Next Generation Low Latency SAN’s

Rupin Mohan
Chief Technologist,
Storage Networking
Hewlett-Packard

Sep 20th, 2015
Santa Clara, CA
Agenda

- Introduction
- How do we measure performance?
- Storage Protocol Comparison
- Current state of the union
- What are the new options?
  - A look into the future
- Summary / Key takeaways
How do we measure performance?

- **IOPS** – I/O’s per second – a measure of the total I/O operations (reads and writes) issued by applications

- **Bandwidth** – a measure of the data transfer rate, or I/O throughput, MB/s

- **Latency** – time taken to complete an I/O request, also known as response time. Usually measured in milliseconds (1/1000 of a second). Going forward, in microseconds (1/1000 of a millisecond)
### Storage Protocol Comparison

<table>
<thead>
<tr>
<th>Attribute</th>
<th>IB</th>
<th>FC</th>
<th>FCoE (DCB Ethernet)</th>
<th>iSCSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth (Gbps)</td>
<td>56</td>
<td>8/16/32</td>
<td>10/25/40/100</td>
<td>10/25/40/100</td>
</tr>
<tr>
<td>Adapter Latency</td>
<td>25 us</td>
<td>50 us</td>
<td>200 us</td>
<td>Wide range(w/o offload / w/o offload)</td>
</tr>
<tr>
<td>Switch Latency per hop</td>
<td>100-200ns</td>
<td>700 ns</td>
<td>200ns to 1 us</td>
<td>200ns to 1 us</td>
</tr>
<tr>
<td>Adapter</td>
<td>HCA</td>
<td>HBA</td>
<td>CNA</td>
<td>NIC</td>
</tr>
<tr>
<td>Convergence - Single L2 network</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Separate Network for management</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>App Changes Req</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Routable</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Kernel Bypass</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Max Frame size</td>
<td>2K</td>
<td>2K</td>
<td>9K</td>
<td>9K</td>
</tr>
<tr>
<td>Maturity</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>
Current state of the union

- Using All Flash 3PAR 7450
- End to end Gen 5 (16Gb) Infrastructure
- Latency at host (end-to-end) around ~0.5 ms
- 75% reduction in latency compared to 8Gb FC
What are the new options?

- Gen 6 Fibre Channel
- Storage protocols on RDMA
  - iSER
  - SMB Direct
- Transport options for RDMA
  - RoCE
  - iWARP
- NVMe over Fabrics
  - NVMe over RDMA
  - NVMe over FC
Gen 6 Fibre Channel

- 32 Gb/s single lane
- 128 Gb/s multi lane
- How is it lower latency?
  - Lower latency through higher clock rate
  - Possible smaller ASIC geometries
RDMA – Remote Direct Memory Access

- Introduction
  - Accelerated IO delivery model, allowing application software to bypass most layers of software and communicate directly with the hardware
  - Requires new programming model: “verbs” rather than “sockets”
  - Protocol options: Block (iSER) / File (SMB Direct)
  - Transport options: RoCE, iWARP, Infiniband

- RDMA Benefits
  - Low latency, also important is latency jitter
  - High Throughput
  - Zero copy capability, OS / Stack bypass
  - Avoid CPU context switching, interrupt coalescing

- Bulk of the work is done by the Target
  - Read operation, translates to a write by Target
  - Write operation, translates to a read by Target
iSER and SMB Direct over Ethernet

- iSER
  - iSCSI extensions over RDMA
  - Mature protocol
  - Limited OS stacks in both functionality and support. Linux (yes), Windows (yes), VMware (?)

- SMB Direct
  - NFS using Direct I/O
  - Mature protocol
  - Limited OS stack support. Linux (yes), Windows (yes), VMware (?)
## RoCE versus iWARP

<table>
<thead>
<tr>
<th>RoCE</th>
<th>iWARP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Needs DCB Switching infrastructure</td>
<td>Does not require DCB Switches</td>
</tr>
<tr>
<td>Routability comes with v2</td>
<td>Routable as it runs over TCP/IP</td>
</tr>
<tr>
<td>End to end congestion management is a big issue.</td>
<td>TCP solves the congestion management issue but adds latency</td>
</tr>
<tr>
<td>Higher cost solution (DCB)</td>
<td>Lower cost solution</td>
</tr>
<tr>
<td>DCB configuration on switches is cumbersome</td>
<td>Switch config simpler</td>
</tr>
<tr>
<td>Lowest Latency</td>
<td>Low Latency</td>
</tr>
</tbody>
</table>
Latency Step Function

- FC, FCoE, iSCSI
  - 0.4 – 2.5 ms

- ISER, SMB3
  - 100 – 350 us

- NVMe
  - 50 – 150 us
  - Leverage RDMA SAN

Today 2015-2017 2017 & beyond
“READ” Response time with increasing I/O size (Lower is better)
NVME OVER FABRICS
NVMe Transports

- Two new fabric transport projects for NVMe
  1. NVMe over Fabrics – Being defined in a subgroup of the NVM Express group
    - NVMe over RDMA
  2. NVMe over FC (FC-NVMe) – New T11 project to define an NVMe over Fibre Channel Protocol
NVMe over Fabrics

- Being defined by a Technical sub-group of the NVM Express group
  - Defined as upper level protocol on top of OFED RDMA verbs
  - Fabric agnostic
    - Supports RDMA fabrics – Ethernet (iWARP, RoCE), Infiniband
    - Support for other fabrics – FC
  - Complies to the NVMe programming model
Key Takeaways

- Fibre Channel is low latency but there is competition
- iSER and SMB Direct are in product development right now
- Gen 6 Fibre Channel is in product development, FC is also working on lowering latency
- NVMe over Fabrics is under standards development now
- Data center networks are key to application latency
Thank You

- Rupin Mohan
  - Rupin.mohan@hp.com
  - +1-774-245-2947