Testing 400G

using 112Gbps SerDes

by XENA Networks

This article covers the quest for Signal Integrity when testing 400G Ethernet using 112Gbps SerDes

introduction

The ever-increasing need for larger and larger bandwidth is driving the development of Ethernet connectivity at 400Gbps, 800Gbps and 1.6Tbps. Currently, 400Gbps is experiencing the strongest growth in deployments in data centres because the standard is in place and products are commercially available. Work on 800Gbps and 1.6Tbps is ongoing and we shall expect to see products and standards at those rates in the near future as well.

For Ethernet, there is currently no technology defined that can support transmission of 400Gbps on a single electrical cable or fibre. Therefore, a 400Gbps connection can be implemented as either 40 x 10Gbps, 16 x 25Gbps, 8 x 50Gbps and 4 x 100Gbps and the optimum choice depends on the cost per bit, the power consumption and the physical space requirement. In general, higher speed lanes wins on all three factors which is why most 400Gbps ports are implemented as either 8 x 50Gbps or 4 x 100Gbps.

At such high line rates, it becomes a challenge to maintain the signal integrity even over fairly short distances. Designers of silicon, connectors, PCBs and full switch modules must perform extensive simulations to ensure good signal integrity throughout the entire signal path and verify the design using advanced test equipment. Interoperability between switch modules from different vendors have to be tested to verify that differences in implementations do not cause signal integrity issues. Thus, it is no longer sufficient to test transmission of packets at Layer 2 and 3 but, your test equipment must also be able to test at the physical Layer 1 level.

Layer 1 is also termed the PHYsical layer (PHY) and is subdivided into three sublayers which are the Physical Coding Sublayer (PCS), the Physical Medium Attachment (PMA) sublayer and the Physical Medium Dependent (PMD) sublayer. The various functions inside these sublayers are integrated with a Serialiser/Deserialiser (SerDes) and therefore, the word SerDes is often used to describe the PHY.

In this White Paper, we start by describing the basics of line, lane, symbol and data rates, as well as modulation formats for 400Gbps Ethernet transmission. Then the key elements of a typical ADC- based SerDes implementation for 100Gbps Ethernet are described, followed by a brief introduction to signal integrity with focus on the challenges relevant for 100Gbps.

Next, the parameters that need to be tested to verify signal integrity are described in detail, along with some practical use

cases for 400Gbps connections. Please note that while the focus of this White Paper is 400Gbps, most of the content is relevant for testing both 800Gbps and 1.6Tbps as well.

Line, lane, symbol rates and data rates

It is easy to get confused by the many terms used for transmission rates for Ethernet. Before we discuss the specifics of implementations and signal integrity issues let's run through the various definitions.

Physical media, lanes and lines

The lowest level is the physical medium which can be either electrical or optical as shown on Figure 1.

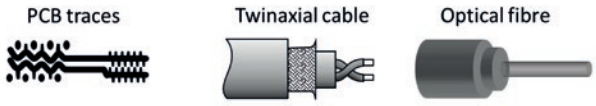

Figure 1: Examples of physical media.

A single lane consists of two physical media connections since the communication is always bidirectional. A line can be a single lane but, above 10Gbps it is common to use multiple lanes per line. See Figure 2 for a simple overview of the definitions of physical medium, lane and line.

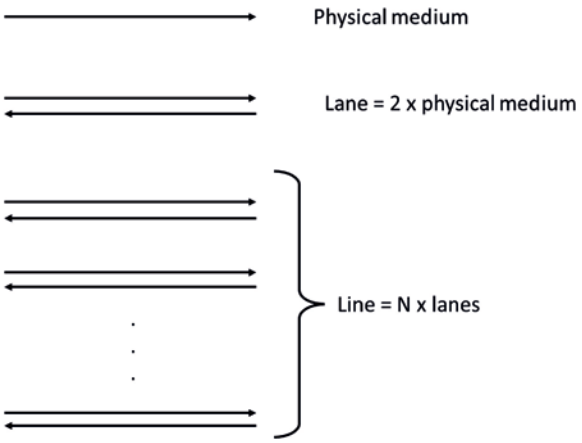

Figure 2: Illustration of physical medium, lane and line.

A lane is sometimes called a SerDes even though the SerDes is actually a function inside the transceiver device.

Symbol and data rates

The way digitally coded information is transported is through symbols. The most common type is binary symbols (bits) where the symbol value can be either "1" or "0" or even combinations of bits like "00", "01", "10", and "11". The symbol rate is simply the number of symbols being sent per second and is measured in Baud.

The data rate is the amount of data sent per second. In the simple case where each symbol period contains only one bit, the data rate is equal to the symbol rate but, in the case where each symbol contains two bits, the data rate is twice the symbol rate. The data rate is measured in bits per second (bps).

100Gbps, 106.25Gbps and 112Gbps rates

The Ethernet data rate is 100Gbps but, in order to transmit this high data rate on a single lane some extra information (overhead) needs to be added to the pure Ethernet data as defined in the IEEE 802.3ck standardisation work. This overhead amounts to 6.25% and the actual data rate of 100Gbps Ethernet per lane is therefore 106.25Gbps.

The standard SerDes used for 100Gbps Ethernet is however defined for 112Gbps and are therefore labelled 112Gbps SerDes. The reason is that the standard SerDes used for transmitting 106,25Gbps Ethernet directly is also used for the Optical Transport Network (OTN) described in ITU-T recommendation G.709. OTN is a transport protocol that adds additional overhead bits on top of its pay-load hence, the lane rate for OTN is defined to 112Gbps rate.

400Gbps connection using 112Gbps SerDes

Figure 3 illustrates one example of an application requiring a high-speed Ethernet connection. The example is a simple connection from one switch port, over a physical external cable to another switch port. Inside the switch, the high-speed connections are implemented using multiple parallel lanes in order to reduce the per lane speed to a manageable speed for the electronics.

On the transmission cable however, it is desirable to transmit as few lanes as possible to reduce the complexity of the cable and connectors. The function of the Serialiser/Deserialiser (SerDes) on Figure 3 is, as the name says, to combine the many lower speed lanes into fewer high-speed lanes at the transmitter and split the high-speed lanes into multiple parallel lower speed lanes at the receiver.

When discussing high-speed Ethernet connections, we will be focusing on the SerDes function.

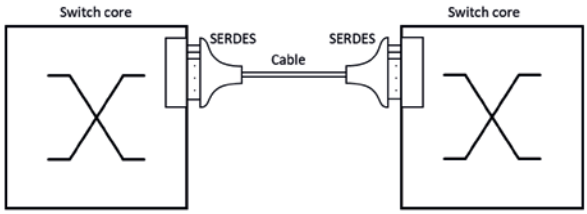

Figure 3: Connection between two switch ports.

Number of lanes on the cable

Although desirable, currently there is no Ethernet technology that can transmit 400Gbps as a single lane and 400Gbps is therefore implemented using multiple parallel lanes at lower speeds. Currently, the highest possible single lane rate is 112Gbps although work on 224Gbps is ongoing. This means we need 4 lanes to implement 400Gbps Ethernet.

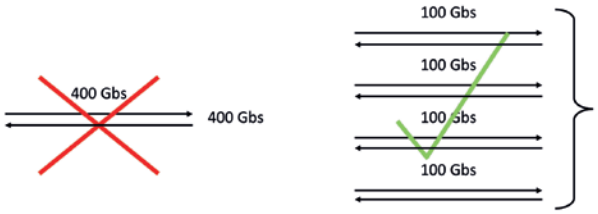

Figure 4: The smallest number of lanes for a 400Gbps connection is 4 bi-directional 100Gbps lanes.

As mentioned in the introduction it is also possible to implement 400Gbps as 8 lanes of 50Gbps. Figure 4 indicates that all the lanes are bidirectional.

Modulation format

A common modulation format used for Ethernet is Non-Return-to-Zero (NRZ) where each symbol period carries information about a single bit – i.e either a "1" or a "0" as illustrated on Figure 5. The fastest single lane NRZ symbol rate defined for Ethernet today is 25.78125 GBaud so, for each 100Gbps connection we would need 4 parallel lanes of 25.7825 GBaud (and 16 for 400Gbps) reaching a data rate of 103.125Gbps as defined in 802.3bj.

However, a new modulation format has recently been defined in 802.3ck using 4-level Pulse Amplitude Modulation (PAM4) with a symbol rate of 53.125 GBaud leading to a data rate of 106.25Gbps on a single lane.

The PAM4 coding scheme is illustrated on Figure 6 and uses four voltage levels, where each level represents a two-bit number. In this way the number of bits transmitted per second is doubled compared to NRZ modulation.

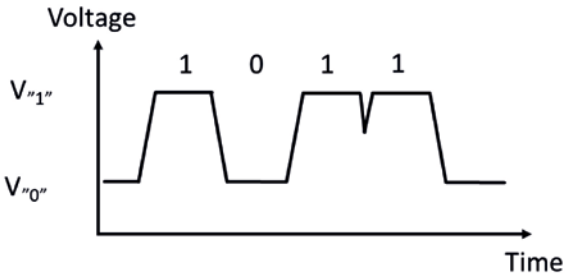

Figure 5: NRZ bit sequence with two voltage levels: $V_{"0"}$ represents a "0" and V"1" represents a "1".

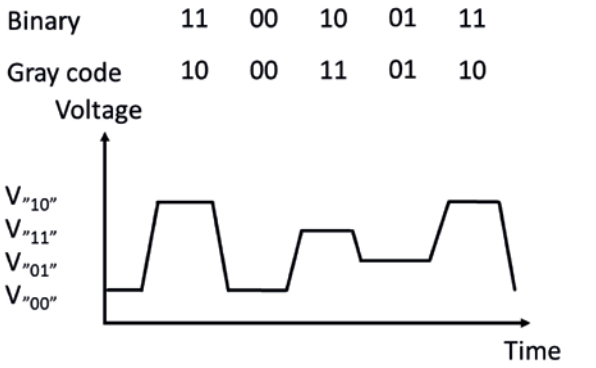

Figure 6: PAM4 Gray-coded bit sequence with four voltage levels that each represent a two-bit value.

As seen from Figure 6 the highest voltage level is representing the value "10" and not "11" as one would intuitively expect. This form of coding is called Gray-coding and is used to reduce bit errors. The Gray code ensure that a shift between two neighbouring voltage levels only leads to a change of 1 bit out of the two in the symbol. This means that such a shift in voltage level can never lead to more than one bit error.

A pre-coding scheme that encodes the PAM4 symbols at the transmitter and decodes them at the receiver is also implemented in order to reduce the possibility for consecutive errors.

The decoding scheme calculates a new value to be transmitted based on the current symbol period and the symbol from the past symbol period using an algorithm like the one shown on Figure 7. At the receiver a similar scheme is used to remove the precoding again.

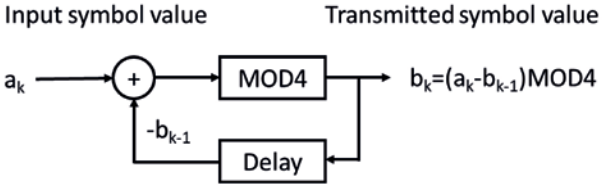

Figure 7: Precoding scheme at the transmitter.

Block-coding and Forward Error Correction (FEC)

As the 100Gbps signals are transmitted they get distorted which eventually leads to bit errors at the receiver. At a lane rate of 100Gbps and a bit error rate (BER) of 10-12 a bit error will occur every 10 seconds. For this reason, the BER requirement for 100Gbps is being increased to 10-14 in the IEEE802.3ck standards work.

In order to obtain these low levels of BER, both block coding as well as Forward Error Correction is required. Essentially, both coding schemes works by adding additional bits to the signal that provides extra information that the receiver can use to correct errors.

For 100Gbps Ethernet as described in 802.3ck the 64b/66b coding is transcoded into a 256b/257b block before the Reed-Solomon Forward Error Correction (RS-FEC) is added. This is done to reduce the amount of overhead used by block coding in order to leave more overhead for a stronger PAM4 FEC. The 256b/256b transcoding groups four 64b/66b block together, removes the two-bit headers and add a single bit header to the whole 256 bits block.

Figure 8: Transcoding four 64b/66b data blocks into one 256b/257b block.

Forward Error Correction is a coding technique used to detect and correct errors in a bitstream by adding redundant bits and error-checking code at the transmitter. At the receiver the FEC decoder uses these extra bits to detect erroneous bits and correct them.

A FEC block consists of n FEC symbols whereof k FEC symbols are the actual data and the remaining (n-k) FEC symbols are code and redundant bits – also denoted parity bits. A FEC is therefore characterised by the ordered pair (n,k).

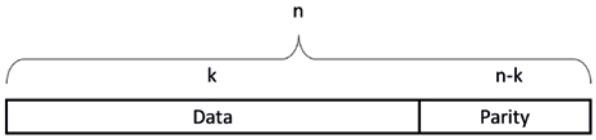

Figure 9: (n,k) FEC codeword.

The amount of errors a given FEC can correct depends on the actual implementation. For 802.3ck 100Gbps Ethernet a Reed-Solomon (RS) (544,514, t=15) FEC is used and the FEC symbols are 10 bit wide. This means that 30 parity FEC symbols are added for every block of 514 data FEC symbols. The t=15 refers to that this FEC can correct up to 15 FEC symbol errors each containing up to 10 bit errors.

The total overhead of the 802.3ck 100Gbps Ethernet is 6.125% meaning that the actual lane rate is 106.125Gbps.

Signal integrity basics

When a signal is transmitted from one switch port to another as shown in Figure 3, the signal is degraded by several factors. The signal level will be reduced due to the inherent resistance of the wires causing gradual loss of signal as the signal moves from the transmitter to the receiver.

Limitations in the bandwidth will lead to Inter-Symbol Interference (ISI) and inductive coupling between electrical lanes as well as connectors lead to cross talk. Impedance mismatches cause reflections (Return Loss). At both the transmitter and receiver jitter can occur and finally the signal will be degraded by thermal noise.

NRZ Eye diagram

Signal integrity is a measure of the quality of the transmitted signal. It is commonly analysed using a so-called eye-diagram which is constructed by overlapping the pulses in a bit sequence as shown for a NRZ-modulated signal on Figure 10. The signals are sampled at the centre of the pulses and if the voltage level is higher than V_{ref} it is interpreted as a “1” and if it is below V_{ref} a “0”.

With no distortion on the transmission the eye is nice and open as on Figure 10. However, in a real system the eye will

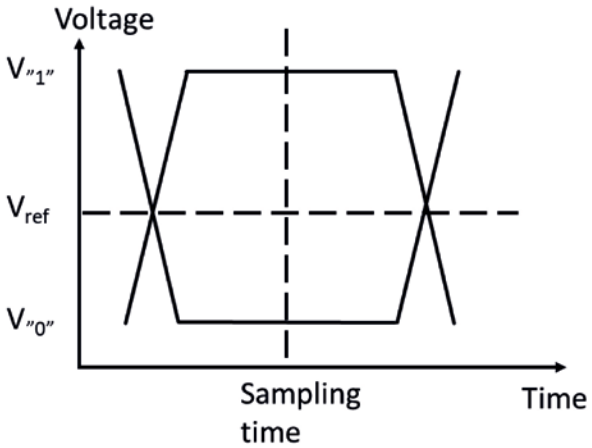

Figure 10: Ideal eye diagram showing reference level as well as sampling time.

look more like the one shown on Figure 11 where some of the "1"s will be mistaken as "0"s and thus lead to bit errors.

In general, impairments to the signal level such as loss and noise will cause the eye to "close" in the vertical direction whereas delay and jitter impairments will cause the eye to "close" in the horizontal direction (See Figure 12).

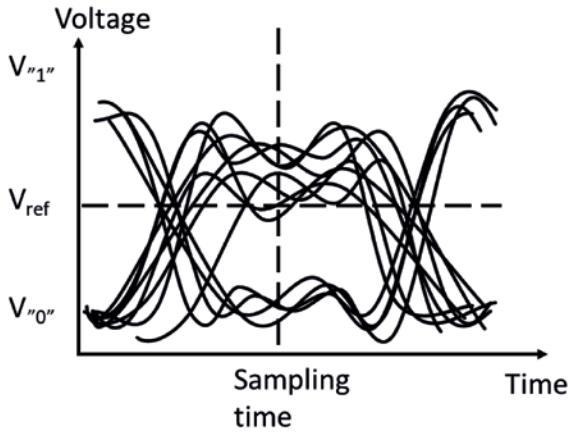

Figure 11: Practical example of eye diagram.

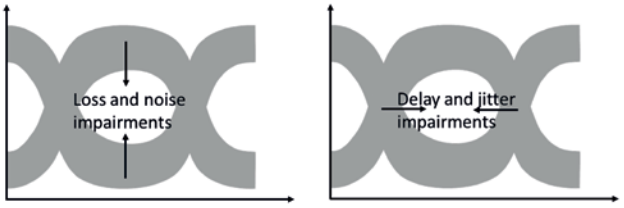

Figure 12: Illustration of eye closure due to Loss and noise (left) and delay and jitter (right).

Signal integrity for 12Gbps with pam4 modulation

In the previous section, the basics of signal integrity for NRZ modulation was discussed. For 112Gbps PAM4 modulated signals however, the situation is a bit more complicated. Since PAM4 has four levels, the eye diagram consists of three eyes

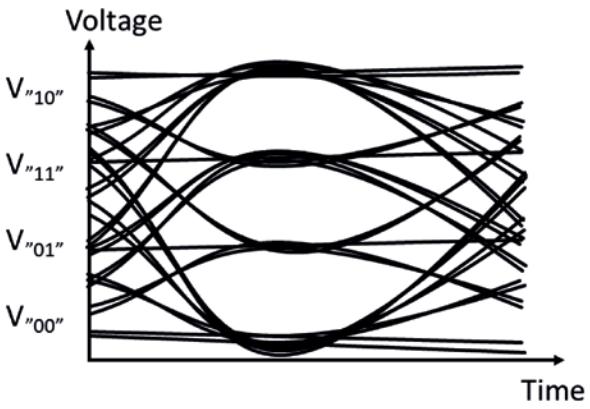

Figure 13: Eye diagram of PAM4 signal.

rather than one as shown on Figure 13. Obviously the PAM4 eyes are much narrower than the NRZ eye which means that PAM4 is less tolerant to distortions than NRZ.

For very high speeds like 112Gbps the pre-FEC eye diagram is likely going to be almost closed and difficult to use for analysing signal integrity. Instead, a vertical slicer eye diagram is used. The vertical slicer eye is constructed by plotting the sampled, recovered signal level for a large number of samples as shown on Figure 14.

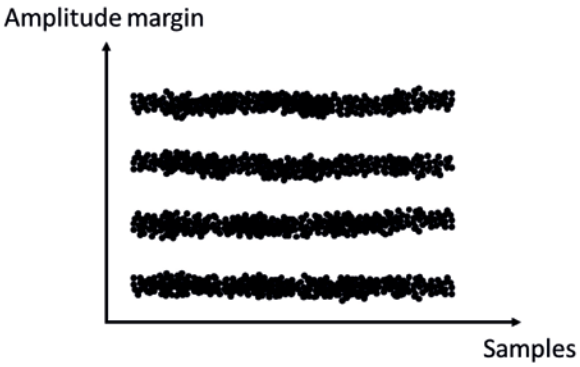

Figure 14: Sampled eye diagram for PAM4 modulated signal.

To understand how the high-speed signals are recovered, we will use the block diagram on Figure 15 to describe the various functions inside the PCS/PMA sub-layers of the receiver. Basically, the single analogue 112Gbps lane is de-serialised into 64 parallel digital 875Mbps lanes. However, high speed SerDes also includes various types of equalisation as well as analogue to digital conversion before the data and clock is recovered.

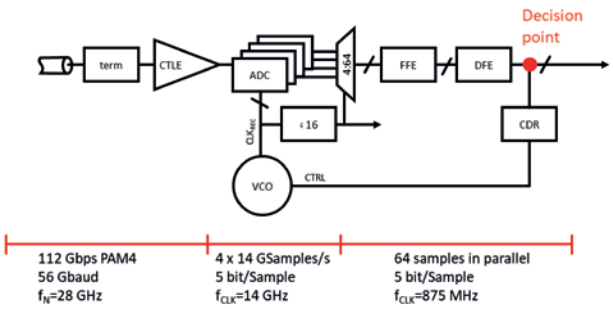

Figure 15: Block diagram of 112Gbps PAM SerDes Rx.

Continuous Time Linear Equaliser (CTLE)

At high data rates the received data stream is severely distorted and requires reconstruction (equalisation) before it can be sampled. In general, higher frequencies will be attenuated more than lower frequencies and the function of the equaliser

is to reverse this frequency response. Only frequencies below the Nyquist sampling frequency (28GHz for 56 GBaud) are relevant to equalise.

One common equaliser approach used in transmit and receive circuits is a Continuous Time Linear Equaliser (CTLE). The CTLE attenuates low-frequency signal components and amplifies components around the Nyquist frequency as shown on Figure 16. The result is that the frequency response becomes almost flat for frequencies below the Nyquist frequency.

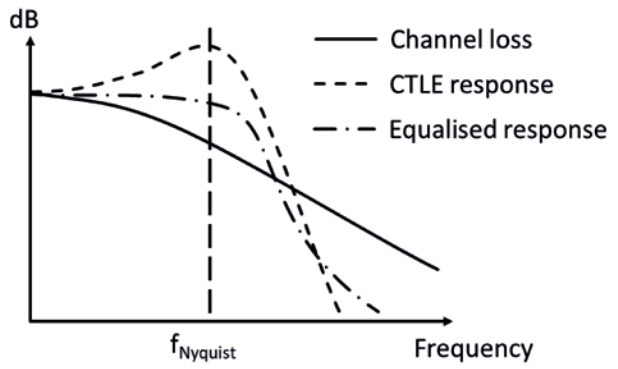

Figure 16: Frequency reponse.

Time interleaved ADCs

The analogue voltage levels of the equalised PAM4 signals are digitised using Analogue-to-Digital-Converters (ADCs) that sample the pulses in the middle of the pulse duration. In order to lower the speed requirement of the ADC it is common to use time interleaved ADCs. In the case shown on Figure 15 four ADCs each sampling at one quarter of the 56 GBaud rate – i.e at 14 GSamples per second. The sampling times are controlled by a voltage-controlled oscillator (VCO) that uses the clock rate derived from the Clock and Data Recovery circuit (CDR).

The ADC's sampling time is phase shifted by 0, 1, 2, and 3 full speed clock periods (0.18 nsec for 56 Gbaud), respectively. Effectively, this means that ADC1 samples bit 0, 4, 8, …, ADC2 samples bits 1, 5, 9, …, ADC3 samples bits 2, 6, 10, … and ADC4 samples bits 3, 7, 11, … See Figure 17 for a timing diagram.

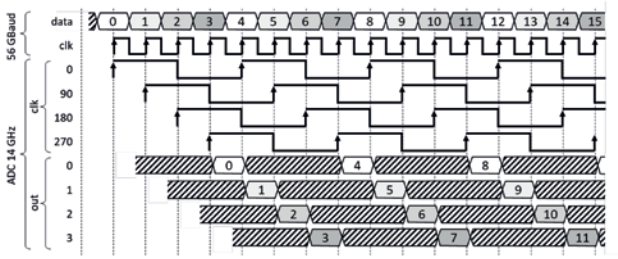

Figure 17: Timing diagram of 4 time-interleaved ADCs.

The digital outputs from the four ADCs are demultiplexed into 64 parallel 875Mbps lanes.

FFE and DFE equalisation

Besides the CTLE which is done at full line rate on the analogue signal another level of equalisation is done on the reduced rate 64 parallel digital signals. These equalisers are typically Feed Forward Equalisers (FFE) and Decision Feedback Equalisers (DFE). These equalisers take a longer bit sequence into account during the equalisation and can thereby reduce Inter-Symbol Interference (ISI).

Clock and Data Recovery (CDR)

The Clock and Data Recovery (CDR) extract the clock frequency and the actual data (bit sequence) from the signal.

How to test for signal integrity at 112Gbps?

Testing the performance of high speed ethernet connections and ports starts by testing signal integrity. As described earlier, signal integrity cannot be measured directly at the analogue high-speed level but has to be extracted from the recovered data. For this reason, the right tool for the task is an Ethernet Traffic Generation and Analysis (TGA) tester with a 112Gbps PAM4 port module and ADC-based SerDes.

Figure 18: Freya-800G-4S-1P module.

Xena's Freya test modules (Figure 18) support testing of 112Gbps SerDes PAM4 based Ethernet at data rates of 100G / 200G / 400G / 800G. Freya is designed for switch, transceiver and PHY design validation & Quality Assurance.

The key features of the module are:

- 4-speeds: 800GE, 400GE, 200GE, & 100GE
- Dual media: QSFP-DD800 & QSFP112
- Supports 112G SerDes (PAM4 112G)
- Test with optics and DACs
- Extensive L1 features
  - Auto-Negotiation & Link Training (AN/LT)
  - Advanced signal integrity view
  - RS-FEC (544,514, t=15)
  - FEC error correction chart
  - PRBS31Q
  - Equalisation controls

Test set-up

Figure 19 shows a simple example where the Device Under Test (DUT) is a single switch port being tested by connecting it to a ValkyrieBay chassis equipped with a Freya-800G-4S-1P test module.

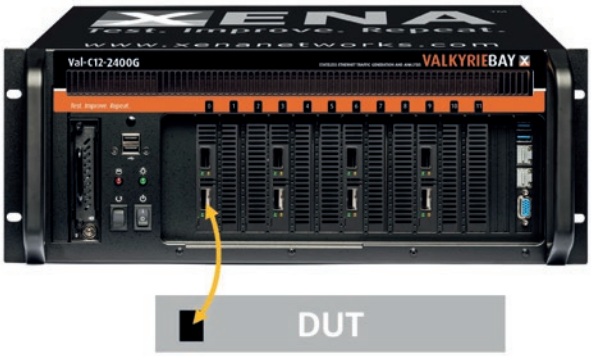

Figure 19: Example of test set-up for single port testing.

Auto Negotiation and Link Training

Auto Negotiation (AN) is a key feature of Ethernet whereby two end-points share information over a link and thereby determine the highest common denominator among the end-point's capabilities.

Link Training (LT) is another process that can take place when a device is connected to a high-speed Ethernet port through a copper cable or backplane. In this case it is important that the characteristics of the transmitted signal are tuned to be optimally carried over the copper cable.

A good starting point for your test is to test the DUT's port and the tester is to enable AN and test if AN can complete with a satisfactory result.

If you are testing over a Direct Attach Cable (DAC) it can furthermore be beneficial to use LT to tune the parameters of the transmitted signal for optimum performance.

Advanced PHY configuration

During testing it may be required to control and monitor the equaliser settings of the SerDes at both transmitter and receiver. This can be used in the case where automatic TX-tuning auto-negotiation is not supported or to test how a transceiver performs using various TX equaliser settings.

PRBS testing

Cables and connectors can be tested using a Pseudo Random Binary Sequence (PRBS) rather than a full Ethernet packet stream. For NRZ modulation a PRBS31 (sequence of 2^{31}-1 bits) is used but, for PAM4 there are two bits per symbol period which means we need an even number of bits in the sequence. For this reason, PAM4 uses a PRBS31Q sequence (where Q stands for quaternary) consisting of two consecutive PRBS31 sequences. The PRBS31Q can be used to check the BER of each physical lane and insert bit errors at the transmitter.

Lane	PRBS Lock	PRBS Bits	PRBS Errors	Error Rate
All	N/A	1.32e+13	342	2.58e-11
0	LOCK	1.66e+12	10	6.04e-12
1	LOCK	1.66e+12	0	<3.02e-12
2	LOCK	1.66e+12	0	<3.02e-12
3	LOCK	1.66e+12	0	<3.02e-12
4	LOCK	1.66e+12	6	3.62e-12
5	LOCK	1.66e+12	325	1.96e-10
6	LOCK	1.66e+12	1	6.04e-13
7	LOCK	1.66e+12	0	<3.02e-12

Figure 20: Example of PRBS test screen.

Pre-FEC Error Distribution diagram

The pre-FEC error distribution diagram plots the amounts of FEC blocks with 0, 1, 2, 3 etc symbol errors before the error correction as illustrated by Figure 21. Also, you can see that the BER was 1.28 10^{-8} before the FEC. After FEC, there are no bit errors which is indicated by the post-FEC of <1.02 10^{-13}.

As mentioned earlier, the maximum amount of errors that the RS(544,514, t=15) FEC can correct is 15 symbol errors.

> **A good starting point for your test is to test the DUT's port and the tester is to enable AN and test if AN can complete with a satisfactory result.**

Figure 21:Example of BER histogram (from XENA Networks … manual).

Signal integrity view

As mentioned previously, for ADC-based SerDes the traditional eye diagram is not useful and instead the signal integrity view shown on Figure 22 is used. From such a plot you can see if the four PAM4 levels are clearly distinguishable.

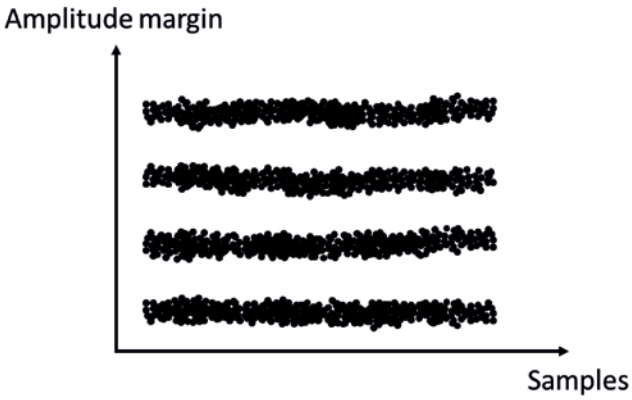

Figure 22: Sampled eye diagram for PAM4 modulated signal.

Layer 2 and Layer 3 testing

For high-speed Ethernet it is crucial to check the signal integrity at Layer 1 as described in this white paper before doing any other tests.

Once a good signal integrity has been established it is of course still important to perform tests of the data formatting functionality (Layer 2) and network path functionality (Layer 3) for which the Freya-800G-4S-1P module provide a wide range of powerful features. However, the description of these tests is outside the scope of this white paper.

Summary

Testing 400Gbps and 800Gbps using 112Gbps SerDes requires focus on the physical layer (Layer 1) because the signals are heavily degraded even after fairly small transmission lengths. Also, the advanced modulation schemes (PAM4) and ADC-based SerDes functions used, require your ethernet test equipment to extract signal integrity measures from blocks inside the PCS/PMA taking several types of equalising and error correction coding into account.

Xena's new Freya series of test modules support all the functionality you need in order to test devices and links at both Layer 1, Layer 2 and Layer 3. For Layer 1 the relevant signal integrity parameters to test for 112Gbps SerDes based systems includes Auto Negotiation, Link Training (for electrical cables), advanced PHY equaliser tuning, PRBS testing, pre-FEC error distribution and the signal integrity view.

For information on Xena's Freya test modules, see
https://xenanetworks.com/freya-ethernet-testing-for-pam4-112gbps-serdes-for-100g-200g-400g-800g/