

# iWARP – Robust, Proven Low Latency Ethernet Clustering

---

## Executive Summary

iWARP (RDMA over Ethernet) is a robust, proven low latency solution for high performance computing, backed by many of the leading 10Gb Ethernet adapter companies. Chelsio's low-latency 10Gb Ethernet solution features full protocol-offload technology, exceeds InfiniBand in application performance, and delivers the highest throughput and low latency needed for high-performance computing applications, along with Ethernet's ubiquity, scalability and cost effectiveness.

## iWARP - The Technology

iWARP (Internet Wide Area RDMA Protocol) can also be called RDMA over Ethernet (or RDMA over CEE (Converged Enhanced Ethernet) depending on whether the Ethernet infrastructure is CEE capable). It is a low-latency Ethernet solution for high performance computing using the RDMA (Remote Direct Memory Access) protocol over TCP/IP, and is standardized through the IETF (Internet Engineering Task Force).

iWARP frees up memory bandwidth and valuable CPU cycles to allow focused processing of HPC applications, by taking advantage of an embedded TCP offload Engine (TOE) and RDMA capabilities for direct application access and kernel bypass. With a number of fabrics converging towards Ethernet as a unified wire, iWARP stands out in providing the lowest cost clustering solution by allowing users to leverage their installed Ethernet networks in the datacenter.

iWARP support is readily available. Users today can choose between two proven software stack solutions: OFED (Open Fabrics Enterprise Distribution) and Microsoft's Network Direct API.

## Ethernet and CEE – Choosing the Right Path

Recently, there has been much confusion surrounding the new Data Center Bridging addition to the IEEE 802.1 standard (DCB – also called DCE, or CEE). This new specification has been deemed necessary as different networks converge over a single physical infrastructure. 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 necessary to support the new Fibre Channel over Ethernet (FCoE) protocol.

While the DCB/CEE specification seems to hold promise, it is not yet ratified by the IEEE and will probably not be until sometime in 2011. When DCB is finally ratified and deployed, iWARP is expected to leverage all the same features to provide enhanced RDMA performance.

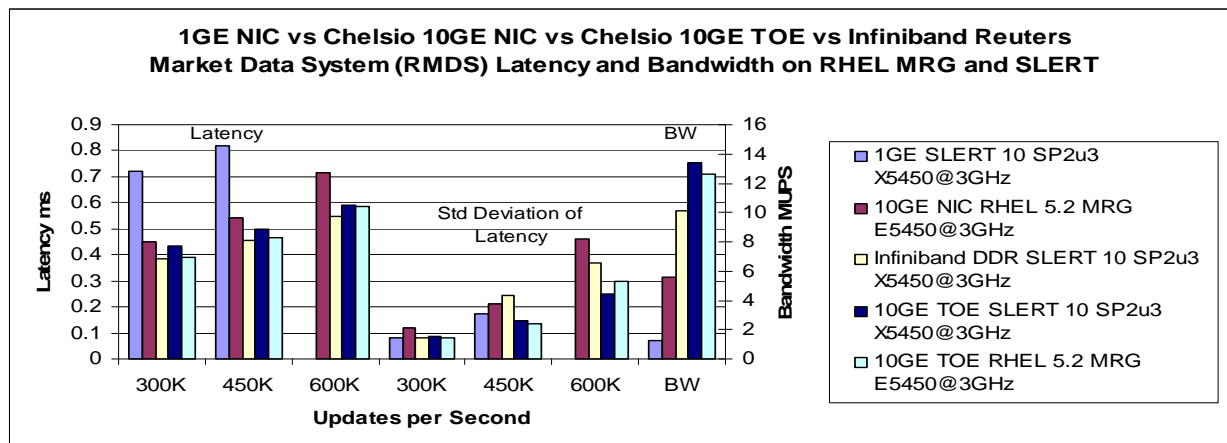
There are two paths for utilizing DCB. One path is called RoCE (simple RDMA over CEE), which is being promoted by a single InfiniBand vendor. While RoCE may be a viable solution in an idealized DCB capable environment, it needs to overcome some very unique technical challenges before it can be deployed in any datacenter.

1. RoCE requires a complete upgrade to a switch platform that supports DCB/CEE (CEE switch ports are currently very expensive).
2. RoCE locks the user into a first generation, proprietary technology provided by only one InfiniBand vendor.
3. Scalability aspects of RoCE are unproven and it remains to be seen how it fares in all-to-all large deployments characteristic of HPC applications. Of particular concern is its sensitivity to congestion, given that the congestion notification part of the DCB suite is optional, unverified and has only been characterized in limited simulation scenarios.
4. RoCE is a Layer-2 protocol that cannot address some large commercial clusters where Layer 3 capabilities are a must-have requirement.
5. RoCE relegates the reliability and congestion handling to software, resulting in poor performance in case of network imperfections and packet drops, and reduced performance benefits. This jeopardizes the whole value proposition and purpose of RDMA.
6. RoCE will continue to need gateways and IT management personnel to be able to achieve a tuned solution – the extra expense hence diminishes Ethernet's TCO benefit.

Going down the other (much less rocky) path of iWARP over DCB (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) bundle that includes iSCSI, TOE, Full-HBA FCoE, VEB, & iWARP. This allows iWARP to be the lowest cost, highest performing solution while proving a variety of vendor and protocol choices. iWARP has several additional benefits:

1. iWARP has proof points and case studies for large clusters.
2. iWARP is multi-vendor.
3. The 30-year old TCP protocol remains the only surefire reliable protocol – and when implemented in silicon as in a TOE or RNIC, provides zero overhead and higher performance.
4. iWARP is built on top of IP, is therefore routable, and hence avoids problems associated with large flat networks, similar to what has been seen in in very large commercial deployments.
5. iWARP is built on top of TCP, and is therefore reliable without requiring a costly onerous lossless network. TCP implementation in silicon (TOE) will shield the application from the CPU cycles and eliminates context swaps necessary to address packet drops or retransmissions, resulting in consistent application performance.
6. iWARP provides all the TCO benefits of Ethernet.

## Performance Benchmarks



The Securities Technology Analysis Center (STAC) measured the performance of the Chelsio NIC using the Reuters Market Data System (RMDS6) benchmark suite, and the results highlighted that Chelsio's 10G adapters provided the **lowest mean latency** and the **lowest standard deviation of latency** ever reported with RMDS using 10G Ethernet, 1G Ethernet or InfiniBand. The complete benchmark results are available at [www.chelsio.com](http://www.chelsio.com).

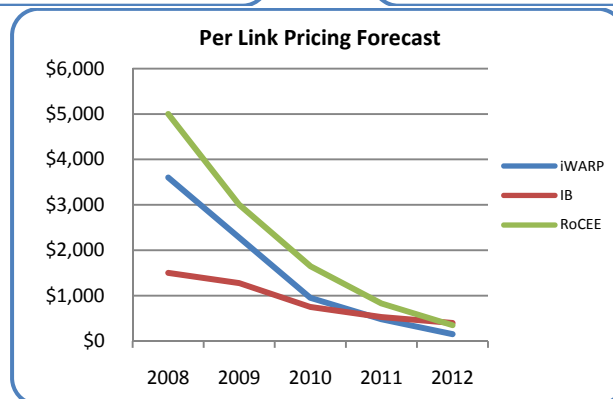
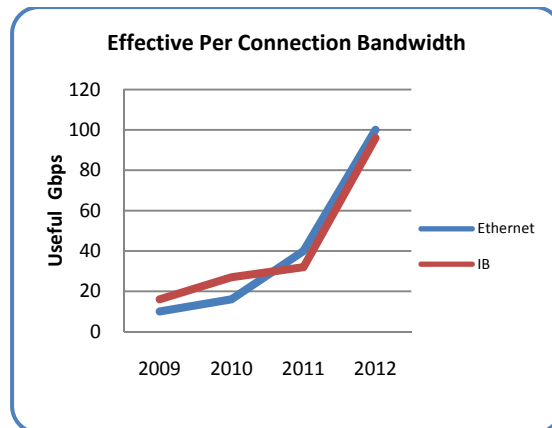
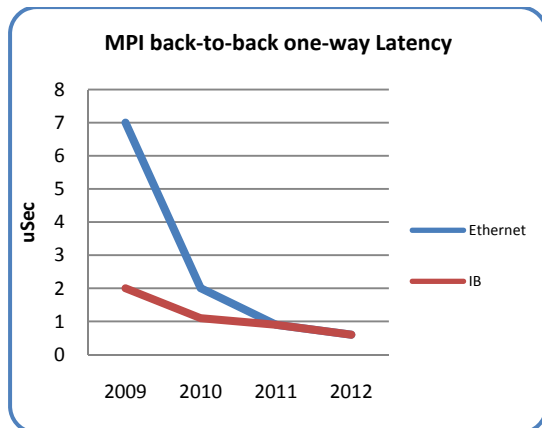
## Performance Roadmap

Benchmarks of state of the art iWARP show a 6usec latency and line rate 10Gb throughput. One must note that benchmarks are only a loose indication of application performance. Based on independent testing by HP & IBM, this latency results in application performance that is at par with DDR-IB (<http://www.chelsio.com/whitepapers.html>).

Chelsio's next generation iWARP enabled adapters will reduce the benchmark latency to 2us in 2010 and to less than 1us in 2011. It is expected therefore that applications using Chelsio's next generation iWARP technology show similar performance to QDR-IB, while benefiting from all other aspects of a Unified Wire (iSCSI, FCoE, TOE, Virtualization, etc...). Chelsio's next generation iWARP release will further enhance the iWARP protocol, closing any gaps with IB.

In addition, iWARP-based solutions have a much shorter path to a LOM solution since they concurrently provide the iSCSI/TOE solutions.

The performance roadmaps of iWARP and IB and a projected pricing comparison are shown in the tables on page 4.



## Summary

iWARP has numerous business and technical advantages over InfiniBand and RoCE:

- **Verified Interoperability:** iWARP solutions go through periodic interoperability tests before they are certified.
- **Multi-Vendor Support:** iWARP is an IETF standard, supported by Intel, Chelsio, and several other companies. Numerous middleware companies now support iWARP.
- **High Performance:** iWARP is projected to provide similar application performance to QDR-IB.
- **Proven:** iWARP has been shown to scale to large node count clusters. It is built on top of TCP which makes it ubiquitous and deployable, everywhere.
- **Low Cost:** iWARP does not require infrastructure/switch upgrades, gateways, or IT personnel training, and it is an additional protocol capability for silicon that will be on the LOM anyway for other reasons, benefiting from a TOE that will be on the LOM for other reasons.
- **Existing Ethernet Infrastructure:** Customers requiring high performance for their applications, such as financial transactions, can maximize their current data center investment.
- **Easy Migration:** It is much easier to convert the existing 40% of the HPC market that is on GbE to 10GbE using iWARP, and to higher speeds once 40Gb and 100Gb Ethernet are available.
- **Supported on Linux:** Chelsio's iWARP is inbox'ed with all distributions and in the upstream kernel.
- **Supported on Windows:** iWARP for Windows under Network Direct is available from Chelsio today (WHQL'ed).

For information on accelerating your cluster with Chelsio's 10G Ethernet, please visit [www.chelsio.com](http://www.chelsio.com).