10Gb Ethernet with TCP Offload – The High Performance Interconnect

Independent Performance Benchmarks of 10 Gbps Ethernet with TOE compared to Infiniband and Myrinet

A Chelsio Communications White Paper
Abstract
This paper presents the results of three independent performance studies involving Chelsio's Terminator TCP Offload Engine. The first study compares 10GigE with TOE, Infiniband, and Myrinet in an identical experimental setup, using micro-benchmark as well as application-level evaluation. The results show that 10GigE with TOE delivers performance that is equal to or better than the exotic interconnects in a computing cluster network environment, and that 10GigE is particularly well-suited for sockets-based applications, yielding performance up to an order of magnitude higher than the competing technologies. The second study performed similar independent measurements, confirming these result and, in addition, demonstrating that basic 10 Gbps Ethernet without TOE may fall short of specialized interconnect performance. The third study focuses on 10 Gbps Ethernet, and by using both micro-benchmarks and application level benchmarks shows that Chelsio's TOE benefits translate into significant latency and bandwidth performance improvements compared to basic NICs, including an order of magnitude increase in Apache Web server performance.

Ethernet TOE vs. Infiniband and Myrinet
This section contains the abstract of the following paper, along with illustrative graphs:


Despite the performance drawbacks of Ethernet, it still possesses a sizable footprint in cluster computing because of its low cost and backward compatibility to existing Ethernet infrastructure. In this paper, we demonstrate that these performance drawbacks can be reduced (and in some cases, arguably eliminated) by coupling TCP offload engines (TOEs) with 10-Gigabit Ethernet (10GigE). Although there exists significant research on individual network technologies such as 10GigE, InfiniBand (IBA), and Myrinet; to the best of our knowledge, there has been no work that compares the capabilities and limitations of these technologies with the recently introduced 10GigE TOEs in a homogeneous experimental testbed. Therefore, we present performance evaluations across 10GigE, IBA, and Myrinet (with identical cluster-compute nodes) in order to enable a coherent comparison with respect to the sockets interface. Specifically, we evaluate the network technologies at two levels: (i) a detailed micro-benchmark evaluation and (ii) an application-level evaluation with sample applications from different domains, including a bio-medical image visualization tool known as the Virtual Microscope, an iso-surface oil reservoir simulator, a cluster file-system known as the Parallel Virtual File-System (PVFS), and a popular cluster management tool known as Ganglia. In addition to 10GigE’s advantage with respect to compatibility to wide-area network infrastructures, e.g., in support of grids, our results show that 10GigE also delivers performance that is comparable to traditional high-speed network technologies such as IBA and Myrinet in a system-area network environment to support clusters and that 10GigE is particularly well suited for sockets-based applications.

The following figures along with the accompanying explanation are extracted from the paper and provide illustrative results for real application workloads.

“Performance of MPI-Tile-IO: MPI-Tile-IO [...] is a tile reading MPI-IO application. It tests the performance of tiled access to a two-dimensional dense dataset, simulating the type of workload that exists in some visualization applications and numerical applications.”
“We evaluate both the read and write performance of MPI-Tile-IO over PVFS. As shown in Figure 11, the 10GigE TOE provides considerably better performance than the other two networks in terms of both read and write bandwidth.”

![MPI-tile-IO over PVFS](image)

**Figure 11 MPI-Tile-IO over PVFS**

“Ganglia [...] is an open-source project that grew out of the UCBerkeley Millennium Project. It is a scalable distributed monitoring system for high-performance computing systems such as clusters and grids.”

“This figure shows the performance of Ganglia for the different networks. As shown in the figure, the 10GigE TOE considerably outperforms the other two networks by up to a factor of 11 in some cases. To understand this performance difference, we first describe the pattern in which Ganglia works. The client node is an end node which gathers all the information about all the servers in the cluster and displays it to the end user. In order to collect this information, the client opens a connection with each node in the cluster and obtains the relevant information (ranging from 2 KB to 10 KB) from the nodes. Thus, Ganglia is quite sensitive to the connection time and medium message latency. As we had seen in Figures 3a and 3b, 10GigE TOE and SDP/GM/Myrinet do not perform very well for medium-sized messages. However, the connection time for 10GigE is only about 60µs as compared to the millisecond range connection times for SDP/GM/Myrinet and SDP/IBA. During connection setup, SDP/GM/Myrinet and SDP/IBA pre-register a set of buffers in order to carry out the required communication; this operation is quite expensive for the Myrinet and IBA networks since it involves informing the network adapters about each of these buffers and the corresponding protection information.”
The full paper can be accessed at:


**Ethernet TOE vs. Ethernet NIC, Infiniband and Myrinet**

This section contains the summary of the following presentation:


- 10 GbE and TOE out performed IB and SDP for socket applications in our test environment.
- Protocol offload, TOE and SDP, offered significant performance improvement
- Further improvement possible with RDMA and zero-copy

The following graph shows the performance of the four different adapters in the IOzone distributed file system application, demonstrating that Ethernet without TOE falls short of specialized inter-connect performance, while TOE lifts it up to pole position.
Though traditional Ethernet based network architectures such as Gigabit Ethernet have suffered from a huge performance difference as compared to other high performance networks (e.g., InfiniBand, Quadrics, Myrinet), Ethernet has continued to be the most widely used network architecture today. This trend is mainly attributed to the low cost of the network components and their backward compatibility with the existing Ethernet infrastructure. With the advent of 10-Gigabit Ethernet and TCP Offload Engines (TOEs), whether this performance gap be bridged is an open question. In this paper, we present a detailed performance evaluation of the Chelsio T110 10-Gigabit Ethernet adapter with TOE. We have done performance evaluations in three broad categories: (i) detailed micro-benchmark performance evaluation at the sockets layer, (ii) performance evaluation of the Message Passing Interface (MPI) stack atop the sockets interface, and (iii) application-level evaluations using the Apache web server. Our experimental results demonstrate latency as low as 8.9 µs and throughput of nearly 7.6 Gbps for these adapters. Further, we see an order-of-magnitude improvement in the performance of the Apache web server while utilizing the TOE as compared to the basic 10-Gigabit Ethernet adapter without TOE.

The following graph shows the Transactions per Second (TPS) performance of the popular Apache Web server using TOE vs. non-TOE. The various bars correspond to different Zipf content popularity distributions, demonstrating 60 to 1000% higher capacity with TOE.
Conclusion

The conclusion of the results compiled here is that the ubiquitous Ethernet technology with TCP/IP offload can now penetrate the high performance computing market and unify the networks by replacing exotic interconnects. The advantages of such unification are numerous, both in connectivity, support and management costs.

For more information about Chelsio Communications and the Terminator architecture, visit the Chelsio web site at www.chelsio.com or send an e-mail to info@chelsio.com.

The full paper can be accessed at: