Seamless Transition to PCIe® 5.0 Technology in System Implementations

PCI-SIG® Educational Webinar Series

English Version: 8am December 9th, 2020 (PST)

Mandarin Version: 9am December 10th, 2020 (CST)

Speakers: English Session

Casey Morrison

VP of Products, Astera Labs, Inc.

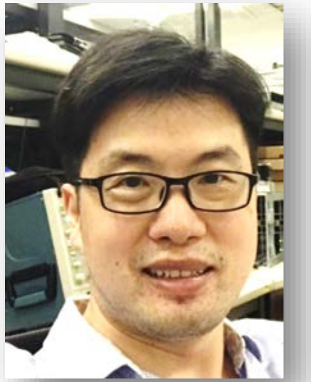
Casey is the VP of Products at Astera Labs and is responsible for defining, validating, and helping customers design-in Astera Labs' semiconductor products and plug-and-play systems (AsteraLabs.com/Products). With 12+ years of experience in high-speed interfaces for data center and wired/wireless communications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

Jonathan Bender

Head of Product Applications, Astera Labs, Inc.

Jonathan is a Member of Technical Staff at Astera Labs and is responsible for development of product applications collateral, board level designs using Astera Labs semiconductor products, supporting key customers, and interop testing Astera Labs' retimers with multiple RC/endpoints in the Cloud-Scale Interop Lab (AsteraLabs.com/InteropLab). His background as a Field Applications Engineer for 6+ years supporting high-speed, clocking, and power ICs provides a system-level insight that helps customers fulfill design goals and constraints utilizing Astera's cutting-edge products.

Speaker: Mandarin Session



Liang Liu

Head of Field Applications Engineering for Asia, Astera Labs, Inc.

Liang is a Member of Technical Staff at Astera Labs and is responsible for supporting customers in China, Taiwan, etc. He has 13+ years of high-speed interface product experiences including system design, application support and product definition. He enjoys addressing the signal integrity challenges on various high-speed interfaces using Astera Labs' retimers (AsteraLabs.com/Products).

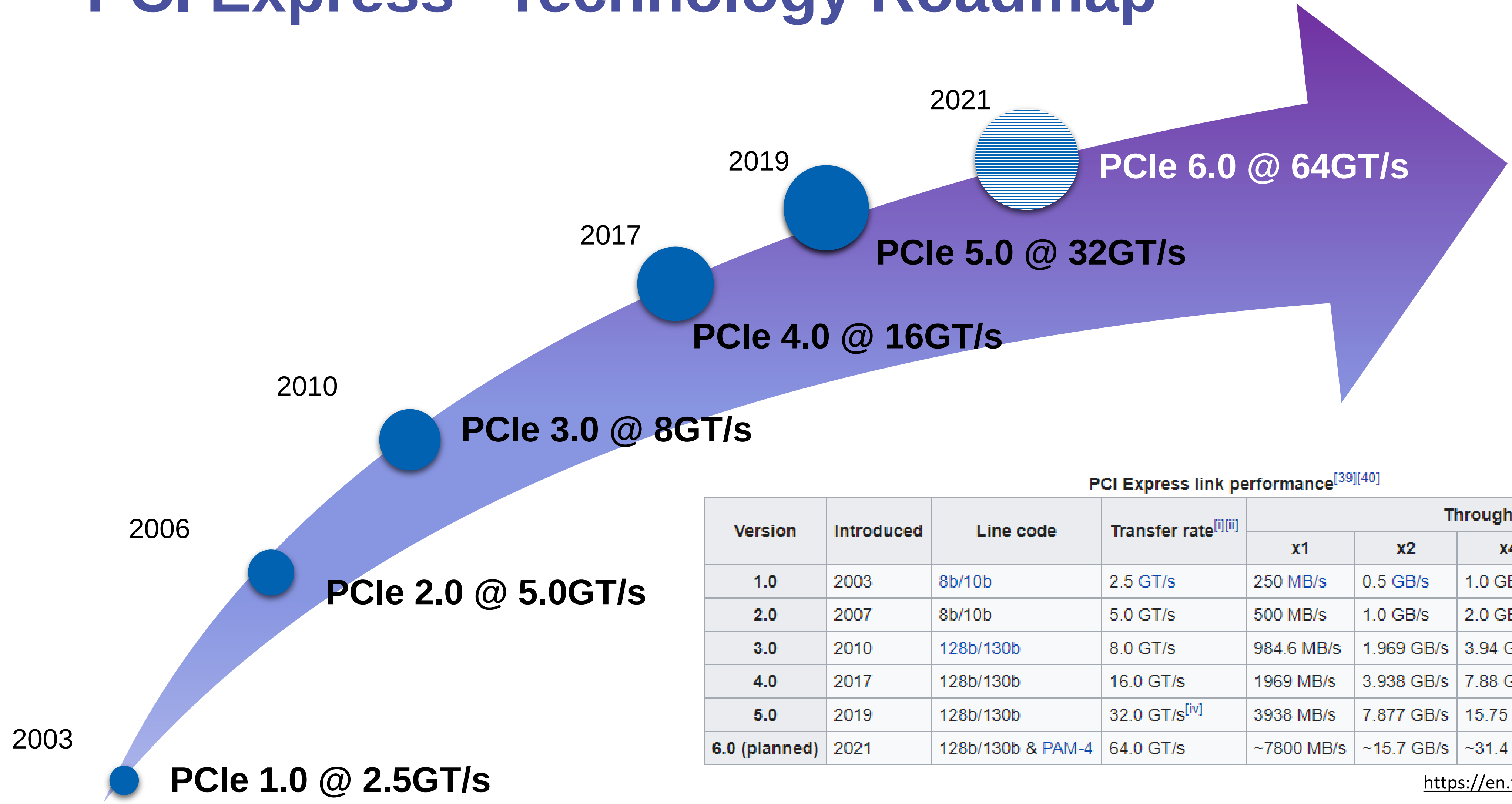
Agenda

Topic	Time
PCI Express® Specifications Overview - Transition from PCIe® 3.0 to PCIe® 4.0 to PCIe® 5.0 architecture	5 min
PCIe® 5.0 Signal and Link Integrity Challenges - Popular system design topologies - Channel loss budget - Reach extension options	20 min
Clocking Considerations and Topologies - PCIe 5.0 technology reference clock considerations - Clocking modes: Common clock, SRNS, SRIS	5 min
Compliance and Interop Testing - Electrical compliance - Interop testing and system robustness	10 min
Q&A	10 min

PCI Express® Specifications Overview

- Transition from PCIe® 3.0 to PCIe® 4.0 to PCIe® 5.0 Architecture
- Transmit and Receive Equalization

PCI Express® Technology Roadmap



PCI Express link performance[39][40]

Version	Introduced	Line code	Transfer rate[i][ii]	Throughput[i][iii]				
				x1	x2	x4	x8	x16
1.0	2003	8b/10b	2.5 GT/s	250 MB/s	0.5 GB/s	1.0 GB/s	2.0 GB/s	4.0 GB/s
2.0	2007	8b/10b	5.0 GT/s	500 MB/s	1.0 GB/s	2.0 GB/s	4.0 GB/s	8.0 GB/s
3.0	2010	128b/130b	8.0 GT/s	984.6 MB/s	1.969 GB/s	3.94 GB/s	7.88 GB/s	15.75 GB/s
4.0	2017	128b/130b	16.0 GT/s	1969 MB/s	3.938 GB/s	7.88 GB/s	15.75 GB/s	31.51 GB/s
5.0	2019	128b/130b	32.0 GT/s[iv]	3938 MB/s	7.877 GB/s	15.75 GB/s	31.51 GB/s	63.02 GB/s
6.0 (planned)	2021	128b/130b & PAM-4	64.0 GT/s	~7800 MB/s	~15.7 GB/s	~31.4 GB/s	~62 GB/s	~124 GB/s

https://en.wikipedia.org/wiki/PCI_Express

PCIe® 5.0 Specification – TX & RX EQ

Equalization Parameter	PCIe 3.0 (8 GT/s)	PCIe 4.0 (16 GT/s)	PCIe 5.0 (32 GT/s)
Modulation	NRZ	NRZ	NRZ
Maximum Bit Error Rate (BER)	1E-12	1E-12	1E-12
Maximum End-to-End Channel Loss	22 dB	28 dB	36 dB
Transmitter			
Differential Output Voltage	800 – 1300 mV	800 – 1300 mV	800 – 1300 mV
Maximum Boost Ratio	8.0 dB	8.0 dB	8.0 dB
Receiver (Behavioral Model)			
CTLE	1st-order Peak frequency: ~4 GHz Max AC boost: ~10 dB	1st-order: Peak frequency: ~6 GHz Max AC boost: ~12 dB	2nd-order Peak frequency: ~14 GHz Max AC boost: ~15 dB
DFE Taps	1	2	3

PCIe® 5.0 Specification – Precoding

- Due to the significant role the **Decision Feedback Equalizer (DFE)** plays in Receiver equalization, **burst errors** are more likely to occur at 32 GT/s compared to 16 GT/s.
- To counteract this risk, PCIe 5.0 introduces **Precoding**. By enabling precoding in the Transmitter and Receiver, the chance of burst errors (and uncorrectable errors) is greatly reduced.
- **Precoding does not impact signal quality** or signal integrity; **it need not be simulated** in IBIS-AMI.
- Any receiver may request Precoding during Link training. With Precoding, the **BER increases by 2x.**

Precoding Circuit Eliminates Burst Errors

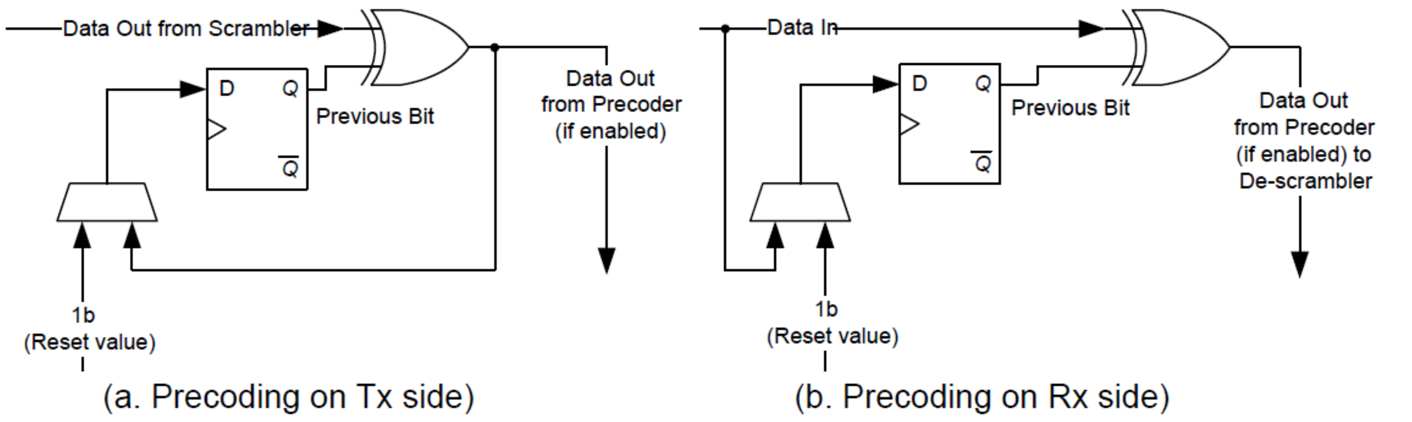


Figure 4-20 Precoding working the scrambler/ de-scrambler

Error Propagation and Burst Errors

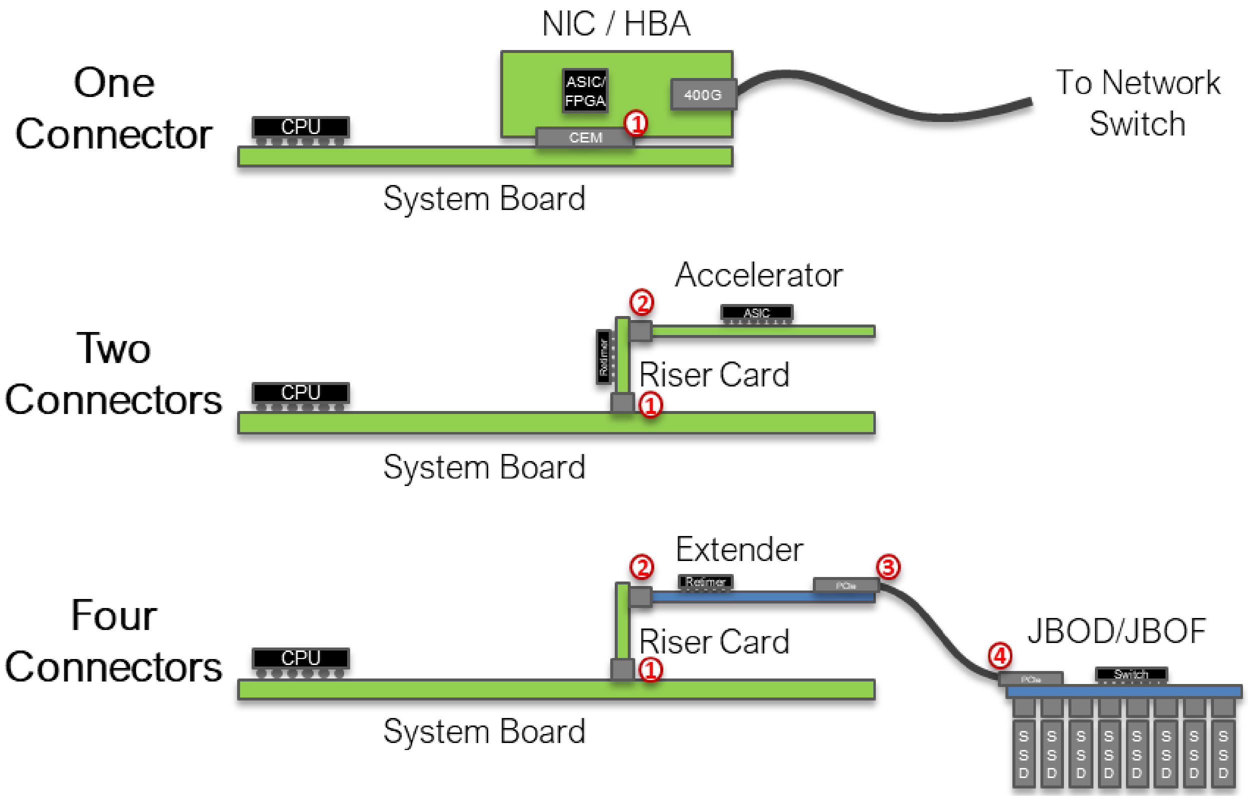
- A DFE uses **past decisions** to help make current decisions about whether the signal is a one or zero
- If a wrong decision is made (i.e. bit error), this can lead to **further wrong decisions** if the weight of the feedback term is strong
- This is **error propagation** which leads to a series (burst) of errors

PCIe® 5.0 Signal and Link Integrity Challenges

- Popular system design topologies
- Channel loss budget
- Reach extension options

PCIe® 5.0 Architecture Topologies

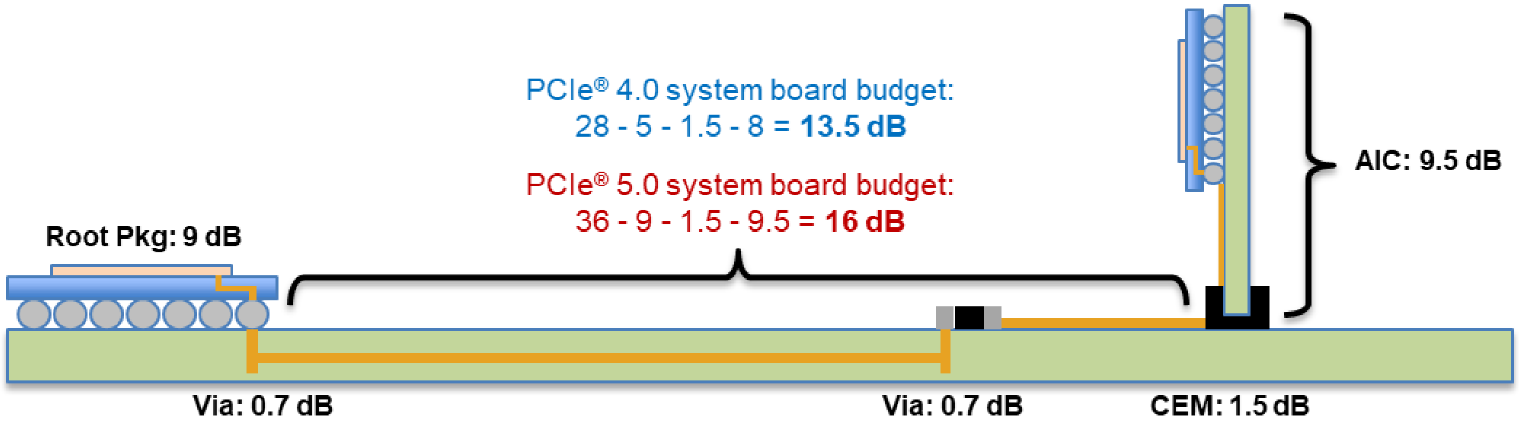
- System designers anticipate various topologies ranging from **one to four connectors**.
- Channel loss budget: 36 dB @ 16 GHz, end-to-end
- Temperature/humidity affects can result in **±10% variation in insertion loss for high-end PCB materials, ±25% variation for main-stream material**
- Some topologies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



PCIe® 5.0 Channel Insertion Loss Budget

PCIe Rev	Nyquist Frequency	Total channel loss budget	Root Package	Non-root Package	CEM connector	Add-in Card (AIC)	Budget for system board
3.0 (8 GT/s)	4 GHz	22 dB	3.5 dB	2.0 dB	1.7 dB	6.5 dB	10.3 dB
4.0 (16 GT/s)	8 GHz	28 dB	5.0 dB	3.0 dB	1.5 dB	8.0 dB	13.5 dB
5.0 (32 GT/s)	16 GHz	36 dB	9.0 dB*	4.0 dB*	1.5 dB†	9.5 dB†	16.0 dB

*$\text{ILfit}_{\text{TX-ROOT-DEVICE}}$ and $\text{ILfit}_{\text{TX-NON-ROOT-DEVICE}}$ parameters in the base specification.
†Based on CEM 5.0 version 0.5.

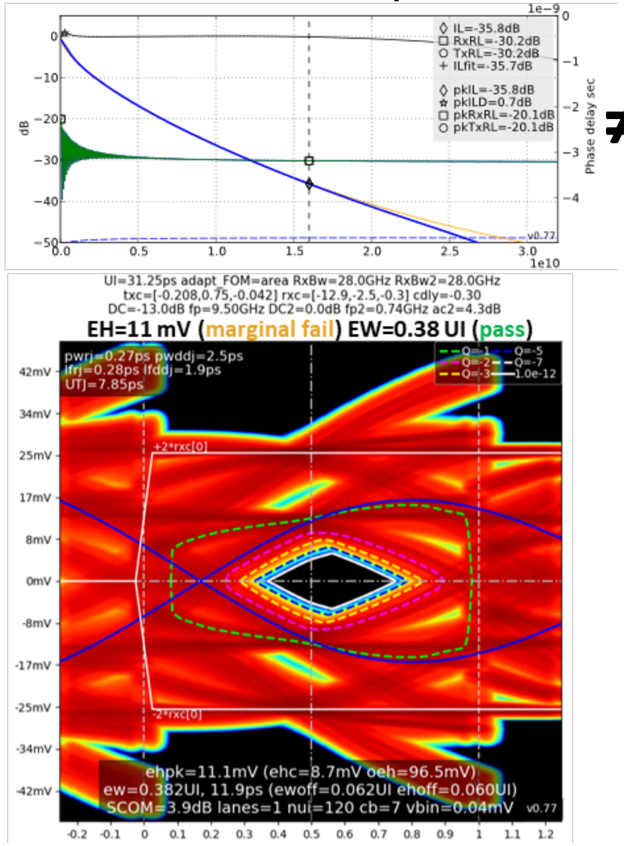


System board budget includes:
- Vias
- Stubs
- AC coupling capacitor
- Microstrip/Stripline trace

Importance of Channel Compliance

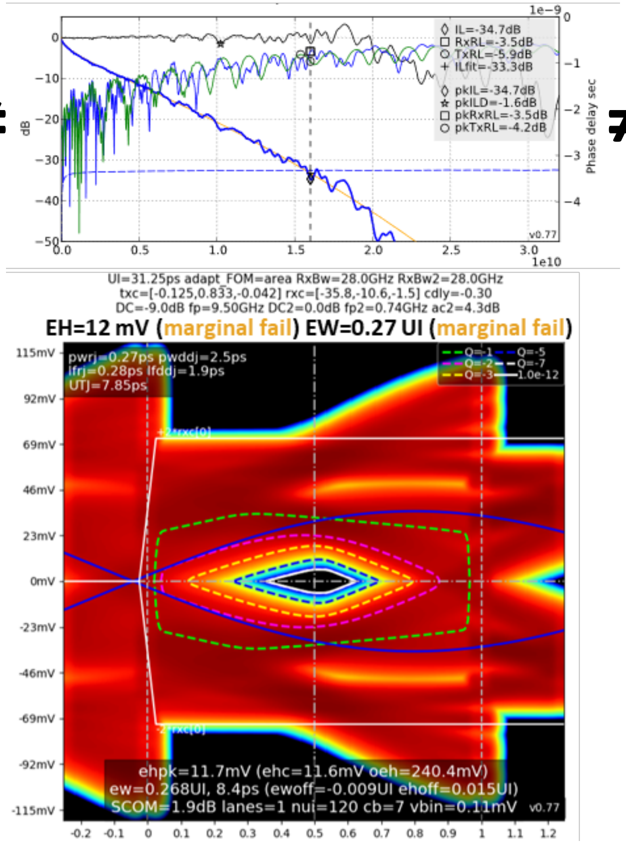
Use a **statistical simulator (e.g. SeaSim)** implementing a **reference transmitter** and **reference receiver** to gauge whether your channel (s-parameter) complies with the PCIe Base Specification.

Example 1: 36 dB channel with smooth loss profile



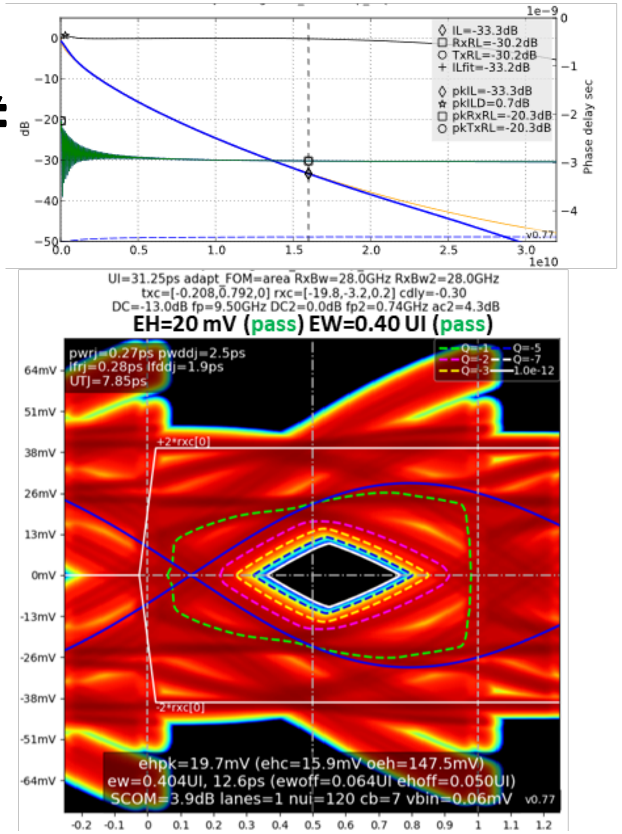
≠

Example 2: 33 dB channel with some reflections



≠

Example 3: 33 dB channel with smooth loss profile



Key Points

- Simple loss budget analysis ≠ channel compliance
- Channels are deemed compliant if EH≥15mV, EW≥0.3UI under worst-case conditions

PCB Material Tradeoffs

- There is no industry standard definition of **Mid-loss**, **Low-loss**, and **Ultra-low-loss**.
- Actual insertion loss will vary depending on specific material properties, routing layer, trace width, copper roughness, stack-up, environment, etc.
- System designers should determine loss numbers which are representative of their use cases.
- Worst-case conditions are usually encountered under high-temperature / high-humidity environments.
- The following values are representative examples:

PCB Material			PCB trace unit length loss – Nominal Conditions (dB/inch)			PCB trace unit length loss – Worst-case Conditions (dB/inch)		
Category	**Nominal-to-worst-case scaling**	**Signal routing type**	**4 GHz**	**8 GHz**	**16 GHz**	**4 GHz**	**8 GHz**	**16 GHz**
Mid-loss	16%	Stripline	0.65	1.16	2.3	0.75	1.35	2.7
		Microstrip	0.69	1.27	2.4	0.80	1.47	2.8
Low-loss	12%	Stripline	0.50	0.85	1.6	0.56	0.95	1.8
		Microstrip	0.58	1.05	1.8	0.65	1.18	2.0
Ultra-low-loss	8%	Stripline	0.35	0.58	1.02	0.38	0.63	1.1
		Microstrip	0.41	0.72	1.15	0.44	0.77	1.2

Key Points

- Not all “Low-Loss” materials are the same
- It’s critical to understand the loss characteristics at worst-case temperature & humidity

Solving Reach Problem – PCB Material

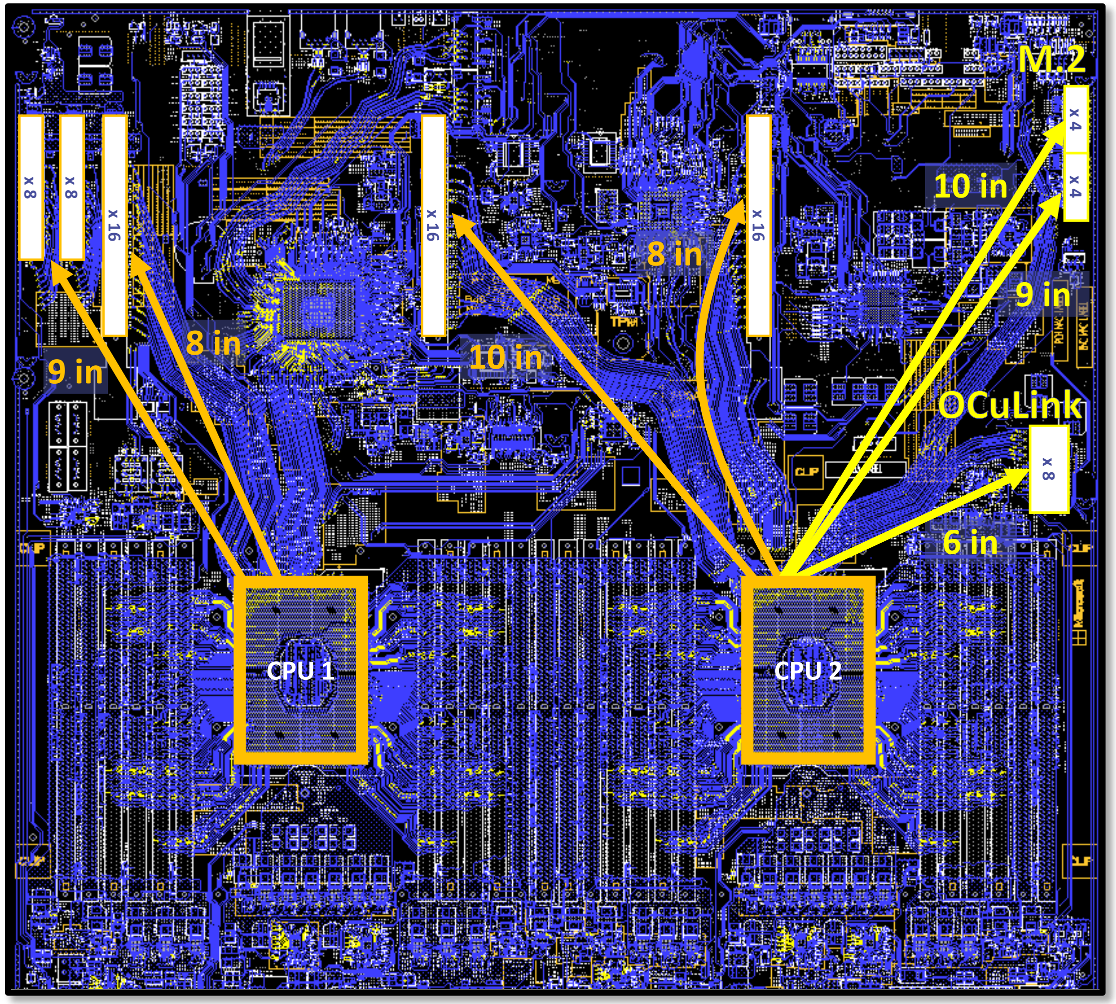
…Based on the noted assumptions, what does this mean for physical System Board channels?

<u>Traditional AIC Topology</u>
(one connector, two vias on system board)

Case	Mid-Loss	Low-Loss	Ultra-Low-Loss
Max system board trace*, 20C	6.3 in	8.2 in	13.3 in
Max system board trace*, 80C	5.5 in	7.3 in	**12.1 in**

*Based on a 36-dB end-to-end channel budget.

Example: Two-Socket Motherboard, OCP Project Olympus



Reference: https://www.opencompute.org/wiki/Server/ProjectOlympus

Solving Reach Problem – Active Components

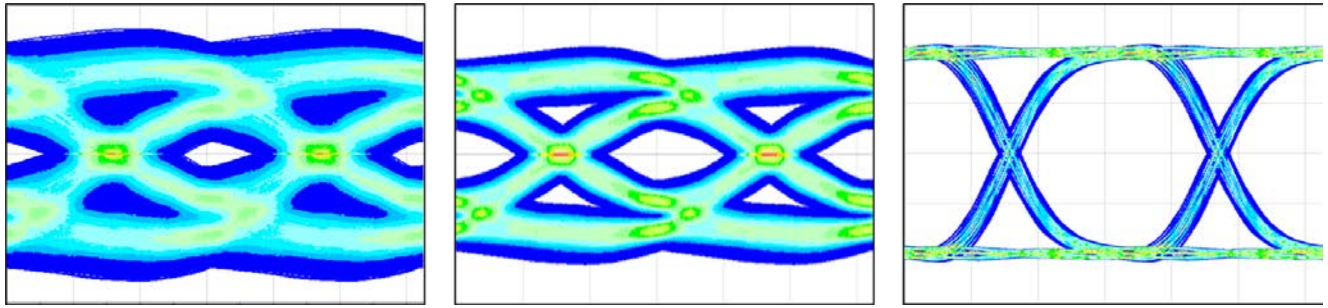
Property	Retimer	Redriver
PCIe protocol participation	Protocol-aware	Protocol unaware
Jitter reduction	Resets entire jitter budget (DDJ, RJ, etc.)	Attenuates DDJ; Amplifies RJ
Equalization capabilities	CTLE, DFE, Tx FIR	CTLE
Adaptation	CTLE, DFE, and Tx FIR automatically adapt to the channel	CTLE setting must be hand-selected based on simulation/experimentation
Diagnostics capabilities	Receiver margining, eye diagram, eye width/height measurement, Link state debug information	None to speak of
Lane-to-Lane skew compensation capabilities	Resets entire skew budget	Does not reset skew budget; may increase total skew
Placement	Anywhere with PCIe-complaint channels on both sides	Not too close to source transmitter; but not too far away
Usage in closed systems (i.e. no open slots)	Recommended; sanctioned use case in PCIe base specification	Highly discouraged. Use at your own risk after extensive simulation and testing
Usage in open systems (i.e. open AIC slot)	Recommended; sanctioned use case in PCIe base specification	Not sanctioned / discouraged

- **Redriver**: A non-protocol-aware, software-transparent extension device [1].
- **Retimer**: A physical layer protocol-aware, software-transparent extension device that forms two separate electrical Link segments [1].
- Active components may be used on a **selective** basis, extending reach for longest channels which require support

Input signal: Attenuated by channel loss

Redriver output: High frequency amplified; uncorrelated jitter persists

Retimer output: Data is recovered and retransmitted; jitter resets



[1] PCIe Base specification

Determining if a Retimer is Required

There are generally two ways to approach this:

A first-order estimate only

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

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

A more thorough, recommended approach

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) ≤ 1E-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 considers other channel characteristics, such as reflections and crosstalk.

Clocking Considerations and Topologies

- Clocking modes: Common clock, SRNS, SRIS
- PCIe® 5.0 reference clock considerations

Clock Architecture

CEM 5.0 will support the following clock architectures:

1. Common Clock
2. SRIS/SRNS (only allowed for bridging to other form-factors for example via CEM 5.0 riser)

Table 2-1. Clocking Architecture Requirements

Clock Architecture	CEM 5.0 Platform	PCIe 5.0 CEM Add-in Card	CEM 5.0 Riser	PCIe 5.0 Retimer
Common Clock	Required	Required	Optional	Required
SRIS	Not Allowed	Not Allowed	Optional	Optional
SRNS	Not Allowed	Not Allowed	Optional	Optional

For CEM 5.0 Platforms and Add-in Cards, common clock is the required clocking architecture. Separate reference clock architectures (SRIS/SRNS) are not CEM compliant.

- CEM-5.0-compliant host platform is required to supply reference clock to the connector.
- CEM-5.0-compliant Add-in Card is required to use the clock supplied at the CEM 5.0 connector.

CEM 5.0 riser implementations have an option to choose either common clock or separate reference clock architectures. The clocking flexibility in riser implementations is allowed to bridge to other form-factors.

Source: PCI Express Card Electromechanical Specification Revision 5.0 Version 0.7

100 MHz REFCLK Considerations

- Jitter limit reduced from 500 fs RMS to **150 fs RMS**
- Note this limit is **after** the PCIe filter function is applied
- This jitter limit includes source jitter and additive jitter of any fanout buffers between the source and the REFCLK receiver

Table 8-18 Jitter Limits for CC Architecture

Data Rate	CC jitter Limit	Notes
2.5 GT/s	86 ps pp	1, 2
5.0 GT/s	3.1 ps RMS	1, 2
8.0 GT/s	1.0 ps RMS	1, 2
16.0 GT/s	0.5 ps RMS	1, 2, 3, 4
32.0 GT/s	0.15 ps RMS	1, 2, 3, 5

- $Jitter_{total} = \sqrt{Jitter_{source}{}^2 + Jitter_{buffer}{}^2}$
- For SRIS systems, rule of thumb is to reduce the RMS jitter limit by $\sqrt{2}$, or to 106 fs
- Maximum down-spread for spread-spectrum clocking (SSC) changed from 5000 ppm to **3000 ppm**

Compliance and Interop Testing

- Electrical compliance
- Interop testing and system robustness

CEM and PHY Test Specifications

Specs in the Review Zone:

- 2020-Aug: PCI Express® Card Electromechanical Specification Revision 5.0, **Version 0.7**
- 2020-Nov: PCI Express® PHY Test Specification Revision 5.0, **Version 0.5**

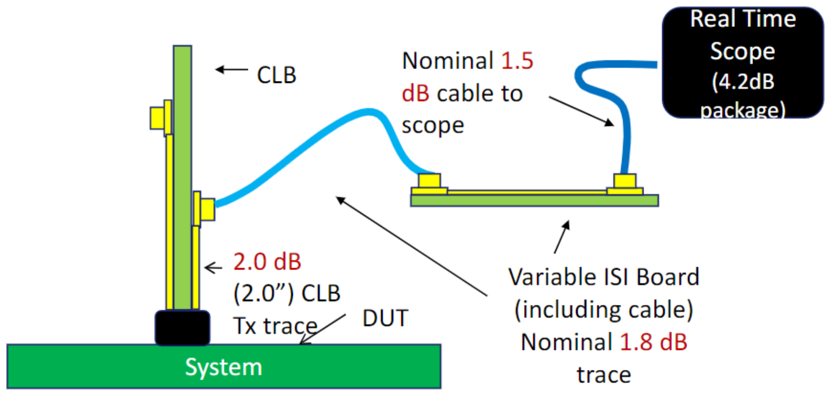
TX Compliance Specs from PCIe® 3.0 to 5.0

TX Compliance Test	Data Rate	Eye Width (1E-12) [ps]	Eye Height (1E-12) [mV]
System Board	32 GT/s	9.688	17.5
	16 GT/s	21.75	19
	8 GT/s	34	41.25
Add-In Card	32 GT/s	10.625	22
	16 GT/s	24.75	23
	8 GT/s	41.25	34

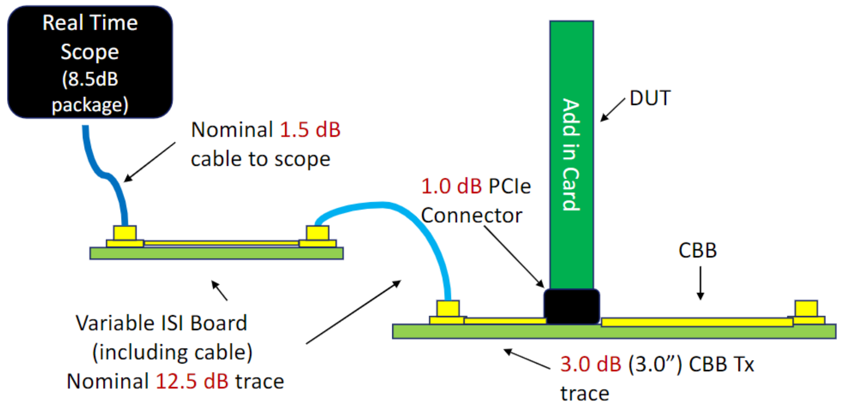
- SigTest Phoenix 5.0.15 (Beta) available for PCIe® 5.0 specification compliance test
- Industry is proving the test solution for TX and RX compliance test

New Developments in TX Compliance Test

System Tx Compliance Test



Add-in Card Tx Compliance Test



- Dual Port test being deprecated for system Tx test
- Using scope embedding instead of ISI board is being explored
- Introducing a system reference clock test is being explored (available on the PHY Test version 0.5 now)

Source: PCIe CEM Previews(2020-Oct)

Interoperability Testing

- Electrical compliance does not guarantee system-level interoperation and robustness
- Test 100s-1000s of A/C and warm reboot cycles to gauge repeatability of normal system operation

Category	Test
Link-up Repeatability Tests	A/C Power Cycle
	Warm Reboot
LTSSM Robustness Tests	Secondary Bus Reset (Hot Reset)
	Link Disable/Enable
	L1 Power State
	Speed Change
	Tx EQ Redo
Special Use-Case Test	Hot Unplug/Plug
In-System Electrical Margin Test	Rx Lane Margining
Application Test	Application Performance Measurement

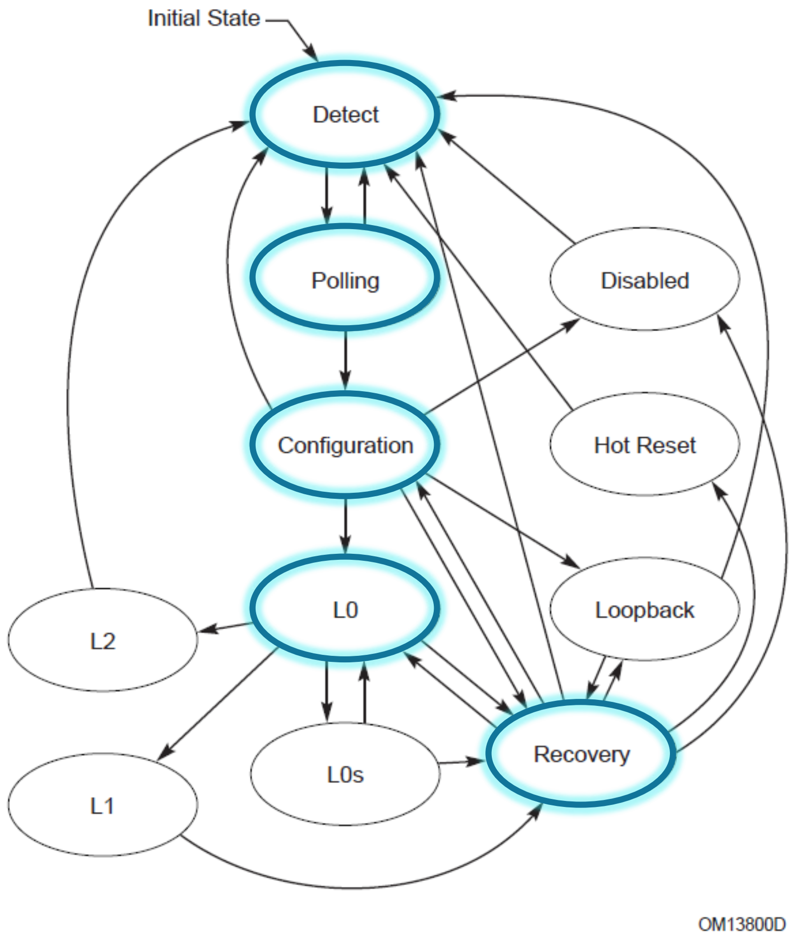


Figure 4-26 Main State Diagram for Link Training and Status State Machine

Interoperability Testing

- Electrical compliance does not guarantee system-level interoperation and robustness
- Test 100s-1000s of **LTSSM stress tests** to ensure corner case operation is robust

Category	Test
Link-up Repeatability Tests	A/C Power Cycle
	Warm Reboot
LTSSM Stress Tests	Secondary Bus Reset (Hot Reset)
	Link Disable/Enable
	L1 Power State
	Speed Change
	Tx EQ Redo
Special Use-Case Test	Hot Unplug/Plug
In-System Electrical Margin Test	Rx Lane Margining
Application Test	Application Performance Measurement

Loopback need not be tested as it is not used in normal Link operation; however it can be useful as an alternate means of measuring electrical margins.

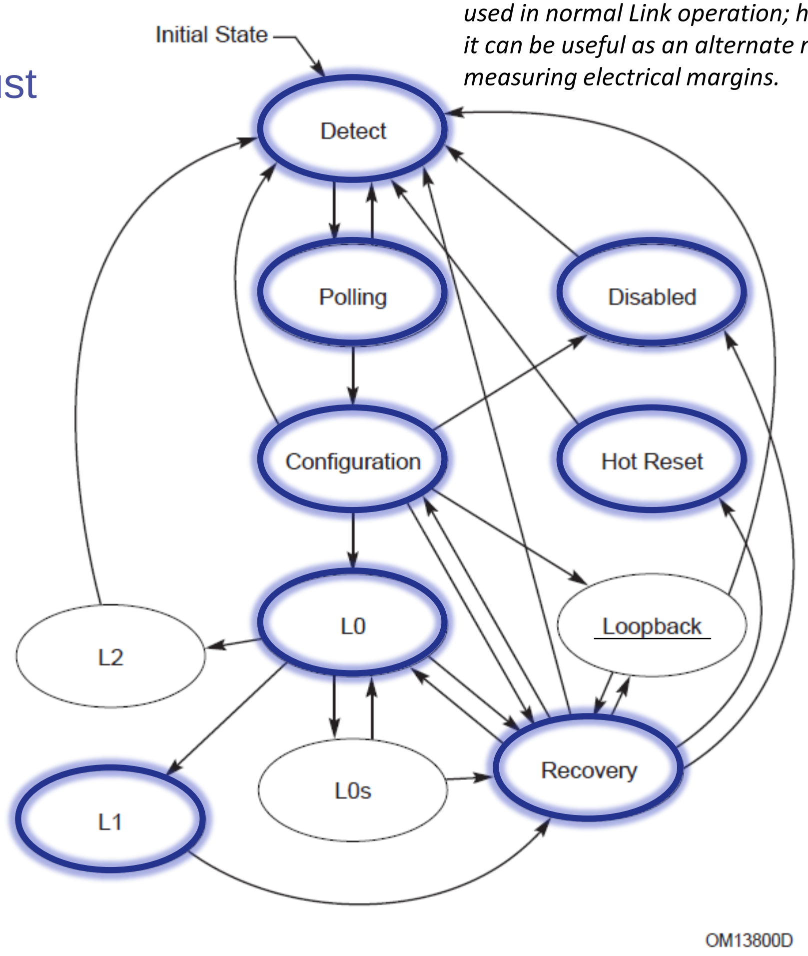


Figure 4-26 Main State Diagram for Link Training and Status State Machine

Interoperability Testing

- Electrical compliance does not guarantee system-level interoperation and robustness
- Test application-specific features, data traffic, and in-system margins

Category	Test
Link-up Repeatability Tests	A/C Power Cycle
	Warm Reboot
LTSSM Stress Tests	Secondary Bus Reset (Hot Reset)
	Link Disable/Enable
	L1 Power State
	Speed Change
	Tx EQ Redo
Special Use-Case Test	Hot Unplug/Plug
In-System Electrical Margin Test	Rx Lane Margining
Application Test	Application Performance Measurement

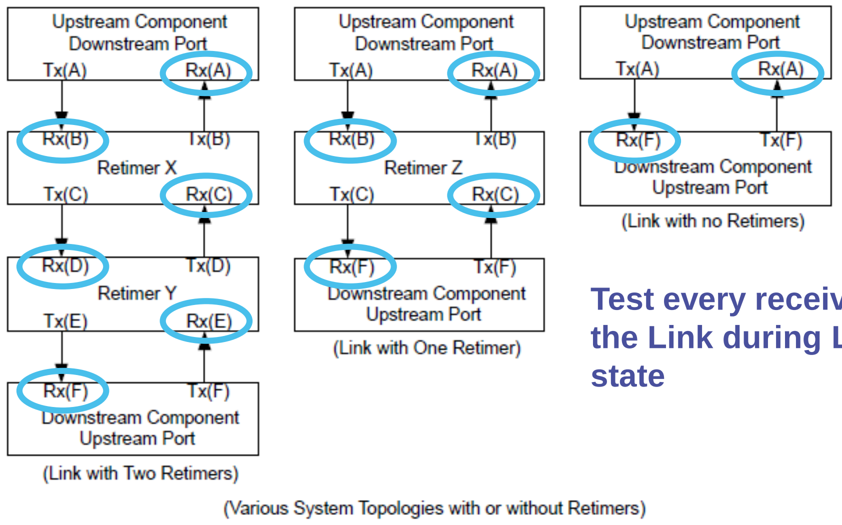


Figure 4-35 Receiver Number Assignment

Test every receiver in the Link during L0 state

Compare measured Rx margins to vendor-specific limits

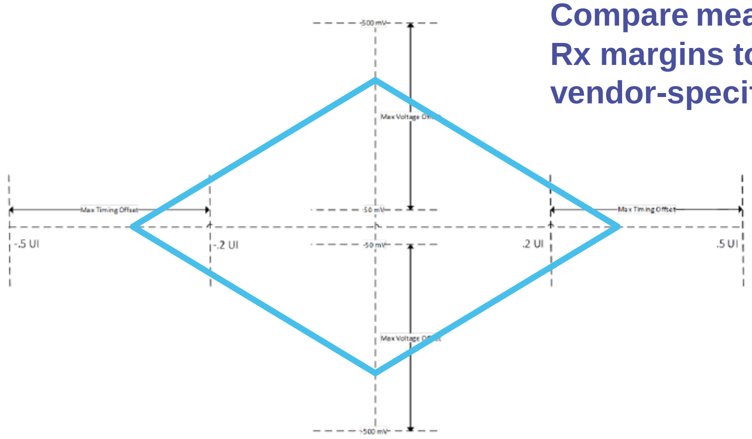


Figure 8-47 Allowed Ranges for Maximum Timing and Voltage Margins

Questions

Thank you for attending the PCI-SIG® Q4 2020 Webinar

For more information please go to www.pcisig.com