High Performance NVMe over 40GbE iWARP

Peer to peer DMA through PCIe bus – No Host Memory Required

Executive Summary

Chelsio first demonstrated NMVe over fabric proof point during IDF 2014, demonstrating end-to-end remote NMVe fabrics model, delivering less than 8µs additional RTT latency along with remote NMVe SSD IOPS same as local NMVe SSD IOPS. Recently NVM Express Organization released version 1.0 of the NVM Express over Fabrics (NVMe-oF) standards and Chelsio supports these standards from day zero with an optimized software stack to deliver an all in-boxed, high performance NVMe-oF solution.

This paper provides an overview of Chelsio NVMe over 40GbE iWARP fabric solution, demonstrating high IOPS and superior throughput numbers in a two machine, back-to-back, target initiator setup environment. This paper also demonstrates an efficient approach to remote storage access made possible by iWARP, which enables the next generation, scalable storage network over standard and cost effective Ethernet infrastructure.

The Chelsio NMVe over RDMA/iWARP Fabric Solution

NVMe over Fabrics specification extends the benefits of NVMe to large fabrics, beyond the reach and scalability of PCIe. NVMe enables deployments with hundreds or thousands of SSDs using a network interconnect, such as RDMA over Ethernet. Thanks to an optimized protocol stack, an end-to-end NVMe solution is expected to reduce access latency and improve performance, particularly when paired with a low latency, high efficiency transport such as RDMA. This allows applications to achieve fast storage response times, irrespective of whether the SSDs are attached locally or accessed remotely across enterprise or datacenter networks.

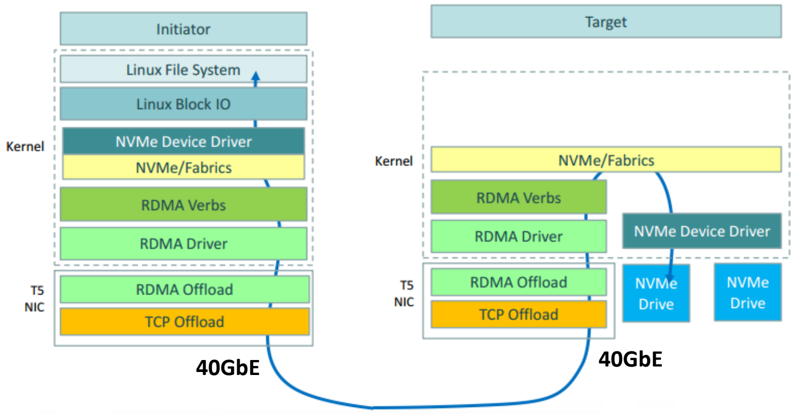

Figure 1 – NVMe over Fabrics Layering

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.

The Terminator 5 (T5) from Chelsio Communications, Inc. is a fifth generation, high performance 4x10/2x40Gbps Ethernet unified wire engine which offers iWARP RDMA, a plug-and-play Ethernet solution for connecting high performance SSDs over a scalable and congestion controlled and traffic managed fabric, with no special configuration needed.

Test Results

The following graphs present READ, WRITE IOPS, Throughput and Latency of Chelsio NVMe over 40GbE iWARP adapters using the **fio** tool. The I/O sizes used varies from 512B to 64KB with an I/O access pattern of random READs and WRITEs. Please note that WRITE numbers are limited by number of SSDs in the test topology.

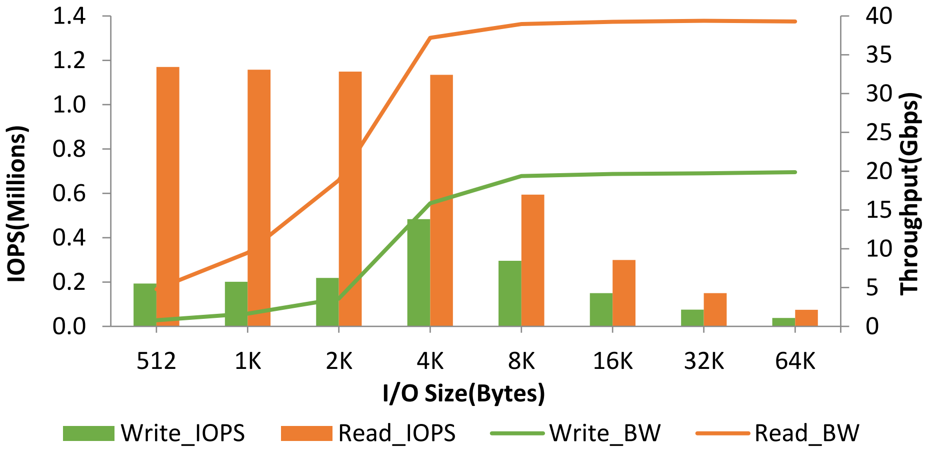

Figure 2 - IOPS & Throughput vs. I/O size

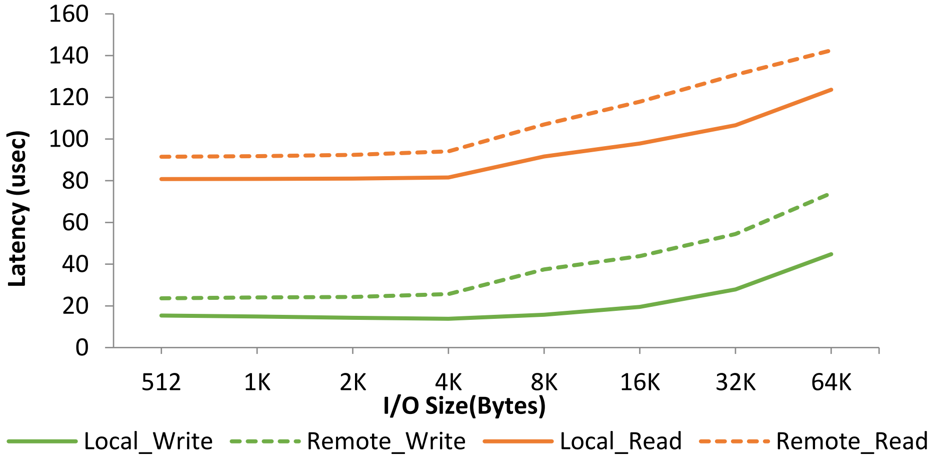

Figure 3 – Latency vs. I/O size (Local vs. Remote)

Test Configuration

The following sections provide the test setup and configuration details.

Topology

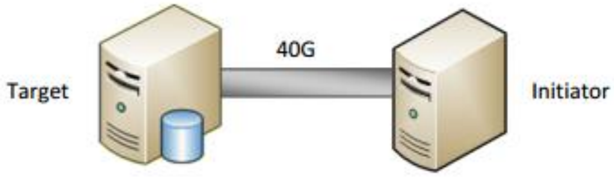

Figure 3 – Test Setup

Network Configuration

The test configuration consists of two machines, **target array** and **initiator**, connected back-to-back using a single 40GbE link, each with 2 Intel Xeon CPU E5-2687W v3 processors clocked at 3.10GHz (HT disabled), 128 GB of RAM and CentOS 7.1 (kernel 4.8.0-rc1) OS. 1 Chelsio T580-CR adapter is installed in each system with latest Chelsio NVMe-oF driver. MTU of 4096 is used.

Storage Topology and Configuration

The initiator connects to the target having 2 NVMe SSD block devices, each of 1.5TB size. Both block devices are used for throughput test, whereas only 1 block device is used for latency test.

Commands Used

WRITE:

```
# fio --rw=randwrite --ioengine=libaio --name=random --size=10000m --direct=1 --
invalidate=1 --fsync_on_close=1 --norandommap --numa_cpu_nodes=0 --
group_reporting --exitall --runtime=60 --time_based --
filename=/dev/nvme0n1:/dev/nvme1n1 --iodepth=32 --numjobs=20 --bs=<value>
```

READ:

```
# fio --rw=randread --ioengine=libaio --name=random --size=10000m --direct=1 --
invalidate=1 --fsync_on_close=1 --norandommap --numa_cpu_nodes=0 --
group_reporting --exitall --runtime=60 --time_based --
filename=/dev/nvme0n1:/dev/nvme1n1 --iodepth=32 --numjobs=20 --bs=<value>
```

Conclusion

This paper showcases the significant performance benefits of Chelsio T5 iWARP RDMA solution for the NVMe specification. The results show that Chelsio's NMVe over iWARP RDMA:

- Reaches approximately 1.2 Million READ IOPS.
- READ throughput reaches line rate from 4K I/O size but WRITE throughput is limited by number of SSDs in the test.
- Adds less than 13 µs latency (tuning continues) for remote versus local NVMe device access.

Related Links

NVMe over 40GbE iWARP RDMA

NVM Express over Fabrics