NVMe over 40GbE iWARP RDMA

Throughput and Latency Benchmark Results

Executive Summary

NVM Express (NVMe), developed by a consortium of storage and networking companies, is an optimized interface for accessing PCI Express (PCIe) non-volatile memory (NVM) based storage solutions.

With an optimized stack, a streamlined register interface and command set designed for high performance solid state drives (SSD), NVMe is expected to provide significantly improved latency and throughput compared to SATA based solid state drives, with support for security and end-to-end data protection.

This paper compares the storage throughput and latency results for regular NIC and RDMA, taken with Chelsio's T580-CR Unified Wire adapter, showcasing Terminator 5 ASIC's seamless support for NVMe. The results show that when RDMA is in use, significant performance gains are obtained. Also, the latency results with RDMA allow realizing the performance potential of NVMe devices.

Overview

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.

In an era of Big Data, massive datacenters, pervasive virtualization and focus on "green" operation and efficiency, RDMA use is rapidly gaining ground. Moreover, RDMA support is natively supported in today's major server operating systems. By providing high level, simplified communication abstractions, such integration further lowers the barrier to realizing the benefits of RDMA, and is further contributing to the acceleration in RDMA adoption.

This paper presents superior throughput results and also demonstrates the low remote storage access latency made possible with iWARP. Such latency numbers allow exploiting the full potential of ultra **low latency SSD** drives, and in combination with the efficient high throughput made possible by iWARP, provide the next generation, scalable storage network over standard, **cost effective Ethernet**.

Test Results

The following graphs compare the READ and WRITE throughput and latency numbers using RDMA Block Device (RBD) and Network Block Device (NBD). The numbers were obtained by varying the I/O sizes using the **fio** tool.

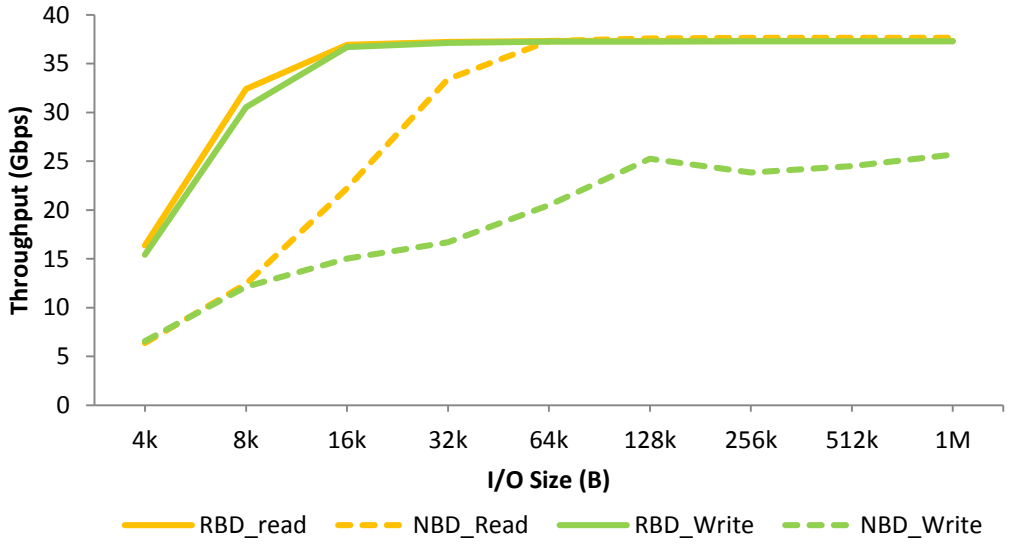

Figure 1 – Throughput vs. I/O size

The READ results reveal that with iWARP RDMA, Chelsio Adapter reaches line rate unidirectional throughput at 1/4th the I/O size needed with NIC. The WRITE results show significantly higher and more consistent performance with iWARP RDMA, reaching line rate throughput at 16 KB I/O size, whereas NIC fails to reach line rate even at higher I/O sizes.

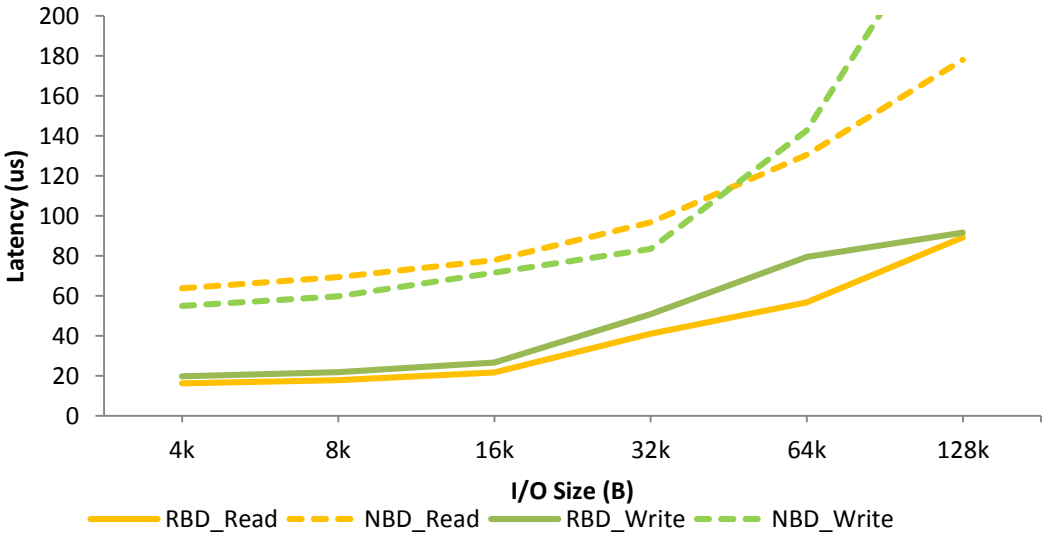

Figure 2 – Latency vs. I/O size

The results show the RTT latency for iWARP RDMA to be 1/3rd of the NIC. Furthermore, the latency for iWARP RDMA increases with a desirably shallower slope than the NIC as the IO size is increased.

Test Configuration

The following sections provide the test setup and configuration details.

Topology

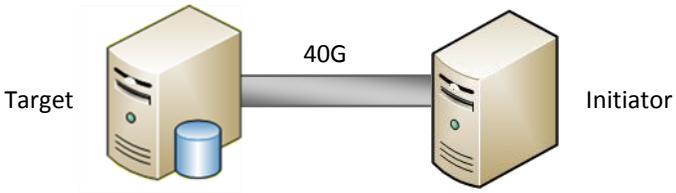

Figure 3 –Test Setup

Network Configuration

The test configuration consists of 2 machines connected back-to-back using a single port: a Target and Initiator, each with 1 Intel Xeon CPU E5-1660 v2 6-core processor clocked at 3.70GHz (HT enabled) and with 64 GB of RAM. Chelsio T580-CR adapter is installed in each system with the latest Chelsio RDMA Block Device Driver and RHEL 6.5 (Kernel 3.18.14) operating system. Standard MTU of 1500B is configured.

I/O Benchmarking Configuration

fio is used to assess the storage capacity of the configuration. The I/O sizes used varied from 4KB to 1MB for throughput test and from 4KB to 128KB for latency test with an I/O access pattern of random READs and WRITEs.

Storage Topology and Configuration

The Initiator connects to the target having 4 *ramdisk* block devices each of 1GB size. All the 4 block devices are used for throughput test, whereas only 1 block device is used for latency test.

Command Used

Throughput:

```
[root@host~]# fio --name=<test_type> --iodepth=32 --rw=<test_type> --size=800m
--direct=1 --invalidate=1  --fsync_on_close=1 --norandommap --group_reporting -
-ioengine=libaio  --numjobs=4 --bs=<block_size> --runtime=30 --time_based --
filename=/dev/rbdi0
```

Latency:

```
[root@host~]# fio --name=<test_type> --iodepth=1 --rw=<test_type> --size=800m -
-direct=1 --invalidate=1  --fsync_on_close=1 --norandommap --group_reporting --
ioengine=libaio  --numjobs=1 --bs=<block_size> --runtime=30 --time_based --
filename=/dev/rbdi0
```

Conclusions

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

- READ throughput reaches line rate at 1/4th the I/O size needed by NIC.
- WRITE throughput is up to 2x of NIC and reaches line rate from 16KB I/O size.
- RTT Latency is drastically lower than NIC.

Chelsio's iWARP RDMA provides a plug-and-play solution for connecting high performance SSDs over a scalable, congestion controlled and traffic managed fabric, with no special configuration needed.

Related Links

The Chelsio Terminator 5 ASIC
iWARP: From Clusters to Cloud RDMA
GPUDirect over 40GbE iWARP RDMA
Lustre over iWARP RDMA at 40Gbps
NVM Express over Fabrics