THE NEXT GENERATION OF ETHERNET

JOHN D'AMBROSIA

Ubiquitous. An adjective that dictionary.com describes as "existing or being everywhere, especially at the same time, omnipresent." This word is also synonymous with Ethernet. Combine this with the Ethernet Alliance's perspective of Ethernet — "From Carrier to Consumer." Now, consider that the Dell'Oro Group reported that a total of 182,817,000 Ethernet ports were shipped in 2006, generating $15.2 billion in revenue.

Now, imagine the challenges facing the body that will define the next generation of that technology. This is the task facing the IEEE P802.3ba Task Force, which was recently approved by the IEEE Standards Association Standards Board to begin developing the standard that will follow 10 Gigabit Ethernet (GbE).

The task force is beginning the development of the next generation of Ethernet at a time when existing networks are already being strained by current capacity requirements. Network engineers forced to deal with this issue are attempting to plan their networks to support anticipated increasing bandwidth requirements. The industry is looking to the next generation of Ethernet for relief from existing and potential future network bottlenecks that are being created as data aggregation drives bandwidth requirements to double approximately every 18 months.

The explosive growth seen in networks is easy to understand: bandwidth consuming content is being utilized by an increasing number of users, who are accessing the content through a growing number of channels at a faster rate. The bandwidth requirements for computing and server applications at the network edge, however, are driven by CPU performance (and Moore's Law), and are growing at a slower rate, doubling approximately every two years.

In previous "next" generations of Ethernet the bandwidth crunch has been solved by making 10x leaps in the medium access control (MAC) data rate and then using the new rate for all applications. However, given the different growth rates of computing and networking, the industry recognized that developing a single new higher speed would not satisfy the individual requirements of the various applications in the ecosystem that Ethernet supports. Therefore, two MAC rates were adopted: 40 Gb/s for computing and server applications, and 100 Gb/s for network aggregation applications. Additionally, it was agreed to develop physical layer specifications targeting the specified application space for each MAC rate. See Table 1.

The task force faces multiple challenges in developing an architectural scheme that will accommodate 40 GbE and 100 GbE. As part of this effort, it must define how it will address the media independent interface (MII) for each rate and all of the respective interfaces. Finally, it must determine how it will address the specifications of seven distinct physical layers (PHYs).

	40 GbE	100 GbE
At least 1 m backplane	√	
At least 10 m copper cable	√	√
At least 100 m OM3 multi-mode fiber	√	√
At least 10 km single-mode fiber		√
At least 40 km single-mode fiber		√

Table 1.

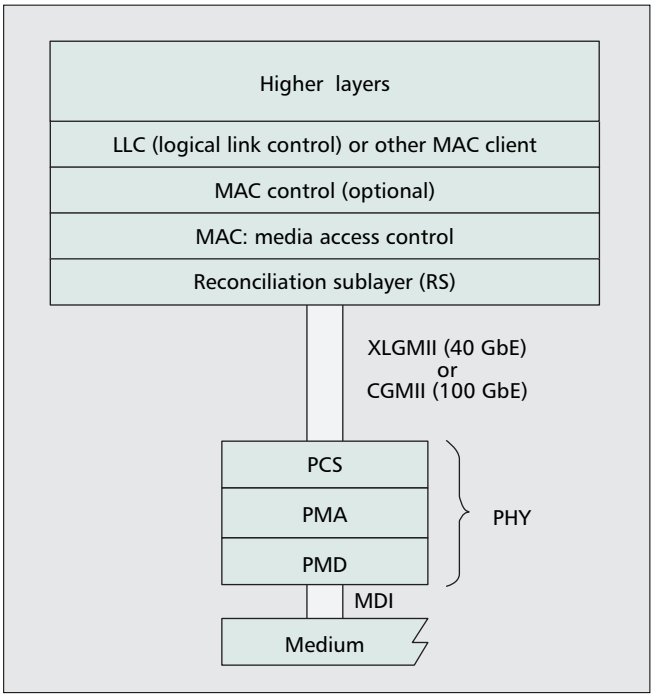

Figure 1. *Typical Ethernet architecture.*

Architectural Needs

Figure 1 shows the typical basic architecture that has been the basis for previous generations of Ethernet. There will be no changes to the 802.3 MAC other than increasing the data rate; the frame format and minimum and maximum frame size will be preserved, and full-duplex operation only will be supported. The reconciliation sublayer (RS) maps the serial bitstream of the MAC into the respective MII. The 40G MII (XLGMII) and 100G MII (CGMII) would provide an interconnection between the MAC and 40 GbE/100 GbE PHYs, respectively, which comprises physical coding sublayer (PCS), physical medium attachment (PMA) sublayer, and physical medium dependent (PMD) sublayer. The PCS sublayer codes the data from the respective MII, which is then serialized by the PMA sublayer. The PMD sublayer is responsible for signaling through the media dependent interface (MDI) and across the medium the PMD is intended to support.

The dual MAC rates and various physical layer specifications that may be implemented today as well as tomorrow will need to be taken into consideration when determining the overall architectural scheme. Two basic approaches are being proposed at this time.

The first approach is similar to the typical approach shown in Fig. 1. A single PCS is used to encode/decode the entire data stream, which would then be combined with the PMA

(Continued on page S10)

(Continued from page S8)

ber of lanes for the respective PMD. New PHYs would then be created based on the PCS/PMA/PMD combinations developed.

The second approach is based on the physical medium entity (PME) aggregation concept developed for the Ethernet in the First Mile project. In this approach the appropriate number of lower-speed parallel PHYs are aggregated together to support the desired rate. This approach is particularly appealing for reuse of 10GbE PHY technology, and given the common packaging technique of four PHYs in a given device, very attractive for 40 GbE PHY specifications.

Determining an architectural scheme for the next generation of Ethernet will be one of the first technical issues for the Task Force to resolve. The Task Force could choose to develop two architectures, which would allow optimization for each specific rate. This decision, however, would burden (in terms of cost and complexity) those who have to support both architectures. Therefore, many are voicing support for a single unified architecture to support both rates.

The PCS, which does the encoding of the data stream, may also be an area of focus. Many discussions to date have assumed that the 64B/66B scheme, which is employed by the 10GBASE-R PCS, will be used. Some, however, have discussed the possibility of using an 8B/10B encoding scheme, but it is not clear where this will lead.

Architecturally, the Task Force must also consider how to ensure the ability to appropriately transport both 40 GbE and 100 GbE over the optical transport network (OTN) found in the wide area network (WAN). Today's OTN is based primarily on 10 Gb/s/λ dense wavelength-division multiplexing (DWDM), with 40 Gb/s/λ DWDM starting to be deployed and 100 Gb/s (ODU4) OTN under development in the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T). The issue for the next generation of Ethernet will be the transport of 40 GbE over the existing 40G (ODU3) OTN. There will be less of an issue with 100 GbE, since the development of ODU4 is happening in parallel with the development of 100 GbE. Potential solutions have been identified, but it is likely that action by both the IEEE and the ITU-T will be necessary to ensure compatibility.

Interfaces

The development of the architecture scheme will also require that the respective MIIs and inter-sublayer interfaces are considered.

While previous generations of Ethernet have had specified MIIs, ultimately the industry developed interfaces that were more practical from an implementation perspective. For example, consider the 10 Gb MII (XGMII), which is 74 pins wide. The IEEE 802.3ae Task Force decided to create the XGMII extender sublayer (XGXS) and the 10 Gigabit attachment unit interface (XAUI) to extend the reach of the interface and reduce the number of pins to 16, four differential pairs in each direction. Since then the industry has reduced this down to a serial 10 Gb solution running over a single differential pair in each direction.

There are similar issues faced when considering the XLGMII and CGMII. The CGMII interface could be as wide as 256 bits in one direction alone, making it very impractical as an off-chip interface. It could be useful, though, as an on-chip interface. One can easily envision a 40 Gb XS (XLGXS) with a 40 Gb AUI (XLAUI) or a 100 Gb XS (CGXS) with a 100 Gb AUI (CAUI). Either of these interfaces could be implemented as an n lane × 10 Gb/s rate. However, it has been suggested that a 10 lane × 10 Gb/s approach might result in an interface that is too wide (10 differential pairs in each direction), so a 25 Gb/s lane rate should also be considered (or work should be started by the industry to make such a solution possible).

Decisions regarding the inter-sublayer interfaces will need to be made. There have been standardized instantiations of optional interfaces, such as XGMII as a PCS interface, and the 10 Gb 16 bit interface (XSBI) as a PMA interface. Historically, there have been no defined PMD interfaces in the IEEE, as this has been left to the industry to resolve. The group will need to choose for each interface whether it will be defined in a physical or abstract manner.

Physical Layer Requirements

General conceptual proposals discussed for the physical layer specifications have primarily focused on a parallel lane approach. The only exception (at this time) is the physical layer specification for 100 GbE over single-mode fiber, where some individuals are proposing a serial solution. Let us consider the various physical layer requirements.

40 GbE, which is targeting computing and server applications, can reutilize existing 10 GbE technologies to minimize the overall R&D investment, which will help improve the adoption rate of 40 GbE. For those end users adopting blade servers, this may be of particular importance. The IEEE 802.3ap™ Backplane Ethernet specification defines two backplane PMDs capable of delivering 10 GbE across a backplane. 10GBASE-KX4 specifies 10 GbE operation over four differential pairs in each direction across a backplane. 10GBASE-KR specifies 10 GbE operation over one differential pair in each direction across a backplane. Therefore, backplanes that are architected based on 10GBASE-KX4 but designed in accordance with the 10GBASE-KR channel model should have an inherent ability to support 40 GbE operation.

Reuse of 10GBASE-KR technology is also being considered as a solution for delivering 40 GbE and possibly even 100 GbE over at least 10 m of copper cabling. Some have suggested, however, that a four-lane approach based on a 25 Gb/s signaling scheme may be a better long-term solution. This approach, however, would necessitate the development of such a signaling scheme since none exists today.

For the multimode fiber physical layer specifications for 40 GbE and 100 GbE, reuse of existing 10 GbE serial technology with parallel fibers and 850 nm VCSEL arrays has generated the most interest. This would offer a low-cost low-power solution in a small footprint. Array technology has matured, with reliability and yields well understood. This is a particularly exciting area, similar in some aspects to years ago when the integration of multiple transistors on a single die began, and holding the potential promise of applying Moore's Law to optical transmission.

There are a multitude of approaches that could form the basis for the PHY specifications for operation at 100 Gb/s over 10 and 40 km of single-mode fiber. For both reaches solutions at both 1310 and 1550 nm have been suggested. Serial and parallel WDM approaches have been suggested. While some may consider a serial solution to be expensive at this time, it is clear that the future specification of a serial solution must be considered in the development of the overall architecture. Parallel approaches suggested have included 10λ × 10 Gb/s, 4λ × 25 Gb/s, and 5λ × 20 Gb/s using either a direct or an electro absorption modulation laser. All of the possible solutions for transmission over 40 km utilize optical amplification, with dispersion compensation also being suggested for the serial approaches.

(Continued on page S15)

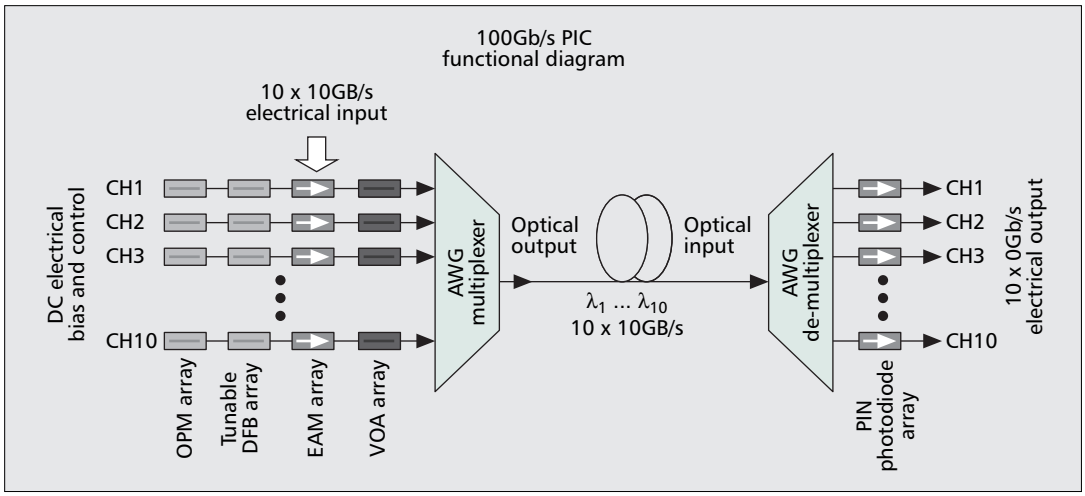

Figure 1.

reduction curve defined by volume manufacturing efficiencies, greater functional integration, and increased device density. This promises ongoing scaling of WDM system functionality and reductions in the "cost per bit" for optical transmission capacity.

For more information and details on PICs, please see the recommended reading below.

Additional Reading

[1] J. Lam *et al.*, "Multi-Function Integrated AWG Devices," *Proc. OFC/NFOEC*, paper OTuD1, 2005.
[2] R. Nagarajan *et al.*, "Large Scale Photonic Integrated Circuits," *IEEE J. Sel. Topics Quantum Elect.*, vol. 11, no. 1, Jan./Feb. 2005, pp. 50–65.
[2] R. Nagarajan et al., "400Gb/s (10-Channel x 40Gb/s) DWDM Photonic Integrated Circuits," *Elect. Lett.*, vol. 41, 2005, pp. 347–49.
[3] R. Nagarajan *et al.*, "Single-Chip 40-Channel InP Transmitter Photonic Integrated Circuit Capable of Aggregate Data Rate of 1.6 Tbit/s," *Elect. Lett.*, vol. 42, 2006, pp. 771–73.
[4] R. Nagarajan *et al.*, "Monolithic, 10 and 40 Channel InP Receiver Photonic Integrated Circuits with On-Chip Amplification," *Proc. OFC/NFOEC*, PD paper, 2007.
[5] M. Allen *et al.*, "Digital Optical Networks Using Photonic Integrated Circuits (PICs) Address the Challenges of Reconfigurable Optical Networks," *IEEE Apps. & Practice*, 2007.

Biography

SERGE MELLE (smelle@infinera.com) is vice president of technical marketing at Infinera Corp., Sunnyvale, California, responsible for market development, technical customer support, and network architecture strategy. Prior to joining Infinera he was vice-president of market development at Nortel Networks, supporting the deployment of major optical networks for service providers in North America, Europe, the Middle East, and Africa. Before joining Nortel, he held business development and product management positions at Pirelli Telecom Systems, where he was involved in the implementation of the industry's first WDM and optical amplifier network deployments. Before this he held product management and engineering positions at EG&G Optoelectronics. He has extensively published in the fields of optics and networking, and holds a B.Sc. degree in physics from Concordia University, Montréal, and an M.A.Sc. degree from the University of Toronto, Canada.

NEXT GENERATION OF ETHERNET/*continued from page S10)*

Summary

The formation of the IEEE P802.3ba Task Force signifies the beginning of the project to define the next generation of Ethernet. While it is unknown what the final solution for 40 GbE and 100 GbE will be, it is clear that the need is here, and the Task Force, as well as the computing and networking industries, have a significant amount of technical work to complete to make the next generation of Ethernet a reality.

To learn more about the IEEE P802.3ba Task Force and details about participation, see http://www.ieee802.org/3/.

Biogrpahy

JOHN D'AMBROSIA, as a scientist at Force10 Networks, focuses on components technology and leads the company's involvement in industry groups. He has been an active participant in the development of Ethernet-related technologies since 1999. At the present time, he is the chair of the IEEE 802.3 Higher Speed Study Group, which is driving the standards development process for the next speed of Ethernet. He served as secretary for the IEEE 802.3ap Backplane Ethernet Task Force, and participated in the development of XAUI for 10 Gigabit Ethernet. He also served as director and secretary for the Ethernet Alliance, and was the chair of the XAUI Interoperability Work Group for the 10 Gigabit Ethernet Alliance. For all of his efforts related to Ethernet, he was recognized by *Network World* in 2006 as part of its "50 Most Powerful People in Networking" list. He also acted as secretary for the High Speed Backplane Initiative and chair of the Optical Internetworking Forum's Market Awareness & Education committee. Prior to joining Force10, he was with Tyco Electronics for 17 years.