





10-GbE EDGES CLOSER TO MAJOR PRODUCT ROLLOUTS AS SILICON PRICES AND POWER CONSUMPTION FALL. BUT IT MAY BE TWO YEARS OR MORE BEFORE WIDESPREAD DEPLOYMENTS START MEETING DEMAND.

BY ANN R THRYFT • CONTRIBUTING TECHNICAL EDITOR

10-GbE in the MAINSTREAM

For several years, next-generation Ethernet capable of 10-Gbps speeds has been on the brink of entering the mainstream. Some areas of the network have for some time been using optical technology, and demand is increasing for 10-GbE (gigabit-Ethernet) data rates as traffic increases. But vastly more complex technology than in previous generations of Ethernet is necessary to run 10-GbE over 100m lengths of copper. That requirement has resulted in expensive MAC (media-access-controller) and switch silicon and costly and inconvenient PHY (physical)-layer-interface chips, limiting the speedy LAN to high-performance applications covering shorter distances within data centers.

This picture is beginning to change, but widespread deployment of 10-GbE is not likely for at least two years (**Figure 1**). In June 2006, the IEEE finalized the 10GBaseT P802.3an spec for operation over 100m copper. Switch and controller chips that meet the spec are now available, but, along with PHY chips, they are expensive and consume too much power. These problems have been especially severe with 10GBaseT PHY-layer chips.

Silicon power consumption must be less than 5W/port for use in commercially available equipment in chips such as switches and NICs (network-interface controllers). “That’s not the case with most of these chips now,” says Alan Weckel, senior analyst for Dell’Oro Group. “In contrast, optical pluggable [transceiver] modules consume less than 1W.” Most 10GBaseT PHY chips currently consume 8 to 10W per port.



Although per-port prices have for some time been falling, lower component prices alone don't necessarily create demand. "It will take time to ramp up volumes in the next couple of years," says Jag Bolaria, senior analyst for The Linley Group. Meanwhile, the cost of optical fiber is decreasing, "and the early stages of 10GBaseT could be the next kicker in the road in getting volumes up."

THE NEED FOR SPEED

Analysts expect a fairly rapid ramp for 10-GbE technology over the next four to five years. Drivers include sheer bandwidth demands because of server-performance increases and other factors, such as virtualization (see **sidebar** "Virtualization and 10-GbE"). The demand for 10-GbE over copper, including 10GBaseT, comes from the fact that copper costs less and is easier to install than fiber, even though 10GBaseT requires unshielded Category 6A or shielded Category 7 cable to meet the spec's maximum distance of 100m. Although fiber will still find use for longer runs, the shorter enterprise and data-center runs need less expensive, 10-Gbps links.

During the next five years, 10-GbE should replace 1-GbE in two major applications that will drive significant port growth, according to Dell'Oro Group. In wiring-closet-switch uplinks, most 10-GbE fixed ports will be uplinks on 24- and 48-port 1-GbE switches, and

AT A GLANCE

▣ The difficulty of implementing 10-GbE (gigabit Ethernet) over 100m copper, which the 10GBaseT spec stipulated in June 2006, has delayed the widespread deployment of 10-GbE. Major drivers for 10-GbE include server virtualization and 1-GbE link aggregation.

▣ 10GBaseT's complex technology has resulted in expensive PHY (physical)-interface chips that run too hot.

▣ Second-generation 10-GbE PHY silicon at the 65-nm-process node, due in 2008 and 2009, will help cut PHY-chip power consumption to approximately 5 to 6W, will improve design, and will lower costs through greater integration.

▣ Manufacturers are developing direct-attachment copper-twinaxial cables for use with smaller-footprint, lower-power SFP+ optical hot-pluggable-transceiver modules, offering an alternative for 10-GbE over distances of 10 to 15m within data centers.

the remaining 10-GbE ports will be 10-GbE-only switches, those finding use at the top of the stack for aggregation. The other major application area is in direct server connections.

Currently, copper finds use mainly between switches and PCs and between switches and servers. The uplinks from wiring-closet switches to data centers

that are currently optical links will remain optical and will have significant potential for volume, says Weckel. Because enterprises are undergoing a major wiring-closet upgrade, the market for switching is increasing, as well. In switch-to-switch connections, the medium for 1-GbE is now optical and will remain optical at 10-GbE. In direct-server connections, the medium is copper, and it will remain copper at the higher speed.

As the number of 1-GbE ports increases, aggregation of those ports becomes a major reason to begin using 10-GbE technology. Without it, the uplink bandwidth is less than the bandwidth of the downstream ports, and blocking occurs. Tier 1 OEMs are selling 1-GbE gear for around \$100/port, says Kamal Dalmia, vice president of marketing for Teranetics. For 10GBaseT switches at their introduction, the price will be approximately \$500/port, a per-gigabit cost of roughly half that of 1-GbE equipment. Another main driver is the need to connect high-performance computing-blade servers in data centers with bandwidth commensurate with their processing speeds. Those servers include eight or 16 processors, each of which is increasingly likely to contain four or even eight cores, ramping up power requirements and speed.

Aside from the combined power-consumption and technology issues in implementing the technology, another major issue could hinder the rate of 10-GbE deployments, says The Linley Group's Bolaria. If the cost of implementing one 10-GbE port is too high, it may make more sense to aggregate links by combining two to four 1-GbE ports; more than that number would be expensive and cumbersome.

MORE COMPLEXITY, POWER

The first 10-GbE standard, IEEE 802.3ae, which originated in 2002, specified several PHY interfaces for optical-transmission media. In 2004, the IEEE 802.3ak-2004 short-reach amendment allowed 10-GbE over 15m coaxial cabling using 10GbaseCX4 PHY-layer chips. Compared with 1000BaseT, 10GBaseT data rates require highly complex digital-signal processing in silicon to deal with echo cancellation and crosstalk cancella-

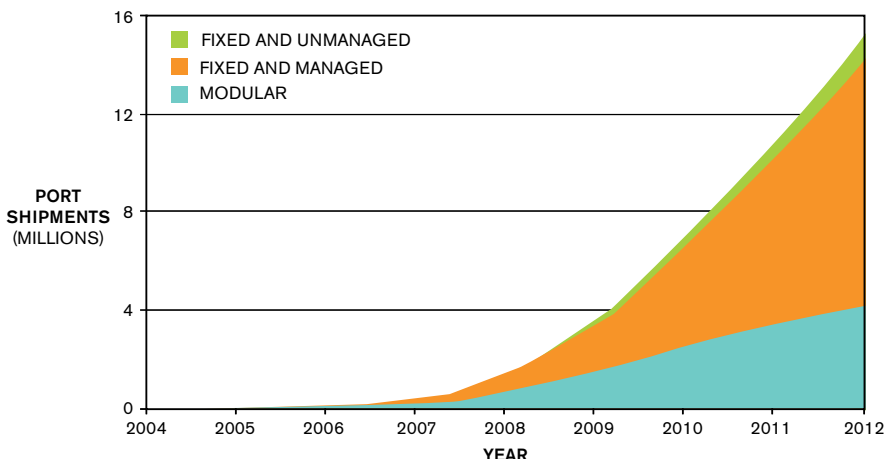


Figure 1 Total 10-GbE port shipments will grow from slightly more than 700,000 in 2007 to 15 million in 2012 (courtesy Dell'Oro Group).



tion, as well as more sophisticated analog circuits. All of this processing to clean up the received signal must take place on links that are 10 times faster than the 1000BaseT spec. The 1000BaseT spec left a wide margin for compensation, but 10GBaseT does not, says Brad Booth, who acted as chairman of the IEEE

10GBaseT task force and is chairman of the Ethernet Alliance.

The 10GBaseT spec includes a subset for a short-reach PHY interface over a maximum 30m of Category 6 four-wire UTP (unshielded-twisted-pair) cable as a means of going to a lower-power operation mode. When the spec's

developers wrote the draft, the concern was that most devices for a 100m range would consume 10 to 12W of power, and those levels of power consumption limit 10-GbE deployment in switch applications, says Booth. In a shorter-reach application with the 30m option, the PHY-layer device could drop to

VIRTUALIZATION AND 10-GbE

One reason for slow 10-GbE (gigabit-Ethernet) deployment is the lack of demand from end devices. Before virtualization, servers needed only a few 1-Gbps connections, and utilization rates weren't 100%. "With virtualization, utilization has increased, and throughput is now higher," says Alan Weckel, senior analyst for Dell'Oro Group. "Now, the demand for a 10-Gbps pipe is growing, and cost is becoming one of the barriers."

As internetworking of computers has increased, dependence on data centers has grown, in turn increasing the move toward virtualization. "Today, everyone is using everyone else's computers, such as for Web-based storage," says Bob Nunn, president and chief executive officer of Fulcrum Microsystems. "But the goal of a virtualized data center won't be achieved without a high-bandwidth interconnect technology that is common to the entire data center." For many, including Nunn, 10-GbE is that interconnect technology.

The increase in storage, server, and switch clustering is behind the move toward virtualization, and latency becomes critical.

"To the user, it must look like one server, and that [scenario] won't happen if server response takes longer than expected," says Kamal Dalmia, vice president of marketing for Teranetics. 10GBaseT increases by a factor of five the density of bits moving in a rack (Figure A). It also halves the cost of bits moving in the rack and increases server-utilization rates by two to four times using virtualization, enabling the network to move bits at a much lower cost, says Matt Rhodes, Teranetics' CEO.

A virtualized data center shares and distributes resources to improve scaling and reduce costs. In conventional arrangements, a local bus interconnects centralized components, whereas in virtualization, fabrics interconnect distributed components, including remote storage. Managing the data during heavy traffic becomes an issue, so switches must incorporate congestion-management and load-balancing features, says Nunn.

In a virtualized server, virtualized guest operating systems run simultaneously on a multicore machine using a "hypervisor," or virtual-machine

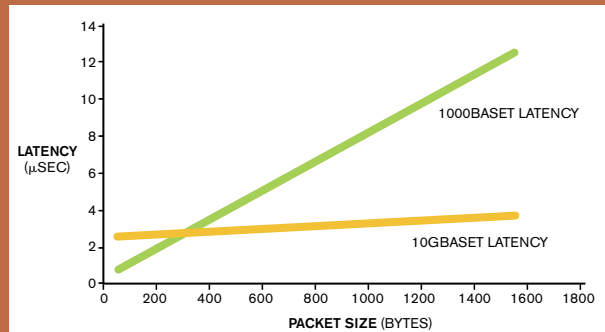


Figure A Even though its PHY-layer latency is greater than that of 1000BaseT, 10GBaseT's higher speed gets packets more quickly to their destinations (courtesy Teranetics).

monitor. Although data-center applications themselves don't scale well across the multiple cores of server CPUs, virtualization changes that scenario, says Steve Pope, co-chief technology officer at Solarflare. "You can have a virtual operating system on each core, each of which runs a virtual application." However, the hypervisor can become a bottleneck when it comes to I/O across those cores.

The Solarstorm vNIC (virtual-network-interface-controller) hypervisor-bypass technology operates on Solarflare's Solarstorm Ethernet controllers. It offloads the hypervisor's I/O burden and provides solid I/O performance to virtualized guest operating systems. Because controller silicon

must provide DMA (direct-memory-access) queues, which can also scale across cores, Solarflare's virtualization architecture provides more than enough virtual interfaces. With advanced virtualization schemes, you can realize significant performance benefits from allocating multiple DMA queues to each core and to each guest operating system, says Pope. Guest operating systems can also bypass the hypervisor to program the virtual interfaces using a set of open-source silicon-vendor-mutual APIs (application-programming interfaces) that Solarflare developed. These APIs let any silicon vendor take advantage of guest-operating-system access.



lower power, which helps eliminate a lot of noise along with many cancellation circuits. The spec also allows the use of Category 6 cable for distances as long as 55m, because 70% of data centers' reach is 55m or less. However, operation over this distance doesn't buy lower power consumption.

Additional PHY-silicon issues are the need for multiple ports to pare down the cost per port, and the inclusion of multispeed ports for backward compatibility with 1-Gbps and even 100-Mbps Ethernet. An alternative method of connecting 10-GbE over copper may be possible using a new form factor in hot-pluggable optical-transceiver modules. The SFP+ (small-form-factor-pluggable-plus) module has a smaller footprint than the previous SFP form factor and consumes less power, allowing greater module density on a line card and offering lower per-port costs. Manufacturers are also developing direct-attachment SFP+ copper-twinaxial cables for distances of approximately 10 to 15m, which are adequate for connections within a data center.

10-GbE-SILICON ISSUES

The process technology for most chips now implementing 10-GbE over copper is 130 or 90 nm. The next generation may go to 65 nm, which is the next-lowest-cost node at which volumes may increase, says The Linley Group's Bolaria. Some in the industry believe that it will take a 45-nm process to reach the lower power consumption required to drive volumes and are trying to integrate some of the analog circuitry necessary for 10GBaseT at that process node. But the high amount of analog circuitry could be a problem in such smaller geometries, says the Ethernet Alliance's Booth. "One alternative may be multichip modules with analog front ends running in one process technology and digital back ends," he says.

Definitions of "high volume" can vary. Huge volumes are on the order of 10 million and 20 million ports, says Bolaria. In 2007, volumes for 10-GbE-switch ports with 10W PHY interfaces were approximately 640,000, and they contained 10W PHY-layer chips. "The next tier would be less than 5W, which will enable [manufacturers to ship] a few million ports," he says. "But, to get to

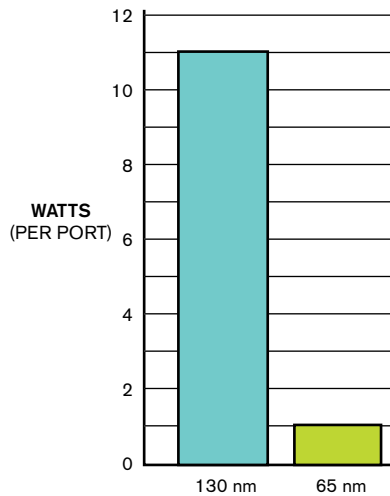


Figure 2 Moving to next-generation process technology will reduce 10GBaseT-silicon power consumption and make possible additional features (courtesy Broadcom).

20 million in port volume, PHY silicon will probably need to consume less than 2W." As ports per chip increase, chip volumes will decline somewhat.

The power budget of a typical first-generation 10-GbE endpoint NIC is approximately 25W, but designers usually want to stay closer to 15W, says Blaine Kohl, vice president of Tehuti Networks. You can build a single-port-10-GbE NIC with a 7W PHY chip, but you can tweak the power allocation even more in some designs. Manufacturers will most likely build single- and

dual-port adapters for endpoints using this year's generation of 10GBaseT PHY silicon. However, switches need PHY chips closer to 3W. By 2010, 3 to 4W 10GBaseT PHY chips will probably be available. At that point, with 2 to 3W controllers, you may be able to integrate a controller chip and a PHY chip in one package. Switches could then adopt a 10GBaseT PHY chip, endpoints could adopt the 10GBaseT package, and manufacturers could be shipping 10GBaseT equipment in volume.

After offering samples of 10-GbE-controller and 10GBaseT PHY chips for a year and a half, Solarflare Communications has just introduced second-generation, 65-nm parts. The SFT9000 10GBaseT PHY chip consumes less than 6W and features multispeed auto-negotiation at speeds as low as 100 Mbps. The SFC4500 10-GbE controller chip consumes 2.2W and features virtualization acceleration. Solarflare expects next year to make available a single-chip LAN-on-motherboard device that will integrate 10GBaseT-PHY and 10-GbE-controller silicon, according to Bruce Tolley, vice president of marketing.

Teranetics' first-generation all-CMOS TN1010 10GBaseT multirate PHY chip supports 1-GbE and 100-Mbps Ethernet. The company's next-generation 65-nm chip will be available this year, with 30 to 40% less power consumption than the current total of about 10W, says CEO Matt Rhodes. Fulcrum Microsystems' single-chip, 24-port FM4000 10-GbE IP (Internet Protocol) Version 4/6 switch/router chips target use in data-center-switching platforms in high-performance-computing, server, and storage-host interconnection, and data-center-aggregation applications. The Layers 2/3/4 chips have full line-rate performance on all ports with a total throughput of 360 million packets/sec. Their 300-nsec latency provides a highly responsive network fabric that exceeds the performance of specialty fabrics, such as InfiniBand and Fibre Channel, and suits high-performance-clustered-computing applications, says Bob Nunn, Teranetics' president and CEO.

Broadcom makes single-chip, 10-GbE switch and controller silicon and optical transceivers. The 65-nm process tech-

MORE AT EDN.COM

- ⊕ For a look at backplane-design issues at 10-Gbps speeds, go to www.edn.com/article/CA6317073.
- ⊕ You can find a discussion of the effects of dispersion on optical-link performance in networks with 10-Gbps links and what electronic-dispersion compensation brings to this problem at www.edn.com/article/CA6317075.
- ⊕ For another article by Ann R Thryft, go to www.edn.com/article/CA6470826.
- ⊕ Go to www.edn.com/080501cs and click on Feedback Loop to post a comment on this article.

nology for these chips requires a significant effort, especially for the switches, because they are large, highly integrated chips with both analog and digital components, says Eric Hayes, director of marketing for the network-switching line of business. The 65-nm BCM56820 switch chip's power consumption is 1W per 10-GbE port—a drastic reduction from the previous 130-nm switch chip's 11W per port (**Figure 2**). “We’re seeing a clear requirement to go to 10-GbE with all of the features and capabilities that were available at 1-GbE, such as security, strong Layer 4 classification for QOS [quality of service], and Layer 3 routing,” he says.

Tehuti Networks based its new SFP+-adapter-reference designs, which include the single-port TN7587-S and dual-port TN7587-D NICs for optical 10-GbE, on the company's single-chip TN3016 10-GbE single- or dual-port controllers. The reference designs include AMCC's (Applied Micro Circuits Corp's) 10-GbE SFP+ QT2025 PHY chip. When you populate the NIC with two SFP+ optical modules, the TN7587-D's power dissipation is 15W. The single- and dual-port controllers dissipate 6 and 7W, respectively.

WHAT'S NEXT?

Many enterprises have just begun to consider 40- and 100-GbE for aggregating 10-Gbps links in data centers. “Even telephone companies are requesting 100-Gbit products,” says The Linley Group's Bolaria. “There's demand for it now in the core of the network.” Because of video-on-demand and Web-based applications with millions of simultaneous users, such as Facebook, Netflix, and YouTube, some industry observers predict that the bandwidth-scaling needs of carriers and ISPs (Internet-service providers) may bypass 40-GbE.

For the first time, however, no aggregation technology is in place for the upcoming lower speed of Ethernet, which could hinder the market, says Dell'Oro Group's Weckel. For the lower speed to be successful, manufacturers must be shipping products using the higher speed, even if not in high volumes. But with 10-GbE to the server, no 40- or 100-GbE line cards exist to provide uplinks to data centers on the

FOR MORE INFORMATION

AMCC www.amcc.com	Intel Corp www.intel.com
Broadcom www.broadcom.com	Linley Group www.linleygroup.com
Dell'Oro Group www.delloro.com	Solarflare Communications www.solarflare.com
Ethernet Alliance www.ethernetalliance.org	Tehuti Networks www.tehutinetworks.net
Fulcrum Microsystems www.fulcrummicro.com	Teranetics www.teranetics.com

aggregation side. “The higher-speed uplink technology must be present for there to be truly widespread adoption,” he says.

The pressure will increase even more when servers begin to appear with 10-GbE interfaces on the motherboard, which may occur with the next generation of Intel server CPUs. Ultimately, 10-GbE may go all the way to the desktop. For many, Ethernet has become the fabric of choice in the network, and that situation is also driving 10-GbE deployment. **EDN**

REFERENCES

- 1 Wheeler, Bob, “10G Ethernet Adoption in Volume Servers,” The Linley Group, 2006, www.ethernetalliance.org/technology/research/10GbE_Adoption_in_Servers.pdf.
- 2 “Moving 10 Gigabit Ethernet into a Volume Platform,” Ethernet Alliance, 2006, www.ethernetalliance.org/technology/white_papers/Moving_10_Gigabit_Ethernet_into_a_Volume_Platform.
- 3 “10GBASE-T: 10 Gigabit Ethernet over Twisted-pair Copper,” Ethernet Alliance, 2007, www.ethernetalliance.org/technology/white_papers/10GBase_T2.pdf.
- 4 Gumanow, Gary, “Solving the Hypervisor Network I/O Bottleneck: Solarflare Virtualization Acceleration,” Solarflare Communications, 2007, www.solarflare.com/technology/documents/SF-101233-TM-5.pdf.

AUTHOR'S BIOGRAPHY

Contributing Technical Editor Ann R Thryft has been writing about technology, including wired and wireless networking, for more than 20 years. You can reach her at athryft@earthlink.net.