND Flash
  - Bit addressable
  - Both material and usage
Storage Class Memories Today

- **3D Xpoint**
  - SSDs not a huge success
    - So far
  - DIMMs show promise
  - 2nd gen still to come (Micron?)
    - Gen 1 is 3D but only 1 cell deep

- **Everspin Spin-transfer Torque MRAM**
  - 1Gb/chip @ 28nm
    - NAND 1.33 Tb/chip
  - DRAM replacement on SSDs

- **Others SciFi**
SCM in Enterprise Storage

- **HPE**
  - Optane AIC in controller
  - 3PAR and Nimble for cache
  - Back-ends still SAS

- **Dell EMC PowerMax**
  - Optane D4800X (dual port)
  - Tier of storage

- Mostly HCI/SDS Optane SSDs
SLC Returns

- Samsung Z-NAND
- Kioxia (AKA Toshiba) XL-FLASH
  - Multi-plane for parallelism
  - 4 KB page
    - 128 KB in 1Tb QLC
  - 128 Gb/die
  - X μsec read latency
- Still flash w/write asymmetry
- SSDs today
  - Flash DIMMs seem passe
Optane DIMMs

- Require latest Xeons
  - Special models for large memory addresses
- OS/Hypervisor support as PMEM
- Complex programming models
All database operations performed in RAM
- Data replicated across nodes (x86)
- AFA/HCI back end for persistence
  - Snapshots
  - Transaction Logs
  - Playback in case

On write:
1. Replicate to 1-n nodes
2. Write to persistent log (typically AFA)
3. ACK
In Memory Database with SCM

- **Much larger capacity/node**
  - 512GB vs 64GB/DIMM
  - 10X latency (SWAG)

- **Lower cost /GB**
  - 2-10X we guess
  - More vs 128GB LRDIMMs
    - 3X cost of 64GB

- **ACK after n-node write**
  - Can be RDMA write
  - Data now persistent
  - Log writes can be aggregated, async
SAP HANA Native Support for Persistent Memory
Officially Supported in SAP HANA 2.3 (April 2018)

**Larger memory capacity with high performance (vs. DRAM & lower tier storage)**

**Lower TCO data storage hierarchy**

**Faster start time delivers less downtime**

**Co-innovation with Intel® leads to first fully optimized major DBMS platform**

**Early Adoption Program** with key partners/customers ongoing

---

**Benefit**

- Process more data in real-time at a lower TCO with improved business continuity

> **3 TB**

- Increased total memory capacity per CPU

**12.5x**

- Improvement in startup time*

---

First major DBMS vendor to officially support Intel Optane DC persistent memory!

[Visit sap.com/persistent-memory](http://sap.com/persistent-memory)
Pure FlashArray//x

- Replaces //m
  - SAS SSDs
  - Expansion via SAS or NVMeoF JBOF
- NVMeoF target on 40Gbps Ethernet
- Full services
Kaminario K2 Composable

- NVMEoF
  - Controller to JBOF
  - Host to array (opt)
- Dynamically assign controllers and flash to virt array
VAST Data Universal Storage System

- 3D XPoint and QLC Flash in HA NVMe-oF JBOFs
- Storage services via stateless servers (metadata in XPoint)
  - File and object  Global Name Space
  - Data reduction  Erasure coding
- Global FTL