Introducing NVMe over 100GbE iWARP Fabrics

Chelsio T6 Based High Performance and Low Latency NVMe-oF Solution

Executive Summary

This paper provides an overview of Chelsio NVMe (Non-Volatile Memory Express) over 100GbE iWARP fabric solution, demonstrating superior throughput numbers in a 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 NVMe over iWARP RDMA Fabrics 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 NMVe SSDs are attached locally or accessed remotely across enterprise or datacenter networks.

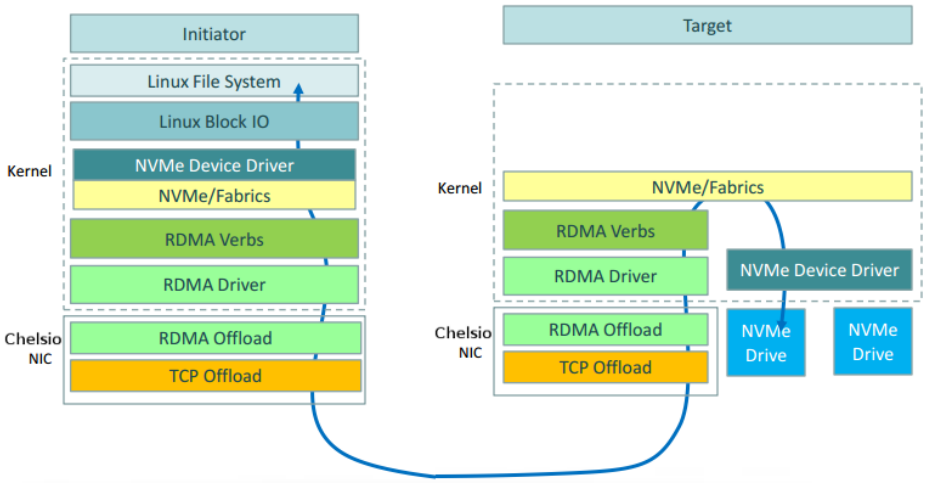

Figure 1 - NVMe over Fabrics with iWARP

The Terminator 6 (T6) ASIC from Chelsio Communications, Inc., a sixth generation, high performance 10/25/40/50/100Gbps unified wire engine offers iWARP RDMA: a plug-and-play Ethernet solution for connecting high performance NMVe SSDs over a scalable, congestion controlled and traffic managed fabric, with no special configuration needed.

T6 enables a unified wire for LAN, SAN and cluster applications, built upon a high bandwidth and low latency architecture along with a complete set of established Networked Block and File Storage and cluster protocols operating over Ethernet (iSCSI, iSER, SMB3.x, iWARP, NVMe-oF, FCoE, NFS,

NFSoRDMA). A unified wire means having the ability to utilize all offload or non-offload protocols at the same time, over the same link, using the exact same firmware, host software and adapter. T6 Ethernet-only networking thus reduces the infrastructure costs in network adapters, cables, switches, rack space, power, equipment spares, management tools, planning, networking staff and installation.

Test Configuration

The following sections provide the test setup and configuration details.

Topology

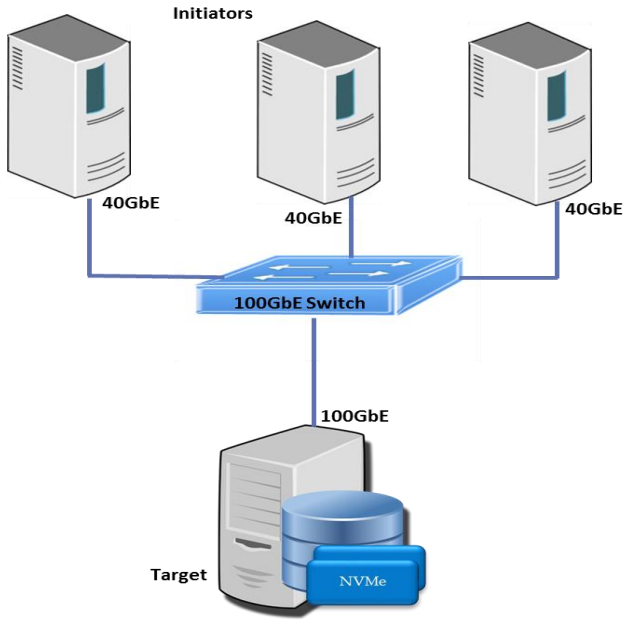

Figure 2- Test Setup

Network Configuration

The setup consists of a target machine connected to 3 initiator machines through a 100GbE switch using single port on each system. The target machine is connected to the switch using a 100GbE link, whereas the link between the initiator machines and switch is 40GbE. MTU of 9000B is used.

- The **target machine** is configured with 2 Intel Xeon CPU E5-2687W v4 12-core processors @ 3.00GHz (HT enabled) and 128GB of RAM. Chelsio T62100-CR adapter is installed in the system with RHEL 7.2 operating system.

- The **initiator machines** are each setup with 1 Intel Xeon CPU E5-1620 v4 4-core processor (HT enabled) @ 3.50GHz and 16GB of RAM. Chelsio T580-LP-CR adapter is installed in each system with RHEL 7.2 operating system.

Storage Configuration

The initiator connects to the target having 3 ramdisk block devices each of 10GB size. Each Initiator uses 4 connections.

Commands Used

WRITE:

```
[root@host~]# fio --rw=randwrite --ioengine=libaio --name=random --size=10000m --
direct=1 --invalidate=1 --fsync_on_close=1 --norandommap --group_reporting --
exitall --runtime=60 --time_based --filename=/dev/nvme0n1 --iodepth=32 --
numjobs=2 --bs=<value>
```

READ:

```
[root@host~]# fio --rw=randread --ioengine=libaio --name=random --size=10000m --
direct=1 --invalidate=1 --fsync_on_close=1 --norandommap --group_reporting --
exitall --runtime=60 --time_based --filename=/dev/nvme0n1 --iodepth=32 --
numjobs=2 --bs=<value>
```

Test Results

The following graph presents READ and WRITE Throughput and CPU percentage performance of Chelsio NVMe over 100GbE iWARP adapters using the **fio** tool. The I/O sizes used varied from 8K to 512K with an I/O access pattern of random READs and WRITEs.

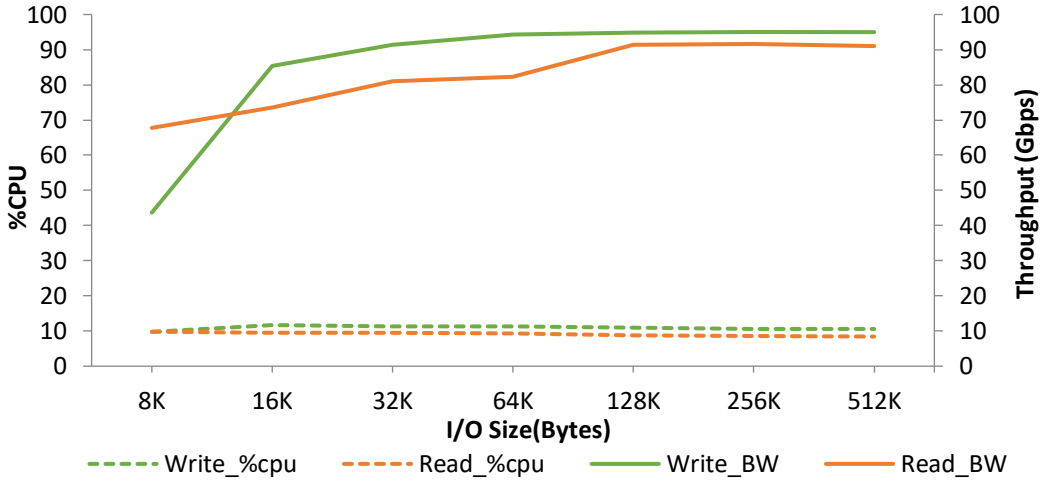

Figure 3 - Throughput & %CPU vs. I/O size

Conclusion

This paper showcases the significant performance benefits of Chelsio T6 based 100G iWARP RDMA solution for the NVMe specification. As evident from the graph, with minimal load on the processing power (<10%), Chelsio's 100Gbe NMVe over iWARP RDMA solution provides line rate throughput across the board. Hence, the T6 10/25/40/50/100GbE Unified Wire adapters:

- Provide concurrent support of NVMe over RDMA Fabrics, Established Networked Block and File Storage delivering high BW, high IOPS and low latency NVMe-oF solution.
- Preserve investment while adopting NVMe-oF.

Related Links

[The Chelsio Terminator 6 ASIC]
[T6 100G NVMe-oF demonstration]
[High Performance NVMe over 40GbE iWARP]
[Concurrent Support of NVMe-oF and Established Networked Block and File Storage]