Mellanox TECHNOLOGIES

WHITEPAPER

March 2009

The Case for Low-Latency Ethernet

The Evolutionary Step for IPC Consolidation over 10 Gigabit Ethernet

1.1 Introduction

The industry momentum behind Fibre over Ethernet (FCoE) sets some significant precedence that raises questions about what is the best approach for server to server messaging (or inter process communication or IPC) using zero-copy send/receive and remote DMA (RDMA) technologies over Ethernet. There are two competing technologies for IPC – InfiniBand and iWARP (based on 10GigE). Applying the same business and technical logic associated with SAN consolidation using FCoE, one would conclude that Low Latency Ethernet (LLE) makes the most sense. Here is why.

1.1 FCoE Traction for SAN Consolidation – The Business Perspective

If Ethernet is the technology for server I/O unification because it is most ubiquitous on the server, one cannot assume it is TCP, UDP and IP all the way, end-to-end in the data center. If it was so, iSCSI with end-to-end Ethernet, from the server to the storage box would have swept the world. But it has not and the reason is there are huge Fibre Channel storage and software investments one just cannot ignore. Hence the emergence and momentum behind Fibre Channel over Ethernet (FCoE) that enables IT managers to unify the I/O on servers to 10GigE and yet maintain seamless connectivity to their Fibre Channel storage equipment while maintaining their Fibre Channel software investments.

1.2 FCoE Traction for SAN Consolidation – The Technical Perspective

Let's look at the technical reasons, compare iSCSI versus FCoE. iSCSI is based on IP networks designed for the LAN and the Internet, relies on TCP to address the issues with traditional Ethernet lossy networks. The reliance on TCP for recovery and flow control results in many overheads, some show up as slow reaction to resolving congestions in the network (as TCP is in the software and depends on available shared CPU cycles) and some show up as expensive, power hungry, non-scalable I/O adapters (with TCP Offload Engines or TOE). Couple that with the Linux communities' adverse reactions to TOE and the lack of TOE value in the virtualized server environments (VMware ESX, Citrix XenServer and other technologies cannot make effective use of TOE), TOE is shrinking to very narrow usage scenarios. That is not all.

iSCSI and TOE made an inherent assumption that Ethernet is not a reliable medium, that is, it is lossy. So, TCP could not be avoided. With the advent of reliable Ethernet, through support of Per Priority Pause, where lossless virtual lanes can be used by storage traffic passing over Ethernet, the importance of TCP in these storage applications further shrinks. Further enhancements to Ethernet being worked through various IEEE workgroups will enable congestion control and management using the layer 2 Ethernet medium, reducing the dependence on TCP further.

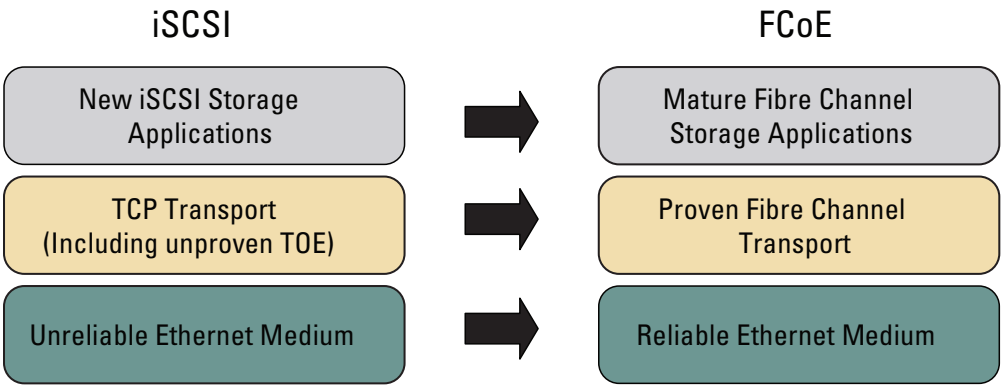


Figure 1: iSCSI versus FCoE

The above enhancements make it logical to apply the Fibre Channel transport over the reliable Ethernet medium, as in FCoE. By doing so, the following benefits become apparent, diminishing the viability of iSCSI:

FCoE preserves investment & skills

- Fibre Channel transport is proven to work for storage
- Less complexity, fewer unknowns
- No expensive, power hungry TOE
- Minimizes changes to OS stacks
- Preserves storage management skills, tools

FCoE is an evolutionary step for SAN consolidation over 10GigE - both from the business and technical perspectives.

1.3 What about LAN and IPC Consolidation

10GigE and LAN go hand in hand, so along with FCoE, that takes care of two critical traffic types in the data center – both FC SAN and Ethernet LAN can be unified over 10GigE adapters on servers. There is a third category of traffic type – the IPC traffic that helps clustered, grid and utility computing and is an ever growing component of service oriented infrastructures where Low-Latency between server nodes directly translates to doing more with fewer servers

(through higher efficiency) and delivering on the promise of "time is money" where every microsecond delay in executing a transaction can result in millions of dollars in losses (as in algorithmic trading, for example). iWARP hasn't been a viable 10GigE low-latency solution for IPC. Applications need to be modified to be iWARP compatible and is a complex, expensive and time consuming process, hence the slow gain in adoption. The latency improvement is still 3x-4x of LLE.

1.4 LLE Versus iWARP for IPC Consolidation – The Business Perspective

Just like there is huge investment and maturity in Fibre Channel technologies for SAN, there is similar investment and maturity in RDMA transport for IPC. If Ethernet is the technology for server I/O unification because it is most ubiquitous on the server, one cannot assume IPC consolidation has to be based

on TCP-based iWARP only. Just like FCoE encapsulates FC data (and therefore maintains the familiarity and maturity of SAN software, interfaces and management compatibility) in Ethernet frames, Low-Latency Ethernet or LLE encapsulates RDMA transport (and therefore maintains the familiarity and maturity of IPC software, interfaces

and management compatibility) in Ethernet frames. This can enable IT managers to unify the I/O on servers to 10GigE and yet maintain seamless interoperability to their IPC and clustering applications while maintaining their software investments (e.g., financial, clustered

database, commercial and academic, high performance computing applications).

Such applications already qualified using the OpenFabrics (www.openfabrics.org) IPC protocol

stack (also available in popular Linux and Windows distributions) over InfiniBand can now be seamlessly deployed over zero-copy send/receive and RDMA Ethernet using LLE. LLE support is available today as a single chip solution in Mellanox ConnectX adapters.

The business advantages of LLE for a data center are:

- No Changes to data center infrastructure
- I/O unification on a single wire over 10 GbE networks
- Continue with existing data center management infrastructure
- Reduction is power and cost savings
- maintain existing and future application compatibility
- significant CapEx and OpEx savings with a single chip solution for I/O unification

1.5 LLE for IPC Consolidation – The Technical Perspective

Let's look at the technical reasons, compare iWARP versus LLE. iWARP is based on IP networks designed for the LAN and the Internet, relies on TCP to address the issues with traditional

Ethernet lossy networks. The reliance on TCP for recovery and flow control results in many overheads, identical to the comparison of iSCSI to FCoE above.

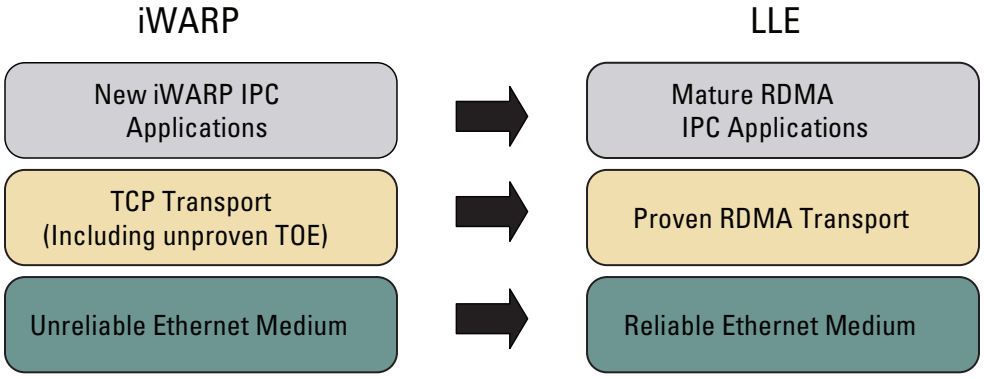


Figure 2: iWARP versus LLE

With the advent of reliable Ethernet, through support of Per Priority Pause, and congestion management and control using the layer 2 Ethernet medium, the proven and efficient RDMA transport becomes the perfect choice for deploying IPC over reliable Ethernet, just like the FC transport for deploying storage over reliable Ethernet. By doing so, the following benefits become apparent, diminishing the viability of iWARP using TOE.

LLE improves performance while preserving investment and skills

- Proven and mature RDMA transport for IPC/server-server communication

- Less complexity, fewer unknowns
- No expensive, power hungry TOE
- Minimizes changes to OS stacks
- Preserves IPC/Server management skills, tools

Like FCoE for SAN consolidation, LLE is an evolutionary step for IPC consolidation over 10GigE – both from the business and technical perspectives.

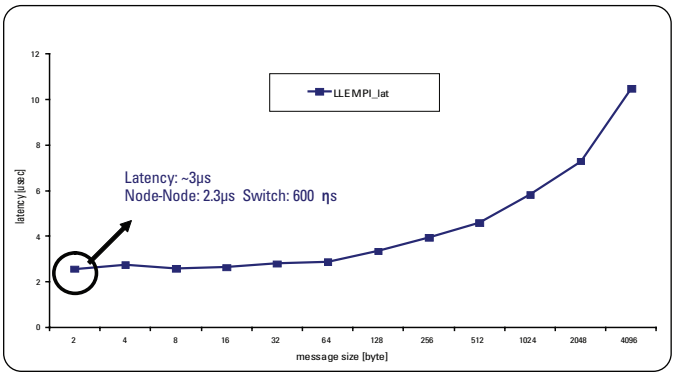
1.6 Advantages of LLE for IPC Clustering

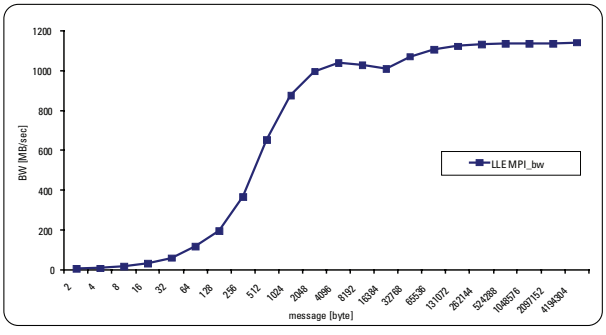
Up until now, we have discussed the reasons for consolidating IPC over 10 Gigabit Ethernet with a mature and efficient RDMA transport. We have compared the iWARP versus LLE with the iSCSI versus FCoE analogy to highlight the need for a better and efficient RDMA transport in today's data center. But for the user, what are the immediate tangible benefits?

LLE provides the most efficient IPC consolidation in the data center

- The lowest latency for IPC clustering (node-to-node: 2.3us)
- High bandwidth for all data center applications
- Low CPU utilization allowing LAN and SAN consolidation on a single adapter
- 3x-4x performance boost compared to the best iWARP solution
- Extends Ethernet pervasiveness with RDMA transport in 10GigE environments
- Mature RDMA applications with no changes to interface or application

Figures 3 & 4 give the performance results running MPI.





Conclusion

Low-Latency Ethernet, with a purpose built and proven RDMA transport, provides the most efficient and light-weight transport over Layer 2 (L2) Ethernet. It ensures interoperability on existing Ethernet infrastructure and takes advantage of virtual-links with per-priority pause support.

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100
Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com