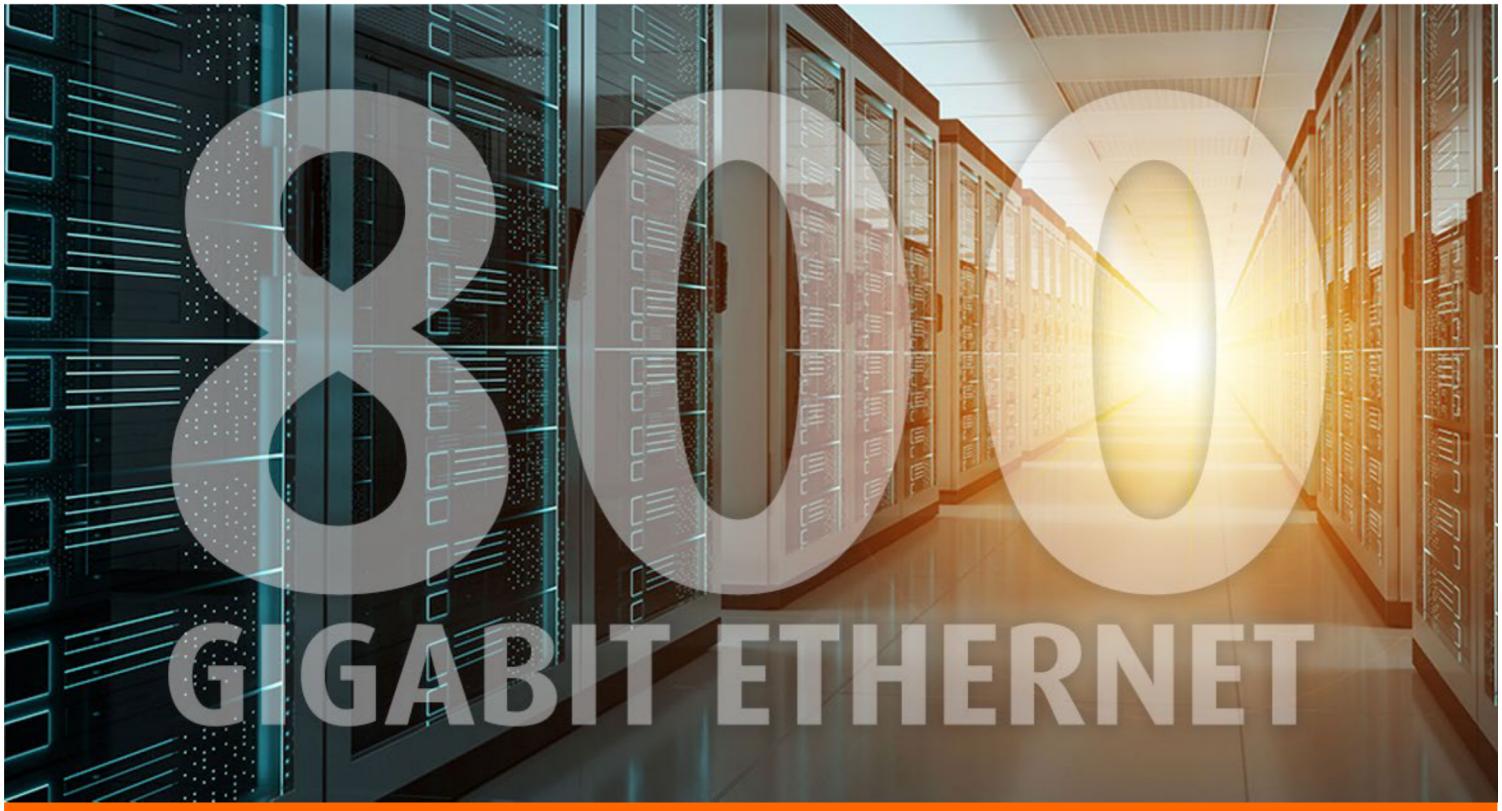
Terabit Ethernet - How?



(PART 2) WHITE PAPER

How Terabit Ethernet works

See also Part 1: “Terabit Ethernet – Why?”

CONTENTS (PART 2*)

*** Visit our website to see part 1 of this White Paper: “Terabit Ethernet – Why?”**

Terabit Ethernet – How?

Part 2: The Technical Challenges facing Terabit Ethernet today

EXECUTIVE SUMMARY

In part 1, we looked at the drivers for Terabit Ethernet and the need to balance power and speed considerations. In part 2, we will look at the technical challenges facing Terabit Ethernet and examine the various technical paths to Terabit Ethernet that are available or emerging today.

Achieving 100GE connectivity almost a decade ago was seen as a major achievement that stretched the limits of available technology. Moving to 400GE required new innovations and approaches that can now pave the path to Terabit Ethernet. These new approaches are enabling 800GE solutions to be delivered already today - well before IEEE standards are available – in response to the overwhelming demand for faster Ethernet solutions.

However, these new approaches make testing the new solutions more challenging. To enable 400GE, 800GE and higher-speed Terabit Ethernet solutions, changes need to be made to the Ethernet physical layer. To date, Ethernet test solutions have mostly needed to focus on the data link layer, but now, extensive testing also needs to be performed on the physical layer. Ethernet testing professionals need to understand how the Ethernet physical layer operates and how this affects the data link layer.

Because changes need to be made to the Ethernet physical layer and because solutions are being delivered before standardization is fully in place, ensuring interoperability becomes a major hurdle. Not just because of the lack of standards, but also because solutions will be operating at the performance edges of current technology capabilities. Trade-offs will need to be made and solutions will be more sensitive to errors. Testing the robustness of Terabit Ethernet solutions in the face of potential errors will become more important.

As we saw in part 1 of this series, Terabit Ethernet will be dictated by the trade-off between speed and power-consumption. Reliable and accurate test equipment will provide the insight required to make these trade-off decisions. Xena Networks is leading the way with 800GE test solutions at the very forefront of developments.

WHAT IS THE PATH TO TERABIT ETHERNET?

While earlier Ethernet generations were driven by cost per bit delivered leading to a focus on higher speed rates, it is now clear that power consumption per bit delivered will also be a major consideration. These dual requirements have also opened new paths to Terabit Ethernet, some of which have already emerged with current 400GE and recent 800GE solutions. But this is just the beginning. Other paths and solutions are also being investigated and innovation is underway in many areas to find new solutions that can meet the strict requirements of Terabit Ethernet.

The potential paths to Terabit Ethernet that are now being paved with 400GE and 800GE are driven by changes at the physical layer. Faster electrical lane speeds, new modulation schemes and the subsequent need for improved Forward Error Correction (FEC) mechanisms are all part of making faster Ethernet speeds possible. Compact transceiver module form factors that can take advantage of these advances and keep power consumption to a minimum are also a key part of making Terabit Ethernet viable.

Compact transceiver module form factors also address another driver for higher speed solutions beyond speed and power consumption, namely space. As data consumption grows, data center space is at a premium. Viable Terabit Ethernet solutions must provide high-speed and power-efficiency, but also enable high port density maximizing the available front panel space in data center switching solutions.

Market demand is driving solutions through industry consortia that cannot wait for IEEE standards to be published. They are providing their own standard specifications that can be used to deliver interoperable solutions today. Thankfully, these efforts are closely aligned with IEEE standardization efforts so investment in available solutions can be made without the risk of major changes in the future.

To understand how Ethernet is changing as we move beyond 100GE, we will look at the innovations that were necessary to progress to 400GE and how these new approaches have been pragmatically re-used to provide 800GE solutions today based on available compact transceiver module form factors. To understand whether these approaches can be extended to 1.6TE and beyond, we need to understand how the Ethernet physical layer works, how it is changing and the challenges that presents.

From 10 times leaps to smaller steps

Up until 10GE, Ethernet speeds increased in factors of 10. Standardization took several years, and industry adoption was relatively slow, as factor of 10 leaps were costly to absorb. The issue is that new Ethernet technology, such as transceivers, switches and other equipment cost more when first introduced. The cost-per-bit delivered can be higher than the current Ethernet speeds available, making it more cost-effective to add more connections at lower speeds than upgrade to the new speed.

After 10GE, 100GE was planned, but some suggested an interim step of 40GE. This initially met stiff resistance, but the economics of cost-per-bit meant that 40GE had a viable role to play and was standardized. Since then, the 10 times leap has been replaced by other factors based on 2 and 2.5.

Beyond 100GE, we can see these alternative multiplication factors in specifications for 200GE, 400GE, 800GE and 1.6TE, which are progressing in factors of 2. These provide stepping-stones that can be exploited on the path to Terabit Ethernet allowing the right balance between speed and power-consumption to be met.

The two-dimensional speed game of 100GE

For 100GE, the determining dimensions of speed were the lane speed and the number of parallel lanes available. The lane speeds are the critical determining factor for both electrical and fiber-based communication, as it is the electrical lane speed connections between the transceiver modules (or equivalent embedded solution) and the switch ASICs that ultimately determine how fast the network performs.

In 100GE implementations, the bits on electrical lanes are represented as two voltage levels, which can be used respectively to symbolize bits “0” and “1”. This is referred to as a Non-Return-to-Zero (NRZ) modulation scheme as depicted in Figure 1. It is called NRZ because the low voltage representing “0” is not zero volts, but a low-voltage level just above zero with “1” represented by a higher voltage level. The modulation is performed electrically and then converted to light for transmission over fiber or transmitted directly on electrical Ethernet cables.

With NRZ, the number of bits transmitted per second is the same as the clock or “baud” rate used to send a “0” or “1” symbol. The baud rate refers to the number of symbols transmitted per second. Each bit is represented by a single symbol (in this case a voltage representing either a “0” or a “1”. So, the baud rate is the same as the line rate for NRZ.

The three-dimensional speed game of Terabit Ethernet

To increase the amount of data that can be transmitted without increasing the baud rate or how many symbols are transmitted per second, the only available option is to consider an alternative modulation scheme to NRZ that allows more bits to be represented by a symbol. For 400GE, the answer is to use 4-level Pulse Amplitude Modulation (PAM4).

With PAM4, instead of using only two voltage levels to represent “0” and “1”, four voltage levels are used to represent pairs of bits, as shown in Figure 1.

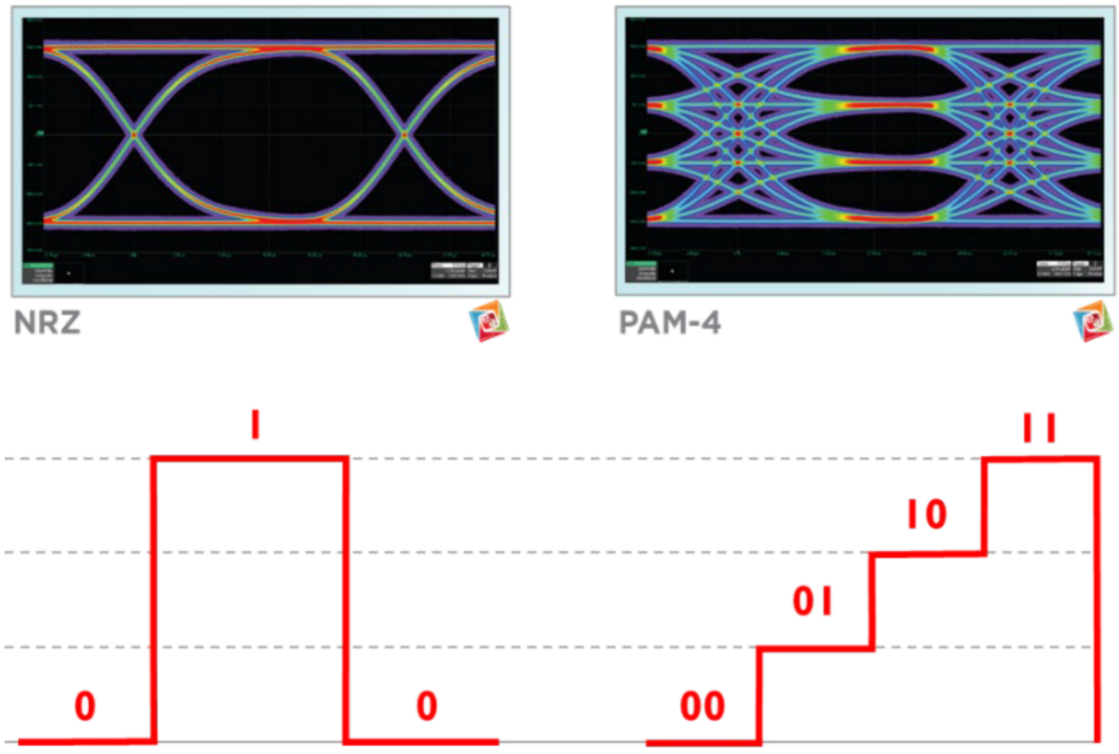


Figure 1: NRZ vs PAM4 encoding

With PAM4, double the number of bits can be transmitted on each clock cycle so now it is possible to transmit 50 Gbit/s of data on a single lane rather than just 25 Gbit/s using NRZ. This can sometimes cause confusion, especially when we need to examine the details of the physical layer implementation. It is therefore helpful to distinguish between the baud rate and the bit rate so that we understand the actual bit-per-second throughput per lane. In the remainder of the document, we will refer to baud rates as Gigabaud-per-second (Gigabaud), whereas the effective bit-per-second throughput will be referenced using Gbit/s. Thus, a PAM4 lane operating at 25 Gigabaud delivers 50 Gbit/s.

Initial 100GE implementations were limited to a baud rate of 25 Gigabaud, but since then, the baud rate has increased to 50 Gigabaud. The combination of higher baud rates per electrical lane and PAM-based modulation opens new paths to reaching Terabit Ethernet. If we also add the third dimension of number of lanes, then a variety of solutions can be provided as can be seen in Figure 2.

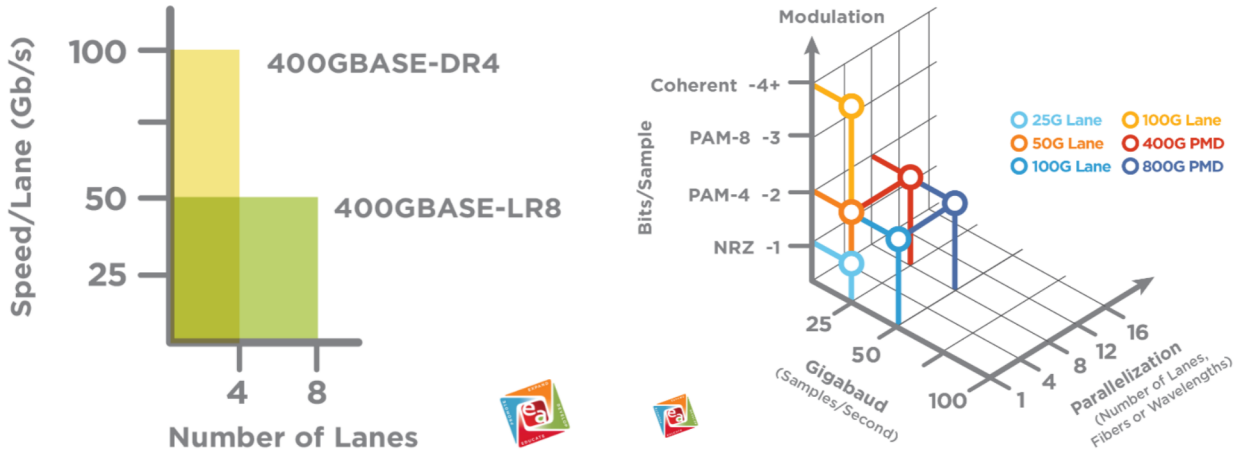


Figure 2: 400GE options and evolution to higher lane speeds

For example, 400GE can be delivered based on four lanes running at 50 Gigabaud delivering 100 Gbit/s for shorter reaches or at half the baud rate using double the number of lanes for longer reach. If we begin to consider other modulation schemes, we could even use lower baud rates. The following table provides an overview of currently available transceiver module options based on combinations of lanes and bit-rates delivered:

Standard	Transceiver	Link Distance	Media Type	Lanes	Encoding
IEEE 802.3bs	400GBASE-DR4	500m	SMF	4×100 Gbit/s	PAM4
	400GBASE-FR8	2km	SMF	8×50 Gbit/s	PAM4
	400GBASE-LR8	10km	SMF	8×50 Gbit/s	PAM4
IEEE 802.3cm	400GBASE-SR8	100m	MMF	8×50 Gbit/s	PAM4
	400GBASE-SR4.2	100m	MMF	8×50 Gbit/s	PAM4
IEEE 802.3cn	400GBASE-ER8	40km	SMF	8×50 Gbit/s	PAM4
100G Lambda	400GBASE-FR4	2km	SMF	4×100 Gbit/s	PAM4
MSA	400GBASE-LR4	10km	SMF	4×100 Gbit/s	PAM4

To understand the meaning of the various transceiver types, the naming is based on the following convention[1] shown in Figure 3:

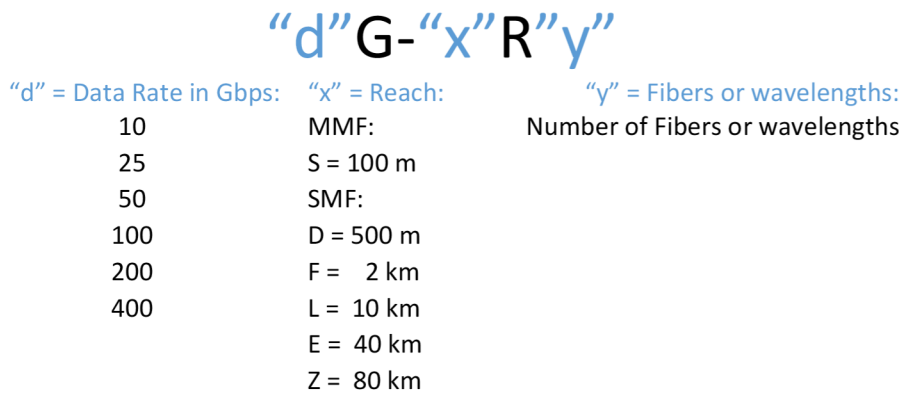


Figure 3: Naming convention for transceivers

Transceiver module form factors are an important consideration, not just from a cost perspective, but also from a compactness and power-consumption perspective, which are important to Terabit Ethernet. Various options are also available here including some highly compact and power-efficient options that are promising for Terabit Ethernet.

QSFP-DD AND OSFP PROVIDE COMPACT LOW-POWER TERABIT ETHERNET TRANSCEIVERS

The first 100GE transceiver developed was based on 10 lanes of 10 Gbit/s as this was the quickest and easiest way to get a 100GE solution to market. However, the CFP form factor was needed as 10 lanes means 20 fibers (10 in and 10 out). This made the module big, expensive, and power-hungry. Smaller CFP transceiver module form factors have been introduced since that time, but the most popular 100GE module transceiver module form factor now is QSFP28 where 4 lanes with a maximum bandwidth of 28 Gbit/s are provided.

The "Q" in QSFP refers to "quad" indicating 4 lanes in each direction for a total of 8 fibers, while the "28" refers to the 28 Gbit/s bandwidth, which is required to enable support for the overhead introduced by various protocols. For example, the bit rate for 25 Gbit/s Ethernet transmission with NRZ is actually 25.78 Gbit/s to accommodate protocol overhead and the use of FEC in some 100GE implementations.

[1] 400GBASE-SR4.2 is a special case where only 4 fibers are used for both transmit and receive hence the ".2" in the naming

The power consumption for QSFP28 is usually less than 3.5W, while CFP power consumption can range from 6W to 24W, which is one of the reasons it is the preferred form factor for 100GE. The availability of a 25 Gbit/s lane rather than a 10 Gbit/s lane thus leads to a compact, high-speed and power-efficient 100GE solution enabling cost-effective, high-port density switch implementations.

As Figure 4 shows, a number of transceiver module form factors based on 1 to 4 and 4+ lanes are in play with respect to Terabit Ethernet:

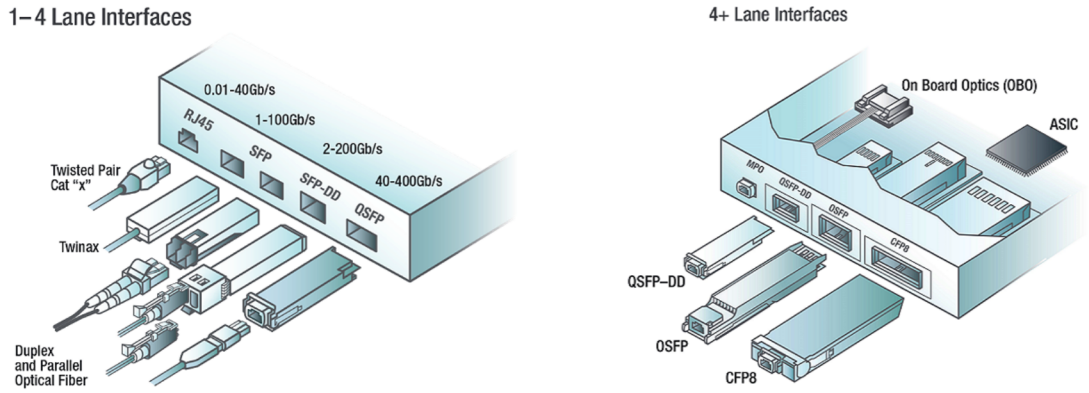


Figure 4: Currently available transceiver module form factor options for 1-4 and 4+ lanes (image courtesy of the Ethernet Alliance)

The first specification of 400GE in 802.3bs was based on 16 lanes of 25 Gbit/s following the same approach as the 10x10 lane 100GE CFP transceiver module form factor. It was quick and used existing technologies. However, deploying 32 fibers to support this approach is simply cost-prohibitive as well as power-hungry making this an unattractive solution.

Right now, the current state of the art is QSFP-DD and OSFP. These provide up to 8x50 Gbit/s electrical inputs for 400GE connectivity and can be extended to 8x100 Gbit/s to support 800GE. OSFP and QSFP-DD currently support all reaches including 400ZR. QSFP-DD is backward compatible to lower speed QSFP modules, such as the QSFP28 transceiver.

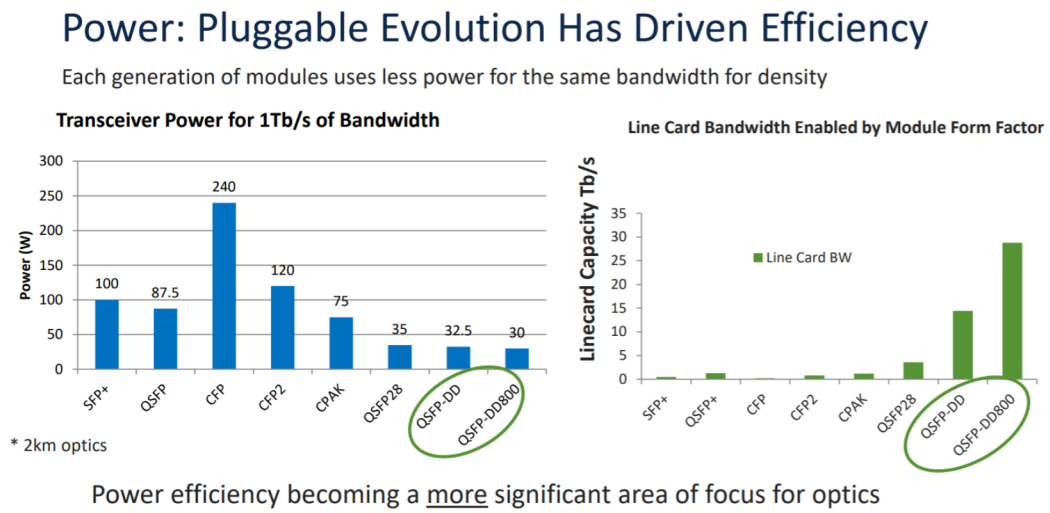


Figure 5: From Cisco presentation " Pluggable Optics – Pros and Cons" at "Co-Packaged Optics – Why, What and How" webinar, OIF in partnership with Lightwave[2]

What is most interesting is that pluggable transceiver module form factors are getting more power efficient. For example, the QSFP transceiver module form factor family provides an order of magnitude better power efficiency compared to other transceiver module form factors as shown in Figure 5.

Considering that power consumption, as well as speed, are major considerations for deployment of Terabit Ethernet solutions, the low power consumption that is possible with QSFP-DD and OSFP is very encouraging. Another encouraging forecast is that the vast majority of pluggable transceivers will be used for short-range applications. In their September 2020 "High-Speed Ethernet Optics" report, LightCounting provided a forecast for 400GE transceiver consumption, as shown in Figure 6.

[2] Source; October 14, 2020 – "Co-Packaged Optics – Why, What and How"

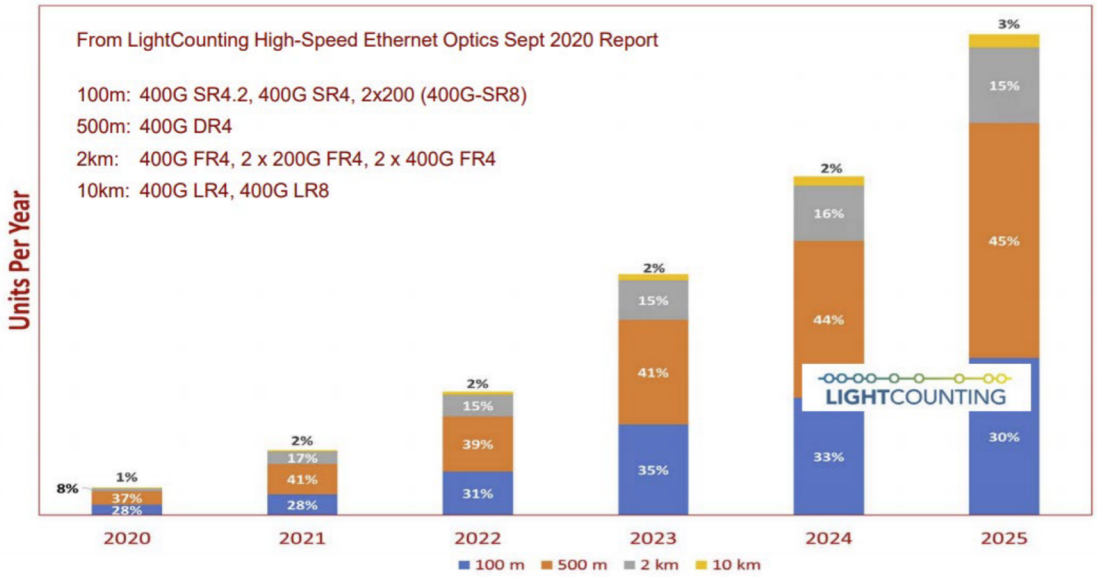


Figure 6: From presentation "The Next Ethernet – Beyond 400 Gb/s" by John d'Ambrosia, FutureWei and IEEE 802.3 working group as part of Lightwave webinar on "Beyond 400G"

As can be seen, 75% of 400GE transceivers are expected to be used in applications requiring less than 500m reach. This is because these solutions can be used for 400GE connectivity, but also break-out to 2x200GE and 4x100GE. The report is only showing optical solutions, but we should expect a similar demand for electrical cables.

FROM 400GE TO 800GE

Despite the fact that 800GE is not standardized by IEEE at this point, the demand for 800GE is so high that solutions are already being made available. This is thanks in large part to the Ethernet Technology Consortium.

The Ethernet Technology Consortium is the new name for the 25 Gigabit Ethernet Consortium, an organization that was founded to accelerate the development of 25GE, 50GE and 100GE using existing and draft specifications to provide solutions to the market as quickly as possible. It has over 45 members including Arista, Broadcom, Cisco, Dell, Google, Mellanox and Microsoft.

On April 6, 2020, the Ethernet Technology Consortium announced the availability of the 800GBASE-R specification for 800GE[3]. But, to truly understand how the solution works and the innovations that make it possible, it is useful to understand how the physical layer is structured and how this changed when moving from 100GE to 400GE. This will also provide insights into the challenges that have been introduced that are driving the need for new Ethernet testing approaches and solutions.

[3] Source: 25 Gigabit Ethernet Consortium Rebrands to Ethernet Technology Consortium; Announces 800 Gigabit Ethernet (GbE) Specification - Ethernet Technology Consortium

Ethernet MAC and physical layers in 400GE

The MAC layer is part of the data link layer in the OSI 7-layer reference model used in the Ethernet/IP protocol and is responsible for data transfer between nodes using Ethernet frames. The MAC is responsible for encapsulating IP packets in the Ethernet frames, transmitting and receiving these frames, handling any pre-ambles and paddings between the frames and protecting against errors using Frame Check Sequence (FCS).

The MAC interfaces to the physical layer through a Media-Independent Interface (MII) link that enables Ethernet to be transmitted over a wide variety of media. The physical layer is responsible for mapping Ethernet frames to the physical medium for transmission and it is here that changes need to be made to enable higher Ethernet speeds.

The Ethernet physical layer is based on three sub-layers, which are described below based on a 400GBASE-DR4 example shown in Figure 7:

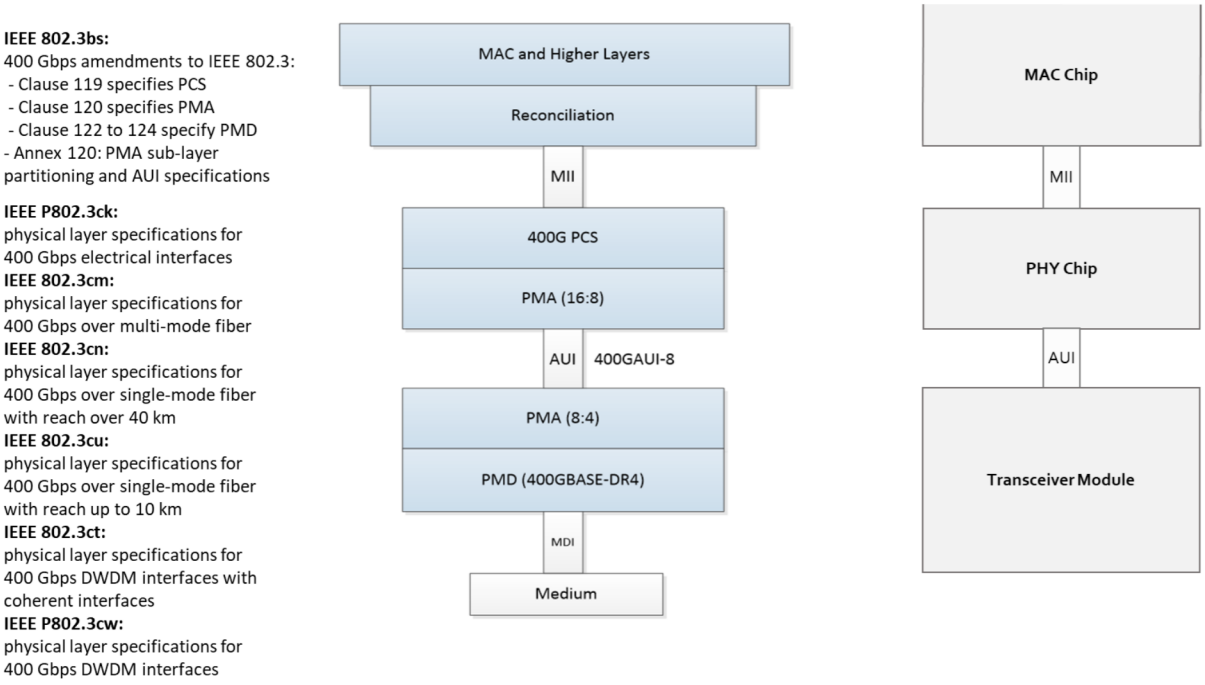


Figure 7: Example 400GE Ethernet MAC and physical layer based on 400GBASE-DR4

The Physical Coding Sublayer (PCS)

In 400GE, a 400 Gbit/s MII (400GMII[4]) link is used, which is an 8 octet, synchronous data interface that is 64-bits wide. The PCS is responsible for encoding and decoding the 64-bit wide 400GMII link to the MAC as well as encoding and decoding of data bits into “code groups” for transmission via the

[4] Also known as the CDGMII interface

PMA sublayer (see below). The interconnection between the MAC and the PCS is a logical interface rather than a physical interface.

The PCS uses block coding to structure data. The 8 octets received from the 400GMII are encoded into a 64-bit block. A synchronization header of two bits is added to form a 66-bit block leading to a 64b/66b block.

In 400GE, PAM4 is used, which means that the Signal to Noise Ratio (SNR) is lower than for NRZ. A strong FEC is therefore needed. In 400GE, the Reed-Solomon RS(544, 514, t=15) is mandatory. It is stronger than the RS(528,514,t=7) FEC that was previously introduced for 100GE using 4x25 Gbit/s. However, 64b/66b encoding is not efficient enough for use with the RS-FEC so the 64b/66b block needs to be transcoded to a more efficient 256b/257b block. The RS-FEC is applied to the 256b/257b block.

The encoded data with FEC is then distributed over 16 PCS lanes running at 26.5 Gbit/s to accommodate the coding and FEC overhead.

The Physical Medium Attachment (PMA) sublayer

The PMA translates between the PCS and the PMD (see below) by mapping received code group bits to data symbols for transmission over the physical medium using PAM4 encoding. As can be seen in Figure 7, this layer can be split allowing a chip-to-chip interface with the top PMA sublayer multiplexing 16 PCS lanes into 8 physical lanes running at 26.5 Gigabaud delivering 50 Gbit/s each. The 8 lanes are connected to the bottom PMA sublayer over an 8 lane 400 Gbit/s Attachment Unit Interface (400GAUI-8). The bottom PMA sublayer retimes the incoming PAM4 signals and then converts back to 16 PCS lanes, which are then multiplexed to the 4 lanes needed for the 400GBASE-DR4 PMD in this example. Each lane operates at 50 Gigabaud delivering 100 Gbit/s.

Physical Medium Dependent (PMD) sublayer

The PMD maps PAM4 data symbols to signal values on the physical medium depending on the type of medium.

Understanding coding overhead and baud rates

Earlier, it was mentioned that the line rate for QSFP28 based Ethernet using NRZ is 25.78 Gbit/s, but why isn't it just 25 Gbit/s? In 400GE, it can be seen that the line rate for 25 Gbit/s connections is actually 26.56 Gbit/s. Why is there a difference?

The difference comes from the block encoding, modulation scheme and FECs used. Traditionally, the overhead associated with 64b/66b block encoding was around 3%[5]. For 25 Gbit/s NRZ, this gives 25.78 Gbit/s as specified in QSFP28 100GE Ethernet applications. In 100GE, the RS(528,514,t=7) was used, but the additional overhead could be accommodated within the same line rate by compressing 4x64b/66b blocks into a 256b/257b block.

[5] From calculation 66/64 = 1,03125.

As we move to 400GE, we need to use PAM4 modulation, which has lower SNR and thus needs a stronger FEC. This adds an additional 3% overhead leading to a line rate of 26.5625 Gbit/s. As we move to higher lane speeds, the line rate becomes multiples of 26.5625, such as 53.125 Gbit/s for delivering 50GE and 106.25 Gbit/s for 100GE. For other protocols, more overhead is needed, which is why there are also references to 28, 56, 112 and 224 Gbit/s line rates, but for Ethernet, the line rates are currently multiples of 26.5625 Gbit/s.

800GE has arrived

The Ethernet Technology Consortium specification introduces a new MAC and PCS layer. Focused on speed, practicality and backward compatibility, the specification re-purposes the existing IEEE 802.3bs standard for 400GE to enable the distribution of data across 8x106 Gbit/s lanes. The 800GE specification from the Ethernet Technology Consortium is achieved by using 2x 400GE PCS layers to connect to a single MAC layer operating at 800 Gbit/s. It allows existing PMA and PMD specs to be re-used and thus allows the solution to also be used for 2x400GE connectivity.

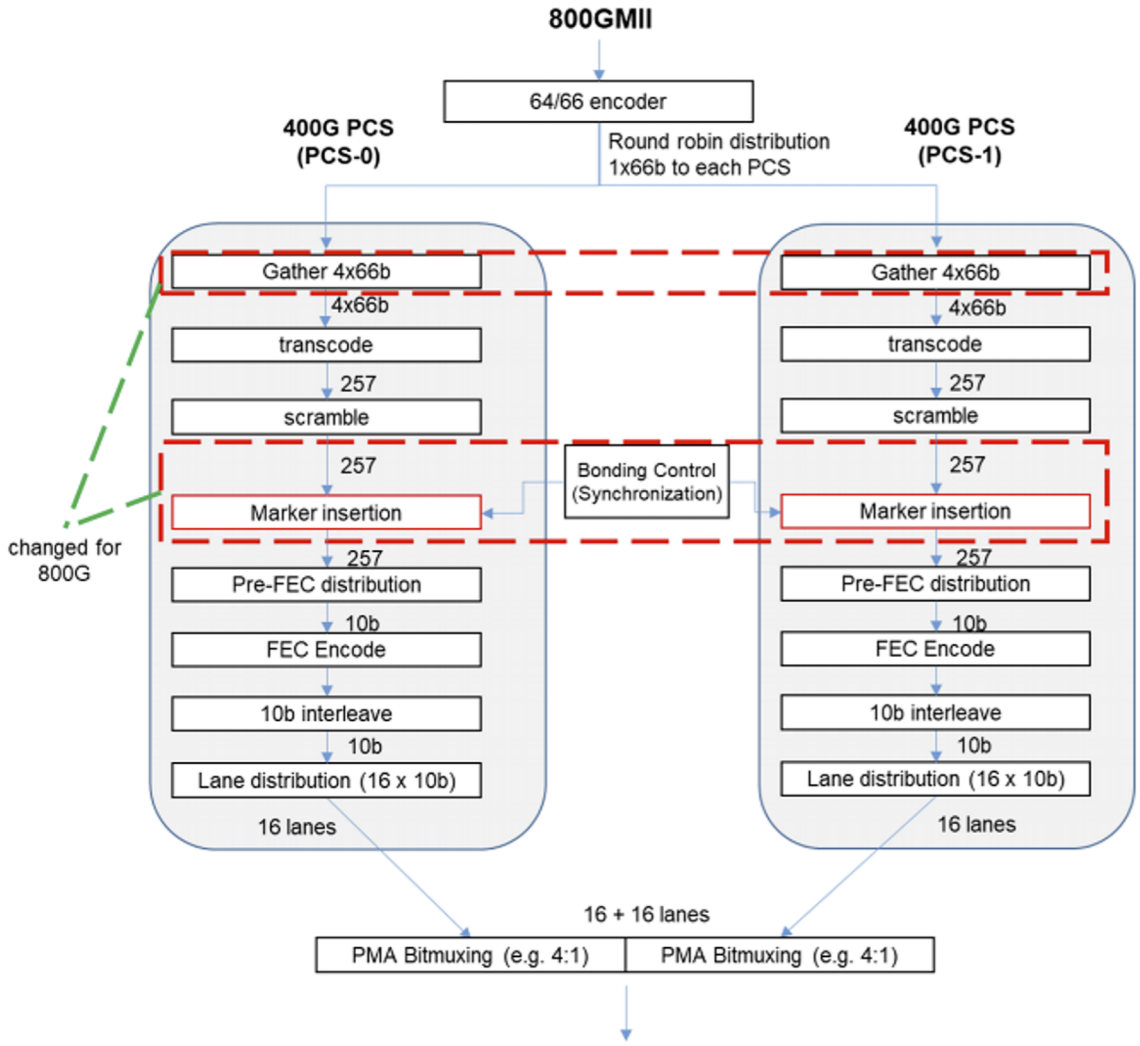


Figure 8: 800GE 8x106Gbit/s PCS transmit flow from Ethernet Technology Consortium[6]

The details of the new 800GE PCS sub-layer transmit flow are shown in Figure 8 with indications of the changes made in the 400GE PCS to support 800GE. In 400GE, a Multi-Lane Distribution (MLD) technique is used to distribute data from the new 64-bit wide 800GMII to the 16x PCS lanes. A 64b/66b block encoder is still used and alternating 66B blocks are sent to each 400G PCS on a round-robin basis. This requires the insertion of alignment markers for every 163,840 blocks encoded, which only need to be slightly altered for 800GE.

Each PCS produces 16 PCS lanes providing a total of 32 PCS lanes to the PMA each operating at 26.5 Gbit/s. The PMA performs a 4:1 multiplexing of the 32 PCS lanes to 8 lanes with PAM4 encoding operating at 53.12 Gigabaud delivering 106 Gbit/s each according to IEEE 802.3-2018 clause 120.

[6] Source: 800G R1.0 Specification (ethernettechnologyconsortium.org)

SPECIFICATIONS FOR 400GE ELECTRICAL AND OPTICAL INTERFACES

This Ethernet Technology Consortium implementation is based on available standards as well as pre-standard drafts from the IEEE 802.3ck working group. The PMD is not defined in the Ethernet Technology Consortium specification with the assumption that 2x400GE PMDs can be used to form an 800GE interface. IEEE 802.3bs specifies a number of different PMDs including 400GE PMDs:

- Clause 122: PMD for 400GBASE-FR8 and 400GBASE-LR8
- Clause 123: PMD for 400GBASE-FR16
- Clause 124: PMD for 400GBASE-DR4

IEEE standards exist for transmission over multi-mode and single-mode fiber as well as DWDM systems using coherent optics and PAM. Working groups are busy completing a specification for transmission over shorter range DWDM systems (IEEE P802.3cw) as well as a 106 Gbit/s per lane electrical standard (IEEE P802.3ck).

The work of the P802.3ck group is interesting as it standardizes 106 Gbit/s operation over electrical links. This opens up opportunities for 400G Direct Attach Cables (DAC), Active Electrical Cables (AEC) and Active Optical Cables (AOC) for up to 2m and 100m for passive and active variants respectively based on only 4 lanes rather than the 8 lane implementations today.

What is more significant is that the availability of a standardized 106 Gbit/s per lane option enables an 8x100 Gbit/s 800GE solution, which is compatible with popular pluggable transceiver module form factors, such as QSFP-DD and OSFP. The Ethernet Tech Consortium have used these draft specifications to provide early 800GE solutions and PHY chips are already becoming available on the market that take advantage of the Ethernet Tech Consortium specifications.

WHY IS INTEROPERABILITY SO IMPORTANT?

The path to cost- and power-efficient Terabit Ethernet requires operating at the bleeding edge of maximum baud rates per lane, PAM4 modulation schemes or potentially even more advanced modulation in the future, operating over as few lanes as possible to enable compact, cost-efficient transceiver module form factors.

However, this comes at a cost. As we have seen, changes need to be made at the physical layer with higher baud rates, potential SNR degradation and the need for advanced FEC implementations. These are all potential sources of implementation error that can make interoperability between vendor solutions difficult.

For example, consider some of the issues affecting Ethernet over electrical connections, such as DAC cables and KR backplanes. For electrical interfaces, before any communication can start, the link end points must be set up with matching configurations. This can be based on a fixed setting, but in some

cases the end points "negotiate" with each other to find the best possible configuration supported by both ends. This is referred to as Auto Negotiation (AN).

Next, to ensure the signal can be transmitted efficiently, link-training is performed where training sequence packets are exchanged. Link-Training (LT) became necessary with 25GbE NRZ signaling and has only increased in complexity and necessity with high-speed PAM4 signals. Because the SNR of PAM4 signals is reduced, errors in detecting symbols are more prominent, especially when transiting electrical links. This is addressed using complex equalization techniques at both the transmitter and the receiver, such as a transmission equalizer, where the amplitude of a sent symbol is adjusted based on both immediately preceding and following symbols. The two end points of the line can automatically adjust the transmission equalizer in the link-training process.

One of the issues that is faced as we move to higher speed implementations is that the transition from auto-negotiation to link-training can fail due to timing mismatches. If too much time elapses, the link will time-out and revert to the auto-negotiation phase again, ultimately repeating continuously.

Using traditional physical layer examination tools and techniques, we can determine if the SERDES is driving the correct electrical signals and within specification parameters, make timing calculations and get a general sense of the physical health of the device under examination. The traditional signal integrity test tools, however, are unable to "show" us the contents of the information exchanged. If there are issues establishing the link, and those issue are the result of auto-negotiation or link-training inconsistencies, the engineer is effectively blind.

This is just one example of the potential interoperability issues that can become more prominent as we move to higher Ethernet speeds. The ability to understand both the context and nature of errors at both the data link and physical layers require a more integrated Terabit Ethernet Testing approach.

IEEE "BEYOND 400GE" STUDY GROUP

IEEE has established a study workgroup named "Beyond 400G" or B400G[7], whose working objectives[8] are to prepare a Project Authorization Request (PAR) and Criteria for Standards Development (CSD) so work on IEEE standards for 800GE and 1.6TE can begin. Typically, this process takes around 5 years, so we should expect the availability of 800GE and 1.6TE standards after 2025.

In the meantime, industry consortia will play a vital role in providing specifications and solutions that are in line with IEEE standardization work but are pre-standard. The Ethernet Technology Consortium is enabling 800GE solutions today, while various Multi-Source Agreement (MSA) are driving solutions like new transceiver modules, such as the QSFP-DD and OSFP MSAs.

Other important initiatives include OIF projects that have influenced Ethernet specifications of the past and will impact work on Ethernet standards for Terabit Ethernet. The OIF currently has two

[7] Source: https://www.ieee802.org/3/B400G/index.html
[8] Source: Agenda and General Information (ieee802.org)

projects for Common Electrical Interfaces (CEI) operating at 112 Gbit/s and 224 Gbit/s respectively. The CEI-112G-LR project was launched in 2017 to enable "high-loss 112G backplane channels" as well as facilitate DAC cable channel links at 112 Gbit/s. At the same time, the CEI-112G-MR project focused on specifications for chip-to-chip interfaces. The CEI-224G project was launched in August 2020 to specify 224 Gbit/s interfaces. The focus of CEI is on enabling interoperability between different vendor solutions and builds upon available standards.

It is therefore clear that the path to Terabit Ethernet will be paved by a close collaboration between industry consortia and standards organizations to enable the availability of pre-standard solutions that are future and backward compatible. However, this also means that implementations will need to be thoroughly tested to ensure that they can interoperate with other pre-standards solutions.

THE IMPLICATIONS FOR TESTING BEYOND 400GE

To date, Ethernet testing of the physical and data link layers have been two separate activities. The tests performed at the data link layer were rarely influenced by physical layer issues. The introduction of FECs in 100GE changed that and with the introduction of the Reed Solomon FEC as a mandatory requirement in 400GE, data link layer testing now requires insight into the physical layer as well.

Measuring SNR and Bit-Error Rates (BER) for Ethernet connectivity has always been important, but takes on an additional significance when moving to PAM4 modulation. At 100GE/NRZ, for many links, BER was not an issue and FEC was not required, but as we move to 400GE and PAM4, BER becomes a bigger issue, even at short ranges, which is why FEC is mandatory.

This is not just an issue of performance, but also of interoperability. As discussed earlier, FEC is required to compensate for the reduction in SNR when moving to PAM4. This challenges Forward Feedback Equalizer (FFE) and Decision Feedback Equalizer (DFE) designs with DFE's struggling to close feedback loops because of the signal speed. This can make it difficult for two different implementations to interoperate if their DFE designs differ. Therefore, testing of these implementations and assurance of interoperability is paramount.

As we saw earlier, auto-negotiation and link-training for electrical Ethernet links is also an issue of concern. While some believe that optical links will replace electrical connections from the server to the Top-of-Rack switch, the availability of higher speed DAC and AEC cables with reaches over 2m or more means that electrical connections will still have a role to play in the future.

Above all, Terabit Ethernet Testing solutions will need to be flexible in accommodating a variety of different configurations with respect to baud rates, number of lanes and modulation schemes for both optical and electrical connections as well as different module formats. There is no one path to Terabit Ethernet, which means that Terabit Ethernet Testing needs to accommodate all relevant paths and options.

XENA'S PATH TO TERABIT ETHERNET TESTING

Ethernet test & measurement solutions need to be at the forefront of technology development to help manufacturers deliver next-generation Ethernet switch, line-card and transceiver solutions. It is not our role to predict which technology will ultimately dominate in the market place, which is why our Ethernet Traffic Generation & Analysis (TGA) solutions are designed for flexibility to support as many options as possible. This also goes for the Freya family of 800GE test modules, which is the newest addition to Xena's Valkyrie product line.

The Freya test modules support both optical transceivers and DAC cables, and there are versions that support either QSFP or OSFP form factors. This makes Freya a versatile test solution for performance and functional testing of 800GE, 400GE, 200GE and100GE network products using 112G SerDes (PAM4 112G) including transceivers, PHYs, switches, routers, NICs, TAPs, packet-brokers, and backhaul platforms.

Figure 9: Freya-800G-1S-1P test module

The Freya test modules are designed for thorough transceiver and PHY testing with comprehensive PCS and PMA layer test including advanced signal integrity views that provide visual information on the quality of the received signal. The test facilities also provide FEC statistics with a Pre-FEC error distribution graph, PRBS-31Q payload test pattern and alarms, user-defined skew insertion per transmit virtual lane, as well as user-defined virtual lane-to-SerDes mapping for testing of the Rx PCS virtual lane re-order function.

They also offer equalization controls for pre-emphasis, transmission attenuation and post-emphasis signal integrity analysis, and the option to auto-tune the receiver equalizer/CTLE. Error injection in the PMA layer is possible, along with link flap single short or repeatable link down events with millisecond precision.

Freya also provides full featured Ethernet test generation and analysis features at all supported rates for advanced Layer 2 and Layer 3 multi streams testing giving the user rich possibilities for configuring traffic profiles matching relevant test conditions. The Freya test modules provide

extensive statistics and graphs making it easy for the user to understand test results. In addition, Freya supports Auto-Negotiation and Link Training (AN/LT) interoperability testing.

The Freya family is based on the Ethernet Technology Consortium specification and the IEEE 802.3ck draft specifications for 800GE with support for RS-FEC (Reed Solomon) (544,514,t=15) according to IEEE 802.3 Clause 119 and Clause 134. Freya modules will fit in a Xena ValkyrieBay chassis and will be your steppings-tone towards future 1.6 Terabit Ethernet testing.

For information on Xena's Freya test modules, see: https://xenanetworks.com/800g-ethernet-pam4/

>> Read also part 1: ''Terabit Ethernet – Why?''

>> Book a consultation with a Xena tech expert