

Just as 10G Ethernet is going through widespread deployment in the datacenter, the discussion has now shifted to even higher speed interconnects—namely 40G and 100G Ethernet.

Background

In July 2006, the IEEE Higher Speed Study Group was formed to look into the next evolutionary step after 10 Gigabit Ethernet. In the past, Ethernet speeds would increase by a factor of 10. However, the next generation jump from 10 Gigabit to 100 Gigabit has proven to be a technological challenge. Some within the IEEE group felt that 100 Gig made sense for Telco's and other backbone network providers, but not as a next step for servers—it was simply more speed and expense than needed for the near future. While the IEEE initially planned to standardize only on 100G Ethernet as the next step after 10G, server vendors initiated a push in early 2007 to include 40G Ethernet in the standard, with the rationale that work used to develop 40G Ethernet will be able to be leveraged into 100G¹, and that 40G would be a better match for server I/O around 2010 than 100G.

In July 2007, the IEEE 802.3ba study group was named, and it is the first standard to include two different Ethernet speeds—the 40 Gbps rate for local server applications, and the 100 Gbps rate for internet backbone—to serve both market needs. In December 2007 the official 802.3ba task force was formed to begin work on the new standard, which is expected to be ratified by 2009² or 2010.

What are the drivers behind the push for 40G and 100G Ethernet?

Similar to the rationale behind 10G Ethernet, the drivers for higher Ethernet speeds of 40G and 100G will be the growth of bandwidth-intensive applications such as high-performance computing, business continuity, virtualization, video on demand, iSCSI, FCoE & NAS storage, video surveillance and voice over IP. Video-based applications in particular will continue to dominate network bandwidth needs.

Are the servers ready?

Currently most servers run a computer expansion interface known as PCIe (PCI Express). PCIe "gen 1" - which with its most common configuration of eight lanes can provide typically about 12.5 Gbps of bandwidth after overheads are subtracted³. We can measure this bandwidth using network interface cards with two 10 GigE ports.

The next generation, PCIe "gen2", doubles the bandwidth to the NIC to approximately 25 GT/s. PCIe "gen2" with its increased bandwidth is a large driver for those pushing for the 40G Ethernet to be included in the 802.3ba standard. "Gen 3" is still in discussion and not defined yet, but is intending to double the bandwidth again – to approx 50 Gbps. Gen3 uses 8GT/s and moves away from 8/10b encoding (announced in 2007).

¹ For example, four lanes of optics for 40G and 10 lanes for 100G would share some common components

² "Spec a year away for 40/100G Ethernet—IEEE group reports progress, but much work ahead," Rick Merritt, EE Times, May 22, 2008.

³ 8 lanes at 2.5G each = 20G. 8b/10b encoding reduces it to 16 Gbps. After overheads and line turn-around time are subtracted, it is reduced to about 12.5 (measured).

Why not just aggregate the links?

Today any speed in Ethernet higher than 10G Ethernet is achieved via link aggregations (LAGs). Link aggregations tend to be complex to configure, and can easily get out of balance and carry less than the promised bandwidth. While today it is possible to create a 40 Gigabit trunk group without 40G Ethernet, an actual 40G Ethernet device will be easier to incorporate in the network. It will be simpler to manage one 40G port and one cable versus four of either. Link aggregation at 10G should be considered a stopgap measure to be used when necessary but not ideal as a long-term solution to higher bandwidth needs in tomorrow's networking environments.

40G and 100G Ethernet vs. SONET

Today, some Telco gear already supports 40 Gig in Sonet OC-768. According to Communications Industry Researchers, 40 G and 100G Ethernet are expected to kill off use of SONET by 2016, and SONET will probably stop at OC-768.⁴

40G and 100G Ethernet vs. InfiniBand

InfiniBand is a low-latency, high-performance interconnect. It's bidirectional, and nearly always uses 4 lanes in each direction (InfiniBand vendors call this 4X). Early SDR (single data rate) products used 2.5 Gbps per lane to provide a 10Gbps connection. InfiniBand vendors describe bandwidth before the required 8b/10b encoding, so 10Gbps connections actually only provide 8 Gbps capacity. Subsequent DDR (double data rate) products use 5 Gbps per lane, providing 16 Gbps capacity, albeit at a reduced distance. Recently Voltaire and Mellanox announced 40G Infiniband (QDR) switch plans at the 2008 International Super Computing (ISC) Show in Dresden, so this should push the capacity to 32 Gbps starting in 2009, but at more reduced distances. The InfiniBand Trade Association also announced a roadmap to much higher speeds, however there is very limited activity behind this announcement.

InfiniBand can boast of very low latency and probably the fastest interconnect option for high-end clusters at the moment. However Infiniband could be considered a market risk since its silicon is currently being manufactured by only a single vendor: Mellanox. By comparison, higher-speed Ethernet will capitalize on the large installed base of Gigabit Ethernet. New 40G and 100G products will become less expensive and more available over time, and will be supported by many silicon and equipment vendors. The Ethernet standard also is universally understood by data center network administrators, so the relative costs for managing and troubleshooting Ethernet are much lower than for a niche fabric such as Infiniband.

Market for 40G and 100G

According to Communications Industry Researchers, 40G Ethernet is expected to garner larger market share than 100G, but as we've seen with 10G this market is difficult to predict. The CIR report projects the market for 40G Ethernet will be worth about \$3.1 billion worldwide by 2016, versus \$1.2 billion for 100G Ethernet.⁵

⁴ "The Path to 100Gbps Networks," Communications Industry Researchers, Inc., January 2008.

⁵ *ibid.*

BLADE Network Technologies' Position on 40G and 100G

We believe that both 40G Ethernet and 100G Ethernet will play important roles in the ongoing evolution of Ethernet. Used internally in a blade server, 40G Ethernet can be run without optics for internal connections, providing economical yet very fast connections for bandwidth-intensive applications such as video. Even today, four lanes at 10 Gbps per lane (utilizing the IEEE 802.3ap 10GBASE-KR standard) is already designed into some blade server midplanes, so we see 40G as being a great fit for blade servers.

Moreover, having 40G or even 100G Ethernet available as an inter-switch connection will benefit blade servers. When used as an aggregation uplink for many blade servers—each connecting into the server backplane with 10G network interfaces—a higher speed 40G uplink will prevent traffic from the servers from being caught in a 10G bottleneck or face the issues discussed earlier surrounding 10G link aggregation.

BLADE Network Technologies (BLADE), the industry's leading supplier of Gigabit and 10G Ethernet network infrastructure solutions residing in blade servers and "scale-out" server and storage racks, has developed a concept and practice called "Rackonomics." In Rackonomics, a rack by rack approach to designing, provisioning and replicating server/compute systems, data networks and storage area networks (SANs) can decrease the total cost of ownership of data center infrastructures. Rackonomics reduces IT complexity, and enables datacenters to scale incrementally. Rackonomics aligns well with higher speed Ethernet; servers are internally connected at 10G, and racks are attached at 40G or 100G.

While BLADE is fully behind the efforts at developing 40G and 100G Ethernet, we can clearly see that "we aren't there yet." As of this writing, neither NIC / MAC silicon nor switch silicon is close at hand for 40G Ethernet. And even if 40G Ethernet uplinks were available today, there would be no devices available to connect to. Once service providers add 40G backbone capacity, all links in the chain can then benefit from an upgrade to 40G Ethernet.

Conclusions

- Servers today are ready to benefit from 40G Ethernet.
- Bandwidth-intensive applications such as video and high performance computing will drive the need for higher speed Ethernet.
- Datacenters are already aggregating 10G uplinks to increase performance, so there is already demand for 40G Ethernet.
- 40Gbps Infiniband (which is actually only 32Gbps) is becoming available, but does not have Ethernet's massive installed base and market competition that will sustain new products and low prices.
- The market for 40G Ethernet will grow faster than 100G Ethernet in the mid-term.
- BLADE will be first with 40 Gig and 100 Gig Ethernet switching in both blade and rack form factors, adding to our long list of firsts (including 10 Gig, Lossless Ethernet capable blade switches, and more). BLADE is actively evaluating technology for 40G and 100G Ethernet.
- We expect 40G Ethernet products to be viable in the server market, and 100G in the network backbone by 2011.
- 40G and 100G Ethernet are only the beginning. Bob Metcalfe – inventor of Ethernet – is now calling for One Terabit Ethernet.

