Next generation Ethernet connect to All FLASH: iSER and NVMeF

Subhojit Roy
Senior Technical Staff Member, IBM
Agenda

- Flash growth and dimensions of change
- Shared SAN storage requirements
- What is iSER?
- What is NVMeF?
- All FLASH on SAN – roadmap
- What are we doing at IBM?
Flash Growth and Dimensions of Change...

<table>
<thead>
<tr>
<th>Workloads</th>
<th>Traditional Enterprise Workloads $\rightarrow$ New age Flash workloads (Tier0: SAP HANA, RT Analytics, Tier1: OLTP, VDI, Social Media Apps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage SW Architecture</td>
<td>Kernel Mode $\rightarrow$ User Mode (SPDK)</td>
</tr>
<tr>
<td>Upper Level Protocols (ULP)</td>
<td>FCP $\rightarrow$ iSER &amp; NVMeoF (Eth &amp; FC)</td>
</tr>
<tr>
<td>Layer2 Interconnect (L2)</td>
<td>FC 8G/16G $\rightarrow$ Eth 10G $\rightarrow$ FC32G $\rightarrow$ Eth RDMA 25/40/50/100G</td>
</tr>
<tr>
<td>Storage Media</td>
<td>HDD $\rightarrow$ SAS SSDs $\rightarrow$ PCIe NVMe $\rightarrow$ PM (3DXP)</td>
</tr>
<tr>
<td>Storage HW Architecture</td>
<td>AF Server $\rightarrow$ AF-HyperConverged $\rightarrow$ AF-Arrays $\rightarrow$ AF Disaggregated Storage</td>
</tr>
</tbody>
</table>

- Flash is driving change
- Change is multidimensional
- Change would not be all at once. One dimension at a time
Shared SAN storage requirements

- Multipath (for high availability)
- Reservation (SCSI Persistent Group Reservation)
- vVols (offload Copy Services to Storage Controller)
- Data movement/migration without involving host CPU (XCOPY, ODX)
- Unmap: Optimize space deallocation on Thin Provisioned storage
- Atomic Compare And Write (CAW): Alternative to SCSI3 Reservations
- Error Handling: Abort Task, Abort Task Set, LUN Reset, Target Reset etc.
What is iSER?

Operating System/Applications

SCSI Layer

FCP

iSCSI

TCP

IP

Fiber Channel Driver

NIC Driver

iWARP Driver

RoCE Driver

Infiniband Adapter Driver

FC HBA

NIC

iWARP rNIC

RoCE rNIC

IB HCA

OFED IB verbs

Software Drivers

Hardware
iSER: Current state of affairs

*iSER is iSCSI with RDMA Data path*

- **Low Latency, Low CPU Utilization**
  (Eliminates copies to/from TCP/IP buffers)
- **No Changes to iSCSI administration**
  (vSphere, Widows, OpenStack work as is)
- **Vendor and Technology Independent**
  (works on iWARP, RoCE & Infiniband HCAs)
- **Works on Standard Ethernet equipment**
  (10G and 25/50/100G switches)
- **Enterprise applications just work!**
  (vVols, Clustering, Multipath etc.)

- **Suitable for All FLASH over High Speed Eth**
  (10, 25, 40, 50, 100 Gbps and beyond)
- **No disruption to administration model**
- **Fits well into Software Defined Storage (SDS) paradigm**
- **Cost Savings**
- **Ideal for shared storage**
  (both for FLASH and HDD)

*iSER is ready for Shared All FLASH SAN storage today!*
What is NVMeF?

Operating System/Applications

- SCSI
- iSCSI
- iSER
- iWARP Driver
- RoCE Driver
- Infiniband Adapter Driver

OFED IB verbs

- iWARP Driver
- RoCE Driver
- IB HCA

Software Drivers

- FCAdapter Driver

Hardware

- iWARP rNIC
- RoCE rNIC
- IB HCA
- iSER

NVMeF
NVMeF: Current state of affairs

Low Latency, Low CPU Utilization (Primarily scuts down on Host s/w stack) — Suitable for All FLASH over High Speed Eth (10, 25, 40, 50, 100 Gbps and beyond)

New Administrative Model — Changes to vSphere, Openstack etc.

Vendor and Technology Independent (works on iWARP, RoCE & Fiber Channel) — Yes and No! Need common user space layer 2

Works on Standard Ethernet equipment (10G and 25/50/100G switches) — Cost Savings

Applications must change to exploit parallelism — Applications need to transform yet

NVMeF is still evolving to adapt to Shared Storage Applications!
All Flash SAN roadmap!

**Technology**
- HDD
- SCSI
- FC
- HDD
- SCSI
- iSCSI/Eth
- HDD/SSD
- SCSI
- FC
- Flash/NVMe
- SCSI
- NVMeoF
- Flash/3DXP
- NVMeoF
- iSER/NVMeF

**Usecase Details**
- SCSI/FC ruled the Enterprise Shared Storage World
- iSCSI started penetration in low end market
- Flash environments used SCSI with FC as an interconnect
- PCIe NVMe Flash debuted as high performance storage
- iSER came as an alternative to FC for connecting external Flash
- NVMeoF specification and technology matured for Tier0 usecase
- NVMeoF/Eth RDMA matured for Shared Storage Usecase

**Timeline For Maturity**
- 1990s – 2010: HDD/SCSI/FC Rule
- 2007–2016: SSD/Flash media evolution, maturity
- 2015–…: NVM evolution
- 2016 to 2020: iSER/SCSI
- 2019/2022: iSER/NVMeF

**Adoption Status**
- < 2010
- 2010–…
- 2012–2016
- 2015–…
- 2016–…
- 2017
- 2019–…
What are we doing at IBM?

- Host Attach over iSER – Linux, VMWare
- Clustering over iSER
### iSER vs Fibre Channel

<table>
<thead>
<tr>
<th>I/O</th>
<th>iSER (40Gb)</th>
<th>Fibre Channel (16Gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read 4KiB</td>
<td>50 (us)</td>
<td>80 (us)</td>
</tr>
<tr>
<td>Write 4KiB</td>
<td>139 (us)</td>
<td>195 (us)</td>
</tr>
<tr>
<td>Read 64KiB</td>
<td>95 (us)</td>
<td>196 (us)</td>
</tr>
<tr>
<td>Write 64KiB</td>
<td>209 (us)</td>
<td>337 (us)</td>
</tr>
</tbody>
</table>

*iSER: Fiber Channel benefits minus the additional costs*
Thank You!