Welcome

OCP ODSA Project Workshop
September 12, 2019
IBM has a long history in supporting open initiatives...

- **2000-2010**
  - IBM leads the open source revolution
  - IBM contributes code to Kubernetes Open Source Project
  - IBM contributes code to Docker Open Source Project
  - Open Stack is launched by IBM & others

- **2016**
  - OpenCAPI Consortium formed. Open memory & I/O interfaces

- **2019**
  - IBM helps establish the Eclipse foundation
  - IBM & Red Hat deliver enterprise Linux Solutions & are amongst the top companies contributing to the Linux kernel
  - OpenPOWER foundation formed as an independent organization to foster innovation around POWER
  - IBM acquires Red Hat

**OPEN CHIP DESIGN**
- Open POWER ISA
- Open Reference Designs
- Open Governance
IBM expands open hardware ecosystem with major contributions to community

OpenPOWER Summit NA August 20, 2019

Open POWER ISA:
Opening POWER Instruction Set Architecture (ISA), inclusive of patent rights.

Open Reference Designs:
Open sourcing a softcore implementation of the POWER ISA as well as reference designs for the architecture-agnostic Open Coherent Accelerator Processor Interface (OpenCAPI) and Open Memory Interface (OMI).

Open Governance:
OpenPOWER Foundation joining the Linux Foundation
Where can I find?

OpenPOWER Foundation
https://openpowerfoundation.org/

OpenPOWER Summit EU 2019 – October 31-November 1
Register Today!
https://events.linuxfoundation.org/events/openpower-summit-eu-2019/

OpenCAPI Consortium
https://opencapi.org/

POWER ISA Softcore
https://github.com/antonblanchard/microwatt

OpenCAPI and OMI Reference Designs – Going live today!
https://github.com/OpenCAPI
OpenCAPI3.0_Device_RefDesign
omi_host_fire
omi_device_ice
Exploiting Composable Heterogeneity through Open Architectures

Jeff Stuecheli
Josh Friedrich
Chiplet Design Opportunities

**Chip disaggregation**
- Cost reduction
- Modularity
- Optimized technology use
- Latency reduction

**System integration**
- More efficient connectivity
- Overcome reticle limit
- Increased system density
- Heterogenous integration
Chiplet Benefits: Cost
Chiplet Benefits: SoC Modularity

POWER9 Processor Family

- Any revision requires new processor tapeout, test pattern generation, and qualification.
- Integration of new IP can create disruptive changes to chip infrastructure.
- Moving any IP to next technology requires full chip to be re-designed for next node.
Chiplet Benefits: Potential Package-Level Modularity

- IP providers deliver tested physical chiplet vs. design IP needing integration
  - Avoids time-consuming & expensive re-integration into SoC
- Durable IP blocks avoid change
Chiplet Benefits: Technology Use Optimization

Cores & Accelerators

- Most Advanced Logic Process
  - Silicon-bound area
  - Base technology elements (logic devices, SRAMs)
  - Power/performance sensitive
  - Functionally resilient to model inaccuracy

PHYs & Analog IP

- Mature Node
  - Pin-bound area
  - Rich technology menu (passives, multi-oxide, etc)
  - Low power density
  - Functionally sensitive to model-to-hardware
Module wiring offers >10x advantage in time of flight over best on-chip transport.
Chiplet Design Opportunities

Chip disaggregation
- Cost reduction
- Modularity
- Optimized technology use
- Latency reduction

System integration
- More efficient connectivity
- Overcome reticle limit
- Increased system density
- Heterogenous integration
Chiplet Benefits: More Efficient Connectivity

- Board level integration requires robust SERDES to handle complex channels, minimize wire counts, etc.

- Chiplet provides opportunity to significantly improve connectivity

- Need a range of options to optimize latency, power, and chip resources
SoC integration is limited by transistors in the reticle. As density improvement has slowed, die size has grown to compensate, but this is approaching its limit. Integration leveraging advanced packaging technologies and chiplet design provides an anecdote.
Chiplet Benefits: Increased Density

Simple opportunity created by exceeding the reticle is to increase compute density, but this approach has limits.
Chiplet Benefits: Heterogenous Integration

Leverage heterogenous chiplets to deliver value
- Right compute for all facets of a job
- Diverse technologies to maintain system balance
- Extreme connectivity through advanced packaging
- Simpler demands on IP chiplet providers vs. delivering full SoC
Chiplet Challenges & Requirements

- Broad range of packaging innovation
- Open PHY & protocol standards
- Cooling & current delivery
- Robust test infrastructure
- Standards + tooling
- Business ecosystem
Chiplet Requirements: Packaging & Silicon Aggregation Innovations

1) **Silicon Area per Module**: Depends on laminate capability, yield, manufacturing economics
2) **End-to-end Intra-connect Latency**: Depends on total silicon radius, internal/external bandwidth
3) **External Bandwidth Escape**: Depends on internal/external Intra-connect split, system interface rqmts
4) **Internal Bandwidth Density**: Depends on internal/external Intra-connect split, system interface rqmts
5) **Granularity of Composability**: Depends on chipset flexibility and composability rqmts

Technology optimization varies depending on silicon, latency, and bandwidth needs:
Chiplet Requirements: Cooling & Current Delivery

- Power/socket has grown significantly over the last decade.

- Power/socket growth will continue to accelerate.
  - Slowing silicon scaling
  - Vmin limitations
  - Dense compute acceleration

- Chiplet design adds significant complexities.
  - Greater than reticle integration
  - Large instantaneous currents from IP activity changes
  - Large # of voltage supplies to support heterogenous IP

- Continued investment in solutions is necessary.
  - Cost-effective cooling solutions
  - Efficient in-package voltage regulation
Chiplet Requirements: Test Innovation

- Yields from standard SoC bond, assembly, and test approaches are not viable for chiplet-based design.

- High reliability server applications present further challenges with additional test post-wafer test sectors.
  - Burn-in
  - System-level test

- Action is needed at all levels to achieve “known-good die” without expensive additional test steps.
  - Microarchitectural redundancy & flexibility
  - Robust circuit design
  - Test section capability: heal, not kill

- Chiplets will also require new test capabilities
  - Probe heterogenous & fine bump pitches
  - Validate 3rd party “black-box” chiplets
  - Identify cross-chiplet interactions (noise, IR, etc)
Chiplet Requirements: Standards & Tooling

• Robust support for today’s SoC ecosystem
  • Standards: VHDL/Veriflog, Oasis/GDS, UPF,
  • Tooling: DRC, LVS, ERC, static timing, noise, power, thermal, etc.

• Similar support needed to support rapid integration of chiplets at package level
  • Power & thermal models for cooling & current delivery
  • Electrical models for noise analysis and signal integrity
  • Mechanical models to study package stress

Thermal Models

Signal Integrity

Mechanical Stress
Chiplet Requirements: Business Models

Opportunities
- Improved solutions
- Increased volume
- Focused investment
- Time-to-market

Challenges
- Supply chain continuity
- Test ownership
- Field support & debug
- Warranty & liability
Chiplet Requirements: Open Interconnect Protocol

- Architecture agnostic
- Asymmetric
- Latency optimized
- Flexible
- Robust, silicon-proven
Open Standards for Servers

• Diss-aggregation of chiplets into customized form factors
  • CPU socket,
  • Memory DIMM: Existing standards difficult specialize
    • Determinism prevents flexibility
  • PCIe: Open standard with long standing compatibility and flexibility
    • high latency in hw and sw
    • Limited power and cooling
  • OMI (Open Memory Interface): Enables flexible memory standards
    • SerDes protocol provided ~5x reduction in host IO overhead
• OAM: Emerging standard has great potential
  • One physical standard, multiple DL/TL.
  • Variability in protocol and topology complicates systems
POWER9 – Acceleration Platform

- Extreme Processor / Accelerator Bandwidth and Reduced Latency
- Coherent Memory and Virtual Addressing Capability for all Accelerators
- OpenPOWER Community Enablement – Robust Accelerated Compute Options

- State of the Art I/O and Acceleration Attachment Signaling
  - PCIe Gen 4 x 48 lanes – 192 GB/s duplex bandwidth
  - 25 G Common Link x 96 lanes – 600 GB/s duplex bandwidth

- Robust Accelerated Compute Options with OPEN standards
  - On-Chip Acceleration – Gzip x1, 842 Compression x2, AES/SHA x2
  - CAPI 2.0 – 4x bandwidth of POWER8 using PCIe Gen 4
  - NVLink – Next generation of GPU/CPU bandwidth
  - OpenCAPI – High bandwidth, low latency and open interface
  - OMI – High bandwidth and/or differentiated for acceleration
POWER9 Family Memory Architecture

Scale Up
Buffered Memory

Superior RAS, High bandwidth, High Capacity
Agnostic interface for alternate memory innovations

Scale Out
Direct Attach Memory

Low latency access
Commodity packaging form factor

OpenCAPI Agnostic Buffered Memory (OMI)

Near Tier
- Extreme Bandwidth
- Low Capacity

Commodity
- Low Latency
- Low Cost

Enterprise
- RAS
- Capacity
- Bandwidth

Storage Class
- Extreme Capacity
- Persistence

Same Open Memory Interface used for all Systems and Memory Technologies
Primary Tier Memory Options

- **DDR4 RDIMM**
  - Capacity: ~256 GB
  - BW: ~150 GB/sec

- **DDR4 LRDIMM**
  - Capacity: ~2 TB
  - BW: ~150 GB/sec

- **DDR4 OMI DIMM**
  - Capacity: ~256 GB → 4 TB
  - BW: ~320 GB/sec

- **BW Opt OMI DIMM**
  - Capacity: ~128 → 512 GB
  - BW: ~650 GB/sec

- **On Module HBM**
  - Capacity: ~16 → 32 GB
  - BW: ~1 TB/sec

- **OMI Strategy**
  - Only 5-10 ns higher load-to-use than RDIMM (< 5 ns for LRDIMM)

- **Same System**
- **Unique System**
DRAM DIMM Comparison

IBM Centaur DIMM

OMI DDIMM

JEDEC DDR DIMM

• Technology agnostic
• Low cost
• Ultra-scale system density
• Enterprise reliability
• Low-latency
• High bandwidth

Approximate Scale
Open Memory Interface (OMI)

- Signaling: 25.6GHz vs DDR4 @ 3200 MHz
  - 4x raw bandwidth per I/O signal
  - 1.3x mixed traffic utilization
- Idle load-to-use latency over traditional DDR:
  - POWER8/9 Centaur design ~10 ns
  - OMI target of ~5-10 ns (RDIMM)
  - OMI target of < 5ns (LRDIMM)
- IBM Centaur: One proprietary DMI design
- Microchip SMC 1000:
  - Open (OMI) design
  - Emerging JEDEC Standard
OpenCAPI Design Goals

- Designed to support range of devices
  - Coherent Caching Accelerators
  - Network Controllers
  - Differentiated Memory
    - High Bandwidth
    - Low Latency
    - Storage Class Memory
  - Storage Controllers

- Asymmetric design, endpoint optimized for host and device attach
  - **ISA of Host Architecture**: Need to hide difference in Coherence, Memory Model, Address Translation, etc.
  - **Design schedule**: The design schedule of a high performance CPU host is typically on the order of multiple years, conversely, accelerator devices have much shorter development cycles, typically less than a year.
  - **Timing Corner**: ASIC and FPGA technologies run at lower frequencies and timing optimization as CPUs.
  - **Plurality of devices**: Effort in the host, both IP and circuit resource, have a multiplicative effect.
  - **Trust**: Attached devices are susceptible to both intentional and unintentional trust violations
  - **Cache coherence**: Hosts have high variability in protocol. Host cannot trust attached device to obey rules.
# OpenCAPI 4.0: Asymmetric Open Accelerator Attach

## Roadmap of Capabilities and Host Silicon Delivery

<table>
<thead>
<tr>
<th>Accelerator Protocol</th>
<th>CAPI 1.0</th>
<th>CAPI 2.0</th>
<th>OpenCAPI 3.0</th>
<th>OpenCAPI 4.0</th>
<th>OpenCAPI 5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Host Silicon</strong></td>
<td>POWER8 (GA 2014)</td>
<td>POWER9 SO (GA 2017)</td>
<td>POWER9 SO (GA 2017)</td>
<td>POWER9 AIO (GA 2020)</td>
<td>POWER10 (GA 2021)</td>
</tr>
<tr>
<td><strong>Functional Partitioning</strong></td>
<td>Asymmetric</td>
<td>Asymmetric</td>
<td>Asymmetric</td>
<td>Asymmetric</td>
<td>Asymmetric</td>
</tr>
<tr>
<td><strong>Host Architecture</strong></td>
<td>POWER</td>
<td>POWER</td>
<td>Any</td>
<td>Any</td>
<td>Any</td>
</tr>
<tr>
<td><strong>Cache Line Size Supported</strong></td>
<td>128B</td>
<td>128B</td>
<td>64/128/256B</td>
<td>64/128/256B</td>
<td>64/128/256B</td>
</tr>
<tr>
<td><strong>Address Translation</strong></td>
<td>On Accelerator</td>
<td>Host</td>
<td>Host (secure)</td>
<td>Host (secure)</td>
<td>Host (secure)</td>
</tr>
<tr>
<td><strong>Native DMA to Host Mem</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Atomics to Host Mem</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Host Thread Wake-up</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Host Memory Attach Agent</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Low Latency Short Msg</strong></td>
<td>4B/8B MMIO</td>
<td>4B/8B MMIO</td>
<td>4B/8B MMIO</td>
<td>128B push</td>
<td>128B push</td>
</tr>
<tr>
<td><strong>Posted Writes to Host Mem</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Caching of Host Mem</strong></td>
<td>RA Cache</td>
<td>RA Cache</td>
<td>No</td>
<td>VA Cache</td>
<td>VA Cache</td>
</tr>
</tbody>
</table>
Summary: Taking a step back...

• Disaggregation trend
  • ~Rack scale: Separate components into pools
    • Memory, CPU, GPU, storage
  • Enables agile deployment
  • Interface Goal: Make ~rack scale appear as board level

• Chiplet trend
  • Example: SOC like, but only build what is needed, buy standard components.
  • Interface Goal: On die like communication (energy, bandwidth)

• Shared goals
  • Heterogeneous Si
  • Efficient flexible protocol layers
    • Probably on application optimized physical layer