Facilitating IP Development for the OpenCAPI Memory Interface

Kevin McIlvain, Memory Development Engineer
IBM

OpenPOWER Summit Europe
RAI Centre | Amsterdam
October 3-4, 2018

Join the Conversation #OpenPOWERSummit
Moral of the Story

OpenPOWER is the best platform to exploit emerging memory technologies!

• Forward thinking architecture
  • Accepting technology scaling challenges and is architected for them!
    • Not just for GPUs and other accelerators…. for memory too!
    • OpenPOWER recognizes memory is at the forefront of transforming system architecture!

• Memory agnostic interfaces
  • OpenCAPI 3.0, 3.1 and 4.0
  • Not tied to a specific media like DDR4
  • Easy to manage technology transitions

• Firmware support for new memory technologies
  • Persistent memory

• Open ecosystems for fast industry research, development and adoption!

We are at the precipice of big changes in the memory subsystem...

OpenPOWER is ready to lead the way!
Evolving Memory Challenges

Von Neumann’s Bottleneck
- Throughput limited by memory bandwidth
  - Processor technology has evolved to help alleviate throughput
    - Caching
    - Prefetch
    - Multithreading

Moore’s law
- Shrinking devices and increasing speed
  - Moore’s Law is Slowing, compute architecture is compensating
    - Increased core count
    - Augment processors with GPUs, FPGAs, specialized accelerators connected with advanced IO
  - Requires more memory bandwidth and density

Not just processors, memory too!
- DRAM scaling difficulties ahead
  - DDR5 will require on chip ECC
  - DDR5 interface will employ DFE on I/O
  - Power and Density Hitting Scaling Limitations
  - Solutions add cost and latency
Evolving Memory Challenges

### Power Processor Trend
- More cores and more threads equals higher performance per socket

### Memory Capacity
- Capacity per socket needs to grow to keep roughly the same amount of memory per core or per thread

### DRAM Die Density
- Flattening out due to scaling issue

Credit: K.H. Kim IBM Research
Evolving Memory Challenges

Challenges

- AI applications require tremendous amounts of data
  - Driving massive increase in memory capacity
- DRAM technology is beginning to face scaling challenges
  - Still require low DRAM latencies
  - Still need to grow capacity and bandwidth
  - DRAM costs are becoming a major inhibitor
- Augment DRAM with emerging memory
  - Still require tiers of traditional DRAM for low latency
  - Need to significantly reduce the cost of high capacity memory
- Heterogeneous memory architectures are the next logical step
  - What is the best path forward?
Evolving Memory Challenges

**Where can we add the most value?**

- Mitigate DRAM scaling challenges
  - Hybrid DIMMs like NVDIMM – N, F, P (H)
  - More SCM closer to the CPU
  - Need = More bits, more bandwidth!

**Challenges**

- DDR4 / DDR5 Interfaces
  - Protocols are specific and not flexible
  - Difficult to manage technology transitions
- NVDIMM – P
  - Specification estimated end of 2018 early 2019
  - Built on top of DDR4 / DDR5
  - What chipsets will support NVDIMM?
    - Competition from 3D-Xpoint
      - Proprietary extension on DDR4 for single SCM technology
      - Restricts innovation and $/GB competition
      - Compromises in BW and capacity, DRAM tier Vs. SCM tier
Evolving Memory Challenges

Value add – OpenCAPI Memory Interface

- Open memory agnostic interface standard
- Embraces the coming diverse memory technologies
- Encourages innovation
  - No dependence on CPU vendors to provide a proprietary route to market
  - Not limited to OpenPOWER CPUs (other CPUs or devices)
- More bandwidth, more bits
  - Fewer pins than JEDEC DDR4 / DDR5
    - ~34 vs. ~150 for DDR4 approx. 4.4x
    - 8 Channels of DDR4 on P9 ~1200pins, OMI is 2x Bandwidth < ½ the pins
  - DDR4 and OMI have equal max read only bandwidth
  - DDR4 and OMI have equal max write only bandwidth
  - OMI can support max read and write bandwidth at the same time
  - Feasible to support 16 OMI channels per processor
    - Perfect for emerging/heterogeneous technologies and near memory acceleration
    - Optimize memory tier bandwidth and capacity for lower TCO
- Industry adoption (drives down costs)
  - JEDEC DDR5 DDIMM – OpenCAPI interfaced buffered DIMM
Evolving Memory Challenges

Moore’s Law Slowing – Von Neumann’s Bottleneck – Changing Architectures

• “With the flexibility of the attach on the memory side and on the compute acceleration side, it really boils down to thinking of the CPU chip as this big switch,” ..... ”this big data switch that’s just one big pile of bandwidth connectivity that’s enabling any kind of memory to talk to any kind of acceleration, and it all plumbs right past the powerful general-purpose processor cores, so you’re pulling that whole compute estate together.” – Jeff Stuecheli
## OpenPOWER Memory Interfaces

### Power 8 (2014 - 2016)
- **All Systems**
  - **DMI – Differential Memory Interface**
    - Fewer pins per channel
    - Agnostic protocol
    - Proprietary
    - Bandwidth 9.6 GT/s IO = 210 GB/s
    - Buffered CDIMM DDR3/4
    - RAS
    - Memory Density
    - Custom DIMM or Buffered Riser with industry standard DIMMs

### Power 9 (2017 - 2018)
- **Scale – Up**
  - **DMI – Differential Memory Interface**
    - Fewer pins per channel
    - Agnostic protocol
    - Proprietary
    - Bandwidth 9.6 GT/s IO = 210 GB/s
    - Buffered CDIMM DDR4
    - RAS
    - Memory Density
  - **Scale – Out**
    - **Direct Attach up to DDR4 2666**
      - Bandwidth 150 GB/s
      - Lower cost

### Power 9 w/Adv. IO (2019)
- **All Systems**
  - **OMI – OpenCAPI Memory Interface**
    - OpenCAPI 3.1
    - Low Latency Optimization
    - Fewer pins per channel
    - Agnostic protocol
    - Open protocol
    - Product differentiation
    - Bandwidth 25 GT/s IO = 350+ GB/s
    - Low cost JEDEC standard DDIMMs

### Power 10 (2020+)
- **All Systems**
  - **OMI – OpenCAPI Memory Interface**
    - OpenCAPI 3.1
    - Low Latency Optimization
    - Fewer pins per channel
    - Agnostic protocol
    - Open protocol
    - Product differentiation
    - Bandwidth 32–50 GT/s IO = 435+ GB/s
    - Low cost JEDEC standard DDIMMs

## OpenPOWER Memory Interfaces

### Best of both worlds
- High bandwidth low pin count interface
- Production differentiation
- Open agnostic protocol
  - rapid development of new memory technologies
- Develop products for OMI with few hurdles
- Standardized memory interface and memory modules
  - JEDEC DDR5 DDIMMs

### Direct attach –
- Industry standard interface
- Specific protocols not very flexible.
OpenPOWER Memory Strategy

Create an open, agnostic, high bandwidth, low latency memory interfaces

- OpenCAPI 3.1 Memory Interface – OMI
- OpenCAPI 3.0 LPC Mode

Support a road map of standardized DDIMMs (Differential DIMMs)

- 1U DDR5 DDIMM currently being standardized through JEDEC
- DDR5 Differential Buffer also being standardized through JEDEC
- Qualification of OpenCAPI DDIMMs supported through OpenPOWER
  - OpenCAPI DDIMM Reference Tester Design
  - OpenCAPI DDIMM Characterization and Qualification Code
  - Verification IP to drive an OpenPOWER Memory Buffer Chip

Support emerging memory technology through the OpenPOWER Ecosystem

- OpenCAPI 3.1 / 3.0 Memory Interface Reference Designs (Initially targeted for Xilinx FPGAs)
  - Allows user to create any type of OpenCAPI 3.1 / 3.0 compliant memory functional unit
    - Hybrid memory controller / Near memory acceleration
    - FPGA based OpenCAPI 3.1 / 3.0 interface design targeting Xilinx FPGA
  - Open Agnostic interface allows much faster time to market
    - No need for new standards. Example NVDIMM-N,F,P
## OpenPOWER Memory Interfaces

### Power 9 w/ adv. IO

<table>
<thead>
<tr>
<th>OMI Bandwidth Considerations (Note 1)</th>
<th>DDIMM DDR4 3200 (Near term)</th>
<th>DDIMM DDR5 3200 (Not far off)</th>
<th>DDIMM DDR5 6400 (Further down the road)</th>
<th>FPGA DDIMM DDR4 2400 (Near term)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DDR Channels</strong></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Channel Width</strong></td>
<td>64</td>
<td>32</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td><strong>Max OC x8 Rd BW</strong></td>
<td>8x25.6Gbps = 204.8Gbps</td>
<td>8x25.6Gbps = 204.8Gbps</td>
<td>8x25.6Gbps = 204.8Gbps</td>
<td>8x25.6Gbps = 204.8Gbps</td>
</tr>
<tr>
<td><strong>Max OC x8 Wr BW</strong></td>
<td>8x25.6Gbps = 204.8Gbps</td>
<td>8x25.6Gbps = 204.8Gbps</td>
<td>8x25.6Gbps = 204.8Gbps</td>
<td>8x25.6Gbps = 204.8Gbps</td>
</tr>
<tr>
<td><strong>Max DDR Rd BW</strong></td>
<td>1x64X3.2GTs = 204.8Gbps</td>
<td>2x32X3.2GTs = 204.8Gbps</td>
<td>2x32X6.4GTs = 409.6Gbps</td>
<td>2x64x2.4GTs = 307.2Gbps</td>
</tr>
<tr>
<td></td>
<td>OC Rd = 100%</td>
<td>OC Rd = 100%</td>
<td>OC Rd = 100% not enough</td>
<td>OC Rd = 100% not enough</td>
</tr>
<tr>
<td></td>
<td>OC Wr = 0%</td>
<td>OC Wr = 0%</td>
<td>OC Wr = 0%</td>
<td>OC Wr = 0%</td>
</tr>
<tr>
<td><strong>Max DDR Wr BW</strong></td>
<td>1x64X3.2GTs = 204.8Gbps</td>
<td>2x32X3.2GTs = 204.8Gbps</td>
<td>2x32X6.4GTs = 409.6Gbps</td>
<td>2x64x2.4GTs = 307.2Gbps</td>
</tr>
<tr>
<td></td>
<td>OC Rd = 0%</td>
<td>OC Rd = 0%</td>
<td>OC Rd = 0%</td>
<td>OC Rd = 0%</td>
</tr>
<tr>
<td></td>
<td>OC Wr = 100%</td>
<td>OC Wr = 100%</td>
<td>OC Wr = 100% not enough</td>
<td>OC Wr = 100% not enough</td>
</tr>
<tr>
<td><strong>Max Simultaneous DDR Rd &amp; Wr BW</strong></td>
<td>NA, Single channel</td>
<td>Rd 102.4Gbps = 50%</td>
<td>Rd 204.8Gbps = 100%</td>
<td>Rd 153.6Gbps = 75%</td>
</tr>
<tr>
<td></td>
<td>Wr 102.4Gbps = 50%</td>
<td>Wr 204.8Gbps = 100%</td>
<td>Wr 204.8Gbps = 100%</td>
<td>Wr 153.6Gbps = 75%</td>
</tr>
</tbody>
</table>

### Take away:
- There is room to exploit the OMI interface and there is ample opportunity for near memory compute

### Note 1:
- Calculations are theoretical maximums for sustained bandwidth for discussion only (does not include protocol and controller overhead for DDR4/5 or OMI)
OpenPOWER Memory Workgroup

What do we need to facilitate IP ecosystem development for OpenPOWER?

Immediate workgroup focus

• Hardware support
  • Reference designs and specifications for OpenCAPI 3.0 and 3.1
  • Verification IP
  • OpenPOWER Partners Provide OpenPOWER Ready Hardware
    • E.g. FPGA based memory DDIMMs for research and development
    • Capable of deploying the reference designs

• Software support
  • Characterization and qualification code
    • Already available for P8 and P9
  • We are at the precipice of persistence and acceleration in memory
    • FW for persistence available in the near future
    • What about acceleration?
Hardware support

- AXI 4 compliant Memory Service Layer (In process)
  - Flexible attach for near memory acceleration and new memory technology
  - Reference designs and specification developed through the workgroup
  - Will support for OpenCAPI 3.0 and 3.1
OpenPOWER Memory Workgroup

Hardware support

- Verification IP (In process)
  - Capability to bring-up and exercise the OpenCAPI Memory Interface without complex processor IP
    - Same IP used to drive the OpenCAPI memory buffer in emulation environments
    - Deployable within an FPGA
  - Reference designs and specification to be documented and provided through the workgroup
OpenPOWER Memory Workgroup

Hardware support

- Hardware tester example design
  - FPGA is host to the OMI Channels
  - FPGA can be host or carrier on the Slim SAS
Software support

- Characterization and qualification code
  - Already available for P8 and P9
  - Near future - test, characterization and qualification code to support the first OMI DDIMMs

- System firmware that is ready for persistent memory
  - Coming in the near future
    - Solves persistency concerns in the software stack

- Bring more discussion into the workgroup
  - We have been hardware focused
  - How do the applications and O/S best utilize these new memory features?
Overview

• Description
  • Forum for Workgroup members to propose, discuss, create and approve modification on OpenPOWER Memory Bus which can be operated as connected with the OpenPOWER Processor Memory Channel which provides a common intermediate interface for multiple types of Memory Controllers on the other side such as DRAM Controller, Flash Controller and other Memory Technology Controllers.

• Scope
  • The OpenPOWER Memory workgroup will be a persistent, standing group that defines, documents, manages and maintains the interface requirements (protocol, timing, etc.) between the OpenPOWER Processor Memory Channel which provides a common intermediate interface to Memory Controllers that attach to Physical Memory. It will also identify which type of memory technology and interface will be compatible, suitable with OpenPOWER Memory Bus.

• Accomplishments
  • OpenPOWER Memory Bus Spec v1.0 for DMI