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

PCle® Ecosystem Perspective

I/O BANDWIDTH DOUBLES EVERY 3 YEARS

<table>
<thead>
<tr>
<th>Year</th>
<th>Bandwidth (GB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>0.13 (PCI)</td>
</tr>
<tr>
<td>1995</td>
<td>0.53 (PCI 2.0)</td>
</tr>
<tr>
<td>1998</td>
<td>1.06 (PCI-X)</td>
</tr>
<tr>
<td>2001</td>
<td>4.2 (PCI-X 2.0)</td>
</tr>
<tr>
<td>2004</td>
<td>8 (x16) (PCI 3.0)</td>
</tr>
<tr>
<td>2007</td>
<td>16 (x16) (PCI 3.0)</td>
</tr>
<tr>
<td>2010</td>
<td>32 (x16) (PCI 4.0)</td>
</tr>
<tr>
<td>2013</td>
<td>64 (x16) (PCI 5.0)</td>
</tr>
<tr>
<td>2019</td>
<td>128 (x16) (PCI 5.0)</td>
</tr>
<tr>
<td>2022</td>
<td>256 (x16) (PCI 6.0)</td>
</tr>
<tr>
<td>2025</td>
<td>256</td>
</tr>
</tbody>
</table>
Introduction – Problem Statement

PCIe® Ecosystem Perspective

I/O BANDWIDTH DOUBLES EVERY 3 YEARS

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, stripline
    - Loss: -14dB to -20dB @ 4GHz
  - Server Configuration:
    - ~20", 2 connectors, via, stripline
    - Loss: -16dB to -26dB @ 4GHz

Typical Client Topology: 2-Connector Server Topology

PCIe 4.0 Server Channel Lengths

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>low loss</td>
<td></td>
<td>18.5</td>
<td>6.5</td>
<td>3.4</td>
<td></td>
<td></td>
<td>13.7</td>
<td>4</td>
<td>17.7</td>
</tr>
<tr>
<td>mid loss</td>
<td></td>
<td>16.5</td>
<td>8.5</td>
<td>5.4</td>
<td></td>
<td></td>
<td>11.4</td>
<td>4</td>
<td>15.4</td>
</tr>
<tr>
<td>mid loss</td>
<td></td>
<td>15.1</td>
<td>9.9</td>
<td>6.8</td>
<td></td>
<td></td>
<td>9.7</td>
<td>4</td>
<td>13.7</td>
</tr>
<tr>
<td>high loss</td>
<td></td>
<td>16.5</td>
<td>8.5</td>
<td>5.4</td>
<td></td>
<td></td>
<td>7.3</td>
<td>4</td>
<td>11.3</td>
</tr>
<tr>
<td>high loss</td>
<td></td>
<td>15.1</td>
<td>9.9</td>
<td>6.8</td>
<td></td>
<td></td>
<td>6.3</td>
<td>4</td>
<td>10.3</td>
</tr>
<tr>
<td>high loss</td>
<td></td>
<td>15.1</td>
<td>9.9</td>
<td>6.8</td>
<td></td>
<td></td>
<td>5.1</td>
<td>4</td>
<td>9.1</td>
</tr>
</tbody>
</table>

Copyright © 2019 PCI-SIG® - All Rights Reserved
**Introduction – Problem Statement**

**Signal Integrity Perspective**

### Doubling Speed, Reduced Signal Reach

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

*ILfit\_TX\_ROOT\_DEVICE and ILfit\_TX\_NON\_ROOT\_DEVICE parameters in the base specification.
†Based on CEM 5.0 version 0.5.

### Example: Two-Socket System Board (from OCP)

**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

**System board budget includes:** vias, stubs, AC coupling capacitor, and microstrip/ stripline trace
Ways to Solve the Signal Integrity Problem

Total Channel Loss [dB] - 36 dB (PCIe 5.0)

Problem Scenario:

Non-Root PKG
Add-in Card
Connector
System Board (Mid-Loss)
Root PKG
Ways to Solve the Signal Integrity Problem

Problem Scenario

Total Channel Loss [dB]

36 dB (PCIe 5.0)

Mid-Loss Material

Over spec limit

Possible Solution: Upgrade PCB Material

Low-Loss Material

~20-30%

Ultra-Low-Loss Material

~30-40%

May not be enough for:
• Base board >8 in.
• Multi-connector
• Cabled topologies

System Board (Mid-Loss)

Root PKG

Add-in Card

Connector

System Board (Low-Loss)

Root PKG

Add-in Card

Connector

System Board (Ultra-Low-Loss)

Root PKG

Add-in Card

Connector

Non-Root PKG

Non-Root PKG

Non-Root PKG

Root PKG

Root PKG

Root PKG

Root PKG

Root PKG
Ways to Solve the Signal Integrity Problem

**Possible Solution: Use a Retimer**
- **Retimer:** Split the channel in two
- **Mid-Loss Material**
- **Possible Solution: Upgrade PCB Material**
- **Low-Loss Material**
- **Ultra-Loss Material**

**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

May not be enough for:
- Base board >8 in.
- Multi-connector
- Cabled topologies

---

Copyright © 2019 PCI-SIG® - All Rights Reserved
### 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:

<table>
<thead>
<tr>
<th>Category</th>
<th>Nominal-to-worst-case scaling</th>
<th>PCB Material</th>
<th>Nominal Conditions - Signal routing type</th>
<th>Nominal Conditions - 4 GHz</th>
<th>Nominal Conditions - 8 GHz</th>
<th>Nominal Conditions - 16 GHz</th>
<th>Worst-case Conditions - 4 GHz</th>
<th>Worst-case Conditions - 8 GHz</th>
<th>Worst-case Conditions - 16 GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid-loss</td>
<td>16%</td>
<td>Stripline</td>
<td>0.65</td>
<td>1.16</td>
<td>2.3</td>
<td>0.75</td>
<td>1.35</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Microstrip</td>
<td>0.69</td>
<td>1.27</td>
<td>2.4</td>
<td>0.80</td>
<td>1.47</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>Low-loss</td>
<td>12%</td>
<td>Stripline</td>
<td>0.50</td>
<td>0.85</td>
<td>1.6</td>
<td>0.56</td>
<td>0.95</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Microstrip</td>
<td>0.58</td>
<td>1.05</td>
<td>1.8</td>
<td>0.65</td>
<td>1.18</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Ultra low-loss</td>
<td>8%</td>
<td>Stripline</td>
<td>0.35</td>
<td>0.58</td>
<td>1.02</td>
<td>0.38</td>
<td>0.63</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Microstrip</td>
<td>0.41</td>
<td>0.72</td>
<td>1.15</td>
<td>0.44</td>
<td>0.77</td>
<td>1.2</td>
<td></td>
</tr>
</tbody>
</table>

### 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)

<table>
<thead>
<tr>
<th>Case</th>
<th>16 GT/s</th>
<th>32 GT/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max system board trace, Nominal conditions</td>
<td>Mid-Loss</td>
<td>Low-Loss</td>
</tr>
<tr>
<td>10.0 in</td>
<td>12.7 in</td>
<td>18.6 in</td>
</tr>
<tr>
<td>Max system board trace, Worst-case (WC) conditions</td>
<td>8.6 in</td>
<td>11.4 in</td>
</tr>
<tr>
<td>Max system board trace, WC and 15% safety margin</td>
<td>5.6 in</td>
<td>7.5 in</td>
</tr>
</tbody>
</table>

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.

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

---

[^1]: PCIe 5.0 Base Specification: Terms of Acronyms

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

[1] PCIe 5.0 Base Specification: Terms of Acronyms
Retimers in a System

![Diagram of Retimers in a System]

- **Retimers**: Used for retiming signals in a system.
- **SERDES**: Serial Data Encoder/Decoder.
- **RETIMER**: Component for retiming signals.
- **UPSTREAM LANE 0**: First upstream lane.
- **DOWNSTREAM LA NE 0**: First downstream lane.
- **ROOT COMPLEX**: The starting point of the upstream lane path.
- **END POINT**: The end of the downstream lane path.
- **PLL**: Phase-Locked Loop for frequency and phase control.
- **CONFIG AND STATUS INTERFACE**: For control and status signals.
- **100-MHz CLOCK**: Clock source for operation.
- **PERST#**: Power-on Reset signal.
- **CLKREQ#**: Clock Request signal.
- **JTAG / SMBus**: Interface for diagnostic and configuration.
- **PIN CONTROL/STATUS**: Interface for pin control and status.

**Notes:**
- Retimers are crucial in maintaining signal integrity across different lanes and operations.
- The diagram illustrates the flow of signals from the root complex to the end point, with retiming and control interfaces as key components.
### Reach Extension Solutions

#### Comparison

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

Channel reach of the PCIe 4.0 ports on a given design

Baseline channel reach on mid-loss material

Note: Illustration only

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

Reach extension by Low-Loss PCB material

Reach extension by Ultra-Low-Loss PCB material

Reach extension by a retimer: up to 28 dB

Port-1 Port-2 Port-3 Port-4, 5, … Port-(n-1) Port-n

Copyright © 2019 PCI-SIG® - All Rights Reserved
1. Scope out the range of trace lengths needed for the system.

Example:

<table>
<thead>
<tr>
<th>Link</th>
<th>Max speed required</th>
<th>Approx. Length</th>
<th>Special topology considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slot 1: x8 for SSDs or accelerator</td>
<td>16.0 GT/s</td>
<td>9 in</td>
<td>Standard AIC</td>
</tr>
<tr>
<td>Slot 2: x8 for SSDs or accelerator</td>
<td>16.0 GT/s</td>
<td>9 in</td>
<td>Standard AIC</td>
</tr>
<tr>
<td>Slot 3: x16 for NIC or GPU/Accelerator</td>
<td>32.0 GT/s</td>
<td>8 in</td>
<td>Standard AIC or Riser</td>
</tr>
<tr>
<td>Slot 4: x16 for NIC or GPU/Accelerator</td>
<td>32.0 GT/s</td>
<td>10 in</td>
<td>Standard AIC or Riser</td>
</tr>
<tr>
<td>Slot 5: x16 for NIC or GPU/Accelerator</td>
<td>32.0 GT/s</td>
<td>8 in</td>
<td>Standard AIC or Riser</td>
</tr>
<tr>
<td>Slot 6: x4 for SSDs</td>
<td>16.0 GT/s</td>
<td>10 in</td>
<td>Internal cable</td>
</tr>
<tr>
<td>Slot 7: x4 for SSDs</td>
<td>16.0 GT/s</td>
<td>9 in</td>
<td>Internal cable</td>
</tr>
<tr>
<td>Slot 8: x8 for SSDs</td>
<td>16.0 GT/s</td>
<td>6 in</td>
<td>Internal cable</td>
</tr>
</tbody>
</table>
Designing with a Retimer

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

Example:

<table>
<thead>
<tr>
<th>Slot</th>
<th>Link</th>
<th>Max speed required</th>
<th>Approx. Length</th>
<th>Special topology considerations</th>
<th>Retimer required?</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Slot 1: x8 for SSDs or accelerator</td>
<td>16.0 GT/s</td>
<td>9 in</td>
<td>Standard AIC</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Slot 2: x8 for SSDs or accelerator</td>
<td>16.0 GT/s</td>
<td>9 in</td>
<td>Standard AIC</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Slot 3: x16 for NIC or GPU/Accelerator</td>
<td>32.0 GT/s</td>
<td>8 in</td>
<td>Standard AIC or Riser</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Slot 4: x16 for NIC or GPU/Accelerator</td>
<td>32.0 GT/s</td>
<td>10 in</td>
<td>Standard AIC or Riser</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>Slot 5: x16 for NIC or GPU/Accelerator</td>
<td>32.0 GT/s</td>
<td>8 in</td>
<td>Standard AIC or Riser</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Slot 6: x4 for SSDs</td>
<td>16.0 GT/s</td>
<td>10 in</td>
<td>Internal cable</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Slot 7: x4 for SSDs</td>
<td>16.0 GT/s</td>
<td>9 in</td>
<td>Internal cable</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>Slot 8: x8 for SSDs</td>
<td>16.0 GT/s</td>
<td>6 in</td>
<td>Internal cable</td>
<td>Yes</td>
<td>3</td>
</tr>
</tbody>
</table>

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.
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).
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
The Path to PCIe® 6.0 Specification

- 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.
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