Chelsio 10GbE iSCSI

Enabling 'No-Compromise' IP SANs For Enterprise Networks

Executive Summary

10GbE-based network convergence helps enterprises address storage infrastructure challenges including optimized expenses, improved performance, and added efficiencies. Ethernet-based storage network convergence enables unification of file and block data, with NFS, CIFS, iSCSI, and the emerging FCoE protocol all running on 10 Gigabit Ethernet and supporting and benefiting from the advances in Ethernet technology.

The Ethernet attached storage market has been growing every year since 2003. In terms of unit shipments, IDC estimates that the iSCSI market grew 40% from 2008 to 2009 vs. 12% and -6% for the NAS and FC markets, respectively. While iSCSI and NAS unit shipments in aggregate grew to exceed FC shipments in 2008, NAS by itself crossed over to exceed FC unit shipments in 2009. By 2012, it is expected that 10GbE iSCSI will be the clear ROI and performance winner vis-à-vis 10GbE FCoE and 8G FC.

- Ethernet economics will drive the switch plus NIC per-port costs for 10GbE iSCSI to $480 in 2011, while FCoE and 8G FC per-port pricing in 2011 is $1,230 and $930 respectively.
- Projected deployment of unified 10GbE NIC as LOM (LAN on motherboard) components in 2012 will drive a dramatic increase in 10GbE iSCSI volumes and a corresponding reduction in per-port price for 10GbE iSCSI to $150.
- iSCSI IOPS and bandwidth performance will grow ten-fold (vs. only two-fold for FC) by 2012.

Targeted at LOM design opportunities, the 4th generation Chelsio Terminator 4 (T4) is a highly integrated, highly virtualized 10GbE controller built around a programmable protocol-processing engine. T4 includes Chelsio's fourth-generation TCP offload (TOE) design, third generation iSCSI design, and second-generation iWARP (RDMA) implementation, and full FCoE offload.

T4 also supports full hardware-based iSCSI protocol offload, which is an essential requirement for 10GbE IP SANs to deliver on the promise of enhanced performance, scalability, and return-on-investment vis-à-vis Fibre Channel, including comprehensive bare metal provisioning and management capabilities that come from hardware-based boot-from-SAN technology.

The hardware-based T4 full iSCSI offload approach provides an OS-agnostic boot-from-SAN. Full-offload iSCSI initiators are required for hypervisors, such as VMware ESX, for iSCSI acceleration and to boot from a SAN.

Copyright 2010. Chelsio Communications. All rights reserved. www.chelsio.com

Chelsio T4 10GbE also support acceleration of Ethernet-based low-latency iWARP (RDMA/TCP)-based storage protocols, including file storage (NFS-RDMA) and Lustre Networking (LNET). T4 implements Chelsio's second-generation 10GbE iWARP (RDMA over TCP/IP) functionality and reduces RDMA latency from T3's already low six microseconds to about two microseconds.

Introduction

The vision of data center fabric convergence depends on the ability of an adapter, switch, and/or storage system to use the same Ethernet physical infrastructure to carry different types of traffic with very different characteristics and handling requirements. For the IT network manager, this equates to installing and operating a single network, while still having the ability to differentiate between traffic types.

Network convergence promises to support both storage and network traffic on a single network. One of the primary enablers of fabric convergence is *10 Gigabit Ethernet*, a technology with sufficient bandwidth and latency characteristics to support multiple traffic flows on the same link. Network convergence also helps enterprises address storage infrastructure challenges including optimized expenses, improved performance and added efficiencies.

- The explosion of data growth across corporations forces IT administrators to deal with complicated and disparate server and storage systems and management, data center space and infrastructure constraints, and new budget pressures associated with environmental factors.
- Deploying dissimilar SAN and NAS storage solutions, each with its own platform and management, results in separate network storage infrastructures. Each infrastructure brings the potential for underutilized and captive storage, requiring multiple data recovery solutions, a variety of data management models, and different teams of people.

In addition to lower complexity and costs, and improved efficiencies and utilization, Ethernet-based storage network convergence enables unification of file and block data, with NFS, CIFS, iSCSI, and the FCoE protocol all running on 10 Gigabit Ethernet and supporting and benefiting from the advances in Ethernet technology.

10GbE iSCSI: The Game Changer for Enterprise-Grade SANs

The Ethernet attached storage market, which has been growing every year since 2003 and is driven by NAS (NFS, CIFS) and iSCSI deployments, is very well positioned to continue its strong growth versus the relatively declining growth for Fibre Channel-enabled storage. The driver in the growth of Ethernet-attached storage is the emergence of unified storage solutions, with iSCSI and NAS support, and the vast support of the Ethernet ecosystem.

According to IDC[1], iSCSI has consistently grown faster than the overall networked storage market.

- In terms of unit shipments, IDC estimates the iSCSI market grew 40%, from 2008 to 2009 vs. 12% and -6% for the NAS and FC markets respectively.

[1] IDC Worldwide Storage Tracker, December 2009

Copyright 2010. Chelsio Communications. All rights reserved. www.chelsio.com

- While iSCSI and NAS unit shipments in aggregate grew to exceed FC shipments in 2008, NAS by itself crossed over to exceed FC in 2009.

At the end of 2009, iSCSI accounted for 20% of networked storage unit shipments, with Fibre Channel accounting for 39% and NAS for 41%.

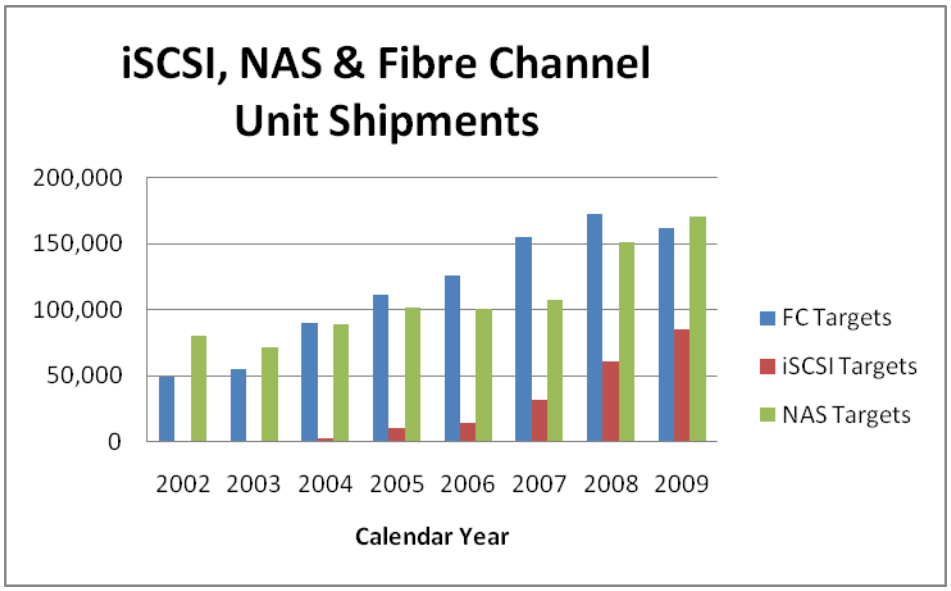

Figure 1: iSCSI, NAS, and Fibre Channel Storage Unit Shipments (IDC, 2009)

iSCSI over 10GbE networks provides performance and efficiency comparable to traditional Fibre Channel SANs and enables deployment of IP SANs for the most demanding of enterprise applications, such as data mining and decision support applications. Thus, 10GbE allows IP SANs, because of their cost, flexibility, and performance advantages, to emerge as a real challenger to Fibre Channel in medium to large enterprise SANs and is the determining factor in iSCSI's projected market share gains at the expense of Fibre Channel.

In the past year, we have also witnessed the emergence of the FCoE protocol. The vision behind FCoE, which enables the transport of Fibre Channel frames over Ethernet, is to allow IT to move to a "converged" 10GbE Ethernet fabric in the data center, while at the same time preserving investments in FC arrays and SAN management expertise. Use of Fibre Channel's upper layers for FCoE allows the leveraging of Fibre Channel software stacks, management tools, and trained administrators. As such, FCoE is geared to enable backward compatibility with existing FC infrastructures and is unlikely to displace or replace iSCSI growth for its sweet-spot use scenarios.

FCoE requires a number of additional capabilities and features, commonly called Data Center Bridging (DCB), in 10GbE NICs, switches, and storage targets, to support the convergence of the Fibre Channel Fabric onto an Ethernet storage network. Thus, it is likely to remain an expensive niche interconnect for the foreseeable future.

Copyright 2010. Chelsio Communications. All rights reserved. www.chelsio.com

Although the new DCB features for Ethernet are not required to support iSCSI (as they are with FCoE), their benefits can also be extended to IP SANs to deliver a more robust, higher performance iSCSI SAN implementation.

SAN Fabric Cost and Performance Comparison

At the end of the day, with multiple competing SAN alternatives, the decision of which SAN fabric to deploy typically comes down to cost and performance, and here is where the true value of a unified 10 GbE-based converged fabric emerges.

SAN Fabric CapEx Model

A comparative analysis of historical and projected data for end-to-end switch plus NIC per-port acquisition costs shows that by 2012 10GbE iSCSI will be the clear ROI winner market vis-à-vis 10GbE FCoE and 8G FC.

- In 2009, end-to-end per-port cost for 10GbE iSCSI was $1,125, while the corresponding costs for 8G FC and 10GbE FCoE were $1,275 and $5,000 respectively.
- Ethernet economics will drive the per-port costs for 10GbE iSCSI to $480 in 2011, while FCoE and 8G FC per-port pricing in 2011 $1,230 and $930 respectively (driven by premium-pricing strategy pursued by FCoE and FC vendors).
- Project deployment of unified 10GbE NIC as LOM components in 2012 will drive dramatic increase in 10GbE iSCSI volumes and a corresponding reduction in per-port price for 10GbE iSCSI to $150.

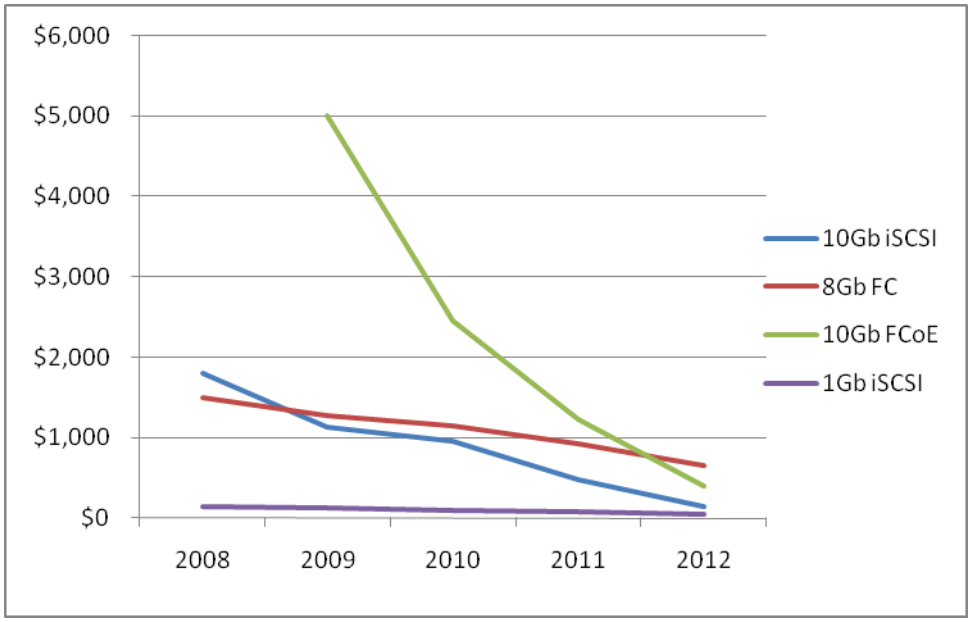

Figure 2. SAN Fabric NIC and Switch Port Acquisition Cost Comparison

Copyright 2010. Chelsio Communications. All rights reserved. www.chelsio.com

Here is a summary of the assumptions behind the above model.

	2008	2009	2010	2011	2012	Assumptions
1Gb iSCSI server port	$0	$0	$0	$0	$0	GbE is a LOM capability
1GbE switch port	$150	$125	$100	$75	$50	
10GbE iSCSI server port	$700	$550	$600	$250	$0	10GbE iSCSI is on LOM in 2012
10GbE switch port (non-CEE)	$1,000	$500	$300	$200	$150	
10GbE cable	$100	$75	$50	$30	$0	LOM and Cat6/7 standardization in 2012 drives 10GbE cable cost to zero
FC HBA	$700	$600	$750	$600	$450	FC will not be a LOM capability
Effective switch port cost for FC-enabled servers	$700	$600	$350	$300	$200	FC-enabled servers also require GbE networking
Effect cable costs for FC-enabled servers	$100	$75	$50	$30	$0	FC-enabled servers require FC and GbE cabling
FCoE Adapter		$1,000	$800	$400	$100	FCoE LOM support (2012) requires separately-priced firmware upgrade
CEE switch port	$5,000	$2,000	$800	$400	$150	Assumes Cisco/Brocade offer DCB as a separately-priced option
FCoE SW Upgrade for CEE switch	$5,000	$2,000	$800	$400	$150	FCoE support for CEE swtich ports requires per-port licensing
CEE cable			$50	$30	$0	Cat6/7 standardization in 2012 drives 10GbE cable cost to zero

Table 1: Assumptions for SAN Fabric Acquisition Cost Comparison Model

10GbE iSCSI vs. Fibre Channel Performance

Quad-port link aggregation and PCIe Gen2 allow maximum per-connection bandwidth for 10GbE iSCSI to be 40Gb/s. In addition, from a roadmap standpoint, the Ethernet market is moving forward aggressively with availability of standard 40G and 100G-based offerings in 2011 and 2012 respectively. In comparison, deployment of 16G FC, the next step in evolution of FC SAN performance is not expected until 2012 (Figure 3).

In summary, significantly higher iSCSI bandwidth performance growth vs. FC makes it a superior SAN fabric for the range of throughput-intensive applications such as back-up, video production, seismic processing, decision support, and data mining.

10GbE iSCSI vs. FC IOPS Performance

Input/Output Operations per Second (IOPS) is a common benchmark for measuring the effectiveness of SANs for supporting transactional applications, such as messaging or on-line transaction processing applications.

Driven by the Ethernet market roadmap, iSCSI IOPS performance is expected to grow ten-fold during 2009-2012 timeframe, while FC IOPS performance is only expected to grow two-fold, making iSCSI the best-in-class offering for transactional applications (Figure 4).

Copyright 2010. Chelsio Communications. All rights reserved. www.chelsio.com

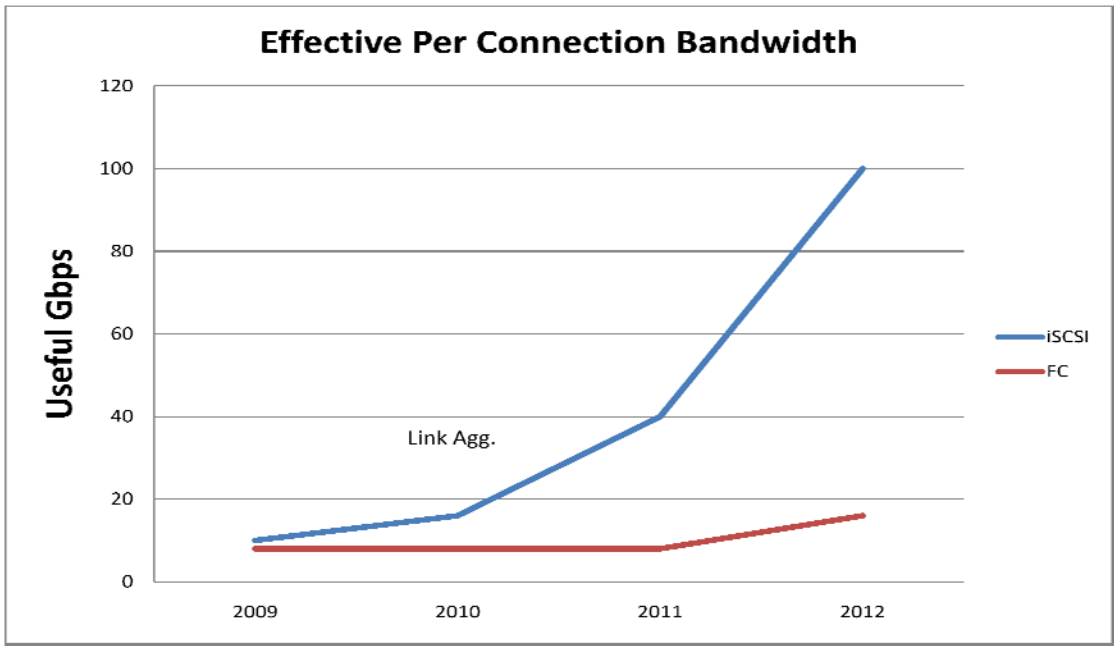

Figure 3. Projected Evolution in Bandwidth Performance for iSCSI and FC

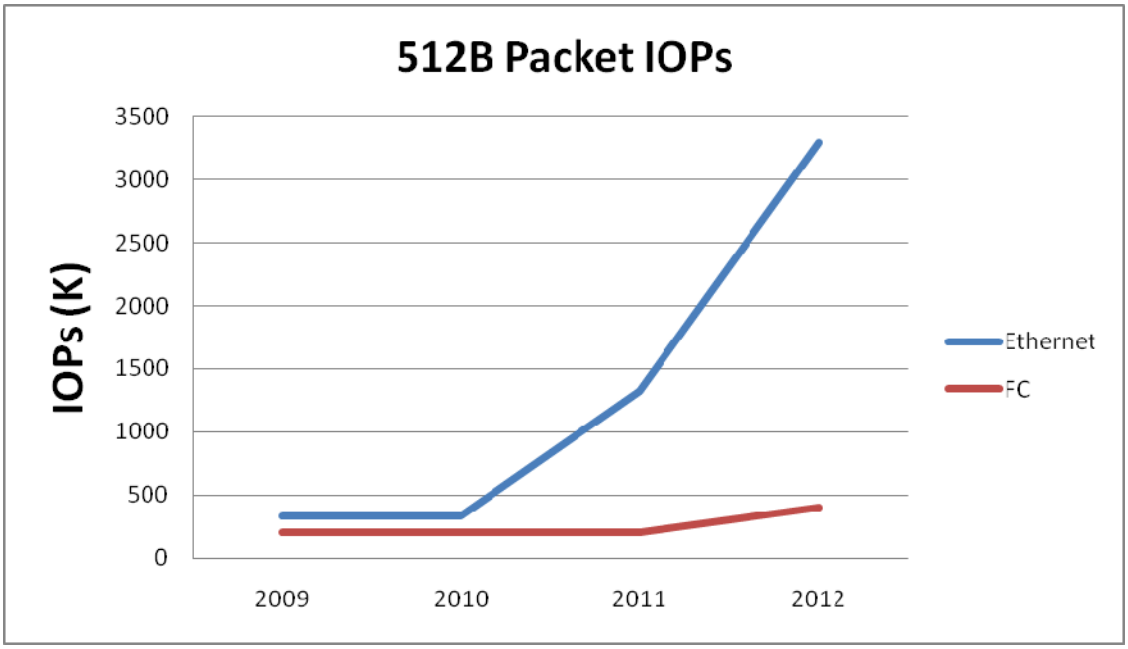

Figure 4. Projected Evolution in IOPS Performance for iSCSI and FC

Copyright 2010. Chelsio Communications. All rights reserved. www.chelsio.com

Chelsio T4 iSCSI Capabilities

Chelsio provides a range of 10GbE adapters, ASICs, and storage solutions which enable the convergence of server networking, storage networking, and cluster computing interconnects onto a single platform and a single fabric. At the root of the Chelsio solution set is the Terminator, a proven, protocol-rich 10GbE ASIC that has secured a broad range of OEM platform wins, including server and storage wins, and is broadly deployed.

Targeted at LOM design opportunities, the 4th generation Chelsio Terminator 4 (T4) is a highly integrated, highly virtualized 10GbE controller built around a programmable protocol-processing engine. T4 includes Chelsio's fourth-generation TCP offload (TOE) design, third generation iSCSI design, and second-generation iWARP (RDMA) implementation, and full FCoE offload.

From a storage standpoint, T4 provides support for both *partial* and *full* iSCSI offload. Leveraging its 10GbE built-in TOE and iSCSI encryption and digest capabilities, T4 boosts iSCSI performance for environments using operating system-specific iSCSI initiators such as the iSCSI initiator within Windows and the Linux *open-iSCSI* initiator.

T4 also supports full hardware-based iSCSI protocol offload, which is an essential requirement for 10GbE IP SANs to deliver on the promise of enhanced performance, scalability, and return-on-investment vis-à-vis Fibre Channel, including comprehensive bare metal provisioning and management capabilities that come from hardware-based boot-from-SAN technology.

With hardware-based full iSCSI offload, SCSI commands issued by the OS are offloaded to the Chelsio T4, converted into TCP/IP packets and transmitted to the iSCSI storage target that stores the disks. To the OS, the remote storage device appears as a locally attached SCSI device. The hardware-based T4 full iSCSI offload approach provides an OS-agnostic boot-from-SAN. Full-offload iSCSI initiators are required for hypervisors, such as VMware ESX, for iSCSI acceleration and to boot from a SAN.

Chelsio iSCSI Software Support

Chelsio's Terminator-based ASIC and adapter solutions provide iSCSI acceleration for a full range of operating environments, including:

- A hardware offloaded iSCSI initiator driver, enabling 10GbE line-rate iSCSI performance for Chelsio unified wire adapters in Windows 2003, Windows 2008-SP2 and Windows 2008-R2 environments.
- Chelsio unified wire adapters support TCP Chimney interface for TOE integration within Windows to enable accelerated performance for the native iSCSI initiator within Windows.
- Support for Chelsio unified wire adapter iSCSI offload capability for the Open-iSCSI initiator arrives out-of-the-box in major Linux distributions.

Future plans include leveraging the full iSCSI offload capability within T4 to offer a SAN boot-enabled 10GbE iSCSI HBA solution for virtualized environments including VMWare Vsphere.

Copyright 2010. Chelsio Communications. All rights reserved. www.chelsio.com

Low Latency Storage Convergence

10GbE iWARP (RDMA over TCP/IP) leverages its Ethernet TCP/IP heritage to support acceleration of Ethernet-based storage protocols, including file storage (NFS over RDMA) and Lustre Networking (LNET). NFSoRDMA allows an NFS client and server to communicate using an RDMA capable network transport, such as iWARP. Benefits of using an RDMA transport for NFS include significantly lower CPU utilization and higher throughput.

Lustre is an object-based, distributed file system, generally used for large scale cluster computing in the research/scientific, oil and gas, manufacturing, rich media and finance markets. LNET provides the communications infrastructure required for Lustre file system deployments.

The Linux NFS-RDMA implementation and LNET run on the OpenFabrics Enterprise Distribution (OFED) RDMA stack. This stack provides the software infrastructure and device driver support for RDMA capable devices. This means that applications such as NFS-RDMA and Lustre LNET can run over iWARP on devices without change.

T4 Implements Chelsio's second-generation 10GbE iWARP (RDMA over TCP/IP) functionality building on the iWARP capabilities of T3, which have been field proven in numerous large, 100+ node clusters, including a 1300-node cluster at Purdue University.

The T4 design reduces RDMA latency from T3's already low six microseconds to about two microseconds. Chelsio achieved this three-fold latency reduction through straightforward increases to T4's pipeline speed and controller-processor speed, demonstrating the scalability of the RDMA architecture established by T3.

Terminator Architecture Advantage

Two approaches to implementing ASIC-based 10GbE iSCSI, iWARP, and TCP/IP protocol processing can be followed. The first uses a system-on-a-chip (SOC) implementation with multiple RISC processor cores and special-purpose engines running TCP/IP and, optionally, iSCSI and iWARP protocols in firmware, while the second utilized by the Chelsio ASIC is characterized by iSCSI, iWARP, and TCP/IP protocols implemented in microcode running on a pipelined VLIW processor implementation.

While both approaches have their advantages, the multiple-RISC SOC implementation also presents a number of scaling problems. First, performing complete protocol processing in firmware running on a single CPU leads to higher latency. Because iSCSI and iWARP operate on top of the TOE, processing these protocols only adds to total latency. Second, some multi-RISC designs cannot achieve 10 Gigabit wire-rate speeds below a minimum threshold number of connections thereby impacting the utility of such approaches for storage applications such as backup and replication.

Conclusion

10GbE iSCSI offers the lowest cost of ownership and the highest performance among SAN fabric alternatives. The Chelsio T4 offers a 3rd generation iSCSI protocol offload design, which is an essential requirement for 10GbE IP SANs to deliver on the promise of enhanced performance, scalability, and return-on-investment vis-à-vis Fibre Channel.

Copyright 2010. Chelsio Communications. All rights reserved. www.chelsio.com