End-to-End System-Level Simulations with Retimers for PCIe Gen5 & CXL: A How-To Guide

Elene Chobanyan, Hewlett Packard Enterprise
Casey Morrison, Astera Labs
Pegah Alavi, Keysight Technologies

Speakers

Elene Chobanyan

Signal Integrity Engineer, Hewlett Packard Enterprise (HPE)

elene.chobanyan@hpe.com | https://www.linkedin.com/in/elene-chobanyan-38463a25/

Elene is a senior signal integrity engineer at Hybrid IT Compute Solutions at Hewlett Packard Enterprise (HPE). She is HPE's primary voting member at PCI-SIG Electrical workgroup, a representative at IPC D24D standard and an active contributor at Gen-Z PHY workgroup. Elene received the B.S. degree in physics and the M.S. degree in electrical and electronics engineering from Tbilisi State University, Tbilisi, Georgia, and PhD. Degree in electrical and computer engineering from Colorado State University, USA. Her current focus areas are high speed server interconnect architecture and design, PCB material characterization techniques, non-volatile and volatile memory subsystems.

Casey Morrison

Head of Product and Applications Engineering, Astera Labs, Inc.

casey.morrison@asteralabs.com | www.linkedin.com/in/casey-morrison | @CaseyTMorrison

Casey is the head of Product and Applications Engineering at Astera Labs and is responsible for defining, validating, and helping customers design-in Astera Labs' semiconductor products and plug-and-play systems. With 12+ years of experience in high-speed interfaces for data center and wired/wireless communications systems, he has a passion for creating chips and systems which help to enable state-of-the-art compute and networking topologies.

Pegah Alavi

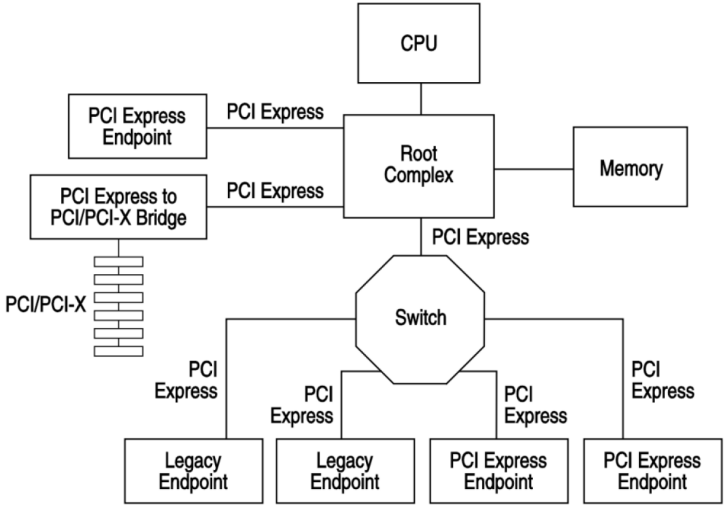
Senior Applications Engineer, Keysight Technologies

pegah_alavi@keysight.com

Pegah is a Senior Applications Engineer at Keysight Technologies, where she focuses on Signal Integrity and High Speed Digital Systems and Applications. Prior to joining Keysight Technologies, Pegah worked on system level modeling of analog and mixed signal circuits in order to best predict the overall systems performance and accurately represent each component.

PCI Express Gen5 / CXL Overview

- PCI-Express Gen5:
  - Base Specification revision 5.0 ratified in 2019
  - Maximum speed: 32 Gbps / Lane / direction; NRZ signaling
  - Single- or multi-lane Links to scale aggregate bandwidth: x1, x2, x4, x8, x16
- Compute Express Link (CXL):
  - Base Specification revision 1.1 ratified in 2019
  - Electrical specifications derived from PCIe 5.0
  - Enables high-speed, efficient interconnect CPU, workload accelerators, and memory
  - Maximum speed: 32 Gbps / Lane / direction; NRZ signaling
- Applications:
  - Servers: CPU-to-network, CPU-to-storage, CPU-to-Accelerator
  - Client compute: CPU-to-peripheral (e.g. GPU)

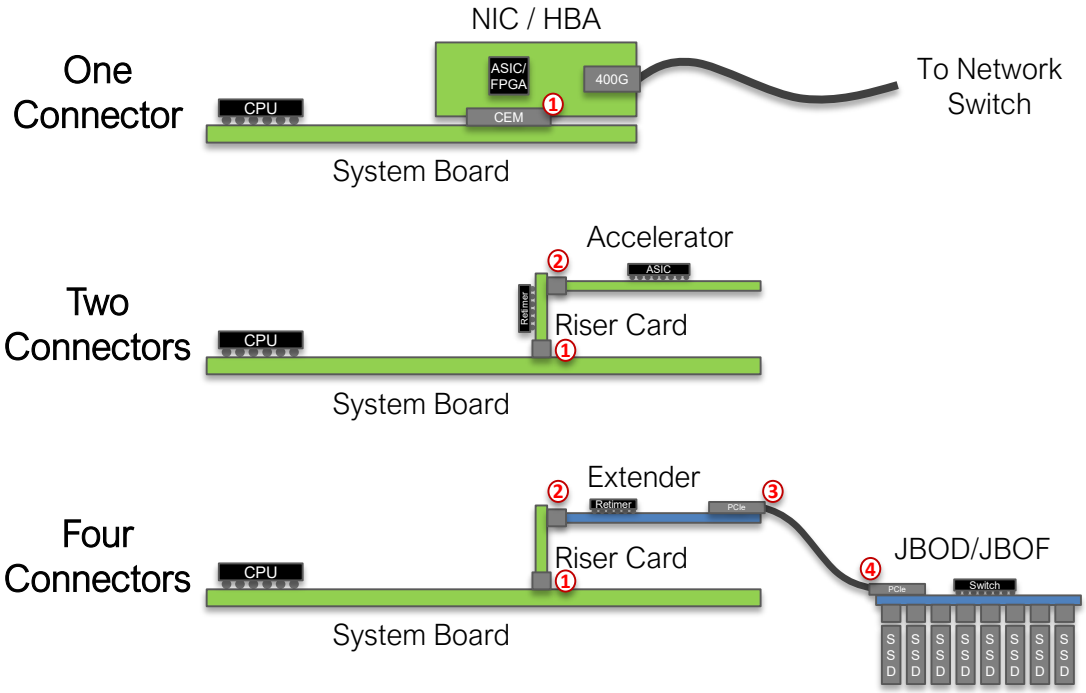


Example PCIe Topology (From Base Specification)

- **Problem Statement:** How can system designers evaluate the tradeoffs between enhanced PCB material and Retimers; and what is the most effective placement of active components in a system to maximize performance while minimizing cost?

PCIe Gen5/CXL Topologies

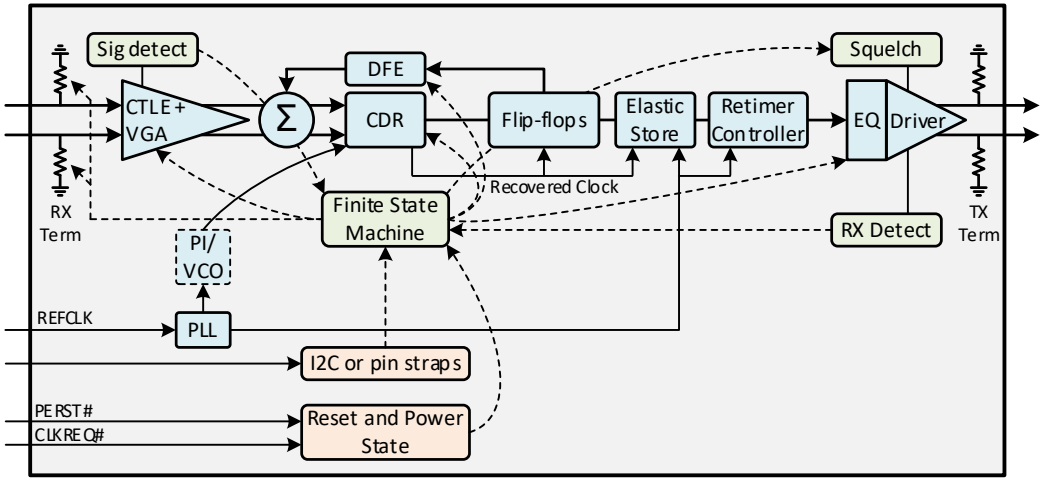
- System designers anticipate various topologies ranging from **one to four connectors**.
- Channel loss budget:
  - PCIe Gen5 & CXL: 36 dB @ 16 GHz, end-to-end
- Temperature/humidity affects can result in **±10% variation in insertion loss for high-end PCB materials, ±25% variation for main-stream material**
- Some topologies exceed the loss budget when accounting for such variations and for **safety margin**
- **Reach extension** via advanced PCB material or a Retimer will be required



What is a Retimer?

- Physical-layer-protocol-aware, software-transparent extension device that forms two separate electrical Link Segments:
  - Root Complex (RC) to Retimer
  - Retimer to Endpoint (EP)
- Mixed-signal analog/digital device—fully recovers data, extracts clock, and retransmits clean data
- Covered in PCIe 5.0 specification (Section 4.3)
- Complies with all PCIe electrical specifications
- Performs Receiver detection and Lane-to-Lane deskew
- Executes Link equalization Phases 2 & 3

Retimer Block Diagram



Approach to System-Level Analysis and Simulation

1. Determine if a Retimer is required	• Use the statistical eye analysis simulator (SeaSim) tool to determine if the topology meets the PCIe channel requirement • Consider the performance degradations associated with **temperature, humidity, and manufacturing variations**. • Consider reserving some **safety margin** for your system
2. Define a simulation space	• Identify **worst-case conditions** (e.g. temperature, humidity, and impedance) • What are the minimum set of parameters (e.g. **transmitter Presets**) which must be varied to adequately analyze system performance and margin?
3. Define the evaluation criteria	• Determine **minimum eye height/width** which is considered a passing result • How many transmitter settings must yield a passing result to have high confidence that there is adequate margin?
4. Execute and analyze results	• Use **IBIS-AMI models** and time domain simulations to analyze each case in the context of the pre-defined evaluation criteria

How to Determine if a Retimer is Required

There are generally two ways to approach this:

Channel Loss Budget Analysis

- Compare the end-to-end channel insertion loss, including RC and EP package losses, against the PCIe channel budget.
- If the topology's channel loss exceeds the PCIe informative specification, then a Retimer is likely required.

Total channel budget	Root package	Non-root package	CEM connector	Add-in Card (AIC)	System Budget[1]
36 dB	9.0 dB	4.0 dB	1.5 dB	9.5 dB	17.5 dB

SeaSim Analysis

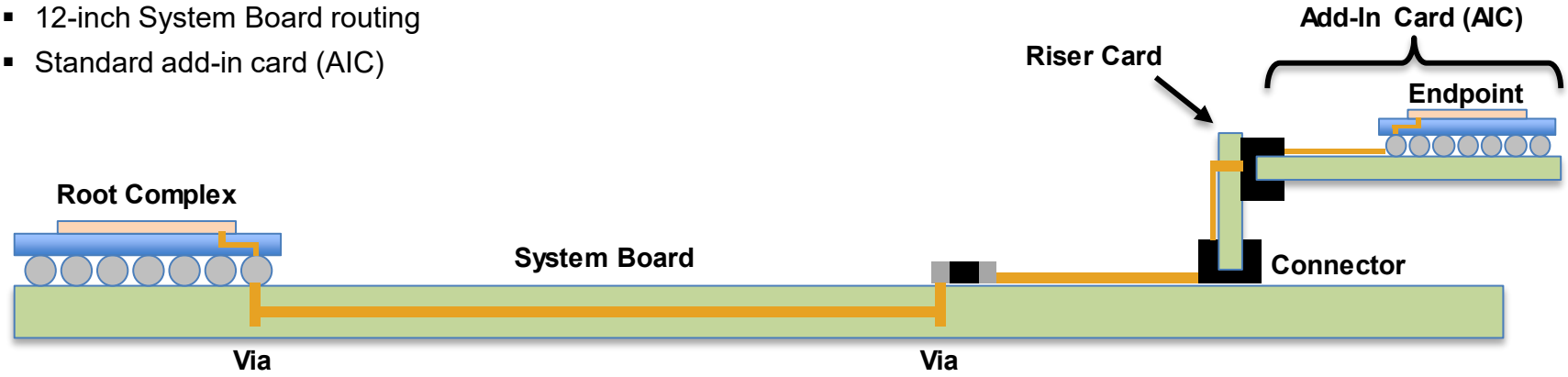
- Simulate channel s-parameter in the Statistical Eye Analysis Simulator (SeaSim) tool to determine if post-equalized eye height (EH) and eye width (EW) meet the minimum eye opening requirements: ≥15 mV EH and ≥0.3 UI EW at Bit Error Ratio (BER) ≤ 10^{-12}.
- If eye opening does not meet reference receiver's requirements, then a Retimer is likely required.
- This methodology is more accurate and preferred to the pure loss budget analysis, as it takes into account other channel characteristics, such as reflections and crosstalk.

[1]System budget includes the baseboard, riser card, the baseboard-to-riser-card, and PCIe card electromechanical (CEM) form factor connectors.

Topology Considered for This Analysis

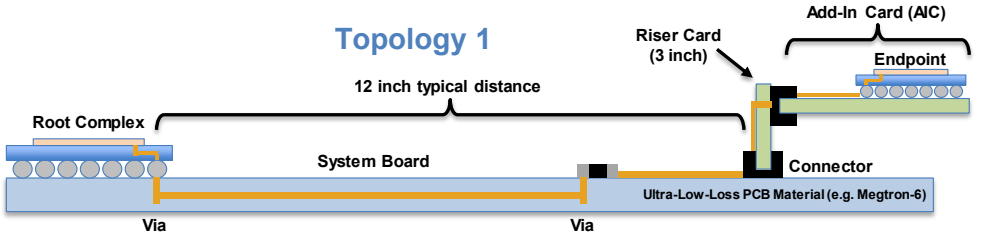
In this analysis, we examine variations of the following topology common in server, storage, and accelerator systems:

- Two connectors
- Three vias
- Riser Card
- 12-inch System Board routing
- Standard add-in card (AIC)

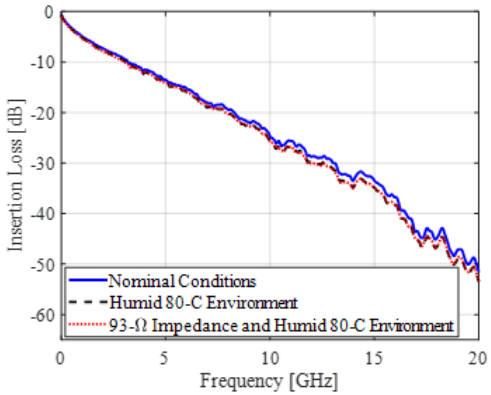


Topology 1: Ultra-Low-Loss PCB, No Retimer

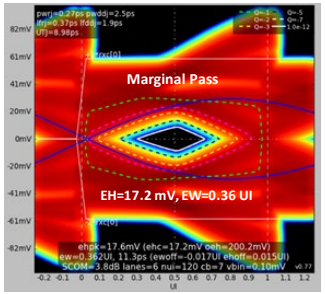
- Using a material like Megtron-6 reduces variations due to temperature/humidity to ~10%
  - ~1 dB/inch at 16 GHz, nominal temp. & humidity
  - ~1.11 dB/inch at 16 GHz, 80C, 75% humidity
  - PCB impedance variations cause further degradation
- Total channel loss **exceeds 36 dB** in each case
- SeaSim analysis shows a marginal pass for nominal conditions, marginal fails for extreme conditions



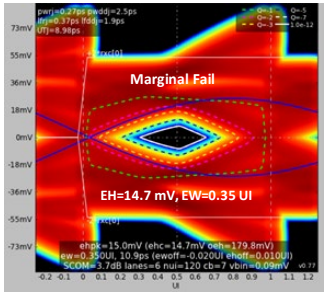
Topology 1 Insertion Loss



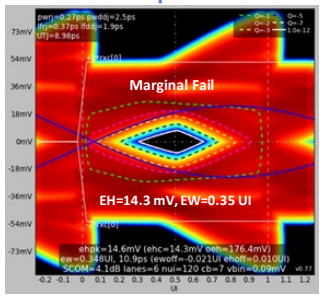
SeaSim Result: Topology 1, Nominal Conditions



SeaSim Result: Topology 1, Humid 80-C Environment



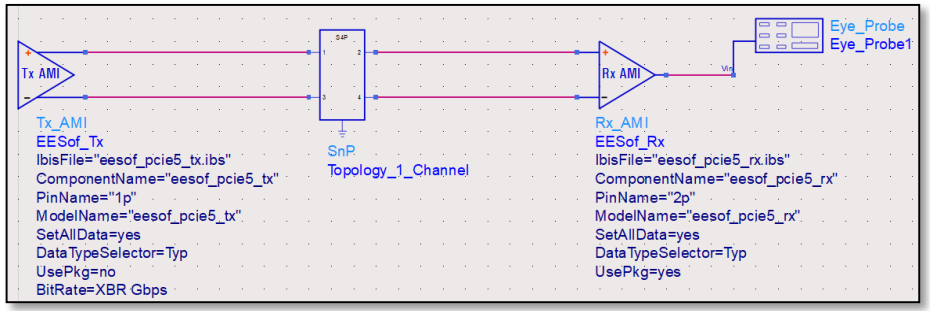
SeaSim Result: Topology 1, Humid 80-C Environment and 93-Ω Impedance



Topology 1: Ultra-Low-Loss PCB, No Retimer

- To further assess Topology 1's viability, an IBIS-AMI simulation is run with **generic TX** and **generic RX** models in Keysight ADS software
- Ten TX Preset settings are simulated, to gauge margin
- **No TX Preset setting yielded passing results** for both eye height and eye width compared to reference receiver minimum eye opening requirements:
  - ≥15 mV EH and ≥0.3 UI EW

Keysight ADS Testbench for Topology 1



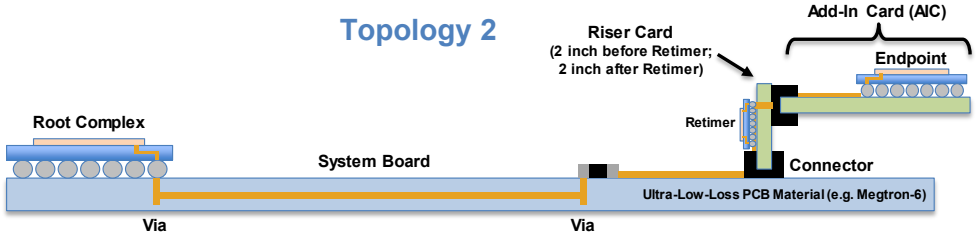
IBIS-AMI Simulation Results for Topology 1

Tx Preset	Case 1.a: Nominal Temperature, Nominal Humidity, and 85-Ω Impedance		Case 1.b: 80C Temperature, High Humidity, and 85-Ω Impedance		Case 1.c: 80C Temperature, High Humidity, and 93-Ω Impedance	
	EH	EW	EH	EW	EH	EW
0	0 mV	0.23 UI	0 mV	0.23 UI	0 mV	0.22 UI
1	3 mV	0.21 UI	2 mV	0.21 UI	2 mV	0.00 UI
2	3 mV	0.24 UI	3 mV	0.22 UI	2 mV	0.00 UI
3	2 mV	0.19 UI	1 mV	0.17 UI	1 mV	0.00 UI
4	2 mV	0.00 UI	2 mV	0.00 UI	2 mV	0.00 UI
5	7 mV	0.28 UI	6 mV	0.26 UI	3 mV	0.21 UI
6	7 mV	0.29 UI	6 mV	0.29 UI	4 mV	0.21 UI
7	8 mV	0.37 UI	7 mV	0.34 UI	3 mV	0.15 UI
8	8 mV	0.36 UI	6 mV	0.36 UI	5 mV	0.21 UI
9	7 mV	0.30 UI	6 mV	0.29 UI	2 mV	0.22 UI

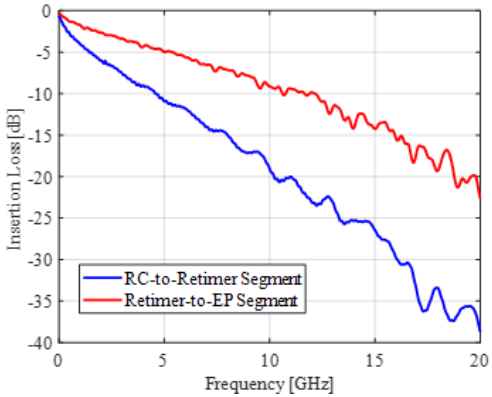
Topology 2: Ultra-Low-Loss PCB, Retimer on Riser Card

- Given Topology 1 fails both evaluation criteria, Topology 2 is proposed which includes a **Retimer on the Riser Card**
- This creates two independent Link segments which are afforded the full channel budget
- Including worst-case temperature, humidity, and impedance, **both Link segments pass both criteria**

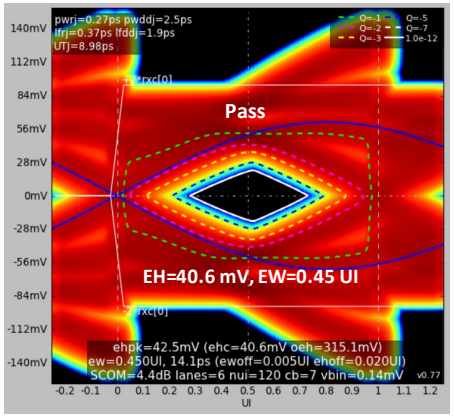
Topology 2



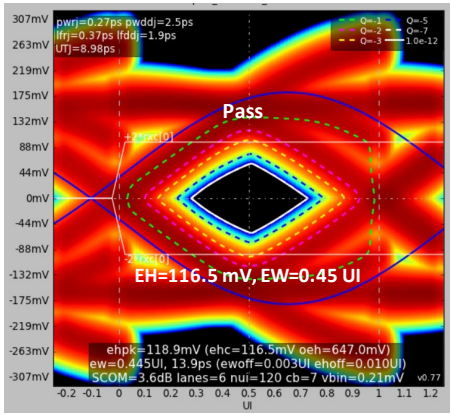
Topology 2 Insertion Loss



SeaSim Result: Topology 2, RC-to-Retimer Segment



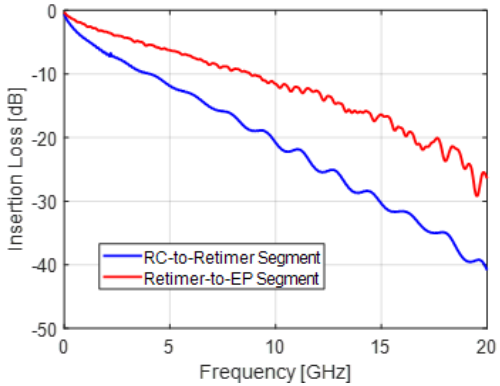
SeaSim Result: Topology 2, Retimer-to-EP Segment



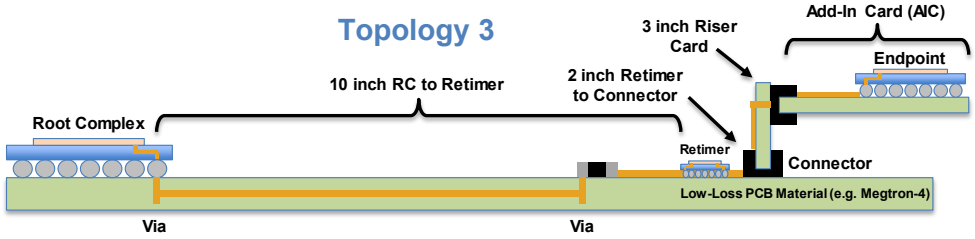
Topology 3: Low-Cost PCB, Retimer on System Board

- Given Topology 2 has ample margin, **a lower-cost system can be constructed with mainstream PCB material**
- Retimer placement is moved to System Board to further improve margin on RC-to-Retimer Link segment
- Including worst-case temperature, humidity, and impedance, **both Link segments pass both criteria**

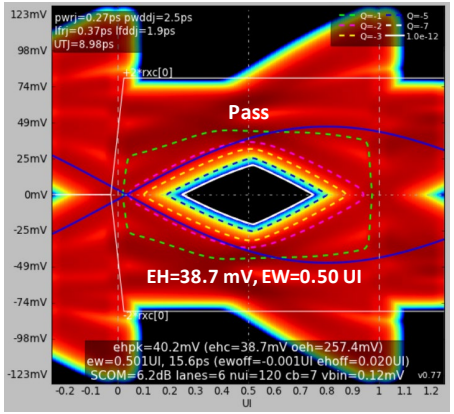
Topology 2 Insertion Loss



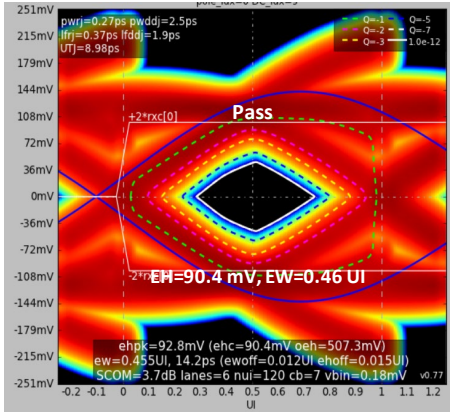
Topology 3



SeaSim Result: Topology 3, RC-to-Retimer Segment



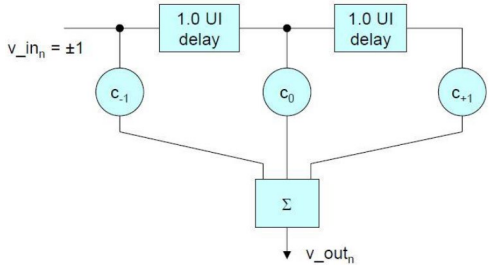
SeaSim Result: Topology 3, Retimer-to-EP Segment



Define a Simulation Space – Independent Variable

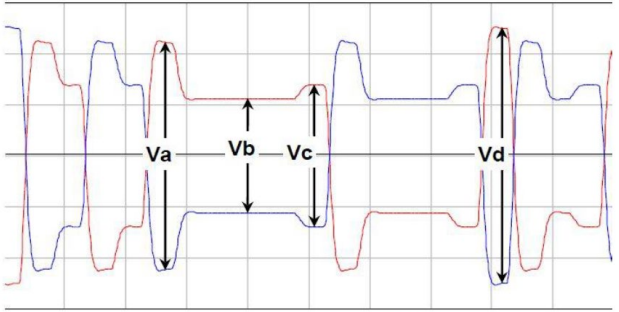
- The main independent variable in PCIe Link simulations is **Transmitter Preset**—pre-defined combinations of pre-shoot and de-emphasis
- Ten such Presets are defined in PCIe

Preset #	Pre-shoot (dB)	De-emphasis (dB)	c_{-1}	c_{+1}	Va/Vd	Vb/Vd	Vc/Vd
P4	0.0	0.0	0.000	0.000	1.000	1.000	1.000
P1	0.0	-3.5 ± 1	0.000	-0.167	1.000	0.668	0.668
P0	0.0	-6.0 ± 1.5	0.000	-0.250	1.000	0.500	0.500
P9	3.5 ± 1	0.0	-0.166	0.000	0.668	0.668	1.000
P8	3.5 ± 1	-3.5 ± 1	-0.125	-0.125	0.750	0.500	0.750
P7	3.5 ± 1	-6.0 ± 1.5	-0.100	-0.200	0.800	0.400	0.600
P5	1.9 ± 1	0.0	-0.100	0.000	0.800	0.800	1.000
P6	2.5 ± 1	0.0	-0.125	0.000	0.750	0.750	1.000
P3	0.0	-2.5 ± 1	0.000	-0.125	1.000	0.750	0.750
P2	0.0	-4.4 ± 1.5	0.000	-0.200	1.000	0.600	0.600



$$v\_out_n = v\_in_{n-1}c_{n-1} + v\_in_n c_n + v\_in_{n+1}c_{n+1}$$

$$|c_{-1}| + |c_0| + |c_{+1}| = 1 \quad c_{+1} \le 0 \quad c_{-1} \le 0$$

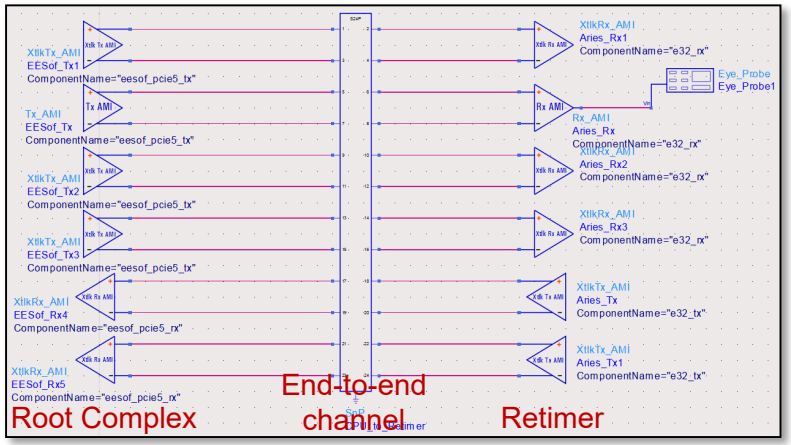


De-emphasis = 20 $\log_{10}$Vb/Va
Preshoot = 20$\log_{10}$Vc/Vb
Boost = 20$\log_{10}$ Vd/Vb

Define a Simulation Space – Testbench Configuration

- In this case, the Root Complex package and Retimer package s-parameters are incorporated into an end-to-end 24-port channel model
  - 3 far-end crosstalk (FEXT) terms
  - 2 near-end crosstalk (NEXT) terms
- The Receiver model used in this analysis (Astera Labs Retimer) has **fully-adaptive equalization** parameters

IBIS-AMI Simulation Testbench: CPU-to-Retimer*



*Note: Similar topology used for Retimer-to-Endpoint

IBIS-AMI Simulation Parameters

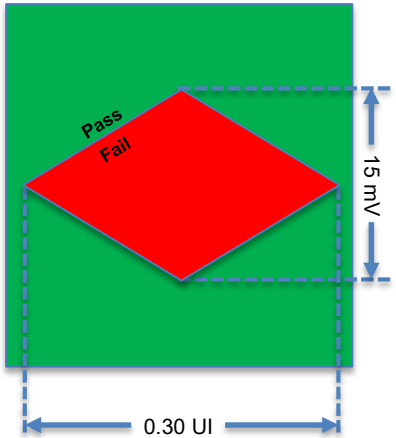
Parameter	Value
Data Rate	32 GT/s
Data Pattern	PRBS23
Total number of bits	500,000 (not including Ignore_Bits)
Crosstalk	Three FEXT aggressors + Two NEXT aggressors
Ignore_Bits	~2.5M *(Note: This is set by the receiver model)*
Simulation type	Time domain (a.k.a. bit-by-bit) *Note: Simulations may be faster running in Statistical Mode, however non-linear behavior may not be adequately represented.*
Bit-by-bit extrapolation	Enabled *Note: Simulations will be faster without this mode enabled, however random jitter (RJ) will not be accounted for as accurately.*

RC Tx Parameters	Retimer Parameters	EP Rx Parameters	Channel Parameters
Model: Keysight PCIe Reference Transmitter *Presets*: 0, 1, ..., 9 *VOD*: 800 mVppd	*Model:* Astera Labs Aries Retimer IBIS-AMI *Receiver*: Automatically-adapted *Transmitter*: Presets 0, 1, …, 9	*Model:* Keysight PCIe Reference Receiver *Receiver*: Automatically-adapted	*Temperature:* 80C *Humidity*: High *Impedance*: 93 Ω

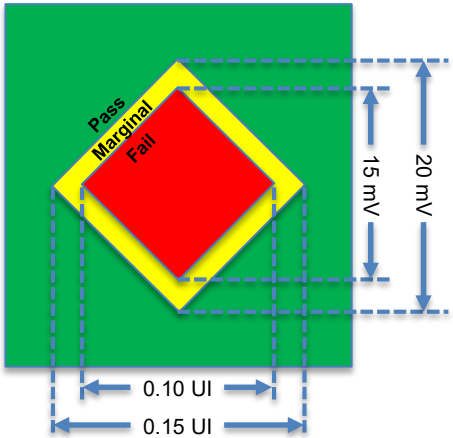
Define Evaluation Criteria

- Bit error rate (BER) is the ultimate gauge of link performance, but an accurate measure of BER is not possible in relatively short, multi-million-bit simulations.
- Instead, this analysis suggests the following pass/fail criteria which consist of two rules:
  1. A link must **meet the receiver's eye height (EH) and eye width (EW)** requirements
  2. A link must meet criterion 1 for **at least half** of Tx Preset settings (≥5 out of 10)
- Criterion 1 establishes that the there is a viable set of settings which results in the desired BER.
- The specific EH and EW required by the receiver is implementation-dependent.
- Criterion 2 ensures that the link has **adequate margin** and is not overly-sensitive to the Tx Preset setting.

Minimum Eye Opening Criteria: Reference Receiver IBIS-AMI Model



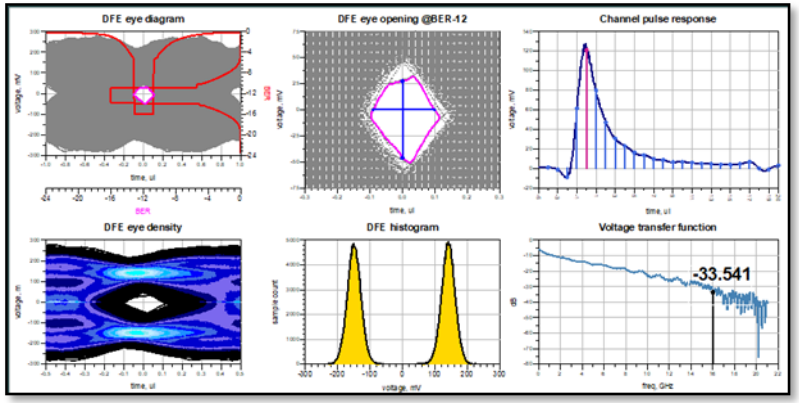
Minimum Eye Opening Criteria: Retimer IBIS-AMI Model



Execute and Analyze the Results – Topology 2

- IBIS-AMI simulations obtained via ADS with reference RC/EP and Astera Labs Retimer models support previous conclusions derived from statistical SeaSim approximation
- Previously marginally passing or failing Topology 1 exhibits high operation margin with a Retimer in the path (Topology 2)

Detailed Data Display for Topology 2, Tx Preset 9

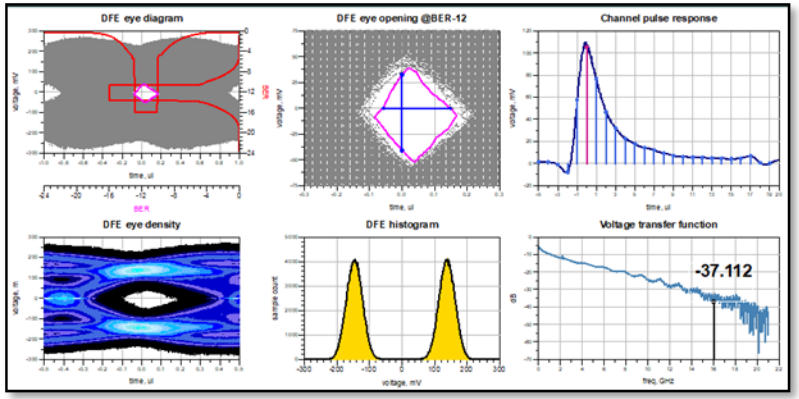


Tx Preset	Topology 2 RC Transmitter to Retimer Receiver Insertion Loss: ~28 dB at 16 GHz *Requirement: EH ≥ 15 mV, EW ≥ 0.15 UI*		Topology 2 Retimer Transmitter to EP Receiver Insertion Loss: ~15 dB at 16 GHz *Requirement: EH ≥ 15 mV, EW ≥ 0.30 UI*	
	EH	EW	EH	EW
0	30 mV	0.10 UI	53 mV	0.38 UI
1	37 mV	0.10 UI	56 mV	0.43 UI
2	19 mV	0.05 UI	55 mV	0.43 UI
3	41 mV	0.10 UI	57 mV	0.45 UI
4	51 mV	0.15 UI	58 mV	0.42 UI
5	77 mV	0.25 UI	70 mV	0.53 UI
6	89 mV	0.26 UI	70 mV	0.52 UI
7	49 mV	0.14 UI	55 mV	0.45 UI
8	68 mV	0.18 UI	61 mV	0.50 UI
9	94 mV	0.25 UI	69 mV	0.50 UI

Execute and Analyze the Results – Topology 3

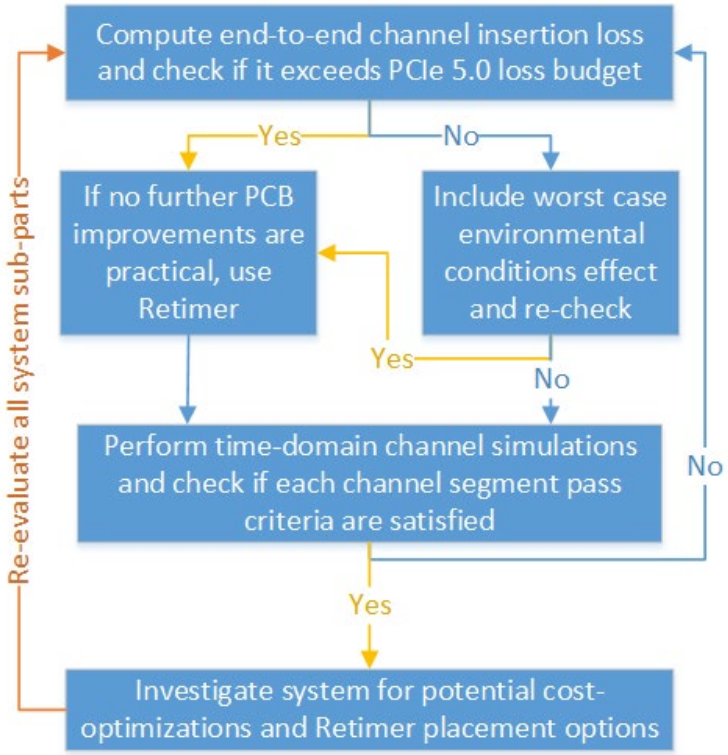
- Topology 2 exhibits high operation margin and allows for further cost-optimization with alternative Retimer placement, by choosing lower performant PCB materials (Topology 3)
- Based on Criterion 2, Topology 3 does not degrade overall system performance and allows for cheaper design

Detailed Data Display for Topology 3, Tx Preset 9



Tx Preset	Topology 3 RC Transmitter to Retimer Receiver Insertion Loss: ~32 dB at 16 GHz *Requirement: EH ≥ 15 mV, EW ≥ 0.15 UI*		Topology 3 Retimer Transmitter to EP Receiver Insertion Loss: ~19 dB at 16 GHz *Requirement: EH ≥ 15 mV, EW ≥ 0.30 UI*	
	EH	EW	EH	EW
0	32 mV	0.08 UI	38 mV	0.42 UI
1	41 mV	0.11 UI	40 mV	0.47 UI
2	37 mV	0.12 UI	40 mV	0.46 UI
3	43 mV	0.12 UI	39 mV	0.46 UI
4	22 mV	0.16 UI	38 mV	0.42 UI
5	70 mV	0.27 UI	54 mV	0.54 UI
6	75 mV	0.28 UI	55 mV	0.56 UI
7	48 mV	0.13 UI	44 mV	0.53 UI
8	61 mV	0.18 UI	49 mV	0.56 UI
9	83 mV	0.29 UI	54 mV	0.57 UI

Conclusions



- PCI Express Retimer and its characteristics/applications have been discussed
- A widely used challenging PCI Express system interconnect architecture topology and its design challenges at 32Gbps speed (PCIe 5.0 or CXL 1.1) has been introduced
- A marginal system performance and failure under stressed environmental conditions has been studied and solved by dividing it into sub-parts via Retimer
- Furthermore, alternative Retimer placement option and system cost-optimization has been studied
- A general practical PCIe 5.0 and CXL system design flow has been established

Thank you!

QUESTIONS?