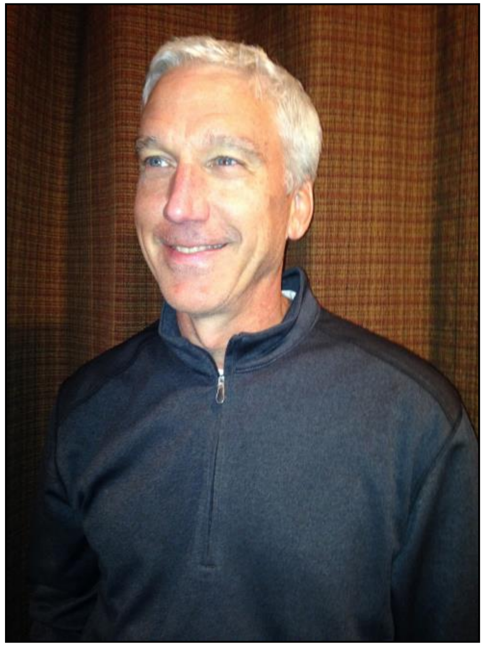
PCI-SIG® Educational Webinar Series 2019

Retimers to the Rescue: PCI Express® Specifications Reach Their Full Potential

Presented by Kurt Lender and Casey Morrison

Meet the Presenters

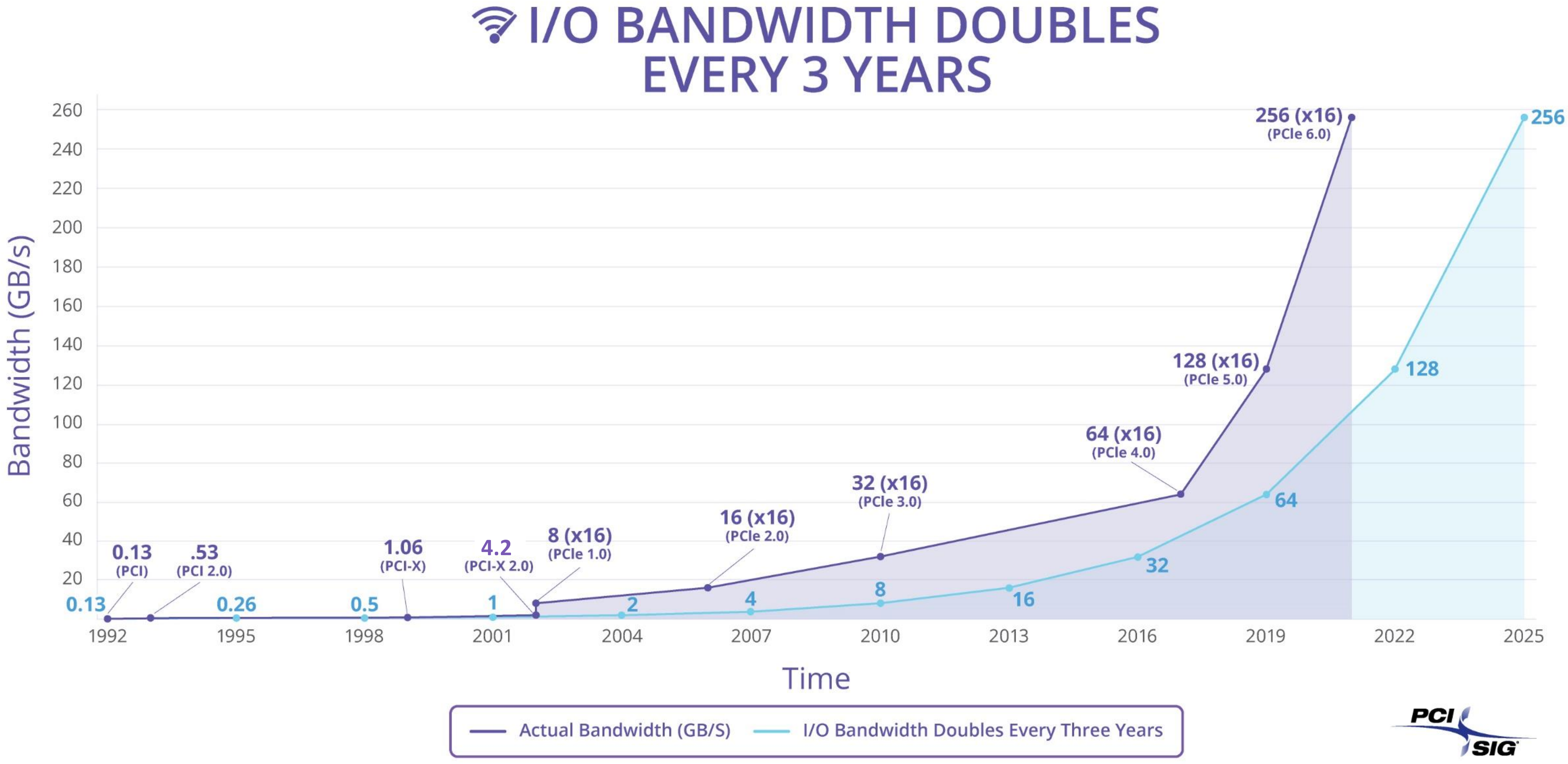


Kurt Lender
PCI-SIG MWG Chair and IHV Enabling Manager, Data Center Group, Intel Corporation

Casey Morrison
Head of Systems and Applications, Astera Labs

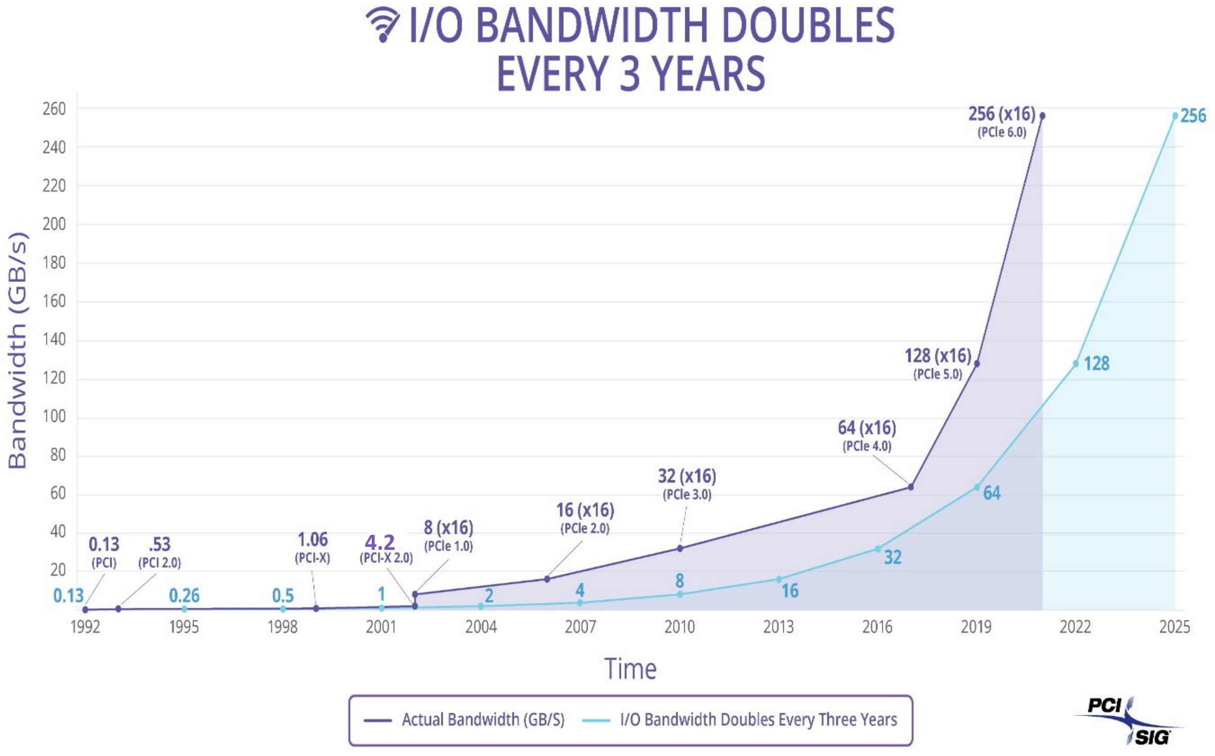
Introduction – Problem Statement

PCIe® Ecosystem Perspective



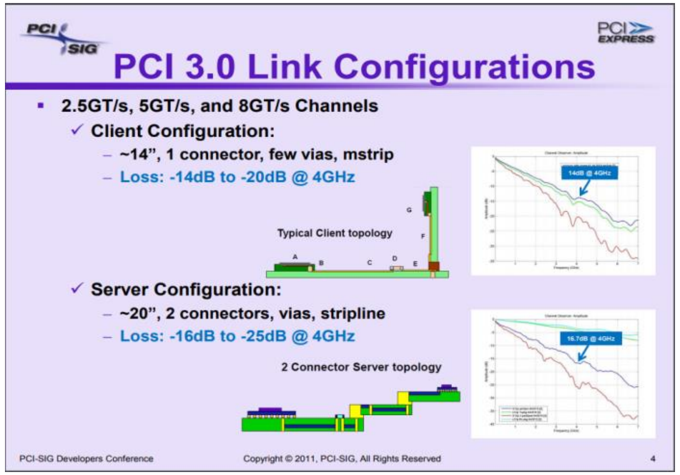
Introduction – Problem Statement

PCIe® Ecosystem Perspective



=

Increasing Gen to Gen Channel Reach Pressures

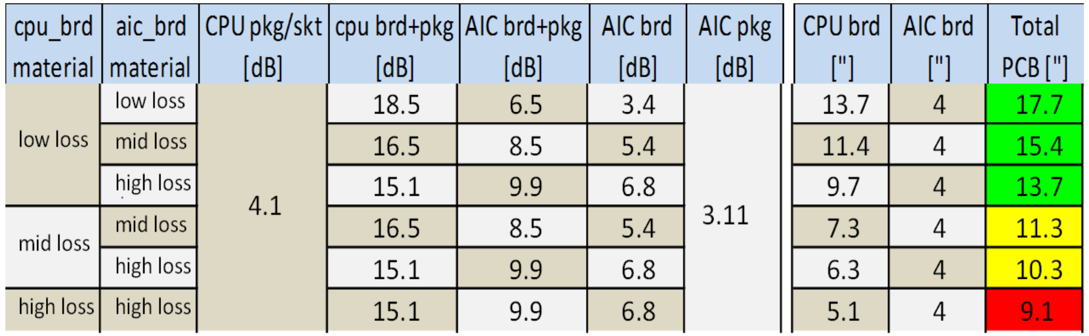


PCI 3.0 Link Configurations

- 2.5GT/s, 5GT/s, and 8GT/s Channels
  - ✓ Client Configuration:
    - ~14”, 1 connector, few vias, mstrip
    - Loss: -14dB to -20dB @ 4GHz
  - ✓ Server Configuration:
    - ~20”, 2 connectors, vias, stripline
    - Loss: -16dB to -25dB @ 4GHz

PCI-SIG Developers Conference Copyright © 2011, PCI-SIG, All Rights Reserved 4

PCIe 4.0 Server Channel Lengths



cpu_brd material	aic_brd material	CPU pkg/skt [dB]	cpu brd+pkg [dB]	AIC brd+pkg [dB]	AIC brd [dB]	AIC pkg [dB]	CPU brd ["]	AIC brd ["]	Total PCB ["]
low loss	low loss	4.1	18.5	6.5	3.4	3.11	13.7	4	17.7
	mid loss		16.5	8.5	5.4		11.4	4	15.4
	high loss		15.1	9.9	6.8		9.7	4	13.7
mid loss	mid loss		16.5	8.5	5.4		7.3	4	11.3
	high loss		15.1	9.9	6.8		6.3	4	10.3
high loss	high loss		15.1	9.9	6.8		5.1	4	9.1

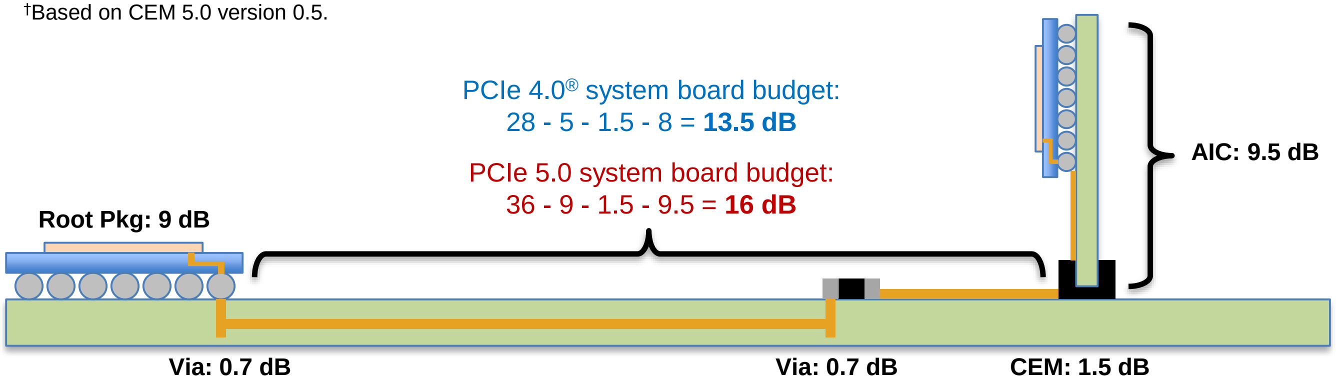
Introduction – Problem Statement

Signal Integrity Perspective

Doubling Speed, Reduced Signal Reach

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

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



PCIe 4.0® system board budget:
28 - 5 - 1.5 - 8 = **13.5 dB**

PCIe 5.0 system board budget:
36 - 9 - 1.5 - 9.5 = **16 dB**

Root Pkg: 9 dB

AIC: 9.5 dB

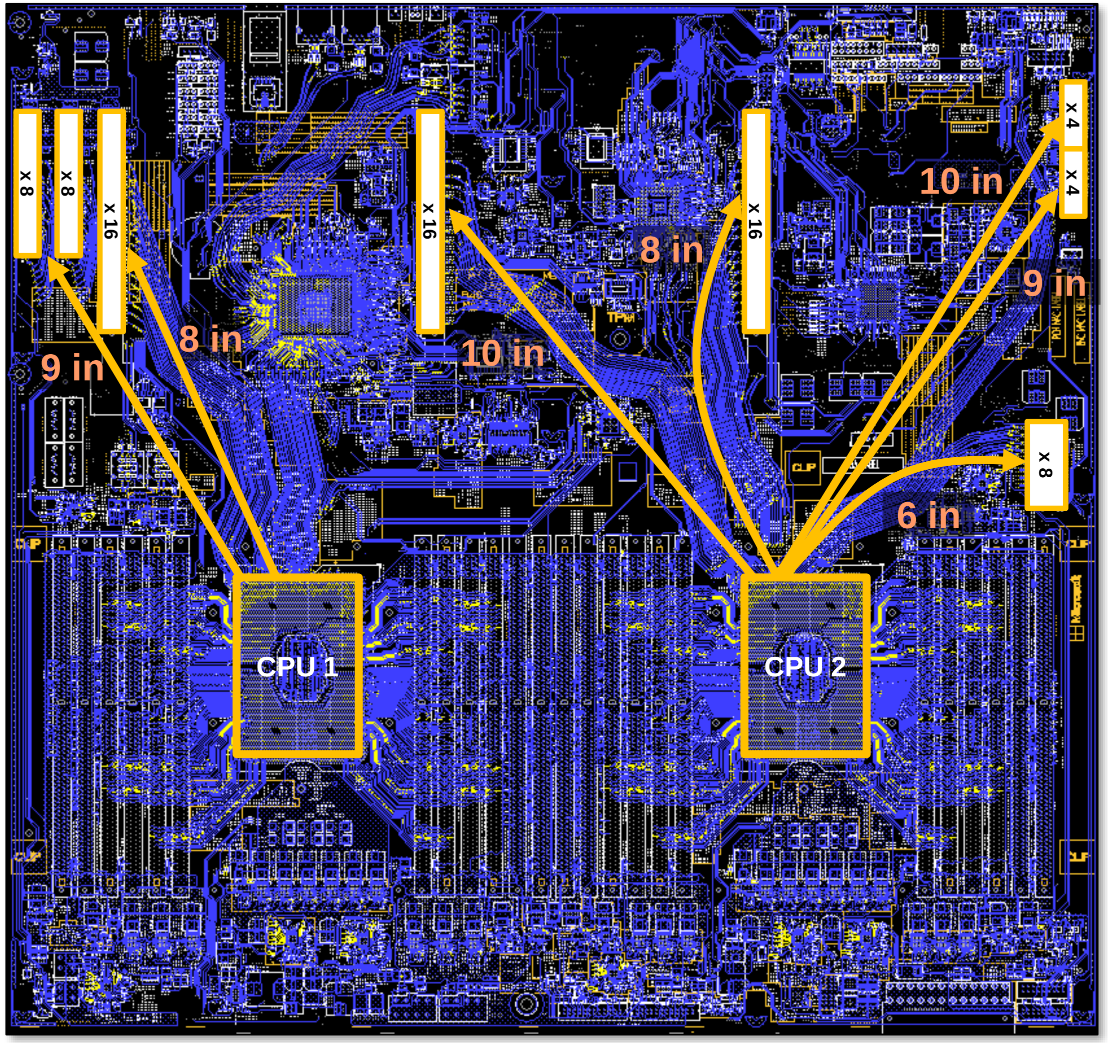
Via: 0.7 dB

Via: 0.7 dB

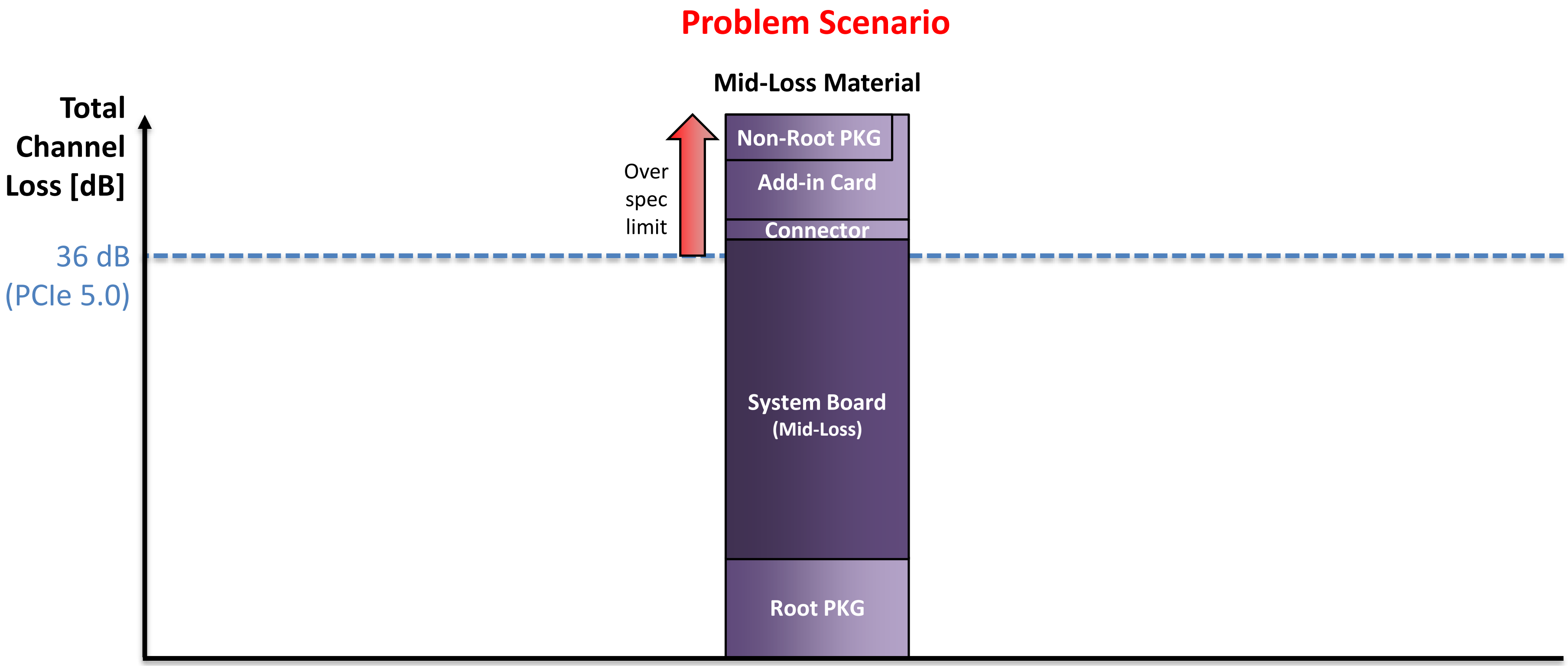
CEM: 1.5 dB

System board budget includes: vias, stubs, AC coupling capacitor, and microstrip/stripline trace

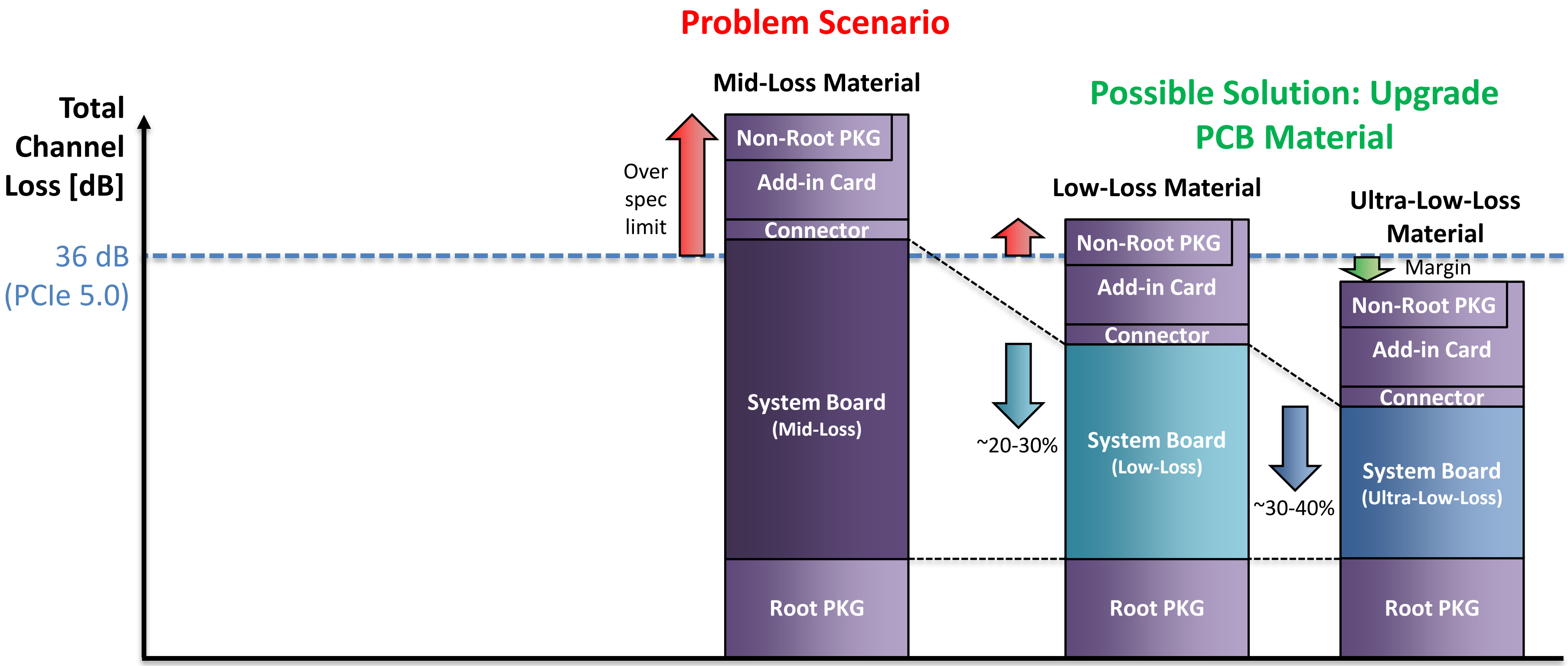
Example: Two-Socket System Board (from OCP)



Ways to Solve the Signal Integrity Problem



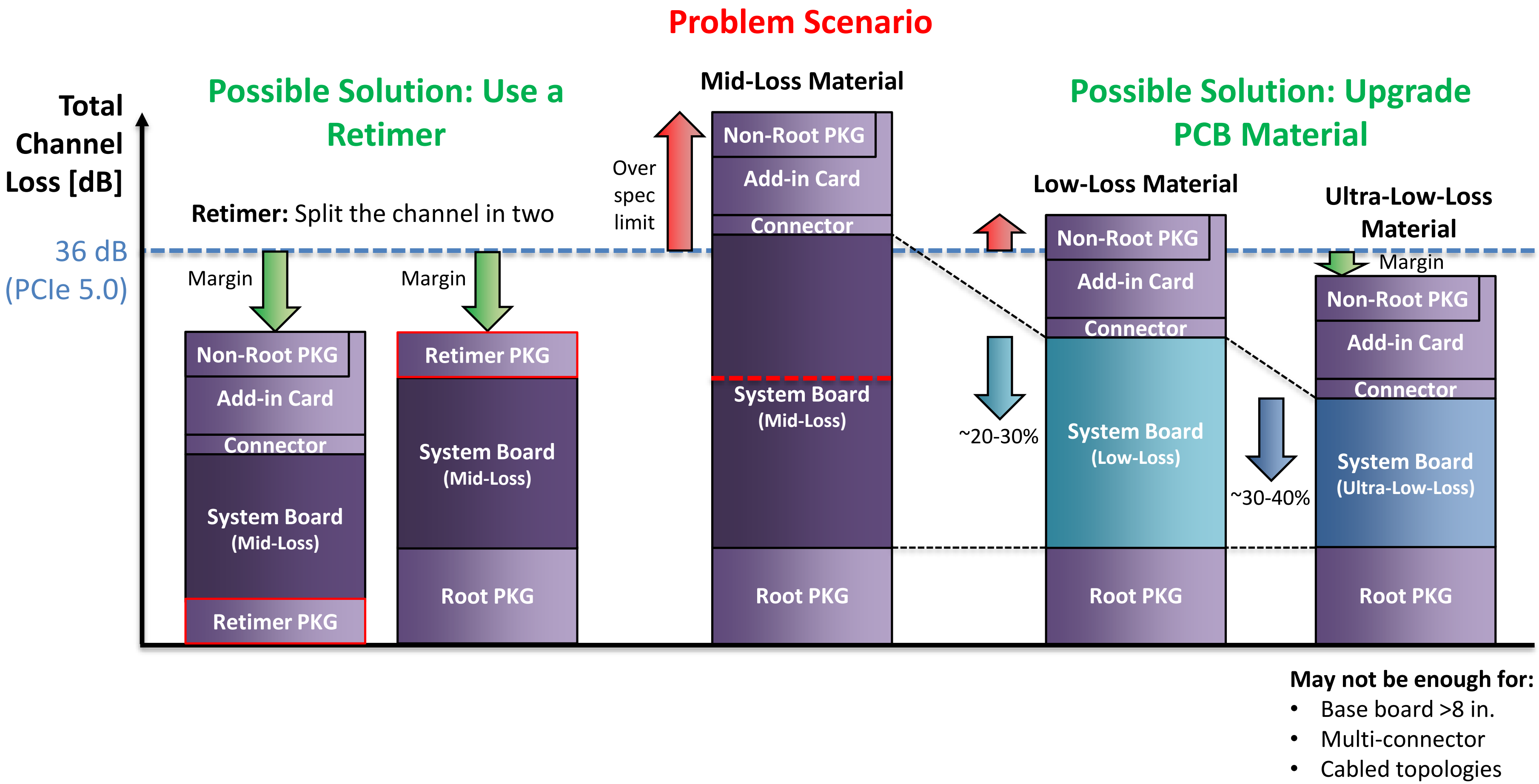
Ways to Solve the Signal Integrity Problem



May not be enough for:

- Base board >8 in.
- Multi-connector
- Cabled topologies

Ways to Solve the Signal Integrity Problem



May not be enough for:

- Base board >8 in.
- Multi-connector
- Cabled topologies

Key Points

- Upgrading PCB material only improves one aspect of total channel: System board
- Even advanced PCB materials may not be enough for longest ports
- Retimers segment the channel into two, creating more margin on each Link segment

PCB Materials

- 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, stackup, environment, etc.
- System designers should determine loss numbers which are representative of their design and use case
- 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

Reach Implications

- A Link which operates “on the edge”—for example, 1E-12 bit error rate (BER)—will enter Recovery at a rate of once every **10 seconds**, and it will Replay a TLP every **1 second** for a x16 Link at 32 GT/s.[1]
- The same analysis shows that if the channel loss is reduced by a few dB, BER can improve significantly.
- System designers want peace of mind, and many employ **Safety Margin** on top of PCIe® channel guidelines.
- **Safety Margin**: self-imposed reduction in channel loss limit to allow for manufacturing variances, simulation-to-measurement correlation mismatches and other unforeseen degradations affecting system performance.

Max Reach for Traditional AIC Topology (one connector + two vias on system board)

	16 GT/s			32 GT/s		
Case	**Mid-Loss**	**Low-Loss**	**Ultra-Low-Loss**	**Mid-Loss**	**Low-Loss**	**Ultra-Low-Loss**
Max system board trace, Nominal conditions	10.0 in	12.7 in	18.6 in	6.2 in	8.6 in	13.5 in
Max system board trace, Worst-case (WC) conditions	8.6 in	11.4 in	17.3 in	5.3 in	7.7 in	12.5 in
Max system board trace, WC *and* 15% safety margin	5.6 in	**7.5 in**	**11.3 in**	3.4 in	**4.8 in**	**7.9 in**

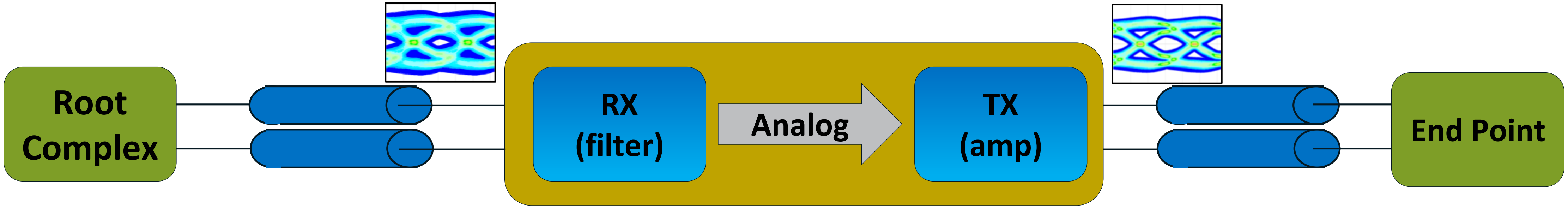
Key Points

- At PCIe 5.0 technology speed, “Low-Loss” material enables ~5-inch system board trace
- Upgrading to “Ultra-low-loss” will enable ~8 inches.

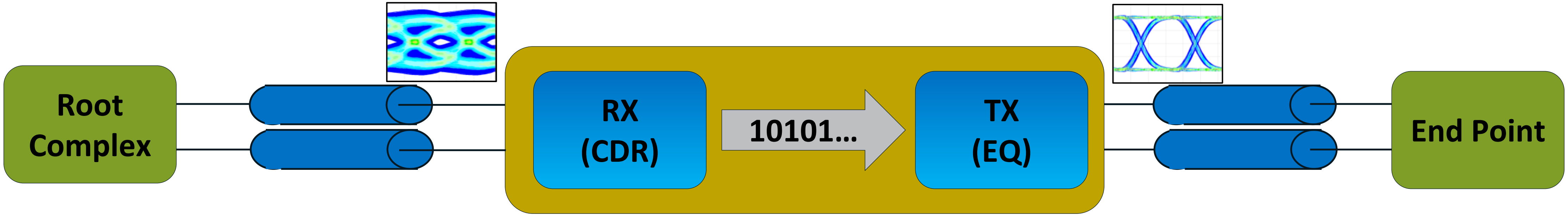
[1] PCI-SIG® DevCon 2019, “Impact of Bit Errors in PCIe 5.0 for Latency-Critical Applications,” https://members.pcisig.com/wg/PCI-SIG/document/13087?downloadRevision=active

Redrivers and Retimers

- **Redriver**
  - Analog signals coming in are filtered and/or amplified
  - Jitter and noise may get worse or at least stay the same



- **Retimer**
  - Analog signals become data inside device, and data is retransmitted
  - Can fully regenerate signals, but at a latency cost

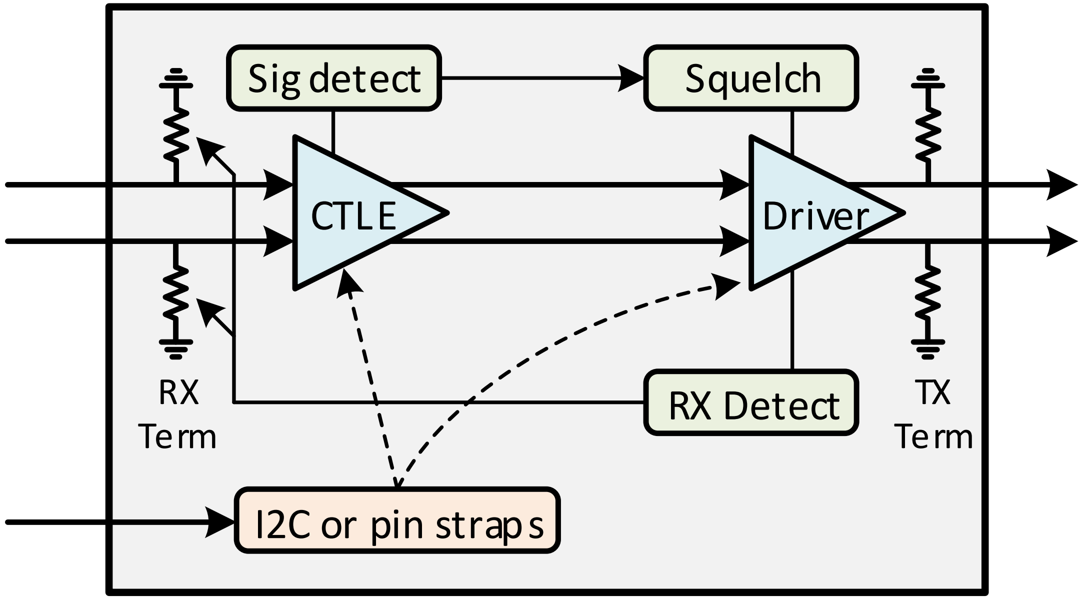


- **“Repeater” is a superset term used to refer to both (caution: use of this term may cause confusion)**

What is a Redriver?

Redriver: Non-protocol-aware software-transparent extension device[1]

- Mostly analog, designed to boost high-frequency portions of a signal
- Data path typically includes a continuous time linear equalizer (CTLE), a wideband gain stage, and linear driver
- Redrivers do not compensate uncorrelated jitter (e.g. RJ, uncorrelated deterministic jitter, etc.)
- Redrivers do not participate in Link EQ
- No formal standard or compliance program



Redriver block diagram

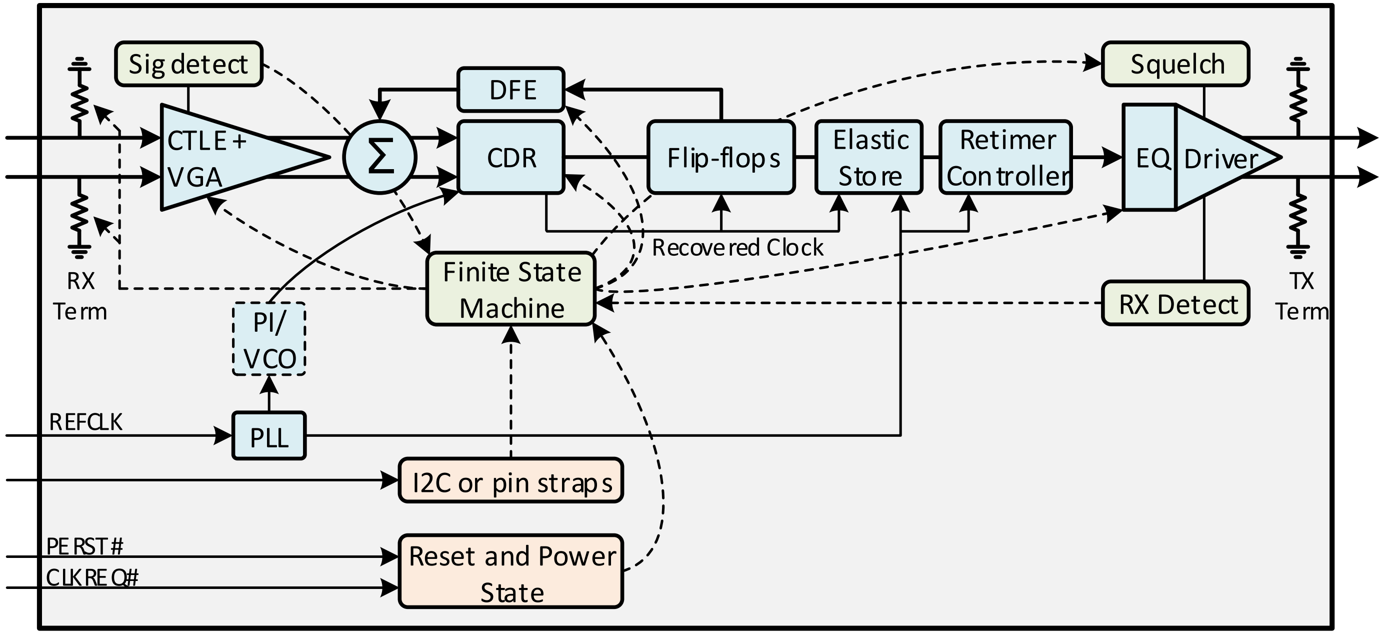
Read more in this PCI-SIG blog paper: https://pcisig.com/pci-express%C2%AE-retimers-vs-redrivers-eye-popping-difference

[1] PCIe 5.0 Base Specification: Terms of Acronyms

What is a Retimer?

Retimer: A physical layer protocol-aware, software-transparent extension device[1]

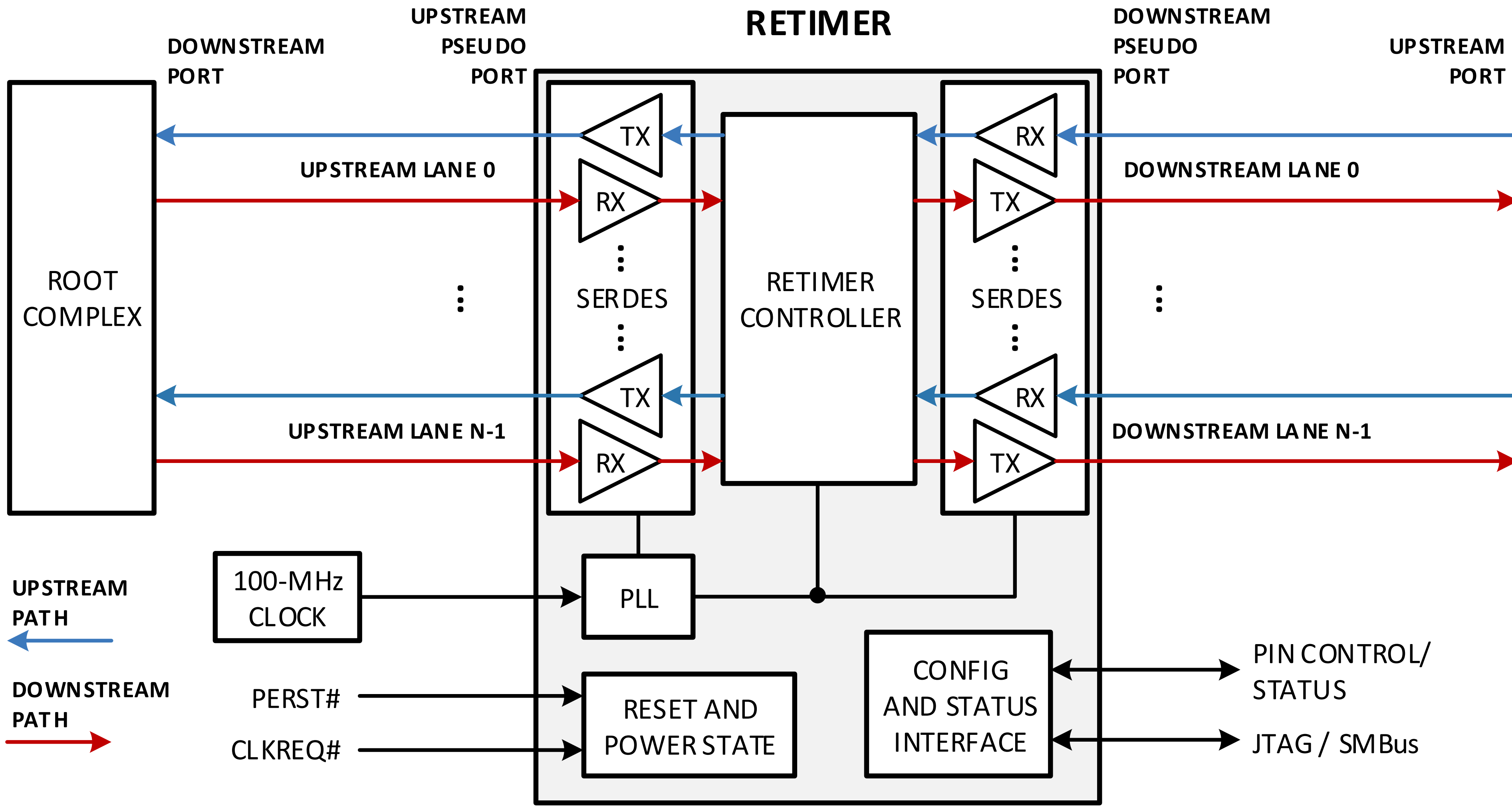
- Covered in PCIe® 4.0 & PCIe 5.0 specifications (Section 4.3)
- Mixed-signal analog/digital device—fully recovers data, extracts clock, and retransmits clean data
- Complies with all PCIe electrical specifications
- Performs Receiver detection and Lane-to-Lane deskew
- Executes Link equalization Phases 2 & 3
- Supports “Equalization to highest rate” and “No equalization needed” PCIe modes



Retimer block diagram

[1] PCIe 5.0 Base Specification: Terms of Acronyms

Retimers in a System



Reach Extension Solutions

Comparison

	PCB Material	Retimer	Redriver
Pros	• Enables modest reach extension for PCIe 4.0 and PCIe 5.0 specifications • No power or latency impact	• Enables 2x to 3x PCIe® channel loss with conventional PCB material • Supported by PCIe 4.0 and 5.0 specs with compliance program	• Enables modest reach extension up to PCIe 4.0 technology but vendor specific • Minimal impact to latency
Cons	• Impacts the cost of the whole PCB • Only enables up to 5-8 inches at 32 GT/s	• Adds power and latency • Impacts BoM cost for select ports	• Not defined in PCIe Base Specification • No formal compliance test • Some impact to power and Bill of Materials (BoM) cost
Comments	• Margin must be kept for AIC and Root Port package • May still require off-board extension for storage	• Usable in open-slot and closed systems	• More viable for closed-slot systems • Some applications up to PCIe 4.0 technology given sufficient amount of testing is performed

Retimer Standards

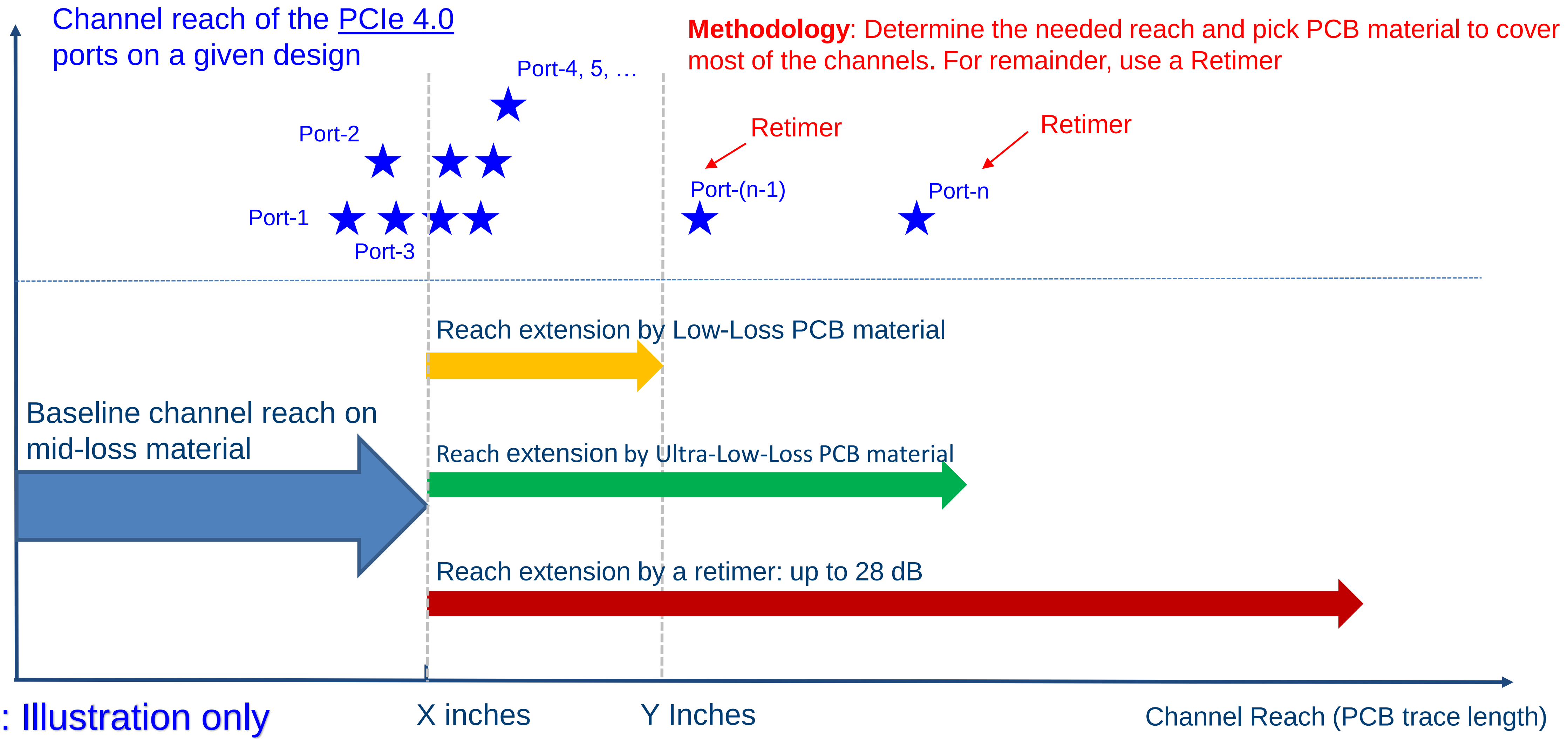
Compliance Testing

- PCI Express® Retimer Test Specification currently at Rev 0.9
- Currently in “FYI testing” phase
- Intent of the Test Specification is to confirm a stand-alone Retimer is compliant to the PCIe 4.0 Base Specification
- Coverage – not all inclusive
  - Electrical Tests
  - Test Macros (Reset, Forwarding, Speed Change, Electrical Idle, etc.)
  - Logical Retimer Tests
  - Interoperability tests
  - Architecture PHY Tests

Designing with a Retimer

High-Level Methodology

Methodology: Determine the needed reach and pick PCB material to cover most of the channels. For remainder, use a Retimer



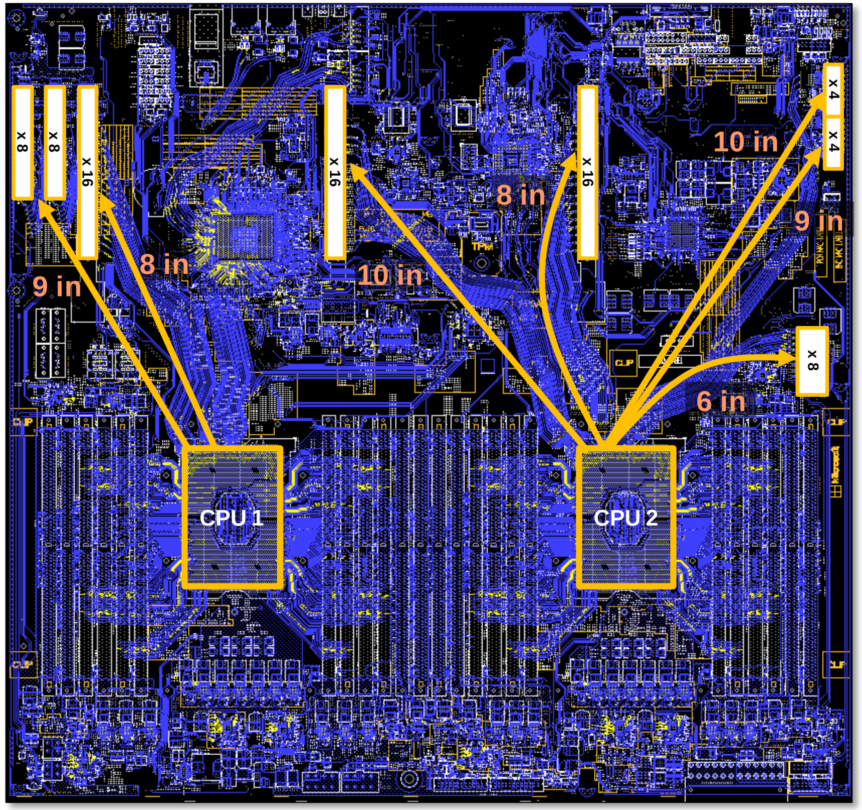
Channel reach of the PCIe 4.0 ports on a given design
Port-4, 5, …
Port-2
Port-1
Port-3
Retimer
Retimer
Port-(n-1)
Port-n
Reach extension by Low-Loss PCB material
Baseline channel reach on mid-loss material
Reach extension by Ultra-Low-Loss PCB material
Reach extension by a retimer: up to 28 dB
X inches
Y Inches
Channel Reach (PCB trace length)

Note: Illustration only

Designing with a Retimer

Example

1. **Scope out the range of trace lengths needed for the system.**



Example:

Link	**Max speed required**	**Approx. Length**	**Special topology considerations**
Slot 1: x8 for SSDs or accelerator	16.0 GT/s	9 in	Standard AIC
Slot 2: x8 for SSDs or accelerator	16.0 GT/s	9 in	Standard AIC
Slot 3: x16 for NIC or GPU/Accelerator	32.0 GT/s	8 in	Standard AIC or Riser
Slot 4: x16 for NIC or GPU/Accelerator	32.0 GT/s	10 in	Standard AIC or Riser
Slot 5: x16 for NIC or GPU/Accelerator	32.0 GT/s	8 in	Standard AIC or Riser
Slot 6: x4 for SSDs	16.0 GT/s	10 in	Internal cable
Slot 7: x4 for SSDs	16.0 GT/s	9 in	Internal cable
Slot 8: x8 for SSDs	16.0 GT/s	6 in	Internal cable

Designing with a Retimer

Example

2. **Chose a combination of PCB material and Retimer to meet system performance and cost requirements**

Example:

Link	Max speed required	Approx. Length	Special topology considerations	Retimer required?			Note
				Mid-Loss	Low-Loss	Ultra-Low-Loss	
Slot 1: x8 for SSDs or accelerator	16.0 GT/s	9 in	Standard AIC	Yes	Maybe	No	1
Slot 2: x8 for SSDs or accelerator	16.0 GT/s	9 in	Standard AIC	Yes	Maybe	No	1
Slot 3: x16 for NIC or GPU/Accelerator	32.0 GT/s	8 in	Standard AIC or Riser	Yes	Yes	Maybe	1
Slot 4: x16 for NIC or GPU/Accelerator	32.0 GT/s	10 in	Standard AIC or Riser	Yes	Yes	Maybe	2
Slot 5: x16 for NIC or GPU/Accelerator	32.0 GT/s	8 in	Standard AIC or Riser	Yes	Yes	Maybe	1
Slot 6: x4 for SSDs	16.0 GT/s	10 in	Internal cable	Yes	Yes	Maybe	3
Slot 7: x4 for SSDs	16.0 GT/s	9 in	Internal cable	Yes	Yes	Maybe	3
Slot 8: x8 for SSDs	16.0 GT/s	6 in	Internal cable	Yes	Maybe	Maybe	3

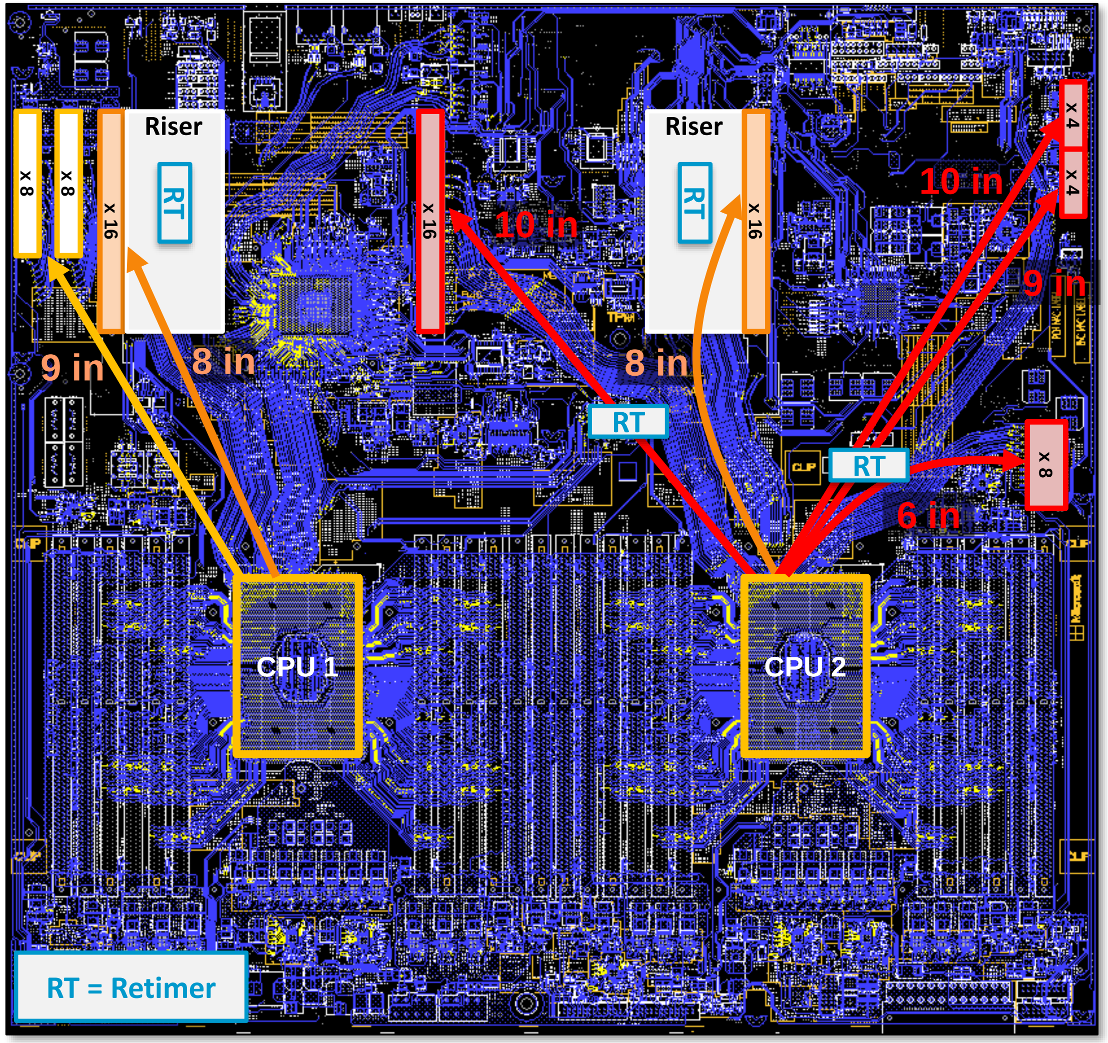
Notes:
1: Need for Retimer depends on whether you want to reserve safety margin or not.
2: With ultra-low-loss material, if a riser card is used, then a Retimer will likely be required on the Riser.
3: Depends on length of cable and number of connectors, but typically >2 connectors will necessitate a Retimer.

Designing with a Retimer

Example

3. **Identify opportunities to group ports requiring a Retimer together to reduce solution size**
   - Multiple x4 and/or x8 Links can utilize a single x16 Retimer, using **bifurcation** as needed.
4. **Determine optimum placement for the Retimer(s)**
   - Place close enough to the slot to allow for a variety of cards and cables to be used, including *passive* riser cards.
   - Consider air flow and routing density.

Bifurcation: Segmenting a xN device (e.g. N=16) into multiple, smaller Links (e.g. x4x4x8).

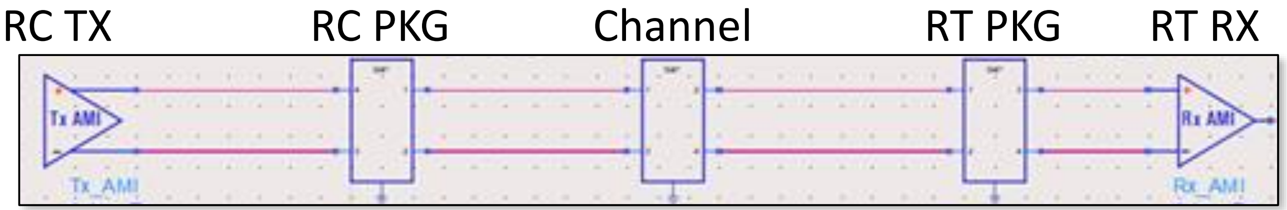


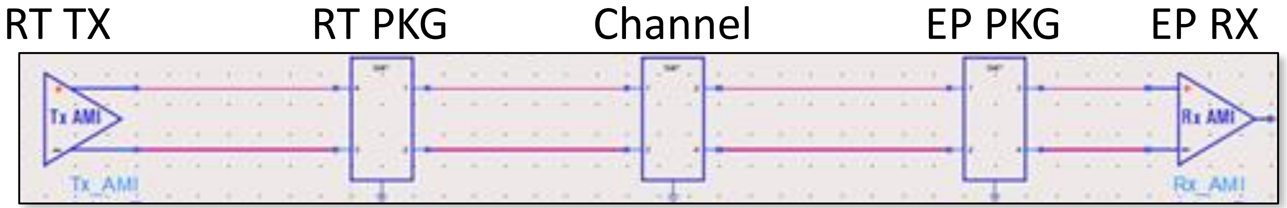
Designing with a Retimer

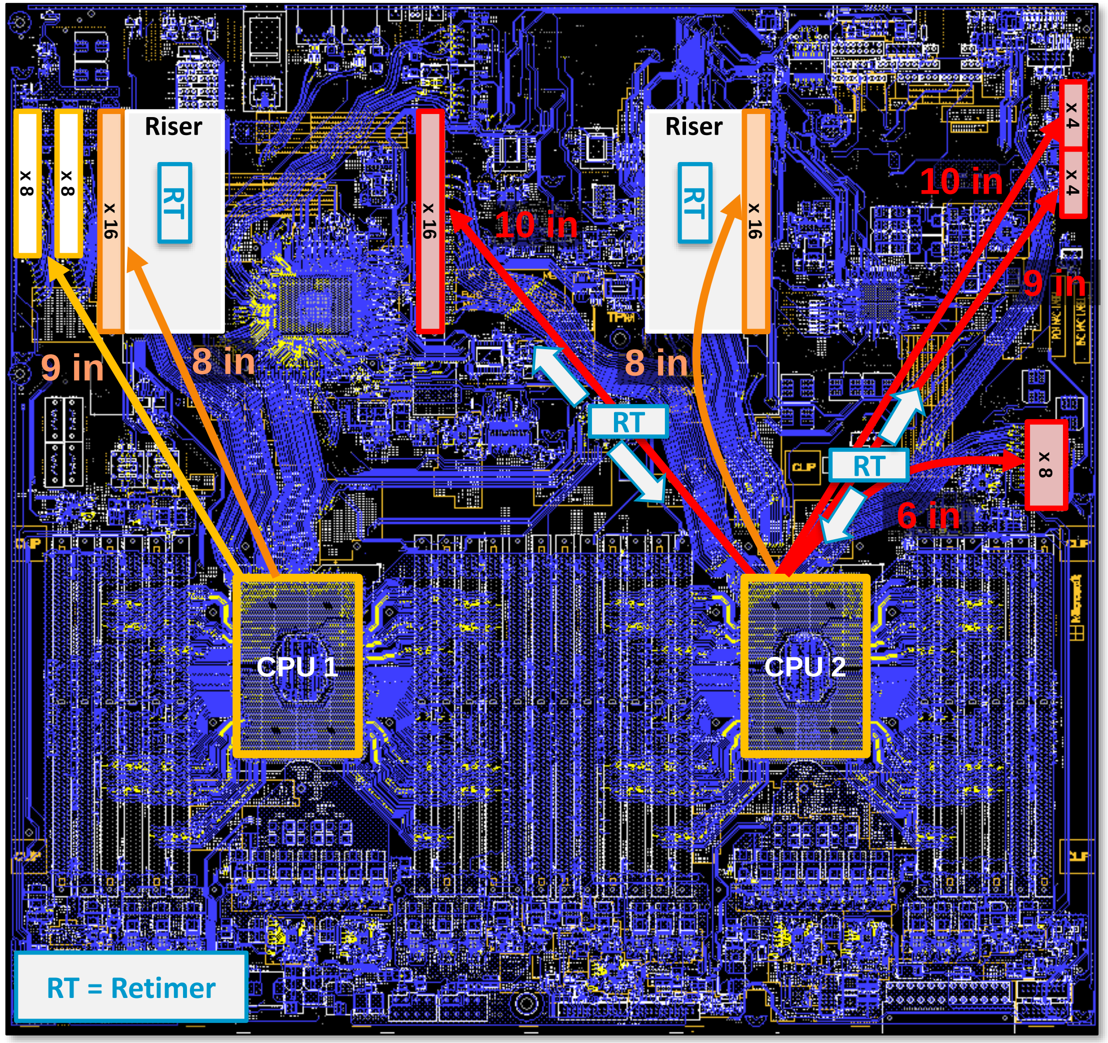
Example

5. **Check Signal Integrity (SI) by running IBIS-AMI simulations, adjust placement as necessary**

   - A Retimer has two Link segments: RC-to-RT and RT-to-EP
   - Each can be simulated independently through SeaSim (to assess the passive channel) or IBIS-AMI (to assess the channel plus RC, RT, and RP).







Retimer Diagnostic Capabilities

Standard Diagnostic Capabilities

- **Slave Loopback**
  - Optional feature in PCIe® Base Spec
  - Allows data to loop back from RC to RT or from EP to RT
- **Receiver Margining**
  - Like any PCIe receiver, Retimers must support Receiver Margining via Control SKP Ordered Sets
  - Eye opening can be assessed on BOTH Pseudo Ports
- **In-Band Register Reads**
  - Read status information from the Retimer via in-band Control SKP Ordered Sets
  - This, unfortunately, requires the Link to be up

Other Possible Diagnostic Capabilities

- **Full Eye Capture**
  - Recording the shape of the eye, beyond just the timing and voltage margin reported by Receiver Margining
- **Protocol Status Reporting**
  - A Retimer is aware of the physical layer protocol events on both the Upstream and Downstream Pseudo Ports.
  - It can record this information and report it to a system controller as needed to facilitate Link debug
  - It can possible generate interrupts to a system controller on important events (e.g. unexpected entry to Recovery)

Future Challenges

- With Speeds increasing, the need for Retimers will continue to increase.
- PCIe 6.0 specification is planning for 64 GT/s using PAM4 signaling and targeting similar channel reach as PCIe 5.0 specification.
- Retimers will need to support the same 64 GT/s PAM4 signaling and operate within the BER constraints required for a low-latency forward error correction (FEC).
- Low latency is key for many emerging PCIe applications: machine learning, artificial intelligence, distributed computing
- Retimers must innovate along with RCs and EPs to keep PCIe Links fast, low-power, and low-latency.

Join PCI-SIG®

PCI-SIG members have access to the PCIe® specification library. If you would like to learn more about joining, please visit the PCI-SIG website: **https://pcisig.com/membership/become-member**

Questions?

Thank You For Attending