Optical Transceiver Interoperability Concern On 25 GbE Links Utilizing 25GAUI C2M ⇔ PSM4/Lane SMF Interface

Vijay Srinivasan
Technical Leader, Intel® EPG Link Applications Engineering, Hillsboro, Oregon

1.0 Interoperability Concern Synopsis

This report is based on a 25GAUI C2M link between an Intel® Ethernet 700 Series Network Adapter and a Switch link partner realized through 25GBASE-LR interface to one lane of a 100G PSM4 transceiver in breakout configuration. Investigation into a link failure led to the observation that stable optical domain clock recovery in a PSM4 lane receiver may not be guaranteed even with compliance to IEEE Std. 802.3-2022 [1] and 100G PSM4 Ver. 2.0 2014 [2] specifications being true for constituent link segments. The goal of this report is to elucidate root cause analysis and issue characterization as an interoperability concern based on a combination of bench measurements and theoretical considerations.

2.0 Link Topology and Applicable Standards

The connection topology follows a typical 25 GbE link commonly observed in server-to-switch applications which is comprised of a 25GAUI C2M optical link involving two signal domain conversions, electrical-optical (E/O) and optical-to electrical (O/E), in each direction through optical transceivers. For analysis, the link is divided into segments, S1-3, consisting of a channel bounded by transmit and receive functions associated with its signal domain. These segments meeting standards-based specification requirements is an accepted minimum threshold and a necessary condition to allow concatenation for constructing a viable end-end link. The link segments and signal domains for the 25GBASE-LR interface to a PSM4/Lane (from QSFP28 100G PSM4 module breakout) studied here is shown in Figure 1 along with reference to standards in Table 1.

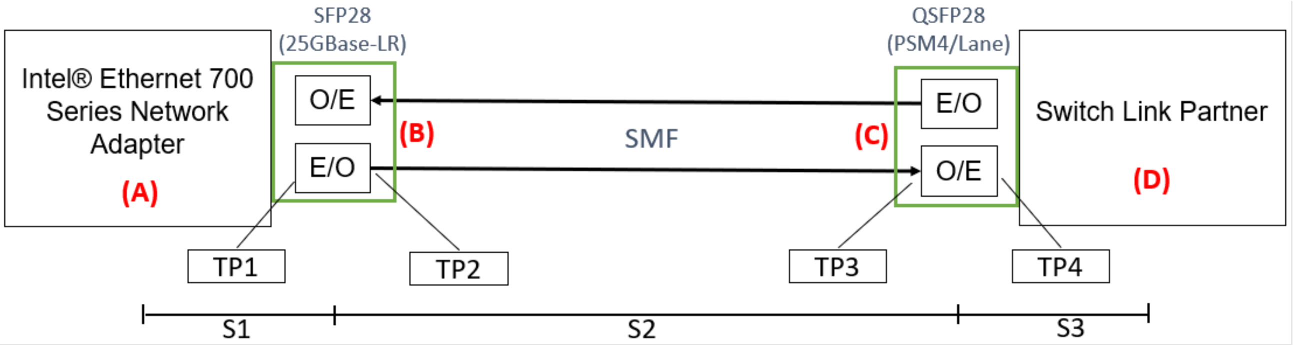


Figure 1. *25GAUI C2M link segments, signal domains, and test point locations*

Link Element	**Node**	**Reference to Standard**
Intel® Ethernet 700 Series Network Adapter (NIC)	A	25GAUI C2M – IEEE Std. 802.3-2022 Annex 109B
25GBASE-LR SFP28 module	B	25GBASE-LR – IEEE Std. 802.3-2022 Clause 114
100G PSM4 QSFP28 module	C	100G PSM4 Version 2.0, September 2014
Switch Link Partner	D	25GAUI C2M - IEEE Std. 802.3-2022 Annex 109B

Table 1. *Reference to standard and clauses applicable to nodes A-D*

3.0 Link Failure Isolation to PSM4/Lane O/E Function

Link issue isolation began with checking link fault indication at either end points of the link. The two nodes A (Intel NIC) and D (Link Partner Switch) were confirmed to be at Remote Fault (RF) and Local Fault (LF) states respectively, which unambiguously associates issue to the signal path direction from NIC transmitter (TX) to Switch receiver (RX). Furthermore, clock and data recovery unit (CDR) lock status indication in the two transceivers and at the Switch receiver diagnostics helped isolate issue to the optical domain and localize point of failure to be within the optical transceiver in the PSM4 module, specifically this module's O/E path. The isolation flow is summarized in Table 2.

Check Points	**Observations**	**Inferences**
MAC Link Fault Status	Intel NIC (RF) Link Partner Switch (LF)	NIC RX CDR Locked (node A) 25GBASE-LR O/E CDR Locked (node B) PSM4/Lane E/O CDR Locked (node C) Issue in NIC TX to Switch RX direction
Switch RX CDR Lock Status	Unstable	Issue in NIC TX to Switch RX direction
25GBASE-LR E/O CDR Lock Status	Locked	Issue not in nodes A and B
PSM4/Lane O/E CDR Lock Status	Unstable	Issue localized to node C O/E function

Table 2. *Link issue localization indicators and inferences*

4.0 PSM4 Transceiver Lane CDR LOL Root Cause Analysis

With issue localized to node C PSM4/Lane O/E CDR Loss of Lock (LOL), failure analysis proceeded with qualification of link element combinations spanning nodes A to D to detect systematic trigger, if any, for observed failure. The link element combinations and their impact are summarized in Table 3.

Link Elements and Combinations			**Impact: PSM4/Lane CDR LOL?**
Node A	**Node B (SFP28 - 25GBASE-LR)**	**Node C (QSFP28 – 100G PSM4)**	
Intel® Ethernet 700 Series Adapter [NIC SKU N1]	Any	Vendor P1	Yes**
		Vendor P2 and P3	No
Intel® Ethernet 700 Series Adapter [NIC SKU N2]	Any	Any	No
Intel® Ethernet 800 Series Adapter [NIC SKU N3]	Any	Any	No
** PSM4/Lane CDR LOL event likelihood and rate observed to be proportional to number of active lanes. **Note:** Link Partner (Node D) confirmed not relevant to failure at preceding stage (Node **C**).			

Table 3. *Interoperability matrix for detecting issue selectivity*

As evident from Table 3, the pairing of NIC SKU N1 with PSM4 module from Vendor P1 is a unique trigger for the interoperability concern. Focus on NIC SKU N1 transmitter (node A) was the logical next step owing to it being the signal source and results from electrical compliance test against IEEE Std. 802.3-2022 Annex 109B 25GAUI C2M for the three Intel NIC SKUs listed in preceding table are captured in Table 4.

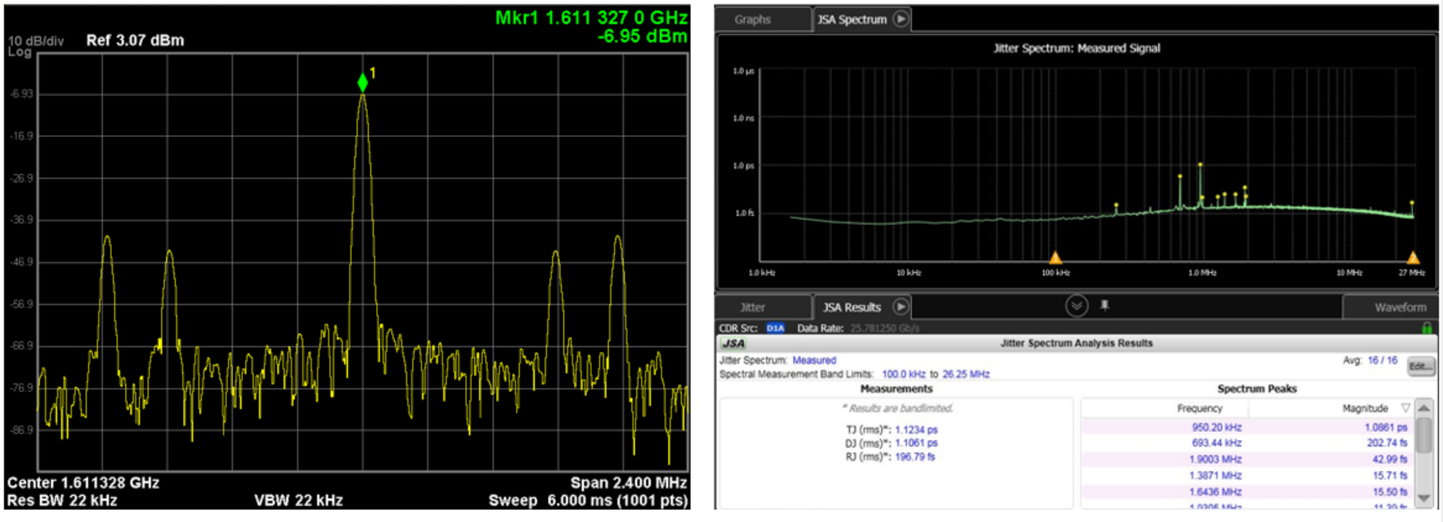
Intel NIC Variant	**Measured Eye Width (UI)**	**Min. Eye Width Specification (UI)**	**IEEE Std. 802.3-2022 Annex 109B Compliant? [Compliance Point: TP1a (Transmitter)]**
NIC SKU N1	0.706 - 0.744	0.46	**Yes**
NIC SKU N2	0.780 - 0.814		**Yes**
NIC SKU N3	0.796 - 0.831		Yes

Table 4. *NIC SKU transmitter conformance test summary*

Data in Table 4 shows all the three NIC SKUs are compliant to 25GAUI C2M transmitter specifications at TP1a along with disparity in measured eye width (EW). Jitter histogram captured with a SQUARE pattern (for better clarity as it eliminates contribution from inter-symbol interference (ISI) component) revealed distinct elevation in periodic jitter (PJ) contribution. The highest levels of PJ could be confirmed to be present on NIC SKU N1 transmitter output exhibiting lowest EW. Transmitter SQUARE pattern output spectrum analysis, Figure 3, showed spectral peaks detectable above the noise floor at frequencies listed in Table 4 which contribute to cumulative PJ accounting for the range of measured EW disparity.

Intel NIC Variant	PJ Components	Amplitude (Relative to -70dBm Noise Floor)	Notes
NIC SKU N1	693KHz	22dB	Switching Regulator Supply Noise
	954KHz	28dB	Unknown Source
	156.25MHz	5dB	PHY Reference Clock Input
NIC SKU N2	156.25MHz	5dB	PHY Reference Clock Input
NIC SKU N3	51KHz	40dB	Unknown Source
	840KHz	20dB	Switching Regulator Supply Noise
	156.25MHz	5dB	PHY Reference Clock Input
Note: Noise floor is in relation to peak level at Nyquist frequency for SQUARE pattern			

Table 4. *NIC SKU transmitter periodic jitter (PJ) frequencies and amplitude*



CDR Src: D1A Data Rate: 25.781250 Gb/s

Jitter Spectrum Analysis Results

Jitter Spectrum: Measured

Spectral Measurement Band Limits: 100.0 kHz to 26.25 MHz

Avg: 16 / 16

Measurements

* Results are bandlimited.

TJ (rms)*: 1.1234 ps
DJ (rms)*: 1.1061 ps
RJ (rms)*: 196.79 fs

Spectrum Peaks

Frequency	Magnitude
950.20 kHz	1.0861 ps
693.44 kHz	202.74 fs
1.9003 MHz	42.99 fs
1.3871 MHz	15.71 fs
1.6436 MHz	15.50 fs

Figure 2. *NIC SKU N1 periodic jitter spectrum analysis*

The PJ disparity between NIC SKUs N1 and N2 was determined to be due to different clock oscillator parts populated on these two NICs and confirmed by systematic part swap exercise on few samples of each SKU. Comparative examination of reference clock output in the two SKUs showed the oscillator in SKU N1 had poorer phase noise profile stemming from feedthrough energy from sources intrinsic to NIC design, such as regulator switching frequency at 693KHz, that elevates PJ and so reducing EW but not significant enough to derail transmitter conformance at TP1a. The largest PJ component measured (~1ps, 954KHz) was also well below module input SJ amplitude limit of 20ps at that frequency (see IEEE Std. 802.3-2022 Clause 114 Table 114-10).

Having established transmitter compliance and the source of jitter disparity between NIC SKUs N1 and N2 the electrical characterization at TP1a (node A) was complete. The next point of examination at the optical output of 25GBASE-LR module (node B) showed no concerns with two different test methods summarized in Table 5.

Test Method	Connections	PASS/FAIL Metric	Result
Link Stability	NIC SKU N1 Port to Port NIC SKU N1 to 25G Switch Port	1 Hour Duration Absence of Link Flap ***AND*** Absence of RS FEC Corrections	PASS
Optical Eye Mask	NIC SKU N1 TX -> Module -> 1km SMF -> Optical Eye Capture (Figure 3.)	1 Hour Duration Pattern 3 (PRBS31) Absence of Mask Hits	

Notes:

1. Tests above included 25GBASE-LR optical modules from three different vendors.
2. Duration of 1 hour >> failure rate (few/second) on link involving NIC-SKU N1 and Vendor P1 PSM4 module.

Table 5. *Characterization of NIC SKU N1 interface to 25GBASE-LR module in isolation*

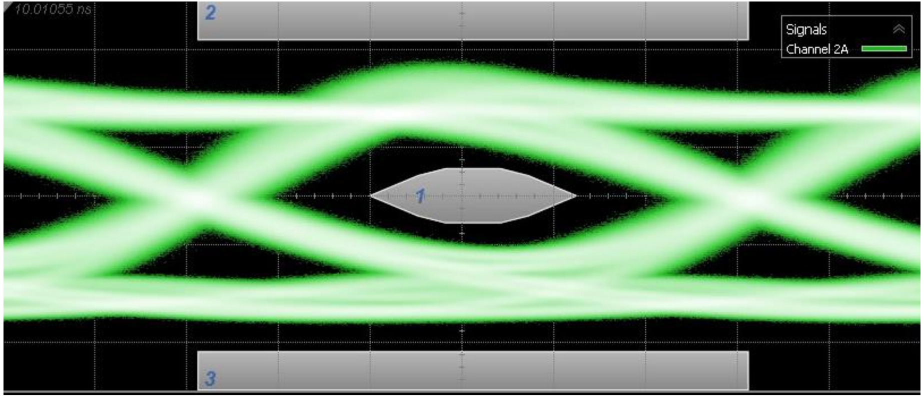


Figure 3. *25GBASE-LR Optical eye mask at Node B (see Table 5, Row 2)*

The passing result in Table 5 allows the following inferences to be drawn when link is established through two 25GBASE-LR modules in sequence:

(a) Elevated PJ associated with NIC SKU N1 transmitter did not impact:
- Electrical-Optical (E/O) conversion stability (TP2, first module interfaced to NIC)
- Optical-Electrical (O/E) conversion stability (TP4, second module interfaced to link partner)

(b) Choice of 25GBASE-LR optical module is not a factor (also implicit in Table 2 for PSM4/Lane CDR LOL event)

Noting that inference (a) above breaks down when the second 25GBASE-LR module is replaced with Vendor P1 PSM4/Lane receiver, few factors differentiating the two cases are summarized in Table 6 below.

O/E Factor	Difference Introduced	Impact to O/E CDR Function
Stressed Receiver Sensitivity (SRS) SJTOL Specification	PSM4: SJ at 200MHz only 25GBASE-LR: SJ at multiple frequencies from < 100KHz to > 100MHz	NIC-SKU N1 TX PJ inside SRS SJTOL limits? 25GBASE-LR → 25GBASE-LR: Yes 25GBASE-LR → PSM4/Lane: No
Single vs multi-lane Intrinsic noise level in module	25GBASE-LR lower than PSM4 Non-monolithic O/E realization with off-chip routing may introduce coupling among PSM4 lanes	Potential for elevated LA output DCD in PSM4 (independent of optical ISI/WDP contributions to DCD)
SMF Reach	PSM4: Up to 500m 25GBASE-LR: Up to 10km	O/E function more robust in 25GBASE-LR
SJ/PJ amplification from 25GBASE-LR E/O CDR	None. Equal, if present, in both 25GBASE-LR→25GBASE-LR and 25GBASE-LR→PSM4/Lane connections tested	Minimal. E/O CDR BW (10 MHz) >> NIC-SKU N1 TX PJ (693KHz, 954KHz)

Table 6. *List of factors differentiating LR and PSM4 modules*

The only factor in Table 6 relating NIC SKU N1 transmitter jitter generation to differing O/E function performance observed with 25GBASE-LR and 100G PSM4 modules is the stressed receiver sensitivity (SRS) sinusoidal jitter tolerance (SJTOL) specification applicable to those modules. An explanation for how difference in SJTOL frequency span is relevant can be derived by considering the theoretical principles governing CDR operation summarized as follows:

1. Widely adopted CDR implementation for ethernet applications employs VCO based architecture (Figure 4.).

2. The classification of CDR type is based on phase detector used ("bang-bang" or "binary" being prominent).

3. The loop bandwidth of the CDR for 25GAUI C2M application is fb/2578, where fb is the baud rate, or 10MHz [3].

4. CDR loop lock status is based on comparing the average phase detector output against a voltage threshold [4].

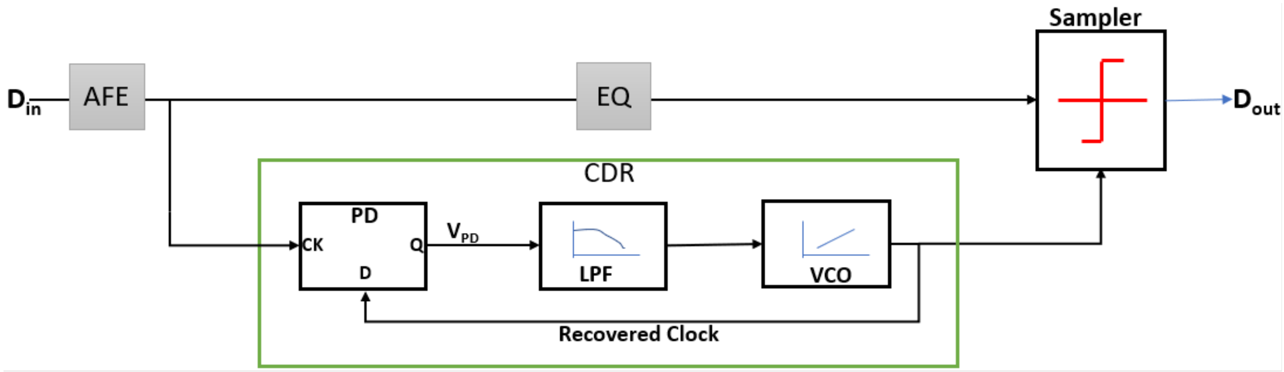


Figure 4. *VCO-based CDR architecture with DFF Bang-Bang Phase Detector (PD)*

5. Inherent delay in CDR loop causes jitter tracking ability to be jitter frequency dependent [3,4,5]:
   a. low frequency jitter (< 0.1*BW) tracked with large amplitude tolerance
   b. high frequency jitter (> 10*BW) tracked with low amplitude tolerance (reduces eye width)
   c. IEEE Std. 802.3-2022 Clause 114.7.10 25GBASE-LR specification for O/E SRS SJTOL reflects (a) and (b) above (same SJTOL specification also applies to 25GBASE-LR E/O function)

Frequency Range	**Sinusoidal jitter (SJ), peak-to-peak (UI)**
f < 100KHz	Not specified
100KHz < f < 10MHz	$5x10^5$ Hz / f
10MHz < f < 10 LB*	0.05

*LB=Loop bandwidth. Upper limit for added SJ should be at least 10*LB of the receiver being tested.

6. Implicit in SRS SJTOL specification is presence of stressor at a single frequency at any given time.

7. Measured NIC SKU N1 transmitter PJ shows multiple frequencies (693KHz, 954KHz, 156.25MHz) simultaneously present which is a condition neither required nor evaluated for SJTOL compliance.

8. Bode plot of a canonical second order loop [3,5,6] in Figure 3., applicable to CDR, is instructive for the role frequency dependent phase shift or delay plays on loop settling behavior [4] where regions A, B, C-D denote low, mid, and high frequency bands.

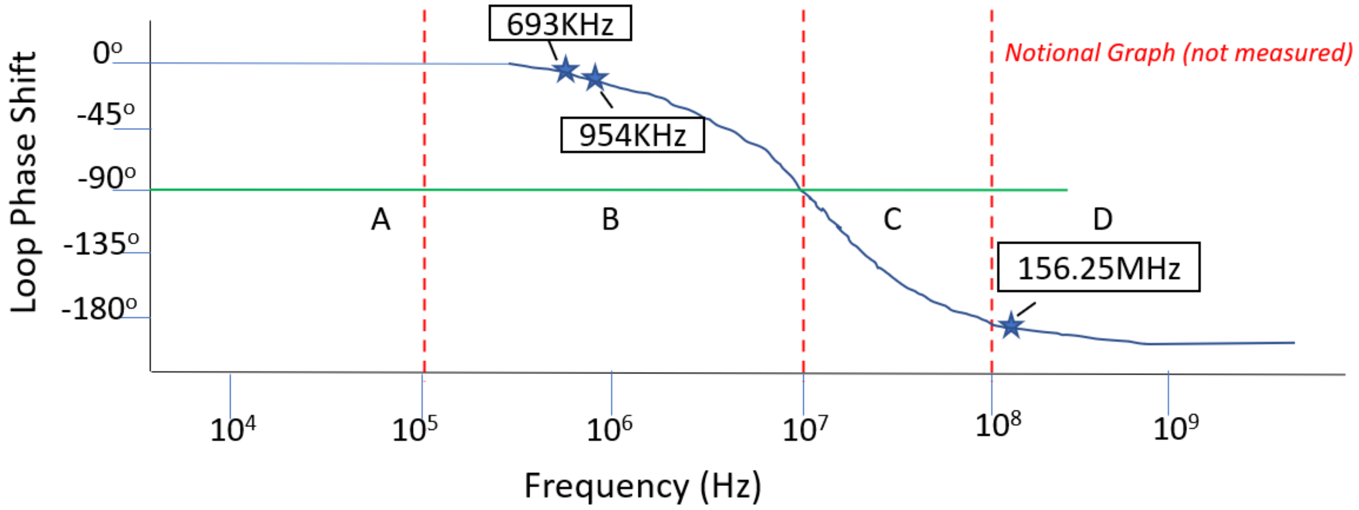


Figure 3. *CDR loop phase shift (delay) variation with frequency*

9. PJ components at 693KHz and 954KHz falling in Region B incurring different phase shifts (delay) may cause CDR loop to "hunt while attempting to track both simultaneously" leading to higher recovered clock jitter potentially leading to sampling errors (BER) [4]. Table 4 shows NIC SKU N3 transmitter jitter PJ also includes multiple components but does not trigger PSM4/Lane O/E CDR LOL when paired with same Vendor P1 module. A key difference is that the two low frequency components at 50KHz and 840KHz are separated by a wider margin. The 50KHz component incurs almost no delay allowing higher frequency component at 840KHz to be tracked robustly with minimal or no "hunting". An advanced SJTOL measurement with pairing of frequencies and their relative offset, with respect to 10MHz loop BW, as variables is planned.

10. The observation in Table 2 specific to PSM4 module from Vendor P1 where incidence and rate of O/E CDR LOL is proportional to number of active lanes is a factor that is independent of NIC SKU N1 transmitter jitter generation. The contribution from adjacent lanes is treated as uncorrelated noise with unknown distribution, intrinsic to Vendor P1 module, which is additive to the PJ stress.

11. The combined effect of CDR loop hunting and intrinsic noise (from 9,10) leads to O/E CDR LOL when interfaced to NIC SKU N1. Interestingly, PSM4 modules from Vendor P2 and P3 do not show CDR LOL with same PJ stress present possibly because of the absence of or reduced intrinsic noise component in those modules.

12. 100G PSM4 SRS SJTOL requirement specified at a single frequency 200MHz may have masked detection of O/E CDR vulnerability in the presence of SJ at other frequencies within loop bandwidth during qualification of Vendor P1 PSM4 module.

The above points combine measured data with theoretical considerations to explain observed PSM4/lane OE/E CDR LOL. Points 9-11 indicate neither NIC-SKU N1 transmitter jitter generation nor Vendor P1 PSM4 module intrinsic noise in isolation may trigger the failure.

5.0 Link recovery in the presence of PSM4/Lane O/E CDR LOL

Intel tests indicate, when module option available, disabling PSM4/Lane O/E CDR restores link with no penalty to link quality. Link quality assessment was based on NIC SKU N1 and Vendor P1 combination and FEC corrected codeword count remaining at zero monitored over a seven-day period. This option, while viable as demonstrated, is subject to sufficient validation and is recommended with Clause 108 RS FEC stipulated in IEEE Std. 802.3-2022 Clause 114 for interfacing with 25GBASE-LR PMD.

6.0 Conclusion

Failure analysis on a 25 GbE optical link involving 25GBASE-LR interface to PSM4/Lane indicates compliance to IEEE Std. 802.3-2022 Annex 109B Host Output, Clause 114 25GBASE-LR and 100G PSM4 Stressed Receiver Sensitivity (SRS) specifications is not sufficient to guarantee stable link across a wide range of vendors and components within the link. This observation offers cause for treatment as an interoperability concern. The key factors for the interoperability concern to manifest are:

1. mismatch between SRS sinusoidal jitter tolerance (SJTOL) requirements between IEEE and PSM4 standards
2. ambiguity related to SJTOL compliance with SJ at multiple frequencies simultaneously present
3. impact of PSM4 lane-lane crosstalk on O/E CDR SJTOL

Interestingly, above factors also apply to E/O function thus contrasting the role of signal conditioning - CTLE and Limiting Amplifier (LA) - may help explain SJTOL performance differences between E/O and O/E respectively, when subject to similar SJ stress.

7.0 Acknowledgement

The author thanks Josh Tsai, Robert Bentley Jr., Benjamin Cheong, and Kevin Cassidy for their contributions in various capacities to this investigation at Intel and Sam Johnson, Intel Link Applications Engineering, for technical discussions and support in facilitating external exposure through Ethernet Alliance High-Speed Networking (HSN) forum.

8.0 References

[1] IEEE Std. 802.3-2022

[2] 100G PSM4 Specification, Version 2.0, September 2014

[3] https://grouper.ieee.org/groups/802/3/bs/public/16_03/debernardinis_3bs_01a_0316.pdf

[4] Behzad Razavi, "Designing BangBang PLLs for Clock and Data Recovery in Serial Data Transmission Systems," in Phase-Locking in High-Performance Systems: From Devices to Architectures, IEEE, 2003, pp.34-45, doi: 10.1109/9780470545492.ch4.

[5] Dongwoo Hong, Chee-Kian Ong and Kwang-Ting Cheng, "BER estimation for serial links based on jitter spectrum and clock recovery characteristics," 2004 International Conference on Test, Charlotte, NC, USA, 2004, pp. 1138-1147, doi: 10.1109/TEST.2004.1387388.

[6] Alan V. Oppenheim. Alan S. Willsky, S. Hamid Nawab, “Signals & Systems 2nd ed.” Pearson Prentice Hall

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.