NVMe over Fabrics for the Next Generation Data Center

Rupin Mohan
Director R&D, CTO (SAN)
Hewlett Packard Enterprise
HPE Storage
Agenda

- Data Center Trends
- The new I/O Stack - NVMe
- Simple, Invisible Fabric
- Performance Data
- Next Steps
Data Center Trends
Disaggregation – What does it mean?

Orchestration

Industry Standard

Hybrid Cloud Enabled

Total Customer Experience

Lowest Total Cost of Ownership

Unlimited Bandwidth

Workload driven

East-West-North-South Traffic

Low Latency

Software Defined

Scalable

Secure
The New I/O Stack
Non-Volatile Memory express (NVMe)

A new storage protocol focused on SSDs, replaces traditional I/O stacks built on SCSI protocol, which is optimized for spinning media. It leverages PCIe instead of SAS/SATA for greater bandwidth, IOPS and reduced latency

<table>
<thead>
<tr>
<th>Designed to move beyond HDDs</th>
<th>Standard interface for Solid State Media</th>
<th>A new protocol</th>
<th>Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Built for high performing storage devices</td>
<td>Leverages PCIe for lower latency, higher IOPS than SATA or SAS</td>
<td>New streamlined protocol and command-set (replaces SCSI)</td>
<td>No, it's not high performance media—it's a protocol to access to it!</td>
</tr>
</tbody>
</table>
A new language for accessing solid state media

Traditional Storage Arrays
1. Storage Controller runs SCSI
2. Front end FC/iSCSI
3. Backend SAS/SATA
4. Software Feature Rich based on SCSI

Hybrid case
1. Storage Controller runs SCSI. Upgraded back end (partial/full)– Controller does SCSI-NVMe translation with NVMe drives in the backend
2. Memory-Driven Flash
3. Software Feature Rich based on SCSI

Next Gen. Storage Arrays
1. Controller runs NVMe
2. Backend NVMe Drives (PCIe, NVMe over Fabrics)
3. Frontend NVMe (FC-NVMe, NVMe over Ethernet)
4. Software Features running NVMe, expect parity in 3 years
# I/O Stack Evolution

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applications</td>
<td>Enterprise Apps taking advantage of SPDK (RDMA)</td>
</tr>
<tr>
<td>OS – Storage Stack</td>
<td>Volume Manager optimized to NVMe, new protocol</td>
</tr>
<tr>
<td>Host Adapter – Driver</td>
<td>FC-NVMe, NVMe over Ethernet (RoCEv2, TCP) – Lim. OS Support</td>
</tr>
<tr>
<td>SAN – Switch</td>
<td>FC, Ethernet switches</td>
</tr>
<tr>
<td>Host Port on Array – Front End Fabric</td>
<td>FC-NVMe, NVMe over Ethernet (RoCEv2, TCP)</td>
</tr>
<tr>
<td>SCM - Cache</td>
<td>3D X-point as read cache (Memory Driven Flash)</td>
</tr>
<tr>
<td>Storage Controller Core</td>
<td>Transition to NVMe, including all features (RC, etc)</td>
</tr>
<tr>
<td>Drives in Head Shell</td>
<td>Partial # of NVMe drives to full cage</td>
</tr>
<tr>
<td>JBOF – for scale</td>
<td>Scale to multiple shelves over PCIe or Switching Fabric</td>
</tr>
</tbody>
</table>

**Management of NVMe Namespaces**
- Redfish/Swordfish API's

---

2019 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
Faster Storage Needs Faster Networks

As drive and controller latency decrease, minimizing software and network latency becomes increasingly important.

SSD NAND technology offers ~100X reduction in latency versus HDD

NVMe* eliminates 20 µs of latency today

Next Gen NVM needs NVMe to deliver 4KB operations in under 10 µs

Source: Flash Memory Summit 2016, Amber Huffman, Chairman NVMe Working Group

Source: Flash Memory Summit 2016, Amber Huffman, Chairman NVMe Working Group

2019 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
Simple Invisible Fabric
NVMe over Fabrics

Shared storage will require NVMe primary arrays to have FABRIC connectivity

- Initially on the **back-end** of the array and on the **front-end** as well
- Back-end always leads front-end in storage development
Which transport wins where and when?

**Front End (in order of priority)**

1. FC-NVMe
   - Time to market and less complexity
   - Upgrade opportunity using existing infrastructure, on-prem
2. NVMe over Ethernet over RoCE v2
   - 25GbE/100GbE could be a driver (speeds & feeds)
3. NVMe over TCP
   - Cloud is going to be big driver here
   - Need full stack offload NIC’s?

**Storage Controller**

1. SCSI
2. NVMe

**Back End (in order of priority)**

1. Low-Mid range arrays - PCIe
2. NVMe over Ethernet over RoCE v2 for scalability over multiple disk shelves
3. InfiniBand
The landscape today....

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Latency</th>
<th>Scalable</th>
<th>Performance</th>
<th>Enterprise Footprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibre Channel</td>
<td>Lower</td>
<td>Yes</td>
<td>High</td>
<td>Reliable Storage Fabric</td>
</tr>
<tr>
<td>RoCEv2</td>
<td>Lowest</td>
<td>Yes</td>
<td>High</td>
<td>Negligible</td>
</tr>
<tr>
<td>iWARP</td>
<td>Medium</td>
<td>Yes</td>
<td>Medium</td>
<td>Negligible</td>
</tr>
<tr>
<td>TCP</td>
<td>High</td>
<td>Yes</td>
<td>Medium</td>
<td>Medium with iSCSI</td>
</tr>
<tr>
<td>InfiniBand</td>
<td>Lowest</td>
<td>Limited</td>
<td>High</td>
<td>None</td>
</tr>
</tbody>
</table>
Performance
NVMe-OF SAN
Next Steps
Key Design Takeaways

- Low latency is key. Intense focus on performance optimizations
- Make the fabric simple and invisible -- Automation / Orchestration
- New storage architectures are in development, across the industry
- NVMe SAN offers significant opportunities to service low latency, high performance disaggregated storage architectures
- Bandwidth and IOPS requirements from applications are going to scale
- Hybrid IT is bringing Cloud and Enterprise closer and in future will get more aligned
- Total customer experience is key
- All this at lowest TCO