Competing Technologies and Architectures for Networked Flash Storage

Ethernet/InfiniBand/OmniPath/PCI/FibreChannel/SAS

Asgeir Eiriksson
Chelsio Communications
Introduction

- API are evolving for optimal use of SSD
- FC and SAS falling behind the speed curve
- Ethernet, IB and OmniPath on same PHY curve
- PCI on different slightly slower PHY curve
- Ethernet, IB, and OmniPath
  - Have different reach
  - Same protocol stack efficiencies
### Introduction: speeds and feeds

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Bandwidth (Gbps)</th>
<th>Reach</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>3, 6, 12</td>
<td>Rack</td>
</tr>
<tr>
<td>Fibre Channel</td>
<td>4, 8, 16, 32</td>
<td>Rack, Data center</td>
</tr>
<tr>
<td>PCI x1/2/4/8/16</td>
<td>8, 16, 32, 64,128</td>
<td>Rack</td>
</tr>
<tr>
<td>Ethernet</td>
<td>1, 2.5, 5, 10, 25, 40, 50, 100</td>
<td>Rack, Data Center, LAN, MAN, WAN</td>
</tr>
<tr>
<td>Infiniband</td>
<td>8, 16, 32, 56, 112</td>
<td>Rack, Data Center</td>
</tr>
<tr>
<td>OmniPath</td>
<td>100, 200</td>
<td>Rack, Data Center</td>
</tr>
</tbody>
</table>
PHY SERDES (single lane) curves

Infiniband and Ethernet
- Same PHY curve
- Same speed curve

<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2015</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>6</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>Infiniband</td>
<td>10</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>Ethernet</td>
<td>10</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>
Traditional Scale Out Storage

Data Center Network

Application Controller

Storage Cluster Network

Storage Controller

Storage Controller

Storage Controller

Disaster Recovery
60-300+ miles
Traditional Scale Out Storage

- Preserves software investment
- Realizes some of the SSD speedup benefits
- Disaster Recovery (DR) requires MAN or WAN
Shared Server Flash

Disaster Recovery 60-300+ miles

Ethernet, Infiniband, or Omnipath Fabric

NVMe Storage Server

NVMe Storage Server

NVMe Storage Server
Shared Server Flash

- Ethernet or IB or OmniPath fabric
- PCIe fabric not sufficient reach or scaling
- RDMA required for sufficient efficiency
  - IB and OmniPath use RDMA
  - Ethernet has RoCE, iWARP and iSCSI with RDMA
- Disaster Recovery (DR) requires MAN or WAN
API

Applications

APIs, Libraries

File System
Lustre, XFS, ext3, ext4, NVMFS

Flash Devices
I/O and Primitives (Atomics, etc.)
Memory, Persistent Memory
API

- Preserve software investment
- Alternatively jump directly to native SSD API
Ethernet, Infiniband, OmniPath

- Infiniband, OmniPath
  - Reliable link layer
  - Credit based flow control
- Ethernet
  - Ubiquitous
  - Pause and Prioritized Pause (PPC) for lossless operation that propagates through some switches and fewer routers
  - Flow Control and Reliability at higher layer e.g. TCP, and IB Transport Layer for RoCE
## Comparing Ethernet Options

<table>
<thead>
<tr>
<th>DCB Required</th>
<th>Reach</th>
<th>IP routable</th>
<th>RDMA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FCoE</strong></td>
<td>√</td>
<td>Rack, LAN</td>
<td>√</td>
</tr>
<tr>
<td><strong>iSCSI</strong></td>
<td>No</td>
<td>Rack, datacenter, LAN, MAN, WAN Wired, wireless</td>
<td>√</td>
</tr>
<tr>
<td><strong>iWARP</strong></td>
<td>No</td>
<td>Rack, datacenter, LAN, MAN, WAN Wired, wireless</td>
<td>√</td>
</tr>
<tr>
<td><strong>RoCEv2</strong></td>
<td>√</td>
<td>Rack, LAN, datacenter</td>
<td>√</td>
</tr>
</tbody>
</table>
Comparing Ethernet Options

- iSCSI, iWARP
  - Use DCB when it is available but not required for high performance

- iSCSI
  - Has RDMA WRITE and accomplishes RDMA READ by using an RDMA WRITE from other end-point
  - Concurrent support for legacy soft-iSCSI
Comparing Ethernet Options

- RDMA bypasses the host software stack
  - RoCE
  - iWARP
  - iSCSI with offload
NVMe over RDMA fabrics

- Bypass
- RDMA
- Bypass
- RDMA
iSER with offload

- Bypass
- RDMA
NVMe over fabrics Option 1

RDMA
- Control Plane on host or ASIC/FPGA/SoC
- Data Plane PCIe-host-PCIe or PCIe only
NVMe over fabrics Option 2

RDMA
- Control Plane Intel/ARM host
- Data Plane PCIe-ASIC/FPGA/SoC-PCIe

Fabric
- Infiniband, OmniPath, or Ethernet RoCEv2/iWARP

Intel/ARM host
Control Plane

ASIC/FPGA/SoC
Data Plane

RDMA NIC
PCIe

RDMA NIC
PCIe

PCIe Fabric
PCIe

PCIe

NVMe SSD

NVMe SSD

NVMe SSD
NVMe over fabrics comparison

- **Option 1**
  - Flexible
  - Extra latency incurred by copy/copies

- **Option 2**
  - Minimizes latency by removing host and host memory system from data path
Summary

- API are evolving for optimal use of SSD
- Ethernet, IB, and OmniPath
  - On same SERDES PHY (single lane) curve
  - Have different reach
  - Same protocol stack efficiencies
Questions?

Asgeir Eiriksson
asgeir@chelsio.com