iSCSI Extensions over RDMA (iSER): A New High-Speed Storage Protocol for Flash Memory

Rob Davis
robd@mellanox.com
The FASTEST Storage Protocol: iSER
The FASTEST Storage: Flash

- What it is: iSCSI With RDMA Transport
  - Runs over Ethernet or InfiniBand at speeds up to 100Gb/s
  - Works with all applications that support SCSI/iSCSI

- Benefits
  - Lets Flash Shine: Highest bandwidth, Highest IOPs, Lowest Latency
  - iSCSI storage features, management and tools (security, HA, discovery...)
  - Faster than iSCSI, FC, FCoE; Easier to manage than SRP

- Ideal for Flash Storage Applications
  - Latency-sensitive workloads; Small, random I/O
    - Databases, Virtualization, VDI
  - Bandwidth-sensitive workloads; Large, sequential I/O
    - Post production, oil/gas
Why iSER? Flash Memory Creates IO Bottleneck

Storage Media Technology

Access Time (micro-sec)

10,000x improvement

Networked Storage

Protocol and Network

Access Time (micro-sec)
iSER Has No TCP/IP Stack
Protocol / Transport Comparison

<table>
<thead>
<tr>
<th>Transport</th>
<th>InfiniBand</th>
<th>Ethernet RoCE</th>
<th>Ethernet TCP</th>
<th>FCoE</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>40/56/100 Gb/s</td>
<td>10/25/40/50/100 Gb/s</td>
<td>10/25/40/50/100 Gb/s</td>
<td>10/40 Gb/s</td>
<td>8/16/32 Gb/s</td>
</tr>
<tr>
<td>RDMA</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Routable</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
iSER Protocol Overview (Read)

- **SCSI Reads**
  - Initiator Send Command PDU (Protocol data unit) to Target
  - Target return data using RDMA Write
  - Target send Response PDU back when completed transaction
  - Initiator receives Response and complete SCSI operation
- **SCSI Writes**
  - Send Command PDU (optionally with Immediate Data to improve latency)
  - Map R2T to RDMA Read operation (retrieve data)
  - Target send Response PDU back when completed transaction
Requirements to Deploy iSER

- Application(s) that can use SCSI/iSCSI
  - All applications that use SCSI-based block storage work with iSER
- OS or Hypervisor that Supports an iSER initiator
  - Today: Linux & VMware ESXi, Oracle Solaris
  - Expected soon: Windows, FreeBSD
- iSER Storage Target (unless Hyper Converged)
  - NetApp, HP SL4500, Oracle ZFS, Violin Memory, Zadara, Saratoga Speed
  - Create in Linux using LIO, TGT, or SCST target
- Network that supports RDMA
  - Adapters support InfiniBand, iWARP or RoCE
  - Switches support InfiniBand or Ethernet DCBx with ECN
Higher Bandwidth and IOPS with Less CPU Utilization than iSCSI
VMWare: iSER over ESXi

iSER/RDMA has 10X Bandwidth Advantage vs TCP/IP and 2.5X IOPs

Test Setup: ESXi 5.0, 2 VMs, 2 LUNS per VM
iSER eliminates storage bottlenecks in VDI deployments
- iSER accelerates the access to cache over RDMA
- 140 Virtual desktops with iSER/RoCE vs. 60 virtual desktops over TCP/IP

Intel 10GbE, iSCSI/TCP

ConnectX3 10GbE, iSCSI/RDMA (iSER)
ROI Comparison Shows iSER Value

<table>
<thead>
<tr>
<th>Interconnect</th>
<th># Virtual Desktop per Server</th>
<th># Servers</th>
<th># Switches</th>
<th>CapEx</th>
<th>CapEx per Virtual Desktop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10GbE with TCP/IP</td>
<td>60</td>
<td>84</td>
<td>2</td>
<td>$3,418,600</td>
<td>$684</td>
</tr>
<tr>
<td>10GbE with RoCE</td>
<td>140</td>
<td>36</td>
<td>1</td>
<td>$1,855,900</td>
<td>$371</td>
</tr>
</tbody>
</table>

iSER Delivers $1.5M CapEx Savings For VDI Deployments

http://www.mellanox.com/related-docs/whitepapers/SB_Virtual_Desktop_Infrastructure_Storage_Acceleration_Final.pdf
iSER OpenStack Support

- **Built-in OpenStack components and management**
  - No additional software required
  - RDMA is already inbox and ready for OpenStack users


Virtual Private Storage Array Adds Mellanox 40 Gigabit Ethernet Technology for iSER RDMA to Deliver Lower Latency and Higher Transactional Rate for Databases

VMworld, San Francisco, CA – August 25, 2014 – Zadara™ Storage today announced a new high performance storage as a service (STaaS) solution offering for private clouds – the Zadara iSER VPSA™ – using a first of its kind Ethernet transport mechanism for exceptional performance at reduced costs. The latest generation of its award-winning Virtual Private Storage Array™ services, iSER VPSA takes advantage of iSCSI Extensions for RDMA (iSER) using 40 Gigabit Ethernet to cut latency and boost application performance. Zadara Storage is the first vendor to offer a storage array supporting this next generation Ethernet-based transport. While all block-based applications running on Zadara iSER VPSA will see significant improvements, latency-sensitive databases will see a marked improvement in their transaction per second (TPS) count.
NetApp

- **Models:**
  - E5500/5600: Hybrid HDD/SSD
  - EF550/560: All-flash

- **Performance:**
  - 530K IOPS
  - 12 GB/s
Saratoga-Speed

- **Model:**
  - AltamontXP: All-flash
  - 40Gb/s Ethernet

<table>
<thead>
<tr>
<th>I/O Type</th>
<th>Read-Write Mix</th>
<th>4KB Blocks</th>
<th></th>
<th>32KB Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bandwidth (KB/s)</td>
<td>IOPS</td>
<td>Bandwidth (KB/s)</td>
</tr>
<tr>
<td>Sequential</td>
<td>100% Read</td>
<td>8,618,900</td>
<td>2,154,725</td>
<td>10,981,200</td>
</tr>
<tr>
<td>Sequential</td>
<td>100% Write</td>
<td>7,758,300</td>
<td>1,939,575</td>
<td>10,654,000</td>
</tr>
<tr>
<td>Sequential</td>
<td>80% Read, 20% Write</td>
<td>2,866,885</td>
<td>716,721</td>
<td>9,400,997</td>
</tr>
<tr>
<td>Random</td>
<td>100% Read</td>
<td>902,709</td>
<td>225,677</td>
<td>6,057,400</td>
</tr>
<tr>
<td>Random</td>
<td>100% Write</td>
<td>822,935</td>
<td>205,734</td>
<td>5,609,700</td>
</tr>
<tr>
<td>Random</td>
<td>80% Read, 20% Write</td>
<td>868,848</td>
<td>217,212</td>
<td>5,592,925</td>
</tr>
</tbody>
</table>
- **Target node**
  - Dual-socket x86 server
  - 4x40GbE NICs
  - iSER LIO target
  - 20xPM953 NVMe drives
- **Initiators**
  - Dual-socket x86 server
  - 1x40GbE NIC
- **Performance**
  - 2.1M – 4K Random Read
  - 17.2GB/s – 128K Seq Read
Micron

- 2.7MIOPS @ 512B
- 1.8MIOPS @ 4K
- 1.2MIOPS @ 8K
- 850KIOPS @ 16K
- 480KIOPS @ 32K
- 15GB/s Max
- "Datasheet" 4K
- 122us read, 43us write
- 1MIOPS @ 8K 50/50
- 471us average
- 2 dual port 40GbE NICs
Conclusions

- iSER gives users the full performance benefits of flash based solutions across a Ethernet or InfiniBand network
- RDMA technology like RoCE enables iSER performance by bypassing the TCP/IP network stack
- A growing number of storage solutions providers support iSER in their Flash based products
Thank you!

Rob Davis
rob@mellanox.com