Enterprise Flash Storage
Annual Update

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

Santa Clara, CA
August 2018

Howard Marks
Chief Scientist
Your not so Humble Speaker

- 30+ years of consulting & writing for trade press
- Occasional blogger at TechTarget
- Chief Scientist DeepStorage, LLC.
  - Independent test lab and analyst firm
- Cohost Greybeards on Storage podcast

Hmarks@DeepStorage.Net  @DeepStorageNet
Agenda

- A brief history lesson
- The shift from SSD to NVMe
- NVMe over fabrics the new lingua franca
- A look in the crystal ball
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 understood
- All Flash Arrays
- Costs close

2018
- Flash is mainstream
- Full data services & data reduction
- Cost effective for most applications
Flash is just the default

- All flash ~$8bil/yr w/12% projected growth
- Disk is still cheaper
  - But being reserved for:
    - Secondary
    - Rich media
- Users are over endurance & deduplication fears
- Shift back to full featured arrays from purpose built AFA
The Great Flash Shortage of 2016-7

- 2008-2015 SSD $/GB −30%/yr
- 2016-2018 maybe 30% total
- Last year I said “Relief to come late 2018/19”
- Supply is easing
  - 96 layer QLC
  - Process improvements
  - New fabs
- Expect 30+% CAGR
Enterprise SSD Evolution

- Further fragmentation
  - Optane/Samsung Z-NAND NVMe
  - 100TB 6gbps SATA
- U.2 across server vendors
  - New form factors:
    - Samsung NGSFF
    - Intel Ruler
Solid State Drive to Solid State Device

- Dropping the HDD form factor
  - M.2 for boot
  - Ruler/NGSFF for hot-swap
  - Better cooling and density
- PCIe replaces SAS/SATA
- PCIe Switch chips vs SAS Expanders
- NVMe replaces SCSI as lingua franca
  - Over PCIe locally
  - Over fabrics
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 \(\approx\) x16 slot
  - Servers and such 2020?

<table>
<thead>
<tr>
<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>2.5G T/s</td>
</tr>
<tr>
<td>PCIe 2</td>
<td>2007</td>
<td>5.0G T/s</td>
</tr>
<tr>
<td>PCIe 3</td>
<td>2010</td>
<td>8.0G T/s</td>
</tr>
<tr>
<td>PCIe 4</td>
<td>2017</td>
<td>16GT</td>
</tr>
</tbody>
</table>
NVMe 101

- Gen1 and 2 PCI SSDs
  - ACHI (SATA command set)
  - Propreataary (Fusion-IO, Verident) with heavy software
- Enter NVM Express
  - A new software protocol for non-volatile memory access
- Lower compute overhead than SCSI
- 64K queues of 64K entries vs SCSI 1 queue of 32 entries
NVMe = Lower Overhead & Latency

- By 2016 NVMe is leading from desktop M.2 to the datacenter
- But limited to internal SSDs
NVMe Over Fabrics (NVMEoF)

- Extends/encapsulates NVMe semantics over
  - Ethernet with RMDA
  - Fibre Channel
  - Infiniband (no products yet announced)
  - TCP
- Adds name spaces and discovery
- 10-50\(\mu\)sec protocol and network overhead
NVMeOF Ethernet Options

- RDMA over Converged Ethernet (ROCE)
- iWARP (Internet Wide-area RDMA Protocol)
- RNICs generally support ROCE or iWARP
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
- Greybeards on Storage
NVMeOF Pioneers

- Apeiron – 40Gbps Ethernet switch in JBOF
- E8 – Dual controller array – basic services
- Mangstor – x86 NVMEoF target
- Excellero – Low CPU SDS, RDMA
Pure FlashArray///x

- Replaces //m SAS SSDs with NVMe flashmodules
- Expansion via SAS or NVMEoF JBOF
- NVMEoF target on 40Gbps Ethernet
- Full services
Dell/EMC PowerMAX

- Should end the “designed from scratch for flash” argument
- All the Symetrix/VMAX software goodness
- NVMe media
- NVMe over fabrics promised
- Scaleout x86 & FICON
NetApp and IBM Go NVMEoFC

- **IBM FlashSystem 9100**
  - 24 flash modules (19.2TB, 384TB net)
  - 16Gbps FC, NVMEoFC*
  - SVC based services

- **NetApp A series AFF**
  - A800 – 48 SSD slots
  - Sub 200μsec latency, 11 millionIOPS
  - Data OnTap services
Standards Progress

- **2014**
  - NVMe™ Base
    - NVMe™ 1.2 – Nov ’14
      - Namespace Management
      - Controller Memory Buffer
      - Host Memory Buffer
      - Live Firmware Update
  - NVMe-oF™ 1.0 May ’16
    - Transport and protocol
    - RDMA binding

- **2015**
  - NVMe™ 1.2.1 May’16

- **2016**
  - NVMe™ 1.3
    - Sanitize
    - Streams
    - Virtualization

- **2017**
  - **NVMe-oF™-1.1**
    - Enhanced Discovery
    - TCP Transport Binding

- **2018**
  - NVMe™ 1.4+
    - IO Determinism
    - Persistent memory Region
    - Multipathing

- **2019**
  - NVMe-MI™ 1.0 Nov’15
    - Out-of-band management
    - Device discovery
    - Health & temp monitoring
    - Firmware Update
  - NVMe-MI™ 1.1
    - SES Based Enclosure Management
    - NVMe-MI™ In-band
    - Storage Device Enhancements

**Legend:**
- Released NVMe™ specification
- Planned release

* Subject to change
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
Kaminario K2 Composeable

- NVMEoF
  - Controller to JBOF
  - Host to array (opt)

- Dynamically assign controllers and flash to virt array
Persistent Memory Now GA

- Scaleable Xeon servers support NVDIMM-N
- Good for software delivered storage
  - Small (8-16GB)
  - Expensive (2-3X DRAM)
- Full OS/Hypervisor Support
  - Windows
  - vSphere
  - Linux
NetList’s HybriDIMM

- Combines DRAM-Flash
- Conceptually like Diablo/Sandisk UltraDimm

- Access:
  - DRAM as std memory
  - Flash w/DRAM buffer as Block storage
  - Flash as persistent memory via Linux Library
  - No special BIOS support needed
  - 128-512GB
Crystal ball section
The Future

- **All PCIe NVMe storage systems**
  - As conventional storage
  - With memory interfaces

- **Next-gen memory (PCM, 3d Xpoint, Etc)**
  - First as write cache in SSD
  - Later as memory
  - Taking a bit longer than expected

- **More persistent memory as memory**
  - Needs application support ala SAP Hana
Storage Class Memory

- As well defined as Software Defined
- For me:
  - Inherently persistent
  - Latency between DRAM and NAND Flash
  - Addressable as memory
    - Not SSD, not NVMe
  - Capacity 4-∞X RDIMM
- Defines material AND implementation
Optane DIMMs Coming Soon

Big and Affordable Memory 128, 256, 512GB
High Performance Storage DDR4 Pin Compatible
Direct Load/Store Access Hardware Encryption
Native Persistence High Reliability

NOW SHIPPING SAMPLES BROAD DEVELOPER ENGAGEMENT
In Memory Databases Today

- 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
Relative Memory Costs

- DRAM: 1
- Optane (P905): 0.12
- Optane (P4800X): 0.41
- NVMe Flash: 0.07
Thank you.
NVMe™ JBOF System Reference Design

1. NVMe JBOF controller
2. Solarflare soft NIC services
   - Over 2,000 vNICs per NIC
   - Hardware firewalls
   - Kernel bypass for acceleration
   - 100% packet surveillance
   - NVMe protocol processing
3. NVMe SSD
4. NVMe SSD tray
5. Just add your storage software stack here

Xilinx and Solarflare have partnered to create a NVMe Just-a-Bunch-of-Flash reference design which makes it quick and easy for storage vendors to perform proof-of-concept on a storage software stack, or go into production with a high-performance flash subsystem with a full suite of services for NVMe fabrics.