Terabit Ethernet - Why?

WHITE PAPER PART 1

Understanding the commercial drivers for Terabit Ethernet

See also Part 2: “Terabit Ethernet – How?”

CONTENTS – PART 1*

*** Visit our website to see part 2 of this White Paper: "Terabit Ethernet – How?"**

Terabit Ethernet – Why?

This is the first of two White Papers exploring Terabit Ethernet. Here in Part 1, we will look at the **commercial** factors driving the development of Terabit Ethernet. In Part 2, we will explore the **technical** drivers for new Terabit Ethernet solutions.

EXECUTIVE SUMMARY

The last decade saw a quadrupling of Ethernet speed rates as the industry accelerated from 100GE to 400GE. Now, just a couple of years later, the era of 800GE is upon us - but even this might not be enough, and therefore plans for 1.6TE and beyond are already underway.

It is becoming clear that "cost-per-bit" is no longer the sole determinant of success for Ethernet speeds. Power consumption is also a major factor, particularly for datacenter operators. The trade-off between speed and power considerations means there are different potential paths to Terabit Ethernet depending on network architectures and budgets.

Due to the multiple options, Terabit Ethernet will require a new generation of test solutions with broader capabilities. In addition to testing Ethernet protocols, test solutions will also be needed that can test Layer 1 performance and interoperability for switches, line cards and transceivers.

Xena's current range of Ethernet Traffic Generation & Analysis (TGA) solutions for 400GE and 800GE are designed to deliver the broad flexibility needed by network equipment vendors, service providers and datacenter operators to develop the next generation of high-speed (800GE & 1.6TbE) Ethernet devices and services.

THE DRIVERS FOR TERABIT ETHERNET ARE DIFFERENT

Ethernet as a technology has always been driven by the need for speed. As the adoption of the Internet and its supporting Ethernet and IP protocols grew, so did the demand for bandwidth and higher speed connections. This demand is still growing and continues to be a major driver for new higher speed Ethernet specifications.

However, as we move beyond 100GE, it is becoming increasingly clear that speed alone is no longer the determining factor. Power consumption is increasingly an important consideration especially for hyperscale datacenters who are now the largest consumers of high-speed Ethernet connections.

The challenge for the industry is to find a way to deliver Terabit Ethernet with both a low cost per bit AND a low power consumption per bit.

Data demand is growing and driving the need for Terabit Ethernet

In 2020, the IEEE 802.3 working group released the results of a year-long study into the drivers for Terabit Ethernet. Entitled "IEEE 802.3™ Industry Connections Ethernet Bandwidth Assessment Part II"[1], or BWA II for short, it is a successor to a previous effort in 2012 that provided the basis for what later became the 400GE standard.

BWA II provides a comprehensive overview of all the factors influencing the need for more bandwidth and higher speed Ethernet connections. It is not a call for a Terabit Ethernet specification, but it does provide the business case basis for future Terabit Ethernet standards.

BWA II predicts that demand for bandwidth will continue to grow, albeit at a slower pace than in the past. The real surprise however, is that the report concludes that current estimates of data consumption will outstrip even what a 1.6TE standard could address, as depicted in Figure 1:

[1] Source: IEEE 802.3 Industry Connections NEA Ad Hoc Ethernet Assessment Part II

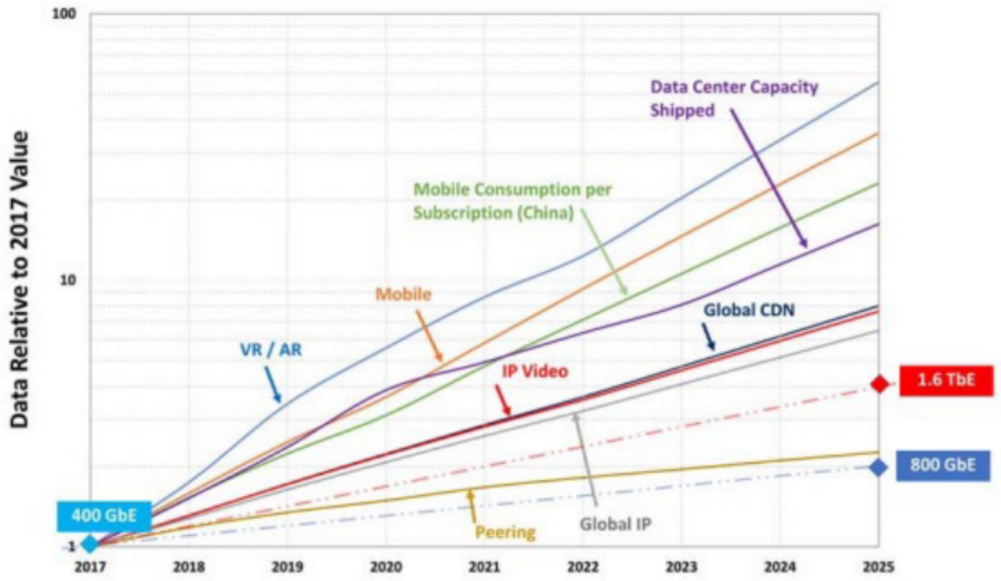


Figure 1: Bandwidth curves showing data consumption relative to 2017 levels

As can be seen in Figure 1, it appears that 800GE will, at no point, meet current demand estimates, while 1.6TE also falls short. These curves are an estimation, but one can clearly see that there is a healthy demand for Terabit Ethernet and even multi-Terabit Ethernet speeds.

One conclusion that could be drawn from studying this chart is that the industry should move to multi-Terabit Ethernet speeds as fast as possible and that the intermediate step of 800GE is superfluous. However, that would be a mistake.

Data consumption is one view that shows the overall demand, but how networks and devices are designed to meet that demand in a cost-effective manner is a different matter. In addition, the growing concern of power-efficiency adds an extra dimension of complexity and Ethernet speeds such as 400G, 800G and 1.6T will all have a part to play, as we will see in later sections.

Understanding Terabit Ethernet demand from an Internet video perspective

To understand the demand for Terabit Ethernet, let's look closer at one of the major data consumers, namely Internet video.

While Internet video is shown separately in Figure 1, it has a knock-on effect on Content Delivery Networks (CDNs), mobile consumption (as this is the preferred viewing device) and data center capacity. As a result, video is having a major impact on network and data center architectures that can influence which path Terabit Ethernet will take.

The BWA II report shows that Internet video (light blue) is by far the largest data consumer, as shown in Figure 2.

Figure 2: Global IP Traffic by Application Type and Global CDN Traffic

The data in Figure 2 is based on input from the annual Cisco VNI report[2]. In Figure 3 we can see which specific kinds of video are driving demand:

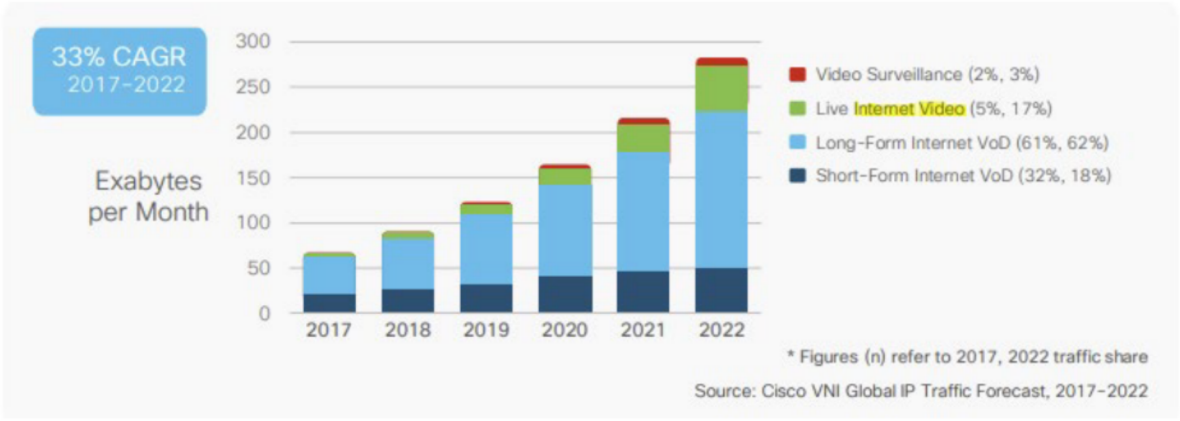


Figure 3: Global internet video traffic by category

Live Internet video is projected to jump from a 5% share in 2017 to a 17% share of Internet video in 2022, which is part of the reason why latency and jitter are becoming more important. The overall growth in Internet video has already impacted the design of communication networks and data center architectures and we can expect video to also influence which Terabit Ethernet path will be taken. It is worth considering the implications that Internet video has on communication networks and data center architectures.

[2] Cisco Visual Networking Index: Forecast and Trends, 2017–2022

Video is affecting network architectures

The first implication is that more data is moving closer to the consumer at the edge of the network. Reducing the distance from the consumer to the provider using content caching makes it easier to minimize the impact of latency and jitter. As shown in Figure 2, CDN traffic is growing in-line with Internet video growth as more video is cached closer to the consumer. This is resulting in more and smaller datacenters being deployed at the edge of the network.

Caching at the edge reduces the backhaul bandwidth requirements but increases the connectivity requirements in the metro network as more traffic stays local. BWA II noted that "metro-capacity of service provider networks is growing faster than core-capacity and will account for a third of total service provider network capacity by 2022".

This change in architecture means we need to rethink where Terabit Ethernet is likely to be used. Traditionally, one would expect high-speed Ethernet to be used on backhaul links to the core network, but with more traffic staying local, we could see 800GE and 1.6TE connections in both metro and access networks.

Video is affecting data center architectures

The second implication of increased video traffic is the need for more bandwidth in data centers. While caching in CDNs close to the consumer relieves the pressure on hyperscale datacenters, only the most popular and frequently used content is cached at the edge. All other content still needs to be provided from hyperscale datacenters. Streaming video is not just bandwidth hungry, it is also storage hungry, especially as we move to higher definition formats, such as 4K and 8K HD.

BWA II used Cisco VNI data to provide an interesting overview of the requirements of various video formats:

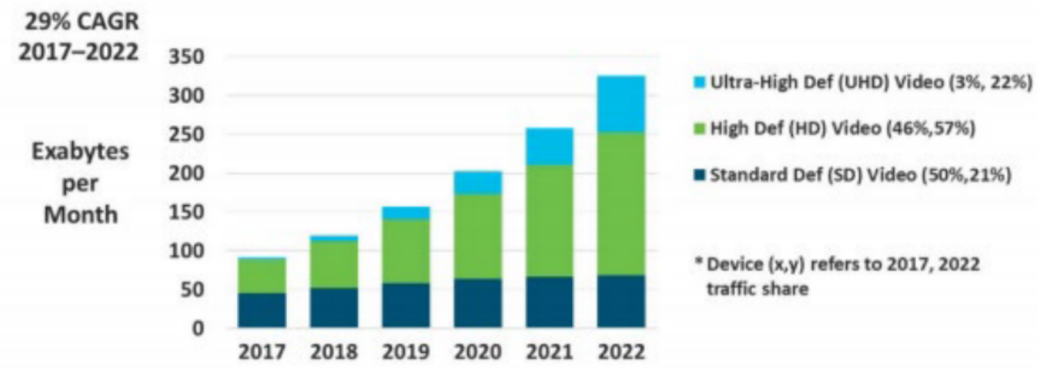


Figure 4: Impact of definition on data growth

As can be seen in Figure 4, Standard Definition (SD) video is not a major driver of data growth. However, the share of High Definition (HD) - which is only expected to grow by 13% (from 46% in 2017 to 57% in 2022) – will cause the amount of data per month consumed to grow by a factor of 5!

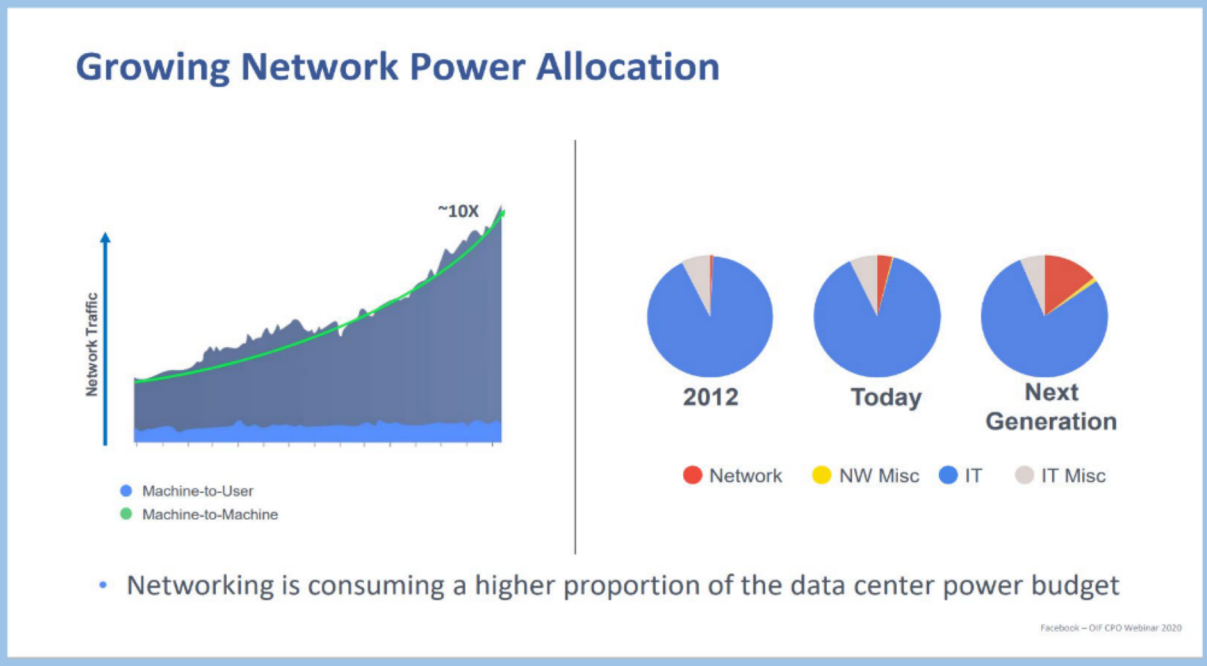
The reason for this increase is that a typical SD video stream requires 2 Mbps, while a HD stream requires between 5 and 7.2 Mbps and a UHD video requires 15 to 18 Mbps. The storage requirements also expand accordingly so two to three times more memory is needed for HD and up to nine times more for UHD. From a datacenter perspective, this means an increase in data exchanged between servers within the datacenter as video files are stored and retrieved from file storage.

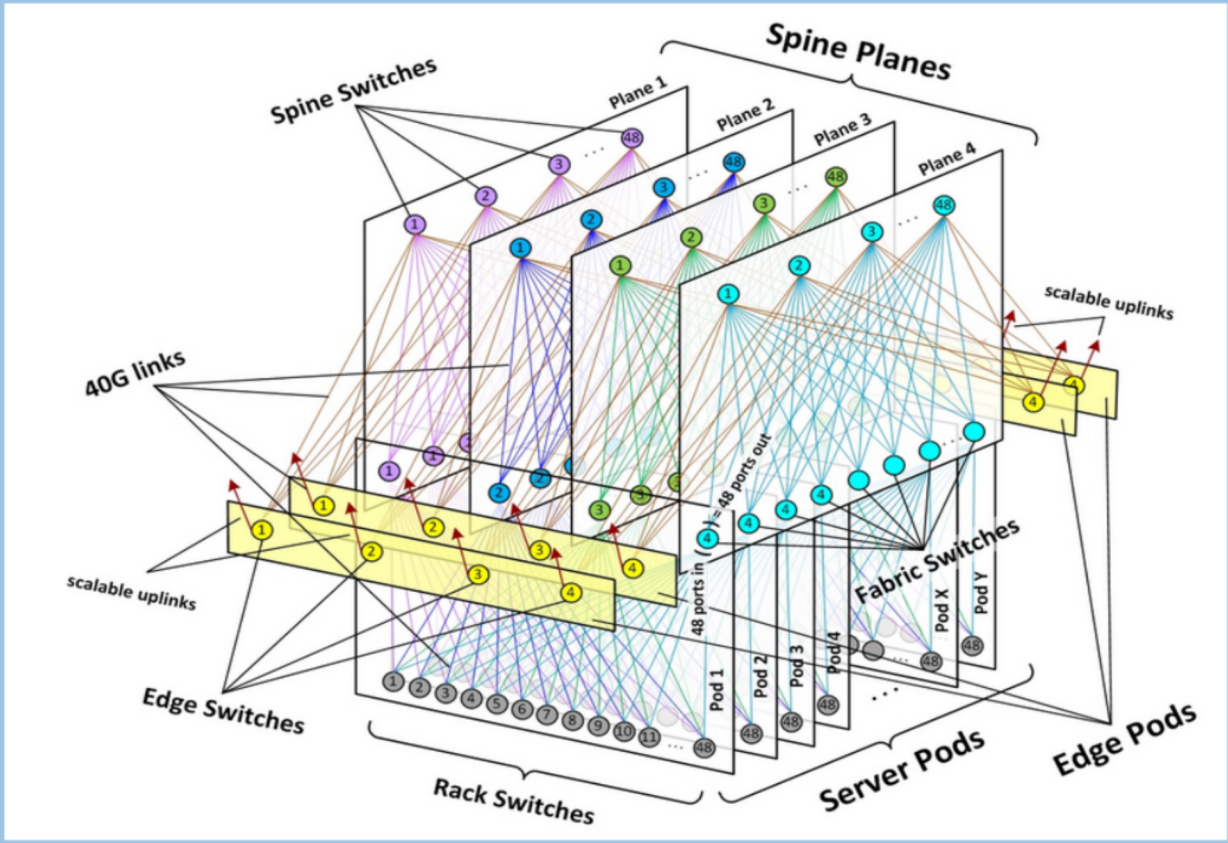
To get an understanding of the impact, look at the evolution of Facebook data center architectures and interconnecting networks since the introduction of video ads and posts in 2013.

THE IMPACT OF VIDEO ON FACEBOOK

In 2013, Facebook introduced video advertisements [1], which was a watershed moment for Facebook in more ways than one. Despite the experts' skepticism, video has proven to be the most popular format of interaction on Facebook, resulting in a dramatic increase in advertising revenue.

However, the introduction of video had profound effects on Facebook's data center infrastructure. As can be seen in the graph shared by Facebook below [2], machine-to-machine traffic (or communications between servers within Facebook) exploded compared to machine-to-user traffic (communications between consumers and Facebook).





Above: Schematic of Facebook data center fabric network topology

This explosive traffic growth forced Facebook to make two profound changes to their data center infrastructure. The first change, made in 2014 very shortly after the introduction of video content meant Facebook abandoned their 3+1 cluster approach to datacenter design in favor of a new fabric architecture as explained in detail in the Facebook engineering blog "Introducing data center fabric, the next-generation Facebook data center network" [3].

A major change Facebook made to the architecture was abandoning large expensive switches aggregating as multiple Top of Rack (ToR) switches to a "Pod" architecture with only 4 fabric switches connecting up to 48 ToRs. This provided a more flexible architecture but required a great deal of high-speed Ethernet interconnection as fabric switches are aggregated in spine planes, which are then serviced by edge switches for communication with other datacenters.

Facebook is an interesting example of how video and machine-to-machine traffic is impacting hyperscale networks. It is driving the need for massive interconnectivity with high-speed Ethernet. All other hyperscale networks are facing similar challenges and, together with Facebook, are actively driving requirements for Terabit Ethernet specifications.

[1] Facebook's Video Ads Risk Alienating Users - WSJ
[2] Facebook presentation, October 14, 2020 – "Co-Packaged Optics – Why, What and How"
[3] Introducing data center fabric, the next-generation Facebook data center network - Facebook Engineering (fb.com)

THE TERABIT ETHERNET POWER CHALLENGE

The Facebook example shows the impact video is having on hyperscale data center architecture and design choices. One consequence of these design choices is a reliance on a dense mesh of high-speed Ethernet connections, which drives the need for Terabit Ethernet connectivity, especially in the fabric and spine (or leaf and spine as others might refer to this design).

But the big challenge is how to increase the capacity of this mesh of Ethernet connections ***without seriously impacting power budgets***.

At a recent EPIC event, Mark Filer, Principal Engineer from Microsoft Azure outlined the problem they faced with Ethernet switches supporting 400GE connectivity[3]. These switches are projected to consume 3 times the power of switches supporting 100GE connectivity, while 400GE optics are expected to consume 3 to 4 times the equivalent power of 100GE optics.

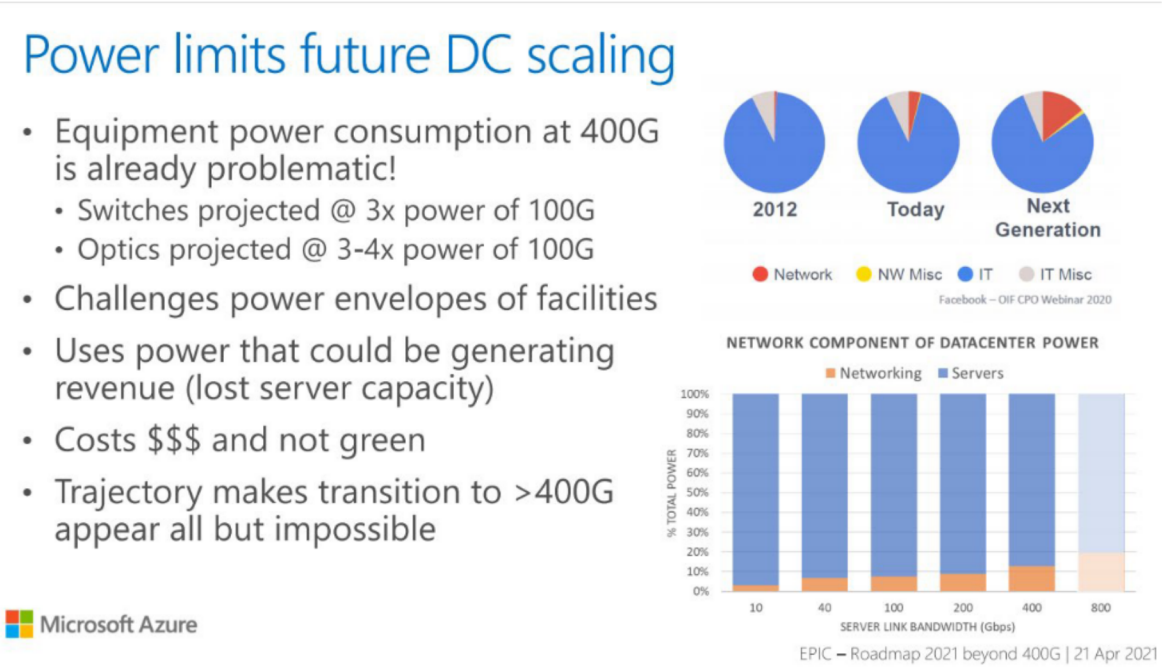


Figure 5: From presentation "What's beyond 400G?" by Mark Filer, Principal Engineer, Microsoft Azure at EPIC 2021 Roadmap beyond 400G event

As can be seen in a slide from Mark Filer's presentation in Figure 5, increasing power consumption in line with speed rates impacts the power envelope of the data center facility. In basic terms, networking is a cost of doing business, because the servers generate revenue. Power for networking is power that could have been used on revenue generating servers! Therefore, reducing power consumption is a business imperative.

However, it is not just a question of revenue. It is a much simpler consideration of whether it is possible to deliver enough power to a data center based on a full mesh of 1.6TE connections. If

[3] Source: EPIC 2021 - Roadmap beyond 400G (epic-assoc.com)

power consumption for switches and optics is increasing at the same rate, then a move to 800G and 1.6TE would require a doubling and quadrupling of the projected power requirements for 400GE respectively. This could mean 1.6TE consuming up to 50% of datacenter power in current facilities.

One answer is to simply build bigger facilities and supply more power, but at some point that becomes impractical.

Many now consider the current path is not sustainable and that a new generation of Terabit Ethernet solutions are required that can not only provide better cost per delivered bit, but also better power consumption per delivered bit.

Understanding tradeoffs in data center architectures and switch designs

Before exploring how we can efficiently increase Ethernet speeds in a power-conscious manner, it is important to consider the trade-offs inherent in data center architectures and switch designs. This was captured quite elegantly in a recent presentation by Cisco Fellow Rakesh Chopra in his keynote presentation at the Ethernet Alliance TEF 21 event[4].

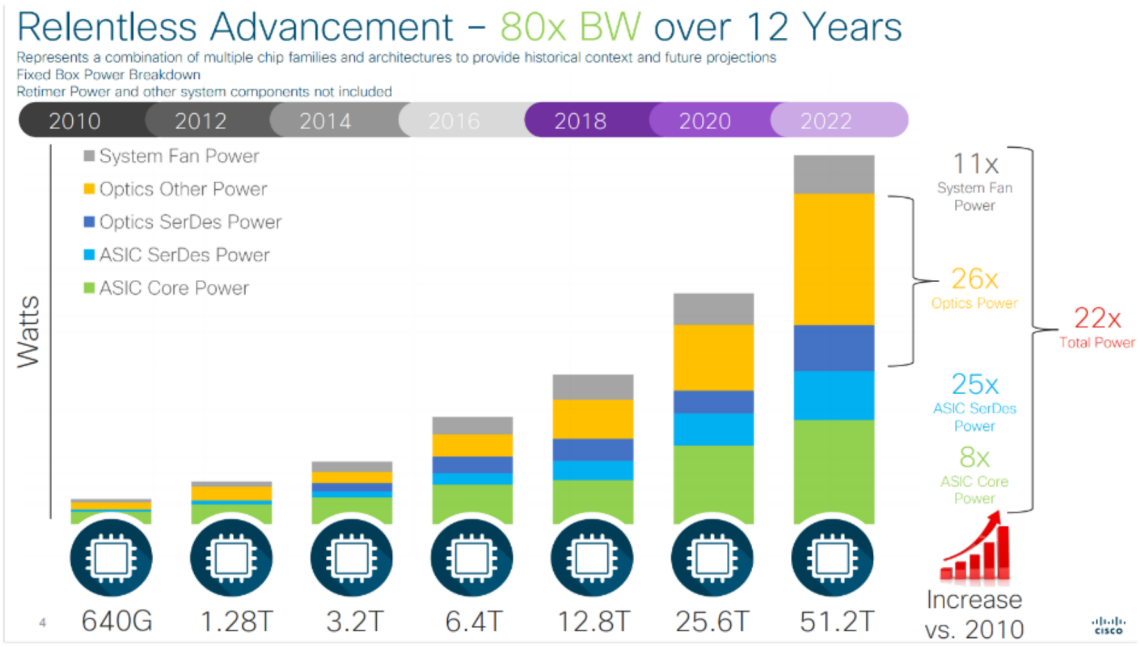


Figure 6: 80 times bandwidth increase for just 22 times power increase.

As Chopra pointed out in Figure 6, power efficiency has improved - the bandwidth of switches has increased by a factor of 80 while power consumption has "only" increased by a factor of 22 during the last decade. But Chopra went on to say that this is not enough: "I think that now power has elevated, in my perspective, to the fundamental thing that we need to address as an industry".

[4] Source: https://ethernetalliance.org/tef-2021-the-road-ahead-recordings-presentations/

Each time the number of ports on a switch, or radix, is doubled, the link speed needs to be reduced by half to meet the same overall switch capacity budget. The switch can only deliver a fixed amount of bandwidth, which needs to be shared by all ports. So, scaling out by increasing the radix is power efficient, but inefficient from a link utilization perspective as it limits the amount of data that could have been delivered to a given port.

Similarly, scaling up by adding an additional network layer increases the cost and total power consumption. For example, the network power per server of a 5-tier network is 3 times higher than a 2-tier network.

Chopra showed the impact of these tradeoffs when considering how a data center architecture can evolve with higher capacity switch configurations:

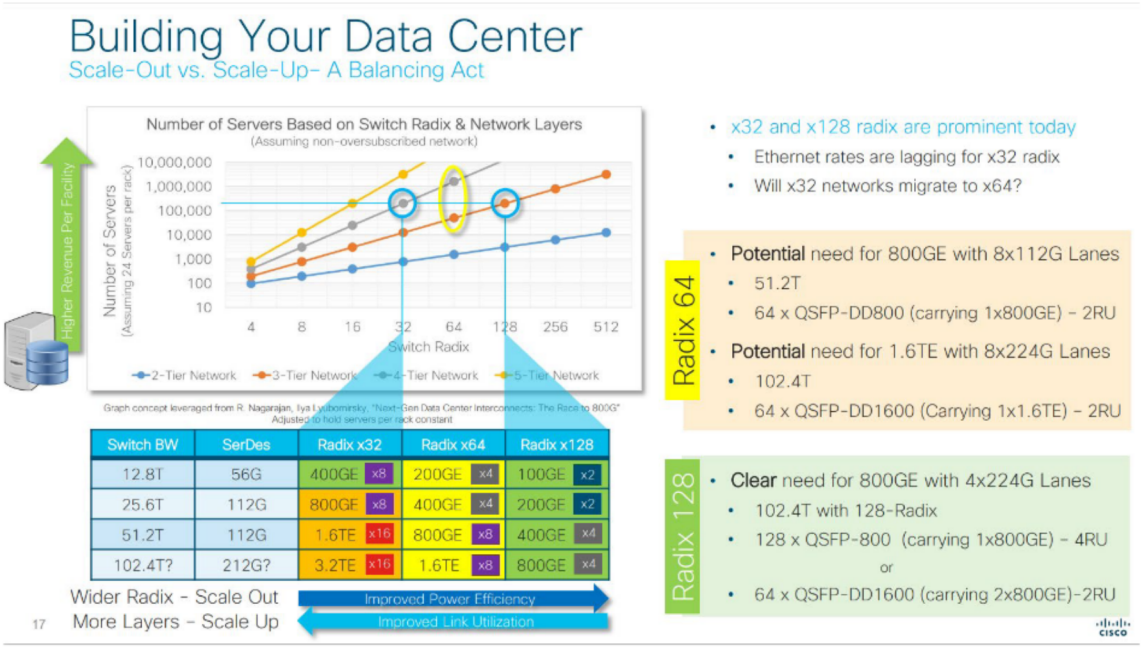


Figure 7: Outlining the need for 800 GE and 1.6 TbE

As can be seen in Figure 7, increasing the radix of the switch to x32, x64 and x128 improves power efficiency but lowers link speeds. This means that there will be a need for 800 GE connectivity based on 112G and 224G SerDes when trying to achieve better power efficiency in datacenters.

THE POWER-SPEED BALANCE & TERABIT ETHERNET

The analysis in Figure 7 provides an interesting counterview of Ethernet speed requirements compared to the earlier data consumption study from the BWA II report in Figure 1. It shows that 800GE can provide a useful option when trying to meet both price-per-bit and power-per-bit goals. The key is understanding the trade-offs in data center architecture and switch designs and how they affect cost and power consumption profiles.

This means that there will be more than one path to Terabit Ethernet as service providers and datacenter operators try to determine the right balance between speed and power efficiency. The availability of different solutions at 400 GbE, 800 GbE and 1.6 TbE will be crucial in providing options to network architects.

But within these different speed rates there will be further options with respect to length, modulation schemes, number of parallel lanes and form factors. This will make Terabit Ethernet more complex, but it will also provide network architects and product developers with more options to create the right path to Terabit Ethernet for specific network and datacenter applications.

It will also have consequences for network architects and equipment manufacturers who want to evaluate the costs and tradeoffs associated with the different design choices as they develop their Terabit Ethernet solutions.

TERABIT ETHERNET TECHNICAL CHALLENGES AND TESTING REQUIREMENTS

Moving beyond 100GE required new approaches and technologies that now must be considered when developing and testing Terabit Ethernet. This includes new modulation schemes and even new optical approaches that lead to a broad variety of potential Terabit Ethernet options.

100GE was considered the end-of-the-road for single-lane solutions based on Non-Return-to-Zero (NRZ) modulation schemes leading to the adoption of 4-level Phase Amplitude Modulation (PAM-4) as a means of increasing the effective bit rate. With NRZ, only two bits can be represented as NRZ is based on just two voltage levels representing a “1” and a “0”. PAM-4, on the other hand, uses 4 voltage levels and can thus represent two bits at each level, as shown in Figure 8. Now, double the number of bits sent with each clock cycle leading to an effective doubling of the number of bits transmitted.

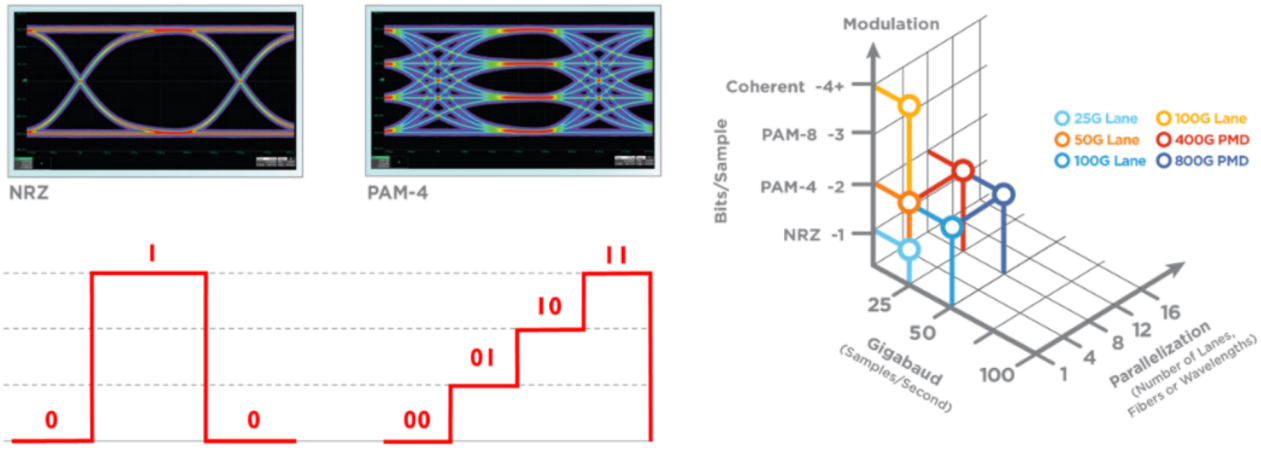


Figure 8: NRZ and PAM-4 encoding and options for achieving 400G and 800G

When this option is combined with the speed of the lane (25 Gbps or 50 Gbps) and the number of parallel lanes used, several different paths to 400G, 800G and 1.6T emerge.

From a product development and design perspective, the challenge becomes how to implement and test PAM-4 ensuring low Bit-Error-Rates (BERs) as the Signal-to-Noise Ratio (SNR) is now reduced. This is important both in achieving good performance, but also in ensuring interoperability with other systems and is a major challenge when moving to Terabit Ethernet speeds.

This new focus on modulation schemes means that Terabit Ethernet testing cannot only focus on Layer 2, but also needs to consider Layer 1 issues. This becomes even more important when we consider the new approaches using Co-Packaged Optics (CPO) where optical and electrical components are packaged together to achieve better cost-per-bit and power-consumption-per-bit.

Terabit Ethernet testing will thus require a broader view encompassing Layer 1 and Layer 2 but will also require a great deal of flexibility in accommodating various combinations of baud rates, modulation schemes and parallelization delivered in a variety of form factors, such as QSFP-DD and OSPF. Flexibility will thus become a key criterion for Terabit Ethernet testing equipment going forward.

XENA'S PATH TO TERABIT ETHERNET TESTING

Test & measurement solutions need to be at the forefront of technology to help first movers develop new products – in this instance terabit switch, line-card and transceiver solutions.

Xena's Ethernet test solutions are designed to be flexible to give engineers all available options in terms of speeds, modulation schemes and form factors.

Xena's 400GE test module (P/N Thor-400G-7S-1P) is a good example. Thor can test seven different Ethernet network speeds: 400GE, 200GE, 100GE, 50GE, 40GE, 25GE and 10GE, plus it supports both the 50G PAM4 PHY and legacy 25G NRZ modulation schemes in a single module.

Figure 9: Xena's Thor-400G-7S-IP test module

The solution provides support for interoperability testing with auto-negotiation and pre/post FEC statistics for PAM4 interfaces. It also provides equalization controls for pre-emphasis, transmission attenuation and post-emphasis signal integrity analysis as well as the option to auto-tune the receiver equalizer/CTLE. (For more info see: Thor 400G 7-Speed Test Module.)

Xena has also announced the Freya 800GE test module, which provides the same flexibility and features including support for both QSFP-DD and OSPF transceiver form factors as well as support for 800G Direct Attached Cable (DAC). Freya is based on 112Gbps SerDes and will be a steppingstone to Xena's next-generation terabit Ethernet traffic generation & analysis solutions.

Both the Thor and Freya modules provide a comprehensive solution for testing both Layer 1 and Layer 2 that will prove essential for network equipment vendors, service providers and datacenter operators in their functional, qualification and interoperability testing efforts.

>> Book a consultation with a Xena tech expert