

10 Gigabit Ethernet In Servers

Benefits and Challenges

By Bob Wheeler
Senior Analyst

January 2005



10 Gigabit Ethernet In Servers: Benefits and Challenges

By Bob Wheeler, Senior Analyst, The Linley Group

This paper examines the state of 10 Gigabit Ethernet (10GbE) technology as it applies to servers. Intended for information technology professionals, the paper examines the benefits and challenges associated with deploying 10GbE today and in the future. We also present our outlook for 10GbE adoption and how it will impact competing technologies for storage and cluster networks.

Benefits of 10 Gigabit Ethernet

Server Consolidation & Virtualization

In a drive to reduce operating costs, corporations are consolidating their servers into fewer but larger data centers. This physical consolidation has the effect of concentrating network traffic. Because most servers already have Gigabit Ethernet (GbE) connections, aggregating the traffic from a server farm requires a higher-bandwidth connection than can be achieved with GbE links. One of the first applications of 10 Gigabit Ethernet in the data center is in removing this bottleneck by connecting servers to a GbE switch with 10GbE uplink ports.

Taking consolidation one step further, blade servers combine multiple servers into a single chassis. Each server blade connects to the system backplane using a GbE connection. An integrated Ethernet switch connects the blades and provides the connections to the local-area network (LAN). With dozens of CPUs generating network traffic within the chassis, GbE connections to the LAN would be massively oversubscribed. By using 10GbE connections to the LAN, blade servers can scale to larger numbers of processors without the network connection becoming a major bottleneck.

Although blade servers are another step in physical consolidation, most applications still reside on a dedicated server. This model of having one application per server has caused a proliferation of servers in corporate networks. It also makes it difficult to scale and balance performance across servers. Virtualization—decoupling the operating system image from the physical system—promises to eliminate these scalability issues. Virtualization should also allow powerful multiprocessor systems to be more fully utilized as CPU resources are balanced across multiple applications.

The combination of consolidation and virtualization will drive server performance to new levels. But as system performance scales, GbE connections to the LAN become a bottleneck. Trunking multiple GbE links is an interim solution, but trunking introduces processing overhead and requires multiple cables (or fibers). As a result, trunking more than four GbE links is uncommon. With 10 times the bandwidth of GbE, 10GbE is the clear answer to increasing LAN connection speeds.

LAN, SAN, and Cluster Convergence

Today, large data centers use specialized networks for storage and for server clusters. Fibre Channel (FC) is the technology of choice for storage-area networks (SAN). FC networks typically operate at 1Gbps or 2Gbps, while 4Gbps products are just becoming available. Because FC SANs are primarily deployed in large data centers, volumes remain small compared with Ethernet-based LANs. IP storage, which uses Ethernet as a storage network, is emerging as an alternative to FC. The key protocol in migrating storage traffic to the LAN is iSCSI. The iSCSI protocol is already supported in shipping versions of Windows, Linux, and Unix operating systems, and iSCSI host-bus adapters (HBA) are available from major vendors.

Server clusters have been connected using proprietary protocols, such as Myrinet, and more recently using InfiniBand. Although InfiniBand is an industry standard, it serves a narrow market and has not achieved high volumes. InfiniBand also has limited cable reach, making it suitable for use only within a data center or a small HPC cluster. On the plus side, InfiniBand provides 10Gbps connections with little delay (or latency), which is critical to parallel-processing applications. Next-generation InfiniBand products will scale to 30Gbps, keeping InfiniBand ahead of Ethernet in peak performance. There are also industry efforts to redirect the remote direct-memory access (RDMA) technology developed for InfiniBand for use over lower-cost Ethernet. Commonly known as iWARP, the protocols for RDMA-over-IP promise to reduce the latency of 10GbE to a level similar to that of InfiniBand. An application-layer interface, iWARP will allow cluster applications to be ported to 10GbE networks.

Although FC and InfiniBand serve their niches well, installing, maintaining, and managing three networks is clearly inefficient. By leveraging new technologies such as iSCSI and iWARP, 10GbE promises to consolidate the LAN, SAN, and cluster into a single network. This convergence of three fabrics into one will be especially important for blade servers due to their physical constraints. The use of lower-cost Ethernet technology should also broaden the market for SAN and cluster capabilities. For these reasons, the convergence of the LAN, SAN, and cluster networks using 10GbE has broad industry support from chip and system vendors alike. Despite this convergence at 10Gbps, we expect the installed base of FC and InfiniBand to continue to be used, but most new 10Gbps installation will eventually move to 10GbE.

Technology Issues in 10GbE Adoption

Protocol Processing

Before 10GbE gains broad adoption, some technology issues must be addressed. For servers, the biggest issue is protocol processing. As any IT manager knows, a network connection consumes a portion of a server's processing power (measured as CPU utilization). In a single-CPU server, a GbE link can consume close to half of the server's processing cycles. If a conventional network-interface card (NIC) architecture was

10G Ethernet in Servers: Benefits and Challenges

simply scaled to 10GbE, the CPU's processing power would be the bottleneck, and throughput would be severely limited.

One major source of processing overhead is the TCP/IP stack. As a result, there have been ongoing efforts to offload some TCP processing from the system CPU onto the NIC hardware. Some of these TCP-offload functions are available today in 10GbE NIC implementations, while others are still under development. Current products support TCP checksum and large-send (or segmentation) offloads, which yield sizable reductions in CPU utilization. These offloads are supported natively in current versions of Windows and through modified stacks supplied by NIC vendors or system OEMs for Linux and Unix. Another way to reduce TCP overhead is by using special large packets known as jumbo frames. Jumbo frames are useful in storage applications that transfer large blocks of data.

Although NIC vendors can offload all TCP processing by bypassing operating system network stacks, end users have been reluctant to use such products unless a major server vendor directly supports them. This support is not an issue for Unix systems, because the server vendor qualifies specific NIC hardware and a modified TCP/IP stack for use on their systems. To address this problem in Windows servers, Microsoft is developing technology known as TCP Chimney. When deployed, TCP Chimney will enable full TCP offload by NIC hardware. TCP Chimney is due for release in the Scalable Networking Pack for Windows Server 2003 during 1H05. Linux environments are problematic; NIC vendors can easily implement modified TCP/IP stacks, but rigorous qualification requirements are lacking as compared with Windows and Unix systems.

TCP offload is not a panacea. TCP offload will deliver large performance improvements in storage applications, but applications that use small data transfers may see little benefit. Another approach is to scale performance by running the network stack on multiple processors. Linux and Unix already allow multiprocessor systems to scale throughput beyond that provided by 100% of a single CPU. Windows, on the other hand, has been limited to running the stack on a single CPU. In parallel with TCP Chimney, Microsoft is removing this limitation using what it calls Receive Side Scaling (RSS).

Another source of processing overhead is data copying. In a conventional networking stack, received packets are stored into operating-system memory and later copied to application memory. This copying consumes CPU cycles and also introduces a delay. For parallel-processing applications that use small buffers, such as distributed databases, data copying is a major performance drain. iWARP enables data to be written directly into application memory, eliminating the extra copy operation. Applications written for InfiniBand clusters can take advantage of iWARP without modification. In current Windows operating systems, Microsoft enables third-party support of iWARP through the Winsock Direct interface. RDMA-enabled NICs, or RNICs, will become available during 2005.

System Interface

When operating at full line rate, 10GbE pushes the limits of current system interfaces. All available 10GbE NICs use a 64-bit 133MHz PCI-X 1.0 interface. This interface can sustain about 7.7Gbps of throughput, not enough for a 10GbE connection. PCI-X 2.0 supports higher clock speeds, but products supporting this version have not been available. In 2005, system and NIC vendors will support a mix of PCI-X 2.0 and PCI Express.

PCI Express comes in different widths, with each “lane” providing nearly 2Gbps of full-duplex throughput. A four-lane, or x4, slot provides almost 8Gbps of bandwidth but still falls short of 10GbE’s full line rate. The next step up is a x8 slot, which provides more than enough throughput for a 10GbE NIC. Servers are already available with at least one PCI Express x8 slot. (Desktop systems have one x16 slot for graphics, replacing AGP.)

Cabling

Another hurdle for 10GbE adoption is cabling and the related physical-layer standards. The original 10GbE standard defined multiple physical layers for fiber media but did not support copper (twisted-pair) cabling. Of these initial specifications, 10GBase-SR and 10GBase-LR are the most popular today. SR supports multimode fiber up to 300m in length, but is limited to 26m over FDDI-grade multimode fiber. Fortunately, 26m is adequate for most data-center applications. Typically used in metro applications, LR supports 10km reach over single-mode fiber but does not support multimode fiber, which is found in most enterprise applications. Another of the initial specifications, LX4, provides 300m reach over FDDI-grade fiber and also supports 10km reach over single-mode fiber. LX4 products are now available, but they carry a price premium over SR.

More-recent efforts have focused on copper cabling and reducing the cost of 300m multimode fiber applications. As an alternative to SR in the data center, the CX4 standard enables 15m reach over InfiniBand-style copper cables. Although CX4 uses large 8-pair shielded cables, it eliminates the need for costly optical modules. Products supporting CX4 have become available during 2004. Currently under development, LRM promises 300m reach over FDDI-grade fiber at a lower cost than LX4. Finally, 10GBase-T will enable the use of twisted-pair (UTP) cabling, but this standard will not be complete until 2006. Unlike 1000Base-T (GbE), 10GBase-T will likely require CAT6 cabling and still may not achieve 100m reach.

Performance Impact of 10GbE

For many of the reasons discussed above, the performance impact of 10GbE will vary greatly by application and implementation. First-generation 10GbE NICs, which implement partial TCP offloads and a PCI-X system interface, deliver peak performance of 6–8Gbps. At large packet sizes, these NICs consume less than 100% of a typical server CPU. Thus, a first-generation 10GbE NIC should deliver 50–100% more throughput than a conventional 4xGbE NICs with a similar level of CPU utilization.

10G Ethernet in Servers: Benefits and Challenges

Second-generation 10GbE NICs with TCP-offload engines (TOE) are becoming available but still use a PCI-X system interface. These TOE NICs should achieve throughput similar to that of first-generation 10GbE NICs while lowering CPU utilization. Third-generation 10GbE NIC products are likely to adopt PCI Express x8 interfaces and should achieve full 10Gbps line rate with large packets. But achieving full-duplex line-rate performance (20Gbps) is likely to consume all of the processing power of two typical server CPUs. Network traffic fluctuates greatly, however, so this situation represents a peak load rather than a sustained concern.

In applications that use small packets, 10GbE provides lower latency than GbE due to its higher line rate. RNICs that implement iWARP can further reduce latency by eliminating memory copies. This should make 10GbE a viable replacement for 10Gbps InfiniBand and proprietary cluster networks. Over time, 10GbE should deliver much better price/performance than InfiniBand through its superior economies of scale.

Pros and Cons of 10GbE

Hardware Cost

Because 10GbE is in the early stages of adoption, prices for both switch ports and NICs have been high. But competition is heating up as vendors try to stake out a position in anticipation of 10GbE ramping to high volume. During 2004, 10GbE ports for modular switches have dropped from more than \$10,000 per port to as low as \$2,500 per port. By mid-2005, we expect to see fixed-configuration 10GbE switches selling for less than \$1,000 per port without optical modules. Some products will integrate CX4 transceivers, eliminating the need for modules altogether.

Standard form factors (such as MSA) are also helping drive competition for pluggable optical modules. XPAK and X2 are the latest module types shipping in volume; 10GbE SR module pricing to OEMs has dropped to only \$350. Although end-user prices for prior-generation modules (XENPAK) have been in the \$3,000 range, XPAK/X2 prices should fall below \$1,000 during 2005.

10GbE NIC pricing has remained relatively high as volumes have trailed those of 10GbE switch ports. During 2004, Intel introduced a conventional 10GbE NIC with SR XPAK module priced at less than \$5,000. We expect TOE NICs to enter the market in 2005 at similar prices. Prices should fall rapidly as volumes increase, as the underlying manufacturing cost of these products does not warrant such high prices. Intel's design, for example, implements the controller in a single chip.

Although 10GbE is not yet shipping in high volumes, no one doubts it will over time. Gigabit Ethernet today is approaching 100 million ports per year, and 10GbE will eventually follow suit. These economies of scale will drive 10GbE cost below that of 10G FC and Infiniband within the next few years.

Cabling

The plethora of sometimes-competing 10GbE physical-layer standards have created confusion and have prevented any one specification from dominating the market. Fortunately, most new 10GbE products employ pluggable optical modules, enabling end users to choose the standard that best meets their needs. For structured cable plants, GbE backbones can be upgraded to 10GbE using 10GBase-LX4 operating over the installed multimode fiber. Within a data center, CX4 offers a low-cost alternative to fiber where 15m reach is adequate. Thus, solutions exist today for initial 10GbE deployments.

10GbE operation over UTP, however, will have to wait until at least 2006. Previous generations of Ethernet have not reached high volume until UTP cabling was supported, and we expect 10GbE to follow the same pattern.

Changes to Software

Because first-generation 10GbE NICs can use standard operating-system TCP/IP stacks, upgrading to 10GbE does not require any software changes. Some software changes may be desirable, however, to achieve optimal performance. Depending on the operating system, NICs that implement partial or full TCP offload may require a vendor-provided TCP/IP stack or a modified or updated OS stack. In the case of Windows Server 2003, the Scalable Networking Pack will be required to add support for TCP Chimney. As discussed above, TCP offload is especially important for file servers and other storage applications that use large blocks of data.

Distributed applications should not require modifications to use iWARP over 10GbE. Because iWARP looks the same to an application as InfiniBand RDMA, applications written for InfiniBand should work with iWARP without modification. An opportunity exists, however, for new applications to take advantage of RDMA's performance as iWARP is adopted.

Conclusions

Thanks to the ubiquity of Ethernet, the adoption of 10GbE is a certainty. The only questions are how quickly 10GbE will be adopted and to what extent it will displace alternative technologies such as FC and InfiniBand. Early adoption of 10GbE is happening where there is a critical need for additional bandwidth. With tens of millions of GbE switch ports shipped in 2004, the most immediate need is in LAN aggregation and backbone switches. With virtually all new servers shipped with GbE connections, there is also a growing need for 10GbE switch ports within the data center. Of the estimated 75,000 10GbE ports shipped in 2004, the vast majority were in switches. Adoption of 10GbE in the data center and in LAN aggregation during 2004 and 2005 sets the stage for broader deployment in LANs beginning in 2006. During 2006, the first 10GbE-over-copper (UTP) products will reach the market, decoupling 10GbE adoption from that of fiber.

10G Ethernet in Servers: Benefits and Challenges

From a server perspective, 2004 saw early adoption of 10GbE NICs in high-performance computing environments. During 2005, we should see early iWARP implementations enter testing in clusters as an alternative to dedicated cluster networks. Storage applications will be next in line to benefit from 10GbE. We expect to see early deployments of 10GbE NICs running the iSCSI protocol in 2005.

In the longer term, 10GbE promises to displace dedicated SAN and cluster networks in large data centers, but some technology challenges continue to stall large-scale adoption. The industry is investing heavily in overcoming these challenges, however, and solutions are beginning to reach the market. TCP offload, iSCSI, and iWARP are key technologies for this converged data-center network. Blade servers will also benefit from fabric convergence, but design cycles for these systems are likely to delay volume shipments until 2007. As it matures, 10GbE will become a key enabler of server consolidation and virtualization, which will demand network bandwidth that can scale with compute power.