Using a PCIe-Based Switch Module to Enhance Enterprise Storage Architecture

Chi-Lin Tom
Memoright
chilintom@memoright-usa.com
Legacy Storage Solution Architecture

2U Server – 16 Slots
12 HDD & 4 SSD

1 SSD = 1TB as Cache
1 HDD = 6TB as Storage
• HBA (Host Bus Adaptor) quickly becomes a bottleneck and contention point.

• Some HDD slots (typically 4) are reserved for SSD as cache and the storage space is reduced.
Legacy Storage Solution Architecture

1 SSD = 512GB
as Cache
1 HDD = 6TB
as Storage

72TB Total Capacity

1 PCIe Switch Module = 4TB
has 4 M.2 NVMe
as Cache

2U Server - 16 Slots
16 HDD & 4 SSD

96TB Total Capacity

1 HDD = 6TB
as Storage
Advantages of Contemporary Storage Solution

- Increase storage space
- Reduce latency
- Improve performance
- Switching is full duplex in parallel with SAS HBA traffic
- Affinity of each M.2 to one or many NUMA CPU cores
- 4 M.2 adds flexibility of resource aggregation
- Provide PCIe pass through to VMs
- No driver needed
Why M.2 PCIe NVMe?

• Higher Bandwidth
  • SATA III SSD: 6Gb
  • Gen 2 x 2: 8Gb
  • Gen 3 x 2: ~16Gb
  • Gen 3 x 4: ~32Gb
• Low latency
• Native PCIe instead of AHCI Mode
<table>
<thead>
<tr>
<th></th>
<th>AHCI</th>
<th>NVMe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Queue Depth</td>
<td>1 command queue; 32 commands per queue</td>
<td>65536 queues; 65536 commands per queue</td>
</tr>
<tr>
<td>Uncacheable register accesses</td>
<td>6 per non-queued command; 9 per queued command</td>
<td>2 per command</td>
</tr>
<tr>
<td>(2000 cycles each)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSI-X and interrupt steering</td>
<td>single interrupt; no steering</td>
<td>2048 MSI-X interrupts</td>
</tr>
<tr>
<td>Parallelism and multiple threads</td>
<td>requires synchronization lock to issue a command</td>
<td>no locking</td>
</tr>
<tr>
<td>Efficiency for 4 KB commands</td>
<td>command parameters require two serialized host DRAM fetches</td>
<td>gets command parameters in one 64 Bytes fetch</td>
</tr>
</tbody>
</table>
Legacy Solution CPU Utilization

- HDD 150 IOPS/Lane
- SSD 80,000 IOPS/Lane
- 72TB HDD Stack + 4TB SSD Cache
- 6Gb/s Throughput
- ~30% of CPU Utilized

Bottleneck

HBA

Flash Memory Summit 2015
Santa Clara, CA
Contemporary Solution CPU Utilization

- 150 IOPS/Lane
- 6Gb/s Throughput
- 96TB HDD Stack
- 1600 MHz
- DRAM
- 4TB Cache
- 90,000 IOPS, 8Gb/s /Lane
- ~75% of CPU Utilized

HBA

Flash Memory Summit 2015
Santa Clara, CA
• Virtualization software often is starved for IOPs and PCIe pass-through can improve the CPU utilization from the average 30% to the desired 75%.
Applications

• Storage Appliance
• EDA
• NAS Gateway
• Others
Application: Storage Appliance

- 1 SSD = 512GB
- 1 HDD = 6TB as Storage
- 1 PCIe Switch Module = 4TB has 4 M.2 NVMe as Cache
- 2U Server – 16 Slots
- 16 HDD & 1 Switch Module
- 96TB Total Capacity
Applications: Electronic Design Automation (current)

Latency: 10-100ms

Bottleneck

Net App NAS

10Gb

10Gb

Server 1

10Gb

Server 2

10Gb

Server N

Ethernet Switch

Node 1 Node 2

Node 3 Node 4

Flash Memory Summit 2015
Santa Clara, CA
Applications: Electronic Design Automation (new)

90 percent of time
One server to get information from itself or next Server /Node

Latency: 0.1-1ms

Applications:
- Electronic Design Automation (new)

Latency: 10-100ms

Bottleneck

Node 1 | Node 2
Node 3 | Node 4

Server 1 (add m.2 as cache)

Server 2 (add m.2 as cache)

Server N (add m.2 as cache)

Net App NAS

Ethernet Switch
## Application: NAS Gateway

<table>
<thead>
<tr>
<th>Access Protocols</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIFS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Management</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Point of Administration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Defined Policies</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Availability</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Way Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QoS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Services</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deduplication</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encryption</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Caching</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Switch Metadata</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRAM Cache</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVMe Switched Fabric</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SAN Integration</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>iSCSI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCoE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AoE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thank You!