The Solid State Storage (R-)Evolution

Michael Krause
Hewlett-Packard
Agenda

- Sustainable Customer Value
  - Performance
  - Availability / Durability / Persistence
  - Serviceability
- Software (R-)evolution
- Hardware (R-)evolution
- Next Generation Persistent Media
Performance

- Multiple performance dimensions
  - Raw performance – bandwidth, latency, IOPS, $$
    - Solid state provides significantly increased performance
  - Watt / IOPS
    - Power consumption heavily dependent upon underlying media
  - HW/SW performance’s ability to unlock stored capacity

- Performance improvements derived from multiple fundamental areas
  - Solid State Media + Interconnect
    - Media attributes play a significant role in solution value beyond just performance
    - Interconnect (IO / memory) + protocol are equally critical
  - Software stack - Volume + Re-architecture
    - Volume stacks achieve significant performance gains transparent to the application
    - New software stacks leverage existing / new techniques to deliver greater gains
  - Collapse of memory / storage hierarchy
    - Drive to simplicity with increased emphasis on data locality management
      - Persistent memory augments / replaces traditional DRAM over time
      - High-speed storage to solid state
      - Infrequently accessed data moves to large capacity bulk storage devices
Storage device characteristics

- Read & write performance vary
- Device latencies vary
- Orders of magnitude of performance improvement

Example Performance Comparison

![Graph](image)

<table>
<thead>
<tr>
<th>Storage Device</th>
<th>Random Read IOPS</th>
<th>Random Write IOPS</th>
<th>Latency µS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP IO Accelerator</td>
<td>97</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>Mainstream SSD</td>
<td>50</td>
<td>10</td>
<td>0.38</td>
</tr>
<tr>
<td>Performance HDD</td>
<td></td>
<td></td>
<td>0.36</td>
</tr>
</tbody>
</table>

@ 2KB
## Projected Flash Performance

<table>
<thead>
<tr>
<th>2.5” (MLC)</th>
<th>800GB</th>
<th>1.6TB</th>
<th>3TB</th>
<th>6TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOPs / MB/s</td>
<td>50K/350</td>
<td>150K/2000</td>
<td>250K/2500</td>
<td>400K/4000</td>
</tr>
<tr>
<td>NAND I/F</td>
<td>133MT/s</td>
<td>200MT/s</td>
<td>400MT/s</td>
<td>400MT/s</td>
</tr>
<tr>
<td>NAND</td>
<td>3x nm</td>
<td>2x nm / 2y nm</td>
<td>1x nm</td>
<td>1y nm</td>
</tr>
<tr>
<td>Interfaces</td>
<td>SAS/SATA 6Gb</td>
<td>SAS 12 Gbps</td>
<td>PCIe 16 GT/s x4</td>
<td></td>
</tr>
</tbody>
</table>

### Current

- Current Year: 2012-2016
- Current Devices: 2.5” (MLC), 3.5” (MLC)
- Current Performance: 800GB 1.6TB 3TB 6TB

### Performance Assumptions
- Performance assumes underlying device/media can saturate interconnect
- Multiple variables could lead to higher or lower performance and capacity than shown
Data availability matters!
- Customers cannot afford to have large data sets / performance critical information off-line
- “No moving parts” != Superior Availability
  - Measurements show only +10% FIT gain using today’s flash

Requires:
- Robust controllers, increased intelligence
- Increased software intelligence
- Long-term, may benefit from improved underlying media

### NAND Type

<table>
<thead>
<tr>
<th>Type</th>
<th>PE Cycles</th>
<th>(Program/Erase)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC</td>
<td>100K</td>
<td></td>
</tr>
<tr>
<td>eMLC</td>
<td>30K</td>
<td></td>
</tr>
<tr>
<td>cMLC</td>
<td>3K</td>
<td></td>
</tr>
<tr>
<td>TLC</td>
<td>500</td>
<td></td>
</tr>
</tbody>
</table>
Serviceability – Solution Impacts

- Adjust capacity, augment functionality, update underlying technology, power capacity / management, …
- Serviceability impacts interconnect selection and associated topologies
- Do interconnect, mechanical form factor, and protocol technologies support hot insertion / removal?
  - Customer process / service availability
- Data leakage
  - Isolation, audit, regulatory, …
- Serviceability dictates the degree customers must either take down a service or require data replication / multi-node solutions to maintain service availability
Agenda

- Sustainable Customer Value
- Software (R-)evolution
  - Dual-Stack
  - Evolution
  - Revolution
- Hardware (R-)evolution
- Next Generation Persistent Media
Major Technology Transition

- Success depends upon one or both:
  - Transparency – the ability to increase solution value with minimal / zero disruption to:
    - Applications
    - Customer culture / operations
    - Management services
    - Cost models relative to delivered solution value
  - Ability to radically increase solution value relative to new “Right turn” opportunities
    - Make possible and practical new solution options
Software – Dual-stack Approach

User Applications
Volume API

Enhanced Management Services

User Applications
New TBD Persistent Memory API

Kernel Applications
Volume API

File System

Class Driver / OS Services

Raw Access Services

Enhanced Management Services

Kernel Applications
New TBD Persistent Memory API

Interface Device Driver / OS Services

Persistent Memory Hardware

User space

Kernel space
Software Evolution

- Volume software stack focused on:
  - Preserving customer investment while transparently inserting new persistent memory technology
  - Building on strong, evolving SCSI ecosystem – *no need to reinvent to deliver excellent customer solution value*
    - All OS, hypervisors, file systems, etc. support SCSI
    - All external storage sans SATA and very new PCIe-attached storage support SCSI
  - Volume software stacks have achieved significant performance gains via simple hardware substitution with minimal-to-no software changes
    - Illustrated via earlier performance data within presentation
  - Use of volume software stack does not mean no new innovation:
    - New Host Controller Interfaces and associated transports / protocols under development
      - SCSI Express, NVM Express, USB, etc.
    - Next slides use SCSI Express to illustrate on-going innovation
SCSI Express Definition

- Multiple standards → one industry name

  - SCSI Express is:
    - SCSI architecture and commands (SAM-5, SPC-4, SBC-3, etc.)
    - SCSI over PCI Express (SOP)
    - PCI Express Queuing Interface (PQI)
    - PCI Express (PCIe)
SCSI Express Architecture

- **Examples**
  - SCSI initiator device: a server with a PCI Express Root Port
  - SCSI target device: an SSD, HDD, HBA, or RAID controller
Future SCSI Express Evolution

- Potential to take advantage of more PCIe capabilities to improve queue / SGL / data buffer access
  - TLP attributes
    - No snoop, Relaxed Ordering, ID-based Ordering, Traffic Class 0 / 7, etc.
  - TLP Processing Hints
    - Steering Tags, Processing Hints
  - Process Address Space ID (PASID)
- More asynchronous events
- More protocol specific information in bridge functions
- Target mode IUs
- MULTIPLE SUCCESS IU to improve efficiency
Software Revolution

- Multiple Persistent Memory APIs under development
  - Proprietary, open source, volume OS

- New API goals:
  - Multiple application insertion points:
    - User application OS Bypass
    - Kernel application direct hardware access
    - Virtualization – Hypervisor Bypass
  - Interconnect Independent
    - PCI Express (PCIe)
    - JEDEC, et. Al. Memory Interconnects
  - Enable memory-storage collapse with new memory semantics
    - PCIe – numerous benefits but some challenges
      - Many processors have non-optimal MMIO performance
      - Non-optimal DMA “caching” techniques may be applied to compensate
    - Memory interconnects
      - Optimal for small / random load-store
      - High processor / memory consumption for large data moves
Software Revolution (cont)

- Potential challenges
  - It takes time to develop / deploy an entirely new software ecosystem
    - Early adopters help pave the way, provide key feedback but comes at a price
  - Media access latency
    - Flash has poor access latency
    - Some media have highly-variable access latency
    - Protocols and / or underlying h/w implementations must compensate – not all h/w is equal nor long-term viable
  - Interconnect, protocol, and media independence
    - May require multiple API iterations to achieve optimal solution
    - May require surfacing numerous physical and protocol attributes
      - Number of attributes may increase complexity, educational challenges, adoption rates
    - May lead to multiple API tailored to specific market needs
  - Underlying hardware will vary widely and will radically evolve
    - Packaging, density, capacity, bandwidth, latency
    - New interconnects and protocols will come to market
    - New persistent media with significantly improved characteristics over today’s flash
Agenda

- Sustainable Customer Value
- Software (R-)evolution
- Hardware (R-)evolution
- Next Generation Persistent Media
Interconnect Evolution

- **SATA**
  - Maximum signaling rate 6Gbps
  - Typically only a x1 link

- **SAS**
  - Maximum signaling rate 12 Gbps
  - x4 link possible, typically x1 or x2 to SSD / HDD

- **PCI Express (PCIe)**
  - Maximum signaling rate 8 Gbps 2012, 16 Gbps (2016)
  - x1-x16 link, x4 to SSD / HDD
  - x4, x8, or x16 to CEM form factor
  - NGFF form factor (thin enclosure focused)

- **Memory**
  - DDR, HBM, Wide IO, HMC, etc.
  - Many variables – a different talk for another day
I/O Interconnect B/W over Time

Tx + Rx GB/s

- SAS 6G x1
- SAS 12G x1
- PCIe G1 x4
- PCIe G2 x4
- PCIe G3 x4
- PCIe G4 x4
PCIe Capabilities to Consider

- Numerous PCIe technologies applicable:
  - No Snoop, Relaxed ordering, ID Ordering (improves performance)
  - TLP Processing Hints (improves performance)
  - Traffic Class 0, 7 for differentiated service at each end over VC0
  - PASID (improves virtualization, enables user-space IO, etc.)
  - Lightweight Notification (reduces overheads, simplifies comms)
  - L1 Substates (improves link power management)
  - PCIe using SFF 8639 – electrical specification for connector
    - SFF 8639 as defined by The SFF Committee
  - NGFF – new thin client-focused form factor
  - And many more…..

- Will cover two in particular
  - Downstream Port Containment (DPC)
  - OCuLink
Downstream Port Containment

- Direct processor access exposes customers to system crash or worst, silent data corruption upon asynchronous communications loss
  - Mechanical detach, controller failure, and so forth.
- DPC isolates, contains, and enables recovery

- Error or Failure in one SSD should not bring down all SSDs
- Asynchronous removal of SSDs is common
New PCIe cable technology under development

Two new x4 cables – implementations may support x1, x2, x4 @ 8GT/s
- Symmetrical external enclosure cable
  - Client, mobility focused
  - Multiple usage models: AV, external storage, peripherals, …
- Symmetrical internal enclosure cable
  - Applicable across market segments
  - Predominantly PCIe-attached storage usage model

Three cable types
- Passive copper - ~1-2 meters
- Active copper - ~2-10 meters
- Active optical – 1-300 meters

Peripheral power support - ~10W
Removes need and expense associated with Refclk
Future proof to support PCIe 4.0 at 16 GT/s required in 2016 time frame to keep up with evolving persistent memory media
Hardware Revolution

- Stacked media
  - TSV technology permeating industry
  - Significant capacity boost with new persistent memory media

- MCM / On-die
  - Creates opportunity to incorporate limited but interesting amount of memory into many solutions

- DIMM
  - Leverages existing package / interconnect to incorporate persistent memory but faces numerous challenges
Hardware Revolution (cont.)

- Express Bay (SFF 8639 connector)
  - Modular storage form factor – excellent serviceability, capacity (TB), etc.
  - Currently limited to x4 PCIe interconnect
    - Still 8-16 GB/s isn’t bad per drive 😊
      - Enclosures will vary from 1-4 drives to many 10s per enclosure
    - Good performance match for new persistent media

- PCIe CEM
  - High-capacity, serviceability challenged form factor
  - MMIO-based memory semantics performance challenged

- Yet to be developed modules
  - New packaging, new solution value-add opportunities
Agenda

- Sustainable Customer Value
- Software (R-)evolution
- Hardware (R-)evolution
- Next Generation Persistent Media
Persistent Memory Barriers

- General Issues
  - $$ / Bit
  - Power consumption
  - Durability
  - Performance vs. persistence

- NAND Future is Bleak
  - Geometry Shrinks
    - Endurance decreases (already at a minimum)
    - Harder to distinguish states in MLC devices
    - Higher probability of errors (read disturbance)
  - Write performance decreasing due to array structure
## The Promise of New Media

<table>
<thead>
<tr>
<th></th>
<th>Memristor</th>
<th>PCM</th>
<th>STTRAM</th>
<th>DRAM</th>
<th>Flash</th>
<th>HDD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density</strong></td>
<td>8-16x DRAM</td>
<td>2012: Very Low</td>
<td>2012: Very Low</td>
<td>Evolving</td>
<td>2012 Low relative to</td>
<td>Superior to DRAM</td>
</tr>
<tr>
<td></td>
<td>Density</td>
<td>2016: 1x DRAM??</td>
<td>2016: 1x DRAM??</td>
<td></td>
<td>DRAM</td>
<td></td>
</tr>
<tr>
<td><strong>Energy per bit</strong></td>
<td>1–3</td>
<td>2–27</td>
<td>0.1</td>
<td>2</td>
<td>10^3–10^4</td>
<td>10^7–10^8</td>
</tr>
<tr>
<td>(pJ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Read time (ns)</strong></td>
<td>~10</td>
<td>20–70</td>
<td>10–30</td>
<td>10–50</td>
<td>25,000</td>
<td>5–8x10^6</td>
</tr>
<tr>
<td><strong>Write time (ns)</strong></td>
<td>20–50</td>
<td>50–500</td>
<td>13–95</td>
<td>10–50</td>
<td>200,000</td>
<td>5–8x10^6</td>
</tr>
<tr>
<td><strong>Retention</strong></td>
<td>Large N Years</td>
<td>Years</td>
<td>Variable:</td>
<td>&lt; Second</td>
<td>Year</td>
<td>Decades</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>micro-sec to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Endurance</strong></td>
<td>&gt;&gt;10^10</td>
<td>10^7–10^8</td>
<td>10^15</td>
<td>&gt;10^18</td>
<td>10^5–10^6</td>
<td>&gt;10^16††</td>
</tr>
<tr>
<td>(cycles)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Caveat: The variance in the published literature is generally large

† Range includes read, write, and erase
†† Hard error rate guarantee for HDDs
Summary – Persistent Memory Market Evolving

- Delivering solid performance gains
- Spawning numerous products / solutions
  - From evolutionary to revolutionary
- Reinvigorating the persistent memory dialog
  - Where and when?
  - Transparent volume vs. right-turn solutions?
  - To what degree will memory-storage collapse?
  - How large must the processor physical address grow?
- And so forth…
Summary – Revolutionary changes are coming

- New protocols, new hardware, new software

- New media will fundamentally change computer and application architecture

- Persistent memory impact will be as profound as mobile technologies on customer experiences

- ....we are only beginning the journey!!
Thank You

Questions???
Performance Comparison

Solid State Enterprise Classification

- Random 100% Read IOPS
- Random 90/10% Mix IOPS
- Random 70/30% Mix IOPS
- Random 100% Write IOPS
- Max Latency ms

Measured performance data is for pre-conditioned flash drives, i.e. steady-state operation.
### SCSI Express - Applicable to Many Use Cases

#### From SSDs to external storage

1. **Standard PCI Express CEM form factor and custom form factor card SSDs**

2. **Express Bay SSDs**

3. **Express Bay HDDs**

4. **RAID controllers (SAS or PCI Express back-end)**

5. **HBAs (SAS, FC, iSCSI, UAS, or PCI Express)**

6. **Interface for internal PCI Express busses in external storage systems (target mode)**