PCIE® 6.0 SPECIFICATION: A HIGH-PERFORMANCE I/O INTERCONNECT FOR ADVANCED NETWORKING APPLICATIONS

Dr. Debendra Das Sharma

Intel Senior Fellow, Intel Corporation
Director, PCI-SIG® Board
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
- Low Power enhancements: L0p
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action
INTERCONNECT: AN IMPORTANT PILLAR OF COMPUTE

Processor interconnects (Load-Store I/O) drive Node/Rack level scaling
a similar system for PC/WS/HH

SMP interconnects enable multi-socket scaling
PCle® technology / CXL enables compute, memory and storage
Attach and dis-aggregation at Rack Level
Typically based on a common PCIe PHY
Networking at Rack level and beyond

Datacenter Interconnect drives performance @ scale
Warehouse scale computing:
compute performance, efficiency

© OpenFabrics Alliance
PCIE® TECHNOLOGY: THE UBIQUITOUS LOAD-STORE I/O INTERCONNECT

- PCIE® technology doubles the data rate with full backwards compatibility every 3 years
- Ubiquitous I/O across the compute continuum: PC, Hand-held, Workstation, Server, Cloud, Enterprise, HPC, Embedded, IoT, Automotive
- One stack / same silicon across all segments with different form-factors; a x16 PCIe 5.0 device interoperates with a x1 PCIe 1.0 device!

<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</td>
<td>64.0 (PAM4, Flit)</td>
<td>2022</td>
</tr>
</tbody>
</table>

(Bandwidth drivers for PCIE technology: Usages driving insatiable demand for compute, memory, storage, and networking bandwidth: PCIE technology is interconnect across these)

PCIE technology continues to deliver bandwidth doubling for six generations spanning 2 decades!
### Key Metrics for PCIe® 6.0 Specification: Requirements

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Rate</td>
<td>64GT/s, PAM4 (double the bandwidth per pin every generation)</td>
</tr>
<tr>
<td>Latency</td>
<td>&lt;10ns adder for Transmitter + Receiver over 32.0 GT/s (including FEC) (We can not afford the 100ns FEC latency as networking does with PAM4)</td>
</tr>
<tr>
<td>Bandwidth Inefficiency</td>
<td>&lt;2 % adder over PCIe 5.0 specification across all payload sizes</td>
</tr>
<tr>
<td>Reliability</td>
<td>0 &lt; FIT &lt;&lt; 1 for a x16 (FIT – Failure in Time, number of failures in 10⁹ hours)</td>
</tr>
<tr>
<td>Channel Reach</td>
<td>Similar to PCIe 5.0 specification under similar set up for Retimer(s) (maximum 2)</td>
</tr>
<tr>
<td>Power Efficiency</td>
<td>Better than PCIe 5.0 specification</td>
</tr>
<tr>
<td>Low Power</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>Plug and Play</td>
<td>Fully backwards compatible with PCIe 1.x specification through PCIe 5.0 specification</td>
</tr>
<tr>
<td>Others</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!
PAM4 SIGNALLING AND ERROR ASSUMPTIONS FOR PCIE® 6.0 TECHNOLOGY

- **PAM4 signaling: Pulse Amplitude Modulation 4-level**
  - 4 levels (2 bits) in same Unit Interval (UI); 3 eyes
  - Helps channel loss (same Nyquist as 32.0 GT/s)

- **Reduced voltage levels (EH) and eye width increases susceptibility to errors**
  - Correlation of errors across Lanes (common source of errors such as power supply noise)
  - Correlation of errors on a Lane due to DFE (burst)
  - FBER: First bit error rate

- **Mitigation:**
  - Gray Coding to reduce errors in each UI
  - Precoding to minimize errors in a burst
  - Forward-error Correct (FEC) + Replay on CRC error

<table>
<thead>
<tr>
<th>Scrambled 2-bit aligned value</th>
<th>Unscrambled 2-bit as well TSO Ordered Sets</th>
<th>Voltage Level</th>
<th>DC-balance Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior to Gray Coding</td>
<td>After Gray Coding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>01</td>
<td>1</td>
</tr>
<tr>
<td>00</td>
<td>00</td>
<td>00</td>
<td>0</td>
</tr>
</tbody>
</table>
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
- Low Power enhancements: L0p
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action
HANDLING ERRORS AND METRICS USED FOR EVALUATION

- **Two mechanisms to correct errors**
  - FEC (Forward Error Correction)
    - Latency and complexity increase exponentially with the number of Symbols corrected
  - Detection of errors by CRC => Link Level Retry (a strength of PCIe® technology)
    - 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

Low latency mechanism w/ FBER of 1E-6 to meet the metrics (latency, area, power, bandwidth)
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
  - Low Power enhancements: L0p
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action
**FLIT MODE: LOW LATENCY AND HIGH LINK EFFICIENCY**

- **Flit** (flow control unit) based: FEC needs fixed set of bytes
- **Error Correction (FEC) 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
  - FEC: 3-way interleaved, single symbol correct – corrects a burst within 3 bytes
  - CRC: RS based – up to 8 byte errors guaranteed detect beyond that 2^64 aliasing
  - Low latency: <2ns FEC correct and CRC check
  - No Sync hdr, no Framing Token (TLP reformat), no TLP/DLLP CRC
  - Improved bandwidth utilization due to overhead amortization
  - Flit Latency: 2ns x16, 4ns x8, 8ns x4, 16ns x2, 32ns 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**
REPLAY IN FLIT MODE

- Once in Flit mode, we are always in Flit mode at all speeds (2.5 GT/s through 64.0 GT/s)
- NOP-Flits do not consume sequence number, not added to Tx Replay buffer, and not replayed
- Replayed Flits start with the Replay Cmd = 00b (w/ Tx sequence number sent) (alternate between 3 Flits with Replay Cmd = 00b followed by x Flits with Ack/ Nak; x = 3 in Normal Flit Exchange, x = 1 in Seq Num handshake)
- Ability to switch from selective Nak to Standard Nak

Example (Normal Flit Exchange Phase):
- B has Ack’d till Flit 10 back to A
- Corrupted Flit 10 but Flit 11 indicates it was NOP – so no replay request
- Corrupted Flit 12; Flit 13 indicates it was payload => B sends 3 consecutive Flits with ‘selective Nak 11’ (another invalid Flit prior to receiving Flit 12 would cause standard Nak)
- A sends Flit 12 with “Explicit Sequence Numbers” in 3 consecutive Flits (12, 15, 15)
- B sees two Flits in error (NOP Flit 15, Payload Flit 16) => B sends “standard Nak 15” in 3 consecutive Flits to A
- A removes Flits 12-15 from its Tx Replay buffer and replays Flits 16, 17, 18 from its Tx Replay buffer with explicit sequence number
KEY PERFORMANCE METRICS WITH PCIE® 6.0 SPECIFICATION:
RETRY PROBABILITY, FIT, BANDWIDTH EFFICIENCY, LATENCY

<table>
<thead>
<tr>
<th>FBER/ Retry Time</th>
<th>10^-5/100ns</th>
<th>10^-5/200ns</th>
<th>10^-5/300ns</th>
<th>10^-5/200ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retry probability per flit</td>
<td>5x10^-6</td>
<td>5x10^-6</td>
<td>5x10^-6</td>
<td>0.048</td>
</tr>
<tr>
<td>B/W loss with go-back-n (%)</td>
<td>0.025</td>
<td>0.05</td>
<td>0.075</td>
<td>4.8</td>
</tr>
<tr>
<td>FIT</td>
<td>4 x 10^-7</td>
<td>4 x 10^-7</td>
<td>4 x 10^-7</td>
<td>4 x 10^-4</td>
</tr>
</tbody>
</table>

Reasonable probability of retry and b/w loss at 10^-6 FBER
FIT is close to 0 due to strong CRC

Bandwidth Scaling with PCIe 6.0 specification at 64.0 GT/s over PCIe 5.0 at 32.0 GT/s w/ 2% DLLP overhead

Better B/W efficiency for small packets (important for networking) in Flit Mode due to CRC amortization and removal of PHY overheads

Bandwidth Scaling with PCIe 6.0 specification at 64.0 GT/s over PCIe 5.0 at 32.0 GT/s w/ 2% DLLP overhead

Better B/W efficiency for small packets (important for networking) in Flit Mode due to CRC amortization and removal of PHY overheads

© OpenFabrics Alliance
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
  - Low Power enhancements: L0p
- Key Metrics and Requirements for PCIe 6.0 Specification – Evaluation
- Conclusions and Call to Action
Existing Methods: L0s, L1, Dynamic Link Width (DLW), Speed Change

- L0s power savings meagre – not supported if Flit mode is negotiated
- L1 offers good power savings – but high entry and exit times
- DLW - width can be modulated with bandwidth demand
  - Cons: High exit latency due to entire Link being in Recovery and Configuration
  - Ex: on a higher bandwidth demand, link retracts for tens of µsecs prior to width increase => traffic stalls for that time
- Speed Change saves bandwidth
  - Cons: Less power savings than DLW and msecs of delay through speed transition

Need a new low-power state (L0p): power consumption proportionate to bandwidth usage, without impacting traffic flow
- **L0p (w/ Flit Mode)** negotiated in Configuration.Complete (TS2)
- **L0p width transition** can be initiated by either side
  - Link Partner either Ack's or Nak's in 1 usec
  - 2 usec without a response: requestor abandons or re-requests
  - On a simultaneous request the one with the higher width wins and "Nak"s its Link Partner's request if L0p.priority is not set
  - If L0p.priority is set, lower width wins

- **Reducing Link width**: (sequence)
  - Control SKP OS on all active Lanes
  - EI0S/ EI on Lanes to be turned off
  - Data Stream on reduced width

- **Increasing Link width**: (sequence)
  - EIEOS/ TS1/ TS2 training on Lanes to be activated while Data Stream continues on active Lanes
  - SDS sequence (>=8G) on Lanes to be activated (Data Stream continues on active Lanes)
  - (Control) SKP OS on all Lanes with new width
  - Data Stream on new width
<table>
<thead>
<tr>
<th>Metrics</th>
<th>Expectations</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Rate</td>
<td>64GT/s, PAM4 (double the bandwidth per pin every generation)</td>
<td>Meets</td>
</tr>
<tr>
<td>Latency</td>
<td>&lt;10ns adder for Transmitter + Receiver over 32.0 GT/s (including FEC) (We can not afford the 100ns FEC latency as n/w does with PAM4)</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 specification 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 10⁹ hours)</td>
<td>Meets</td>
</tr>
<tr>
<td>Channel Reach</td>
<td>Similar to PCIe 5.0 specification 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 specification</td>
<td>Design dependent – expected to meet</td>
</tr>
<tr>
<td>Low Power</td>
<td>Similar entry/ exit latency for L1 low-power state Addition of a new power state (L0p) to support scalable power consumption with bandwidth usage without interrupting traffic</td>
<td>Design dependent – expected to meet; L0p looks promising</td>
</tr>
<tr>
<td>Plug and Play</td>
<td>Fully backwards compatible with PCIe 1.x specification through PCIe 5.0 specification</td>
<td>Meets</td>
</tr>
<tr>
<td>Others</td>
<td>HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform</td>
<td>Expected to Meet</td>
</tr>
</tbody>
</table>

Meets or exceeds requirements on all key metrics
CONCLUSIONS AND CALL TO ACTION

- PCIe® 6.0 specification completed
- We met the challenges on multiple fronts
  - New signaling with PAM4: 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 PAM4 at lower speeds
  - We have exceeded or met the requirements
  - PCIe 6.0 specification is the reflection of the combined innovation capability of 900+ members with a track record of delivering flawlessly against challenges for more than two decades!!
- Plan for products … Expect Networking to be an early adopter
THANK YOU

Dr. Debendra Das Sharma

Intel Senior Fellow, Intel Corporation

Director, PCI-SIG® Board