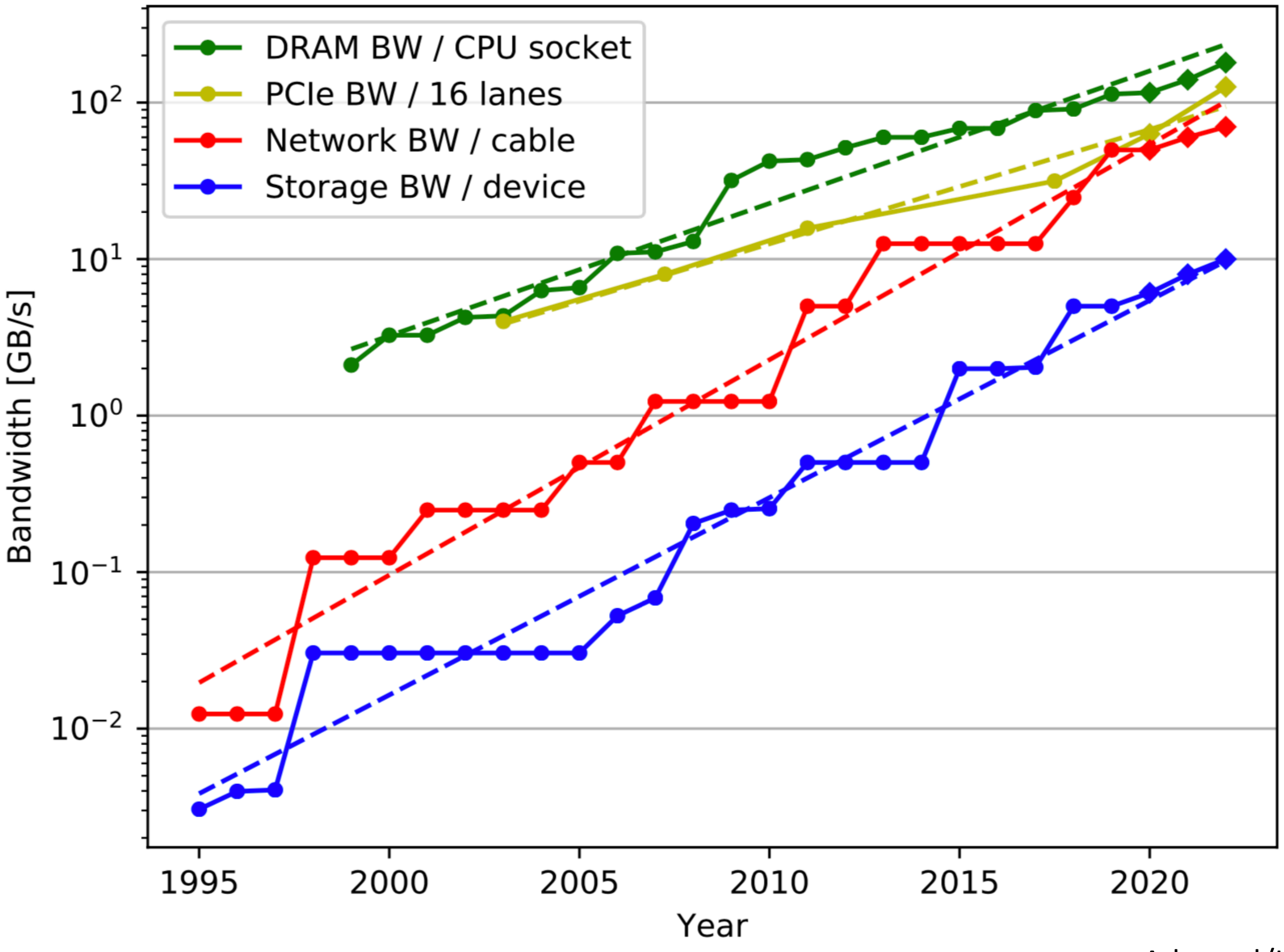
Distributed Memory on POWER 10

H. Peter Hofstee, IBM

hofstee@us.ibm.com

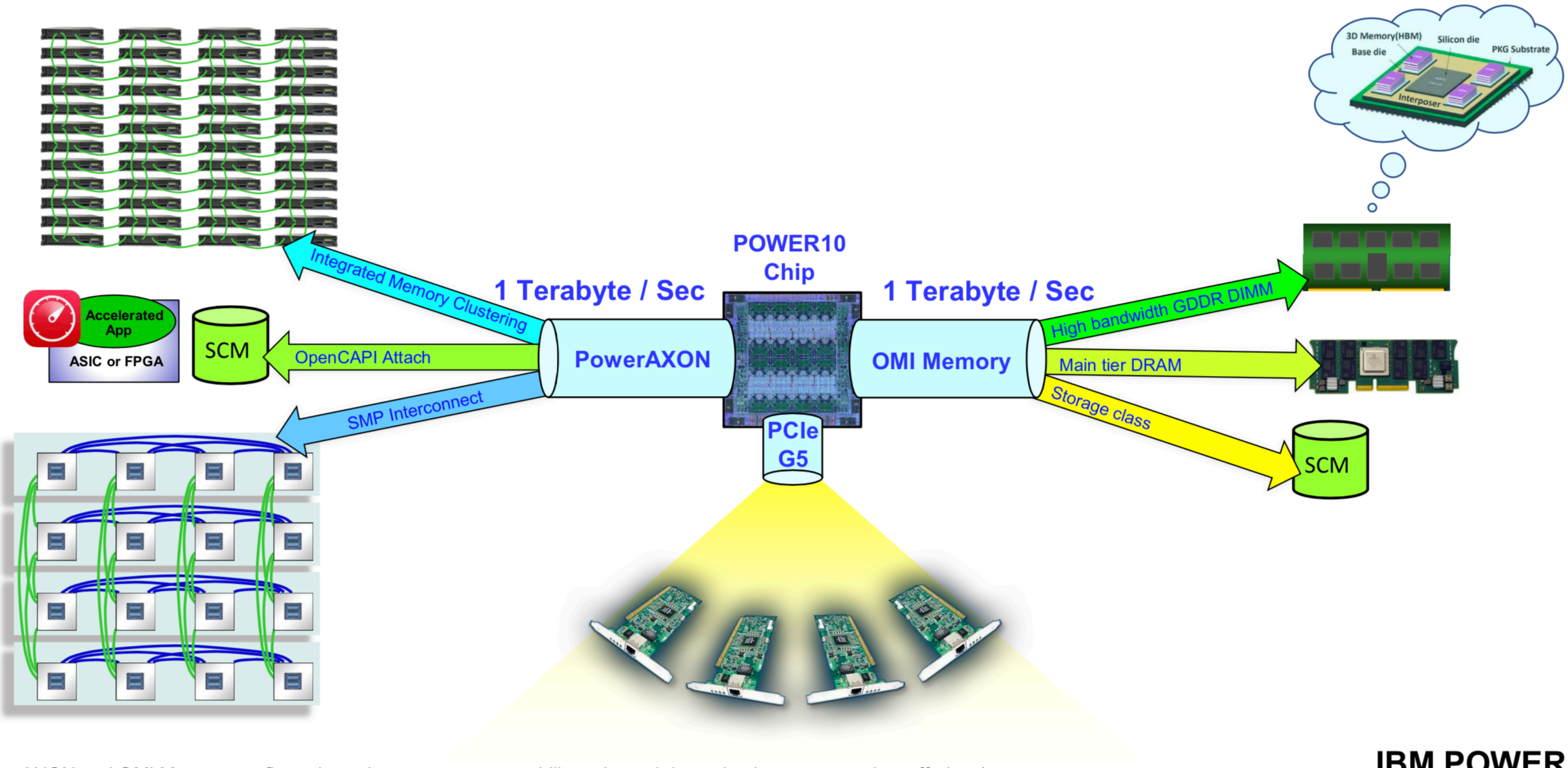
Agenda

- Bandwidth Trends
- OpenCAPI on POWER 10
- OpenCAPI Memory Disaggregation
  - POWER9 “Thymesisflow”
  - POWER10 “Memory Inception”



Adapted/Updated from Sandisk Blog

System Composability: POWER 10



(PowerAXON and OMI Memory configurations show processor capability only, and do not imply system product offerings)

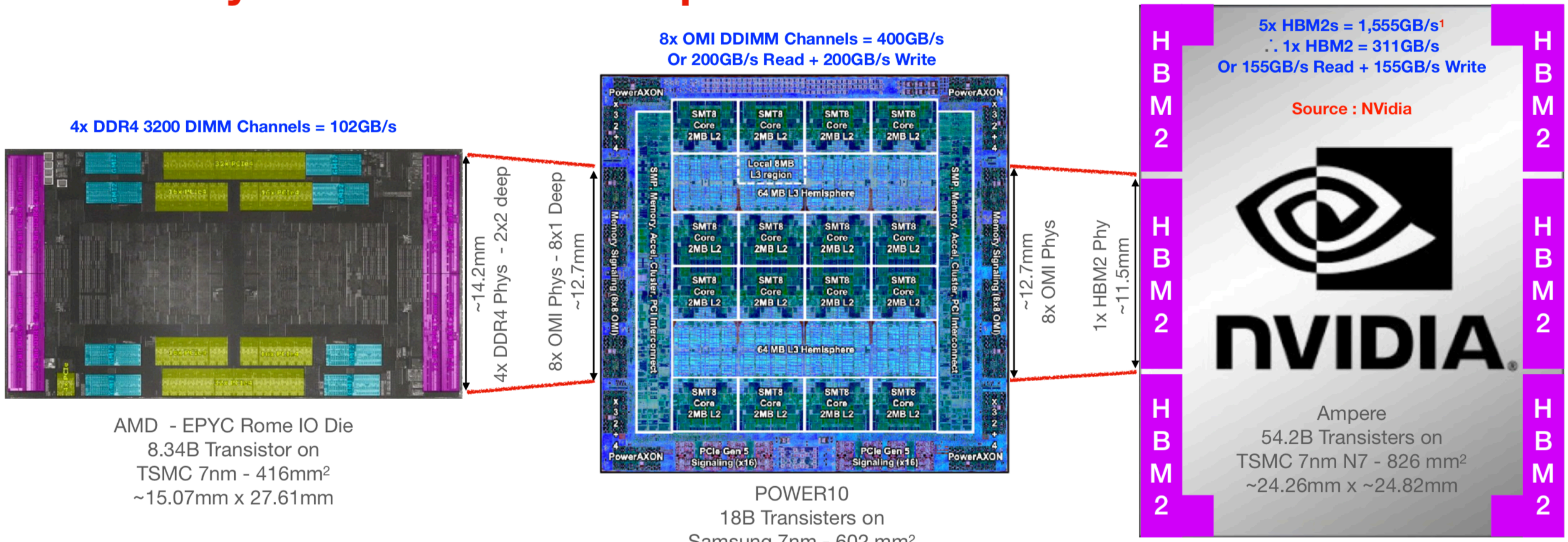
IBM POWER10

W. Starke & B. Thompto, Hot Chips 32, 2020

The OMI Advantage

OpenCAPI

Memory Bandwidth AND Depth at LOW Cost



PowerAXON
SMT8 Core 2MB L2
Local 8MB L3 region
64 MB L3 Hemisphere
SMP, Memory, Accel, Cluster, PCI Interconnect
Memory Signaling (8x8 OMI)
PCIe Gen 5 Signaling (x16)
HBM2
NVIDIA

4x DDR4 3200 DIMM Channels = 102GB/s

~14.2mm
4x DDR4 Phys - 2x2 deep

8x OMI DDIMM Channels = 400GB/s
Or 200GB/s Read + 200GB/s Write

8x OMI Phys - 8x1 Deep
~12.7mm

~12.7mm
8x OMI Phys

1x HBM2 Phy
~11.5mm

5x HBM2s = 1,555GB/s[1]
∴ 1x HBM2 = 311GB/s
Or 155GB/s Read + 155GB/s Write

Source : NVidia

AMD - EPYC Rome IO Die
8.34B Transistor on
TSMC 7nm - 416mm^2
~15.07mm x 27.61mm

POWER10
18B Transisters on
Samsung 7nm - 602 mm^2
~24.26mm x ~24.82mm

Ampere
54.2B Transisters on
TSMC 7nm N7 - 826 mm^2
~24.26mm x ~24.82mm

7.2GB/s / mm of Die Edge
Up to 36GBytes/mm of Die Edge

31.5GB/s / mm of Die Edge
Up to 81GBytes/mm of Die Edge*

27.0GB/s / mm of Die Edge
Up to 0.7GBytes/mm of Die Edge

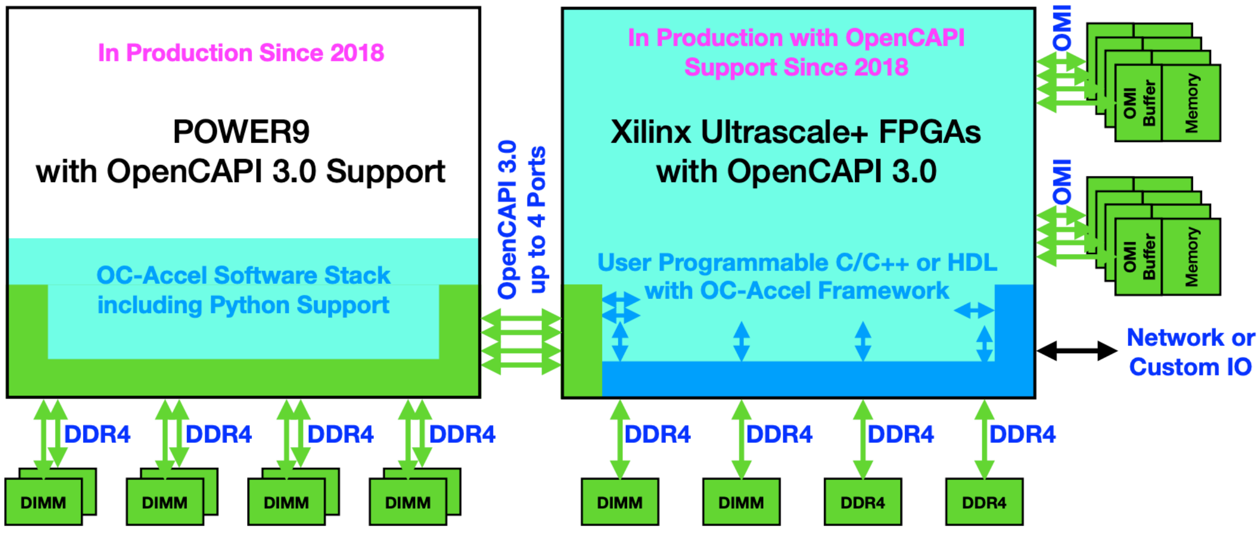
* Higher with different Media - e.g. 1.9TBytes/mm with BittWare 250-HMS

To Scale = 20pts : 1mm

Allan Cantle, 2020 OpenPOWER Summit NA

Great Concept……..The Reality?

In Production Today with OpenCAPI



In Production Since 2018

POWER9 with OpenCAPI 3.0 Support

OC-Accel Software Stack including Python Support

OpenCAPI 3.0 up to 4 Ports

In Production with OpenCAPI Support Since 2018

Xilinx Ultrascale+ FPGAs with OpenCAPI 3.0

User Programmable C/C++ or HDL with OC-Accel Framework

OMI

OMI Buffer

Memory

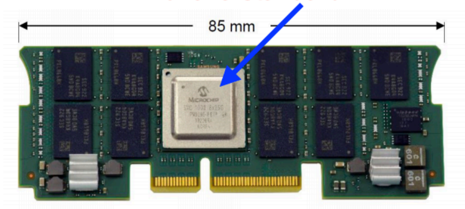
Network or Custom IO

Up to 8x DDR4 DIMMS @ 170GB/s Total

Up to 4x DDR4 DIMMS @ 85GB/s Total

Microchip OMI to DDR4 Buffer Chip

with <10ns latency adder over a standard RDIMM



1U DDIMM Format
72b DDR4 3200

DDR4 OpenCAPI Memory Interface OMI DDIMM

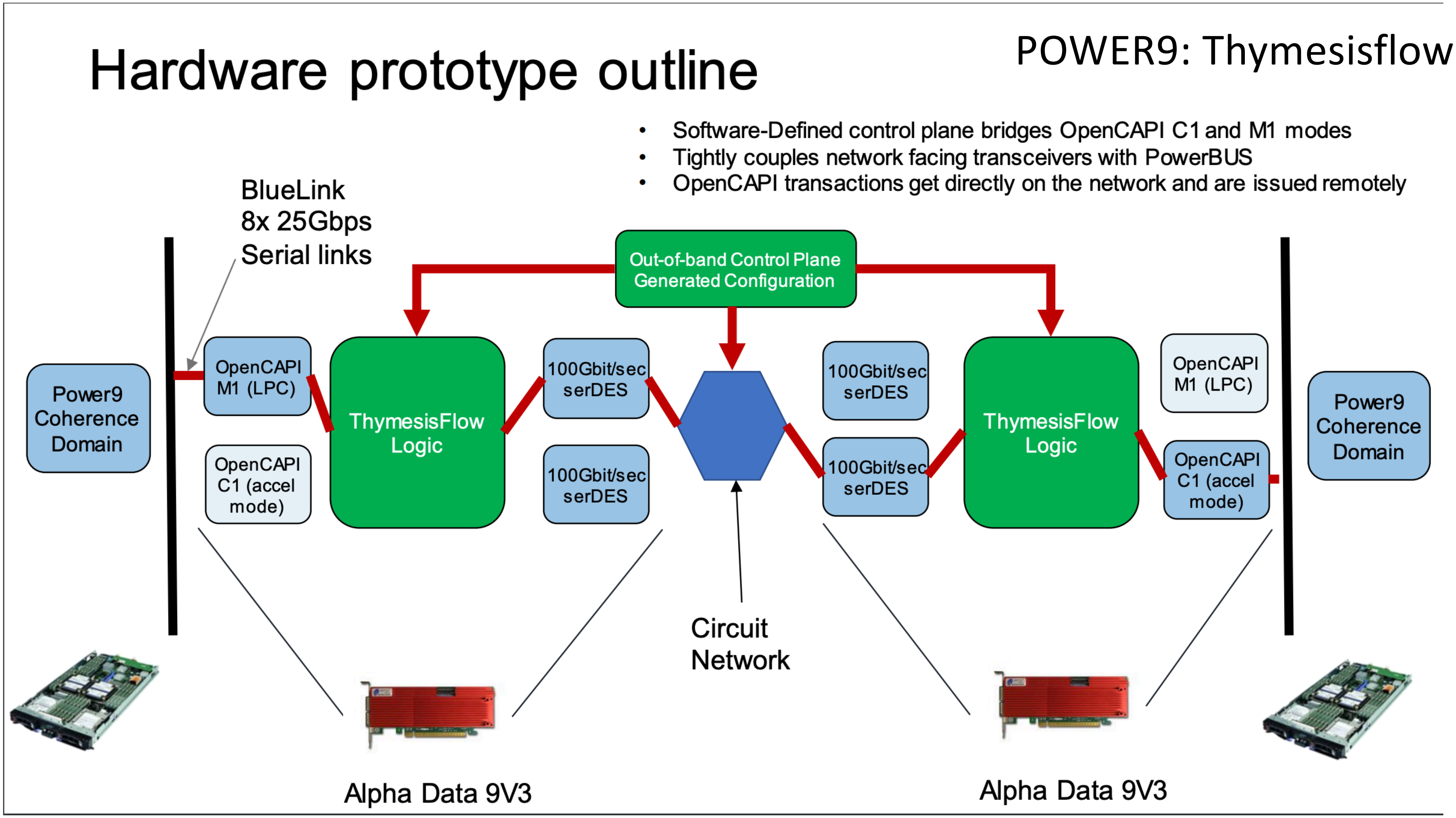
Introduced in mid 2019

Maxeler Lightning Talk on FPGA Application acceleration of Memory Bound Problems with OMI BoF Panel at 1:20pm CDT Track 2A

OpenCAPI Acceleration Framework - OC-Accel - Presentation at 10:35am CDT Track 2A

All RTL & Software is proven & fully Open Sourced

Allan Cantle, 2020 OpenPOWER Summit NA



Hardware prototype outline

POWER9: Thymesisflow

- Software-Defined control plane bridges OpenCAPI C1 and M1 modes
- Tightly couples network facing transceivers with PowerBUS
- OpenCAPI transactions get directly on the network and are issued remotely

BlueLink
8x 25Gbps
Serial links

Out-of-band Control Plane Generated Configuration

Power9 Coherence Domain

OpenCAPI M1 (LPC)

OpenCAPI C1 (accel mode)

ThymesisFlow Logic

100Gbit/sec serDES

100Gbit/sec serDES

Circuit Network

100Gbit/sec serDES

100Gbit/sec serDES

ThymesisFlow Logic

OpenCAPI M1 (LPC)

OpenCAPI C1 (accel mode)

Power9 Coherence Domain

Alpha Data 9V3

Alpha Data 9V3

Christian Pinto, 2020 OpenPOWER Summit NA (and Micro 53, Oct 2020)

https://github.com/OpenCAPI/ThymesisFlow

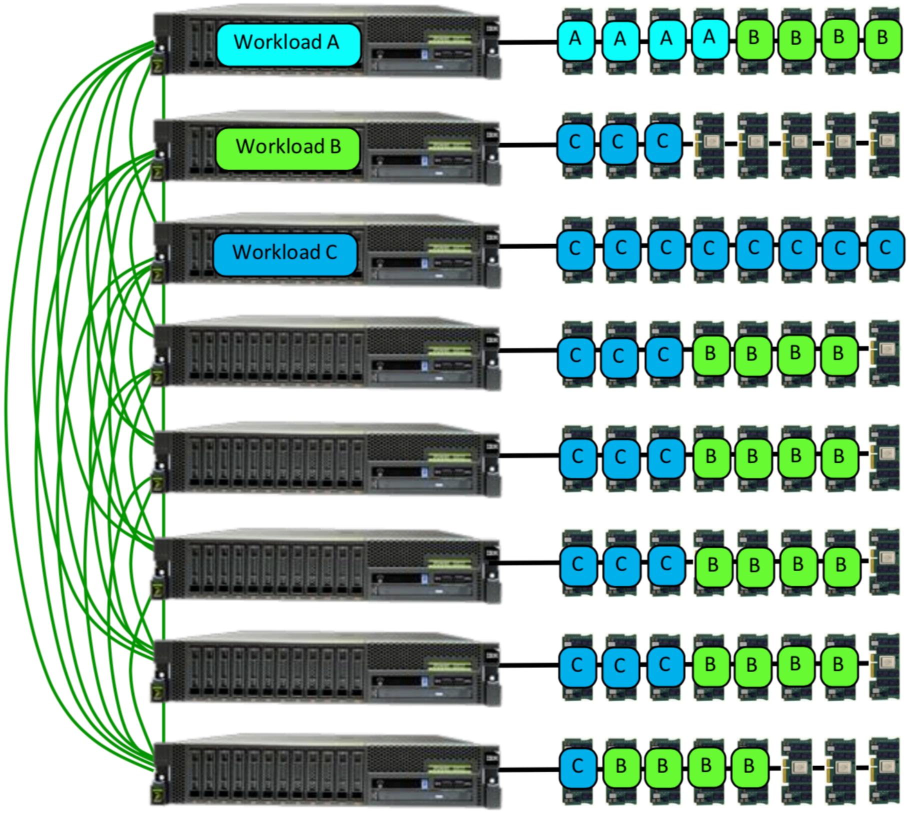
Use case: Share load/store memory amongst
directly connected neighbors within Pod
Unlike other schemes, memory can be used:

- As low latency local memory
- As NUMA latency remote memory

Example: Pod = 8 systems each with 8TB
Workload A Rqmt: 4 TB low latency
Workload B Rqmt: 24 TB relaxed latency
Workload C Rqmt: 8 TB low latency plus
16TB relaxed latency

All Rqmts met by configuration shown

POWER10 2 Petabyte memory size enables
much larger configurations

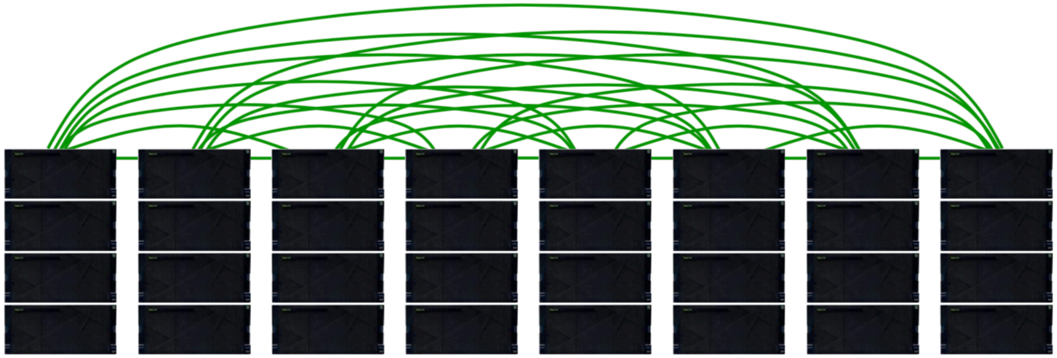


(Memory cluster configurations show processor capability only, and do not imply system product offerings)

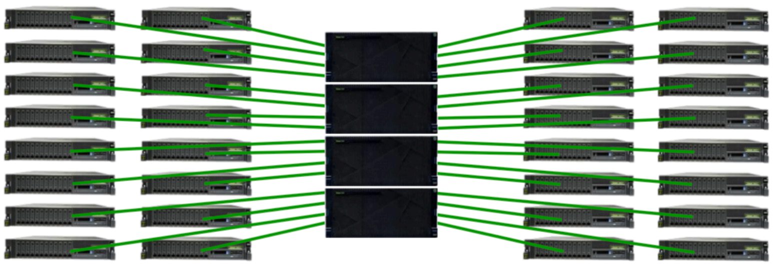
W. Starke & B. Thompto, Hot Chips 32, 2020

Memory Clustering: Enterprise-Scale Memory Sharing

Pod of Large Enterprise Systems
Distributed Sharing at Petabyte Scale



Or Hub-and-spoke with memory server
and memory-less compute nodes

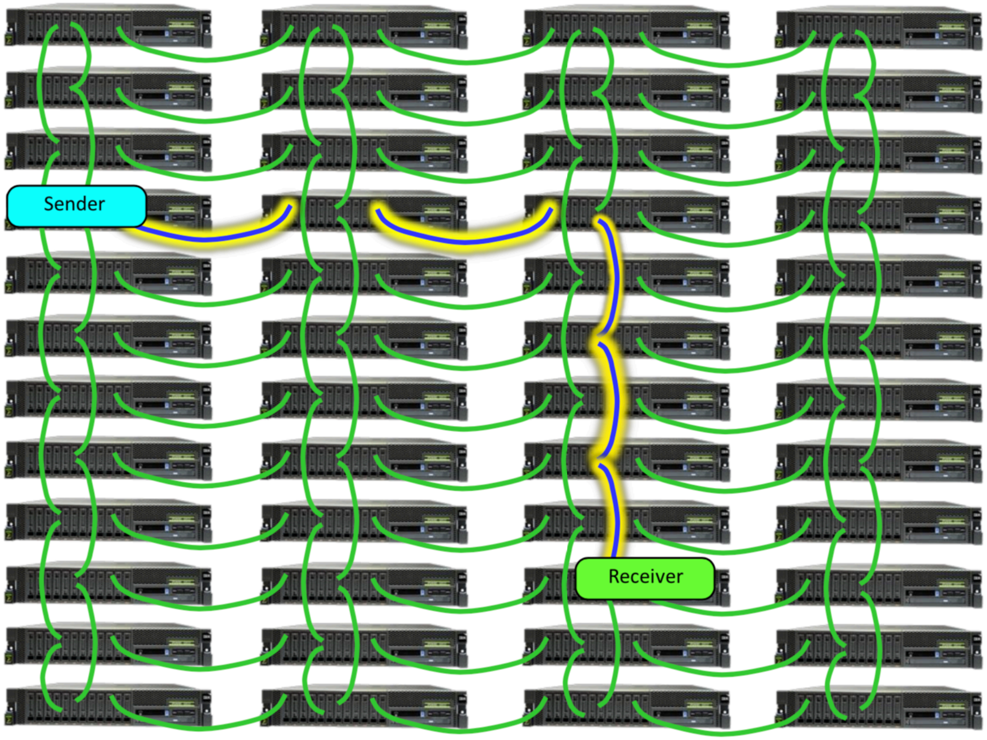


(Memory cluster configurations show processor capability only, and do not imply system product offerings)

Memory Clustering: Pod-level Clustering

Use case: Low latency, high bandwidth messaging scaling to 1000's of nodes

Leverage 2 Petabyte addressability to create memory window into each destination for messaging mailboxes



(Memory cluster configurations show processor capability only, and do not imply system product offerings)

IBM POWER10

W. Starke & B. Thompto, Hot Chips 32, 2020

References

- William Starke & Brian Thompto, “IBM’s POWER10 processor”, Hot Chips 32, Aug 16-18 2020

- https://events.linuxfoundation.org/openpower-summit-north-america/program/schedule/
  - Allan Cantle, “OpenCAPI, A Memory-Centric Fabric for a Data-Centric World”, Keynote 2020 OpenPOWER Summit NA, Sep 15, 2020
  - Christian Pinto, “Thymesisflow, A Hardware/Software Open Framework for Software-Defined Memory Disaggregation based on OpenCAPI”, 2020 OpenPOWER Summit NA, Sep 15, 2020