Critical 100 Gigabit Ethernet Testing: Equalization, DP-QPSK, and Crosstalk Compliance

Ransom Stephens, Ph.D.

Introduction: Back to the Future

In the last decade, high speed serial links have all but replaced parallel interconnects. Problems caused by skew and crosstalk outweighed parallel systems ability to increase data rate by simply adding another lane.

Now, as design engineers polish the remaining rough edges from the third generation of serial technologies like USB3 and PCI-Express Gen 3, the lure of combining the advantages of serial technology [1] – like noise-reducing differential signaling, jitter-resistant embedded clocks, and eye-opening equalization – with parallel scalability is too much to resist.

The robustness of high speed serial combines with parallel scalability and fiber optic reach in 40 and 100 GbE (GbE = Gigabit per second Ethernet). On the electrical side, the 40 and 100 Gb/s transmission rates are predominantly four and ten parallel lanes of 10 Gb/s each, respectively, but it's inevitable that 40/100 GbE designs will incorporate electrical lanes at rates of 20-28 Gbit/s. To this end, the Optical Internetworking Forum Common Electrical I/O (OIF-CEI [2]) working group has supplemented the 40/100 GbE specification [3] to include electrical lanes up to 686 mm long at 20-28 Gb/s.

The historic problems of parallel technology, namely skew and crosstalk, have been addressed. Skew has been pushed up the protocol stack to a level where it's no longer a layout issue. Crosstalk, on the other hand, can't be shuffled under the rug so conveniently. The 40/100 GbE specification [4] prescribes maximum levels of integrated crosstalk noise and minimum interference tolerance.

Some of the usual High Speed Serial (HSS) test issues – clock recovery, equalization, jitter and noise – carry over to 40/100 GbE with extra complications. For many HSS designers, the addition of fiber optics presents further complications.

This paper concentrates on how 40/100 GbE deals with the challenges presented by layering back-to-the-future parallel design on top of modern HSS technology. Our primary focus is on the electrical side of the technology, right up to the electro-optical interface at 25 Gb/s.

We begin with an overview of the serial/parallel geometry, electrical/optical mix, and distance requirements of 40 and 100 GbE. This takes us to the problem of skew, how embedded clocking works with multiple lanes, nuances of de-emphasis and equalization at these rates, the electro-optical interface and a clever alternative to wavelength division multiplexing, DP-QPSK, and then, finally, how crosstalk is specified and tested. We conclude with test ideas that can unleash a designer's creative genius.

Overview of 40 and 100 GbE

Figure 1 shows key 100 GbE geometries and Table 1 lists the primary attributes for different categories of 40 and 100 GbE. Short reach designs incorporate 10 × 10 Gb/s on shielded cables, medium reach has 10 Gb/s signals on each of ten multimode fibers, and both long and extended reach have 4 × 25 Gb/s signals Wavelength Division Multiplexed (WDM) onto one single mode fiber.

Notice that 40 GbE includes a backplane formulation. No backplane formulation for 100 GbE has been specified, but lack of specification will not prevent design engineers from building them. It's inevitable that implementations of 100 GbE shown in Figure 1c will require 4 x 25 Gb/s chip-to-chip and/or chip-to-module between the SER/DES and optoelectronic components over at least several cm of Printed Circuit Board (PCB).

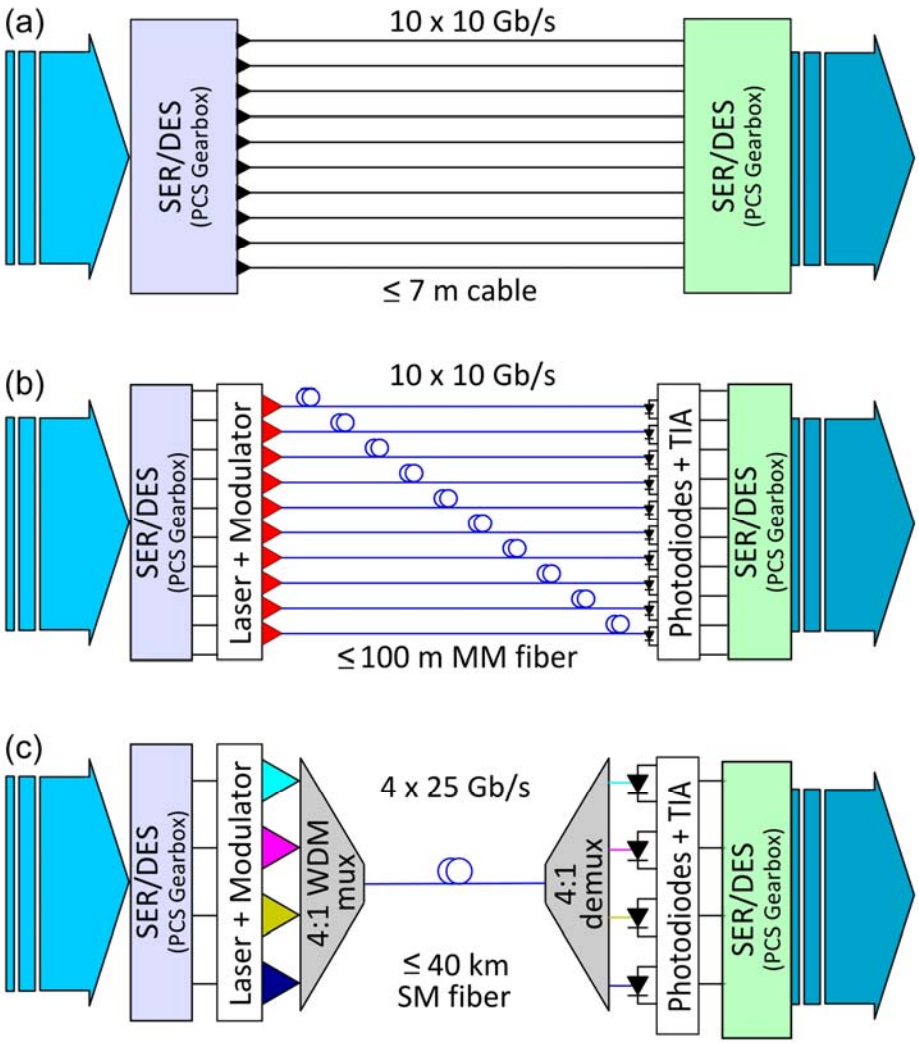

Figure 1: Diagrams of the 100 GbE topologies.
(a) 10 cable lanes, (b) 10 fiber optic lanes at 10.3125 Gb/s each, (c) 4 lanes at 25.78125 Gb/s each.

RATE	CONFIGURATION	MEDIUM	MAX RANGE
40 Gb/s	4 × 10 Gb/s	Backplane	1 m
		4 shielded cables	7 m
		4 multimode fibers	100 m
		WDM on 1 single mode fiber	10 km
100 Gb/s	10 × 10 Gb/s	10 shielded cables	7 m
		10 multimode fibers	100 m
	4 × 25 Gb/s	WDM on 1 single mode fiber	10 km
		WDM on 1 single mode fiber	40 km

Table 1: Summary descriptions of the eight 40 and 100 GbE geometries.

Skew

Skew is the variation of the propagation times of parallel lanes. In traditional parallel technology, skew had to be limited to a small fraction of a bit period to prevent data from being misaligned at the receiver.

In 40/100GbE, skew is mitigated by shuffling the problem from the physical medium level of the protocol stack to the Physical Coding Sublayer (PCS). Data is multiplexed and assigned to parallel lanes by a "PCS gearbox." The PCS gearbox at the transmitter encodes alignment markers in the data so that the receiver gearbox can reconstruct the original signal in the presence of scores of bit periods of physical skew. This scheme tolerates nanoseconds of skew which means that the length of parallel physical lanes can vary by feet, not millimeters. Therefore, skew is not a big problem at the physical layer of 40/100 GbE systems.

Clocking

As discussed in *AN-25* [1], much of the power of high speed serial technology is in the receiver's ability to recover a clock signal from the data. By using a recovered clock, both the waveform and the clock used to distinguish logic levels carry the same low frequency jitter. Hence, jitter frequencies below the clock recovery bandwidth shouldn't cause errors.

Each receiver in a 40/100 GbE system recovers an independent clock from logic transitions in the signal it receives. Similarly, each output lane can be driven with a recovered clock. The result is that there is no requirement that the parallel lanes have fixed phase or frequency relationships.

The ability of a receiver to recover a clock embedded in the signal is predicated on two issues: First, since the clock is recovered by sampling the timing of transitions in the data, the signal must have sufficient transition density. Transition density is the number of logic transitions per bit transmitted. If there is insufficient transition density, the receiver will lose clock lock. Second, to prevent baseline drift, the signal must carry a balanced signal (an equal number of 1s and 0s) over the bandwidth of the clock recovery circuit. That is, the system must have 50% mark density.

To maintain sufficient transition density and even mark density, the PCS gearbox applies 64B/66B encoding and self-synchronized scrambling to the data. This scheme guarantees balanced mark density and at least one transition in every 64 bits.

De-Emphasis and Equalization

At the physical layer, we can think of each lane as an independent High Speed Serial system. The signals are low-swing, AC coupled, differential signals with de-emphasis at the transmitter and equalization at the receiver.

De-emphasis is a form of equalization applied at the transmitter. It is defined as the ratio between the peak-to-peak amplitude following a logic transition to the amplitude during a non-transition bit, the Voltage Modulation Amplitude (VMA), Figure 2. Since the levels of three symbols – that in question plus those following and preceding – determine the level of each transmitted symbol, this type of de-emphasis is called a "three tap filter." By increasing the voltage swing at transitions, the high frequency components of the signal are boosted. Since the transmission path is essentially a (very) complex filter whose gross characteristic is its low-pass nature, boosting high frequencies at the transmitter helps open the eye at the receiver.

$$\text{De-emphasis} = 20\log\left(\frac{V_{PP}}{\text{VMA}}\right)$$

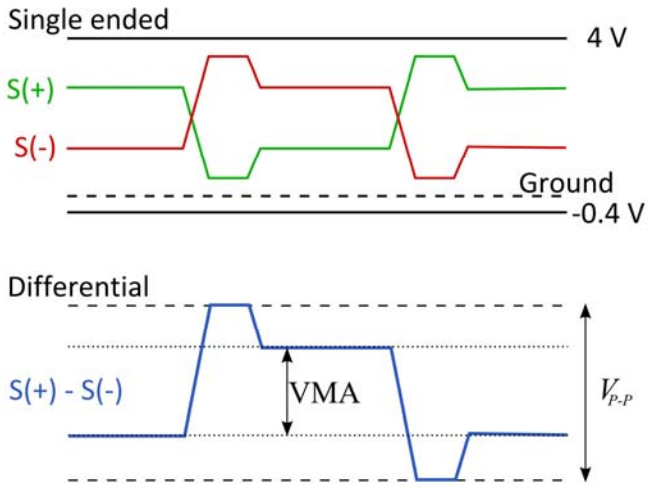

Figure 2: Transmitter de-emphasis

To maximally compensate the frequency response of the channel, transmitter de-emphasis is programmable. During initialization, a known training sequence is transmitted. By communicating with the receiver through a back channel, the transmitter uses an iterative process to determine the most effective de-emphasis scheme.

Receiver equalization schemes are not specified in 40/100 GbE but are likely to include both a Continuous Time Linear Equalizer (CTLE), which can be implemented as an analog filter, and a nonlinear, digital Decision Feedback Equalizer (DFE). The equalizer parameters are tuned on the training sequence during initialization.

By including an initialization training sequence, the authors of the specification have invited implementation of exotic, proprietary, adaptive equalization schemes.

The electro-optical interface and DP-QPSK modulation

A crucial, not to mention clever, approach to the electro-optical interface is Dual Polarization Quadrature Phase Shift Keyed optical modulation. At radio frequencies, QPSK involves varying the phase of each cycle of the carrier. At optical frequencies, a factor of ten-thousand higher than radio frequencies, it's different. If we could vary the phase at optical frequencies we'd be talking about a million GbE not 100 GbE.

In optical QPSK, the carrier, a laser beam, is split in two. The two beams have a fixed phase relationship, Figure 3. Each beam enters a solid state interferometer (e.g., a Mach-Zehnder). Each beam enters a different leg of an interferometer. An NRZ signal is applied to one leg of the interferometer. The electrical signal varies the index of refraction of that leg. The light in that leg then lags behind the beam in the other leg. When they are recombined at the far end of the interferometer, the result is a small phase shift in the optical carrier. The shifted beams from the two interferometers are then combined, each with one of four well-defined phases – a QPSK signal.

By applying QPSK encoding to each of the two planar light polarizations separately, four signals can be transmitted on one optic fiber: DP-QPSK. The resulting topology is much like Figure 1c but with the optical WDM multiplexer and demultiplexer replaced by a DP-QPSK modulator and demodulator respectively.

Figure 4 shows a test configuration for 4 × 25 Gb/s DP-QPSK opto-electronics. Independent data patterns are applied to each input of the DP-QPSK modulator and four error checkers measure the respective Bit Error Rates (BERs) of the demodulated outputs; BER < 10^{-12} is required.

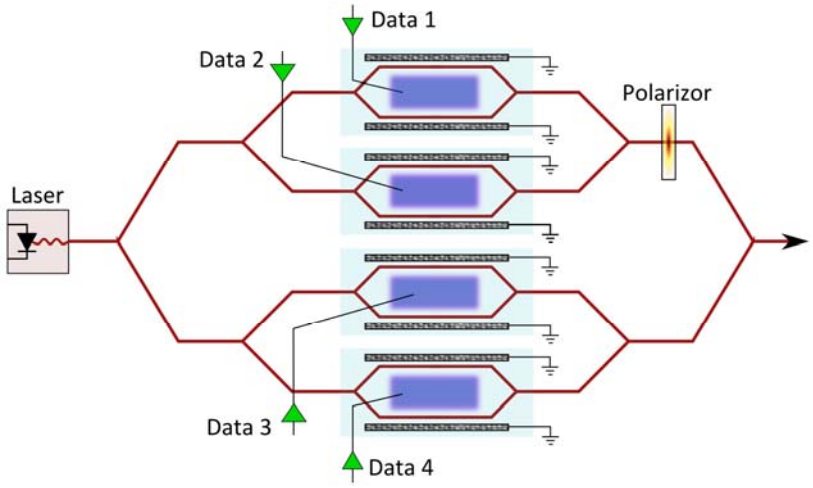

Figure 3: Dual Polarization-Quadrature Phase Shift Keyed modulation of four optical signals onto one fiber.

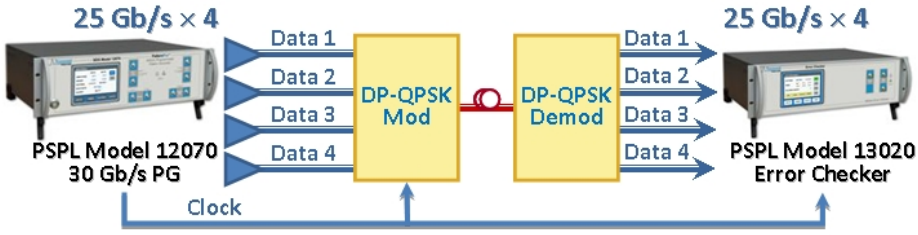

Figure 4: DP-QPSK Testing.

Crosstalk

Crosstalk occurs when a signal is affected by the transmission of data on neighboring transmission lanes. In 40/100 GbE vernacular, the signal lane is called the *victim* and the interfering lanes are *disturbers* (a.k.a., aggressors).

Crosstalk is caused by the coupling of electromagnetic radiation between the victim and disturbers. Electromagnetic radiation is generated whenever an electric field changes. Since data signals at 10 and 25 Gb/s have frequent, sharp transitions, there are a lot of changing electrical fields. The faster the transition, the greater the crosstalk amplitude and the shorter its duration. The eye-diagrams to the left in Figure 5, a and c, show victims whose disturbers have fast rise/fall times. To the right, Figure 5b and d, the disturbers have slower rise/fall times. The time-delay position of the crosstalk pulse is determined by the phase relationships between disturber and victim. In Figure 5a and b the disturbers are half a bit-period out of phase with the victim and in Figure 5c and d the victim and disturbers are in phase. Multipath interference smears the eye-diagram trajectories which mimics random jitter and noise.

In systems where each lane is driven by an independent clock, jolts of crosstalk appear smeared across the victim's eye diagram. Most jitter and noise analysis techniques mistake crosstalk for Random Jitter (RJ). But crosstalk, unlike RJ, is bounded and deterministic.

In any case, crosstalk simply can't be shuffled higher in the protocol stack the way that skew is.

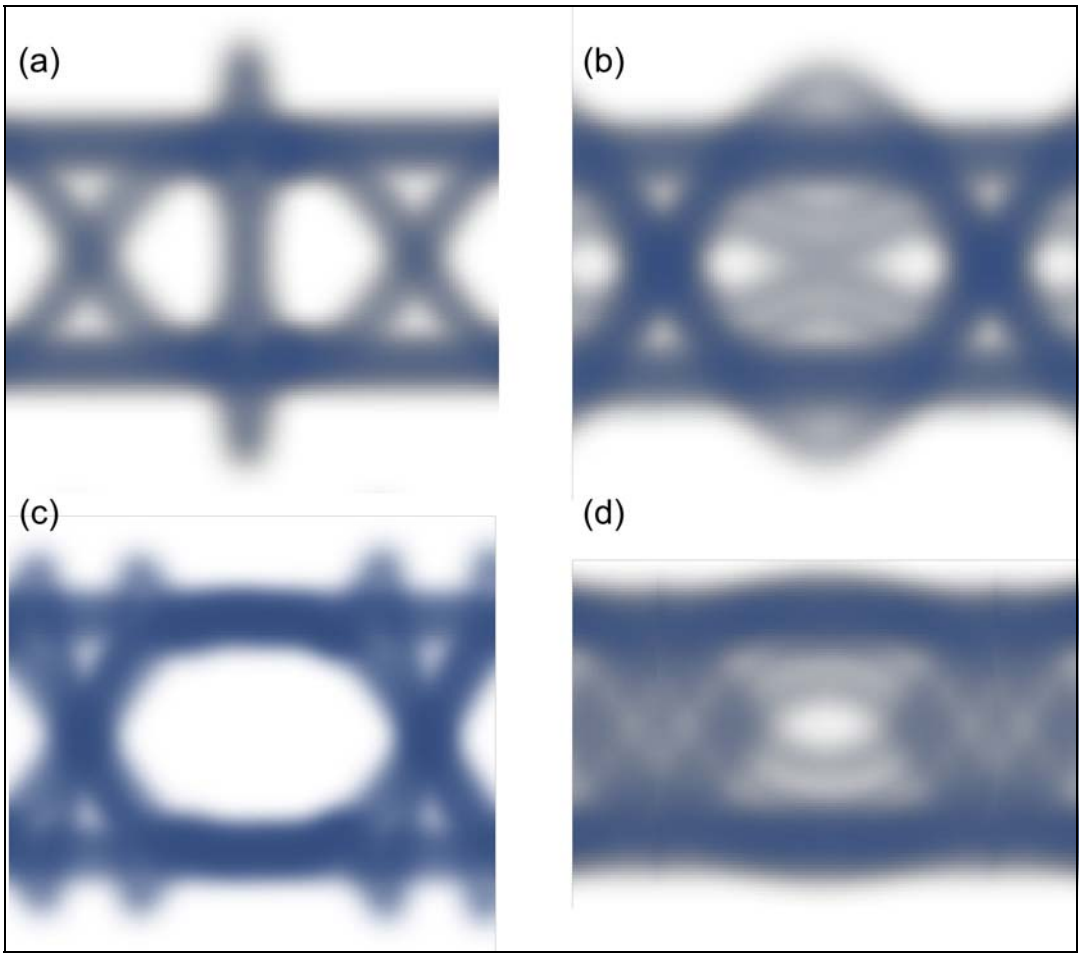

Figure 5: Crosstalk characteristics for phase-locked victim+disturbers. Fast rise times on the left (a and c), slow on the right (b and d). The relative phase of the disturber is at the center of the eye on top (a and b) and in the crossing point on the bottom (c and d).

Crosstalk is usually quantified in decibels as the ratio of the disturber power observed on the victim lane, that is, the disturber's crosstalk power, to the disturber power on its own lane.

There are two categories of crosstalk. Near End Crosstalk (called NEXT) is the ratio of the power picked up by the victim to the transmitted disturber power with both measured at the transmitter. Far End Crosstalk (FEXT) is the ratio of power measured on the victim lane at the receiver to the power of the disturber at the transmitter.

Since there are as many as 9 and no fewer than 3 disturbers, crosstalk on a single lane is given by adding the individual ratios of crosstalk power to disturber power for each disturber. For a 4 × 25 Gb/s system, the Multi-Disturber FEXT on lane 0 is given in decibels by the following equation:

$$\mathrm{MDFEXT}(f) = -10\log\left[\sum_{i=1}^{3}\frac{P_{0i}(f)}{P_i(f)}\right].$$

Integrated Crosstalk Noise

$$\sigma_{nx} = \sqrt{2\Delta f \sum_{n=1}^{N} \frac{A_{nt}^2}{f_n} W_{nt}(f_n) 10^{-MDNEXT(f_n)/10}} \quad \text{and} \quad \sigma_{fx} = \sqrt{2\Delta f \sum_{n=1}^{N} \frac{A_{ft}^2}{f_n} W_{ft}(f_n) 10^{-MDFEXT(f_n)/10}}$$

To calculate the RMS crosstalk noise, a minimal receiver response, $W(f)$, is applied to measurements of MDFEXT and MDNEXT measured over a wide span of frequencies.

The near end, σ_{nx}, and far end, σ_{fx}, RMS crosstalk noise is then given by the equations listed above, where A is the peak disturber differential output amplitude, Δf is the frequency step over which MDFEXT(f) and MDNEXT(f) are measured, and W is the weighting function of a minimal receiver response that incorporates simple filter and transfer functions.

The total Integrated Crosstalk Noise (ICN) is:

$$\sigma_{nx} = \sqrt{\sigma_{nx}^2 + \sigma_{fx}^2} \ .$$

Rather than specify inflexible cut-off values for both channel insertion loss and receiver crosstalk, a combination of the two is specified. The combination permits larger, if more complicated, ranges of both insertion loss and crosstalk which enables greater receiver design freedom.

Figure 6 shows the allowed values of ICN as a function of channel insertion loss measured at the data-rate clock frequency.

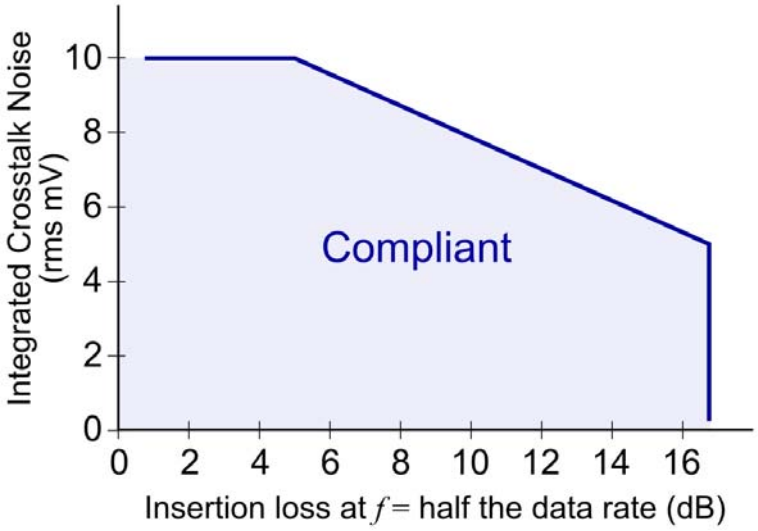

Figure 6: The Integrated Crosstalk Noise (ICN) template as a function of insertion loss measured at the single-lane data-rate clock frequency.

Interference Tolerance Testing

The *interference tolerance test* is essentially a stressed receiver tolerance test that incorporates crosstalk (c.f., *AN-25* [1]).

Two tests are performed to determine the receiver's ability to tolerate a combination of jitter and worst case interference. Both tests stress the receiver with the same levels of applied Sinusoidal Jitter (SJ), Random Jitter (RJ), the shortest permitted rise/fall times, and maximum Integrated Crosstalk Noise (ICN).

In addition, Test 1 includes the maximum allowed MDFEXT and moderate insertion loss. Test 2 includes moderate MDFEXT and maximum insertion loss. To vary the insertion loss in each test, two separate reference channels are required. The low and high loss channels are specified by their insertion loss frequency profile, $\mathrm{IL}(f)$. Figure 7 shows the test setup.

The tests require independent crosstalk signals on every disturber lane and each lane must pass both tests independently.

The disturber signals should emulate real-world signals as much as practical in voltage swing and rise/fall times. It's nice for the disturbers to be properly tuned de-emphasis but if they don't, that's okay as long as the specified RMS FEXT and ICN are achieved.

It is important that the applied disturber crosstalk signals be uncorrelated.

If more than one lane carries the same signal – for example, if the output of a pattern generator is split and applied to two lanes – then those two channels will have the same waveform. Since crosstalk is caused by radiation from transitions, the synchronized transitions will interfere with each other. Constructive interference might present a worst-case scenario but even this is not obvious. If the disturber-victim phase differences are such that jolts of crosstalk occur at the crossing point, there will be little effect on the bit error rate, Figure 5c and d. If they occur in the center of the eye, Figure 5a and b, then it's an unrealistic stress on the system [7]. Further, it's not necessary for each generator to operate from the same clock. Since recovered clocks can be used independently on each lane, they don't necessarily have a well defined phase relationship.

The 40/100 GbE specification requires that disturbers carry either scrambled idle codes or the standard 2^{31}-1 Pseudo-Random Binary Sequence (PRBS31). I recommend using a pattern developed by the OIF-CEI group [3], the "CID Jitter Tolerance pattern." The PRBS31 is inadequate for a few reasons, the most egregious being that the longest string of Consecutive Identical (CID) bits it carries is 31, but the 64B/66B encoding plus scrambling can result in CIDs of 64. Clearly, 64 straight logic ones or zeros is a demanding stress to clock recovery circuits and should be tested.

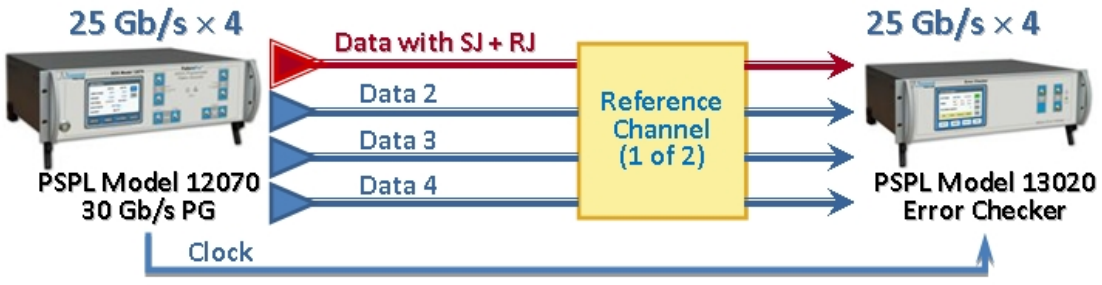

Figure 7: Interference tolerance test setup for 4 × 25 Gb/s 100 GbE with a four channel pattern generator [5], and four channel error checker [6].

Since PRBS31 is over two billion bits long, even at 28 Gb/s it takes 76 ms to repeat a cycle. It takes hours to generate enough cycles to provide the stress extremes that occur, however rarely, when periodic, sinusoidal, and random jitter combine with inter-symbol interference.

The CID Jitter Tolerance Pattern was composed by comparing transition densities and mark densities of different segments of the PRBS31 pattern with the most stressful segments of common scrambled OC-768 (from 40 Gb/s SONET/SDH) frames. It is far shorter and includes runs of 72 CIDs with 10328 or more bits from the standard PRBS31 sequence in a way that guarantees mark density balance.

The OIF-CEI test pattern better represents the worst case symbol sequences to which receivers will be subject in the field. Plus, a shorter pattern makes tests easier to repeat and verify which is a huge advantage when compliance testing reverts to diagnostic testing. Whichever pattern you choose should be repeated at least a million times. Receivers pass the test if they meet the key criterion: Bit Error Rate < 10^{-12}.

The Test Tools

In the back-to-the-future parallel world of 100 Gigabit Ethernet, the challenges presented by modern high speed serial testing have the added complication of an electro-optical interface and, on center stage, crosstalk. In addition to standard bench tools – like high bandwidth oscilloscopes – 40/100 GbE compliance and diagnostic testing requires high performance BER testing. Crosstalk testing requires pattern generators with some sophistication, in particular:

1. ***Multiple channels*** *with* ***independent*** *outputs*
2. ***Adjustable*** *inter-channel* ***phases***
3. *Differential outputs with* ***variable amplitudes****, at least 500 mV to 2 V*
4. ***Applied jitter stresses*** *including: Sinusoidal Jitter (SJ) and Random Jitter (RJ)*
5. ***PRBS31*** *patterns plus the ability to generate* ***user defined patterns*** *1-2 Mbit long*
6. ***Variable crossing point*** *to apply Duty-Cycle Distortion (DCD)*
7. *Adjustable* ***rise/fall times***

The BER < 10^{-12} requirement should be checked for each of the tests described in this paper, but you might not need a complete Bit Error Ratio Tester (BERT). A standard BERT is depicted in Figure 8 consisting of a high fidelity pattern generator and error checker. Figure 9 shows Picosecond Pulse Labs' modular system that has all of the sophistication but can be configured to suit your specific needs.

For the interference tolerance test and crosstalk system debugging, you'll need ten independent pattern generators for the 10 × 10 Gb/s 100 GbE and four for the 4 × 25 Gb/s 100 GbE or 4 × 10 Gb/s 40 GbE implementations. Since each lane has to be tested independently, you may not need more than one error checker for each test station.

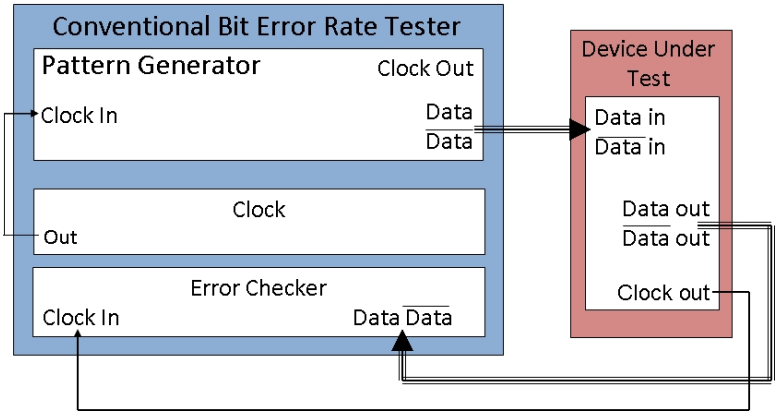

Figure 8: Conventional single channel BERT loop-back test configuration.

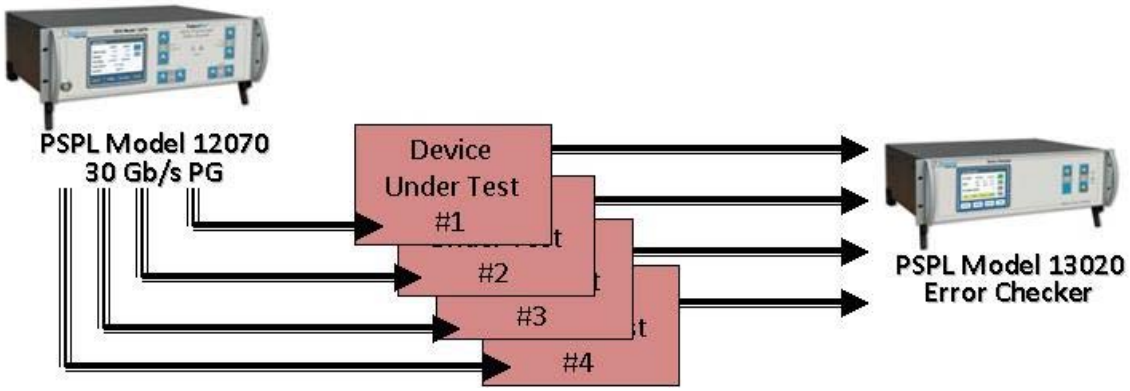

Figure 9: Modular, configurable multi-channel BERT system consisting of a PSPL 12070 multi-channel pattern generator [5] and PSPL 13020 multi-channel error checker [6].

You may not need an error checker at all. Cyclic Redundancy Checks (CRC) and Forward Error Correction (FEC) are integral features of 40/100 GbE so in many cases the system you're testing can count its own errors. The obvious advantage of a modular approach is that you can add equipment as it's needed to reduce initial and overall costs.

Testing electro-optic interfaces like DP-QPSK modulators or WDM Mux/Demux, on the other hand, requires error checking on every lane. Since test time scales with the number of error checkers, it's worth considering the cost-time tradeoffs when setting up your lab.

Crosstalk BER Root Cause Analysis

Diagnostic testing is usually more interesting than compliance testing. By developing a deeper understanding of how crosstalk affects your design you can take your system beyond the specification in quality and price.

Have another look at Figure 5. It illustrates the extreme problems of crosstalk in an idealized way. The three disturbers in Figure 5 are in phase with each other as well as phase-locked to the victim. In Figure 5a and b the disturbers are perfectly out of phase with the victim. Since the electromagnetic interference jolt occurs in the eye center, this configuration takes the highest BER hit. In Figure 5c and d, the

disturbers and the victim are all in phase and the interference lands on the crossing point leaving the center of the eye comparatively undisturbed.

In a real system, the phases of the parallel lanes need not be fixed and, even if they are, won't be so conveniently placed as they are in Figure 5. Consequently, finding the root causes of crosstalk problems is notoriously difficult. However, by mapping the BER as it varies with different inter-lane phases and test pattern sequences, you can isolate problems.

The idea is to scan the relative phases of each lane while measuring their individual BERs. The most sensitive lanes will have higher overall BERs. BER peaks will occur at the most problematic relative phase combinations. You can probe the causes of the BER peaks by applying different test patterns with the phases fixed. For example, a test pattern with low transition density can isolate clock recovery problems in the presence of crosstalk. Similarly, low mark density test patterns probe a receiver's sensitivity to drift.

If the system you're designing will have lanes operating on a common clock, you can increase design margin by tuning inter-lane skew to avoid phase relationships that lead to BER peaks. The maximum skew variation needn't exceed a bit period, so this is a fine adjustment that shouldn't impact the physical-coding sublayer and how it deals with the larger issue of skew.

Crosstalk root-cause analysis can also be used to enhance your designs. Recall that the 40/100 GbE specification includes an initialization sequence of known patterns that can be used to train adaptive equalizers. If your receiver needs to operate with transmitters and cable/backplane assemblies over which you have no control, testing your receiver response under all disturber phase and pattern relationships will assure interoperability and may inspire you to create a proprietary crosstalk-canceling equalization scheme.

Conclusion

To summarize, 40 and 100 GbE put a parallel layer of complexity over the success of High Speed Serial data technology. Most of the test issues common to third generation HSS resurface with the added complexity of crosstalk.

The 40/100 GbE specification combines insertion loss and crosstalk tolerance to allow greater design freedom at the expense of more complex compliance testing. With that complexity come greater challenges and opportunities for designers to perfect their designs.

References

[1] "Introduction to Precision Analysis of High-Speed Serial Systems and Components," By Ransom Stephens, Application Note 25, published by Picosecond Pulse Labs, December-2010: http://picosecond.com/objects/AN-25.pdf

[2] The Optical Internetworling Forum Common Electrical I/O working group and their work can be found here: http://www.oiforum.com/

[3] OIF CEI specifications: http://www.oiforum.com/public/documents/OIF_CEI_03.0.pdf

[4] The latest draft of the 40/100 GbE specification – at the time this note was written – can be found here: http://standards.ieee.org/getieee802/download/802.3ba-2010.pdf

[5] For information about the PSPL Model 12070 high rate mutli-channel Pattern Generator, visit: http://www.picosecond.com.

[6] For information about the PSPL Model 13020 multi-channel Error Checker, visit: http://www.picosecond.com.

[7] "Characterizing, anticipating, and avoiding problems with crosstalk," Ransom Stephens and Al Neves, DesignCon2006: http://ransomsnotes.com/CrosstalkAnalysisDesignCon2006-RansomStephensAndAlNeves-v1.0.pdf