OpenCAPI Memory Interface

A data-centric approach to server design
Industry Landscape

- Key changes occurring in our industry
  - Historical microprocessor technology continues to deliver far less than the historical rate of cost/performance improvement per generation – Running out of steam
  - New advanced memory technologies changing the economics of computing

- Companies realizing need for **accelerated computing** with a coherent **high performance bus** to meet today’s computational demand
Evolution of System Architecture

Yesterday's Plumbing
Tomorrow's Differentiation

OpenCAPI Northbound

OpenCAPI & PCI

Yesterday's Plumbing
Tomorrow's Differentiation

CPU/ Accelerator Bandwidth

System bottleneck

© 2018 IBM Corporation
Accelerated Computing and High Performance Bus

- Attributes driving Accelerators
  - Emergence of complex storage and memory solutions
  - Introduction of device coherency requirements (IBM’s introduction in 2013)
  - Growing demand for network performance and network offload
  - Various form factors (e.g., GPUs, FPGAs, ASICs, etc.)
    - All are relevant in the modern data center
    - No single FF can address everything

- Driving factors for a high performance bus - Consider the environment
  - Increased industry dependence on hardware acceleration for performance
  - Hyperscale datacenters and HPC are driving need for much higher network bandwidth
    - 100 Gb/s -> 200 Gb/s -> 400Gb/s are emerging
  - Deep learning and HPC require more bandwidth between accelerators and memory
  - New memory/storage technologies are increasing the need for bandwidth with low latency
Two Bus Challenges

1. Coherent high performance bus needed
   - Hardware acceleration will become commonplace
   - But, if you are going to use Accelerators, you need to get data in/out very quickly
     - Today’s system interfaces are insufficient to address this requirement
     - Traditional I/O architecture results in very high CPU overhead when applications communicate with I/O or Accelerator devices
   - Systems must be able to integrate multiple memory technologies with different access methods, coherency and performance attributes

2. These challenges must be addressed in an open architecture allowing full industry participation
   - Need to be architecture agnostic to enable the ecosystem growth and adaption
   - Establish sufficient volume base to drive cost down
   - Support broad ecosystem of software and attached devices
OpenCAPI Approach

• Two bus challenges
  1. Coherent high performance bus
     → Needed to define a new technology
  2. A need to make this ‘open’
     → Needed to establish open community

• Approach taken to define this new technology
  • Bottom’s up design approach for advanced capabilities and performance
    • Clean sheet of paper with no incumbent overhead
    • Architecture and electrical view point

• OpenCAPI bus architecture was Born
Comparison of Memory Paradigms

**Main Memory**
- Processor Chip
- DLx/TLx
- DDR4/5

**Example: Basic DDR attach**
- DDR4/5
- DLx/TLx
- Data

**Emerging Storage Class Memory**
- Processor Chip
- DLx/TLx
- Data
- SCM

**Storage Class Memories have the potential to be the next disruptive technology..... Examples include ReRAM, MRAM, Z-NAND...... All are racing to become the defacto**

**Tiered Memory**
- Processor Chip
- DLx/TLx
- Data
- DDR4/5
- DLx/TLx
- Data

**Storage Class Memory tiered with traditional DDR Memory all built upon OpenCAPI 3.1 & 3.0 architecture.**
**Still have the ability to use Load/Store Semantics**

**OpenCAPI 3.1 Architecture**
- Ultra Low Latency ASIC buffer chip adding +5ns on top of native DDR direct connect!!

- OpenCAPI 3.1 & 3.0 architecture.
OpenCAPI Memory

• Signaling ➔ AXON @25.6GHz vs DDR4 @ 3200 MHz
  – 4x bw per signal IO
• Idle latency over traditional DDR
  – POWER8/9 Centaur design ~10 ns
  – OpenCAPI target of ~5 ns

• Centaur ➔ One proprietary design
• OpenCAPI ➔ Open
POWER9 – Dual Memory Subsystems

Scale Out
Direct Attach Memory

- 8 Direct DDR4 Ports
  - Up to 140 GB/s of sustained bandwidth
  - Low latency access
  - Commodity packaging form factor
  - Adaptive 64B / 128B reads

Scale Up
Buffered Memory

- 8 Buffered Channels
  - Up to 230GB/s of sustained bandwidth
  - Extreme capacity – up to 8TB / socket
  - Superior RAS with chip kill and lane sparing
  - Compatible with POWER8 system memory
  - Agnostic interface for alternate memory innovations
Scale Out
Direct Attach Memory
2 Socket SMP

Scale Up
Buffered Memory
16 Socket SMP
## Comparison of IBM CAPI Implementations

<table>
<thead>
<tr>
<th>Feature</th>
<th>CAPI 1.0</th>
<th>CAPI 2.0</th>
<th>OpenCAPI 3.0</th>
<th>OpenCAPI 3.1</th>
<th>OpenCAPI 4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processor Generation</strong></td>
<td>POWER8</td>
<td>POWER9</td>
<td>POWER9</td>
<td>Power9 Follow-On</td>
<td>Power9 Follow-On</td>
</tr>
<tr>
<td><strong>CAPI Logic Placement</strong></td>
<td>FPGA/ASIC</td>
<td>FPGA/ASIC</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>CAPI Logic Placement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interface</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lanes per Instance</td>
<td>PCIe Gen3 x8/x16 8 Gb/s</td>
<td>PCIe Gen4 2 x (Dual x8) 16 Gb/s</td>
<td>Direct 25G x8 25 Gb/s</td>
<td>Direct 25G x8 25 Gb/s</td>
<td>Direct 25G x8 25 Gb/s</td>
</tr>
<tr>
<td>Lane bit rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Address Translation on CPU</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Native DMA from Endpoint Accelerator</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Home Agent Memory on OpenCAPI Endpoint with Load/Store Access</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>NA</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Native Atomic Ops to Host Processor Memory from Accelerator</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Host Memory Caching Function on Accelerator</strong></td>
<td>Real Address Cache in PSL</td>
<td>Real Address Cache in PSL</td>
<td>No</td>
<td>NA</td>
<td><strong>Effective Address Cache in Accelerator</strong></td>
</tr>
</tbody>
</table>

*Remove PCIe layers to reduce latency significantly*
**Centaur buffer**
4 channel DDR3/4 51 GB/sec peak  
DMI interface 20 GB/sec rd, 10 GB/sec w

**OpenCAPI thin buffer**
DDR4 @3.2 25 GB/sec peak  
OMI interface 22 GB/sec rd + 22 GB/sec wr
## Proposed POWER Processor Technology and I/O Roadmap

<table>
<thead>
<tr>
<th>Year</th>
<th>Architecture</th>
<th>Cores</th>
<th>Process Technology</th>
<th>Micro-Architecture</th>
<th>Process Technology Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>POWER7</td>
<td>8</td>
<td>45nm</td>
<td>New Micro-Architecture</td>
<td>New Process Technology</td>
</tr>
<tr>
<td>2012</td>
<td>POWER7+</td>
<td>8</td>
<td>32nm</td>
<td>Enhanced Micro-Architecture</td>
<td>New Process Technology</td>
</tr>
<tr>
<td>2014</td>
<td>POWER8</td>
<td>12</td>
<td>22nm</td>
<td>New Micro-Architecture</td>
<td>New Process Technology</td>
</tr>
<tr>
<td>2016</td>
<td>POWER8 w/ NVLink</td>
<td>12/24</td>
<td>14nm</td>
<td>Enhanced Micro-Architecture</td>
<td>With NVLink</td>
</tr>
<tr>
<td>2017</td>
<td>P9 SO</td>
<td>12/24</td>
<td>14nm</td>
<td>New Micro-Architecture</td>
<td>Direct attach memory</td>
</tr>
<tr>
<td>2018</td>
<td>P9 SU</td>
<td>12/24</td>
<td>14nm</td>
<td>Enhanced Micro-Architecture</td>
<td>Buffered Memory</td>
</tr>
<tr>
<td>2019</td>
<td>P9 w/ Adv. I/O</td>
<td>12/24</td>
<td>14nm</td>
<td>Enhanced Micro-Architecture</td>
<td>New Memory Subsystem</td>
</tr>
</tbody>
</table>

### Sustained Memory Bandwidth

- **POWER7**
  - Up To 65 GB/s
- **POWER8**
  - Up To 210 GB/s
- **POWER9**
  - Up To 150 GB/s
- **POWER10**
  - Up To 435 GB/s

### Standard I/O Interconnect

- **POWER7**
  - PCIe Gen2
- **POWER8**
  - PCIe Gen3
- **POWER9**
  - PCIe Gen4 x48
- **POWER10**
  - PCIe Gen5

### Advanced I/O Signaling

- **POWER7**
  - N/A
- **POWER8**
  - N/A
- **POWER9**
  - N/A
- **POWER10**
  - 32 & 50 GT/s

### Advanced I/O Architecture

- **POWER7**
  - N/A
- **POWER8**
  - CAPI 1.0
- **POWER9**
  - CAPI 2.0, OpenCAPI3.0, NVLink2.0
- **POWER10**
  - CAPI 2.0, OpenCAPI4.0, NVLink

---

Statement of Direction, Subject to Change
JOIN TODAY!

www.opencapi.org