Building High Performance IP SANs with Chelsio's 10GbE iSCSI Solution

A Chelsio Communications White Paper

By Tom Burniece

The purpose of this white paper is to summarize requirements for a complete storage target processing solution for iSCSI at 10Gbps, providing maximum performance over the full range of storage applications.

Background

The prospect of massively-scalable storage and computing elements linked in a high-performance, low-cost fabric presents a compelling vision for solving the enterprise's needs for storage capacity, application performance, and network reach. These capabilities would enable computing and storage resources to be deployed and managed efficiently, providing high availability and business continuity across long distances. All of this will become a reality for the first time with iSCSI SANs -- Storage Area Networks that use the SCSI protocol delivered over low-cost 10Gbps Ethernet-based TCP/IP networks.

Although, iSCSI technology has been available for three years, it has garnered only three to four percent of the SAN market. Instead, Fiber Channel is the incumbent high-end SAN technology, providing optimized SCSI-compatible transport, high bandwidth (4Gbps), very low latency, and minimal load on the application CPUs. However, Fibre Channel is an expensive technology because it's largely limited to storage networks and requires specialized switches, host-bus adapters (HBAs), and management software. In contrast, iSCSI uses standard Ethernet switches, does not need special HBAs, and the server driver software is free. As a result, many predict that once Ethernet's practically-available bandwidth exceeds Fibre Channel's, iSCSI will take over.

What's been missing in that scenario is a cost-competitive 10Gbps Ethernet (10GbE) infrastructure and a complete iSCSI storage target processing solution that places minimal load on the application computing resources. That day has finally arrived with Chelsio's fully integrated 10GbE iSCSI target processing solution for 10GbE SANs using fiber or copper wiring. In this paper we will discuss how this combination delivers world-class storage performance and compares to Fibre Channel and proprietary interconnects such as InfiniBand and Myrinet.

The Value of 10GbE IP SANs

The value of iSCSI stems from its leverage of commodity Ethernet with its massive scale economies, widespread administrative expertise and management tools. Because of Ethernet, IP SANs enable reducing the deployment and administration costs for SANs – dramatically.

Heretofore, the major obstacle to adoption of iSCSI has been the lack of performance for storage I/O on the existing 100Mbps and Gigabit Ethernet infrastructure in certain applications. However, as 10GbE costs decline and it penetrates the data center, that changes. A 10GbE-based iSCSI SAN offers over twice the bandwidth of 4Gbps Fibre Channel SAN, along with the ease and flexibility of IP, and eliminates the specialized and expensive Fiber Channel equipment and administration. 10GbE-based IP SANs are already comparable in cost to 4Gbps Fiber Channel SANS and will ultimately be much lower.

There are a number of major applications for such a cost-effective IP Storage solution:

1. Enterprise-wide storage consolidation, where multiple application servers can access virtualized pools of shared storage for maximum efficiency and availability

2. SAN extension over IP-based WANs with both asynchronous and synchronous replication across very long distances for rapid disaster recovery and business continuance
3. SAN/NAS consolidation, where both block and file level access can be accomplished across a single unified fabric, using IP-based gateways to a consolidated back-end iSCSI SAN
4. Grid computing and clustering, where RDMA and multi-pathing provide very low latency and high availability, with enormous scalability

Market Development Scenario

Some of this is already starting to happen on a smaller scale on 100Mbps and 1Gbps Ethernet infrastructure, especially in the SMB (small and medium business) market. Buyers need to consolidate their storage from DAS (direct attached storage) to SANs and can't afford to invest in Fibre Channel. As a result, IDC estimates that iSCSI SAN revenue will grow rapidly from ~$300 million in 2005 to ~$2.9 billion by 2008. A complete, low-cost 10Gbps iSCSI storage target solution could accelerate that market growth.

Often in emerging markets, early vendors provide limited solutions to the overall problem. For a complete storage target processing solution for iSCSI at 10Gbps, offloading TCP processing from the CPU is not enough: it must eliminate unnecessary data movement into and out of CPU memories and also reduce the latency through acceleration and optimization of all protocols.

With the emergence of such a complete iSCSI target processing solution at 10Gbps, even high-throughput applications, such as video streaming, image processing, and backup, that previously were the exclusive territory of Fibre Channel, will become candidates for iSCSI. In addition, low latency applications, like clustering and grid computing, that have used Infiniband and Myrinet, will move to iSCSI. As a result, iSCSI will be well on its way to becoming a pervasive SAN interconnect for storage arrays and storage-aware devices, such as virtualization engines, as well as for blade servers. In addition, truly effective, high-performance backup, replication, archiving, and disaster recovery systems will be enabled, providing unprecedented levels of data availability and business continuity.

This will fulfill the long-term vision of converging networking, storage, and high-performance computing onto a common network infrastructure, based on Ethernet.

Requirements for an Integrated Solution for 10GbE iSCSI Target Processing

The purpose of this white paper is to summarize requirements for a complete storage target processing solution for iSCSI at 10Gbps, providing maximum performance over the full range of storage applications. Thus a complete storage target processing solution for iSCSI at 10Gbps comprises:

Hardware:

- TCP/IP offload
- iSCSI protocol acceleration
- RDMA
- Copper, as well as fibre, interconnects

Software:

- Sockets-based target stack
- Compatible software architecture for 1G/10G support
- Security
- High availability
- Scalability
- Interoperability
- Ease of Management

We will look at each of these in more detail.

TCP/IP offload In a networked device using a 10GbE link, the need for offloading the TCP/IP-protocol-processing burden from the CPU is acute, since otherwise it consumes all the cycles of a 20GHz processor just to drive a 10Gbps TCP/IP bi-directional link. A dedicated TCP offload engine (TOE), integrated into the network interface, can perform this task.

Early approaches to TOE design addressed only the TCP processing overhead issues, especially in the TCP Fast Path, and did not handle the exception issues in the TCP Main Path, such as connection establishment, re-transmissions, and out-of-order data. At 10Gbps these exceptions cause severe performance problems due to the increased likelihood of packet drop and subsequent re-ordering. In fact, interrupt handling and process scheduling are the dominant issues in running a TCP/IP software stack at 10Gbps. As a result, a high-performance 10Gbps TOE solution must offload all of these exceptions in hardware, including packet reassembly, at both ends of the wire.

iSCSI Protocol Acceleration Similar to TCP/IP offload, iSCSI protocol acceleration is needed to offload CPU-intensive iSCSI operations at 10Gbps, such as iSCSI header and data digest, Zero Copy, Direct Data Placement (DDP) and Protocol Data Unit (PDU) delineation. Executing these functions in hardware can result in very high performance for both high throughput and high transaction rate applications, with very low CPU utilization for protocol processing.

RDMA Remote Direct Memory Access (RDMA) is the mechanism of offloading data-copy operations from the application processor by allowing data to be transferred directly from the memory of one node on a network to the memory of another node on that network. RDMA with wire-speed latency is a key ingredient in clustering. Both Fibre Channel and Infiniband have inherent RDMA capability, but up until now Ethernet has not. As a result, the IETF has developed a new RDMA standard called iWarp (Internet Wide Area RDMA Protocol). iWARP runs RDMA over TCP/IP. Incorporating iWarp in iSCSI will allow direct array-to-array replication, as well as much more efficient sharing of storage in blade server systems and in grid computing and clustering applications

The iWarp spec combines many of the off-load concepts above, in order to eliminate the three major sources of network overhead in Ethernet: TCP/IP processing, intermediate buffer copies, and application context switching. It does this through a combination of TCP offload, OS kernel bypass, and RDMA. It is capable of offloading over 90% of the data transmission overhead from a server's CPU, yielding Infiniband-level latencies on Ethernet. iWarp also offers CRC-level reliability on TCP/IP. Extensions to iWarp are already being considered, such as a Datamover Architecture (DA) protocol and iSCSI-over-RDMA (iSER) that would be layered between iWarp and iSCSI, in order to provide port-to-port RDMA, using the DA interface. iSER is also being proposed as an extension to Infiniband.

Copper and Fibre Interconnects There are two copper-cabling standards for 10Gbps Ethernet: CX4, for which products are available today, and 10GBase-T, for which products are expected in 2006.

The CX4 standard uses the same 4-wire twin axial cabling as Infiniband, at about half the cost of fibre; and 10GBase-T uses unshielded twisted-pair (UTP) wiring, the most pervasive and least expensive Ethernet wiring installed today. A CX4 connection will only reach 50 feet at 10Gbps, while 10GBase-T is expected to reach 100 meters. Thus a low-cost fibre option will also still be needed in order to support longer distance 10GbE runs. Some CX4 hardware is already available for 10GbE with port costs comparable to Fibre Channel and it is expected to decrease rapidly. On a per-bandwidth basis, 10GbE CX4 hardware is already less expensive than 4Gbps Fibre Channel hardware and iSCSI costs will reduce even further.

Sockets-based Stack Using a socket-based stack provides seamless integration of iSCSI processing solutions into storage systems, using standard drivers. Sockets are the standard data transport abstraction in Ethernet that establishes the communication link between a specific pair of TCP or UDP ports, across operating systems such as Linux, UNIX, and BSD.

Common software architecture for 1GbE and 10GbE iSCSI
A common software architecture simplifies the migration path from 1GbE iSCSI to 10GbE iSCSI by enabling storage system providers to incorporate 10GbE support with no additional investment in software.

Security IP storage facilitates putting storage data on the same network with normal Ethernet traffic, unlike Fibre Channel SANs. As a result, the IETF included IPsec (Internet Protocol Security) encryption in the iSCSI specification.

However, most iSCSI implementations to date have used a separate, dedicated Ethernet network for iSCSI SAN. As a result, IPsec has not yet been heavily implemented in iSCSI SANs. In addition to IPsec, support for CHAP (Challenge Handshake Authentication Protocol), iSNS (Internet Storage Naming Service), and LUN (Logical Unit Number) access control security are also basic requirements for security in a complete iSCSI solution.

High availability The ability to survive a single point of failure through some form of redundancy allows an application to continue to run without interruption, or at least roll-back quickly to a good state and restart. This is accomplished through some combination of multi-pathing, concurrent copies of the data, and an automatic failover mechanism. The combination of 10GbE iSCSI and iWarp, along with Microsoft's MPIO (Multi-Path I/O) or Linux's built-in, multi-pathing drivers, should provide wire-speed, multiple-path, synchronous and asynchronous replication over long distances at low cost. This will enable iSCSI to bring very high availability to all enterprise SANs that was previously only available in the high-end with Fibre Channel and FCIP (Fibre Channel over IP) or iFCP (Internet Fibre Channel Protocol).

Scalability A complete iSCSI Solution needs to scale to thousands of connections with performance that will meet the needs of both high-transaction applications, like Exchange, SQL Server, and Oracle, as well as high-throughput applications like video streaming.

Interoperability In order for iSCSI to achieve its destiny as the pervasive SAN interconnect, as well as an enable networking and storage convergence, it is critical that it be fully interoperable across operating systems, applications, and heterogeneous systems.

Ease of Management One of the goals of iSCSI is to enable the same people who manage Ethernet networks to also manage storage, rather than require specialists, as happens today with Fibre Channel. While this is not necessarily as easy as it sounds, it needs to be a requirement for a complete iSCSI solution.

Chelsio's 10GbE iSCSI Processing Solution for Storage Arrays

The only vendor to deliver a complete, integrated solution to all of the requirements above for 10GbE iSCSI is Chelsio. We will examine their solution point-by-point:

TCP/IP Offload Chelsio's approach to a TOE is a full hardware implementation of the entire IETF RFC standards' TCP/IP stack, bypassing all software processing between the network interface and the application layer, including connection setup and teardown, timer management, retransmissions, and other exception handling. It does this with a pipelined, data-flow architecture, configured as a single VLIW processor and a memory switch that presents a large packet interface to the application, without breaking the dynamics of TCP on the wire. There is no caching in the pipeline and the FIFO buffering between stages of the pipeline is matched to the off-chip memory latency and bandwidth, so the processing rate is the same, no matter how many connections are in the pipeline. The result is it can provide four to seven times the performance per CPU of a software stack, even without eliminating data copies. When that is also implemented, the CPU utilization for packet processing drops by another factor of five. The Chelsio TOE is capable of handling 64,000 connections at once, with a set-up and tear-down rate of 3 million connections per second.

Chelsio's initial "Terminator" TOE chip has demonstrated line-rate performance before saturating the PCI-X. This is an aggregate throughput of 7 Gbps with a measured user-to-application latency of just 9 µsec and that is with standard 1,500B Ethernet frames. In fact, the Chelsio TOE architecture eliminates the need for Jumbo frames, which can cause 30%--50% performance degradation in high transaction rate applications, due to buffer conflicts.

Chelsio's full TCP/IP offload approach does not need connection state coordination with the host stack, which is a major issue with partial offload approaches. In addition, it can hide the limitations of the buffering resources in the switch and compete directly with Infiniband and Myrinet on a latency perspective.

iSCSI Protocol Acceleration Chelsio's iSCSI protocol acceleration is based on its capability to parse and split the headers for iSCSI and compute the header and payload separately. As a result, it can recover and delineate the PDUs on receive, perform Direct Data Placement for the recovered PDU, do Zero Copy, and offload CRC generation / checking, all in hardware. The combination of TCP offload and iSCSI protocol acceleration dramatically improves storage I/O performance and frees up the other CPU resources to do more useful work higher up in the stack, such as application layer data integrity checks, using the offloaded CRC, which is much more powerful than the weak Internet checksum protection. For example, the combination of Payload CRC and Zero Copy on receive has been measured to provide a 16 times improvement in performance per CPU cycle than a pure software stack. For iSCSI payloads the actual measured I/O performance is 854 Mbytes per second and 534k I/Os per second.

RDMA Chelsio currently supports iWarp v0.7 in hardware on PCI-X, enabling very low-latency array-to-array replication, and will soon support the entire iWarp 1.0 suite, including iSER.

Copper and Fibre Interconnects Chelsio currently supports CX4 copper, as well as MMF-850nm and SMF-1310nm fiber. iSCSI on CX4 can also be used in a DAS (Direct Attached Storage) configuration, substituting for UltraSCSI 320. Chelsio will also support 10GBase-T, when it is available.

Sockets-based Stack Chelsio has a highly optimized, sockets-based iSCSI target stack with support for manual iSNS discovery, one-way and mutual CHAP, LUN masking and access control security, ERL 0 (Error Recovery Level Zero), and the IETF RFC 3720 "must-level" features. This commercial-grade target stack is supported across the Chelsio product line, ensuring compatibility from 1Gbps to 10Gbps, plus seamless integration with existing storage applications. It runs on Linux 2.6.12 and has been certified on Microsoft WHQL iSCSI for Windows. It can be used in either standard Sockets mode or in an accelerated mode, using a Sockets API. Future releases will include support for iSCSI ERL 1 & 2, and MPIO.

In addition, Chelsio will be introducing a sockets-based iSCSI initiator stack later this year, running concurrently with the above target stack. Until then, the HBAs use open-source readily available initiators.

Common software architecture for 1GbE and 10GbE iSCSI Chelsio's GbE and 10GbE iSCSI adapter and controller products utilize the same software architecture, providing storage array vendors with a smooth migration path with no additional software investment.

Security Today, most iSCSI SANs are on dedicated Ethernet networks, so IPsec is not yet needed, and Chelsio does not have IPsec offload in its chip today. However, Chelsio will deliver IPsec in 2006. Once IPsec is integrated into a complete iSCSI solution, the convergence of WANs, LANs and SANs onto shared IP networks will start to become a reality. Chelsio already supports CHAP, iSNS, IQN (iSCSI Qualified Name), and LUN access control security in its iSCSI solution.

High Availability Chelsio's unique single-processor TOE approach assures that full bandwidth is available for each connection, with the added feature that the amount of allocated bandwidth can also be "dialed-in" to individual connections for optimal traffic management. Combined with iWarp and multi-pathing support, Chelsio's 10GbE iSCSI target processing solution will provide synchronous, as well as asynchronous, replication over long distance for very high availability.

Scalability With up to 64 thousand connections capability, the Chelsio iSCSI Solution will easily scale up in network size, as well as beyond 10Gbps in network speed.

Interoperability Chelsio products have been fully tested for operation with QLogic and Adaptec iSCSI HBAs, plus open-iSCSI initiator stacks from Microsoft, Cisco, and UNH.

Ease of Management Chelsio provides iSCSI MIBs with 64-bit counters, an SDK, programming guide and porting guide.

The Chelsio Terminator product family

Chelsio's current iSCSI product line includes:

Model	*Ethernet*	*Ports*	*Media*	*Bus*	*Protocols*
T204	GbE	4	Copper and Fiber	PCI-X	TCP, iSCSI
T210	10GbE	1	Fiber		TCP, iSCSI
T210-CX	10GbE	1	CX4 copper	PCI-X	TCP, iSCSI
N210	10GbE	1	Fiber		TCP – non TOE

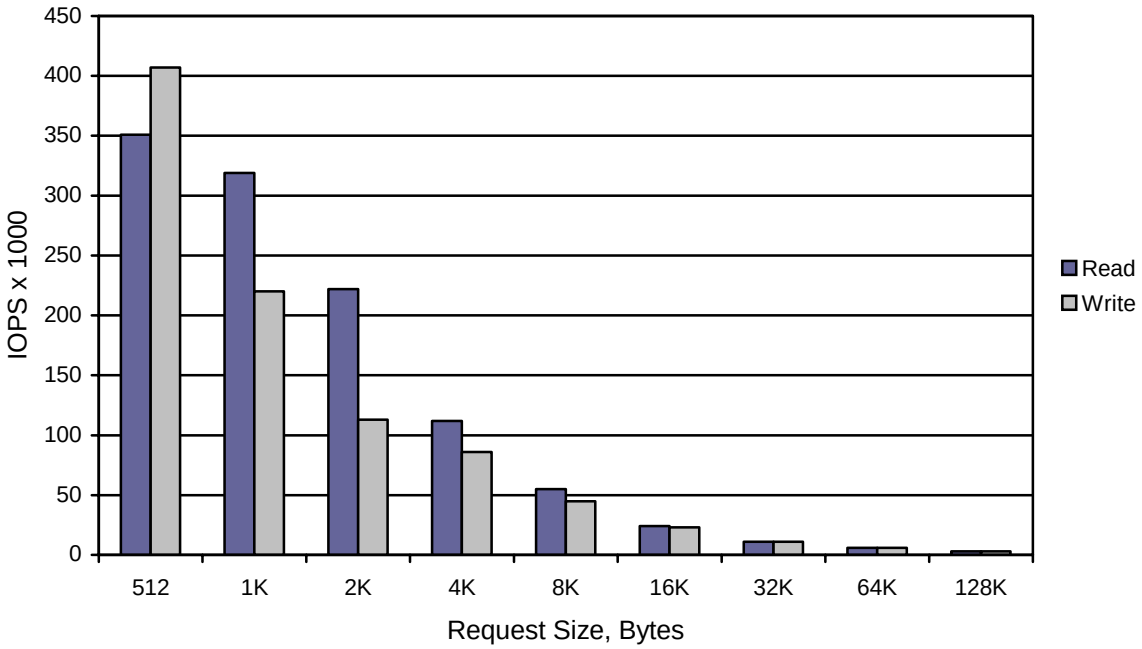
All are based on the Terminator ASIC, which is now in its second-generation, and are optimized for storage OEM applications. Each of the adapter and controller products are supported by the same commercial-grade iSCSI target-mode stack. More data on each of these products can be found on the Chelsio website at http://www.chelsio.com.

iSCSI Performance

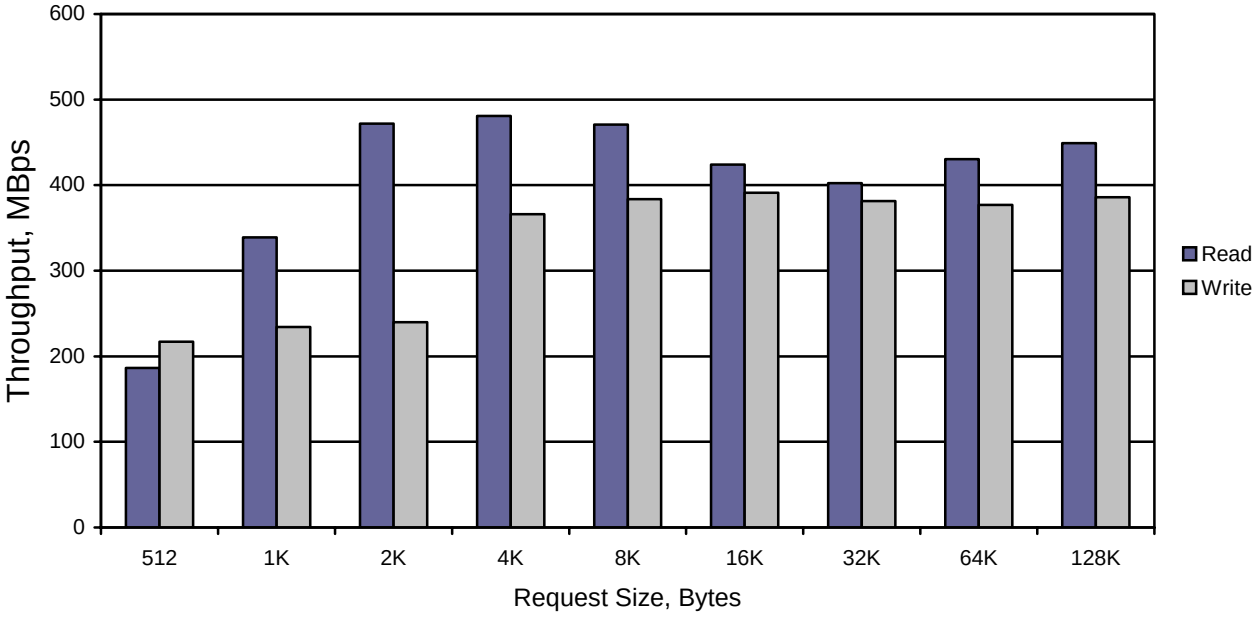
Chelsio's 10GbE iSCSI solutions deliver very high transaction-rate performance for applications like OLTP, messaging, and CRM, as well as wire-speed through-put performance for application like video streaming, backup, and HPC. In addition, its low latency and RDMA capability make it ideal for synchronous and asynchronous replication across distance for disaster recovery and business continuance

In the tests below, the Chelsio T204 was deployed in a link-aggregated configuration in an iSCSI target mode server, powered by dual 2.4 GHz Opteron processors, running the Chelsio iSCSI target stack on a RAM disk. The target mode server was connected to 16 Windows 2003 servers running the Microsoft iSCSI initiator and the industry-standard Iometer tool was used for measuring 512-byte initiator reads and writes. Measurements on the T204 showed an internal throughput of up to 3.9 Gbps with less than 20% CPU utilization, using standard 1500 byte Ethernet frames.

The T204 iSCSI target read and write IOPS performance versus request size is shown below. This is very comparable to Fibre Channel performance.

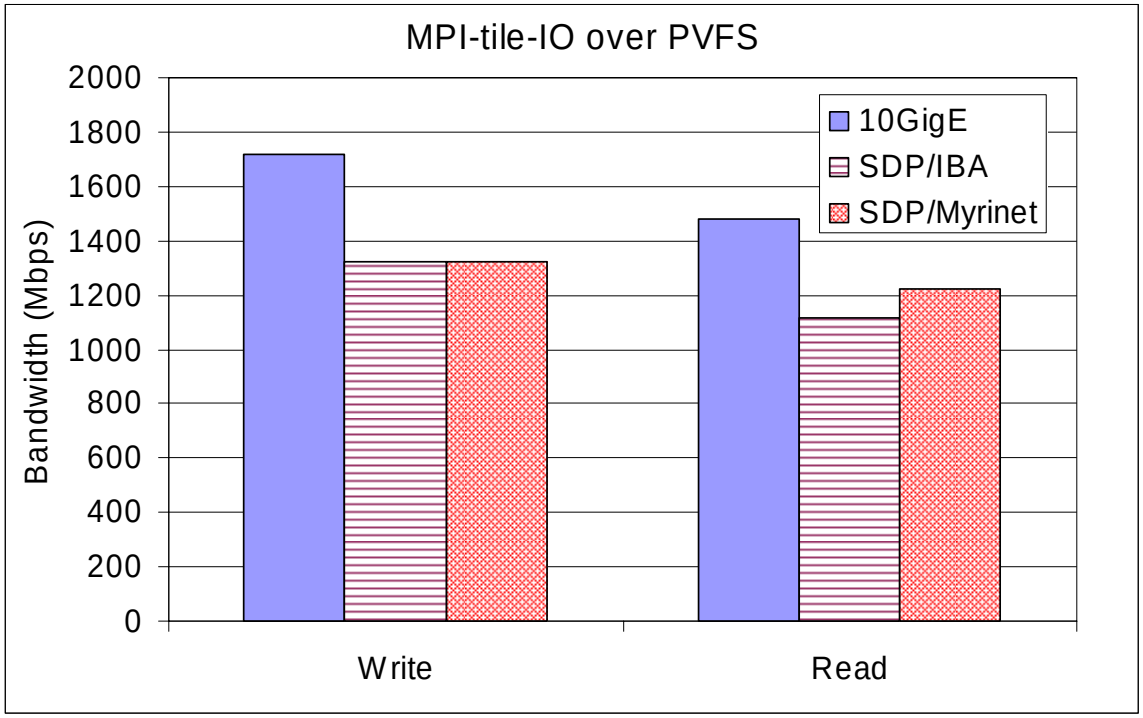


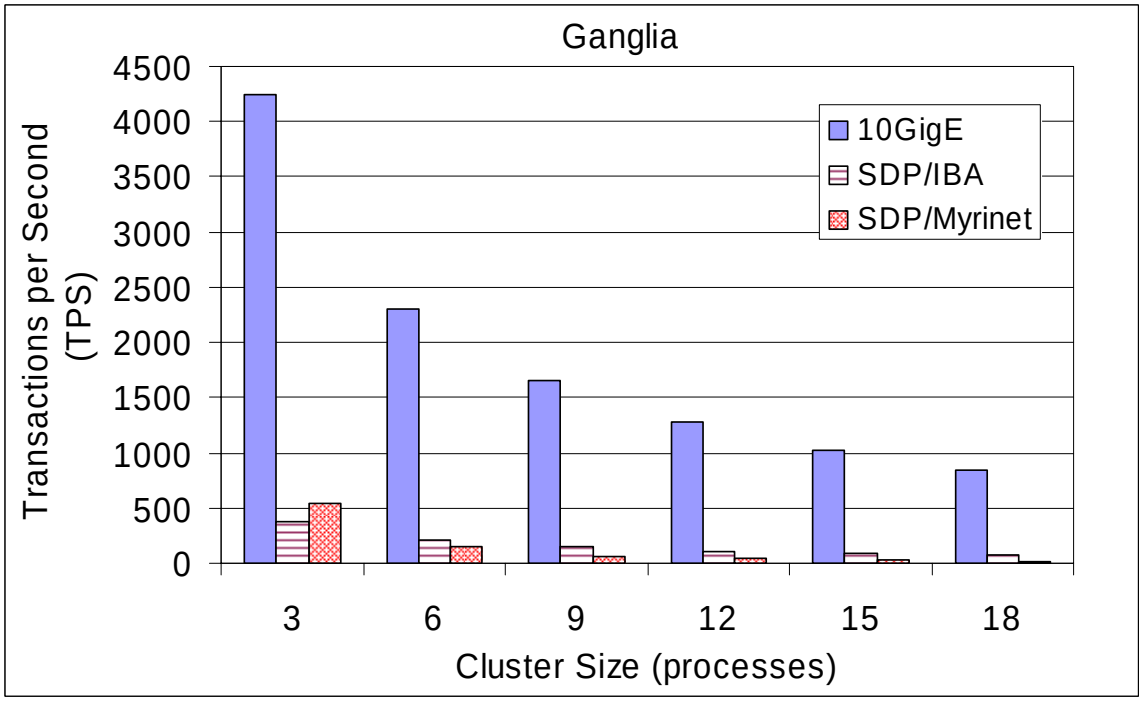
The T204 iSCSI target through-put read and write tests showed consistent wire-speed performance across different request sizes, as shown below. This exceeds 4Gbps Fibre Channel performance.



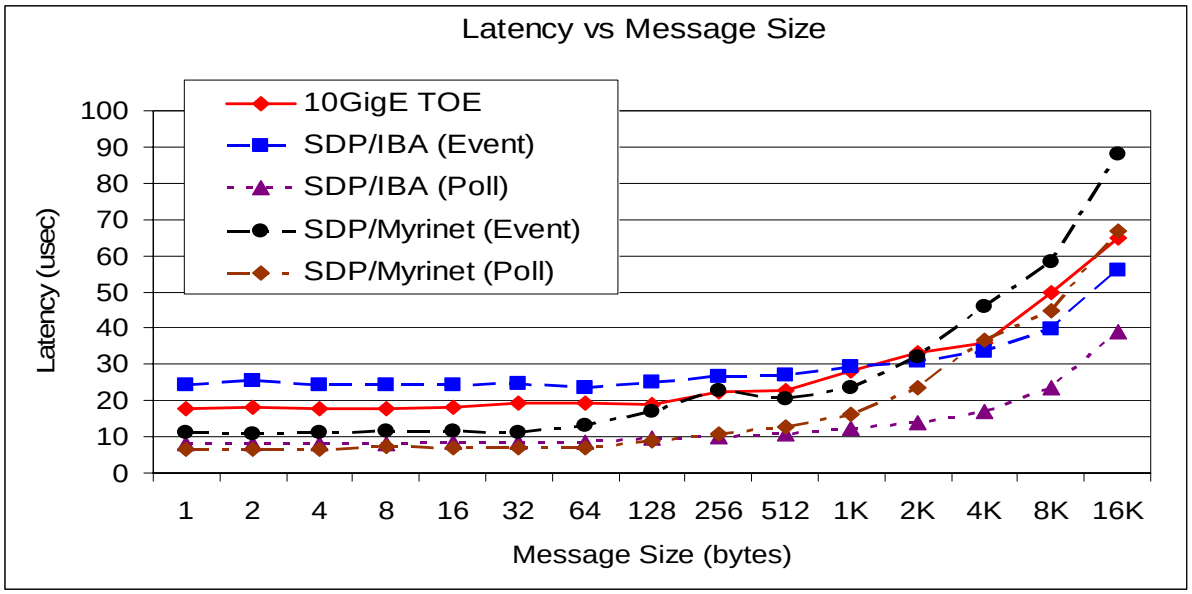
In addition, three independent performance studies have been published on Chelsio's 10GbE solution
(see: http://www.chelsio.com/technology/Chelsio_10GbE_TOE_Perf_Bmarks.pdf)

Some selected results from these tests, comparing Chelsio with Infiniband and Myrinet in two high-performance computing applications, are shown below:

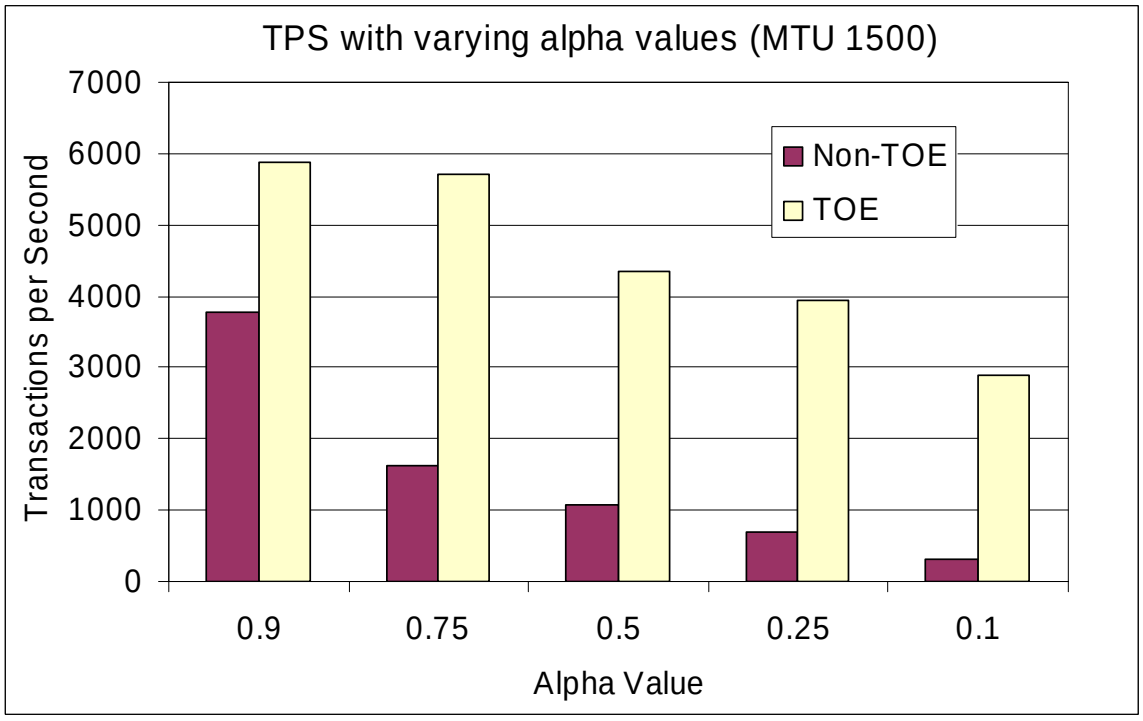




Head-to-head latency tests, for a 4-node cluster with dual 32-bit Xeon 3GHz processors running Red Hat 9.0 Linux kernel 2.4-25smp over a 10GbE switch with a single connection, are shown below:



In addition, the Transaction per Second (TPS) performance of the Apache Web Server application with the Cheliso 10GbE TOE, versus no TOE, was measured as shown below:



Based on these and other published test results, Chelsio's iSCSI solution can clearly provide world-class performance results in high transaction rate and high throughput applications, as well as in latency critical storage applications, such as synchronous replication.

Conclusions

Chelsio has developed the first complete 10GbE iSCSI target processing solution and it is showing results that enable iSCSI to challenge Fibre Channel, Infiniband, and Myrinet in high-end applications, as well as to bring about the long-anticipated convergence of networking and storage into a unified Ethernet-based fabric

Tom Burniece is an independent consultant in Silicon Valley with over 30 years of experience in the storage industry, specializing in strategy formulation, business development, technology evaluation, and due diligence