100G Kernel and User Space NVMe/TCP Using Chelsio Offload

Boosting Software-Defined Storage Performance While Reducing Hardware Costs

Key Take-aways

- Remote, disaggregated, networked NVMe-oF storage with performance comparable to local storage.
- Reduced CPU overhead, resulting in cheaper CPUs, and more cores left over for host storage software.
- Better effective network utilization and server storage I/O performance.
- Respond to dropped or reordered packets at silicon speed via TOE and isolate host and network performance from each other.
- Boost performance productivity while reducing costs and complexity.
- Enable more affordable entry level solutions, or higher performing scalable solutions.
- Free up CPU resources to run your software defined storage and application software.

The NVMe over Fabrics (NVMe-oF) specification extends the benefits of NVMe to large fabrics beyond the reach and scalability of traditional in-server physical PCIe. NVMe/TCP is a technology that facilitates NVMe-oF over existing standard datacenter IP networks. It provides the following advantages over other legacy storage networks and Fabric transports like RDMA (RoCE) and Fibre Channel:

- TCP/IP is a Robust and stable protocol
  - TCP/IP has been an IETF standard (RFC 793, 791) for over 40 years.
  - Well-known, Inherent accuracy, reliability, and scalability.
- TCP/IP is Easy to use
  - Plug-and-Play compatibility.
  - Lower set-up time.
  - No application changes.
- TCP/IP costs less to deploy
  - No additional switches/hardware is required. Compatible with existing data center infrastructure and network tools.
  - Enables a decoupled server and switch upgrade cycle and a brownfield strategy for datacenter deployments.
  - End-user can purchase more compute servers for the same investment amount.
  - Leverage existing TCP network management expertise to reduce costs.
- Available in the latest Linux kernels and Storage Performance Development Kit (SPDK).

This paper presents the significant performance benefits of the Chelsio T6 100GbE NVMe/TCP Offload solution in both Kernel and User modes. Chelsio T6 adapters deliver line-rate throughput and more than 2.9 million IOPs at the 4K I/O size. In addition, with only a 8.85 µs delta latency between remote and local storage access, the Chelsio solution proves to be the best-in-class in providing the next generation, scalable storage network over standard and cost-effective Ethernet infrastructure with an efficient processing path.

Chelsio T6 TCP Offload (TCP/IP Offload Engine) is fully capable of offloading TCP/IP processing of Kernel and User space SPDK NVMe/TCP target I/O to hardware at 100Gbps. Thus, it provides low latency, high throughput Ethernet solution for connecting high-performance NVMe SSDs over a scalable, congestion-controlled, and traffic-managed fabric.

The unique ability of a *TOE* to perform the full transport layer functionality in hardware is essential to obtaining tangible benefits. The vital aspect of the transport layer is process-to-process communication in user space. This means that data passed to the *TOE* comes directly from the application process. The data delivered by the *TOE* goes directly to the application process resulting in less server CPU overhead.

SPDK[1] has been designed to extract maximum server and storage I/O performance by moving all the necessary software drivers to user space. By moving the driver software to user space and changing from kernel interrupts, locks, and I/O path software bottlenecks, application performance is enhanced. The benefits of TOE and SPDK include scalable high-performance with low latency; for user space storage applications like NVMe/TCP target and software-defined storage. NVMe/TCP Offload (Kernel and User Mode) are part of Unified Wire Packages available via the Chelsio website.

Test Results

Chelsio's state-of-the-art *TCP Offload* allows for the entire TCP/IP state to run on the NIC itself, including connection set-up and tear-down, and all the exception handling, thus saving considerable host server resources. The following graph presents IOPs and throughput results of SPDK NVMe/TCP Offload Target with SPDK Kernel NVMe/TCP (regular NIC) hosts using Null Block devices. The results are collected using the **fio** tool with I/O sizes varying from 4 KBytes to 256 KBytes with an access pattern of random READs and WRITEs.

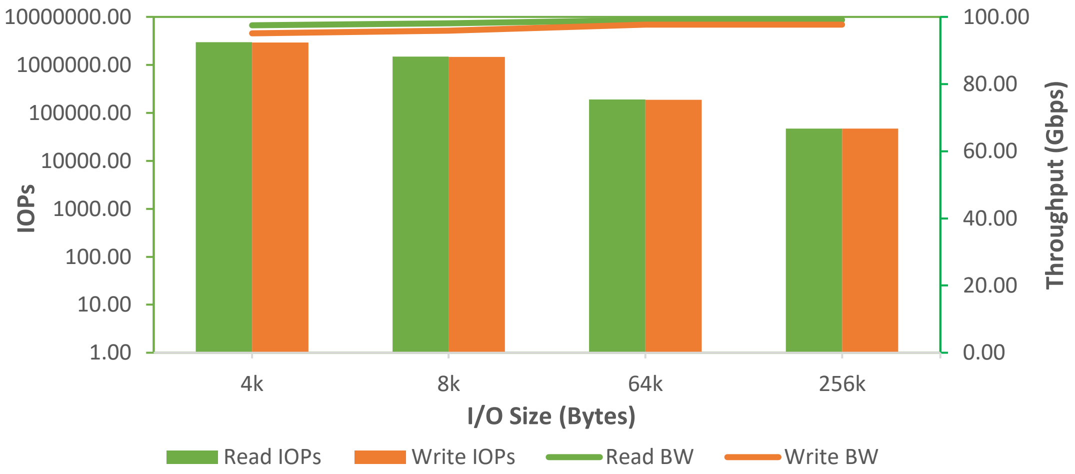

Figure 1 – SPDK NVMe/TCP Offload Target IOPs and Throughput vs. I/O size

[1] The Storage Performance Development Kit (SPDK) is a set of tools and libraries for writing high performance, scalable, user-mode storage applications. More information can be found at www.spdk.io.

The following graph shows IOPs and throughput results of Kernel space NVMe/TCP Offload Target & hosts using Null Block devices. The results are collected using the **fio** tool with I/O size varying from 4 KBytes to 256 KBytes with an access pattern of random READs and WRITEs.

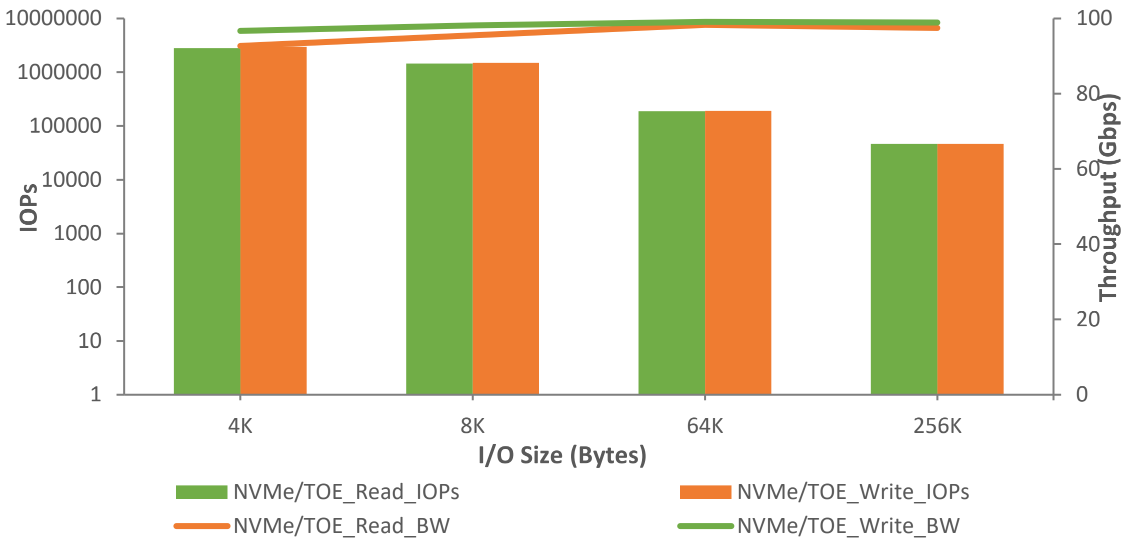

Figure 2 – Kernel NVMe/TCP Offload Target IOPs and Throughput vs. I/O size

The above graphs show how the TCP Offload-enabled T6 solution delivers line-rate READ and WRITE throughput for both Kernel and User space NVMe targets. With the T6 NVMe/TCP Offload, READ and WRITE IOPs reach 2.9 Million at 4K I/O size while using less server CPU.

The following graph compares the CPU consumption per Gbps of Kernel space NVMe/TCP Offload and NVMe/TCP Targets for both READ and WRITE operations.

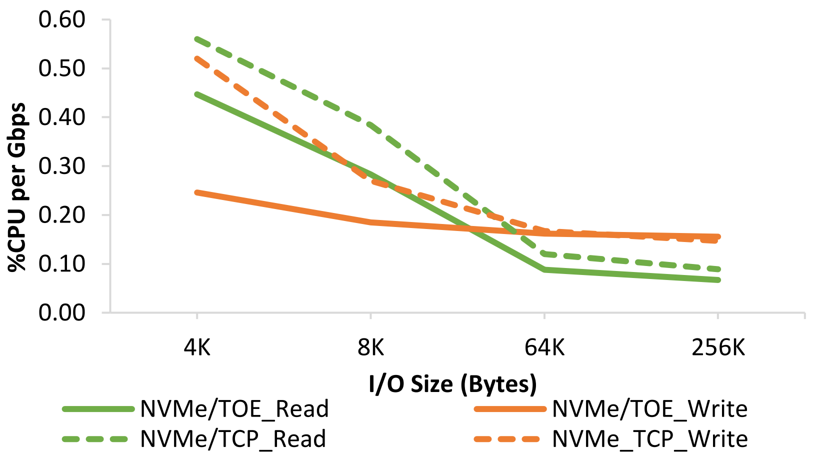

Figure 3 – Kernel NVMe/TCP Offload, NVMe/TCP Target % CPU/Gbps vs. I/O size

Figure 3 shows the NVMe/TCP Offload solution consumes significantly less CPU per Gbps (up to 50%) compared to NVMe/TCP. This is one of the most essential benefits for a hardware offloaded

TCP solution, resulting in a lower cost bill of materials. For example, in the testing for this paper, about two cores per socket can be saved for line-rate performance at 4KB using the TCP running on the card relative to software TCP running on the host CPU.

Since the TCP stack is running on the NIC, the TOE can isolate the host application's performance from the performance spikes caused by network traffic. The TOE eliminates the need for the application to be swapped in to retransmit or reorder a packet as this is done by the NIC. This results in more efficient use of the CPU.

To measure system jitter handling capabilities, we used 1024 iperf TCP Connections to generate traffic, and then we run one instance of netperf TCP_RR test to measure the mean latency and its variance in the NIC only case and the TCP Offload only case. In the NIC to NIC case, we measure 4615 µsec and in the TCP Offload to TCP Offload case we measure 4253 µsec. The real benefit of TCP Offload is shown when we examine the standard deviation of the latency. Though the average latency for TCP offload is only 8% lower, the standard deviation is 60% lower than NIC to NIC. As a result, TCP Offload handles jitter much better and therefore can make traffic move smoother and more predictive.

	Average Latency (µsec)	**Standard Deviation**
NIC <-> NIC	4615	5240
TOE <-> TOE	4253	2160

A NVMe/TCP Offload solution delivers a fully ordered, reliable data stream to the host with less server CPU overhead. Chelsio's TCP Offload solution is required to achieve 100 Gb/s and higher line-rate throughput with minimal CPU usage.

The following table presents the 4K Random I/O latency observed between the NVMe storage local to the server and the latency observed when the NVMe storage is accessed remotely over NVMe-oF, in cases of using transports TCP, TCP offload, and for comparison purposes RDMA (in this case iWARP).

	Read			**Write**		
Target <-> host	**Local**	**Remote**	**Delta**	**Local**	**Remote**	**Delta**
Kernel TCP <-> Kernel TCP	109.15	130.61	21.46	24.43	44.65	20.22
Kernel TOE <-> Kernel TCP	109.15	126.18	17.03	24.43	42.67	18.24
Kernel TOE <-> Kernel TOE	109.15	124.95	15.8	24.43	40.84	16.41
SPDK NIC <-> SPDK NIC*	105.31	126.87	21.57	20.08	39.9	19.1
SPDK TOE <-> SPDK NIC*	105.31	114	8.69	20.08	29.65	8.85
SPDK iWARP <-> SPDK iWARP*	105.31	110.88	5.57	20.08	27.4	6.6

* FIO command run with the SPDK FIO Plugin

When using SPDK with NVMe/TCP offload to a NIC running SPDK alone, the observed Read latency delta is 8.69 microseconds! Very close to that observed for RDMA! This demonstrates the local like performance of remote distributed storage using the Chelsio T6 TCP Offload enabled and SPDK solution.

Conclusion

This paper showcases the server CPU savings and the local-like performance capabilities of remote storage access using the Chelsio T6 100G NVMe/TCP Offload solution. The Chelsio T6 enables the NVMe storage devices to be shared, pooled, and managed more effectively across a low latency, high-performance network, and CPU server savings.

The test result proof points in this paper show that NVMe/TCP Offload:

- Delivers line-rate 99 Gbps throughput for both READ and WRITE.
- Reaches 2.9 Million IOPs at an I/O size of 4K.
- Adds only 8.85 µs latency for remote NVMe device access compared to local access.
- Provides significant CPU savings compared to NVMe/TCP.

TCP Offload improves performance for all TCP applications while freeing up CPU resources for application processing. This means all storage and networking traffic runs over a single 25/100Gb network, rather than building and maintaining multiple networks, resulting in significant acquisition and operational cost savings.

Using a Chelsio Offload-enabled adapter and the Unified Wire Software package available as part of the Chelsio solution, users can create and maintain a true Converged Fabric-based server cluster for software-defined storage and other applications.

Key take-aways:

- ✓ Remote, dis-aggregated, networked NVMe-oF storage with local like performance
- ✓ Reduced CPU overhead: fewer components and cost, or more work done for the same cost
- ✓ Better effective network utilization and server storage I/O performance
- ✓ Boost performance productivity while reducing costs and complexity
- ✓ Enable more affordable entry-level solutions or higher-performing scalable solutions
- ✓ Free up CPU resources to run your software-defined storage and application software

Call to Action

Contact Chelsio to arrange a trial evaluation of our T6 Unified Wire for NVMe/TCP and other Server Storage I/O Network acceleration needs with your applications and software at sales@chelsio.com. Learn more about Chelsio T6 Unified Wire and related technologies, along with technology, product, and business financial benefits by contacting us and visiting www.chelsio.com.

Related Links

The True Cost of Non-Offloaded NICs
We put the iWARP in NVMe-oF
100G SPDK NVMe over Fabrics
NVMe-oF with iWARP and NVMe/TCP