Lustre over iWARP RDMA at 40Gbps

Comparing Performance over Ethernet and IB-FDR

Executive Summary

Lustre is a scalable, secure and highly-available cluster file system that addresses extreme I/O needs, providing low latency and high throughput in large computing clusters. Lustre can use and benefit from RDMA, similarly to other storage protocols. This paper compares the performance of Lustre RDMA over 40Gbps Ethernet and FDR InfiniBand, showing nearly identical performance. Unlike IB, iWARP provides a high performance RDMA transport over standard Ethernet gear, with no special configuration needed or additional management costs. Thanks to its hardware TCP/IP foundation, it provides low latency and all the benefits of RDMA, with routability to scale to large clusters and long distances.

Chelsio's Terminator 5 ASIC with RDMA over Ethernet (iWARP) provides high performance with low latency, while concurrently enabling a full suite of networking and storage protocols, including user space IO with WireDirect, full offload of TCP/IP and UDP/IP, iSCSI and FCoE, all traffic managed and firewalled.

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 for Lustre, 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.

Copyright 2014. Chelsio Communications Inc. All rights reserved.

Test Results

The following graphs compare Lustre READ and WRITE throughput over iWARP and IB-FDR, at different I/O sizes using the **fio** tool.

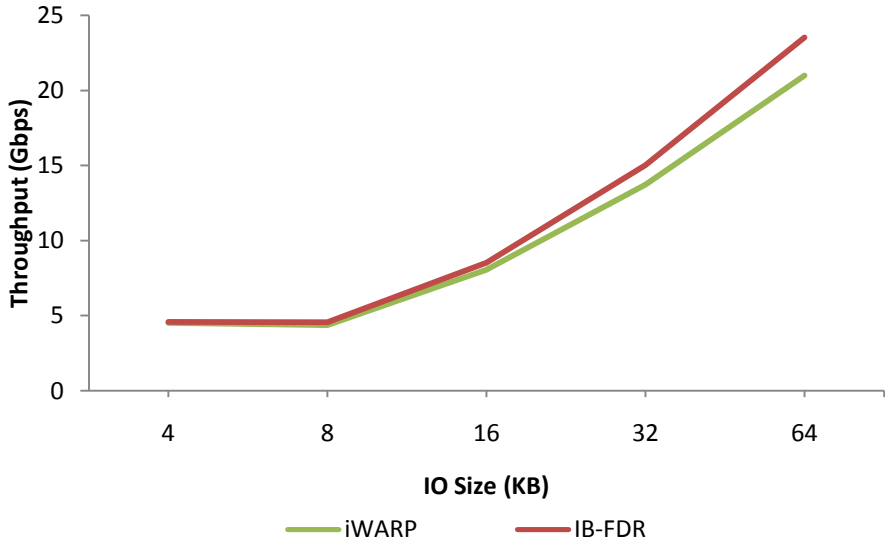

Figure 1 – READ Throughput vs. IO size

The READ throughput numbers show 40 Gbps iWARP delivering nearly identical performance with IB-FDR (56 Gbps) over the range of interest.

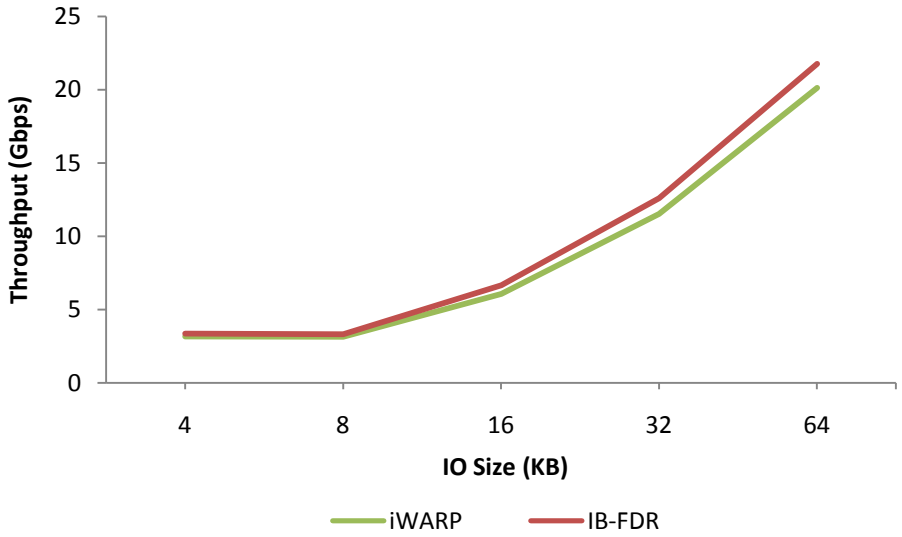

Figure 2 – WRITE Throughput vs. IO size

The WRITE results confirm the equality between the two transports, with nearly the same performance despite the theoretical bandwidth advantage of IB (56G vs. 40G for one port).

Copyright 2014. Chelsio Communications Inc. All rights reserved.

The following graph compares Lustre READ and WRITE throughput over a wider IO range, up to 4MB, as a percentage of wire rate, to focus on the performance curve of the two transports.

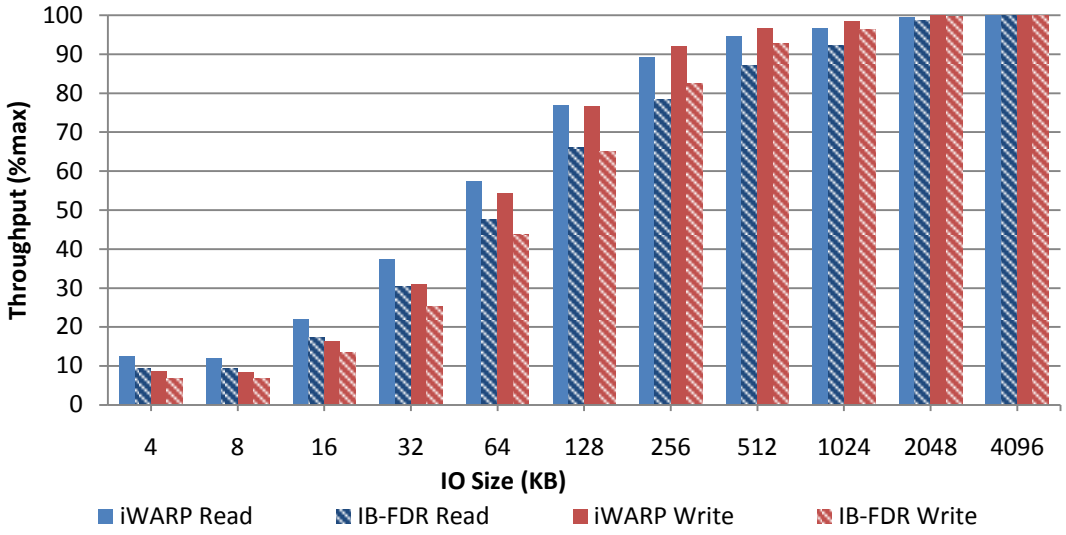

Figure 3 – Throughput as % of max vs. IO size

The results show that Lustre over iWARP RDMA at 40Gbps provides highly competitive performance compared to IB-FDR, and a superior performance curve in relation to the IO size.

This result is just one more in a series of studies of the two fabrics that consistently affirm this conclusion, that iWARP is a no compromise, drop-in Ethernet replacement for IB.

Test Configuration

The following sections provide the test setup and configuration details.

Topology

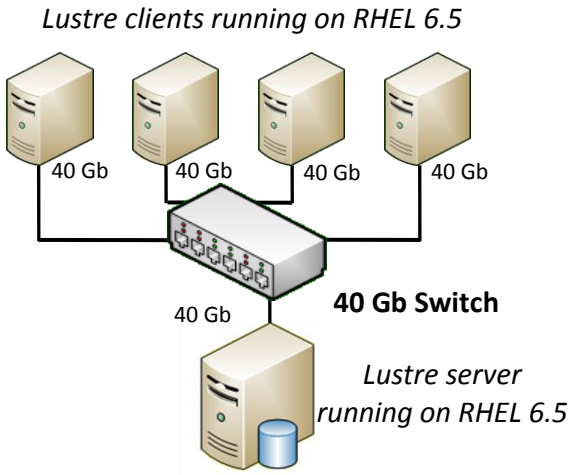

Figure 4 – Lustre Setup

Copyright 2014. Chelsio Communications Inc. All rights reserved.

Network Configuration

The test configuration consists of a Lustre server machine connected to 4 clients through a 40Gb switch using a single port. The standard MTU of 1500B is used.

The **Server** and each of the 4 **Initiator** machines are each configured with 2 Intel Xeon CPU E5-2687W v2 8-core processors running at 3.10GHz and 64 GB of RAM Installed with RHEL6.5 (kernel v2.6.32-431.17.1.el6_lustre-2.5.2) operating system.

In the Chelsio setup, a T580-LP-CR adapter is installed in each system with Unified Wire v2.10.1.0, whereas in the InfiniBand setup, a Mellanox MCX353A-FCBT ConnectX-3 adapter is used.

Storage Topology and Configuration

The storage array exposes 1 *ramdisk*. Each of the 4 clients runs 4 instances of fio tool.

I/O Benchmarking Configuration

fio was used to assess the I/O capacity of the configuration. The I/O sizes used varied from 4KB to 4096KB.

Command Used

```
[root@host]# fio --name=<write/read> --iodepth=1 --rw=<write/read> --size=1000m
--direct=1 --numjobs=1 --bs=<Block Size> --runtime=30 --time_based=30 --
directory=<mounted share>
```

Conclusion

This paper compares the performance of Lustre RDMA over Chelsio's T5 iWARP RDMA adapters and the latest IB-FDR adapters. The performance results show that iWARP at 40GbE is on par with IB-FDR, while utilizing standard Ethernet infrastructure, with no special configuration or management needed. Thanks to the resulting cost and management savings, iWARP is the most cost effective high performance RDMA transport available today.

Chelsio's T5 iWARP RDMA over Ethernet is shipping at 40Gbps, and is part of a high performance Unified Wire alternative to esoteric interconnects such as InfiniBand, enabling simultaneous operation of RDMA, NIC, TOE, iSCSI and FCoE over Ethernet. With superior performance across the board, as recognized in independent studies, Chelsio's adapters remain "great all-in-one adapters".

Related Links

The Chelsio Terminator 5 ASIC
NFS/RDMA over 40Gbps Ethernet
iWARP: From Clusters to Cloud RDMA
40Gb Ethernet: A Competitive Alternative to InfiniBand

Copyright 2014. Chelsio Communications Inc. All rights reserved.