The 802.3z Gigabit Ethernet Standard

By Howard Frazier

This column considers standards development in the IEEE 802.3z Gigabit Ethernet Task Force, part of the IEEE 802.3 CSMA/CD Working Group. IEEE Std 802.3z-1998 was formally approved by the IEEE Standards Board on June 25th, 1998. The author is responsible for the leadership of the Gigabit Project and serves as the Task Force chair of IEEE 802.3z.

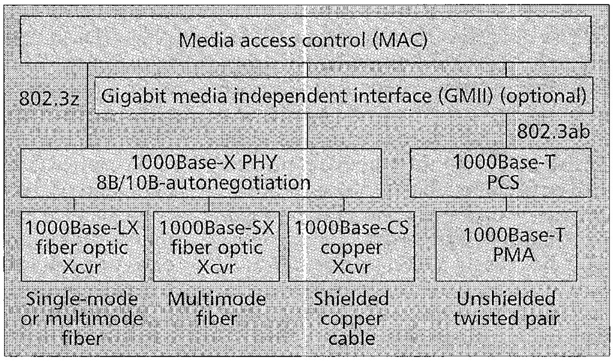
These perspectives, which will appear in each issue of IEEE Network, *are aimed to provide the reader with highlights of IEEE 802 activities to enable better dissemination of these standards into marketable products as well as seek new ideas to be brought into the IEEE 802 arena. I hope the readers find this column useful and look forward to providing future installments. Please send suggestions, criticisms, and requests for future articles to jcarlo@ti.com, IEEE 802 chair.*

The standard for gigabit Ethernet, IEEE Std 802.3z, extends the operating speed of the world's most popular local area network to 1 billion bits per second (1000 Mb/s) for interconnecting high-performance switches, routers, and servers in the backbone of local area networks. Maintaining backward compatibility with the over-100-million-node installed base of 10 Mb/s and 100 Mb/s was a key requirement. Throughout the two-and-a-half-year standard development cycle, the IEEE P802.3z Gigabit Task Force was guided by the overall objective of increasing the data rate of Ethernet technology by embracing the advancing state of the art in network design and physical-layer signaling technology. Soon to be published as a supplement to ANSI/IEEE Std 802.3 (fifth edition), IEEE Std 802.3z represents the next logical step in the evolution of Ethernet local area networks.

As a supplement to ANSI/IEEE Std 802.3, the standard for Gigabit Ethernet had a solid base to build on. IEEE Std 802.3z consists of a series of updates to the 802.3 base standard plus a group of new clauses (chapters) which describe the characteristics unique to 1000 Mb/s operation. The standards produced by the IEEE 802.3 Working Group address the two lowest layers of the International Organization for Standardization (ISO) seven-layer reference model, the data link and physical layers. Within these two layers, the specification is divided into sublayers, separated by well defined interfaces. Figure 1 presents a block diagram identifying the various components of IEEE Std 802.3z.

Starting from the top of the diagram, the media access control (MAC) sublayer describes the algorithms used to control the transmission and reception of packets on an Ethernet network. IEEE Std 802.3z encompasses both the newer full-duplex MAC and the classic carrier sense multiple access with collision detection (CSMA/CD) MAC. The full-duplex MAC is described in IEEE Std 802.3x, and allows point-to-point simultaneous transmit and receive communications with a dedicated full data rate per station. The classic Ethernet CSMA/CD is based on a shared Ethernet where all stations share the same data rate. While most implementations will take advantage of the contention-free access and flexible topologies permitted by full-duplex operation, the P802.3z specification also extends the CSMA/CD MAC to work at 1000 Mb/s. A technique known as *carrier extension* was added to the CSMA/CD MAC to overcome the inherent limitation of the collision detection algorithm which mandated that the round-trip delay between any two stations could not exceed the time required to transmit the smallest allowable frame. An optional feature called *frame bursting* was defined to improve the throughput of gigabit CSMA/CD LANs.

When operating in full-duplex mode, the MAC relies on the fact that the underlying serial communications link inherently supports simultaneous transmission and reception without interference between the transmitted and received signals. Essentially, the CSMA/CD protocol is disabled, and since there is no round-trip delay constraint, carrier extension is not necessary. Furthermore, the span of any given link may be as long as the underlying physical layer permits. Full-duplex operation was introduced into IEEE Std 802.3 by the P802.3x project, which also specified a mechanism for link-level flow control. In the event of transient congestion on a link, a receiver can inhibit the transmission of frames by emitting a Pause frame, which instructs the transmitter to withhold further transmissions until some amount of time transpires.



■ Figure 1. *Gigabit Ethernet Layer diagram.*

The optional gigabit media-independent interface (GMII) definition allows MAC and physical-layer (PHY) implementations from different vendors to interoperate, and provides a means by which the future 1000BASE-T PHY can be attached to the MAC. The GMII also provides a convenient division between the high-density digital logic functions associated with the MAC, and the high-speed mixed signal functions associated with the PHY. This partition allows appropriate design and fabrication processes to be applied to each of these functions.

The GMII delivers 8-bit octets to the physical coding sublayer (PCS) on the transmit path, and accepts 8-bit octets from the PCS on the receive path at a rate of 125 million octets (1 billion bits) per second. The 1000BASE-X PCS borrows heavily from the NCITS T11 Fibre Channel standard, using the same 8B/10B code set. The P802.3z Task Force adapted the Fibre Channel code to gigabit Ethernet after stripping the coding down to the bare essentials. Starting with a mature and well understood code will save vendors time in the product development process and also speed the standards process.

In a fashion similar to the 100BASE-X PCS defined in IEEE Std 802.3u, the 1000BASE-X PCS uses a continuous signaling scheme, wherein an active signal is always present on the medium. The idle period between packets is filled with a special pattern of symbols which contain a sufficient number of transitions between the logic 1 and logic 0 states so that a receiver's phase-locked loop can maintain clock recovery on a continuous basis. Another special sequence of symbols is used to delimit the beginning and end of a packet.

The 10-bit symbols produced by the PCS are serialized by the physical medium attachment (PMA) sublayer. Commercially available Serializer/Deserializer (SerDes) components first developed for Fibre Channel can be used for gigabit Ethernet. Only modest additional testing requirements are needed, mainly due to the fact that gigabit Ethernet operates at a clock rate of 1.25 Gb/s (1 Gb/s data rate) while Fibre Channel operates at a clock rate of 1 Gb/s (800 Mb/s data rate). This reuse of off-the-shelf components has contributed to the rapid development and deployment of gigabit Ethernet systems. The PMA sublayer is also responsible for recovering a clock reference from the received data stream. The expansion which results from encoding 8-bit octets into 10-bit symbols requires a signaling rate of 1.25 Gbaud at the serial interface to the medium.

The PCS includes a function referred to as *auto negotiation*, which is a link startup and initialization procedure first defined in IEEE Std 802.3u for 100BASE-T. Within IEEE Std 802.3z, auto negotiation is used to select between the CSMA/CD and full-duplex operating modes, and to select whether the Pause flow control mechanism is enabled or disabled on a link-by-link basis.

At the bottom of the diagram are the 1000BASE-SX and 1000BASE-LX fiber optic transceiver specifications. The 1000BASE-SX specification for short-wavelength laser transceivers supports multi-mode fiber optic links at distances up to 275 m using 62.5 μ fiber, and 550 m using 50 μ fiber. Once again, Fibre Channel provided the starting point for the 1000BASE-SX specification, but in addition, the 1000BASE-SX specification embraces VCSELs as well as the older CD style of laser developed for the Fibre Channel market.

1000BASE-LX supports longer distances using higher-cost components, spanning 550 m on 62.5 μ or 50 μ fiber, and up to 5 km on single-mode fiber. The 1000BASE-LX laser transmitter is optimized for single-mode fiber, and requires a mode-conditioning patch cord to support multimode fiber optic cable. The patch cord mitigates an effect known as differential mode delay by altering the launch characteristics of a 1000BASE-LX laser transmitter so that the resulting optical beam more closely resembles the overfilled launch pattern produced by a light-emitting diode (LED). Multimode fiber modal bandwidth is specified under the launch conditions produced by an LED. Through use of the mode-conditioning patch cord, a 1000BASE-LX link can be characterized on the basis of the rated modal bandwidth of a multimode fiber. Both 1000BASE-SX and 1000BASE-LX specify the familiar duplex SC optical connector, eliminating the most common installation problem encountered in fiber optic networks, the misconnection of the transmitting and receiving fibers.

IEEE Std 802.3z also includes a specification for a transceiver technology referred to as 1000BASE-CX, which supports shielded copper cables links spanning 25 m. The SerDes component which makes up the PMA sublayer is designed to drive this cable directly, which makes 1000BASE-CX an economically attractive choice for short-distance interconnections, for instance, between devices located within the same rack or within a computer room or telephone closet.

A new project within the IEEE 802.3 Working Group, referred to as 1000BASE-T, is chartered to develop a PHY specification which will support 1000 Mb/s operation on four pairs of category 5 UTP cabling, at a maximum link distance of 100 m. This project is being conducted in the P802.3ab Task Force. 1000BASE-T will take advantage of recent advances in silicon process technology which permit complex high-speed digital signal processing algorithms to be implemented cost effectively. A 1000BASE-T PHY will transmit its signal on all four pairs of wire simultaneously, thus reducing the data rate on each pair to 250 Mb/s. The use of a five-level pulse amplitude modulation scheme further reduces the signaling rate on each pair. Hybrids and digital echo cancellation are used to achieve full-duplex communication.

Network administrators will actively embrace gigabit Ethernet because the familiar management attributes that have been in use in 10 and 100 Mb/s Ethernet networks have been retained. The attribute definitions have been updated to reflect the fact that statistics counters tick 10 times faster for gigabit Ethernet, and certain attributes have been extended to embrace the new physical layers, but the "look and feel" of Ethernet from the network manager's point of view has been preserved.

Biography

HOWARD FRAZIER (hfrazier@cisco.com) is employed by Cisco Systems, Inc. within the Workgroup Business Unit. He is chair of the IEEE 802.3z Gigabit Task Force, which developed the Gigabit Ethernet standard. Previously, he was chair of the IEEE 802.3u 100BASE-T Task Force, which developed the standard for Fast Ethernet. Prior to joining Cisco he was employed by Sun Microsystems, Inc. He graduated from Carnegie-Mellon University with a B.S.E.E. in 1983.

ATM FORUM/from page 7

proposed which uses the Chernoff method (which in turn uses the Markov, or Chebyshev, inequality) and assumes that local delays at switches are independent and gamma distributed [1].

Other brief discussions centered around the issues at a VC-to-VP aggregation point. One issue was how to determine the QoS of the VP relative to the QoS of the VCs. It has been documented that no general solution is known to this problem. Another issue was how VP sources should handle explicit forward congestion indication (EFCI); the problem is that source behavior 12 requires sources to reset the EFCI state. At a VC-VP boundary, this means that EFCI information from the VC control loop would be lost. The solution proposed for this is to reflect the EFCI state of the incoming data cells back to the VC source, stopping short of a full virtual source/virtual destination implementation.

A number of joint sessions have been held with network management, SAA, and RBB groups. The work with network management centered on how to count valid and invalid RM cells, and how to log traffic descriptors. The work with SAA centered around ABR API issues to provide an interface to query and set ABR parameters. Another issue was how to request SCR and MBS parameters given the mean and PCR of a video stream. The work with RBB centered around simplification of traffic parameters for residential users, and issues of shared access over asymmetric links such as cable.

Several performance contributions on GFR, IP over ATM, and VS/VD were presented in recent meetings.

References

[1] J. Kenney, Ed., "ATM Traffic Management Living List," ATM Forum LTD-TM-01.07, Apr. 1998.

[2] ATM Forum Traffic Management, *The ATM Forum Traffic Management Specification Version 4.0*, Apr. 1996; available as ftp://ftp.atmforum.com/pub/approved-specs/af-tm-0056.000.ps.

Biography

SHIVKUMAR KALYANARAMAN (shivkuma@ecse.rpi.edu) is an assistant professor at the Department of Electrical, Computer and Systems Engineering at Rensselaer Polytechnic Institute in Troy, NY. He received a B.Tech degree from the Indian Institute of Technology, Madras, India in July 1993, followed by M.S. and and Ph.D. degrees in Computer and Information Sciences at the Ohio State University in 1994 and 1997, respectively. His research interests include multimedia networking, traffic management in ATM networks and Internet, Internet pricing, and performance analysis of distributed systems. He is a co-inventor in two patents (the ERICA and OSU schemes for ATM traffic management), and has coauthored several papers and ATM forum contributions in the field of ATM traffic management. He is a member of IEEE-CS and ACM. His homepage is: http://www.ecse.rpi.edu/Homepages/shivkuma