PCI Express: Delivery Bandwidth for OCP

Al Yanes / President / PCI-SIG
PCI-SIG® Snapshot

Organization that defines the PCI Express® (PCIe®) I/O bus specifications and related form factors.

- 750+ member companies located worldwide

Creating specifications and mechanisms to support compliance and interoperability.

- Australia
- Austria
- Belgium
- Brazil
- Bulgaria
- Canada
- China
- Czech Republic
- Denmark
- Finland
- France
- Germany
- Hong Kong
- Hungary
- India
- Ireland
- Israel
- Italy
- Japan
- Malaysia
- Norway
- Russia
- Singapore
- Slovak Republic
- South Korea
- Sri Lanka
- Sweden
- Switzerland
- Taiwan
- The Netherlands
- Turkey
- United Kingdom
- United States
PCI Express 4.0

PCI-SIG continues its solid reputation of delivering **low cost, high-performance, low-power specifications** for multiple applications and markets.

- **PCI Express 4.0 Specification** – (16GT/S)
  - Finalized and **published October 2017**
  - Includes new performance enhancements
  - Maintains position as the interconnect of choice for the expansive storage market and the backbone for the fast growing cloud ecosystem
PCI Express 4.0

- PCIe 4.0 Key Functional Enhancements
  - Lane Margining at the Receiver - Allows systems to determine how close to “the edge” each lane is operating
  - Expanded Tag and Credits
    - Allows both tags/credits to service devices for future usage
    - 10 bit Extended Tags support up to 1024 transactions
    - Scaled Flow Control supports larger credits
  - PCIe 4.0 Electrical
    - Maintains backward compatibility with installed base of PCIe devices
    - Limited channel reach: approx. 12” one connector (including 4” add-in card)
    - Longer channels require retimers or lower loss channel

- PCIe 4.0 Adoption
  - Numerous vendors have 16GT/s PHYs in silicon
  - Major IP vendors offering 16GT/s controllers
  - Dozen 16GT/s solutions at a recent PCI-SIG Compliance Workshop
  - Several member companies have exhibited 16GT/s demos
Power Efficient Performance

- **Delivers Scalable Performance**
  - Width scaling: x1, x2, x4, x8, x12, x16,
  - Frequency scaling: Five generations
    - 2.5 and 5 GT/s with 8b/10b encoding
    - 8 and 32 GT/s with 128b/130b encoding

- **Low Power (Active/Idle)**
  - Rich set of Link and Device States
    - L0s, L1, L1-substates, L2/L3
    - D0, D1, D2, D3_hot/cold
    - Platform-level power optimization hooks: Dynamic Power Allocation, Optimized Buffer, Flush Fill, Latency Tolerance Reporting
    - Active power – 5pJ/b, Standby power: 10uW/Lane

- **Vibrant ecosystem with IP Providers**
PCI Express 4.0 Channels

- End to end loss target $\approx 28$ dB
- Root Package loss $\approx 5$ dB
- Add-in Card Package loss $\approx 3$ dB
- Total Add-in Card $\approx 8.0$ dB
- Connector $< 1$ dB
Form Factors for PCI Express

- **BGA**
  - 16x20 mm ideal for small and thin platforms

- **M.2**
  - 42, 80, and 110mm lengths, smallest footprint of PCI Express® (PCIe®) connector form factors, use for boot or for max storage density

- **U.2 2.5in (aka SFF-8639)**
  - 2.5in makes up the majority of SSDs sold today because of ease of deployment, hotplug, serviceability, and small form factor
  - Single-Port x4 or Dual-Port x2

- **CEM Add-in-card**
  - Add-in-card (AIC) has maximum system compatibility with existing servers and most reliable compliance program. Higher power envelope, and options for height and length

Source: Intel Corporation
SRIS (Separate Refclk Independent SSC Architecture)

- Challenge: PCIe specification did not support independent clock with SSC
  - SATA* cable ~ $.50
  - PCIe cables include reference clock > $1 for equivalent cable

- PCIe base specification 3.0 ECNs approved
  - Requires use of larger elasticity buffer
  - Requires more frequent insertion of SKIP ordered set
  - Requires receiver changes (CDR)
  - Second ECN updates Model CDRs

- SRIS will create a number of new form factor opportunities for PCIe
  - OCuLink*
  - Lower cost external/internal cabled PCIe
  - Next-generation of PCI-SIG cable specification
RAS Features

- PCIe® architecture supports very high-level set of Reliability, Availability, Serviceability (RAS) features
  - All transactions protected by CRC-32 and Link level Retry, covering even dropped packets
  - Transaction level time-out support (hierarchical)
  - Well defined algorithm for different error scenarios
  - Advanced Error Reporting mechanism
  - Support for degraded link width / lower speed
  - Support for hot-plug
NVM Express™ Driving PCIe SSDs in Data Center

Data Center SSD Units by Interface

Data Center SSD total GB by Interface

Source: Forward Insights Q1’15
PCI Express – 5.0 Specification

- PCI Express 5.0 Specification 32GT/S NRZ:
  - Applications such as artificial intelligence, machine learning, gaming, visual computing, storage and networking
  - High-end networking solutions (i.e. 400Gb Ethernet and dual 200Gb/s InfiniBand solutions)
  - Accelerator and GPU attachments for high-bandwidth solutions
  - Constricted form factor applications that cannot increase width and need higher frequency to achieve performance
  - Continued use of L1 Sub-states to constrain power consumption during transmission idle periods
Ethernet Evolution

Market Adoption of Ethernet Speeds

Ethernet’s “success” in providing cost-effective and reliable solutions, soon expanded into new markets.

Service Provider applications started deploying Ethernet due to customer requests.

Service Provider applications started driving Ethernet’s new higher speed rates.

Controller and Adapter Market Port Forecast

Source: Dell’Oro Research Q4'15
GPUs and PCIe Bandwidth

Open Compute Project’s Olympus Hyperscale GPU Accelerator chassis

- Flexible PCIe topology
- GPGPU-to-Host via high-BW PCIe links
- Peer-to-peer without Host interaction
  - GPGPU peer-to-peer via NVLink
  - GPGPU peer-to-peer to IB NICs via x16 PCIe

Source: Open Compute Project
Timeline for PCIe 5.0 specification by 1H/2019

- Changes limited to primarily speed upgrade
  - Protocol already supports higher speed via extended tags and credits
  - Existing PHYs in the industry already run at 28GHz / 56GHz
- Specification process enhanced to accelerate development
  - PCIe 5.0 version 0.5 released Q4 2017
    - Max pad to pad loss target is expected to be around 35 dB
    - Focus on connector studies
    - Pre-encoding to reduce DFE burst errors
    - SRIS – potentially mandatory
    - EIEOS definition at 32 GT/s change to sequences of 32 ones and zeros
  - PCIe 5.0 version 0.7 target release April 2018
 PCIe 5.0 Delivering 32GT/s

- **Supports 400Gb Ethernet Solutions**
  - 400Gb = 50GB
  - 50GB in both directions

- **Full duplex**
  - 128/130 bit encoding with 1.5% overhead
  - x16 ~64GB/s sufficient to support 400Gb Ethernet solutions (64GB > 50 GB)
  - Total Full Duplex = ~128GB

- CEM connector targeted to be backwards compatible for add-in cards
- **Targeted Release in 1H 2019**

<table>
<thead>
<tr>
<th></th>
<th>RAW BIT RATE</th>
<th>LINK BW</th>
<th>BW/LANE/WAY</th>
<th>TOTAL BW X16</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCIe 1.x</td>
<td>2.5GT/s</td>
<td>2Gb/s</td>
<td>250MB/s</td>
<td>8GB/s</td>
</tr>
<tr>
<td>PCIe 2.x</td>
<td>5.0GT/s</td>
<td>4Gb/s</td>
<td>500MB/s</td>
<td>16GB/s</td>
</tr>
<tr>
<td>PCIe 3.x</td>
<td>8.0GT/s</td>
<td>8Gb/s</td>
<td>~1GB/s</td>
<td>~32GB/s</td>
</tr>
<tr>
<td>PCIe 4.0</td>
<td>16GT/s</td>
<td>16Gb/s</td>
<td>~2GB/s</td>
<td>~64GB/s</td>
</tr>
<tr>
<td>PCIe 5.0</td>
<td>32GT/s</td>
<td>32Gb/s</td>
<td>~4GB/s</td>
<td>~128GB/s</td>
</tr>
</tbody>
</table>
PCI-SIG History

I/O BANDWIDTH DOUBLES
Every 3 Years

Bandwidth (GB/s)

0.13 (PCI)
0.26
0.5
1
2
4
8
16
32 (x16) (PCIe 3.0)
64 (x16) (PCIe 4.0)
128 (x16) (PCIe 5.0)


Time

PCI-SIG BANDWIDTH 1992-2019

Actual Bandwidth (GB/S)  I/O Bandwidth Doubles Every Three Years
Summary

- PCIe technology continues to evolve to exceed industry bandwidth requirements
  - PCIe 5.0 with 32GT/s ideal for artificial intelligence, machine learning, gaming, visual computing, storage and networking
  - Growing need for increased bandwidths in GPUs attachments and accelerators