40 Gigabit Ethernet and 100 Gigabit Ethernet

Technology Overview

June 2010

Authors:
John D'Ambrosia, Force10 Networks
David Law, 3COM
Mark Nowell, Cisco Systems

1. This work represents the opinions of the authors and does not necessarily represent the views of their affiliated organizations or companies.

Ethernet Alliance | 3855 SW 153rd Drive | Beaverton, OR 97006
www.ethernetalliance.org

Executive Summary

The IEEE Std 802.3ba™-2010 40 Gb/s and 100 Gb/s Ethernet amendment to the IEEE Std 802.3™-2008 Ethernet standard was approved on June 17, 2010 by the IEEE Standards Association Standards Board. The IEEE P802.3ba Task Force developed a single architecture capable of supporting both 40 and 100 Gigabit Ethernet, while producing physical layer specifications for communication across backplanes, copper cabling, multi-mode fibre, and single-mode fibre. This white paper provides an overview of 40 and 100 Gigabit Ethernet underlying technologies.

Introduction

For more than 30 years, Ethernet has evolved to meet the growing demands of packet-switched networks. It has become the unifying technology enabling communications via the Internet and other networks using Internet Protocol (IP). Due to its proven low cost, known reliability, and simplicity, the majority of today's internet traffic starts or ends on an Ethernet connection. This popularity has resulted in a complex ecosystem between carrier networks, enterprise networks, and consumers creating a symbiotic relationship between its various parts.

In 2006, the IEEE 802.3 working group formed the Higher Speed Study Group (HSSG) and found that the Ethernet ecosystem needed something faster than 10 Gigabit Ethernet. The growth in bandwidth for network aggregation applications was found to be outpacing the capabilities of networks employing link aggregation with 10 Gigabit Ethernet. As the HSSG studied the issue, it was determined that computing and network aggregation applications were growing at different rates. For the first time in the history of Ethernet, a Higher Speed Study Group determined that two new rates were needed: 40 gigabit per second for server and computing applications and 100 gigabit per second for network aggregation applications.

The IEEE P802.3ba 40 Gb/s and 100 Gb/s Ethernet Task Force was formed in January 2008 to develop a 40 Gigabit Ethernet and 100 Gigabit Ethernet draft standard. Encompassed in this effort was the development of physical layer specifications for communication across backplanes, copper cabling, multi-mode fibre, and single-mode fibre. Continued efforts by the Task Force led to the approval of the IEEE Std 802.3ba-2010 40 Gb/s and 100 Gb/s Ethernet amendment to the IEEE Std 802.3-2008 Ethernet standard on June 17, 2010 by the IEEE Standards Board.

This white paper provides an overview of the IEEE Std 802.3ba-2010 40 Gb/s and 100 Gb/s Ethernet standard and the underlying technologies.

The 40 Gigabit and 100 Gigabit Ethernet Objectives

The objectives that drove the development of this standard are listed below with a summary of the physical layer specifications provided in Table 1.

- Support full-duplex operation only
- Preserve the 802.3 / Ethernet frame format utilizing the 802.3 media access controller (MAC)
- Preserve minimum and maximum frame size of current 802.3 standard
- Support a bit error rate (BER) better than or equal to 10^{-12} at the MAC/ physical layer service interface
- Provide appropriate support for optical transport network (OTN)
- Support a MAC data rate of 40 gigabit per second
- Provide physical layer specifications which support 40 gigabit per second operation over:
  - at least 10km on single mode fibre (SMF)
  - at least 100m on OM3 multi-mode fibre (MMF)
  - at least 7m over a copper cable assembly
  - at least 1m over a backplane
- Support a MAC data rate of 100 gigabit per second
- Provide physical layer specifications which support 100 gigabit per second operation over:
  - at least 40km on SMF
  - at least 10km on SMF
  - at least 100m on OM3 MMF
  - at least 7m over a copper cable assembly

	40 Gigabit Ethernet	100 Gigabit Ethernet
At least 1m backplane	✓	
At least 7m copper cable	✓	✓
At least 100m OM3 MMF	✓	✓
At least 150m OM4 MMF	✓	✓
At least 10km SMF	✓	✓
At least 40km MMF		✓

Table 1 - Summary of Physical Layer Specifications for IEEE 802.3ba

The 40 Gigabit Ethernet and 100 Gigabit Ethernet Architecture

The IEEE Std 802.3ba-2010 amendment specifies a single architecture, shown in Figure 1, that accommodates 40 Gigabit Ethernet and 100 Gigabit Ethernet and all of the physical layer specifications under development. The MAC layer, which corresponds to Layer 2 of the OSI model, is connected to the media (optical or copper) by an Ethernet PHY device, which corresponds to Layer 1 of the OSI model. The PHY device consists of a physical medium dependent (PMD) sublayer, a physical medium attachment (PMA) sublayer, and a physical coding sublayer (PCS). The backplane and copper cabling PHYs also include an auto-negotiation (AN) sublayer and a forward error correction (FEC) sublayer.

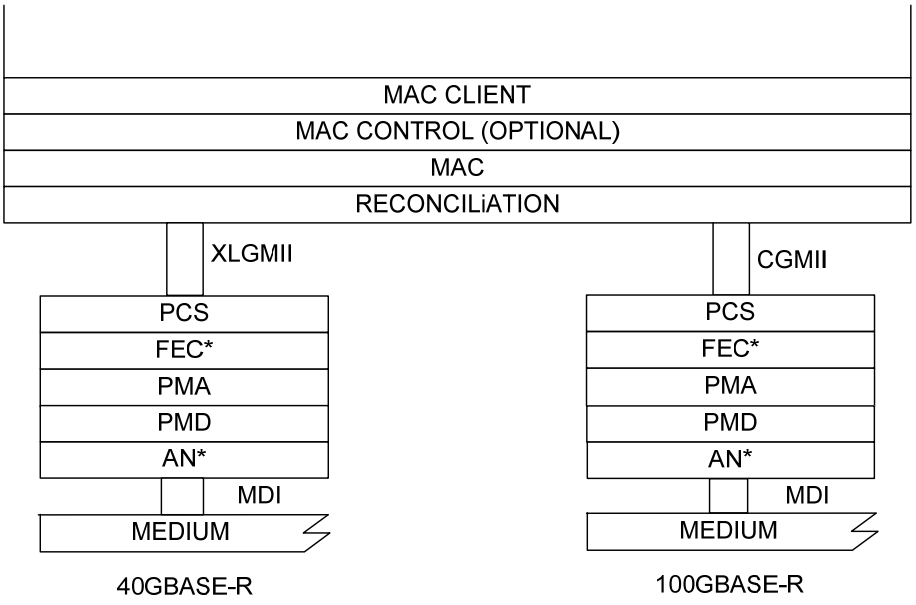

Figure 1 - IEEE 802.3ba Architecture

The Physical Coding Sublayer (PCS)

As shown in Figure 1, the PCS translates between the respective media independent interface (MII) for each rate and the PMA sublayer. The PCS is responsible for the encoding of data bits into code groups for transmission via the PMA and the subsequent decoding of these code groups from the PMA. The Task Force developed a low-overhead multilane distribution scheme for the PCS for 40 Gigabit Ethernet and 100 Gigabit Ethernet.

This scheme has been designed to support all PHY types for both 40 Gigabit Ethernet and 100 Gigabit Ethernet. It is flexible and scalable, and will support any future PHY types that may be developed, based on future advances in electrical and optical transmission. The PCS layer also performs the following functions:

- Delineation of frames
- Transport of control signals
- Ensures necessary clock transition density needed by the physical optical and electrical technology
- Stripes and re-assembles the information across multiple lanes

The PCS leverages the 64B/66B coding scheme that was used in 10 Gigabit Ethernet. It provides a number of useful properties including low overhead and sufficient code space to support necessary code words, consistent with 10 Gigabit Ethernet.

The multilane distribution scheme developed for the PCS is fundamentally based on a striping of the 66-bit blocks across multiple lanes. The mapping of the lanes to the physical electrical and optical channels that will be used in any implementation is complicated by the fact that the two sets of interfaces are not necessarily coupled. Technology development for either a chip interface or an optical interface is not always tied together. Therefore, it was necessary to develop an architecture that would enable the decoupling between the evolution of the optical interface widths and the evolution of the electrical interface widths.

The transmit PCS, therefore, performs the initial 64B/66B encoding and scrambling on the aggregate channel (40 or 100 gigabits per second) before distributing 66-bit block in a round robin basis across the multiple lanes, referred to as “PCS Lanes,” as illustrated in Figure 2.

The number of PCS lanes needed is the least common multiple of the expected widths of optical and electrical interfaces. For 100 Gigabit Ethernet, 20 PCS lanes have been chosen. The number of electrical or optical interface widths supportable in this architecture is equivalent to the number of factors of the total PCS lanes. Therefore, 20 PCS lanes support interface widths of 1, 2, 4, 5, 10 and 20 channels or wavelengths. For 40 Gigabit Ethernet 4 PCS lanes support interface widths of 1, 2, and 4 channels or wavelengths.

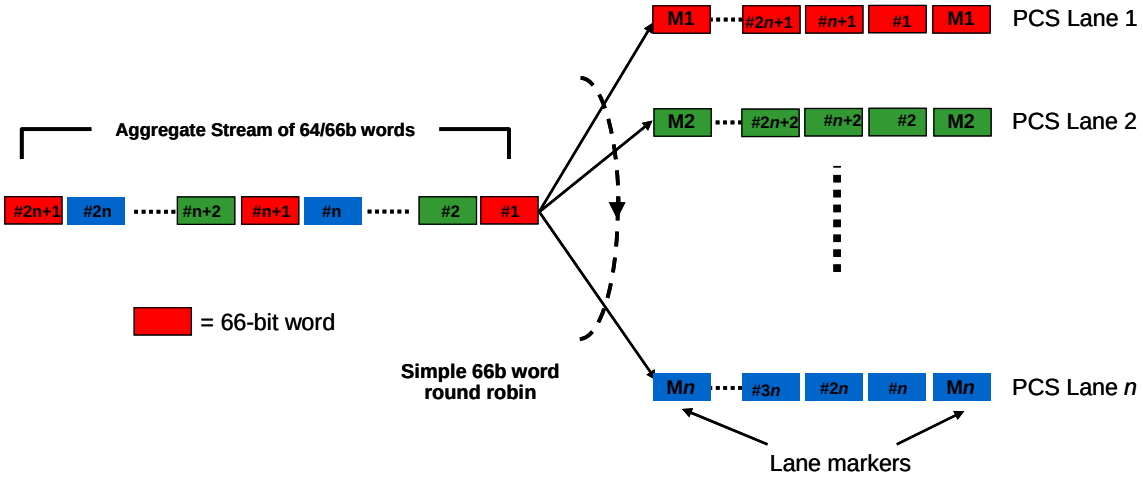

Figure 2 – PCS Multilane Distribution Concept

Once the PCS lanes are created they can then be multiplexed into any of the supportable interface widths. Each PCS lane has a unique lane marker, which is inserted once every 16,384 blocks. All multiplexing is done at the bit-level. The round-robin bit-level multiplexing can result in multiple PCS lanes being multiplexed into the same physical channel. The unique property of the PCS lanes is that no matter how they are multiplexed together, all bits from the same PCS lane follow the same physical path, regardless of the width of the physical interface. This enables the receiver to be able to correctly re-assemble the aggregate channel by first de-multiplexing the bits to re-assemble the PCS lane and then re-align the PCS lanes to compensate for any skew. The unique lane marker also enables the de-skew operation in the receiver. Bandwidth for these lane markers is created by periodically deleting inter-packet gaps (IPG). These alignment blocks are also shown in Figure 2.

The receiver PCS realigns multiple PCS lanes using the embedded lane markers and then re-orders the lanes into their original order to reconstruct the aggregate signal.

Two key advantages of the PCS multilane distribution methodology are that all the encoding, scrambling and de-skew functions can all be implemented in a CMOS device (which is expected to reside on the host device), and minimal processing of the data bits (other than bit muxing) happens in the high speed electronics embedded with an optical module. This will simplify the functionality and ultimately lower the costs of these high-speed optical interfaces.

The PMA sublayer enables the interconnection between the PCS and any type of PMD sublayer. A PMA sublayer will also reside on either side of a retimed interface, referred to as “XLAUI” (40 gigabit per second attachment unit interface) for 40 Gigabit Ethernet or “CAUI” (100 gigabit per second attachment unit interface) for 100 Gigabit Ethernet.

Figure 3 illustrates the general architecture for 100 Gigabit Ethernet, as well as examples of two other architectural implementations:

- 100GBASE-LR4, which is defined as 4 wavelengths at 25 gigabit per second per wavelength on single-mode fiber.
- 100GBASE-SR10, which is defined as 10 wavelengths across 10 parallel fiber paths at 10 gigabit per second on multi-mode fiber.

These two implementations will be used to illustrate the flexibility needed to support the multiple PMDs being developed for 40 and 100 Gigabit Ethernet.

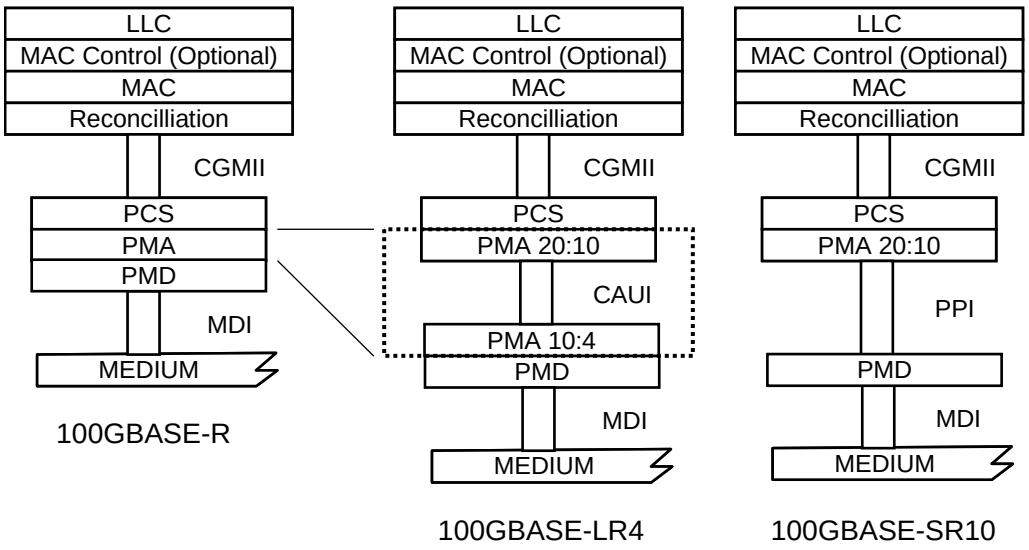

Figure 3 - Illustrations of 100GBASE-R Architectures

As described in the previous section, 20 PCS lanes are used for 100 Gigabit Ethernet. In the example implementation shown in the middle diagram of Figure 3, the two PMA sublayers are interconnected by the CAUI electrical interface., which is based on a 10 lane wide interface at 10 gigabit per second per lane. In this implementation the PMA sublayer at the top of the CAUI multiplexes the 20 PCS lanes into 10 physical lanes. The PMA sublayer at the bottom of the CAUI performs three functions. First, it retimes the incoming electrical signals. After the retiming the electrical lanes are then converted back to 20 PCS lanes, which are then multiplexed into the 4 lanes needed for the 100GBASE-LR PMD.

The implementation of the 100GBASE-SR10 architecture, however, is different. In this implementation a host chip is directly connected to an optical transceiver that is hooked up to 10 parallel fiber paths in each direction. The PMA sublayer resides in the same device as the PCS sublayer, and multiplexes the 20 PCS lanes into the ten electrical lanes of the parallel physical interface (PPI), which is the non-retimed electrical interface that connects the PMA to the PMD.

In summary, the high level PMA functionality of multiplexing still exists but the actual implementation is dependent on the specific PMD being used.

40 Gigabit Ethernet and 100 Gigabit Ethernet Interfaces

The various chip interfaces in the IEEE Std 802.3ba-2010 amendment are illustrated in Figure 3. The IEEE Std 802.3ba-2010 amendment specifies some interfaces as logical, intra-chip, interfaces, as opposed to a physical, inter-chip, interfaces as they have been in the past. A logical interface specification only describes the signals and their behavior. A physical interface specification also describes the electrical and timing parameters of the signals.

The inclusion of logical interfaces supports system on a chip (SoC) implementations where various cores, implementing the different sublayers, are supplied by different vendors. The provision of an open interface specification through the IEEE Std 802.3ba-2010 amendment will help integrate these cores into a SoC in the same way that chips from different vendors can be integrated to build a system. While a physical interface specification is sufficient to specify a logical interface, there are cases where the interfaces are unlikely to ever be implemented as a physical interface, making the provision of electrical and timing parameters unnecessary.

There are three defined chip interfaces that have a common architecture for both speeds. The MII is a logical interface that connects the MAC to a PHY and the AUI is a physical interface that extends the connection between the PCS and the PMA. The naming of these interfaces follows the convention found in 10 Gigabit Ethernet, IEEE Std 802.3ae, where the 'X' in XAUI and XGMII represents the Roman numeral 10. Since the Roman numerals for 40 are 'XL' and the Roman numeral for 100 is 'C', the same convention yields XLAUI and XGMII for 40 gigabit per second and CAUI and CGMII for 100 gigabit per second. The final interface is the parallel physical interface (PPI), discussed in further detail below, which is the physical interface for the connection between the PMA and the PMD for 40GBASE-SR4 and 100GBASE-SR10 PMDs.

40 Gigabit Attachment Unit Interface (XLAUI) and 100 Gigabit Attachment Unit Interface (CAUI)

The XLAUI, which supports the 40 gigabit per second data rate, and CAUI, which supports the 100 gigabit per second data rate, are low pin count physical interfaces that enables partitioning between the MAC and sublayers associated with the PHY in a similar way to XAUI in 10 Gigabit Ethernet. They are self-clocked, multi-lane, serial links utilizing 64B/66B encoding. Each lane operates at an effective data rate of 10 gigabit per second, which when 64B/66B encoded, results in a signaling rate of 10.3125 gigabaud per second.

The lanes utilize low-swing AC-coupled balanced differential signaling to support a distance of approximately 25 cm. In the case of XLAUI, there are four transmit and four receive lanes of 10 gigabit per second, resulting in a total of 8 pairs, or 16 signals. In the case of CAUI, there are 10 transmit lanes and 10 receive lanes of 10 gigabit per second, resulting in a total of 20 pairs or 40 signals.

These interfaces can serve as chip to chip interfaces. For example, they are used to partition system design between the largely digital-based system chip and more analog-based portions of the PHY chip, which are often based on different technology. In addition, while there is no mechanical connector specified for

XLAUI and CAUI in the IEEE 802.3ba amendment; these interfaces have also been specified for chip-to-module applications for pluggable form factor specifications, enabling a single host system to support the various PHY types through pluggable modules. The pluggable form factor specifications themselves are beyond the scope of IEEE 802.3 and are being developed by other industry organizations.

Parallel Physical Interface (PPI)

The PPI is a physical interface for short distances between the PMA and PMD sub-layers. It is common to both 40 Gigabit Ethernet and 100 Gigabit Ethernet; the only differentiation is the number of lanes. The PPI is a self-clocked, multi-lane, serial links, utilizing 64B/66B encoding. Each lane operates at an effective data rate of 10 gigabit per second, which when 64B/66B encoded results in a signaling rate of 10.3125 gigabaud per second. In the case of the 40 Gigabit Ethernet, there are 4 transmit and 4 receive lanes of 10 gigabit per second, in the case of the 100 Gigabit Ethernet there are 10 transmit lanes and 10 receive lanes of 10 gigabit per second.

Physical Media Dependent (PMD)

Different physical layer specifications for computing and network aggregation applications have been developed. For computing applications, physical layer solutions will cover distances inside the data center for up to 100m for a full range of server form factors, including blade, rack, and pedestal configurations. For network aggregation applications, the physical layer solutions include distances and media appropriate for data center networking, as well as service provider inter-connection for intra-office and inter-office applications. A summary of the physical layer specifications being developed for each MAC rate is shown in Table 2.

	40 Gigabit Ethernet	100 Gigabit Ethernet
At least 1m backplane	40GBASE-KR4	
At least 7m copper cable	40GBASE-CR4	100GBASE-CR10
At least 100m OM3 MMF	40GBASE-SR4	100GBASE-SR10
At least 150m OM4 MMF	40GBASE-SR4	100GBASE-SR10
At least 10km SMF	40GBASE-LR4	100GBASE-LR4
At least 40km SMF		100GBASE-ER4

Table 2 - IEEE 802.3ba Physical Layer Specifications

BASE-CR and 40BASE-KR4 Physical Layer Specifications

The 40GBASE-KR4 PMD supports backplane transmission, while the 40GBASE-CR4 and 100GBASE-CR10 PMD support transmission across copper cable assemblies. All three of the PHYs leverage the Backplane Ethernet 10GBASE-KR architecture, developed channel requirements and PMD.

The architecture for the PHY types is shown in Figure 4. All three PHYs use the standard 40GBASE-R and 100GBASE-R PCS and PMA sublayers. The BASE-CR and 40GBASE-KR4 PHYs also include an auto-negotiation (AN) sublayer and an optional FEC sublayer.

The BASE-CR and 40GBASE-KR4 specifications also leverage the channel development efforts of the Backplane Ethernet project. The channel specifications for 10GBASE-KR were developed to ensure robust transmission at 10 gigabit per second. The 40 Gigabit Ethernet and 100 Gigabit Ethernet PHYs apply these channel characteristics to 4 and 10 lane solutions. The BASE-CR specifications also leverage the cable assembly specifications developed in support of 10GBASE-CX4. For 40GBASE-CR4, two connectors have been selected: the QSFP+ connector, which will support a module footprint that can support either copper-based or gigabit per second optic-based modules. The 10GBASE-CX4 connector has also been selected, which will enable an upgrade path for those applications that are already invested in 10GBASE-CX4

The effective data rate per lane is 10 gigabit per second, which when 64B/66B encoded results in a signaling rate of 10.3125 gigabaud per second. Thus, the 40GBASE-KR4 and 40GASE-CR4 PMDs support transmission of 40 Gigabit Ethernet over 4 differential pair in each direction over either a backplane or twin axial copper cabling medium, while the 100GBASE-CR10 PMD will support the transmission of 100 Gigabit Ethernet over 10 differential pair in each direction for at least 7m over a twin axial copper cable assembly.

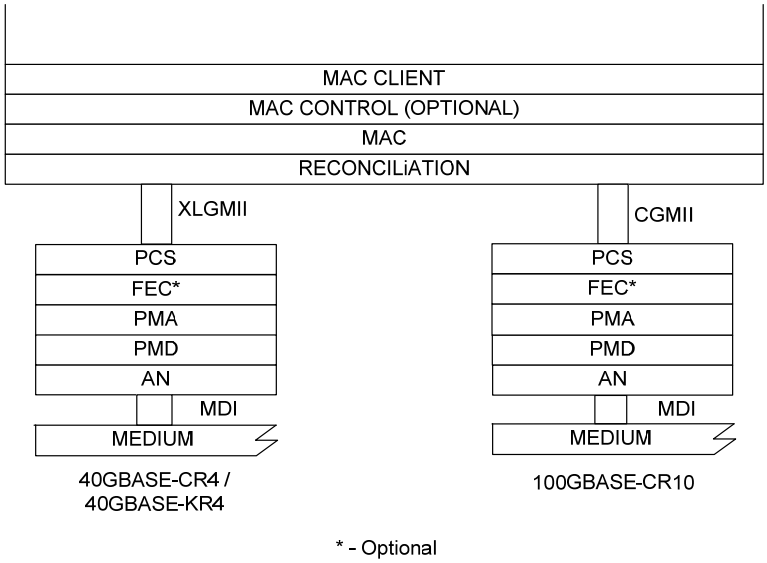

Figure 4 – Backplane and Copper Cable Architecture

BASE-SR, BASE-LR, and BASE-ER Physical Layer Specifications

All of the optical PMDs share the common architecture shown in Figure 5. While they share a common architecture, the PMA sublayer plays a key role in transmitting and receiving the number of PCS lanes from the PCS sublayer to the appropriate number of physical lanes that are required per the PMD sublayer and medium.

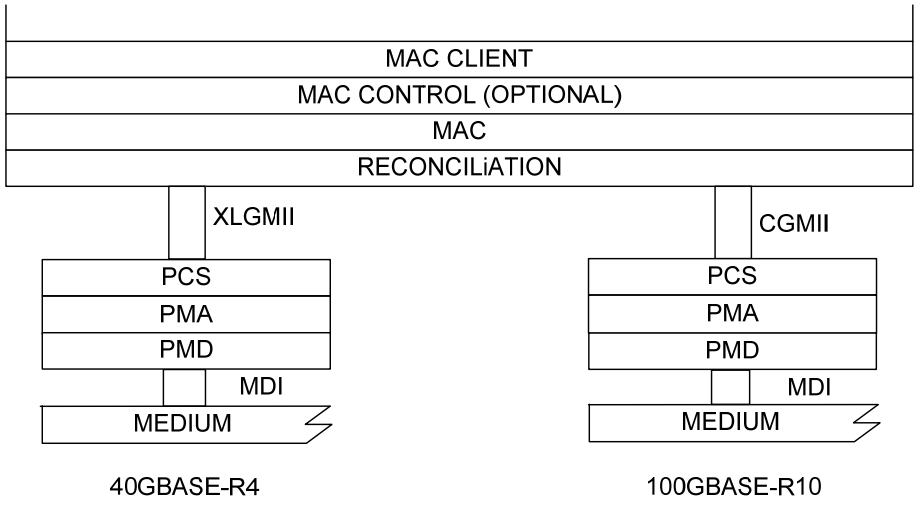

Figure 5 – 40GBASE-R and 100GBASE-SR Architecture

The different optical PMDs are:

- 40GBASE-SR4 and 100GBASE-SR10 PMD - based on 850nm technology and supports transmission over at least 100m OM3 parallel fibers and at least 150m OM4 parallel fibers. The effective rate per lane is 10 gigabit per second, which when 64B/66B encoded results in a signaling rate of 10.3125 gigabaud per second. Therefore, 40GBASE-SR4 supports transmission of 40 Gigabit Ethernet over a parallel fibre medium consisting of 4 parallel OM3 fibers in each direction, while the 100GBASE-SR10 PMD supports the transmission of 100 Gigabit Ethernet over a parallel fibre medium consisting of 10 parallel OM3 fibers in each direction.
- 40GBASE-LR4 - based on 1310nm, coarse wave division multiplexing (CWDM) technology and supports transmission over at least 10km on SMF. The grid is based on the ITU G.694.2 specification using wavelengths of 1270, 1290, 1310, and 1330nm. The effective data rate per lambda is 10 gigabit per second, which when 64B/66B encoded results in a signaling rate of 10.3125 gigabaud per second. This will help provide maximum re-use of existing 10G PMD technology. In this way, the 40GBASE-LR4 PMD supports transmission of 40 Gigabit Ethernet over 4 wavelengths on each SMF in each direction.
- 100GBASE-LR4 - based on 1310nm, dense wave division multiplexing (WDM) technology and supports transmission over at least 10km on SMF. The grid is based on the ITU G.694.1 specification using wavelengths of 1295, 1300, 1305, and 1310nm. The effective data rate per lambda is 25 gigabit per second, which when 64B/66B encoded results in a signaling rate of 28.78125 gigabaud per second. In this way, the 100GBASE-LR4 PMD supports transmission of 100 Gigabit Ethernet over 4 wavelengths on each SMF in each direction.

- 100GBASE-ER4 - based on 1310nm, WDM technology and supports transmission over at least 40km on SMF. The grid is based on the ITU G.694.1 specification using wavelengths of 1295, 1300, 1305, and 1310nm. The effective data rate per lambda is 25 gigabit per second, which when 64B/66B encoded results in a signaling rate of 28.78125 gigabaud per second. Therefore, the 100GBASE-LR4 PMD supports transmission of 100 Gigabit Ethernet over 4 wavelengths on each SMF in each direction. To achieve the 40km reaches, it is anticipated that implementations will include semiconductor optical amplifier (SOA) technology.

Conclusion

Ethernet has become the unifying technology enabling communications via the Internet and other networks using IP. Its popularity has resulted in a complex ecosystem between carrier networks, data centers, enterprise networks, and consumers with a symbiotic relationship between the various parts.

While symbiotic in nature, the different applications in the Ethernet ecosystem are growing at different rates: server and computing applications are growing at a slower pace than network aggregation applications. This divergence in growth rates spurred the introduction of two higher rates for the next generation of Ethernet: 40 Gigabit Ethernet for server and computing applications and 100 Gigabit Ethernet for network aggregation applications. This will enable Ethernet with its proven low cost, known reliability, and simplicity, to continue to evolve and be the ubiquitous connection for traffic on the Internet.

About the Ethernet Alliance

The Ethernet Alliance was formed by companies committed to the continued success and expansion of Ethernet technologies. By providing a cohesive, market-responsive, industry voice, the Ethernet Alliance helps accelerate industry adoption of existing and emerging IEEE 802 Ethernet standards. It serves as an industry resource for end users and focuses on establishing and demonstrating multi-vendor interoperability. As networks and content become further intertwined, the Ethernet Alliance works to foster collaboration between Ethernet and complementary technologies to provide a totally seamless network environment. To learn more, please go to www.ethernetalliance.org.

Glossary

- AN - Auto-negotiation
- CAUI – 100 gigabit per second Attachment Unit Interface
- CGMII – 100 gigabit per second Media Independent Interface
- CWDM – Coarse Wave Division Multiplexing
- FEC - Forward Error Correction
- HSSG – Higher Speed Study Group
- IEEE 802.3 Standard – the Ethernet Standard
- IEEE P802.3ba – the project that developed the amendment to the Ethernet Standard for 40Gb/s and 100 Gb/s Ethernet
- IEEE Std 802.3ba-2010 – the approved amendment to the Ethernet Standard for 40Gb/s and 100 Gb/s Ethernet
- IP - Internet Protocol
- MAC – Media Access Control Layer
- MDI - Medium Dependent Interface
- MII - Media Independent Interface
- MLD - Multilane Distribution
- OTN - Optical Transport Network
- PCS - Physical Coding Sublayer
- PHY - Physical Layer Devices Sublayer
- PMA - Physical Medium Attachment Sublayer
- PMD - Physical Medium Dependent Sublayer
- PPI - Parallel Physical Interface
- RS - Reconciliation Sublayer
- SoC – System on a Chip
- WDM – Wave Division Multiplexing
- XLAUI - 40 gigabit per second Attachment Unit Interface
- XLGMII – 40 gigabit per second Media Independent Interface