EVOLUTION OF INTERCONNECTS AND FABRICS TO SUPPORT FUTURE COMPUTE INFRASTRUCTURE

Dr. Debendra Das Sharma
Intel Fellow and Director I/O Technologies and Standards
Intel Corporation
AGENDA

- Mega-Trends in compute landscape
- Interconnects and Fabrics - an important pillar of compute
- Evolution of Interconnects and Fabrics
- Future Directions
• Insatiable demand for compute, storage, and data movement
• Innovative applications leading to more demand which in turn leading to more innovations
• Interconnect is an important pillar of compute
  • Compute, storage/ memory, interconnect, software, process technology, security
Explosion of data leading to rapid innovations. Move faster, Store more, Process everything seamlessly, efficiently, and securely.
DELIVERING PERFORMANCE IN DATA CENTER

Optical Modules
- High-bandwidth connectivity at 100G and beyond

Leaf Switch
- P4-programmable scale-out fabric with uncompromising performance

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

Data Center as a Computer – Interconnects are key to driving warehouse scale efficiency!
CHALLENGES AT SCALE

- Expanding hybrid-clouds
- Congested data flow
- Incompatible security and software
- Dynamic shifts in workload needs
- Stranded resources
THE VISION

Seamless
Edge to cloud experience

Predictable and secure
services anywhere

Optimized TCO
Hardware and Software
## TAXONOMY, CHARACTERISTICS, AND TRENDS OF INTERCONNECTS

<table>
<thead>
<tr>
<th>Category</th>
<th>Type and Scale</th>
<th>Current Data Rate/ Trend</th>
<th>PHY Latency (Tx + Rx)</th>
<th>Other Characteristics</th>
</tr>
</thead>
</table>
| Latency Tolerant  | Networking Data Center Scale             | 56/ 112 GT/s-> 224 GT/s (PAM4)         | 100+ ns w/ FEC (20ns+ w/o FEC)         | • Narrow Lane count (4 or 8)  
• Backplane usage w/ cables & retimers                                                  |
| Latency Sensitive | Load-Store I/O (PCIe/ CXL / SMP cache coherency)  
Node level (moving to sub-Rack level)   | 32 GT/s (NRZ) -> PCIe Gen6 64 GT/s (PAM4) | <10ns (Tx+ Rx: PHY-PIPE)  
0-1ns FEC overhead                                                               | • 200-300 Lanes per CPU socket  
• Low-cost and HVM  
• Socket & mother board (12” reach)  
• Backwards compatibility (PCIe/ CXL) – single standard for all usages  
• Area/ power sensitive  
• Reliability (FIT << 1; Failure in Time – number of failures in 10⁹ hours)  
• SMP coherency and memory access uber-latency sensitive  |

Latency Sensitive I/O moving to PAM-4: innovation needed to meet latency, area, and cost challenges for viability
- Ability to directly access memory (CPU, I/O)
- Tightly coupled – small scale – Fabric through PCIe
- 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
  - Hardware based link level replay on error
- Timeout and Error reporting hierarchy
  - Hardware based error containment guarantees

Load-store I/O evolving to meet memory innovations and expanding to Rack level for resource pooling
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
- Drivers: Networking, XPU, Memory, Alternate Protocol – need to keep w/ compute cadence

Six generations of evolution spanning 2 decades! Need to keep KPIs in-tact!

<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>

© OpenFabrics Alliance
<table>
<thead>
<tr>
<th>Metrics</th>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Rate</td>
<td>64 GT/s, PAM4 (double the bandwidth per pin every generation)</td>
</tr>
<tr>
<td>Latency</td>
<td>&lt;10ns adder for Transmitter + Receiver (including Forward Error Correct, FEC) for PCIe CXL / Memory/ SMP coherency interconnects need less than 1ns adder (Ld/St can not afford the 100ns FEC latency as networking does – okay at DC-level scale)</td>
</tr>
<tr>
<td>Bandwidth Inefficiency</td>
<td>&lt;2 % adder over PCIe 5.0 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 $10^9$ hours)</td>
</tr>
<tr>
<td>Channel Reach</td>
<td>Similar to PCIe 5.0 under similar set up for Retimer(s) (maximum 2)</td>
</tr>
<tr>
<td>Power Efficiency</td>
<td>Better than PCIe 5.0</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 through PCIe 5.0</td>
</tr>
<tr>
<td>Others</td>
<td>HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform</td>
</tr>
</tbody>
</table>

64.0 GT/s PAM-4 is a major inflection point for Load-store I/O - PCIe 6.0 is on track to meet each of these metrics!
PCIE 6.0 - A LOW-LATENCY APPROACH

- Light-weight FEC + Low-latency Link level replay
- A combination of $10^{-6}$ FBER (First Burst Error Rate) with a 3-way interleaved single symbol correct FEC keeps retry rate low
- Spec defined mechanisms for low-latency replay
- Strong CRC (Cyclic Redundancy Check) for low FIT
- Flit (Flow-Control Unit) mode results in better link efficiency than before!

Low-latency, low-power, backward-compatible doubling of bandwidth journey continues!
CXL: A NEW CLASS OF INTERCONNECT

- Heterogenous computing and disaggregation
- Efficient resource sharing
- Shared memory – efficient access
- Enhanced movement of operands and results
- Memory bandwidth and capacity expansion
  - Memory tiering and different memory types
  - In-memory processing

CXL Enabled Environment

© OpenFabrics Alliance
**CXL APPROACH**

**Coherent Interface**
Leverages PCIe with 3 mix-and-match protocols

**Low Latency**
.Cache and .Memory targeted at near CPU cache coherent latency (<200ns load to use)

**Asymmetric Complexity**
Eases burdens of cache coherent interface designs
CXL 1.1 USAGE MODELS

Caching Devices / Accelerators

- Processor
- DDR

CXL

- PROTOCOLS: CXL.io, CXL.cache

- USAGES: Smart NIC, NIC atomics, PGAS, RDMA

Accelerators with Memory

- Processor
- DDR

CXL

- PROTOCOLS: CXL.io, CXL.cache, CXL.memory

- USAGES: GP GPU, Dense computation

Memory Buffers

- Processor
- DDR

CXL

- PROTOCOLS: CXL.io, CXL.memory

- USAGES: Memory BW expansion, Memory capacity expansion, Storage class memory

© OpenFabrics Alliance
Switching for fan-out and pooling
Managed Hot-plug flows to move resources
Persistence flows for supporting persistent memory
  - Covers larger latency memory in addition to DRAM
Type-1 and Type-2 device (accelerator) assigned to one host
Type-3 device (memory) can be pooled across multiple hosts
Fabric Manager for managing resources
Software API for devices
Enhanced security: DTLB, device authentication, link encryption
  - Working with DMTF, PCI-SIG for synergies
  - Spans devices and switch

Dis-aggregated System with CXL optimizes resource utilization delivering lower TCO and power-efficiency
FUTURE DIRECTIONS

- **Composable Disaggregated Infrastructure at Rack level**
  - Heterogenous compute/ memory, storage, networking fabric resources
  - connected through high bandwidth, low-latency Load-Store Interconnect
  - delivering almost-identical performance per watt as independent servers
  - w/ multiple domains w/ shared memory, message passing, atomics

- **Synergy between Networking and Load-store**
  - Expect boundaries to be fungible
  - Fabric Manager, Multi-head, multi-domain, Atomics support, Persistence flows, Smart NIC with optimized flows to access system memory without involving host, VM migration

- **Challenges:**
  - Latency: NUMA optimization, low-latency switch
  - Bandwidth demand: higher rate helps
  - Power Efficiency
  - Blast Radius – containment and QoS
  - Scaling: Moore’s law and Dennard-scaling
  - Copper-Optical transition point
  - Software!
THANK YOU

Dr. Debendra Das Sharma
Intel Fellow and Director I/O Technologies and Standards
Intel Corporation