Building Enterprise-Class Storage Using 40GbE

Unified Storage Hardware Solution using T5

Executive Summary

This white paper focuses on providing benchmarking results that highlight the Chelsio T5 performance advantage across the full range of storage networking protocols including iSCSI, iSCSI over RDMA (iSER), Windows Server 2012 SMB 3.0, Lustre over RDMA, NFS over RDMA and FCoE. Benchmarking results show T5 to be an ideal fit for high-performance storage networking solutions across the range of storage protocols. Furthermore, thanks to using the routable and reliable TCP/IP as a foundation, T5 allows highly scalable and cost effective installations using regular Ethernet switches.

Internet SCSI (iSCSI) testing demonstrates that Chelsio T5 adapters provide consistently superior results, lower CPU utilization, higher throughput and drastically lower latency, with outstanding small I/O performance compared to Intel's XL710 "Fortville" server adapter running at 40Gbps.

iSCSI over RDMA testing demonstrates that iSER over T5 iWARP (RDMA/TCP) achieves consistently superior results in throughput, IOPS and CPU utilization in comparison to InfiniBand-FDR. Unlike IB, iWARP provides RDMA transport over standard Ethernet gear, with no special configuration needed or added management costs.

Similarly, a performance comparison of Chelsio's T580-CR RDMA-enabled adapter and Intel's XL710 40GbE server adapter when used with SMB 3.0 file storage protocol in Windows Server 2012 R2 shows T580 exhibiting a better performance profile in both NIC and RDMA operation, with the results clearly proving that when RDMA is in use, significant performance and efficiency gains are obtained for SMB.

A comparison of Lustre RDMA and NFS over RDMA using T5 40GbE iWARP and InfiniBand-FDR shows nearly identical performance. Unlike IB, iWARP provides a high-performance RDMA transport over standard Ethernet gear, with no special configuration needed or added management costs. Thanks to its hardware TCP/IP foundation, it provides low latency and all the benefits of RDMA, with routability to scale to large clusters and long distances. As a result, iWARP can be considered a no-compromise, drop-in Ethernet replacement for IB.

FCoE performance results for T5 show that a T5-enabled FCoE target reaching nearly 4M IOPS and full 40G line-rate throughput at I/O sizes as low as 2KB. As a result, T5 can be considered the industry's first high-performance full FCoE offload solution.

Introduction

At a time of exponential data growth, data center storage systems are being pushed to their performance limits and capacity capabilities. With a predominant level of enterprise data stored in centralized file and block-based storage arrays, networking performance between servers and storage systems has become critical to overall data center performance and efficiency. Growing acceptance of hybrid and all-flash arrays has put even more pressure on this connection, as systems with built-in solid-state drives (SSDs) can pump enough data to saturate traditional network connections many times over.

All of this has merged to make the network between storage systems and servers the main bottleneck in modern data centers. However, 40 Gbps Ethernet (40GbE) hardware-accelerated storage networking protocols, such as iSCSI offload, iWARP (RDMA/TCP) and TCP/IP Offload Engine (TOE) can alleviate this bottleneck, while also allowing the cost-of-ownership advantages of pervasive Ethernet networking. 40GbE storage protocol offload technologies allow drastically lower data transfer latency, significantly higher CPU and overall storage system efficiency. Additionally, both 40GbE storage offload protocols provide the best network bandwidth, IOPS, and latency available today, allowing hybrid and all-flash arrays to deliver their greatest performance impact.

Terminator 5 ASIC

The Terminator 5 (T5) ASIC from Chelsio Communications, Inc. is a fifth generation, high-performance 2x40Gbps unified wire engine which offers storage protocol offload capability for accelerating both block (iSCSI, FCoE) and file (SMB, NFS) level storage traffic. Furthermore, T5 storage protocol support is part of a complete, fully virtualized unified wire offload suite that includes FCoE, RDMA over Ethernet, TCP and UDP sockets and user space I/O.

By leveraging Chelsio's proven TCP Offload Engine (TOE), 40GbE offloaded iSCSI over T5 enjoys a distinct performance advantage over a regular L2 NIC. Unlike FC and FCoE, 40GbE iSCSI runs over regular Ethernet infrastructure, without the need for specialized FC fabrics, or expensive DCB enabled switches and Fibre Channel Forwarder switches. By using IP, it is also routable over the WAN and scalable beyond a local area environment. Finally, the TCP transport allows reliable operation over any link types, including naturally lossy media such as wireless.

T5 also includes enhanced data integrity protection for all protocols, and particularly so for storage traffic, including full end-to-end T10-DIX support for both 40GbE iSCSI and FCoE, as well as internal data path CRC and ECC-protected memory.

Finally, thanks to integrated, standards based FCoE/iSCSI and RDMA offload, T5 40GbE-based adapters are high-performance drop-in replacements for Fibre Channel storage adapters and InfiniBand RDMA adapters. In addition, unlike other converged Ethernet adapters, the Chelsio T5 based 40GbE NICs also excel at normal server adapter functionality, providing high packet processing rate, high throughput and low latency for common network applications.

40GbE iSCSI[1]

This section summarizes iSCSI performance results for Terminator 5 (T5) ASIC running over a 40Gbps Ethernet interface.

Using a T5-enabled iSCSI target, the benchmarking results show 40Gb line-rate performance using standard Ethernet frames with I/O sizes as small as 2KB, and more than 3M IOPS for 512B and 1KB I/O sizes. These results clearly show iSCSI to be an ideal fit for very high-performance storage networking solutions. Furthermore, thanks to using the routable and reliable TCP/IP as a foundation, iSCSI allows highly scalable and cost effective installations using regular Ethernet switches.

The following graph plots the performance results, showing how line-rate 40Gbps can be achieved even at I/O sizes as small as 2KB for unidirectional transfer and 4KB for bidirectional transfer.

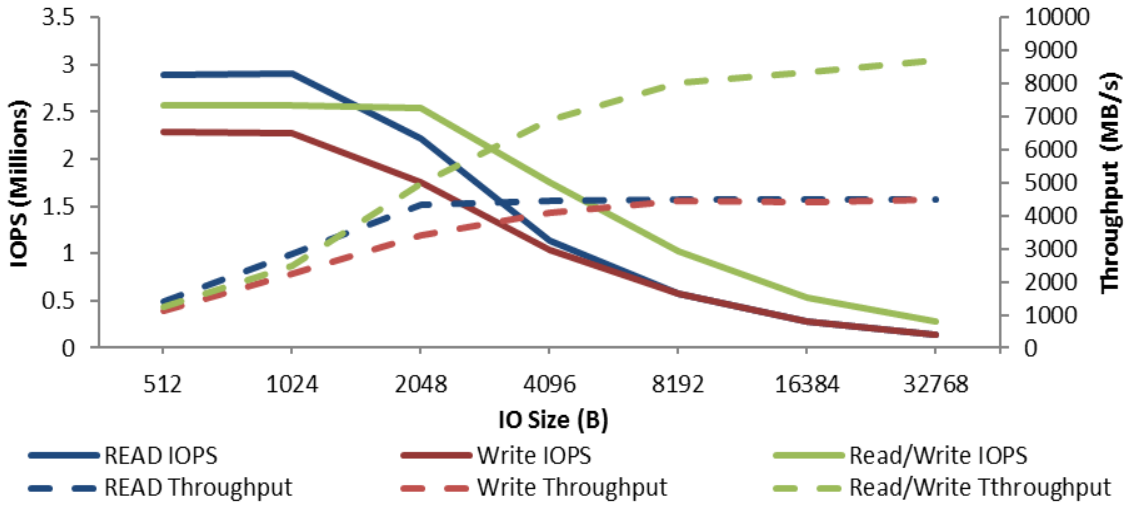

Figure 1 – iSCSI Throughput and IOPS

Chelsio T580-CR vs. Intel Fortville XL710[2]

This section presents iSCSI performance results comparing Chelsio's T580-CR and Intel's latest XL710 "Fortville" server adapter running at 40Gbps. The iSCSI testing demonstrates that Chelsio T580-CR provides consistently superior results, lower CPU utilization, higher throughput and drastically lower latency[2], with outstanding small I/O performance. The T580 notably delivers line-rate 40Gbps at the I/O sizes that are more representative of actual application use.

[1] These results are excerpted from Chelsio white paper iSCSI at 40Gbps, which details benchmark testing results and test hardware/software configuration.

[2] This section is excerpted from Chelsio white paper Linux NIC and iSCSI Performance over 40GbE which details benchmark testing, including throughput, CPU utilization, latency and IOPS performance, and test hardware/software configuration.

The following graphs compare the single port iSCSI READ and WRITE Throughput and CPU Utilization numbers for the two adapters respectively, obtained by varying the I/O sizes using the **iometer** tool. The first graph shows Chelsio's T580-CR performance to be superior in both CPU utilization and throughput, reaching line-rate at ¼ the I/O size needed for Intel's Fortville XL 710. The second graph shows that Chelsio's adapter provides higher efficiency, freeing up significant CPU cycles for actual application use.

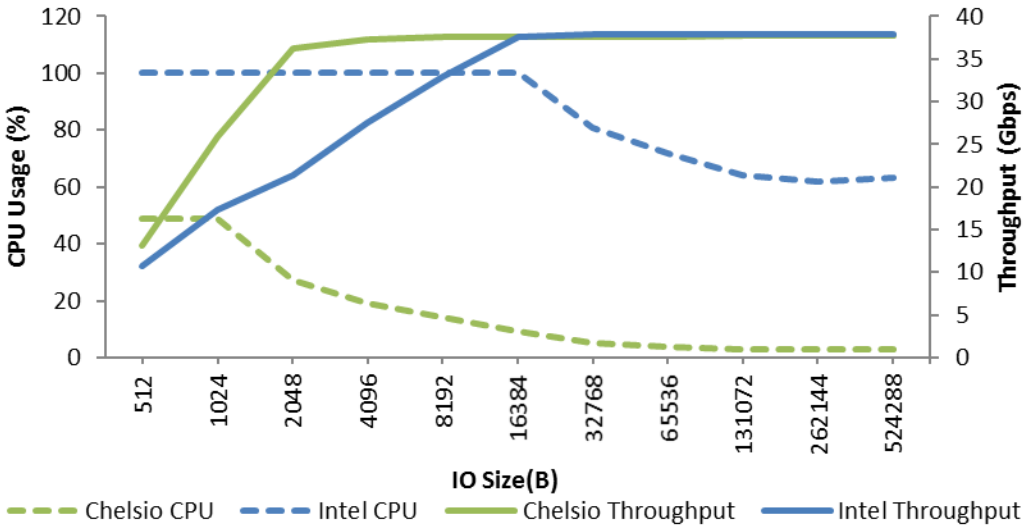

Figure 2 – READ Throughput and %CPU vs. I/O size

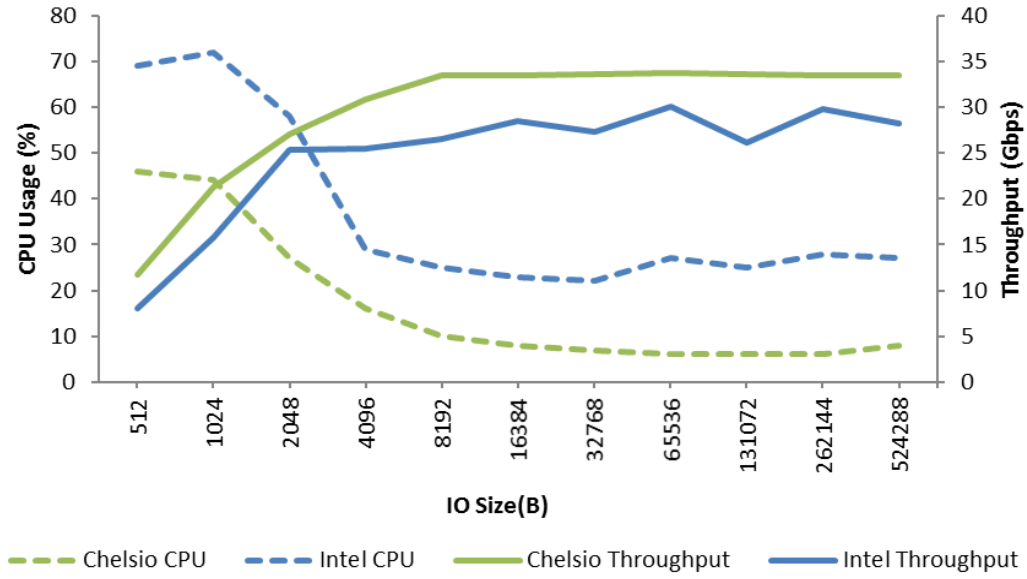

Figure 3 – WRITE Throughput and %CPU vs. I/O size

Linux iSER Performance: 40GbE iWARP vs. InfiniBand FDR[3]

The iSCSI Extensions for RDMA (iSER) protocol is a translation layer for operating iSCSI over RDMA transports, such as iWARP/Ethernet or InfiniBand. This section summarizes iSER performance results, comparing iWARP RDMA over 40Gb Ethernet and FDR InfiniBand (IB). The results demonstrate that iSER over Chelsio's iWARP RDMA adapters achieves consistently superior results in throughput, IOPS[3] and CPU utilization when compared to IB. Unlike IB, iWARP provides the RDMA transport over standard Ethernet gear, with no special configuration needed or added management costs. Thanks to its hardware offloaded TCP/IP foundation, iWARP provides the high-performance, low latency and efficiency benefits of RDMA, with routability to scale to large datacenters, clouds and long distances.

The following graphs compare the unidirectional iSER READ and WRITE throughput and CPU usage numbers of the Chelsio iWARP RDMA and Mellanox IB-FDR adapters, varying I/O sizes using the **fio** tool. The results cited below reveal that iSER over iWARP RDMA achieves significantly higher numbers throughout, with outstanding small I/O performance. In addition, the READ results expose a bottleneck that appears to prevent the IB side from saturating the PCI bus, even at large I/O sizes.

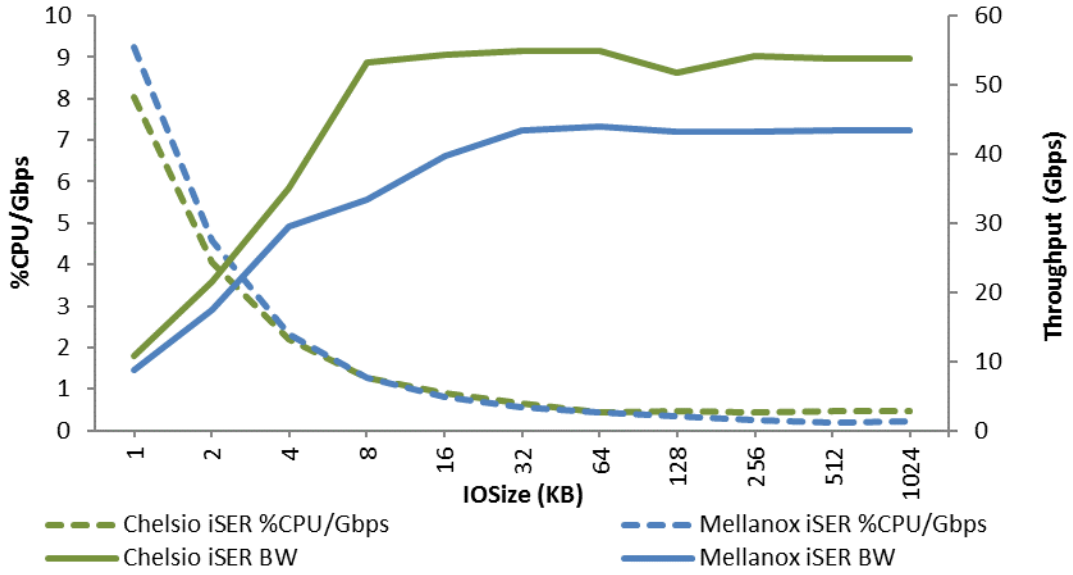

Figure 4 – READ Throughput and %CPU/Gbps vs. I/O size (2 ports)

[3] This section is excerpted from Chelsio white paper Linux iSER Performance which details comparative iSER benchmarking including throughput, IOPS, and CPU utilization dimensions.

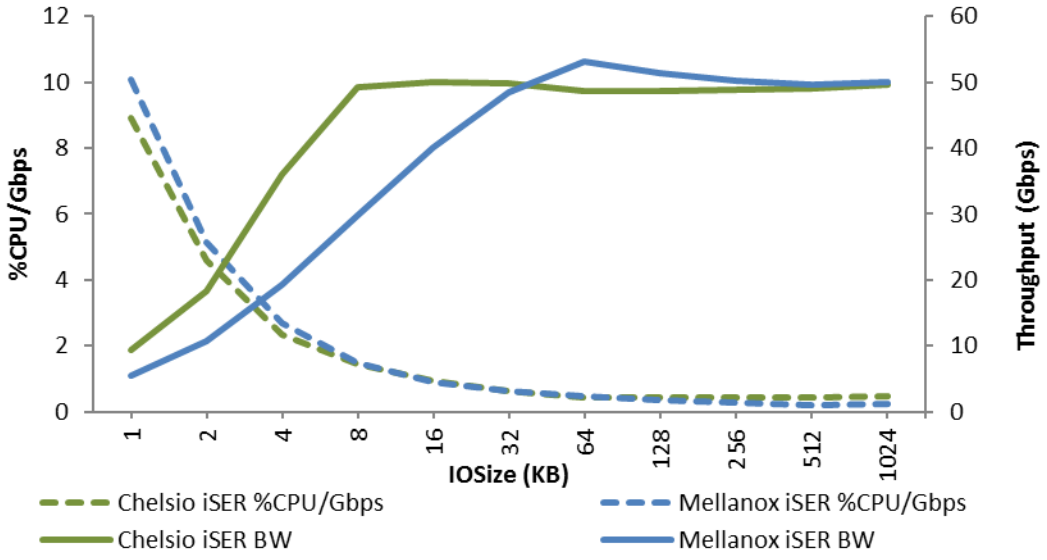

Figure 5 – WRITE Throughput and %CPU/Gbps vs. I/O size (2 ports)

Windows SMB 3.0 Performance at 40Gbps[4]

A notable feature of Microsoft Windows Server 2012 R2 is the release of SMB 3.0 with SMB Direct (SMB over RDMA), which seamlessly leverages RDMA-enabled network adapters for improved storage bandwidth, latency and efficiency. One of the main advantages of the SMB 3.0 implementation is that once the RDMA-capable network adapter driver is installed, all its features are automatically enabled and made available to the SMB application.

This section provides benchmark testing data that demonstrate the benefits of SMB Direct through a performance comparison of Chelsio's T580-CR RDMA-enabled adapter and Intel's XL710 40GbE server adapter. While the T580-CR exhibits a better performance profile in both NIC and RDMA operation, the results clearly show that when RDMA is in use, significant performance and efficiency gains are obtained.

The following graphs compare the throughput and IOPS benchmark results for the two adapters at different I/O sizes. The results reveal that Chelsio's adapter provides significantly higher and more consistent performance, reaching line-rate unidirectional throughput at 8KB I/O size. The difference is particularly clear when RDMA kicks in as the I/O size exceeds 4KB, with up to 2x the performance of the Intel NIC in bidirectional transfers. The Intel adapter also exhibits large variability in the same test environment, which is symptomatic of performance corners.

[4] This section is excerpted from Chelsio white paper Windows SMB 3.0 Performance at 40Gbps.

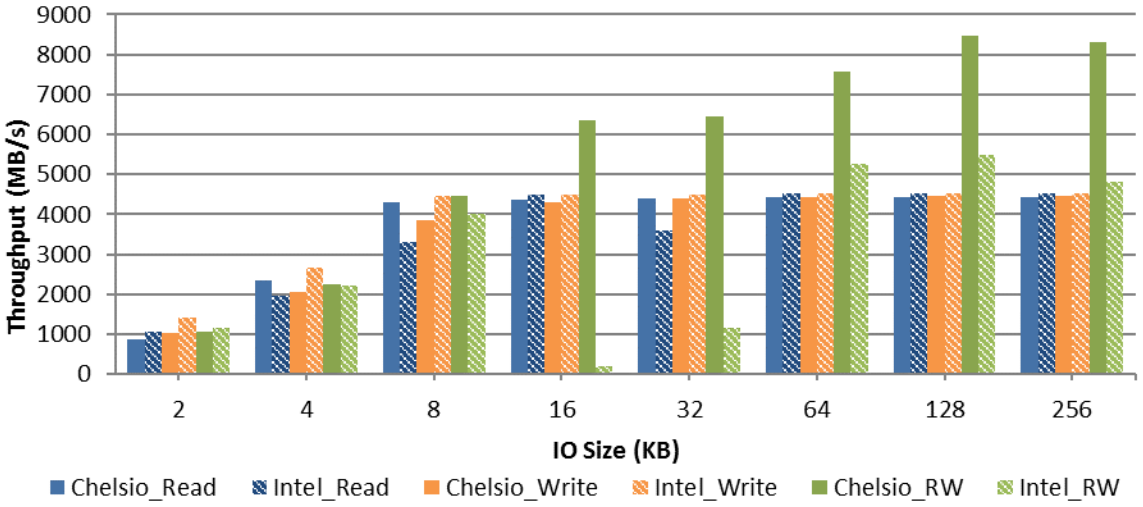

Figure 6 – RDMA and NIC Throughput Comparison for SMB 3.0

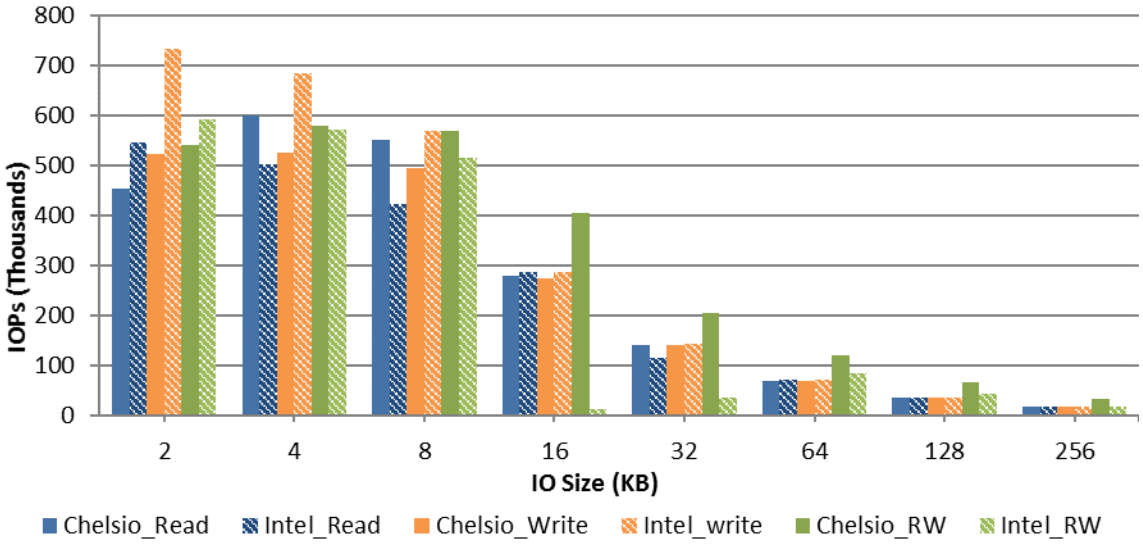

Figure 7 – RDMA and NIC IOPS Comparison for SMB 3.0

The tests results therefore establish that SMB significantly benefits from an RDMA transport in increased performance and efficiency compared to regular NICs. These benefits are automatically and transparently enabled when using Windows Server 2012 R2, which minimizes the user's management and configuration burden.

In addition to performance and efficiency gains, SMB over iWARP benefits from greatly improved data integrity protection, thanks to iWARP's end-to-end payload CRC (in lieu of simple checksums for the standard NIC). Furthermore, being especially designed for storage networking, T5 incorporates additional reliability features, including internal datapath CRC and ECC-protected memory. For these reasons, T5 adapters have been selected for Microsoft's Cloud Platform System, a scalable, turnkey private cloud solution.

Lustre over iWARP RDMA at 40Gbps[5]

Lustre is a scalable, secure and highly-available cluster file system that addresses extreme I/O needs, providing low latency and high throughput in large computing clusters. Lustre can use and benefit from RDMA, just like other storage protocols. This section compares the performance of Lustre RDMA over 40Gbps Ethernet and FDR InfiniBand, showing nearly identical performance. Unlike IB, iWARP provides a high-performance RDMA transport over standard Ethernet gear, with no special configuration needed or additional management costs. Thanks to its hardware TCP/IP foundation, it provides low latency and all the benefits of RDMA, with routability to scale to large clusters and long distances.

The following graphs compare Lustre READ and WRITE throughput over iWARP RDMA and IB-FDR, at different I/O sizes using the **fio** tool. The READ throughput numbers show 40 Gbps iWARP delivering nearly identical performance with IB-FDR (56 Gbps) over the range of interest.

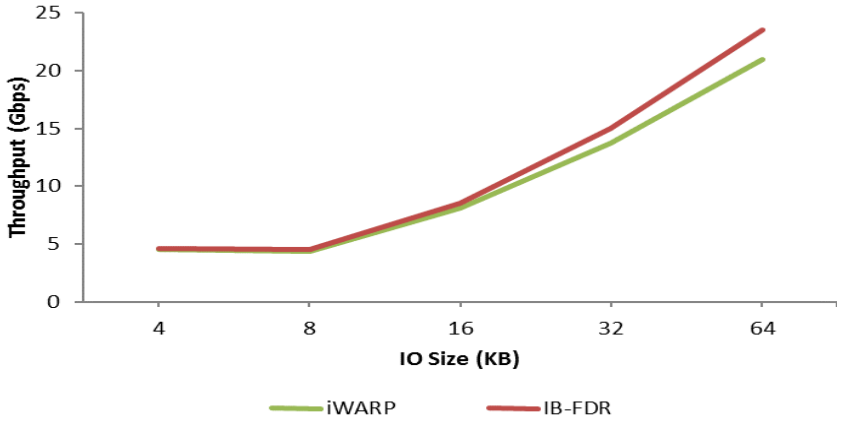

Figure 8 – READ Throughput vs. I/O size

The READ throughput numbers show 40 Gbps iWARP RDMA delivering nearly identical performance with IB-FDR (56 Gbps) over the range of interest. The WRITE results confirm the equality between the two transports, with nearly the same performance despite the theoretical bandwidth advantage of IB (56G vs. 40G for one port).

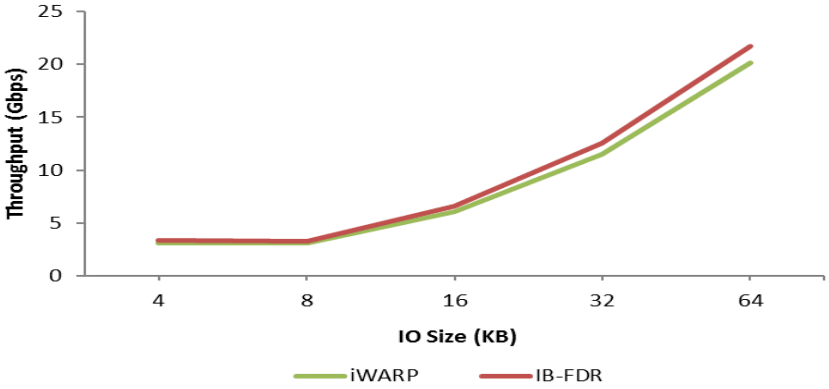

Figure 9 – WRITE Throughput vs. I/O size

[5] This section is excerpted from Chelsio white paper Lustre over iWARP RDMA at 40Gbps.

These results show that Lustre over iWARP RDMA at 40Gbps provides highly competitive performance compared to IB-FDR, and a superior performance curve in relation to the I/O size. This result is just one more in a series of studies of the two fabrics that consistently affirm this conclusion; that iWARP is a no-compromise, drop-in Ethernet replacement for IB.

NFS over RDMA at 40Gbps[6]

NFS over RDMA is an exciting development for the trusted, time proven NFS protocol, with the promise of high-performance and efficiency brought in by the RDMA transport. The unprecedented price and adoption curve of 40Gbps Ethernet, along with the focus on high efficiency and high-performance in an era of big data and massive datacenters, are driving the interest in performance optimized transports, such as RDMA. RDMA is also particularly interesting when mated to high throughput, low latency SSDs, where it allows extracting the most performance out of these new devices.

This section presents early performance results for NFS over 40Gbps iWARP RDMA, using Chelsio's Terminator 5 (T5) ASIC and InfiniBand-FDR. The following graphs compare NFS READ and NFS WRITE throughput and CPU idle percentage for iWARP RDMA and FDR InfiniBand at different I/O sizes using the **iozone** tool.

The results show that iWARP RDMA at 40Gbps and IB FDR (56Gbps) provide virtually the same NFS performance, in throughput and CPU utilization, although the latter is theoretically capable of higher wire rate. In addition, thanks to its TCP/IP foundation, iWARP RDMA allows using standard Ethernet equipment, with no special configuration and without requiring a fabric overhaul or additional acquisition and management costs. This result is another that affirms this conclusion; shown in all previous studies of the two fabrics, which makes iWARP RDMA a no compromise, drop-in Ethernet replacement for IB.

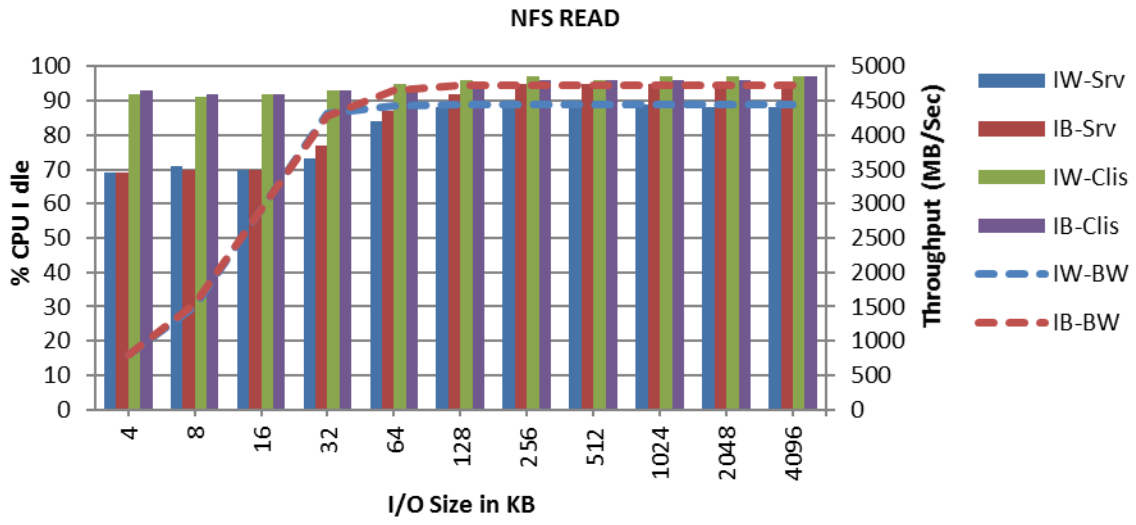

Figure 10 – READ Throughput & CPU Idle % vs. I/O Size

[6] This section is excerpted from Chelsio white paper NFS/RDMA over 40Gbps Ethernet.

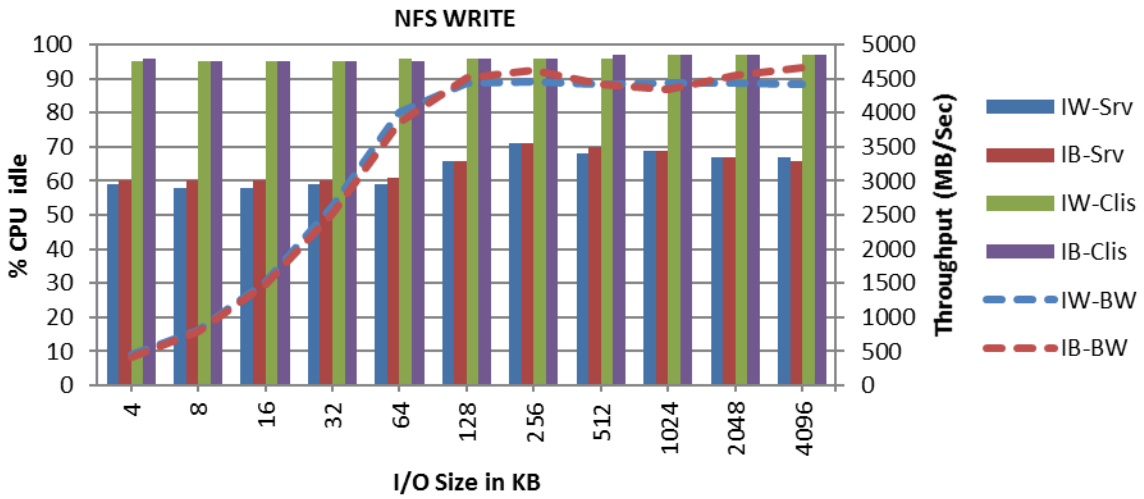

Figure 11 – WRITE Throughput & CPU Idle % vs. I/O Size

FCoE at 40Gbps with FC-BB-6[7]

This section reports FCoE performance results for Chelsio's Terminator 5 (T5) ASIC running at 40Gbps. T5 is the industry's first high-performance full FCoE offload solution with FC-BB-6 VN-to-VN and SAN management software support.

The results using a single FCoE target running over a Chelsio T580-CR Unified Wire Network adapter shows I/O numbers reaching nearly 4M per second, and line-rate throughput starting at I/O sizes as low as 2KB. The following graph plots the performance results, showing how line-rate 40Gbps is achieved at I/O sizes as small as 2KB, with peak IOPS nearing 4M/sec for READ and 2.5M/sec for WRITE.

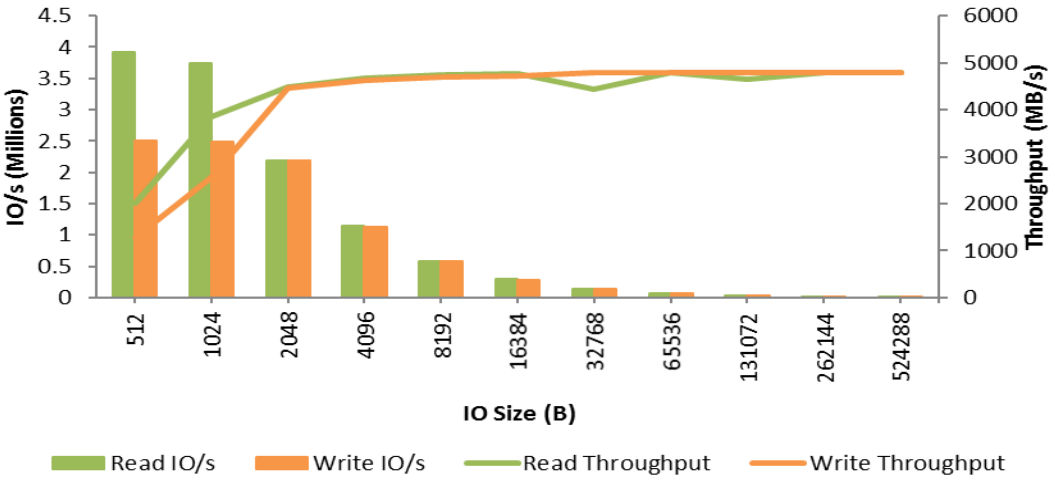

Figure 12 – Throughput and IOPS vs. I/O size

[7] This section is excerpted from Chelsio white paper FCoE at 40Gbps with FC-BB-6

Conclusion

Benchmarking results show Chelsio T5 to be an ideal fit for high-performance storage networking solutions across the full range of storage protocols. Internet SCSI (iSCSI) and iSCSI over RDMA (iSER) testing demonstrates that Chelsio T5 adapters provide consistently superior results, lower CPU utilization, higher throughput and drastically lower latency. Furthermore, thanks to using the routable and reliable TCP/IP as a foundation, T5 allows highly scalable and cost effective installations using regular Ethernet switches.

In addition, due to its hardware TCP/IP foundation, T5 provides low latency and all the benefits of RDMA for high-performance deployments such as for Windows Server 2012 SMB Direct, NFS over RDMA and Luster over RDMA, with routability to scale to large clusters and long distances. As a result, iWARP can be considered a no-compromise, drop-in Ethernet replacement for IB.

Related Links

The Chelsio Terminator 5 ASIC
T5 Offloaded iSCSI with T10-DIX
iSER: Frequently Asked Questions
SMBDirect Latency on Windows Server 2012 R2
iSCSI Heritage and Future