100G iSCSI - A Bright Future for Ethernet Storage

Tom Reu
Consulting Application Engineer
Chelsio Communications
Presentation Outline

- Company Overview
- iSCSI Overview
- iSCSI and iSER
- Innovations
- Summary
iSCSI Timeline

- RFC 3720 in 2004
  - Latest RFC 7143 in April 2014
- Designed for Ethernet-based Storage Area Networks
  - Data protection
  - Performance
  - Latency
  - Flow control
- Leading Ethernet-based SAN technology
  - In-boxed Initiators
  - Plug-and-play
  - Closely tracks Ethernet speeds
    - Increasingly high bandwidth

- 10 GbE, IEEE 802ae 2002
  - First 10 Gbps hardware iSCSI in 2004 (Chelsio)
- 40/100 GbE, IEEE 802.3ba 2010
  - First 40Gbps hardware iSCSI in 2014 (Chelsio)
  - First 100Gbps hardware available in Q3/Q4 2016
iSCSI Trends

- iSCSI Growth
  - FC in secular decline
  - FCoE struggles with limitations
- Ethernet flexibility
  - iSCSI for both front and back end networks
- Convergence
  - Block-level and file-level access in one device using a single Ethernet controller
  - Converged adapters with RDMA over Ethernet and iSCSI consolidate front and back end storage fabrics
- Hardware offloaded
  - 40Gb/s (soon to be 50Gb/s & 100 Gb/s) aligns with migration from spindles to NVRAM
  - Unlocks potential of new low latency, high speed SSDs
- Virtualization
  - Native iSCSI initiator support in all major OS/hypervisors
  - Simplifies storage virtualization
iSCSI Overview

- High performance
  - Zero copy DMA on both ends
  - Hardware TCP/IP offload
  - Hardware iSCSI processing
- Data protection
  - CRC-32 for header
  - CRC-32 for payload
  - No overhead with hardware offload

- Why Use TCP?
  - Reliable Protection Protocol
    - retransmit of load/corrupted packets
    - guaranteed in-order delivery
    - congestion control
    - automatic acknowledgment
iSER Overview

- iSER - iSCSI Extensions for RDMA
- Used to operate iSCSI over RDMA transports such as iWARP/Ethernet or Infiniband
- iSER reach options
  - SCSI over iWARP over TCP/IP
  - SCSI over RoCEv2/IB over UDP/IP
- Requires RDMA NICs (RNICs) on both sides
# Introduction: Speeds and Feeds

<table>
<thead>
<tr>
<th>Bandwidth (Gbps)</th>
<th>Reach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ethernet</strong></td>
<td></td>
</tr>
<tr>
<td>iWARP</td>
<td>1, 2.5, 5, 10, 25, 40, 50, 100</td>
</tr>
<tr>
<td>iSCSI</td>
<td></td>
</tr>
<tr>
<td>RoCEv3</td>
<td></td>
</tr>
<tr>
<td><strong>Infiniband</strong></td>
<td>8, 16, 32, 56, 112</td>
</tr>
</tbody>
</table>
Advanced Data Integrity Protection

- Above and beyond iSCSI CRC-32
- Data Integrity Field (DIF) protects against silent data corruption with 16b CRC
  - Adds 8-bytes of Protection Information (PI) per block
- Data Integrity Extension (DIX) allows this check to be done between application and HBA
- T10-DIF+DIX provide a full end-to-end data integrity check
  - iSCSI CRC-32 handoff possible
- T5 supports hardware offloaded T10-DIF+DIX for iSCSI (and FCoE)

iSCSI Layering - Target

- **Backend**
  - PSCSI
  - Block
  - File
  - Ramdisk
  - Null I/O Ramdisk

- **SCSI Layer**
  - CTL
  - SCST
  - Chelsio SCSI Target
  - LIO

- **Transport Layer**
  - Chelsio FCoE Target
  - Chelsio iSCSI Target
  - LIO iSCSI Target

- **Lower Layer Driver**
  - RDMA Verbs
    - RDMA CM
    - iWARP CM
    - RDMA Driver

- **T5 Network Controller**
  - PDU FCoE Offload
  - PDU iSCSI Offload
  - RDMA Offload

- **Host TCP/IP Stack**
  - LIO iSCSI Acceleration

- **NIC**
  - TCP Offload
Quick succession of Ethernet speeds requires no SW API modifications for the networking controller.
iSCSI Performance at 40Gbps

- Storage array with 64 targets connected to 6 initiator machines through 40 Gbps switch
  - Targets are ramdisk null-rw
  - Each initiator connects to 6 targets
- Iometer configuration on initiators
  - Random access pattern
  - 50 outstanding IO per target
  - 8 worker threads, one per target
  - IO size ranges from 512B to 32KB
iSCSI vs iSER scaling

- Chelsio T5 supports iSCSI and iSER concurrently
  - 2x40GE/4x10GE support
  - A storage target using T5 can connect to iSCSI and iSER initiators concurrently
  - The iSCSI hardware can support hardware initiators and software initiators concurrently
  - Full TCP/IP offload
  - Full iSCSI offload or iSCSI PDU offload
iSCSI vs iSER scaling

- Chelsio’s iSCSI and iSER implementations scale equally well
  - iSCSI and iSER share the same hardware pipeline
    - Protocols interleave at packet granularity
    - Same hardware is used to implement DDP for iSCSI and iSER
    - Same hardware is used to segment iSCSI and iSER payload
    - Same hardware is used to insert/check CRC for iSCSI and iSER
    - Same hardware TCP/IP implementation
    - Same end-to-end latency for iSCSI and iSER
- Operation mode is dynamically selected on a per-flow basis
iSCSI vs iSER Performance Comparison

- Use performance numbers for the Chelsio T5 that is a 4x10GE/2x40GE device that supports iSCSI offload, and iSER concurrently
  - 2x40GE performance limited by PCIe 8x Gen3
- In addition supports concurrently FCoE offload, NVMe over iWARP RDMA fabric, and regular NIC operation
Performance iSCSI/iSER Offload

*iSCSI Initiators with T580-CR adapters*

iSCSI/iSER Target running on RHEL 6.5 (3.6.11)
Performance iSCSI 2x40GE offload

2-Port iSCSI Target

I/O Size(Bytes)

Throughput (GB/s)

CPU%

Write_BW  Read_BW  Write_CPU  Read_CPU
Performance 1x40G iSER

![Graph showing Performance 1x40G iSER](image)
iSER/iWARP vs iSER/FDR IB

100Gb - What does it bring to iSCSI?

- Support for 100 GbE iSCSI with LOW CPU Utilization
- 100 GbE will have Excellent Support for NVMe devices
- Chelsio iSCSI processing efficiency will be on-par with processing efficiency already achieved with iWARP
Summary

- iSCSI is a mature protocol with wide industry support
- iSCSI Native initiator in-boxed in all major operating systems/hypervisors
  - Back-end & front-end applicability, virtualization
- Hardware offloaded iSCSI shipping at 40 Gb and soon shipping at 25, 50, 100 Gb
  - High IOPs and throughput
  - Low Latency
  - At 100Gb on both the initiator and target side, we will be able to transmit and receive exactly ONE iSCSI PDU within one TCP segment
- An iSCSI SAN is cheaper and easier to deploy than an iSER SAN
  - iSCSI has a “built-in” second source
    - Software-only solution is CRITICAL for enterprise OEMs
    - iSER has interoperability issues
- For those customers who want it, Chelsio supports ISER (over iWARP) too
More information

www.chelsio.com

www.chelsio.com/whitepapers
for all available White Papers

To contact Sales, sales@chelsio.com
To contact Support, support@chelsio.com
Questions
Thank You!