PCI EXPRESS® 6.0: A LOW LATENCY, HIGH BANDWIDTH, HIGH RELIABILITY AND COST-EFFECTIVE INTERCONNECT WITH 64.0 GT/S PAM-4 SIGNALING

Dr. Debendra Das Sharma
Intel Fellow and Director of I/O Technologies and Standards
Data Center Group
Director, PCI-SIG Board and co-chair CXL Board Technical Task Force
Agenda

• Introduction to PCI Express® and Load-Store Interconnects
• Evolution of Data Rates in PCI Express
• Key Metrics and Requirements for PCIe 6.0
• PAM-4 and Error Assumptions/ Characteristics
• Error Correction and Detection: FEC, CRC, and Retry
• Flit Mode
• Key Metrics and Requirements for PCIe 6.0 – Evaluation
• Conclusions
Cloud Computing Landscape Today

Silicon Photonics
High-bandwidth connectivity at 100G and beyond

Programmable Switching
P4-programmable scale-out fabric with uncompromising performance

Rack of Servers
Programmable infrastructure acceleration for demanding data movement with Smart NIC

Core/ Edge Network & Inter-DC Network

Hyperscale Data Center and Edge

Spine Switch

Data Center as a Computer – Interconnects are key to driving warehouse scale efficiency!
### Taxonomy, Characteristics, and Trends of Interconnects

<table>
<thead>
<tr>
<th>Category</th>
<th>Type and Scale</th>
<th>Data Rate/Characteristics</th>
<th>PHY Latency (Tx + Rx)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency Tolerant</td>
<td>Networking / Fabric</td>
<td>56/112 GT/s-&gt; 224 GT/s (PAM4)</td>
<td>100+ ns w/ FEC (20ns+ w/o FEC)</td>
</tr>
<tr>
<td>(Narrow, fast)</td>
<td>Data Center Scale</td>
<td>4-8 Lanes, cables/backplane</td>
<td></td>
</tr>
<tr>
<td>Latency Sensitive</td>
<td>Load-Store I/O Arch. Ordering</td>
<td>32 GT/s (NRZ) -&gt; PCIe Gen6 64 GT/s (PAM4)</td>
<td>&lt;10ns (Tx+ Rx: PHY-PIPE) 0-1ns FEC overhead</td>
</tr>
<tr>
<td>(Wide, high speed)</td>
<td>(PCIe/ CXL / UPI)</td>
<td>Hundreds of Lanes: Power, Cost, Si-Area, Backwards Compatible, Latency, On-board -&gt; cables/backplanes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Node level -&gt; sub-Rack</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Latency Sensitive Load-Store I/O moving to 64.0 GT/s using PAM-4: innovations on track to meet latency, area, and cost challenges
Load-Store I/O: 101

• Ability to directly access memory (CPU, I/O)
• Memory mapped into system memory space
  – Coherent or Non-coherent access
  – Accesses across PCIe non-coherent
  – Accesses across CXL can be either
• Some form of ordering or cache coherency
  – PCIe: Producer-Consumer Ordering Semantics
• Transactions are guaranteed to be delivered and completed in a reasonable time
  – no dropped packets, no software based retry
  – typically hardware based link level replay on error
• Timeout and Error reporting hierarchy
  – Hardware based error containment guarantees

(Load-Store I/O (PCIe, CXL, SMP coherency) based on a common PCIe PHY => PCIe needs to stay Low-latency with 0-latency add generationally)

Device A
Write Data
Write Flag

Device B
Read Flag
Read Data

(Producer Consumer Ordering Model: Reading updated Flag guarantees reading updated Data)
PCI EXPRESS: LAYERED PROTOCOL

- PCI compatibility, configuration, driver model
- PCIe architecture enhanced configuration model
- Split-transaction, packet-based protocol
- Credit-based flow control, virtual channels
- Logical connection between devices
- Reliable data transport services (CRC, Retry, Ack/Nak)
- Physical information exchange
- Interface initialization and maintenance
- Market segment specific form factors
- Evolutionary and revolutionary
PCI-SIG®: An Open Industry Consortium

Founded in 1992
Organization that **defines the PCI Express® (PCIe®)** I/O bus specifications and related form factors
830+ member companies located worldwide
Creating specifications and mechanisms to support compliance and interoperability
Agenda

• Introduction to PCI Express® and Load-Store Interconnects
• Evolution of Data Rates in PCI Express
• Key Metrics and Requirements for PCIe 6.0
• PAM-4 and Error Assumptions/ Characteristics
• Error Correction and Detection: FEC, CRC, and Retry
• Flit Mode
• Key Metrics and Requirements for PCIe 6.0 – Evaluation
• Conclusions
Evolution of PCI-Express

- Double data rate every gen in ~3 years
- Full backward compatibility
- Ubiquitous I/O: PC, Hand-held, Workstation, Server, Cloud, Enterprise, HPC, Embedded, IoT, Automotive, AI
- One stack / silicon, multiple form-factors
- Different widths (x1/ x2/ x4/ x8/ x16) and data rates fully inter-operable
  – a x16 Gen 5 interoperates with a x1 Gen 1!
- PCIe deployed in all computer systems since 2003 for all I/O needs

PCIe continues its impressive run of doubling bandwidth for six generations spanning 2 decades!

<table>
<thead>
<tr>
<th>PCIe Specification</th>
<th>Data Rate(Gb/s) (Encoding)</th>
<th>x16 B/W per dirn**</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>2.5 (8b/10b)</td>
<td>32 Gb/s</td>
<td>2003</td>
</tr>
<tr>
<td>2.0</td>
<td>5.0 (8b/10b)</td>
<td>64 Gb/s</td>
<td>2007</td>
</tr>
<tr>
<td>3.0</td>
<td>8.0 (128b/130b)</td>
<td>126 Gb/s</td>
<td>2010</td>
</tr>
<tr>
<td>4.0</td>
<td>16.0 (128b/130b)</td>
<td>252 Gb/s</td>
<td>2017</td>
</tr>
<tr>
<td>5.0</td>
<td>32.0 (128b/130b)</td>
<td>504 Gb/s</td>
<td>2019</td>
</tr>
<tr>
<td>6.0 (WIP)</td>
<td>64.0 (PAM-4, Flit)</td>
<td>1024 Gb/s (~1Tb/s)</td>
<td>2021*</td>
</tr>
</tbody>
</table>

* - Projected  ** - bandwidth after encoding overhead
Bandwidth Drivers for PCIe 6.0

- Device side: Networking (800G in early 2020s), Accelerators, FPGA/ASICs, Memory
- Alternate Protocols on PCIe
- As the per socket compute capability grows at an exponential pace, so does I/O needs — we have already added a lot of Lanes per socket (currently 128 Lanes) => speed has to go up
- But .. we need to meet the cost, performance, power metrics as an ubiquitous I/O with hundreds of Lanes in a platform

(Artificial Intelligence • High-performance • High-bandwidth
Automotive • High-performance • Reliability • Availability • Serviceability
Enterprise Servers • Redundancy/failover • Ubiquity • Power savings
PC/Mobile/IoT • Faster performance • Power efficiency • Low latency
Storage • Faster data transfer • Better user experience • Ubiquity
Cloud • Scalable architecture • Increased performance • Reduced TCO

(New Usage Models: Cloud, AI/Analytics, Edge)

New Usage models are driving bandwidth demand – doubling every three years
### Key Metrics for PCIe 6.0: Requirements

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Rate</strong></td>
<td>64 GT/s, PAM4 (double the bandwidth per pin every generation)</td>
</tr>
<tr>
<td><strong>Latency</strong></td>
<td>&lt;10ns adder for Transmitter + Receiver over 32.0 GT/s (including FEC)</td>
</tr>
<tr>
<td></td>
<td>(We can not afford the 100ns FEC latency as networking does with PAM-4)</td>
</tr>
<tr>
<td><strong>Bandwidth Inefficiency</strong></td>
<td>&lt;2 % adder over PCIe 5.0 across all payload sizes</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>0 &lt; FIT &lt;&lt; 1 for a x16 (FIT – Failure in Time, number of failures in 10^9 hours)</td>
</tr>
<tr>
<td><strong>Channel Reach</strong></td>
<td>Similar to PCIe 5.0 under similar set up for Retimer(s) (maximum 2)</td>
</tr>
<tr>
<td><strong>Power Efficiency</strong></td>
<td>Better than PCIe 5.0</td>
</tr>
<tr>
<td><strong>Low Power</strong></td>
<td>Similar entry/ exit latency for L1 low-power state</td>
</tr>
<tr>
<td></td>
<td>Addition of a new power state (L0p) to support scalable power consumption with bandwidth usage without interrupting traffic</td>
</tr>
<tr>
<td><strong>Plug and Play</strong></td>
<td>Fully backwards compatible with PCIe 1.x through PCIe 5.0</td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td>HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform</td>
</tr>
</tbody>
</table>

Need to make the right trade-offs to meet each of these metrics!
Agenda

- Introduction to PCI Express® and Load-Store Interconnects
- Evolution of Data Rates in PCI Express
- Key Metrics and Requirements for PCIe 6.0
- PAM-4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Key Metrics and Requirements for PCIe 6.0 – Evaluation
- Conclusions
**PAM-4 Signaling at 64.0 GT/s**

- **PAM4 signaling:**
  - Pulse Amplitude Modulation 4-level
  - 4 levels (2 bits) encoded in same Unit Interval (UI)
  - 3 eyes
  - Helps channel loss (same Nyquist as 32.0 GT/s)
- **Reduced eye height (EH) and width (EW):**
  - Increases susceptibility to errors
  - 3 eyes in same UI
- **Gray Coding to help minimize errors in UI**
- **Precoding to minimize errors in a burst**
- **Voltage levels define encoding (Tx/ Rx)**

### Voltage Level

<table>
<thead>
<tr>
<th>Voltage Level</th>
<th>2- Bit Encoding</th>
<th>DC Balance Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>-Vtx</td>
<td>+3</td>
</tr>
<tr>
<td>01</td>
<td>-Vtx/3</td>
<td>+1</td>
</tr>
<tr>
<td>10</td>
<td>+Vtx/3</td>
<td>-1</td>
</tr>
<tr>
<td>11</td>
<td>+Vtx</td>
<td>-3</td>
</tr>
</tbody>
</table>

### Encoding per UI (2bit) | Tx Voltage | Rx Voltage (V) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>-Vtx</td>
<td>V &lt;= Vth1</td>
</tr>
<tr>
<td>01</td>
<td>-Vtx/3</td>
<td>Vth1 &lt; V &lt;= Vth2</td>
</tr>
<tr>
<td>11</td>
<td>+Vtx/3</td>
<td>Vth2 &lt; V &lt;= Vth3</td>
</tr>
<tr>
<td>10</td>
<td>+Vtx</td>
<td>V &gt; Vth3</td>
</tr>
</tbody>
</table>
Error Assumptions and Characteristics w/ PAM-4

Parameters of interest: FBER and error correlation within Lane and across Lanes

- **FBER** – First bit error rate
  - Probability of the first bit error occurring at the Receiver
- Receiving Lane may see a burst propagated due to DFE
  - The number of errors from the burst can be minimized
    - Constrain DFE tap weights - balance TxEQ, CTLE and DFE equalization
- **Correlation of errors across Lanes**
  - Due to common source of errors (e.g., power supply noise)
  - Conditional probability that a first error in a Lane => errors in nearby Lanes
- BER depends FBER and the error correlation in a Lane and across Lanes
Agenda

- Introduction to PCI Express® and Load-Store Interconnects
- Evolution of Data Rates in PCI Express
- Key Metrics and Requirements for PCIe 6.0
- PAM-4 and Error Assumptions/Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Key Metrics and Requirements for PCIe 6.0 – Evaluation
- Conclusions
Handling Errors and Evaluation Metrics

- Two mechanisms to correct errors
  - Correction through FEC (Forward Error Correction)
    - Latency and complexity increases exponentially with the number of Symbols corrected
  - Detection of errors by CRC => Link Level Retry (a strength of PCIe)
    - Detection is linear: latency, complexity and bandwidth overheads
    - Need a robust CRC to keep FIT << 1 (FIT: Failure in Time – No of failures in $10^9$ hours)

- Metrics: Prob of Retry (or b/w loss due to retry) and FIT

- Need to use both means of correction to achieve:
  - Low latency and complexity
  - Retry probability at acceptable level (no noticeable performance impact)
  - Low Bandwidth overhead due to FEC, CRC, and retry

Need to keep FEC correction latency low (2ns) to meet the performance needs of Load/Store I/O
Our Approach: Light-weight FEC and Retry

• Light-weight FEC & strong CRC
• FEC gets to a reasonable retry rate
• Keep latency (including retry) low
• We are better off retrying a packet with $10^{-6}$ (or $10^{-5}$) probability with a retry latency of 100ns
  – better than always paying a FEC latency impact of 150ns+ in networking

Low latency mechanism w/ FBER of $10^{-6}$ to meet the metrics (latency, area, power, bandwidth)
Agenda

• Introduction to PCI Express® and Load-Store Interconnects
• Evolution of Data Rates in PCI Express
• Key Metrics and Requirements for PCIe 6.0
• PAM-4 and Error Assumptions/Characteristics
• Error Correction and Detection: FEC, CRC, and Retry
• Flit Mode
• Key Metrics and Requirements for PCIe 6.0 – Evaluation
• Conclusions
Flit Mode with PCIe 6.0

- Flit (flow control unit) based
- FEC needs fixed set of bytes
- Correction in flit => detection (CRC) in flits => Retry at flit level
- Lower data rates also in Flit Mode if enabled
- Flit size: 256B
  - 236B TLP, 6B DLP, 8B CRC, 6B FEC
  - No Sync hdr, Framing Token or per packet CRC
  - Improved b/w utilization w/ overhead amortization
  - Flit accumulation Latency:
    - 2ns x16, 4ns x8, 8 ns x4, 16 ns x2, 32 ns x1
    - Ack/ credit slot => low Latency/ low storage

Low latency improves performance and reduces area
Replay in Flit Mode

- Flit with NOP-only TLPs not replayed unless the subsequent flit also had an uncorrectable error
- On a replay, the Transmitter can choose to skip over the NOP Flits
- All replayed flits have the Replay Cmd = 11b (w/ Tx sequence number sent)
• Single Symbol Correct interleaved FEC plus 64-b CRC works really well for raw FBER of 1E-6 even with high Lane correlation
  – Retry probability per flit is 5 x 10^-6
  – B/W loss is 0.05% even with go-back-n
  – FIT is almost 0
  – Can mitigate the bandwidth loss significantly by adopting retry only the non-NOP TLP flit

FBER 1E-6 meets the performance goals with a light-weight FEC
PCIe 6.0 Flit Mode Bandwidth at 64.0 GT/s

- Bandwidth increase = 2X (BW efficiency of flit mode) / (BW efficiency in non-flit mode)

- Overall we see a >2X improvement in bandwidth (benefits most systems)
  - Efficiency gain reduces as TLP size increases
  - Beyond 512 B (128 DW) payload goes below 2

- Bandwidth efficiency improvement in flit mode due to the amortization of CRC, DLP, and ECC over a flit (8% overhead) – works out better than sync hdr, DLLP, Framing Token per TLP, and 4B CRC per TLP overheads in PCIe 5.0

Bandwidth Efficiency improvement causes > 2X bandwidth gain for up to 512B Payload in 64.0 GT/s flit mode
Latency Impact of Flit Mode

- Flit accumulation in Rx only (Tx pipeline).
- FEC + CRC delay expected to be ~ 1-2 ns.
- Expected Latency savings due to removal of sync hdr, fixed flit sizes (no framing logic, no variable sized TLP/ CRC processing) is not considered in Tables here.
- With twice the data rate and the above optimizations, realistically expect to see lower latency except for x2 and x1 for smaller payload TLPs –worst case ~10ns adder.

<table>
<thead>
<tr>
<th>Data Size (DW)</th>
<th>TLP Size (DW)</th>
<th>Latency in ns for 128b/130b @ 32.0GT/s</th>
<th>Latency in ns in Flit Mode @ 64.0 GT/s</th>
<th>Latency Increase due to accumulation (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>6.09375</td>
<td>18</td>
<td>11.90625</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>10.15625</td>
<td>20</td>
<td>9.84375</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>14.21875</td>
<td>22</td>
<td>7.78125</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>22.34375</td>
<td>26</td>
<td>3.65625</td>
</tr>
<tr>
<td>32</td>
<td>36</td>
<td>38.59375</td>
<td>34</td>
<td>-4.59375</td>
</tr>
<tr>
<td>64</td>
<td>68</td>
<td>71.09375</td>
<td>50</td>
<td>-21.09375</td>
</tr>
<tr>
<td>128</td>
<td>132</td>
<td>136.09375</td>
<td>82</td>
<td>-54.09375</td>
</tr>
<tr>
<td>256</td>
<td>260</td>
<td>266.09375</td>
<td>146</td>
<td>-120.09375</td>
</tr>
<tr>
<td>512</td>
<td>516</td>
<td>526.09375</td>
<td>274</td>
<td>-252.09375</td>
</tr>
<tr>
<td>1024</td>
<td>1028</td>
<td>1046.09375</td>
<td>530</td>
<td>-516.09375</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Size (DW)</th>
<th>TLP Size (DW)</th>
<th>Latency in ns for 128b/130b @ 32.0GT/s</th>
<th>Latency in ns in Flit Mode @ 64.0 GT/s</th>
<th>Latency Increase due to accumulation (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>0.380859375</td>
<td>1.125</td>
<td>0.744140625</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.634765625</td>
<td>1.25</td>
<td>0.615234375</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>1.088671875</td>
<td>1.375</td>
<td>0.486328125</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>1.396484375</td>
<td>1.625</td>
<td>0.228515625</td>
</tr>
<tr>
<td>32</td>
<td>36</td>
<td>2.412109375</td>
<td>2.125</td>
<td>-0.287109375</td>
</tr>
<tr>
<td>64</td>
<td>68</td>
<td>4.443359375</td>
<td>3.725</td>
<td>-1.318359375</td>
</tr>
<tr>
<td>128</td>
<td>132</td>
<td>8.505859375</td>
<td>5.125</td>
<td>-3.380859375</td>
</tr>
<tr>
<td>256</td>
<td>260</td>
<td>16.63085938</td>
<td>9.125</td>
<td>-7.505859375</td>
</tr>
<tr>
<td>512</td>
<td>516</td>
<td>32.88085938</td>
<td>17.125</td>
<td>-15.75585938</td>
</tr>
<tr>
<td>1024</td>
<td>1028</td>
<td>65.38085938</td>
<td>33.125</td>
<td>-32.25585938</td>
</tr>
</tbody>
</table>

Meets or exceeds the latency expectations.
Agenda

• Introduction to PCI Express® and Load-Store Interconnects
• Evolution of Data Rates in PCI Express
• Key Metrics and Requirements for PCIe 6.0
• PAM-4 and Error Assumptions/Characteristics
• Error Correction and Detection: FEC, CRC, and Retry
• Flit Mode
• Key Metrics and Requirements for PCIe 6.0 – Evaluation
• Conclusions
### Key Metrics for PCIe 6.0: Evaluation

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Expectations</th>
<th>Evaluation (Trend)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Rate</td>
<td>64 GT/s, PAM4 (double bandwidth per pin every generation)</td>
<td>Meets (must do)</td>
</tr>
<tr>
<td>Latency</td>
<td>&lt;10ns adder for Transmitter + Receiver over 32.0 GT/s (including FEC)</td>
<td>Exceeds (Savings in latency with &lt;10ns for x1/ x2 cases)</td>
</tr>
<tr>
<td>Bandwidth Inefficiency</td>
<td>&lt;2 % adder over PCIe 5.0 across all payload sizes</td>
<td>Exceeds (getting &gt;2X bandwidth in most cases)</td>
</tr>
<tr>
<td>Reliability</td>
<td>0 &lt; FIT &lt;&lt; 1 for a x16 (FIT – Failure in Time, failures in 10^9 hrs)</td>
<td>Meets</td>
</tr>
<tr>
<td>Channel Reach</td>
<td>Similar to PCIe 5.0 under similar set up for Retimer(s) (maximum 2)</td>
<td>Meets</td>
</tr>
<tr>
<td>Power Efficiency</td>
<td>Better than PCIe 5.0</td>
<td>Design dependent – expected to meet</td>
</tr>
<tr>
<td>Low Power</td>
<td>Similar entry/ exit latency for L1 low-power state</td>
<td>Design dependent – expected to meet; L0p looks promising</td>
</tr>
<tr>
<td></td>
<td>Addition of a new power state (L0p) for scalable power consumption with bandwidth usage</td>
<td></td>
</tr>
<tr>
<td>Plug and Play</td>
<td>Fully backwards compatible with PCIe 1.x through PCIe 5.0</td>
<td>Meets</td>
</tr>
<tr>
<td>Others</td>
<td>HVM, cost-effective, scales to hundreds of Lanes in platform</td>
<td>Expected to Meet</td>
</tr>
</tbody>
</table>

On track to meet or exceed requirements on all key metrics
PCIe 6.0 is at Rev 0.7 level; Rev 0.9 imminent

Very challenging in multiple fronts

- New signaling with PAM-4: tradeoff around errors/ correlation, channels, performance/ area, and circuit complexity to double the bandwidth

- Metrics (latency, bandwidth efficiency, area, cost, power) which are significantly more challenging than what other standards have done with PAM-4 at lower speeds
  - e.g., 100+ ns FEC latency on other standards vs our single digit ns latency targets; 12+% bandwidth inefficiency in other standards vs <2% inefficiency targets for us)

- We are on track to exceed or meet the requirements

- Need to continue to do due diligence though analysis, simulations, and test silicon characterization to ensure we have a robust specification

- We have the combined innovation capability of 800+ members with a track record of delivering flawlessly against challenges for more than two decades – we will deliver this time also!!

The journey continues …
THANK YOU

FOR YOUR CONTINUED SUPPORT
TOGETHER WE WILL CONTINUE TO
BUILD GREAT PRODUCTS!!