WHITE PAPER

Chelsio

High-Performance Networking for Optimized Hadoop Deployments

Chelsio Terminator 4 (T4) Unified Wire adapters deliver a range of performance gains for Hadoop by bringing the Hadoop cluster networking into optimum balance with the recent improvements in server and storage performance, while minimizing the impact of high-speed networking on the server CPU. The result is improved Hadoop Distributed File System (HDFS) performance and reduced job execution times.

Executive Summary

Typical Hadoop scale-out cluster servers utilize TCP/IP networking over one or more Gigabit Ethernet network interface cards (NICs) connected to a Gigabit Ethernet network. However, the latest genera tion of commodity servers offers multi-socket, multicore CPU technology like Intel's Nehalem, which outstrips the network capacity offered by GbE networks. With advances in processor technology, this mismatch between server and network performance is predi cted to grow. Similarly, while Solid State Disks (SSDs) are evolving to offer equivalent capacity-per-dollar of Hard Disk Drives (HDDs), they are also being rapidly adopted for caching and for use with medium-sized datasets. The advances in storage I/O performance offered by SSDs exceeds the performance offered by GbE networking, which makes network I/O increasingly the most common impediment to improved Hadoop cluster performance.

10 Gigabit Ethernet has the potential of bringing the Hadoop cluster netwo rking into balance with the recent improvements in performance brought by server CPUs and advances in storage technology. Balancing the performance of server network I/O improves the efficiency of every server in the cluster, thus improving the performance of the entire Hadoop infrastructure by removing critical bottlenecks. *However, to achieve optimum balance, the network I/O gains delivered by 10GbE must come with optimal efficiency so that the impact of high-speed network I/O on the server CPU is minimized*.

The Chelsio T420 10GbE Unified Wired Adapter utilizes Chelsio's Terminator 4 (T4), a highly integrated 10GbE ASIC chip built around a programmable protocol processing engine. The T4 ASIC represents Chelsio's fourth generation TCP offload engine (TOE) design and second generation iWARP (RDMA) implementation. The T4 demonstrates better performance across a range of benchmarks than that for Terminator 3 (T3), while running the same microcode that has been field-proven in very large clusters.

Chelsio's 10GbE adapters improve network performance by leveraging an embedded TOE, which offloads TCP/IP stack processing to the server NIC. Used primarily with high-speed interfaces such as

10GbE, the TOE frees up memory bandwidth and valuable CPU cycles on the server, delivering the high throughput and low latency needed for cluster networking applications, while leveraging Ethernet's ubiquity, scalability, and cost-effectiveness.

Chelsio 10GbE TOE is aimed at increasing the ability to move data into and out of a server in multi-core processing server environments and has proven to be ideal for environments where servers are concurrently required to run CPU as well as network I/O-intensive tasks such as Hadoop cluster networking. A key benefit of the Chelsio 10GbE TOE implementation is that it maintains full socket streaming semantics enabling applications, such as HDFS, using the sockets programming model to leverage its performance capabilities without modification.

The Chelsio T4 second generation iWARP design builds on the RDMA capabilities of T3 while leveraging the embedded TOE capabilities, with advanced techniques to reduce CPU overhead, memory bandwidth utilization, and latency by combining offloading of TCP/IP processing from the CPU, eliminating unnecessary buffering, and dramatically reducing expensive operating system calls and context switches – thereby moving data management and network protocol processing to the T420 10GbE Unified Wire Adapter. The OpenFabrics software stack that is fully integrated into the flavors of Linux distributed by Novell and Red Hat fully supports 10GbE iWARP.

Independent benchmarking has determined a whole range of performance advantages of Chelsio 10GbE solutions compared to GbE for Hadoop cluster environments. Chelsio 10GbE adapters were found to lead to network and storage I/O balance in Hadoop cluster nodes and while clearly complementing the use of SSDs in dramatically improving HDFS sequential and random write performance, and reducing job execution times.

Introduction -- The Big Data Imperative

The growing number of people, devices, and sensors that are now interconnected by digital networks has revolutionized our ability to generate, communicate, share and access data. In 2010, more than four billion people (or 60 percent of the world's population) were connected by cell phones and 12 percent of those were using smart phones, whose market penetration was growing at 12 percent per year. Today, more than 30 million networked sensors are deployed in the transportation, automotive, utilities, and retail sectors. The number of sensors is increasing at a rate of 30 percent per year. This has created a large data management problem, which is now referred to as "The Big Data Imperative."

"Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. As technology advances over time, the size of data sets that qualify as big data will also increase. The definition will also depend on the sizes of datasets common in a particular industry and the types of software tools available. Hadoop is being widely accepted as a solution to the problem.

The Apache Hadoop open source software addresses the problems associated with big data in two key ways. First, it provides a highly scalable distributed file system, called the *Hadoop Distributed File System* (HDFS), which is used for storing, managing, and securing very large datasets. Second, the Hadoop *MapReduce* framework provides a powerful programming model capable of harnessing the computing power of several commodity servers into a single high-performance computing cluster capable of efficiently analyzing large datasets.

Hadoop Overview and Benefits

Hadoop is a powerful, fault-tolerant platform for managing, accessing, and analyzing very large datasets. Hadoop and its underlying HDFS distributed file system have been proven to scale up to 2,000 nodes in a data management scale-out cluster, and beyond in a range of leading Web 2.0 and cloud computing environments such as Yahoo, Facebook, Amazon and EBay.

HDFS is designed to be fault-resilient and self-repairing with minimal or no operator intervention for node failover. A fundamental assumption of Hadoop is that it is more efficient to perform computations on nodes that locally store the data involved rather than move the data to an arbitrary set of compute clients over a network. This premise is a major architectural driver of HDFS, and it has resulted in very good Hadoop performance and linear price-performance scalability using commodity scale-out server clusters.

Hadoop is an open source project with a broad-based, dynamic developer and user community. Hadoop benefits from this worldwide network, rapidly advancing in capability and software quality, several steps ahead of competing Data Management paradigms and many times faster than legacy solutions.

Hadoop Distributed File System

HDFS is a highly scalable, distributed file system that provides a single, global namespace for the entire Hadoop cluster. HDFS comprises DataNodes supporting direct-attached storage (DAS), which store data in large 64 or 128 MB "chunks" to take advantage of sequential I/O capabilities of disks and minimize latencies resulting from random seeks. HDFS provides data redundancy through block replication. The default replication factor is three, which translates to three complete copies of the data being available at all times.

The HDFS *NameNode* is at the center of HDFS. It manages the file system by maintaining a file metadata image that includes file name, location, and replication state. The NameNode has the ability to detect failed DataNodes and replicate data blocks on those nodes to surviving DataNodes.

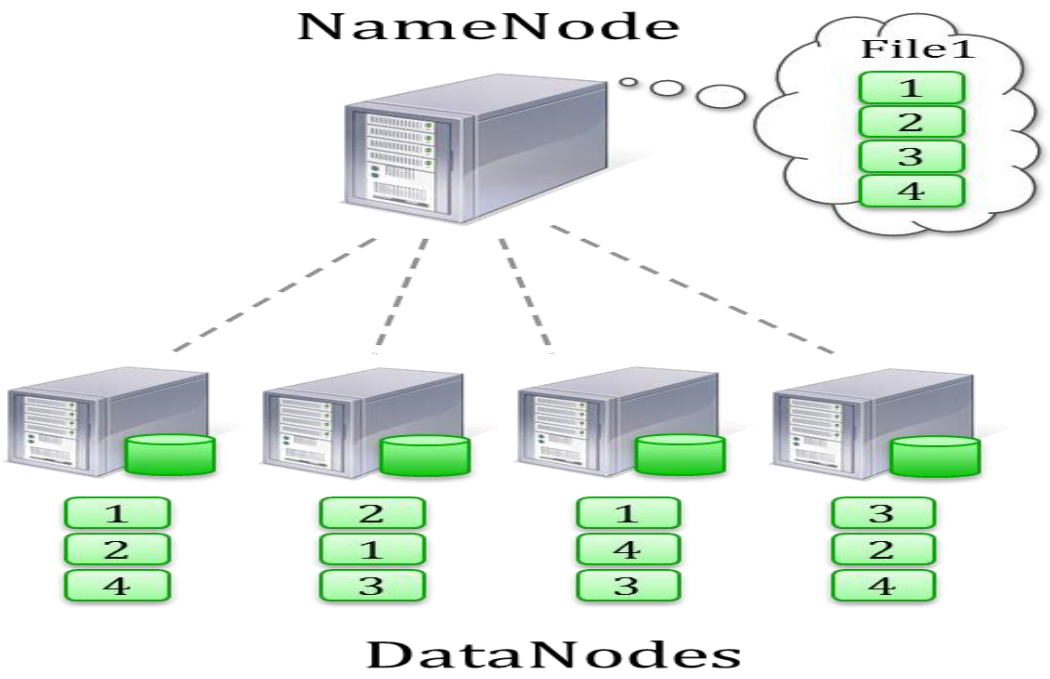

Figure 1. The HDFS Architecture

MapReduce

The Hadoop MapReduce algorithm was originally developed by Google for large-scale data analysis and provides the foundation of Hadoop's capabilities for processing very large datasets very quickly. At a top-level, MapReduce provides a method for allocating a job across the servers in a Hadoop cluster.

At a top-level, the Map function involves reading the raw data and preparing it for the Reduce phase. The Map phase output consists of sets of key-value pairs that are determined by the MapReduce program and sorted by key. For example, if the input is a text file, a key pair might consist of the record offset in the data file (similar to a row ID number) and the text in the row. The reduce function receives

map output from all the servers in the cluster, performs a shuffle and sort operation, and combines that output to complete the job. Output is written to the HDFS.

The Map and Reduce functions are run in parallel across the servers in the cluster, and they are managed by the JobTracker, which receives the request to run the MapReduce program. It then divides the job up into smaller Map and Reduce tasks. These tasks are assigned to different servers in the cluster, based on the location of the data required by the tasks, and are executed in parallel. The JobTracker monitors the progress of all the tasks. The results are written to an output file in HDFS (or a set of output files) when all tasks are complete.

The Hadoop model of assigning tasks to servers that already have the relevant input data storage in their DAS storage or "pushing computation to storage" is a critical aspect of the Hadoop value proposition for quickly processing large datasets. In comparison, legacy approaches that require loading-in data into compute servers (which can take days for petabyte-level datasets) are infeasible for big data applications.

Hadoop Network I/O Challenges

Because Hadoop deployments can have very large infrastructure requirements, hardware and software choices made at design time have a significant impact on performance and return on investment (ROI) of Hadoop clusters. Specifically, Hadoop cluster performance and ROI are highly dependent on network architecture and technology choices. While Gigabit Ethernet is the most commonly deployed network, it provides less than ideal bandwidth for many Hadoop workloads and for the whole of range of I/O-bound operations comprising a Hadoop job. Gigabit Ethernet is also incompatible with the latest advances in processor and storage I/O technologies.

Hadoop I/O Bound Operations

The initial Hadoop data loading operation has a direct impact on Hadoop cluster scalability and availability. As discussed earlier, Hadoop input data is stored in multiple files with each file stored on a separate DataNode. The gauge of parallelism in a Hadoop job is directly related to the number of input files.

Each file is triple replicated from server-to-server upon input for protection against node failures. Therefore, the ability to rapidly and reliably distribute large data sets across cluster servers is directly related to the speed of the network infrastructure used to distribute that data and enable replication.

Between the Map and Reduce operations, the data is 'shuffled' between cluster servers. This involves moving all outputs of the map operation with the same key to the same reducer server. At this point, the data network is again the critical path. End-to-end performance and latency of the network interconnect directly impact the shuffle phase of a data set reduction.

Following the Map/Reduce functions, data is combined for output and reporting. This requires another network operation as the outputs from each "reducer" node are moved onto a single reporting node.

Again, network interconnect performance can directly impact the performance of the Hadoop cluster, especially if the Hadoop cluster is running multiple data reductions.

Workflow Phase	Network I/O Challenge
Initial data loading	Speed of data distribution to cluster nodes
Data Replication	Initial and on-going data replication performance
Shuffle Phase	Performance of data movement to Reducer server(s)
Combine Phase	Performance of data movement from Reducer server(s) to Reporting node

Table 1: Hadoop Network I/O Challenges

CPU and Storage I/O Advancements

In nearly all cases, a Hadoop job will encounter bottlenecks reading data from disk or from the network (I/O-bound) as well as processing data (CPU-bound). Most teams looking to build a Hadoop cluster don't yet know the profile of their workload and often the first jobs that an organization runs with Hadoop are far different than the jobs that Hadoop is used for as they become proficient with it. For these reasons it makes sense to invest in a Hadoop cluster that is balanced in CPU, network and disk I/O performance when the team is unfamiliar with the types of jobs they are going to run.

Typical Hadoop scale-out cluster servers utilize TCP/IP networking over one or more Gigabit Ethernet network interface cards (NICs) connected to a Gigabit Ethernet network. However, the latest generation of commodity servers offers multi-socket, multicore CPU technology like Intel's Nehalem, which outstrips the network capacity offered by GbE networks. With advances in processor technology, this mismatch between server and network performance is expected to grow. Similarly, while Solid State Disks (SSDs) are evolving to offer equivalent capacity-per-dollar of Hard Disk Drives (HDDs), they are seeing rapid adoption for caching and for use with medium-sized datasets. The advances in storage I/O performance offered by SSDs surpasses the performance offered by GbE networking, which makes network I/O increasingly the most common impediment to improved Hadoop cluster performance.

10 Gigabit Ethernet has the potential of bringing the Hadoop cluster networking into balance with recent performance improvements created by server CPUs and advances in storage technology. Balancing the performance of server network I/O improves the efficiency of every server in the cluster, thus improving the performance of the entire Hadoop infrastructure by removing critical bottlenecks.

However, to achieve optimum balance, the network I/O gains delivered by 10GbE must be accompanied by optimal efficiency so that the impact of high-speed network I/O on the server CPU is minimized.

In addition to achieving higher bandwidth, enterprises can gain a significant performance boost for Hadoop clusters by utilizing 10GbE networking technologies such as TCP Offload Engines (TOE) and

Remote Direct Memory Access (RDMA). These technologies cut the CPU cycles needed to run the communication tasks, resulting in improved efficiency and faster job execution. This makes the choice of the specific 10GbE server networking option for Hadoop cluster servers a strategically logical one.

Chelsio Terminator 4 (T4) 10Gbe Unified Wire Adapters: Proven for High-Performance Cluster Networking

The Chelsio T420 10GbE Unified Wired Adapter utilizes Chelsio's Terminator 4 (T4), a highly integrated 10GbE ASIC chip built around a programmable protocol processing engine. The T4 ASIC represents Chelsio's fourth generation TCP offload engine (TOE) design, third generation iSCSI design, and second generation iWARP (RDMA) implementation. In addition to full TCP and iSCSI offload, T4 supports full FCoE offload. The T4 data flow processing architecture also provides for wire-speed operation at small to large packet sizes regardless of the number of TCP connections. The T4 demonstrates better performance across a range of benchmarks than that for Terminator 3 (T3) while running the same microcode that has been field-proven in very large clusters.

For the server connection, the T4 ASIC includes a PCI Express v2.0 ×8 host interface. With support for the 5Gbps Gen2 data rate, the PCIe interface provides up to 32Gbps of bandwidth to the server. On the network side, T4 integrates four Ethernet ports that support GbE as well as 10GbE operation.

Chelsio's 10GbE adapters improve network performance by leveraging an embedded TOE, which offloads TCP/IP stack processing to the server NIC. Used primarily with high-speed interfaces such as 10GbE, the TOE frees up memory bandwidth and valuable CPU cycles on the server, delivering the high throughput and low latency needed for cluster networking applications, while leveraging Ethernet's ubiquity, scalability, and cost-effectiveness.

Chelsio 10GbE TOE is aimed at increasing the ability to move data into (and out of) a server in multi-core processing server environments. TOE has proven to be ideal for environments where servers are concurrently required to run CPU as well as network I/O-intensive tasks such as high-performance computing (HPC) cluster environments, market data environments as well as for big data applications such as Hadoop clusters networking.

A key benefit of the Chelsio 10GbE TOE implementation is that it maintains full socket streaming semantics, enabling applications, such as HDFS, using the sockets programming model to leverage its performance capabilities unmodified.

RDMA/TCP (iWARP): Cost-Effective, High-Performance Cluster Networking

The Chelsio T4 second generation iWARP design builds on the RDMA capabilities of T3 while leveraging the embedded TOE capabilities, with advanced techniques to reduce CPU overhead, memory bandwidth utilization, and latency by a combination of offloading TCP/IP processing from the CPU, eliminating unnecessary buffering, and dramatically reducing expensive operating system calls and context switches – thereby moving data management and network protocol processing to the T420 10GbE Unified Wire Adapter. The OpenFabrics software stack that is fully integrated into the flavors of Linux distributed by Novell and Red Hat fully supports 10GbE iWARP.

RDMA and DCB – Selecting the Optimum Path

The IEEE has been developing standards collectively referred to as "Data Center Bridging" (DCB) as an addition to the 802.1 standard. It provides enhancements to standard Ethernet in the form of congestion notification, priority flow control (PFC), enhanced transmission selection (ETS), and discovery and capability exchange (DCBX). DCB is required to support the new Fibre Channel over Ethernet (FCoE) protocol as well as is deemed necessary as different networks converge over a single physical infrastructure.

When DCB is fully deployed, iWARP is expected to leverage all the same features to provide enhanced RDMA performance over DCB. Another proposed path for RDMA networking over DCB is the RDMA over CEE or RoCE protocol being promoted by a lone InfiniBand vendor. While RoCE may be a viable solution in an idealized DCB capable environment, it needs to overcome a range of challenges before it can be feasibly deployed as cluster network interconnect in the datacenter:

- RoCE requires a complete upgrade to a new switch platforms that supports DCB/DCB
- RoCE locks the user into a first generation, proprietary technology provided by a lone InfiniBand vendor
- Scalability of RoCE is unproven and it remains to be seen how it fares in all-to-all large deployments characteristic of cluster networking applications
- RoCE will continue to need gateways and IT management personnel to be able to achieve a tuned solution – these extra expenses diminish Ethernet's TCO benefit

Going down the established route of iWARP (with or without DCB) has numerous advantages. iWARP allows the use of legacy switches, is an established IETF standard, and is field proven. Numerous RNIC vendors are on their third and fourth generation technology (Chelsio, Intel) with others supporting the standard (Broadcom and QLogic). In addition, Chelsio is delivering a full unified-wire (CNA) solution that includes iSCSI, TOE, Full-HBA FCoE, VEB, and iWARP. This allows iWARP to be the lowest cost, highest performing solution, while providing a wide variety of vendor and protocol choices.

Chelsio 10GbE TOE Performance for Hadoop[1]

This section details the performance advantages of Chelsio 10GbE TOE solutions compared to GbE for Hadoop cluster environments using hard disk drive (HDDs) or solid-state disks (SSDs) as storage devices. The Chelsio 10GbE TOE was found to lead to network and storage I/O balance in Hadoop cluster nodes and clearly complements the use of SSDs in dramatically improving HDFS sequential and random write performance, and reducing job execution times.

Sequential Write Performance: The DFSIO file system benchmark (a part of the Hadoop distribution)

[1] The section is based on benchmarking of Chelsio T3 TOE performance is for Hadoop Distributed File System (HDFS) available at: http://nowlab.cse.ohio-state.edu/publications/conf-papers/2010/sur-masv-dc10.pdf. Benchmarking of Chelsio T4 Unified Wire is underway and will be published in an updated version of this paper.

was used to characterize the sequential I/O performance of HDFS. For Hadoop cluster configurations using hard disk drives, the study found that using GbE networking, HDFS was on average network I/O bound. The use of Chelsio 10GbE TOE shifted the bottleneck to storage I/O. In spite of the system being storage I/O bound, the Chelsio 10GbE TOE delivered up to 100% sequential performance gains over GbE. The use of SSDs in HDFS alleviates the storage I/O bottlenecks found using HDD testing. In this case, the Chelsio 10Gbe TOE exhibited 250%+ sequential write performance gains over GbE.

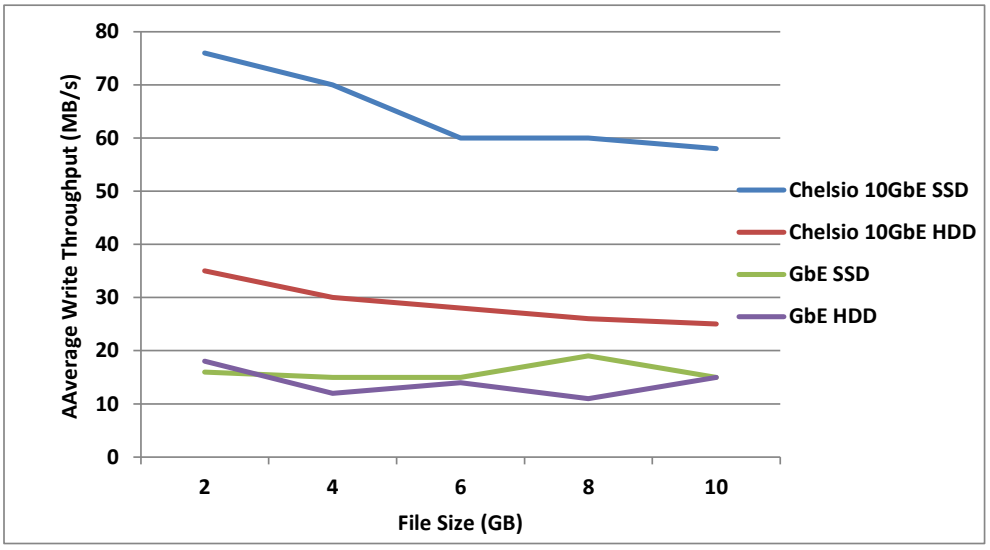

Figure 2. HDFS Sequential Write Performance

Random Write Performance: The Hadoop RandomWriter benchmark was used to measure HDFS random read/write performance using simulated MapReduce jobs for both HDDs and SSDs. The storage I/O bottleneck limited the gains in execution time for Chelsio 10GbE TOE to 30% versus GbE, while for the SSD case, the Chelsio 10Gbe TOE exhibited up to 300% reduction in execution time.

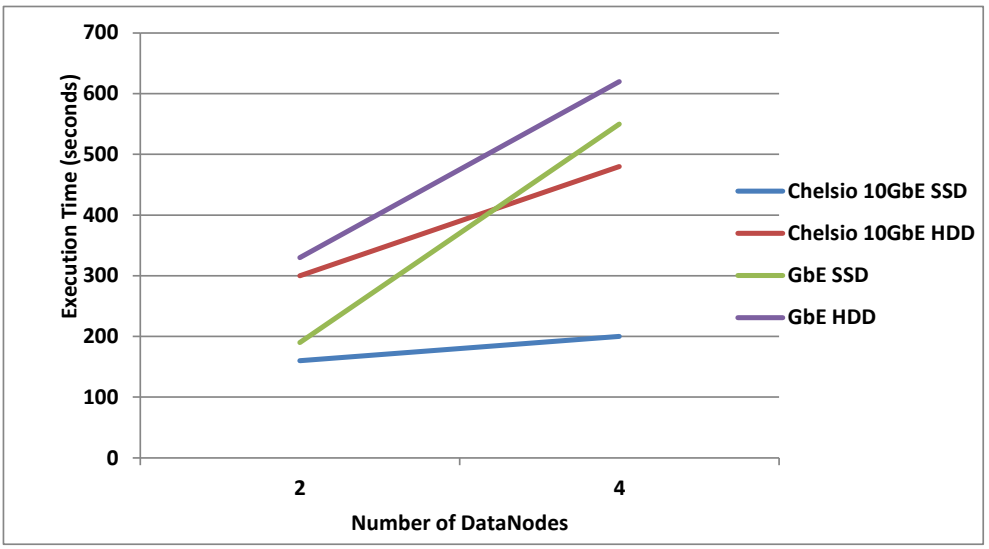

Figure 2. HDFS Random Write Performance

Conclusion

10 Gigabit Ethernet has the potential of bringing the Hadoop cluster networking into balance with the recent improvements in performance brought by server CPUs and advances in storage technology. However, to achieve optimum balance, the network I/O gains delivered by 10GbE must come with optimal efficiency so that the impact of high-speed network I/O on the server CPU is minimized. The Chelsio T420 10GbE Unified Wired Adapter utilizes Chelsio's Terminator 4 (T4), a highly integrated 10GbE ASIC that represents Chelsio's fourth generation TCP offload engine (TOE) design and second generation iWARP (RDMA) implementation. The T4 demonstrates better performance across a range of benchmarks than that for Terminator 3 (T3) while running the same microcode that has been field-proven in very large clusters.

Chelsio 10GbE TOE is aimed at increasing the ability to move data into (and out of) a server in multi-core processing server environments and has proven to be ideal for environments where servers are concurrently required to run CPU as well as network I/O-intensive tasks such as Hadoop cluster networking. The Chelsio T4 second generation iWARP design builds on the RDMA capabilities of T3 while leveraging the embedded TOE capabilities, to move data management and network protocol processing to the T420 10GbE Unified Wire Adapter.

Chelsio 10GbE solutions deliver a range of performance advantages over GbE networking for Hadoop cluster environments. Independent testing determined that Chelsio 10GbE adapters lead to network and storage I/O balance in Hadoop cluster nodes and clearly complement the use of SSDs in dramatically improving HDFS sequential and random write performance, and reducing job execution times.