Linux iSER Performance

iWARP RDMA over 40Gb Ethernet vs. IB-FDR

Executive Summary

The iSCSI Extensions for RDMA (iSER) protocol is a translation layer for operating iSCSI over RDMA transports, such as iWARP/Ethernet or InfiniBand. This paper presents iSER performance results comparing iWARP RDMA over 40Gb Ethernet and FDR InfiniBand (IB). The results demonstrate that iSER over Chelsio's iWARP RDMA adapters achieves consistently superior results in throughput, IOPS and CPU utilization when compared to IB. Unlike IB, iWARP provides the RDMA transport over standard Ethernet gear, with no special configuration needed or additional management costs. Thanks to its hardware offloaded TCP/IP foundation, iWARP provides the high performance, low latency and efficiency benefits of RDMA, with routability to scale to large datacenters, clouds and long distances.

Overview

The Terminator 5 (T5) ASIC from Chelsio Communications, Inc. is a fifth generation, high-performance 2x40Gbps/4x10Gbps server adapter engine with Unified Wire capability, allowing offloaded storage, compute and networking traffic to run simultaneously. T5 also provides a full suite of high performance stateless offload features for both IPv4 and IPv6. In addition, T5 is a fully virtualized NIC engine with separate configuration and traffic management for 128 virtual interfaces, and includes an on-board switch that offloads the hypervisor v-switch.

Remote DMA (RDMA) is a technology that achieves unprecedented levels of efficiency, thanks to direct system or application memory-to-memory communication, without CPU involvement or data copies. With RDMA enabled adapters, all packet and protocol processing required for communication is handled in hardware by the network adapter, for high performance. iWARP RDMA uses a hardware TCP/IP stack that runs in the adapter, completely bypassing the host software stack, thus eliminating any inefficiencies due to software processing. iWARP RDMA provides all the benefits of RDMA, including CPU bypass and zero copy, while operating over standard, simple Ethernet.

Thanks to the integrated, standards based FCoE/iSCSI and RDMA offload, T5 based adapters are high performance drop-in replacements for Fibre Channel storage adapters and InfiniBand RDMA adapters. This paper demonstrates this fact for iSER, by comparing performance over T5 40GbE iWARP and IB-FDR 56Gbps equipment. Thanks to the common API in the shared OFED architecture, no application changes are needed to switch between the two transports.

Test Results

The following graphs compare the unidirectional iSER READ and WRITE throughput, IOPs and CPU usage numbers of the Chelsio iWARP RDMA and Mellanox IB-FDR adapters, varying the I/O sizes using the **fio** tool.

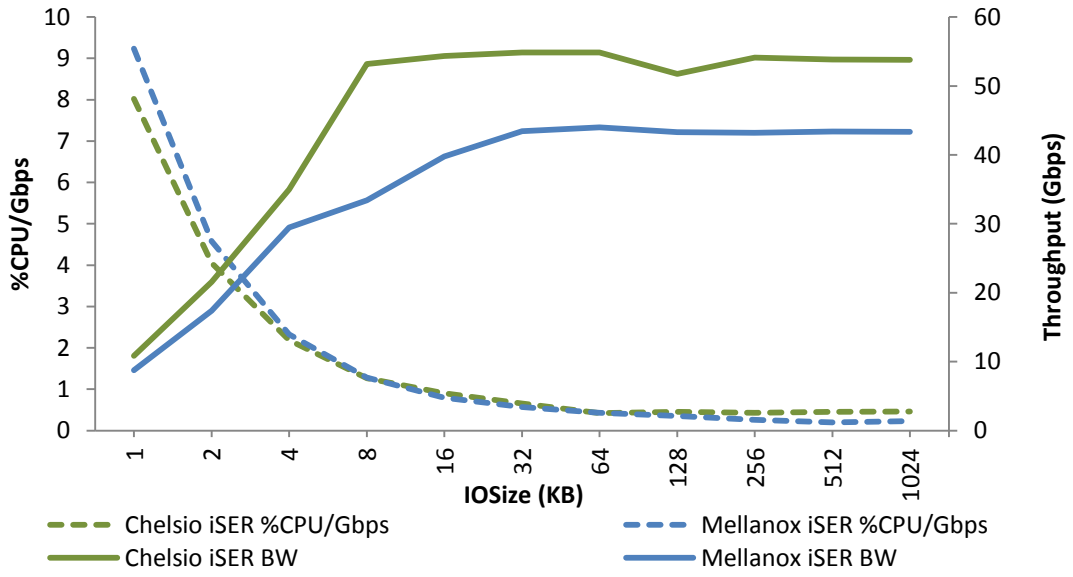

Figure 1 – READ Throughput and %CPU/Gbps vs. I/O size (2 ports)

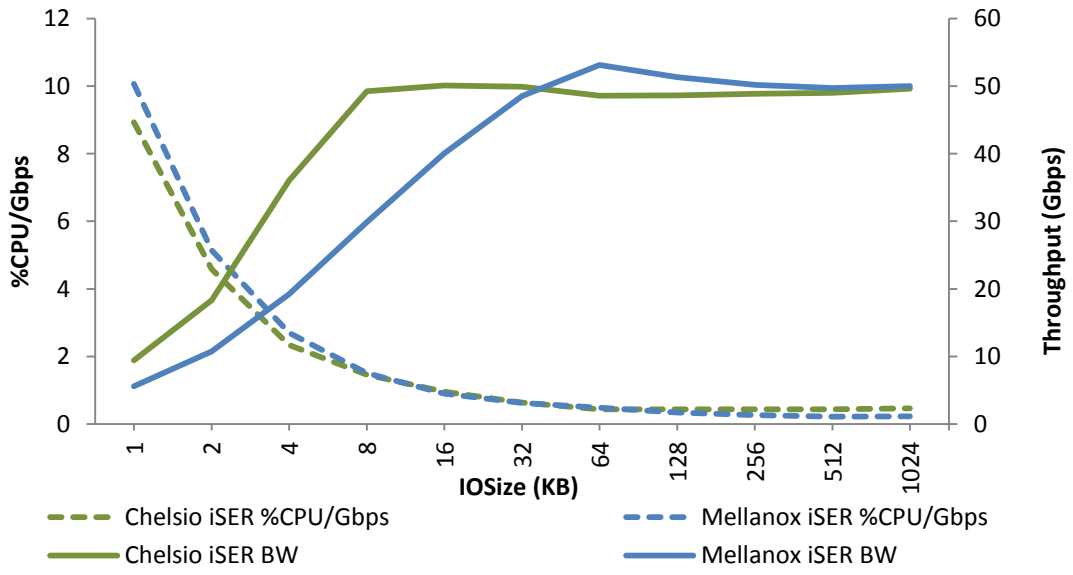

Figure 2 – WRITE Throughput and %CPU/Gbps vs. I/O size (2 ports)

The results above reveal that iSER over IWARP RDMA achieves significantly higher numbers throughout, with outstanding small I/O performance. In addition, the READ results expose a

bottleneck that appears to prevent the IB side from saturating the PCI bus, even at large I/O sizes.

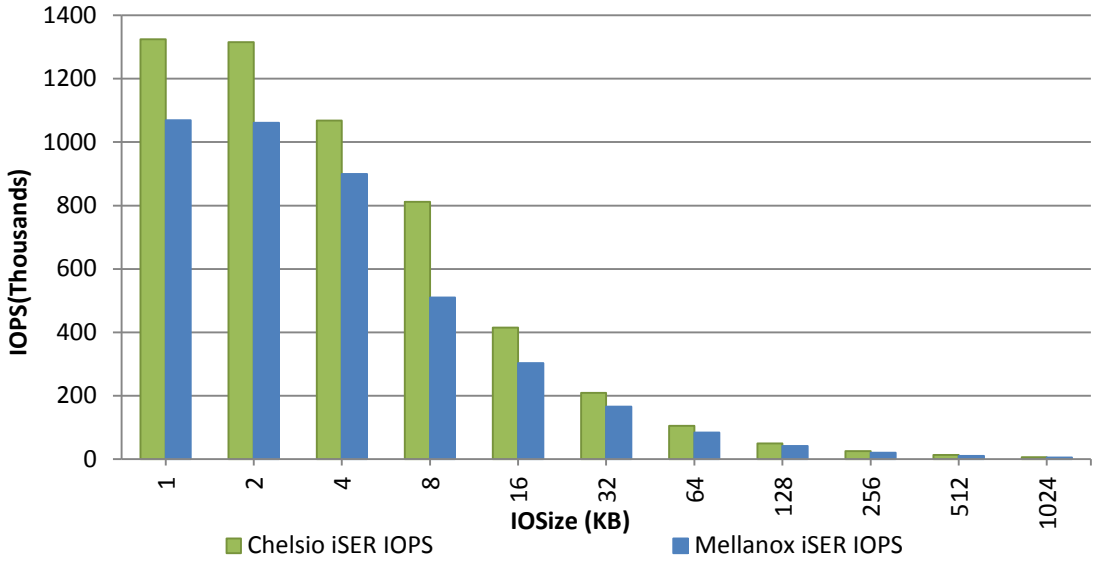

Figure 3 – READ IOPS vs. I/O size

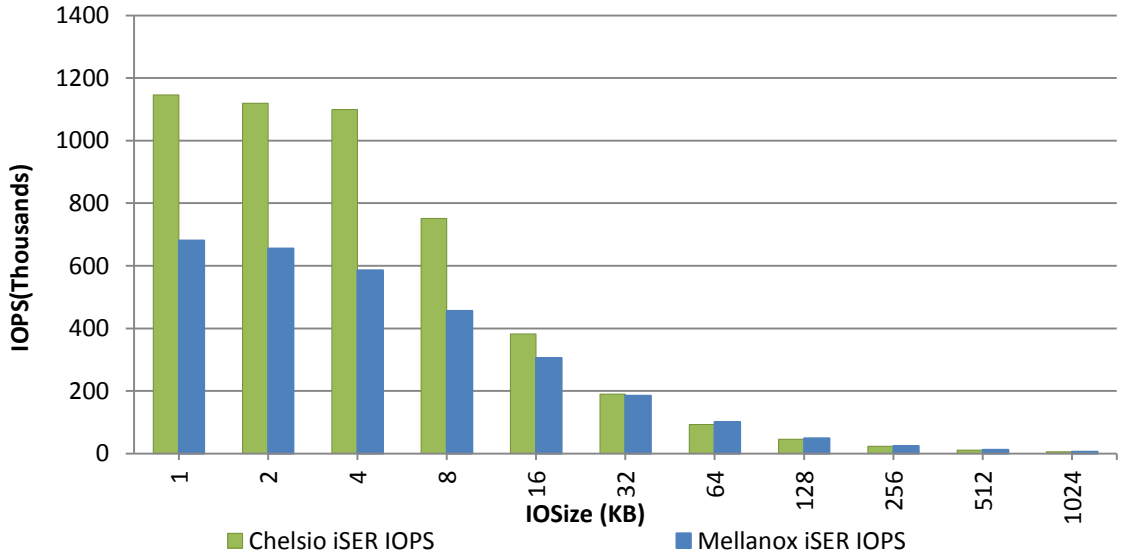

Figure 4 – WRITE IOPS vs. I/O size

Chelsio's iSER over IWARP RDMA provides performance that is superior to InfiniBand throughout the range of study, with the highest difference in the particularly interesting small IO range, up to 16KB. Furthermore, Chelsio's CPU usage per Gbps is mostly lower than Mellanox, establishing the maturity and efficiency of the whole iWARP RDMA solution.

Test Configuration

The following sections provide the test setup and configuration details.

Topology

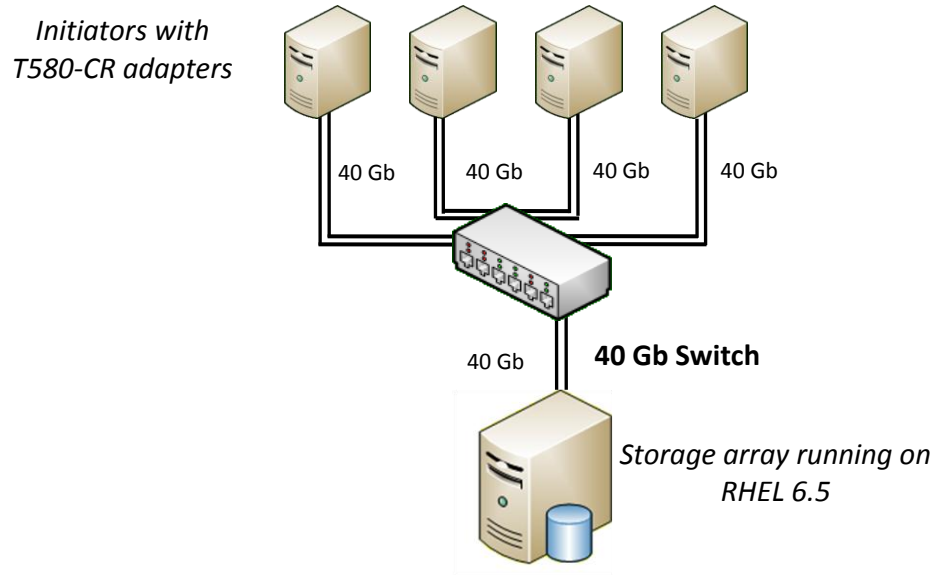

Figure 5- Storage Array connected to 4 Initiators using a 40Gb Switch

Storage Topology and Configuration

The Chelsio iSER over iWARP RDMA setup consists of a target storage array connected to 4 initiator machines through a 40Gb switch using two ports on each system. The InfiniBand setup consists of a target storage array connected to 4 initiator machines through an InifiniBand switch using single port on each system. Standard MTU of 1500B is used.

- **The storage array** is configured with 2 Intel Xeon CPU E5-2687W v2 8-core processors running at 3.10GHz (HT enabled) with 64 GB of RAM.

  In the Chelsio setup, a T580-CR adapter is installed and configured with inbox drivers. The setup uses RHEL 6.5 (3.16.0) operating system with TGT target driver.

  In the InfiniBand setup, a Mellanox MCX353A-FCBT Connect-X adapter is installed with Mellanox OFED driver v2.3-1.0.1. The setup uses RHEL 6.5 operating system with TGT target driver.

- Each of **the initiator machines** is configured with 2 Intel Xeon CPU E5-2687W v2 8-core processors running at 3.10GHz with 64 GB of RAM.

  In the Chelsio setup, a T580-CR adapter is installed with RHEL 6.5 (3.16.0) operating system and inbox open iSCSI Initiator.

In the InfiniBand setup, a Mellanox MCX353A-FCBT Connect-X adapter is installed with open iSCSI Initiator and RHEL6.5 operating system.

In the Chelsio setup, the storage array is configured with 16 targets (8 per port), each configured with 1 ramdisk LUN. Each Initiator port connects to 2 targets.

In the InfiniBand setup, the storage array is configured with 16 targets, each configured with 1 ramdisk LUN. Each Initiator port connects to 4 targets.

I/O Benchmarking Configuration

fio is used to assess the I/O capacity of the configuration. The I/O sizes used varied from 1KB to 1024KB with an I/O access pattern of random READs and WRITEs.

Commands Used

```
[root@host]# fio --name=rand_read --iodepth=1 --rw=randread --size=800m --direct=1 --numjobs=16 instances=4 --bs=<4k-1024KB> --runtime=30 --time_based

[root@host]#fio --name=rand_write --iodepth=1 --rw=randwrite --size=800m --direct=1 --numjobs=16 instances=4 --bs=<4k-1024KB> --runtime=30 --time_based
```

Conclusion

This paper provided iSER performance results comparing Chelsio's T5 iWARP adapter running over 40GbE to IB FDR. The benchmark results show that:

- Chelsio's T580-CR Unified Wire Network adapter T5 delivers superior iSER bandwidth and IOPS performance over the range of study, with significantly higher numbers over the particularly interesting small I/O sizes.
- Chelsio's iSER solution is capable of saturating the PCI bus with READ operations, unlike the IB alternative.
- Chelsio's iSER solution provides consistently higher efficiency in CPU utilization per Gbps of throughput.

These results show that iSER over Chelsio's iWARP RDMA can achieve high performance without the need for a new fabric that is not compatible with the large Ethernet installed base.

Chelsio's T5 iWARP RDMA over Ethernet is shipping at 40Gbps, and is part of a complete high performance Unified Wire alternative to esoteric interconnects, such as InfiniBand, enabling simultaneous operation of RDMA, NIC, TOE, iSCSI and FCoE over Ethernet.

Related Links

The Chelsio Terminator 5 ASIC
Lustre over iWARP RDMA at 40Gbps
NFS/RDMA over 40Gbps Ethernet
SMBDirect 40 GbE iWARP vs 56G Infiniband
iWARP: Ready for Data Center and Cloud Applications