GPUDirect over 40GbE iWARP RDMA

High Performance CUDA Clustering with Chelsio's T5 ASIC

Executive Summary

NVIDIA's GPUDirect technology enables direct access to a Graphics Processing Unit (GPU) over the PCI bus, shortcutting the host system and allows for high bandwidth, high message rate and low latency communication. When married to iWARP RDMA technology, high performance direct access to GPU processing units can be expanded seamlessly to Ethernet and Internet scales.

With iWARP RDMA, network access to the GPU is achieved with both high performance and high efficiency. Since the host CPU and memory are completely bypassed, communication overheads and bottlenecks are eliminated, resulting in minimal impact on host resources, and translating to significantly higher overall cluster performance.

This paper provides an early view of benchmark results that illustrate the benefits of GPUDirect RDMA using Chelsio's T580-CR Unified Wire adapter running at 40Gbps. The results show large improvements in throughput and latency when GPUDirect RDMA is in use, compared to standard server NIC. Furthermore, T5 iWARP is shown to provide 30% higher throughput in comparison to InfiniBand over Ethernet (RoCE).

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, enabling 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. 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.

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 Ethernet networks.

This paper compares performance results of GPUDirect over RDMA with T5 40G iWARP adapter, and provides a comparison to Mellanox's CX-3 Pro 40G RoCE adapter.

Copyright 2015. Chelsio Communications Inc. All rights reserved.

Test Results

The following graphs compare throughput and latency results over iWARP RDMA, with GPUDirect enabled and disabled at different I/O sizes using the **openmpi** tool. The graphs focus on the default I/O size range of operation for GPUDirect.

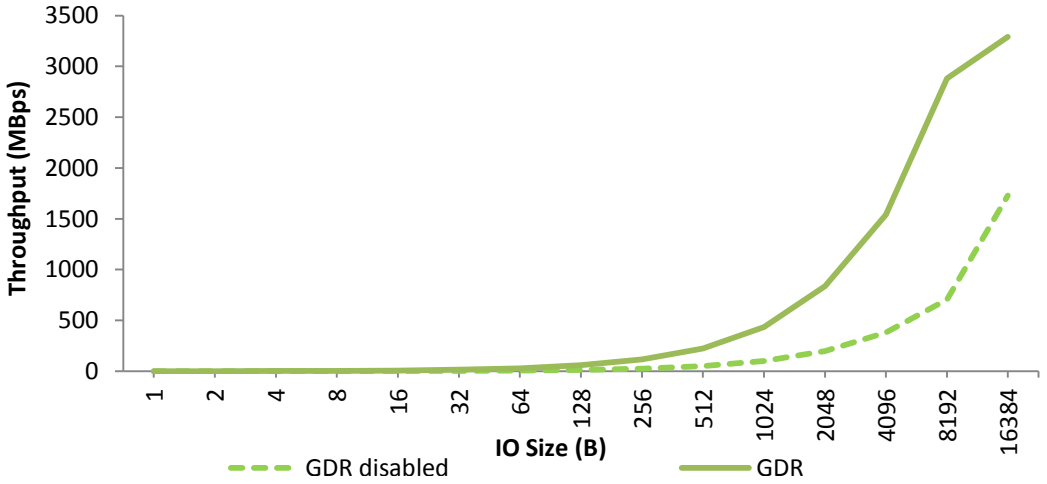

Figure 1 – GDR enabled/disabled Throughput vs. I/O size

The above results clearly show up to 4x the throughput with GPUDirect RDMA enabled than disabled.

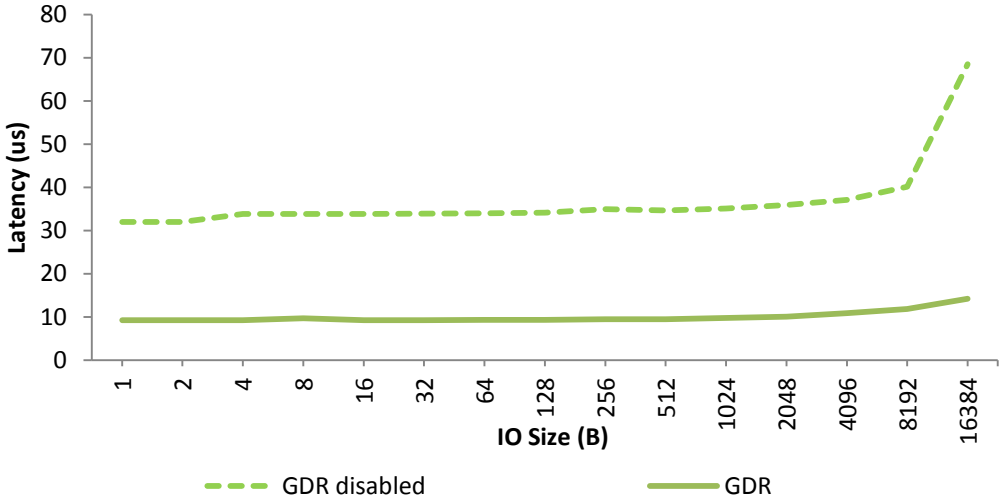

Figure 2 – GDR enabled/disabled Latency vs. I/O size

Using GPUDirect RDMA results in a drastic reduction in latency for the openmpi test application, from more than 50usec to below 10usec over most of the range of interest. Benefits of such magnitude are rare and are a testament to the power of RDMA in this context.

Copyright 2015. Chelsio Communications Inc. All rights reserved.

The following graph compares openmpi throughput results over iWARP and RoCE, with GPUDirect enabled and disabled, at different I/O sizes.

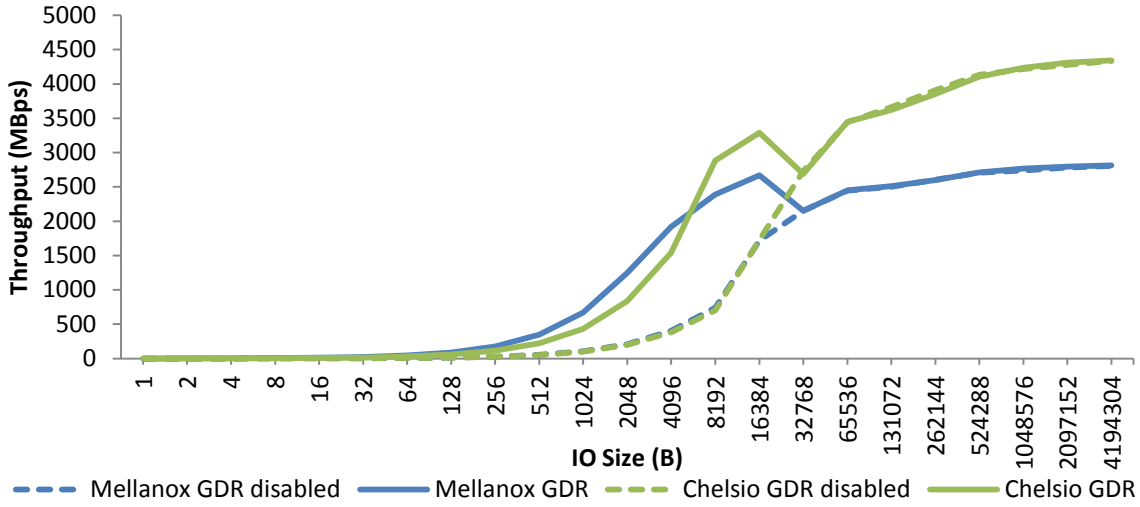

Figure 3 – GDR enabled/disabled Throughput vs. I/O size

The above results clearly show that Chelsio T5 iWARP achieves line rate throughput whereas RoCE fails to reach line rate and plateaus at 40% lower than the capacity, with the difference persisting into the large I/O range, where GDR is disabled.

Test Configuration

The following sections provide the test setup and configuration details.

Topology

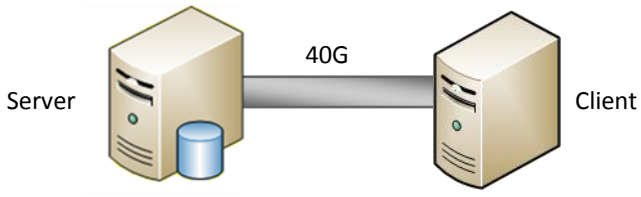

Figure 4 –Test Setup

Network Configuration

The test configuration consists of 2 machines connected back-to-back using single port: a Server and Client, each with 2 Intel Xeon CPU E5-2687W v3 10-core processors clocked at 3.10GHz, with 128GB of RAM and RHEL6.5 operating system. Standard MTU of 1500B is configured. One T580-CR, MT27520 ConnectX-3 Pro, Tesla K20X GPU adapter is installed in each system with the latest Chelsio GPUDirect RDMA driver, CUDA v6.5, OpenMPI v1.8.4 (with CUDA support), OSU micro benchmarking tools v4.4.1 and nVIDIA peer memory driver.

Copyright 2015. Chelsio Communications Inc. All rights reserved.

I/O Benchmarking Configuration

openmpi was used to assess the I/O capacity of the configuration. The I/O sizes used varied from 1B to 4MB for throughput tests and 1B to 8KB for latency tests.

Command Used

```
[root@host]# /opt/ompi-1.8.4-gdr/bin/mpirun --allow-run-as-root -mca
btl_openib_want_cuda_gdr <0/1> -np 2 -host <host1,host2> -npernode 1 -mca btl
openib,sm,self -mca btl_openib_if_include cxgb4_0:1 -mca
btl_openib_receive_queues P,131072,64 -x CUDA_VISIBILE_DEVICES=0 /root/osu-
micro-benchmarks-4.4.1/mpi/pt2pt/<osu_bw/osu_latency> -d cuda D D
```

Conclusion

This paper highlighted the benefits of using Chelsio's T580-CR iWARP RDMA adapter along with NVIDIA's GPUDirect technology in delivering dramatically lower latency and higher throughput for mission-critical scientific and HPC applications. The results also show iWARP RDMA to provide significantly higher bandwidth than competing RDMA over Ethernet protocols.

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 2015. Chelsio Communications Inc. All rights reserved.