PCIe® 6.0 Specification: A High-Performance Interconnect for Storage Networking Challenges

Dr. Mohiuddin Mazumder
Co-chair, PCI-SIG® Electrical Work Group
Senior Principal Engineer
Intel Corporation
Agenda

- Background
- Key Metrics and Requirements for PCIe® 6.0 Specification
- PAM4 and Error Assumptions/ Characteristics
- Error Correction and Detection: FEC, CRC, and Retry
- Flit Mode
- Ordered Sets handling with high error rate
- Low Power enhancements: L0p
- Area reduction: Shared Credits
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action
PCle® Interconnect for Storage

- ~70-80% of data-center SSDs use PCIe as the interface because of its:
  - High performance – data rate and scalable widths (x1, x2, x4, x8, x16)
  - Low latency – directly connects storage device to the host
  - Extended RAS (Reliability, Availability, and Serviceability) features
  - Standard form factors (e.g., m.2, u.2, Add-In-Card, and EDSFF)
  - Low power
  - Reduced Total Cost of Ownership
- Examples of usages that benefit most from PCIe (NVMe®) SSDs:
  - Database, AI/ML, HPC, Virtualization, Edge Computing, Automotive, Gaming, 3D Graphics

**Enterprise SSD Capacity Shipment Forecast by Interface**

NVMe technology is the leading interface for SSDs

- **PCIe (NVMe) SSD Capacity (TBs)**

PCIe® architecture delivers a high performance, low-latency interconnect between the storage SSDs and the host CPU/switch
Examples of PCIe® Form Factors and Usages

- **Ethernet Network Adapter**
- **FPGA Accelerator**
- **Storage**

**CEM Cards**

**CEM Connectors**

**BGA**

**M.2**

**mSATA SSD**

**m.2 2242 SSD**

22 mm wide

42 mm long

**U.2 2.5in (SFF-8639)**

Many form factors and applications drive the bandwidth scaling of PCIe technology.
Progression of PCIe® Technology Bandwidth

PCIe technology continues to deliver bandwidth doubling for six generations spanning two decades!

<table>
<thead>
<tr>
<th>PCIe Specification</th>
<th>Data Rate (GT/s) (Encoding)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>2.5 (8b/10b)</td>
<td>2003</td>
</tr>
<tr>
<td>2.0</td>
<td>5.0 (8b/10b)</td>
<td>2007</td>
</tr>
<tr>
<td>3.0</td>
<td>8.0 (128b/130b)</td>
<td>2010</td>
</tr>
<tr>
<td>4.0</td>
<td>16.0 (128b/130b)</td>
<td>2017</td>
</tr>
<tr>
<td>5.0</td>
<td>32.0 (128b/130b)</td>
<td>2019</td>
</tr>
<tr>
<td>6.0 (WIP)</td>
<td>64.0 (PAM-4, Flit)</td>
<td>2021*</td>
</tr>
</tbody>
</table>
PCI Express® Architecture Advantages

- Single PHY standard
- Low power and high performance
- Alternate protocol support
- Doubling the bandwidth for sixth generation with full backwards compatibility
- A variety of standard form factors
- A robust and mature compliance and interoperability program

Backwards compatibility enable broad ecosystem and makes PCIe architecture a low-cost I/O for diverse applications
### Key Metrics for PCIe 6.0 Specification: Requirements

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Expectations</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) (PCIe usages cannot 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!
Sensitivity to noise (xtalk, reflection, and device-related) is a key challenge
Error Assumptions and Characteristics with PAM4

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 on the FBER and the error correlation in a Lane and across Lanes**
Handling Errors and Metrics Used for Evaluation

- 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 enable low FEC latency (<2ns) to meet the performance needs of performance critical Load/Store I/O*
Our Approach: Light-Weight FEC and Retry

- Light-weight FEC, strong CRC, and keep the overall latency (including retry) low so that the Load/Store applications do not suffer latency penalty.
- We are better off retrying a packet with $10^{-6}$ (or $10^{-5}$) probability with a retry latency of 100ns vs having a FEC latency impact of 100ns with a much lower retry probability.

Low latency mechanism with FBER of 1E-6 to meet the metrics (latency, area, power, bandwidth)
### Flit Encoding PCIe® 6.0 Architecture: Low-latency with High Efficiency

- **Flit (flow control unit) based**: FEC needs fixed set of bytes.
- **Correction in flit => CRC (detection)** in flits => Retry at flit level.
- **Lower data rates will also use the same flit once enabled**.
- **Flit size**: 256B
  - 236B TLP, 6B DLP, 8B CRC, 6B FEC
  - No Sync hdr, no Framing Token (TLP reformat), no T(DL)LP CRC
  - Improved bandwidth utilization due to overhead amortization.
- **Flit Latency**: 2ns x16, 4ns x8, 8 ns x4, 16 ns x2, 32 ns x1
  - Guaranteed Ack and credit exchange => low Latency, low storage.
- **Optimization**: Retry error flit only with existing Go-Back-N retry.

---

**Low latency improves performance and reduces area**
Electrical Improvements to Achieve Low-latency

- First Bit Error Rate (FBER) < 10^-6
  - Pad-to-pad channel loss < 32 dB at 16 GHz
  - Significant crosstalk and reflection reduction
  - Reference clock and CDR improvement
  - Jitter reduction compared to PCIe® 5.0 technology ~ 2x
  - Improvement in Reference Equalization: Tx 2nd pre-cursor, improved CTLE peaking and bandwidth, and 16-tap DFE

- Minimize Burst Error Probability
  - PAM4 precoding
  - Gray coding
  - Limits on DFE taps
### Pad-to-Pad Loss and System Routing Length

<table>
<thead>
<tr>
<th>Loss Parameters</th>
<th>PCIe® 5.0 Specification Rev 1.0 (dB)</th>
<th>PCIe 6.0 Specification Rev 0.9 (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pad-to-Pad Loss at 16 GHz</td>
<td>-36</td>
<td>-32</td>
</tr>
<tr>
<td>Root Complex (RC)</td>
<td>-9.0</td>
<td>-8.0</td>
</tr>
<tr>
<td>Add-in-Card (AIC)</td>
<td>-9.5</td>
<td>-8.5</td>
</tr>
<tr>
<td>System</td>
<td>-17.5</td>
<td>-15.5</td>
</tr>
</tbody>
</table>

13” system routing requires -32 dB pad-to-pad loss support and PCB loss of 1.0 dB/in
Summary of Key Electrical Changes in PCIe 6.0 Technology

- 64 GT/s PAM4 requires FEC
- Raw BER (pre-FEC and before any DFE burst error): 1e-06
- Pad-to-pad loss: -32 dB at 16 GHz
- Reference Package Models: ~3-6 dB improvement in reflection and xtalk
- Ref CLK Rj RMS: 100 fs (clean), 150 fs in system simulations
- 4-tap Tx equalization with a 2nd Tx pre-cursor – New Preset Table for 64 GT/s
- Tx precoding and gray coding mandatory
- ~2x improvement in silicon jitter
- Reference Rx: Improved CTLE and 16-tap DFE
- Rx eye mask (Top eye: 0.10 UI and 6.0 mV)

A holistic approach to improve Electrical, Logical, and Protocol layers was key to achieve a low-latency PAM4 solution
Assessment of Channel Reach at 64 GT/s

**TX**
- Rev0.9 Jitter Specification
- Fixed best TxEq: Pre2, Pre1, Post1 -> 0.04,-0.2,0
- Tx SNDR: 34 dB
- Tx $R_{TERM}$: 45 Ohm (to account for DC loss)
- Voltage swing: 0.8V (1.0V for NEXT)
- 2 FEXT / 3 NEXT
- Rise/Fall Time: 0.2 UI

**Channel**
- Inductive coil based Rev0.7 Package models
- PCB length: 4” - 14”, AIC length: 4”
- Best available CEM connector
- BB and AIC impedance variation: low, nom, high -> 9 cases
- Directions: NRC to RC

**RX**
- CTLE: Rev 0.9 Spec
- DFE: 16-tap, h1/h0: 0.55, 10-bit Quantization (~1mV)
- Rx $R_{TERM}$: 50 Ohm

**Other**
- BER 1e-6
- Eye Mask: Top Eye 6.0 mV / 0.10 UI to maintain
Eye Height vs. System Routing Length

Top EH Spec Limit = 6 mV

CTLE Index 7: DC Gain = -9 dB
CTLE Index 11: DC Gain = -5 dB

h1/h0 Spec Limit = 0.55

Limits on DFE tap coefficients imposed to minimize DFE error burst probability
Eye Width vs. System Routing Length

Top EW Spec Limit = 10% UI

h1/h0 Spec Limit = 0.55

Compliant channel solutions are feasible for ≤13” system and ≤4” AIC routing
PCIe® Technology Compliance Process

PCI-SIG® Specs

Describe
- Device requirements
  - Base and CEM specs

C&I Test Specs

Define
- Test criteria based on spec requirements
  - Test Definitions
  - Pass/Fail Criteria

Test H/W & S/W

Validate
- Test criteria
  - Compliance
  - Interoperability

Test Tools and Procedures

Clear Test Output Maps
- Directly to Test Spec

Predictable path to design compliance

Test and various form-factor Specs get developed within 1-2 years of Base Spec completion
PCle 6.0 Specification: Key Messages

- PCle 6.0 architecture can meet the needs of storage interconnect solution in the foreseeable future
- 64 GT/s PAM4 - doubles the bandwidth with backwards compatibility
- 64-bit CRC, a light FEC, FBER < 1e-06, and link level retry with low (~ 1e-05) retry probability enable low-latency (< 2 ns for x16) and high reliability (FIT << 1)
- PCle 5.0 architecture-like channel reach is feasible with improvements in circuits and channels
Please take a moment to rate this session.

Your feedback is important to us.