April 2008

Mellanox TECHNOLOGIES

The Case for InfiniBand over Ethernet

The Evolutionary Step for IPC Consolidation over 10 Gigabit Ethernet

1.0 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). If one were to apply the same business and technical logic behind the initial success of FCoE, one would conclude that InfiniBand over Ethernet (IBoE) 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. Because 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 and maintain 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.

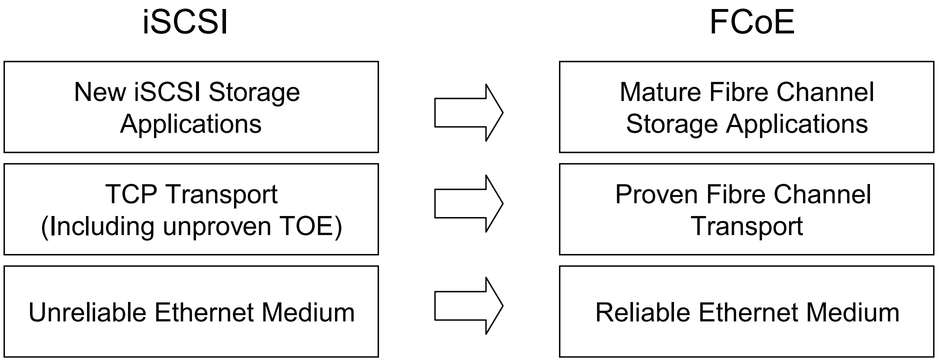
iSCSI
FCoE
New iSCSI Storage Applications
Mature Fibre Channel Storage Applications
TCP Transport (Including unproven TOE)
Proven Fibre Channel Transport
Unreliable Ethernet Medium
Reliable Ethernet Medium

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 mgmt 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 millisecond delay in executing a transaction can result in millions of dollars in losses (as in algorithmic trading, for example).

1.4 IBoE 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 InfiniBand 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 familiar and maturity of SAN software, interfaces and management compatibility) in Ethernet frames, InfiniBand over Ethernet or IBoE encapsulates IB data (and therefore maintains the familiar 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 and maintain 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 IBoE. IBoE support is available today in Mellanox ConnectX adapters.

1.5 IBoE for IPC Consolidation – The Technical Perspective

Let's look at the technical reasons, compare iWARP versus IBoE. 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.

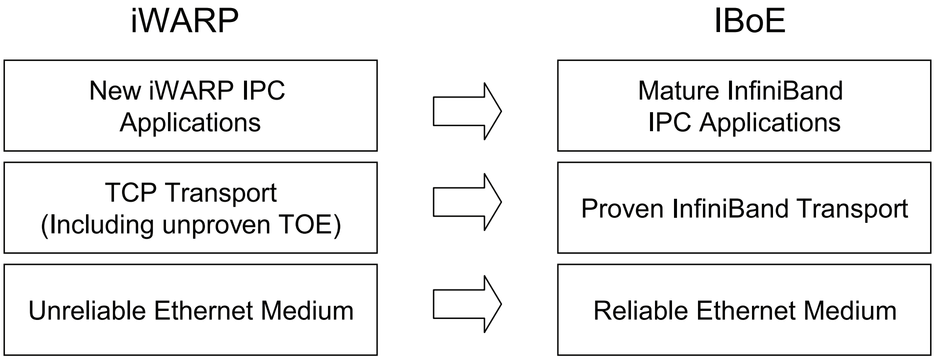
iWARP
IBoE
New iWARP IPC Applications
Mature InfiniBand IPC Applications
TCP Transport (Including unproven TOE)
Proven InfiniBand Transport
Unreliable Ethernet Medium
Reliable Ethernet Medium

Figure 2: iWARP versus IBoE

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 InfiniBand 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:

IBoE: preserves investment and skills

- Proven to work for IPC/server-server communication
- Less complexity, fewer unknowns
- No expensive, power hungry TOE
- Minimizes changes to OS stacks
- Preserves IPC/Server mgmt skills, tools

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