IBM’s Next Generation POWER Processor

Hot Chips
August 18-20, 2019

Jeff Stuecheli
Scott Willenborg
William Starke
# Proposed POWER Processor Technology and I/O Roadmap

## Focus of 2018 talk

<table>
<thead>
<tr>
<th>Year</th>
<th>Architecture</th>
<th>Cores</th>
<th>Technology</th>
<th>Micro-Architecture</th>
<th>Process Technology</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></td>
<td>POWER7+</td>
<td>8</td>
<td>32nm</td>
<td>Enhanced Micro-Architecture</td>
<td>New Process Technology</td>
</tr>
<tr>
<td>2012</td>
<td>POWER8</td>
<td>12</td>
<td>22nm</td>
<td>New Micro-Architecture</td>
<td>New Process Technology</td>
</tr>
<tr>
<td>2014</td>
<td>POWER8 w/ NVLink</td>
<td>12/24</td>
<td>14nm</td>
<td>Enhanced Micro-Architecture With NVLink</td>
<td>New Process Technology</td>
</tr>
<tr>
<td>2016</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></td>
<td>P9 SU</td>
<td>12/24</td>
<td>14nm</td>
<td>Enhanced Micro-Architecture</td>
<td>Buffered Memory</td>
</tr>
<tr>
<td>2018</td>
<td>P9 AIO</td>
<td>12/24</td>
<td>14nm</td>
<td>Enhanced Micro-Architecture</td>
<td>New Memory Subsystem</td>
</tr>
<tr>
<td>2020</td>
<td>P10</td>
<td>TBA</td>
<td>TBA</td>
<td>New Micro-Architecture</td>
<td>New Process Technology</td>
</tr>
</tbody>
</table>

**Sustained Memory Bandwidth**

- **Up To 65 GB/s**
- **Up To 210 GB/s**
- **Up To 150 GB/s**
- **Up To 210 GB/s**
- **Up To 650 GB/s**
- **Up To 800 GB/s**

**Standard I/O Interconnect**

- PCIe Gen2
- PCIe Gen3
- PCIe Gen4 x48
- PCIe Gen5

**Advanced I/O Signaling**

- N/A
- N/A
- 20 GT/s 160GB/s
- 25 GT/s 300GB/s
- 25 GT/s 300GB/s
- 32 & 50 GT/s

**Advanced I/O Architecture**

- N/A
- N/A
- CAPI 1.0
- CAPI 1.0, OpenCAPI3.0, NVLink
- CAPI 2.0, OpenCAPI3.0, NVLink
- CAPI 2.0, OpenCAPI4.0, NVLink
- TBA

**Statement of Direction, Subject to Change**
## Proposed POWER Processor Technology and I/O Roadmap

### Focus of today’s talk

<table>
<thead>
<tr>
<th>Year</th>
<th>Architecture</th>
<th>Cores</th>
<th>Process Technology</th>
<th>Micro-Architecture</th>
<th>Interconnect</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>POWER7</td>
<td>8</td>
<td>45nm</td>
<td>New Micro-Architecture</td>
<td>PCIe Gen2</td>
</tr>
<tr>
<td></td>
<td>POWER7+</td>
<td>8</td>
<td>32nm</td>
<td>Enhanced Micro-Architecture</td>
<td>PCIe Gen2</td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>POWER8</td>
<td>12</td>
<td>22nm</td>
<td>New Micro-Architecture</td>
<td>CAPI 1.0, NVLink</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Enhanced Micro-Architecture</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>POWER8 w/ NVLink</td>
<td>12/24</td>
<td>14nm</td>
<td>New Micro-Architecture With NVLink</td>
<td>PCIe Gen4 x48</td>
</tr>
<tr>
<td>2017</td>
<td>P9 SO</td>
<td>12/24</td>
<td>14nm</td>
<td>Direct attach memory</td>
<td>PCIe Gen4 x48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>New Process Technology</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>P9 SU</td>
<td>12/24</td>
<td>14nm</td>
<td>Enhanced Micro-Architecture</td>
<td>PCIe Gen4 x48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Buffered Memory</td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td>P9 AIO</td>
<td>12/24</td>
<td>14nm</td>
<td>Enhanced Micro-Architecture</td>
<td>PCIe Gen5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>New Memory Subsystem</td>
<td>32 &amp; 50 GT/s</td>
</tr>
<tr>
<td>2021</td>
<td>P10</td>
<td>TBA</td>
<td>New Micro-Architecture</td>
<td>New Process Technology</td>
<td>TBA</td>
</tr>
</tbody>
</table>

### Sustained Memory Bandwidth
- **POWER7**: Up To 65 GB/s
- **POWER7+**: Up To 65 GB/s
- **POWER8**: Up To 210 GB/s
- **POWER8 w/ NVLink**: Up To 210 GB/s
- **P9 SO**: Up To 150 GB/s
- **P9 SU**: Up To 210 GB/s
- **P9 AIO**: Up To 650 GB/s
- **P10**: Up To 800 GB/s

### Standard I/O Interconnect
- **POWER7**: PCIe Gen2
- **POWER7+**: PCIe Gen2
- **POWER8**: PCIe Gen3
- **POWER8 w/ NVLink**: PCIe Gen4 x48
- **P9 SO**: PCIe Gen4 x48
- **P9 SU**: PCIe Gen4 x48
- **P9 AIO**: PCIe Gen5

### Advanced I/O Signaling
- **POWER7**: N/A
- **POWER7+**: N/A
- **POWER8**: N/A
- **POWER8 w/ NVLink**: 20 GT/s 160GB/s
- **P9 SO**: 25 GT/s 300GB/s
- **P9 SU**: 25 GT/s 300GB/s
- **P9 AIO**: 25 GT/s 300GB/s

### Advanced I/O Architecture
- **POWER7**: N/A
- **POWER7+**: N/A
- **POWER8**: CAPI 1.0, NVLink
- **POWER8 w/ NVLink**: CAPI 2.0, OpenCAPI3.0, NVLink
- **P9 SO**: CAPI 2.0, OpenCAPI3.0, NVLink
- **P9 SU**: CAPI 2.0, OpenCAPI4.0, NVLink
- **P9 AIO**: CAPI 2.0, OpenCAPI4.0, NVLink

**Statement of Direction, Subject to Change**
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</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>2020</td>
<td>P9 AIO</td>
<td>12/24</td>
<td>TBA</td>
<td>Enhanced Micro-Architecture</td>
<td>New Memory Subsystem</td>
</tr>
<tr>
<td>2021</td>
<td>P10</td>
<td>TBA</td>
<td>New Micro-Architecture</td>
<td>New Process Technology</td>
<td></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 800 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**: 20 GT/s, 300GB/s
- **POWER10**: 32 & 50 GT/s

**Advanced I/O Architecture**
- **POWER7**: N/A
- **POWER8**: CAPI 1.0, NVLink
- **POWER9**: CAPI 2.0, OpenCAPI3.0, NVLink
- **POWER10**: CAPI 2.0, OpenCAPI4.0, NVLink

**Up To**
- 210 GB/s PCIe Gen4 x48
- 250 GB/s CAPI 2.0, OpenCAPI3.0, NVLink
- 300 GB/s CAPI 2.0, OpenCAPI4.0, NVLink
- TBA

Statement of Direction, Subject to Change
Future Evolution of System Architecture

OMI Memory

Yesterday's Plumbing
Tomorrow's Differentiation

Emerging JEDEC Buffers

PowerAXON & PCI

Yesterday's Plumbing
Tomorrow's Differentiation

PCI Gen N+1

SMP / OpenCAPI / NVLink
POWER9 – Premier 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
THE WORLD’S TWO MOST POWERFUL SUPERCOMPUTERS
BUILT FOR THE AI ERA
WITH OPEN COLLABORATION
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 Design Goals

- Low Latency and High Bandwidth
  - Fixed width DL CRC
  - Aligned TL
  - Aligned Data
  - Separately pipelined control/tag vs data
    - Compromise in switching capability
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

<table>
<thead>
<tr>
<th>Memory Type</th>
<th>Capacity</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR4 RDIMM</td>
<td>~256 GB</td>
<td>~150 GB/sec</td>
</tr>
<tr>
<td>DDR4 LRDIMM</td>
<td>~2 TB</td>
<td>~150 GB/sec</td>
</tr>
<tr>
<td>DDR4 OMI DIMM</td>
<td>~256 GB → 4 TB</td>
<td>~320 GB/sec</td>
</tr>
<tr>
<td>BW Opt OMI DIMM</td>
<td>~128 → 512 GB</td>
<td>~650 GB/sec</td>
</tr>
<tr>
<td>On Module HBM</td>
<td>~16 → 32 GB</td>
<td>~1 TB/sec</td>
</tr>
</tbody>
</table>

**OMI Strategy**

- **72b ~2666 MHz bidi**: DDR4 RDIMM (8b)
- **72b ~2666 MHz bidi**: DDR4 LRDIMM (8b)
- **8b ~25G diff**: DDR4 OMI DIMM (8b, BUF)
- **8b ~25G diff**: BW Opt OMI DIMM (8b, BUF)
- **1024b**: On module Si interposer
- **16b**: On Module HBM

### Key Points
- **Same System**: Only 5-10ns higher load-to-use than RDIMM (< 5ns for LRDIMM)
- **Unique System**: Unique system considerations

© 2019 IBM Corporation
Final Addition to the POWER9 Processor Family

Processor Chip Details
- 728 mm² (25.3 x 28.8 mm)
- 8 Billion Transistors
- Up to 24 SMT4 Cores
- Up to 120 MB eDRAM L3 cache

Semiconductor Technology
- 14nm finFET
- Improved device performance
- Reduced energy
- eDRAM
- 17 layer metal stack

High Bandwidth Signaling
- 25 GT/s low energy differential
  - PowerAXON, OMI memory
- 16 GT/s low energy differential
  - Local SMP
- 16 GT/s PCIe Gen4

Open Memory Interface (OMI)
- 16 channels x8 at 25 GT/s
- 650 GB/s peak 1:1 r/w bandwidth
- Technology Agnostic
- Offered w/ Microchip DDR4 buffer (410 GB/s peak bandwidth)

PowerAXON 25 GT/s Attach
- Up to 16 socket glue-less SMP (4x24 SMP added to 3x30 local)
- Up to x48 NVIDIA NVLINK GPU attach
- Up to x48 OpenCAPI 4.0 coherent accelerator / memory attach

Industry Standard I/O Attach
- x48 PCIe Gen 4 at 16 GT/s
- Up to x16 CAPI 2.0 coherent accelerator / storage attach

The Bandwidth Beast
Advanced I/O (AIO)

2 TB/s Raw Signaling Bandwidth
Shared by 6 Attach Protocols
Unleashing the Bandwidth: POWER9 AIO + NVIDIA GPU

Potential System Offering

OMI Memory

OpenCAPI

PCIe Gen4

OMI Memory

OMI Memory

NVLink

NVLink

SMP

Statement of Direction, Subject to Change
# 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>Functional Partitioning</td>
<td>Asymmetric</td>
<td>Asymmetric</td>
<td>Asymmetric</td>
<td>Asymmetric</td>
<td>Asymmetric</td>
</tr>
<tr>
<td>Host Architecture</td>
<td>POWER</td>
<td>POWER</td>
<td>Any</td>
<td>Any</td>
<td>Any</td>
</tr>
<tr>
<td>Cache Line Size Supported</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>Address Translation</td>
<td>On Accelerator</td>
<td>Host</td>
<td>Host (secure)</td>
<td>Host (secure)</td>
<td>Host (secure)</td>
</tr>
<tr>
<td>Native DMA to Host Mem</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Atomics to Host Mem</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Host Thread Wake-up</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Host Memory Attach Agent</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Low Latency Short Msg</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>Posted Writes to Host Mem</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Caching of Host Mem</td>
<td>RA Cache</td>
<td>RA Cache</td>
<td>No</td>
<td>VA Cache</td>
<td>VA Cache</td>
</tr>
</tbody>
</table>
POWER9 Connectivity Variants

POWER9 Scale Out
Direct Attach Memory
2 Socket SMP

- x24 Accelerator
- 4 x DDR4 memory
- Interconnect

POWER9 Scale Up
DMI Buffered (Centaur) Memory
16 Socket SMP

- x48 PowerAXON
- 4 x DMI buffered memory
- Interconnect

POWER9 Advanced I/O
OMI Buffered Memory
16 Socket SMP

- x48 PowerAXON
- 8 x OMI buffered memory
- Advanced Interconnect

- OMI Bandwidth
- OpenCAPI function
- NVLINK function

3x bandwidth per mm² as DDR signaling

6x bandwidth per mm² as DDR signaling
DRAM DIMM Comparison

IBM Centaur DIMM

OMI DDIMM

- 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

8x25G Open Memory Interface (OMI) Serial DDR4 Smart Memory Controller

INCREASED MEMORY BANDWIDTH
Enables 4x memory channels vs. x72 DDR4

MEDIA INDEPENDENCE
Single OMI interface provides for multiple media types

LOWER SOLUTION COSTS
Reduced silicon, IP and package costs for CPUs and SoCs
Microchip’s SMC 1000 8x25G features an innovative low latency design that delivers less than four ns incremental latency over a traditional integrated DDR controller with LRDIMM. This results in OMI-based DDIMM products having virtually identical bandwidth and latency performance to comparable LRDIMM products.

6X bandwidth / PHY area advantage gives POWER9 AIO bandwidth of 16 DDR ports
The Bandwidth Beast
POWER9 with Advanced I/O (AIO)

OMI Memory

PowerAXON

NVLink

Thank You!