40 and 100 Gigabit Ethernet Overview

Abstract

This paper takes a look at the main forces that are driving Ethernet bandwidth upwards. It looks at the standards and architectural practices adopted by the different segments, how the different speeds of Ethernet are used and how its popularity has resulted in an ecosystem spanning data centers, carrier networks, enterprise networks, and consumers.

Make Your Network Mobile

Overview

There are many reasons driving the need for higher bandwidth Ethernet, however, the main reason is our insatiable appetite for content. The definition of content in itself has evolved over time – where once the majority of traffic on an Ethernet network may have been occasional file transfers, emails and the like, today technology is allowing us to push and receive richer content such as voice, video and high definition multimedia. Similarly, mechanisms for delivering content have evolved over time to reflect this demand. While there were a few technologies competing for LAN dominance in the early days of networks, Ethernet has become the clear choice. The same could not be said for WAN technologies, where TDM and ATM networks dominated for quite some time. It is only in the last few years that Ethernet is being adopted both in the core of carrier networks and more recently in the last mile.

Ethernet is a technology that is well understood by network managers as it is affordable, reliable, simple and scalable. For these reasons and others, most traffic generated in the world today likely starts and ends on Ethernet – and because of this you could say that today, Ethernet is everywhere!

This paper takes a look at the main forces that are driving Ethernet bandwidth upwards. It looks at the standards and architectural practices adopted by the different segments, how the different speeds of Ethernet are used and how its popularity has resulted in a complex ecosystem between carrier networks, enterprise networks, and consumers.

Driving the Need for Speed

Ethernet in the Enterprise and Data Center

Data center virtualization, which includes storage and server virtualization, is all about the efficient use of resources. In the data center this is multifaceted. On the one hand data center managers are trying to bring power, cooling and space utilization under control, while on the other hand they are trying to maximize the use of computing, networking and storage.

Even though computing and storage resources have been separable for some time now, there have been a couple of serious contenders for connectivity options, namely Fibre Channel and iSCSI. While iSCSI can be deployed over the same Ethernet infrastructure as computing resources, Fibre Channel requires that network managers create a dedicated network to connect their computing and storage solutions. Initial iSCSI technology allowed for connectivity of 1 Gbps. This was then dwarfed by Fibre Channel which at the time provided speeds of 4 Gbps and then later 8 Gbps. For some time Fibre Channel was the preferred medium, with servers being connected to 1 Gbps Ethernet and 4/8 Gbps Fibre Channel. That was true, at least until 10 GbE was introduced. (See Figure 1)

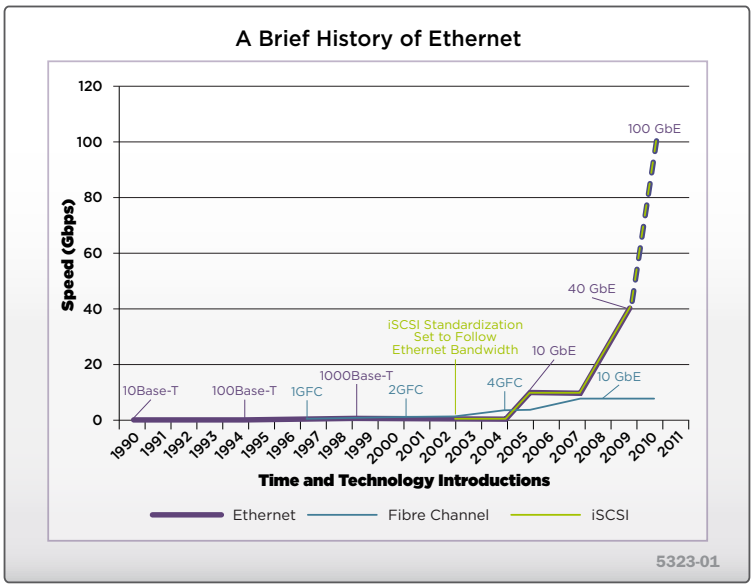

Figure 1

As server technology evolves, the number of servers per rack is increasing. Blade servers are an example of this, where a server chassis containing 8 or more servers is capable of pushing out over 10 Gbps of traffic to the network. To add to this, advances in virtualization are increasing the density of virtual servers per physical server creating yet more traffic on the wire. As a result network managers are having to migrate to 10 GbE server interfaces. This in itself is an order of magnitude increase in traffic at the edge of the network and so logically the core and aggregation layers in the network must scale proportionally.

Now that 10 GbE is becoming more mainstream, data center network managers are reconsidering its use for not just LAN connectivity but also the integration of SAN traffic over the same physical Ethernet network. This transition not only lowers the number of physical connections required in the network by a factor of 10, but also reduces overall complexity and cost. By using iSCSI over Ethernet, a dedicated Fibre Channel network is often no longer needed.

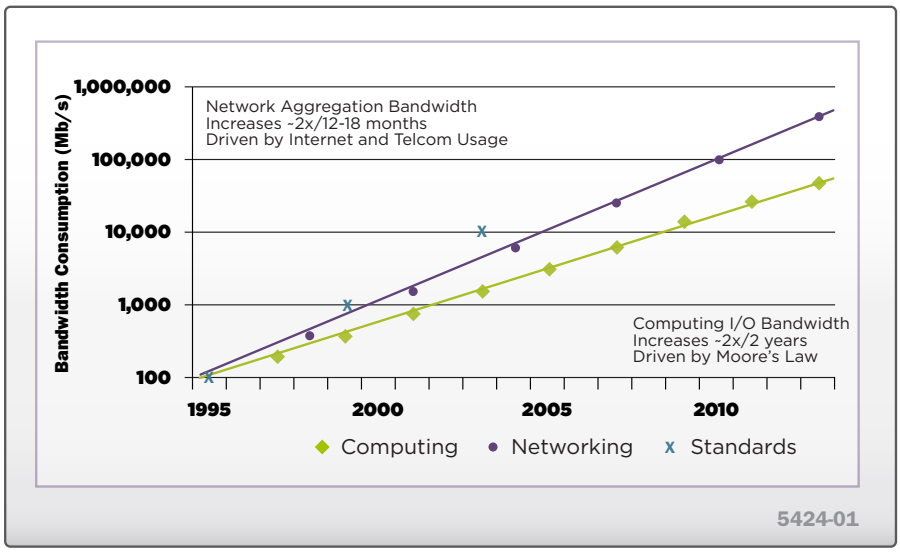

Figure 2

Bandwidth Trends

With performance on a Moore's Law curve (see Figure 2[2]) and therefore doubling approximately every 24 months[1], 40 GbE will be the next speed for network aggregation. The IEEE Higher Speed Study Group (HSSG), which was formed to analyze the requirements for next generation Ethernet, discovered that there was clear divergence between the bandwidth growth rates of computing applications and network aggregation bandwidth.

Given these trends, 10 GbE will likely become commonplace at the network edge with 40 GbE in the aggregation layer (see Figure 3). 100 GbE will, at least for the moment, find its place primarily as a WAN medium, due mostly to the expense of the required optics which would be prohibitive for data center and enterprise applications. As the cost of deploying 100 GbE declines, it will find its way into the core of the Data Center as inter-switch links, and for providing inter-building and inter-campus connections for enterprise networks.

By implementing 40 GbE in the core today (see Figure 3), network managers stand to realize further reductions in cost, configuration and cabling complexity, where a 40 GbE implementation, for example, would reduce core cabling by as much as a quarter. Today data center managers are looking to reduce oversubscription ratios in the network layers, choosing (ideally) to have a non-blocking architecture in the core and around 2.4:1 in the aggregation layers. The right switching architecture coupled with 40 GbE will help achieve the desired results.

[1] Ethernet Alliance; Overview of Requirements and Applications for 40 Gigabit and 100 Gigabit Ethernet.
[2] Ibid.

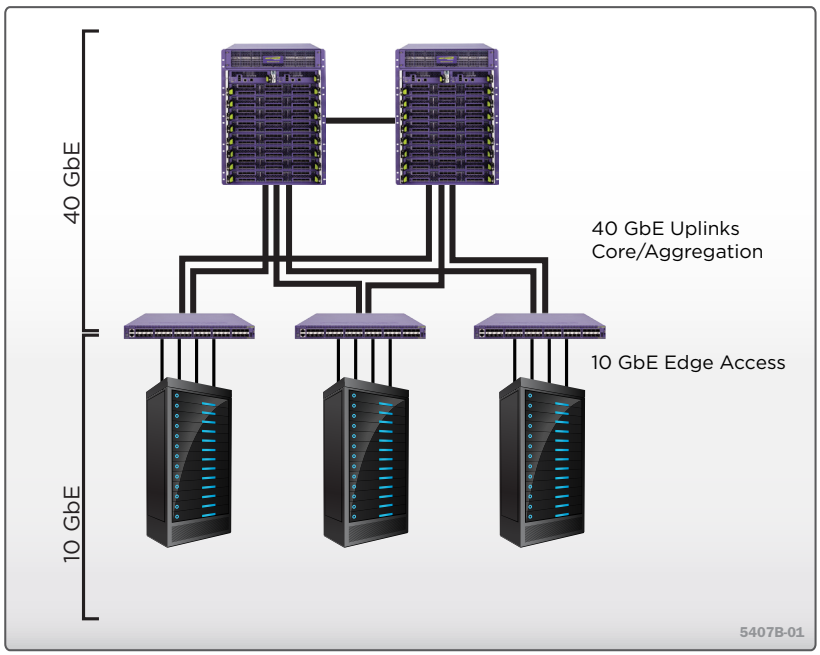

Figure 3

40 GbE or 10 GbE LAG?

As discussed in more detail later in this paper, 40 GbE and 100 GbE are based on the 10 GbE specifications – consisting of 4 or 10 lanes of 10 GbE. Given this, data center managers are faced with two options: take the plunge and implement 40 GbE and/or re-use existing 10 GbE equipment and implement a Link Aggregation Group (LAG) of four 10 GbE circuits. There is really no right or wrong here; however the latter does carry extra overhead and will not provide the same raw data throughput as a single 40 GbE link. Further, and as mentioned earlier, the cabling and switching equipment required for a single 40 GbE circuit is a quarter of a 4 x 10 GbE configuration. On the flipside a LAG implementation of 4 x 10 GbE does make provisions for resilience, depending on configuration.

40 GbE LAG Switching

Today manufacturers are designing core and aggregation switches with 40 GbE and 100 GbE in mind and while proprietary stacking methodologies are usually capable of handling much higher capabilities, there still remains the need to have greater switching capabilities for connecting Top-of-Rack (ToR) switches together where stacking cabling cannot be used. One method of implementation is to LAG 2 or more 40 GbE circuits together redundantly as seen in Figure 4 below. Following this methodology will probably (for most applications) provide adequate bandwidth for traffic at the aggregation layer.

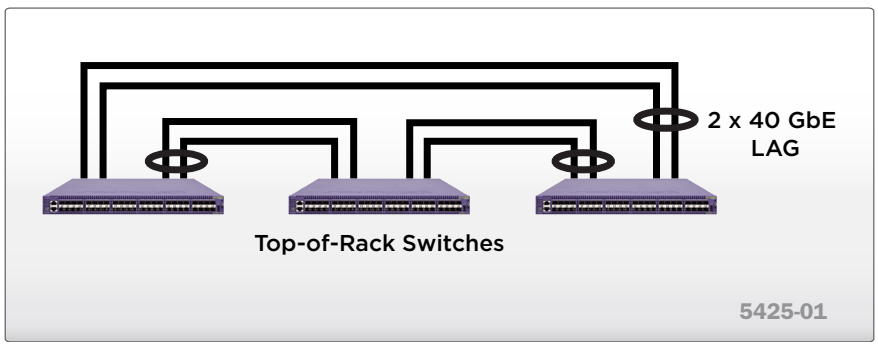

Figure 4

Ethernet in Carrier, Operator and Service Provider Networks

One of the biggest challenges that carriers face is, if bandwidth demands continue to grow faster than Moore's law! Carriers and service providers alike are generally seeing a doubling of bandwidth across their backbones approximately every 12 to 18 months. Much like the enterprise and data center evolution, this increase in bandwidth consumption is attributed to a number of factors.

- More and more people are gaining access to the Internet as both availability of Internet connections increases and the price of access falls.
- The profile of traffic over the Internet has evolved to include more media rich content such as IPTV and HDTV, video conferencing, video on demand, digital photography, telephony and more.
- Cloud computing is increasing in popularity as applications become decentralized.
- Businesses and consumers are realizing the benefits of remote storage and backups as a means of providing resilience for peace of mind and even cost savings.
- Net books, tablets, and smart phones are increasing the load on mobile broadband networks as demand for mobile Internet access increases beyond the occasional email to more into rich content.

As good as this might sound for business, this order of magnitude increase in demand has a compounding effect and is causing serious SLA challenges regarding how to handle aggregation and backbone bandwidth. Internet Service Providers (ISPs), Multiple System Operators (MSOs), and Content Delivery Network Providers (CDNs) all see sustained growth in bandwidth demands. Many are being forced to migrate away from the ailing ATM/TDM infrastructures to Ethernet. Their issue is not whether they have to migrate, but whether they should migrate to 10, 40 or 100 GbE in the backbone.

By providing the ability to increase bandwidth by up to a factor of ten carriers, operators, ISPs and others have turned to Ethernet technology once again (see Figure 5) to provide the balance between cost effectiveness and performance.

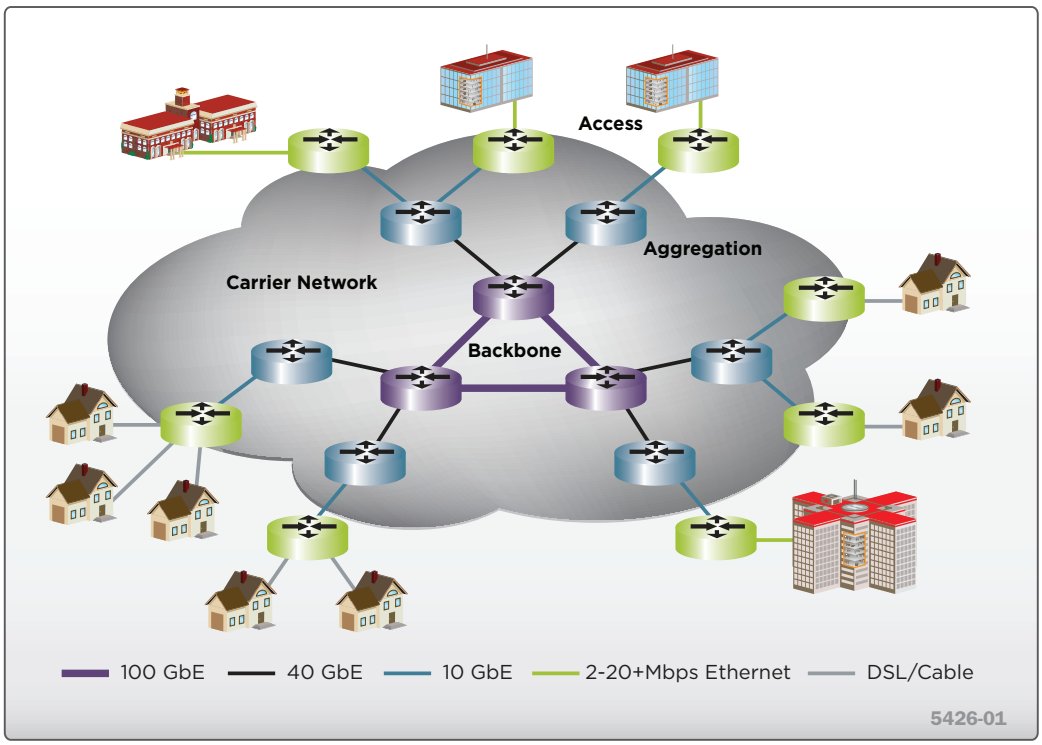

Figure 5

The 40/100 GbE Architecture

The underlying architecture for both 40 GbE and 100 GbE is based on the 10 GbE architecture; however, in order to obtain the higher data rates, the IEEE P802.3ba task force developed a low overhead scheme referred to as 'Multilane Distribution' (MLD) which essentially consists of a number of parallel links or lanes. This Physical Coding Sublayer (PCS) supports both 40 GbE and 100 GbE PHY types and provides a platform for future PHYs.

The PCS allocates traffic into the individual lanes and leverages the same 64B/66B encoding used in 10 Gigabit Ethernet, where each 66-bit word is distributed in a round robin basis into the individual lanes (See Figure 6[3]). The lanes are then fed to the PHY or physical interface for transmission through the appropriate medium.

For a 40 GbE circuit, the PCS would divide the bit stream that has been passed down from the MAC controller into 4 PCS Lanes each consisting of 10 Gbps streams. These Lanes are then handed down to the Physical Medium Attachment or PMA layer. The PMA contains functions for transmission and reception of the lanes.

Tables 1 and 2 in the appendix describe the different media and architectures that have been or are under development for 40 GbE and 100 GbE. Depending on the PMA type, the individual Lanes are transmitted either over copper or fiber pairs or passed through DWDM optics that would transmit each lane over four or more optical wavelengths. For example:

40GBASE-SR4 consists of 4 fibers in each direction for TX and RX each carrying a 10 Gbps signal.

40GBASE-LR4 describes a 40 GbE circuit over Single Mode Fiber (SMF) which would consist of a single fiber in each direction for TX and RX each carrying 4 wavelengths of 10 Gbps.

100 GbE is simply a continuation of this principle consisting of 10 fibers in each direction or 10 wavelengths in each direction.

100BASE-LR4 and ER4, however, are a little different. Here four lanes of TX and RX are used, each containing a bit stream of 25 Gbps. This new standard allows for fewer optical devices or wavelengths and simpler fiber assemblies resulting in reduced costs for both components.

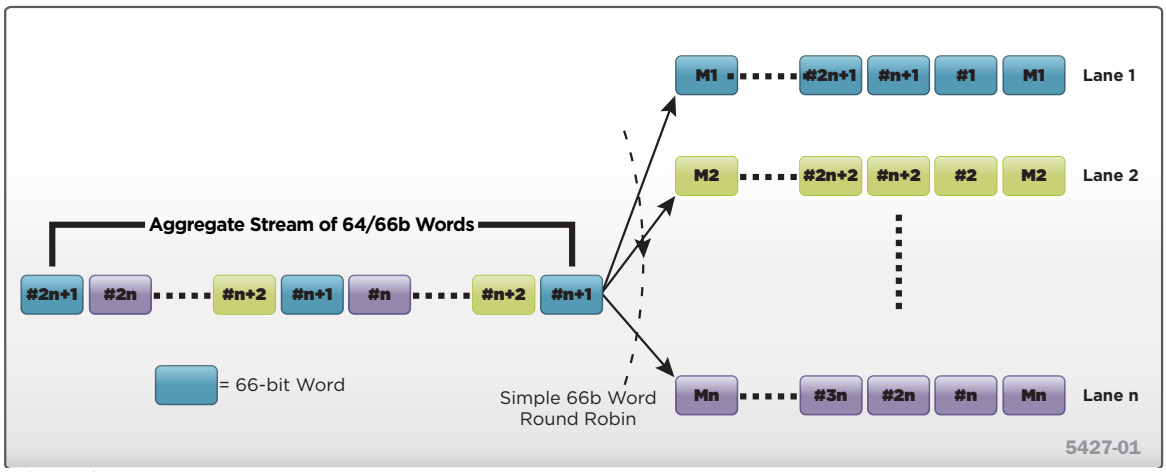

Figure 6

[3] Ibid.

Conclusion

Ethernet's success to date has been attributed to its affordability, reliability, simplicity, and scalability. As a result, most networking traffic today begins and ends its journey on Ethernet.

The advent of both 40 GbE and 100 GbE have underlined Ethernet's success by providing it with a roadmap for years to come in the enterprise, the data center and in an increasing number of applications in carrier networks.

Extreme Networks® has made the transition from 10 GbE to high density 40 GbE simple and cost effective with the BlackDiamond® and Summit® series of switches. Provisions have been made on new products such as the BlackDiamond X8 for high density and cost effective 100 GbE without the need for a forklift upgrade to the infrastructure.

Further Reading

- Overview of Requirements and Applications for 40 Gigabit Ethernet and 100 Gigabit Ethernet Technology Overview White Paper (Archived 2009-08-01) – Ethernet Alliance
- 40 Gigabit Ethernet and 100 Gigabit Ethernet Technology Overview White Paper – Ethernet Alliance
- IEEE 802.3 standards processes: http://www.ieee802.org/

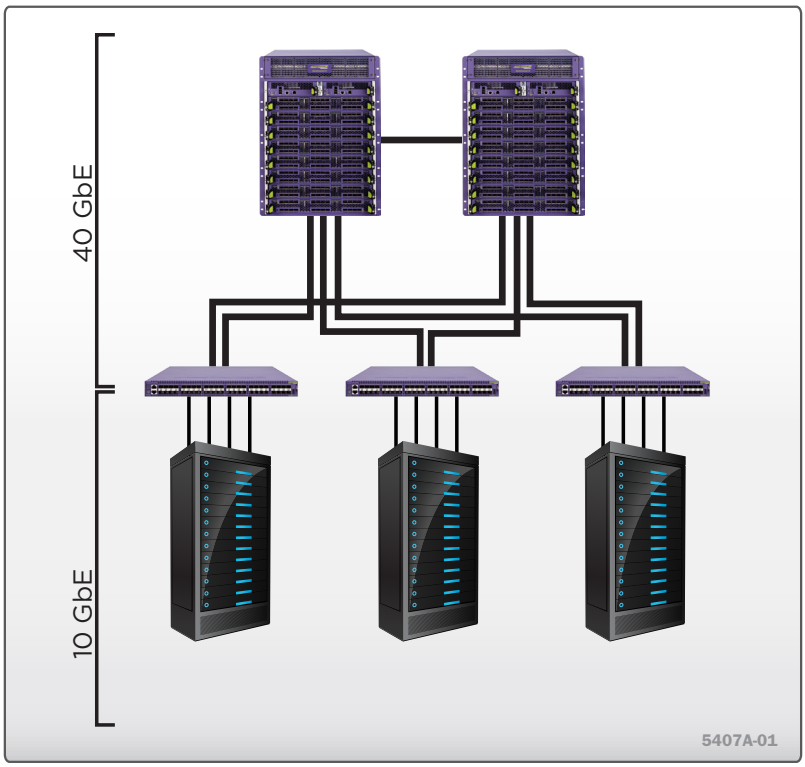

Figure 7

Appendix A: Physical Media

Tables 1 and 2 below show a summary of the physical layer specifications for both the 40 GbE and 100 GbE MAC rates. The Ethernet Alliance defines different physical layer specifications for computing and network aggregation scenarios, where computing within the data center covers distances of up to 100 meters and network aggregation covers a variety of physical layer solutions to cover service provider interconnection, inter and intra-office applications.

Table 1. 40 GbE Physical Media

Media	Distance	Form Factor	Transmission Media
Backplane			
40Gbase-KR4	At Least 1m	Controlled impedance (100 Ohm) traces on a PCB with 2 connectors and total length up to at least 1m.	4 x 10 Gbps backplane transmission
Copper			
40Gbase-CR4	At Least 7m	QSFP passive copper cable assembly	4 x 10 Gbps, 8 differential pair twin axial cable
Fiber			
40Gbase-SR4	At Least 100m	QSFP optical module with LC connectors CFP optical module with LC or SC connectors QSFP active optical fiber assembly	4 x 10 Gbps on 8 parallel OM3 ribbon fiber, 850 nm
40Gbase-LR4	10KM	Connectors	

Table 2. 100GbE Physical Media

Media	Distance	Form Factor	Transmission Media
Copper			
100GBase-CR10	**At Least 7m copper cable**	**12SFP CXP passive copper cable assembly**	**10 x 10 Gbps, 20 differential pair twin axial cable**
Fiber			
100GBase-SR10	At Least 100m	12SFP CXP active optical fiber assembly CXP optical module with MTP or MPO connectors CFP optical module with MTP or MPO connectors	10 x 10 Gbps on 20 parallel OM3 ribbon fiber, 850 nm
100GBase-LR4	At Least 10KM	CFP optical module with LC or SC connectors	4 x 25 Gbps DWDM on SMF pair, 1295-1310 nm
ER4	40km	Connectors	nm

Appendix B: Future Serial 40GbE

The IEEE-SA is currently working on the development of the IEEE P802.3bg standard which describes Physical Layer and Management Parameters for Serial 40 Gbps Ethernet Operation Over Single Mode Fiber.

Appendix C: Network Design Considerations

Given the way 40 GbE and 100 GbE have been architected, there are a number of ways that equipment manufacturers can suggest architectural designs. The idea is to leverage the n x 10 GbE fibers. For example, in the case of 40Gbase-SR4, the specification leverages 4 x 10 Gbps fibers in each direction or four pairs of fibers. From this a number of physical and MAC architectures could be applied.

1. Point-to-point: Four 10 Gbps lanes can be combined (as defined above) to provide native 40 GbE with an aggregate of 40 Gbps throughput. Here the interface would have its own single MAC controller. This point-to-point 40 GbE deployment helps reduce the number of interfaces to be managed between the Ethernet switches and routers.
2. Point-to-multipoint: Using a single 40 GbE interface at one end and splitting out the individual fiber pairs in the cable to achieve a 1 to 4 assembly. By implementing a MAC controller on each optical pair the assembly can be used to aggregate four 10 GbE streams from four aggregation switches into a single interface on a core switch thus reducing complexities and costs. By using such high density interfaces, the number of switching layers in a network could be reduced by eliminating top of rack switches or blade switches and directly attaching servers to end of row switches. This architecture would be very attractive to data center network managers who are trying to reduce the number of switching layers in their networks.

Appendix D: Other 40G and 100G Architectures

In addition to 40GbE Ethernet, there are two other technology families within 40G technology. The other two are InfiniBand and OC-768 POS. Neither of these two technologies stand to compete with 40 or 100 GbE because they lack the attributes of Ethernet, but they do have a place in today's industry. The focus of this paper is on 40 GbE and 100 GbE but a brief overview of the other two 40G technologies is given below.

40G InfiniBand

InfiniBand makes use of a switched fabric topology, as opposed to a hierarchical switched network like Ethernet. It is a switched fabric communications link primarily used in high-performance computing. InfiniBand leverages point-to-point bidirectional serial links for the purpose of connecting processors to high-speed peripherals such as disks. The InfiniBand architecture specification defines a connection between processor nodes and high performance I/O nodes such as storage devices.

OC-768 Packet Over SONET/SDH (POS)

OC-768 is defined as Packet Over SONET/SDH. In the early days of telecommunications networks, SONET established itself as the backbone of choice for most carriers. In this respect it is found in most facilities-based carrier and operator networks to aggregate slower-speed SONET.

OC-768 currently provides the fastest transmission speeds of up to 39,813.12 Mbps with a data rate of 37.584 Gbps. OC-768 uses Dense Wavelength Division Multiplexing (DWDM) to carry multiple channels of data on a single optic fiber.

Acronyms

- DWDM – Dense Wavelength Division Multiplexing
- GbE – Gigabit Ethernet , usually preceeded by 1, 10, 40, or 100
- Gbps – Gigabit per Second
- HSSG – Higher Speed Study Group
- ITU – International Telecommunications Union
- IEEE – Institute of Electrical and Electronic Engineers
- IETF – Internet Engineering Task Force
- MAC – Media Access Control
- PCS – Physical Coding Sublayer
- PMA – Physical Medium Attachment
- PMD – Physical Medium Dependent
- PHY – Physical Layer Device
- SMF/MMF – Single Mode Fiber/Multi Mode Fiber
- CWDM – Coarse Wave Division Multiplexing
- IEEE 802.3 Standard – the Ethernet Standard
- IEEE P802.3ba – the proposed amendment to the Ethernet Standard for 40 Gbps and 100 Gbps
- Ethernet
- IP – Internet Protocol
- MAC – Media Access Control Layer
- MLD – Multilane Distribution

Make Your Network Mobile

Corporate and North America
Extreme Networks, Inc.
3585 Monroe Street
Santa Clara, CA 95051 USA
Phone +1 408 579 2800

Europe, Middle East, Africa and South America
Phone +31 30 800 5100

Asia Pacific
Phone +65 6836 5437

Japan
Phone +81 3 5842 4011

extremenetworks.com