10 Gigabit Ethernet: Bridging the LAN/WAN Divide

Maen Artimy

I. INTRODUCTION

THE introduction of 10 Gigabit Ethernet standard extends the IEEE 802.3 family of protocols to an operating speed of 10 Gbps. This provides a significant increase in bandwidth while retaining compatibility with the installed base of 802.3 networks. In addition to maintaining the principles of Local Area Networks (LANs), the new 10 Gigabit Ethernet expands the Ethernet application to include Wide Area Networks (WANs).

Specifications for 10 Gigabit Ethernet (10GbE) are published in a collection of standards beginning with IEEE 802.3ae-2002. Other standards include 802.3ak-2004, 802.3an-2006, 802.3ap-2007, and 802.3aq-2006. The 802.3ae-2002 and 802.3ak-2004 amendments have been consolidated into the IEEE 802.3-2005 standard. The other amendments will be consolidated into IEEE Std 802.3-2008, which has yet to be published.

This document provides an overview of current 10 Gigabit Ethernet technologies and their main applications.

II. ARCHITECTURE

10 Gigabit Ethernet ensures compatibility with previous generations of Ethernet by preserving the MAC interface, frame format, and frame size. However, it supports the full-duplex mode of operation only. Thus, 10 Gigabit Ethernet does not require the carrier-sensing multiple-access with collision detection (CSMA/CD) protocol for medium access.

Figure 1 shows the 10 Gigabit Ethernet architecture. 10GbE is a layer 1 and 2 technology that uses the IEEE 802.3 MAC sublayer, connected through a 10 Gigabit Media Independent Interface (XGMII) to Physical Layer entities such as 10GBASE-SR, 10GBASE-LX4, 10GBASE-CX4, 10GBASE-LRM, 10GBASE-LR, 10GBASE-ER, 10GBASE-SW, 10GBASE-LW, 10GBASE-EW, and 10GBASE-T. A rate control mode is added to the MAC to adapt the data rate to Synchronous Optical Network/Synchronous Digital Hierarchy (SONET/SDH) data rate for WAN-compatible applications.

The IEEE 802.3-2005 [1] defines two Physical Layer Device (PHY) types. LAN PHYs operate at 10 Gbps, and WAN PHYs operate at SONET STS-192c/SDH VC-4-64c rate. Their specifications allow for a nominal network extent of up to 40 km using optical fibre plant. The IEEE 802.3-2005 standard also supports operation over a twinaxial cable assembly for wiring closet and data centre applications. Support for UTP cabling plant and long multimode fibre is provided in IEEE 802.3an-2006 standard [2] and IEEE 802.3aq-2006 standard [3], receptively.

The standard specifies a family of PHY implementations. The generic term, 10 Gigabit Ethernet, refers to any use of the 10 Gbps IEEE 802.3 MAC (the 10 Gigabit Ethernet MAC) coupled with any IEEE 802.3 10GBASE physical layer implementation. The following terms refer to specific family of physical layer implementations:

- 10GBASE-X: refers to physical layer implementations based upon 8B/10B data coding method. The 10GBASE-X family includes 10GBASE-LX4 and 10GBASE-CX4.
- 10GBASE-R: refers to physical layer implementations based upon 64B/66B data coding method. The 10GBASE-R family includes 10GBASE-SR, 10GBASE-LR, and 10GBASE-ER and 10GBASE-LRM.
- 10GBASE-W: refers to physical layer implementations based upon STS-192c/SDH VC-4-64c encapsulation of 64B/66B encoded data. The 10GBASE-W family includes 10GBASE-SW, 10GBASE-LW, and 10GBASE-EW.
- 10GBASE-T: refers to a physical layer implementation based upon 64B/65B data coding placed in a low density parity check (LDPC) frame for transmission on four-pair, twisted-pair copper cabling.

All 10GBASE-R and 10GBASE-W PHY devices share a common PCS specification. The 10GBASE-W PHY devices also require the use of the WAN Interface Sublayer (WIS), which provides the capability to transmit and receive IEEE 802.3 MAC frames within the payload envelope of a SONET/SDH frame.

Table I specifies the correlation between the PHY terms and the corresponding components. The following subsections describe the functions of 10GbE sublayers.

A. Reconciliation Sublayer (RS) and 10 Gigabit Media Independent Interface (XGMII)

10 Gigabit Ethernet specifies an optional 10 Gigabit Media Independent Interface (XGMII) as a synchronous interface between the MAC and physical layers. The XGMII is a 74-signal wide interface that includes 32-bit data path, 4 control bits, and one clock signal for each direction. The data path is organized into four lanes with each lane conveying a data octet or control character on each edge of the associated clock. The XGMII supports only the 10 Gbps MAC data rate in full duplex operation. To support 10GBASE-W PHYs at the STS-192/VC-4-64c line rate of 9.95328 Gbps, inter-packet gap Idle

© 2008 Internetworking Atlantic Inc.

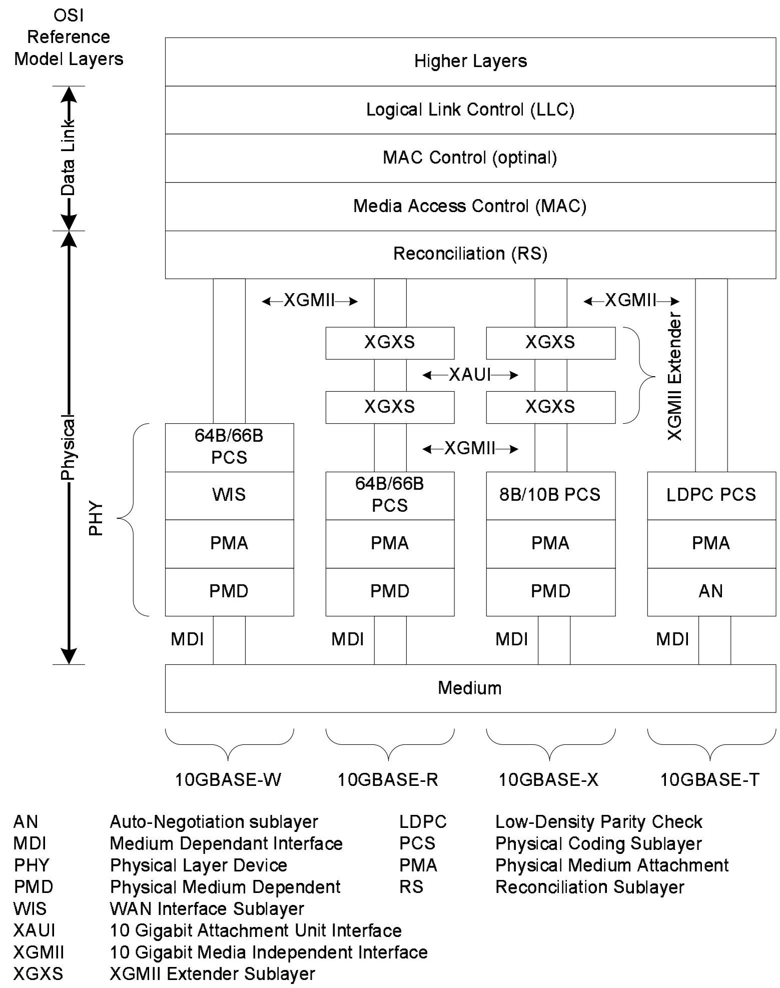

Fig. 1. The 10 Gigabit Ethernet Architecture

control characters are added.

The XGMII provides a uniform interface to the Reconciliation Sublayer (RS) for all 10 Gbps PHY implementations. The RS provides conversion services between MAC serial data stream and the parallel data paths of the XGMII. The 10 Gbps Physical Coding Sublayer (PCS) is specified to the XGMII, so if not implemented, a conforming implementation shall behave functionally as if the RS and XGMII were implemented.

The XGMII is primarily intended as a chip-to-chip (integrated circuit to integrated circuit) interface implemented on a printed circuit board or as a logical interface between ASIC logic modules within an integrated circuit. This interface provides media independence so that an identical media access controller may be used with all 10GBASE PHY types.

B. XGMII Extender Sublayer (XGXS) and 10 Gigabit Attachment Unit Interface (XAUI)

The XGMII Extender may optionally be used to extend the operational distance of the XGMII across a circuit board with reduced pin count. The XGMII Extender is comprised of an XGXS at the RS end (DTE XGXS), an XGXS at the PHY end (PHY XGXS) and a XAUI between them, as shown in Figure 1.

The source XGXS converts bytes on an XGMII lane into a self clocked, serial, 8B/10B encoded data stream. Each of the four XGMII lanes is transmitted across one of four XAUI differential signal pairs. To accommodate both data and overhead associated with the coding, the XAUI operates at rate of 3.126GBaud per lane.

The fewer, self-clocked signals of the XGMII Extender increase the XGMII's electrically limited distances from approximately 7 cm to to approximately 50 cm.

C. Physical Coding Sublayer (PCS) and Physical Medium Attachment (PMA) sublayer

The are three Physical Coding Sublayer (PCS) and Physical Medium Attachment (PMA) pairs, the 10GBASE-X PCS is attached to the 10GBASE-X PHY family, and the 10GBASE-R may be attached through the PMA sublayer to a LAN (10GBASE-R) PMD sublayer supporting a data rate of 10 Gbps or it may be attached to a WAN (10GBASE-W) PMD through the WIS and PMA sublayers. When attached to a WAN sublayer, this PCS adapts the data stream to the WAN data rate. The 10GBASE-T PCS is intended for use over balanced twisted-pair structured cabling systems. 10GBASE-T

TABLE I
MANDATORY COMPONENTS OF THE 10 GIGABIT ETHERNET PHYS

Description	850 nm PMD	1310 nm PMD	1550 nm PMD	1310 nm WDM PMD	1310 nm Serial MMF PMD	64/66B PCS & PMA	8B/10B PCS & PMA	LDPC PCS & PMA	WIS	Serial PMA	4-Lane Electrical PMD
10GBASE-SR	M					M				M	
10GBASE-SW	M					M			M	M	
10GBASE-LR		M				M				M	
10GBASE-LW		M				M			M	M	
10GBASE-ER			M			M				M	
10GBASE-EW			M			M			M	M	
10GBASE-LX4				M			M				
10GBASE-CX4							M				M
10GBASE-LRM					M	M				M	
10GBASE-T								M			

signalling requires four pairs of balanced cabling that support links of up to 100 metres.

The 10GBASE-X PCS provides synchronization and encoding/decoding functions from/to the 32 data bits and 4 control bits of the XGMII to/from four parallel lanes conveying 10-bit code-groups each (8B/10B), for communication with the underlying PMA.

The 10GBASE-R PCS provides all services required by the XGMII, including encoding/decoding of 64 data bits (from two consecutive XGMII transfers) to/from 66-bit blocks (64B/66B), transferring encoded data to/from the PMA in 16 bit transfers, and when connected to a WAN PMD, deleting/inserting idles to compensate for the rate difference between the MAC and PMD.

In 10GBASE-T PCS utilizes low density parity check (LDPC) mechanism to encode groups of 64 data bits from XGMII into a 65-bit blocks (64B/65B).

The PMA allows the PCS or WIS to support a range of serial bit-oriented physical media by performing functions such as serialization/deserialization of 16-bit data and clock transmission/recovery.

D. WAN Interface Sublayer (WIS)

The WAN Interface Sublayer (WIS) is used in a 10GBASE-W PHY to generate Ethernet data streams that may be mapped directly to SONET/SDH streams at the physical level, without requiring MAC or higher-layer processing.

A 10GBASE-W interface is not compatible with SONET or SDH interface since it does not conform to the optical, electrical, and logical requirements specified by SONET or SDH. With the exception of sustained data rate, a 10GBASE-W PHY does not appear different from a PHY without a WIS to the MAC layer. However, a 10GBASE-W interface may interoperate only with another 10GBASE-W interface.

The WIS maps the encoded Ethernet data received (transmitted) from (to) the PCS into a frame format compatible with SONET and SDH networks by implementing a minimal number of the standard SONET overhead fields and functions.

E. Physical Medium Dependent (PMD) sublayer

Several PMDs were specified to meet several media and distance objectives. A general naming convention for the PMDs is shown in Table II.

A 1310nm serial PMD is specified to meet 2km and 10km objectives over single-mode fibre (SMF). A 1550 nm serial solution is specified to meet the 40km SMF objective. An 850 nm PMD is specified to achieve a 65-meter objective over multimode fibre using serial 850 nm transceivers.

Additionally, two versions of the Wide Wavelength Division Multiplexing (WWDM) PMDs are specified, a 1310 nm version over single-mode fibre for a distance of 10km and a 1310 nm PMD for a distance of 300m over multimode fibre.

Table III lists the maximum achievable distances for each fibre PMD.

The IEEE 802.3an standard specifies 10 GbE over Unshielded Twisted-Pair (UTP) cabling, with a maximum length of 100 meters. This maximum distance can be reached over Category 6a UTP cabling. Other types of cabling can also be used, but they support shorter distances. Table IV lists the possible distance for each type of copper cable.

III. OPTICAL MODULES

In general, optical modules contain a transmitter, a receiver and a multiplexer that changes the line rate. The module can be used in any networking system where an optical fibre is terminated and the data received is converted into an electrical signal. A 10-Gig Ethernet module includes also a management interface that is used to configure the module and obtain information about its capabilities. A 10-Gig Ethernet optical module may also include some or all of the physical layer functions, including SONET framer and possibly an Ethernet MAC layer.

Most optical modules are designed to meet the specifications set by multi-source agreements (MSA) among manufactures. The MSAs cover the implementation details left out by IEEE standard such as package dimensions and pinout. These agreements are intended to establish compatible sources of a pluggable fibre optic modules. The relevant MSAs for 10GbE are XENPAK, X2, XPAK, XFP and SFP+.

XENPAK MSA was released on May 7th, 2001 and it covered all PMD types defined by the original IEEE 802.3ae. A XENPAK module can be hot-plugged through the front panel of a typical device and secured using screws. The module has an electromagnetic interference (EMI) gasket to reduce emissions. XENPAK is based on the parallel interface, XAUI, and it requires an SC duplex fibre optic connector,

TABLE II
GENERAL NAMING CONVENTION FOR 10 GIGABIT ETHERNET PHYSICAL INTERFACES

Prefix	First Suffix = Media Type	Second Suffix = PHY Encoding Type	Third Suffix =
10GBASE-	C = Copper (twinaxial) S = Short L = Long E = Extended Z = Ultra extended T = Copper (UTP)	R = LAN PHY W = WAN PHY X = LAN PHY	4 = 4 WWDM wavelengths or 4 XAUI lanes M = Multimode

TABLE III
FIBRE OPTIC CABLING MAXIMUM DISTANCES

Fibre Type	62.5μm MMF		50μm MMF		SMF		
Nominal Wavelength (nm)	850	1300 WWDM	850	1300 WWDM	1310	1550	1300 WWDM
10GBASE-S	33m		300m				
10GBASE-L					10km		
10GBASE-E						40km	
10GBASE-LX4		300m		300m			10km
10GBASE-LRM	220m		220m				

TABLE IV
COPPER CABLING TYPES AND DISTANCES

Cable type	Maximum distance
Twinax	15m
Class E / Category 6	55m to 100m
Class E / Category 6: unscreened	55m
Class E / Category 6: screened	100m
Class E_A / Augmented Category 6	100m
Class F	100m

which results in lower port density, but provides better head dissipation.

X2 MSA was released on July 31st, 2002. The MSA specifies a module that is physically 40% smaller than XENPAK, but maintains the electrical I/O specification defined by XENPAK MSA. X2 is electrically compatible with XENPAK MSA. and uses the same electrical connector as XENPAK supporting four wire XAUI. X2 is initially focused on optical links to 10 kilometres and is targeted at 10 Gbps enterprise, storage and telecom applications that do not require the thermal capacity provided by XENPAK.

Similar to X2, XPAK MSA uses 70-pin XAUI electrical interface connector. However, XPAK reduces the size and adds the capability of double-sided mounting on a printed circuit board to offer increased port densities (up to 20 devices can be installed on a single 17-inch PCB). The single-sided mounting capability of the XPAK form factor also enable PCI card applications. The MSA accommodates all optical and copper PMDs and support links of up 10 kilometres in its initial implementations.

XFP MSA specifies a hot-pluggable module that supports multiple data rates (OC-192, 10G FC, G.709, and 10 Gigabit Ethernet) using a 10 Gigabit serial electrical interface (XFI), that is not standardized by the IEEE. An early XFP module prototype was demonstrated in March, 2002. XFP allows 16 transceivers on a typical 19” rack with 23mm pitch (width) density while consuming less than 1/3 the power and size of an MSA with parallel interface. This small size is possible due to locating the majority of electronic signal processing on the host board rather than within the module.

SFP+ is a transceiver module specified by the ANSI group T11 for 8.5 Gbps and 10 Gbps Fibre Channel and Ethernet applications. Similar to 1-4 Gbps SFP modules, the SFP+ form factor is 30% smaller than the XFP form factor and moves the signal processing functions to the host board. The benefits include higher port density (48 ports), lower power modules, and lower system costs through better integration of IC functions at the host board level.

IV. APPLICATIONS

Although 10 Gigabit Ethernet can be regarded as an upgrade path for existing LAN environments, it can also present a key solution for several other applications that demand high data rates, low latency, and long reach.

A. Bandwidth aggregation

As the demand for Gigabit connectivity to the desktop rises, 10GbE will be used as a high-speed interconnection between buildings in campus environments. The technology is ideal for aggregating traffic from wiring closet switches and local servers within the building. However the technology must be economically feasible to replace the alternative solution, link aggregation of multiple 1G links using the IEEE 803.ad standard.

B. High Performance Cluster Computing (HPCC)

Clusters of tens, hundreds, or even thousands of computing nodes are becoming common technological tools for research, simulations and modelling, digital image rendering, and scientific applications. Fundamental requirements for HPCC are high speed and low latency communications. In current clustering environments, nodes have 1GbE connections to Gigabit switches with 10GbE providing interconnection between switches. As 10GbE nodes become available they will require direct connections to Ethernet switches at 10Gbps speeds and switches will be interconnected with multiple 10GbE links or the future 40 Gigabit Ethernet, IEEE 802.3ba. With its applicability to existing network cable plants, 10GBaseT is expected to become the optimal high-speed networking solution for highly consolidated enterprise computing and HPCC centres.

C. Storage Area Network (SAN)

The iSCSI protocol aims to be the dominant SAN technology replacing Fibre Channel and other, often-proprietary, fabric interconnect technologies. Although it has a wider user base, Ethernet-based iSCSI lacked the capacity to move large amounts of data. The introduction of 10GbE will improve iSCSI position by offering equivalent or superior data carrying capacity at similar latencies.

As 10GbE NICs appear in SAN servers and Network Attached Storage (NAS) that exhibit 6-7 Gbps performance, the networks should scale proportionally by adopting 10GbE infrastructure. Moreover, 10GbE can support applications such as business continuance/disaster recovery, remote back-up, and storage on demand.

D. Internet Service provider (ISP)

Some LAN PMDs, such as 10GBASE-ER, will extend metro networks over distances beyond those of 1GbE networks. The WAN PHY enables IP/Ethernet switches to be attached to the SONET/SDH infrastructure. This feature enables Ethernet to use SONET/SDH for layer-1 transport across a WAN backbone while remaining an asynchronous link protocol. As a result, two 10GbE devices connected via SONET/SDH can behave as though they are directly attached to each other over a single Ethernet link. This allows a packet to travel back and forth between fibre-optic or copper Ethernet networks and Time Division Multiplexing (TDM) networks without any re-framing or protocol conversion.

10GbE offer some advantages over exiting packet over SONET (POS) solutions. Using WAN PHY, IP packets can be sent over SONET/SDH without the need to strip off the Ethernet frame and re-encapsulate the packet into a PPP or HDLC frame. ISPs can have similar diagnostics and management features to those found in SONET networks using built-in Ethernet capabilities and Simple Network Management Protocol (SNMP). In terms of reliability, 10GbE networks can implement link aggregation to provide a layer-2 link redundancy that outperforms the POS linear Automatic Protection Switching (APS) by supporting load-balancing, diverse media, and $N + 1$ protection, albeit at a higher switching time.

E. Server consolidation

Combined with virtualization, 10GbE offers a single server the bandwidth needed to replace several servers doing different jobs. 10GbE can offer a low-latency network for Remote Direct Memory Access (RDMA) support, which is critical in the server-to-server communication typically associated with clustering and SANs.

The advantage of reaching new distances is providing organizations the ability to re-locate data centres in more cost-effective locations up to 40 km away from their campuses or support multiple campus locations within the 40 km distance. As a result, organizations will have more options for server and data centre consolidation strategies.

F. High-bandwidth applications

A 10GbE LAN can support the continuous growth of bandwidth-hungry applications such as streaming video, medical imaging, centralized applications, and high-end graphics. The bandwidth that 10 Gigabit backbones provide also enables the next generation of network applications to flourish: telemedicine, telecommuting, distance learning, digital videoconferencing, HDTV, video-on-demand, and extreme Internet gaming.

V. Conclusion

10 Gigabit Ethernet represents a natural evolution of the IEEE 802.3 standard. By maintaining interoperability with the vast, ubiquitous, base of Ethernet networks, 10GbE offers a direct migration path to higher performance levels at potentially lower total cost-of-ownership.

Nevertheless, the 10GbE wide-scale deployment has been hampered by factors such as the diversity of fibre optic modules (six MSAs were introduced in five years), their high cost, and high power consumption. Moreover, the delayed support of UTP cabling has established the perception of the 10GbE as a technology limited to point-to-point trunking between Ethernet switches.

As the price per port declines and with the advent of 10GBASE-LRM and 10GBASE-T PHYs, 10GbE is being adopted as a cost-effective technology with applications that benefit data centres, carriers, and enterprises. 10GbE is predicted to be a major force behind the Ethernet switch market growth, which will provide an annual revenue that surpasses $19 billion in 2011 [4].

References

[1] IEEE Std. 802.3-2005, IEEE Standard for Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements. Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications, 2005.

[2] IEEE Std. 802.3an-2006, Amendment 1: Physical Layer and Management Parameters for 10 Gb/s Operation, Type 10GBASE-T, 2006.

[3] IEEE Std. 802.3aq-2006, Amendment 2: Physical Layer and Management Parameters for 10 Gb/s Operation, Type 10GBASE-LRM, 2006.

[4] http://www.delloro.com/news/2007/ES020107.htm

Maen Artimy, Ph.D., is the senior network designer at Internetworking Atlantic Inc. Email: maen.artimy@internetworking-atlantic.com