Power Systems

IBM

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

	POWER7 Architecture		POWER8 Architecture		POWER9 Architecture			POWER10
	2010 POWER7 8 cores 45nm New Micro-Architecture New Process Technology	2012 POWER7+ 8 cores 32nm Enhanced Micro-Architecture New Process Technology	2014 POWER8 12 cores 22nm New Micro-Architecture New Process Technology	2016 POWER8 w/ NVLink 12 cores 22nm Enhanced Micro-Architecture With NVLink	2017 P9 SO 12/24 cores 14nm New Micro-Architecture Direct attach memory New Process Technology	2018 P9 SU 12/24 cores 14nm Enhanced Micro-Architecture Buffered Memory	2020 P9 AIO 12/24 cores 14nm Enhanced Micro-Architecture New Memory Subsystem	2021 P10 TBA cores New Micro-Architecture New Process Technology
Sustained Memory Bandwidth	Up To 65 GB/s	Up To 65 GB/s	Up To 210 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 Gen2	PCIe Gen3	PCIe Gen3	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen5
Advanced I/O Signaling	N/A	N/A	N/A	20 GT/s 160GB/s	25 GT/s 300GB/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 , NVLink	CAPI 2.0, 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

	POWER7 Architecture		POWER8 Architecture		POWER9 Architecture			POWER10
	2010 POWER7 8 cores 45nm New Micro-Architecture New Process Technology	2012 POWER7+ 8 cores 32nm Enhanced Micro-Architecture New Process Technology	2014 POWER8 12 cores 22nm New Micro-Architecture New Process Technology	2016 POWER8 w/ NVLink 12 cores 22nm Enhanced Micro-Architecture With NVLink	2017 P9 SO 12/24 cores 14nm New Micro-Architecture Direct attach memory New Process Technology	2018 P9 SU 12/24 cores 14nm Enhanced Micro-Architecture Buffered Memory	2020 P9 AIO 12/24 cores 14nm Enhanced Micro-Architecture New Memory Subsystem	2021 P10 TBA cores New Micro-Architecture New Process Technology
Sustained Memory Bandwidth	Up To 65 GB/s	Up To 65 GB/s	Up To 210 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 Gen2	PCIe Gen3	PCIe Gen3	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen5
Advanced I/O Signaling	N/A	N/A	N/A	20 GT/s 160GB/s	25 GT/s 300GB/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 , NVLink	CAPI 2.0, OpenCAPI3.0, NVLink	CAPI 2.0, OpenCAPI3.0, NVLink	CAPI 2.0, OpenCAPI4.0, NVLink	TBA

Statement of Direction, Subject to Change

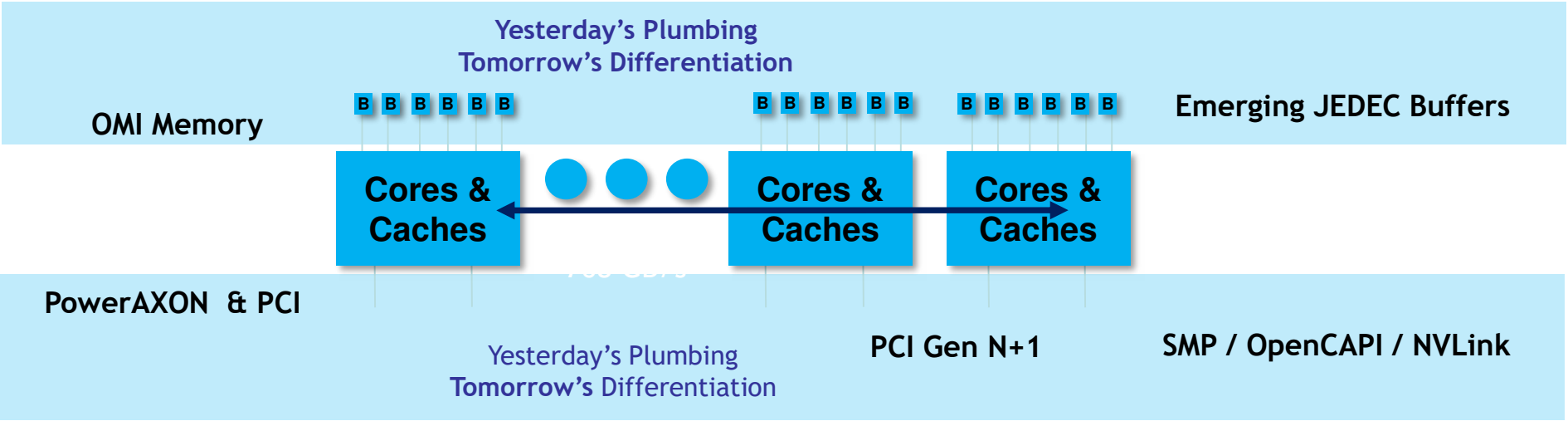
Proposed POWER Processor Technology and I/O Roadmap

Looking forward

	POWER7 Architecture		POWER8 Architecture		POWER9 Architecture			POWER10
	2010 POWER7 8 cores 45nm New Micro-Architecture New Process Technology	2012 POWER7+ 8 cores 32nm Enhanced Micro-Architecture New Process Technology	2014 POWER8 12 cores 22nm New Micro-Architecture New Process Technology	2016 POWER8 w/ NVLink 12 cores 22nm Enhanced Micro-Architecture With NVLink	2017 P9 SO 12/24 cores 14nm New Micro-Architecture Direct attach memory New Process Technology	2018 P9 SU 12/24 cores 14nm Enhanced Micro-Architecture Buffered Memory	2020 P9 AIO 12/24 cores 14nm Enhanced Micro-Architecture New Memory Subsystem	2021 P10 TBA cores New Micro-Architecture New Process Technology
Sustained Memory Bandwidth	Up To 65 GB/s	Up To 65 GB/s	Up To 210 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 Gen2	PCIe Gen3	PCIe Gen3	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen5
Advanced I/O Signaling	N/A	N/A	N/A	20 GT/s 160GB/s	25 GT/s 300GB/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 , NVLink	CAPI 2.0, OpenCAPI3.0, NVLink	CAPI 2.0, OpenCAPI3.0, NVLink	CAPI 2.0, OpenCAPI4.0, NVLink	TBA

Statement of Direction, Subject to Change

Future Evolution of System Architecture

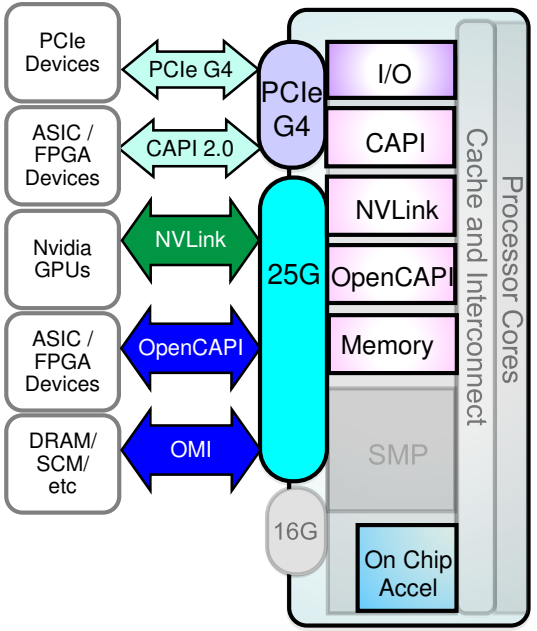


POWER9 – Premier Acceleration Platform

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

POWER9 PowerAccel

- **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

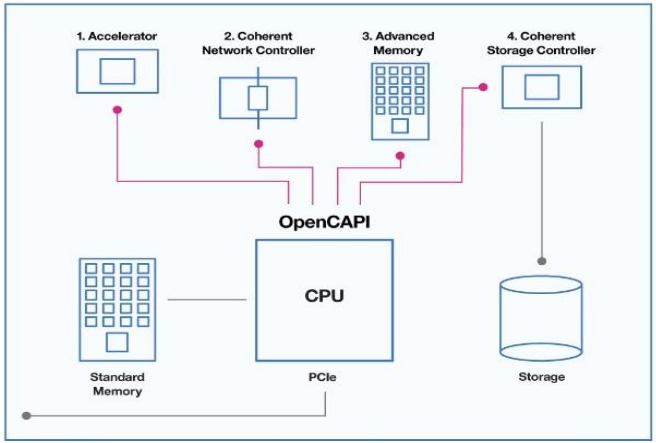
OpenPOWER™

Mellanox TECHNOLOGIES

NVIDIA

OpenCAPI Design Goals

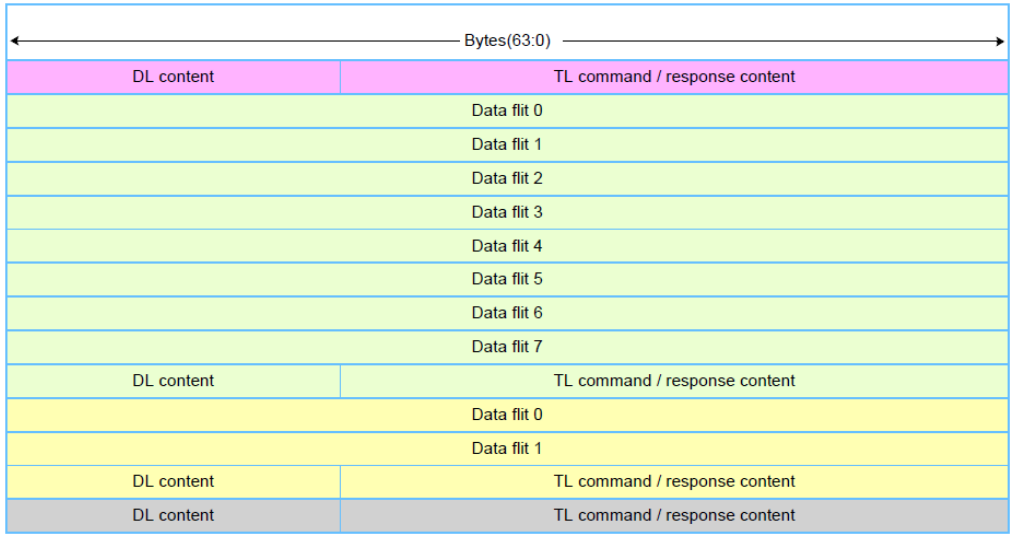
- Designed to support range of devices
  - Coherent Caching Accelerators
  - Network Controllers
  - Differentiated Memory
    - High Bandwidth
    - Low Latency
    - Storage Class Memory
  - Storage Controllers



- 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

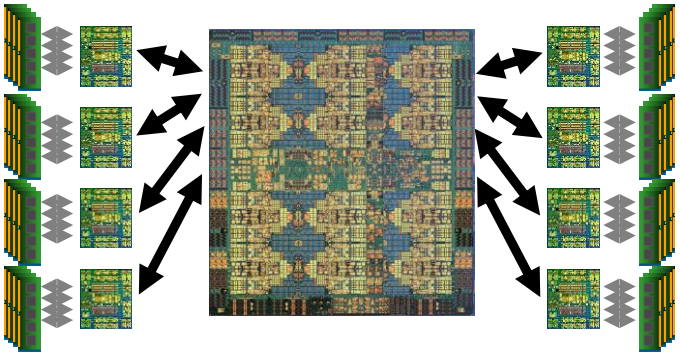
- Low Latency and High Bandwidth
  - Fixed width DL CRC
  - Aligned TL
  - Aligned Data
  - Separately pipelined control/tag vs data
    - Compromise in switching capability

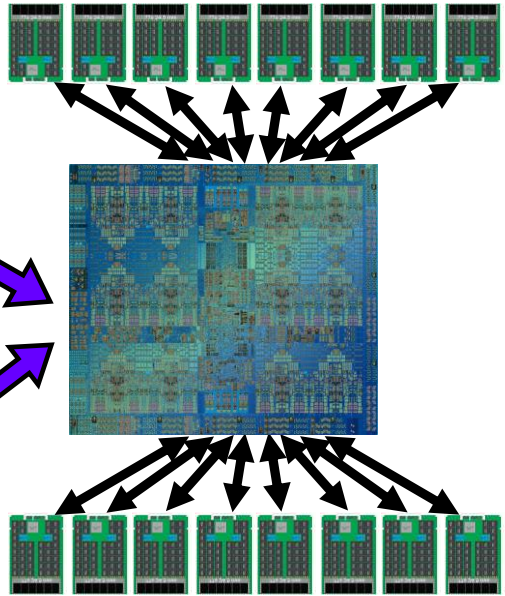


POWER9 Family Memory Architecture

Scale Up
Buffered Memory

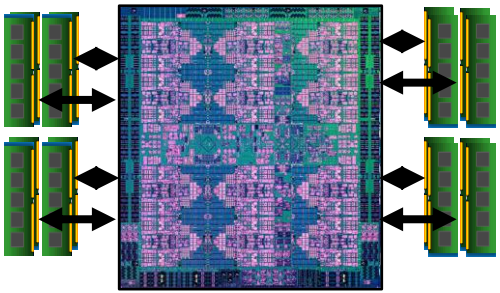
Superior RAS, High bandwidth, High Capacity
Agnostic interface for alternate memory innovations





Scale Out
Direct Attach Memory

Low latency access
Commodity packaging form factor

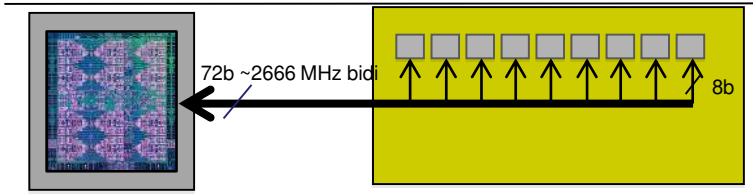


OpenCAPI Agnostic Buffered Memory (OMI)

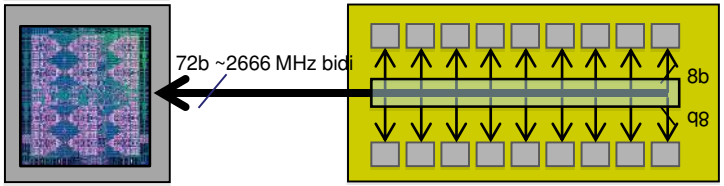
Near Tier	Commodity	Enterprise	Storage Class
Extreme Bandwidth	Low Latency	RAS	Extreme Capacity
Low Capacity	Low Cost	Capacity	Persistence
		Bandwidth	

Same Open Memory Interface used for all Systems and Memory Technologies

Primary Tier Memory Options



DDR4 RDIMM
Capacity ~256 GB
BW ~150 GB/sec

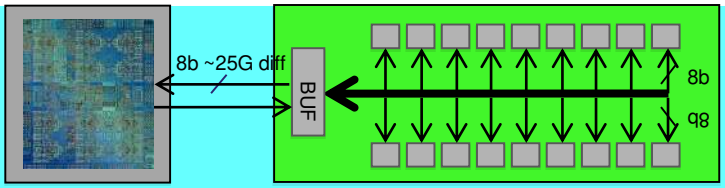


DDR4 LRDIMM
Capacity ~2 TB
BW ~150 GB/sec

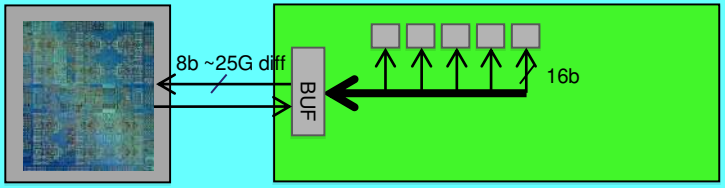
Same System

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

OMI Strategy

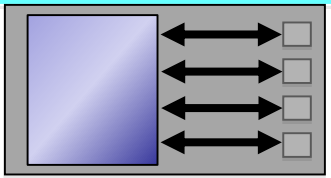


DDR4 OMI DIMM
Capacity ~256GB→4 TB
BW ~320 GB/sec



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

Same System



1024b
On module
Si interposer

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

Unique System

Final Addition to the POWER9 Processor Family

Processor Chip Details

- 728 mm^2 (25.3 x 28.8 mm)
- 8 Billion Transistors
- Up to 24 SMT4 Cores
- Up to 120 MB eDRAM L3 cache

Semiconductor Technology

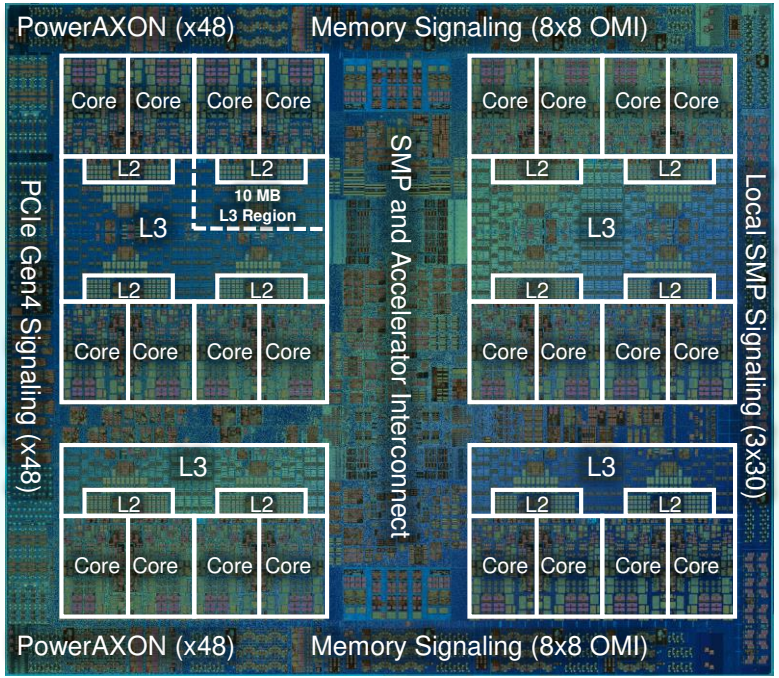
- 14nm finFET
- Improved device performance
- Reduced energy
- eDRAM
- 17 layer metal stack

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

The Bandwidth Beast

Advanced I/O (AIO)



2 TB/s Raw Signaling Bandwidth
Shared by 6 Attach Protocols

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

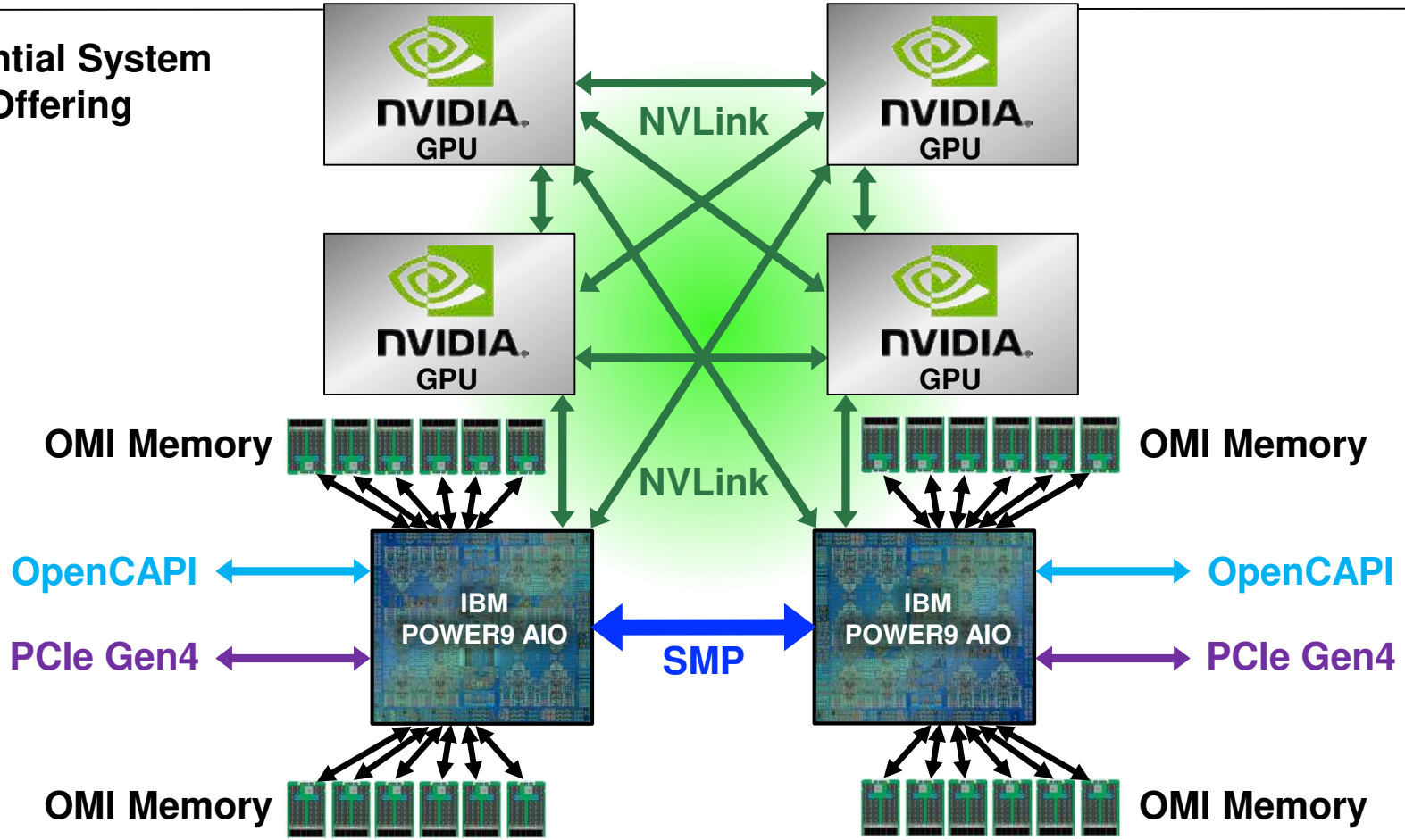
- Up to 16 socket glue-less SMP (4x24 SMP added to 3x30 local)
- Up to x48 NVIDIA NVLINK GPU attach
- Up to x48 OpenCAPI 4.0 coherent accelerator / memory attach

Industry Standard I/O Attach

- x48 PCIe Gen 4 at 16 GT/s
- Up to x16 CAPI 2.0 coherent accelerator / storage attach

Unleashing the Bandwidth: POWER9 AIO + NVIDIA GPU

Potential System Offering



Statement of Direction, Subject to Change

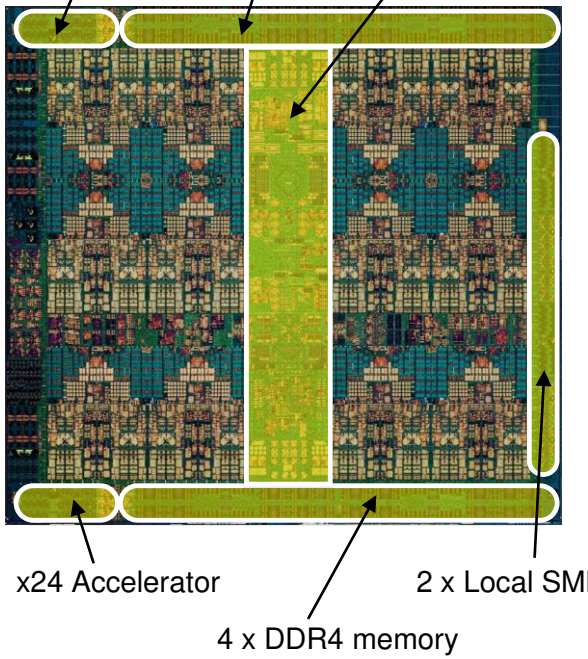
OpenCAPI 4.0: Asymmetric Open Accelerator Attach

Roadmap of Capabilities and Host Silicon Delivery

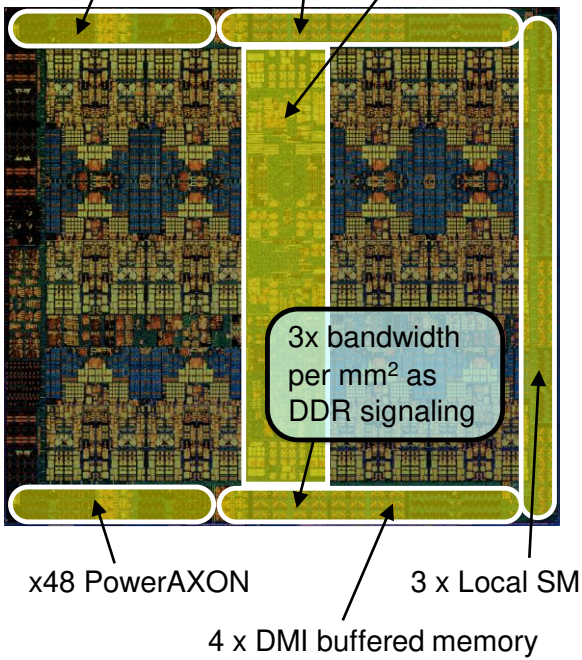
Accelerator Protocol	CAPI 1.0	CAPI 2.0	OpenCAPI 3.0	OpenCAPI 4.0	OpenCAPI 5.0
First Host Silicon	POWER8 (GA 2014)	POWER9 SO (GA 2017)	POWER9 SO (GA 2017)	POWER9 AIO (GA 2020)	POWER10 (GA 2021)
Functional Partitioning	Asymmetric	Asymmetric	Asymmetric	Asymmetric	Asymmetric
Host Architecture	POWER	POWER	Any	Any	Any
Cache Line Size Supported	128B	128B	64/128/256B	64/128/256B	64/128/256B
Attach Vehicle	PCIe Gen 3 Tunneled	PCIe Gen 4 Tunneled	25 G (open) Native DL/TL	25 G (open) Native DL/TL	32/50 G (open) Native DL/TL
Address Translation	On Accelerator	Host	Host (secure)	Host (secure)	Host (secure)
Native DMA to Host Mem	No	Yes	Yes	Yes	Yes
Atomics to Host Mem	No	Yes	Yes	Yes	Yes
Host Thread Wake-up	No	Yes	Yes	Yes	Yes
Host Memory Attach Agent	No	No	Yes	Yes	Yes
Low Latency Short Msg	4B/8B MMIO	4B/8B MMIO	4B/8B MMIO	**128B push**	128B push
Posted Writes to Host Mem	No	No	No	**Yes**	Yes
Caching of Host Mem	RA Cache	RA Cache	No	**VA Cache**	VA Cache

POWER9 Connectivity Variants

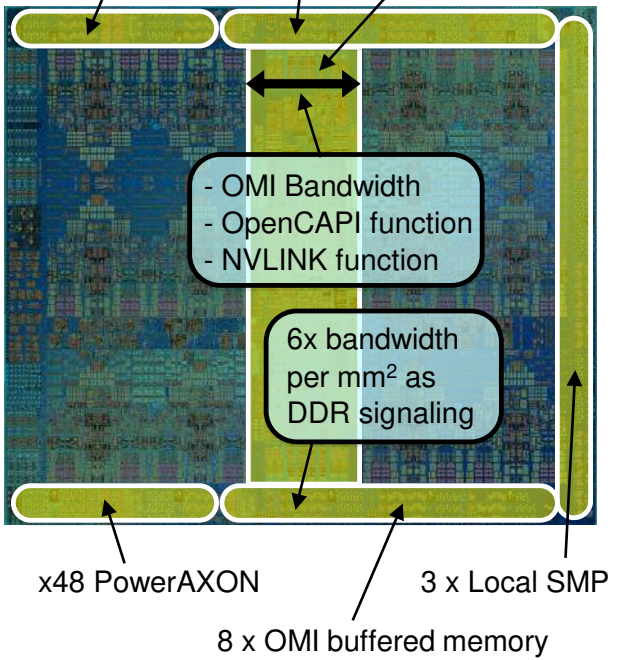
POWER9 Scale Out
Direct Attach Memory
2 Socket SMP



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



POWER9 Advanced I/O
OMI Buffered Memory
16 Socket SMP



DRAM DIMM Comparison

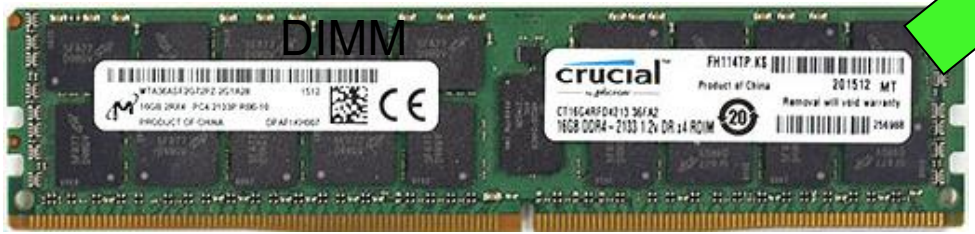
IBM Centaur DIMM

OMI DDIMM

Ultra-scale

Economy

JEDEC DDR DIMM



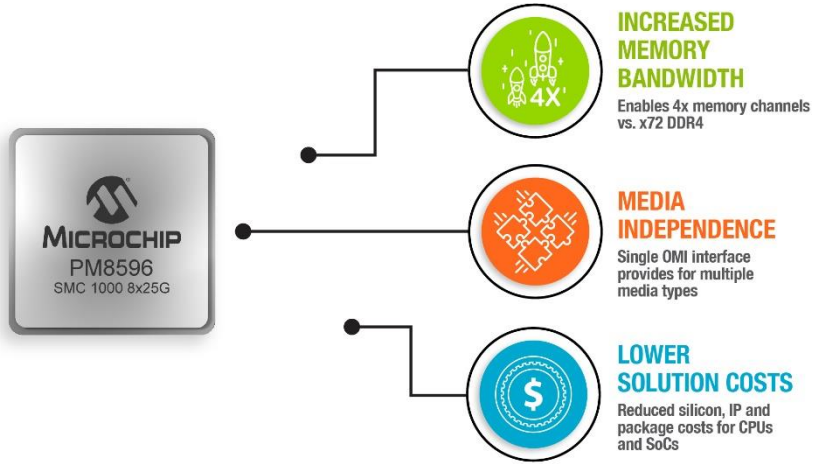
- Technology agnostic
- Low cost
- Ultra-scale system density
- Enterprise reliability
- Low-latency
- High bandwidth

Approximate Scale

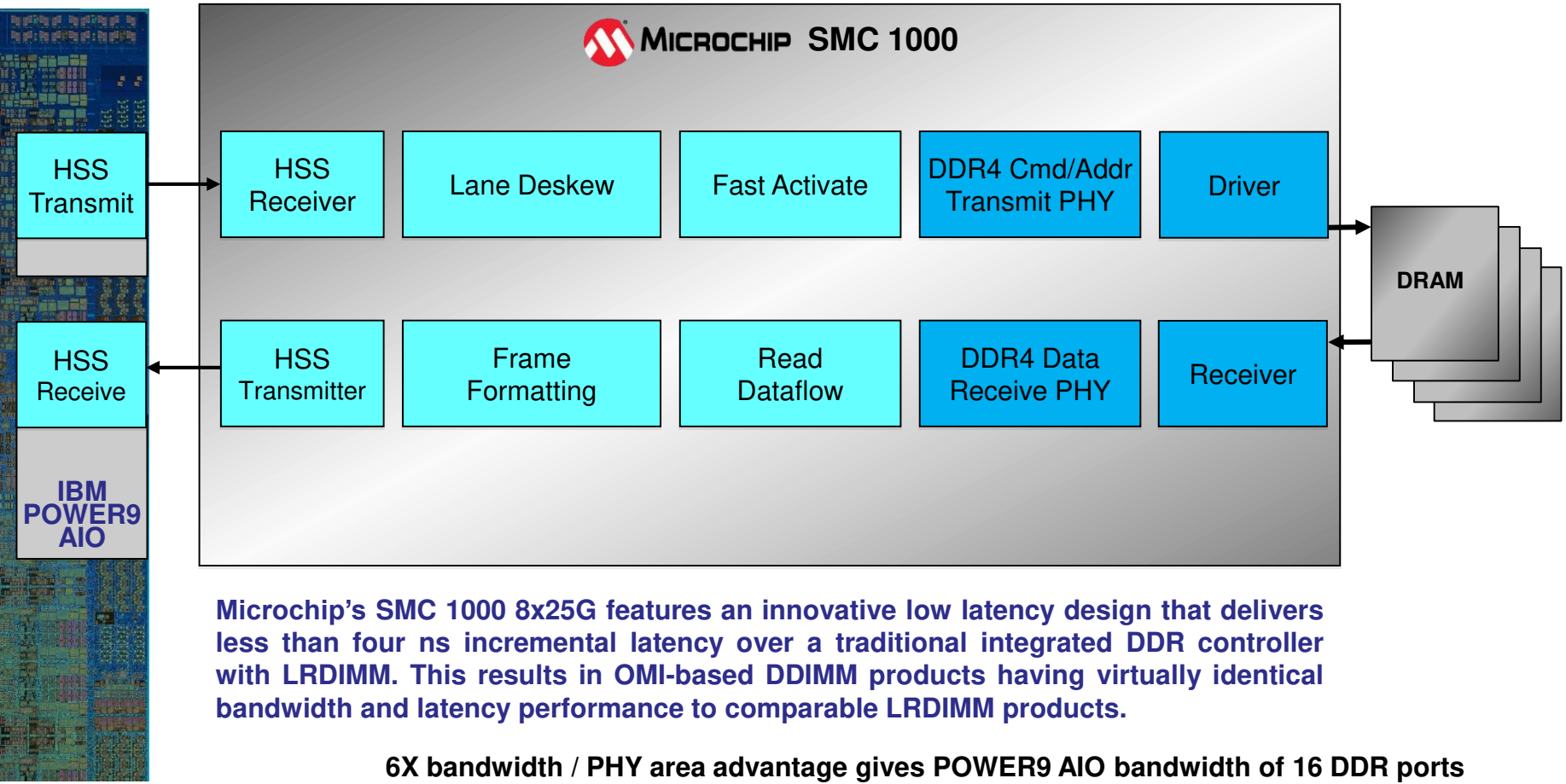
Open Memory Interface (OMI)

- 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



OMI Memory Latency



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

Final Addition to POWER9: Stay Tuned for POWER10

PowerAXON

NVIDIA GPU

NVIDIA GPU

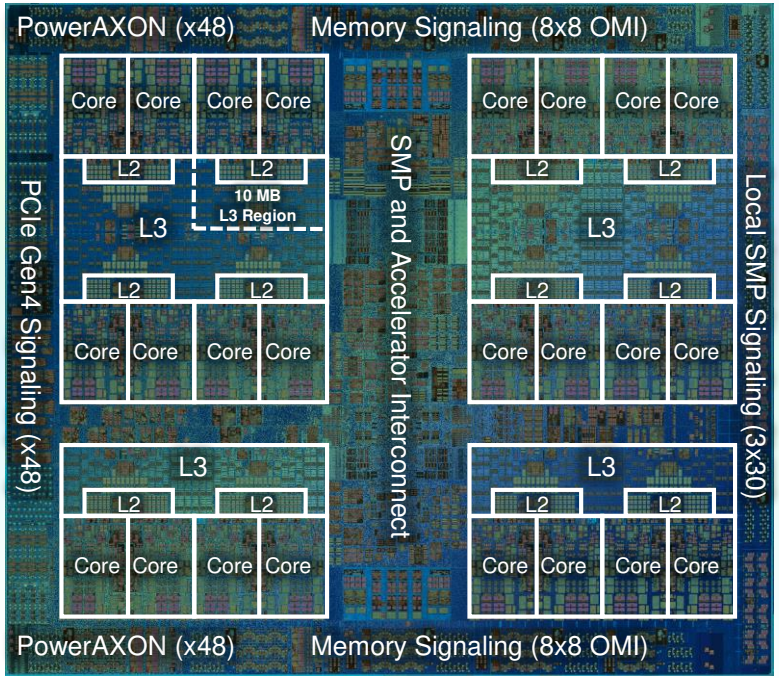
NVLink

OpenCAPI

OpenPOWER™

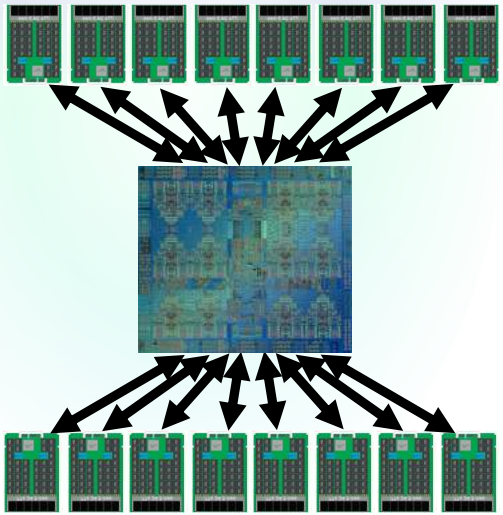
The Bandwidth Beast

POWER9 with Advanced I/O (AIO)



OMI Memory

MICROCHIP



Thank You!