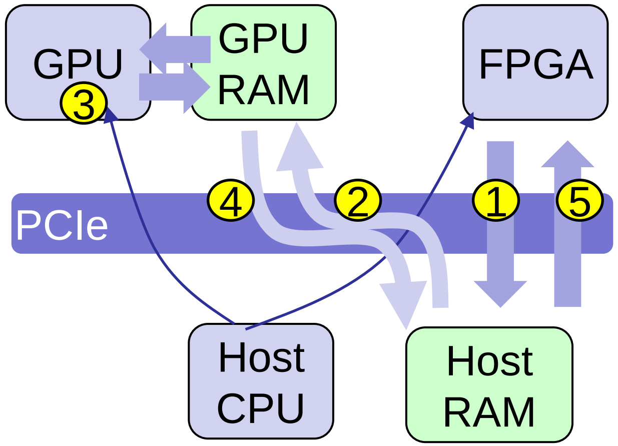
1 Adapters

2 Hardware Accelerators

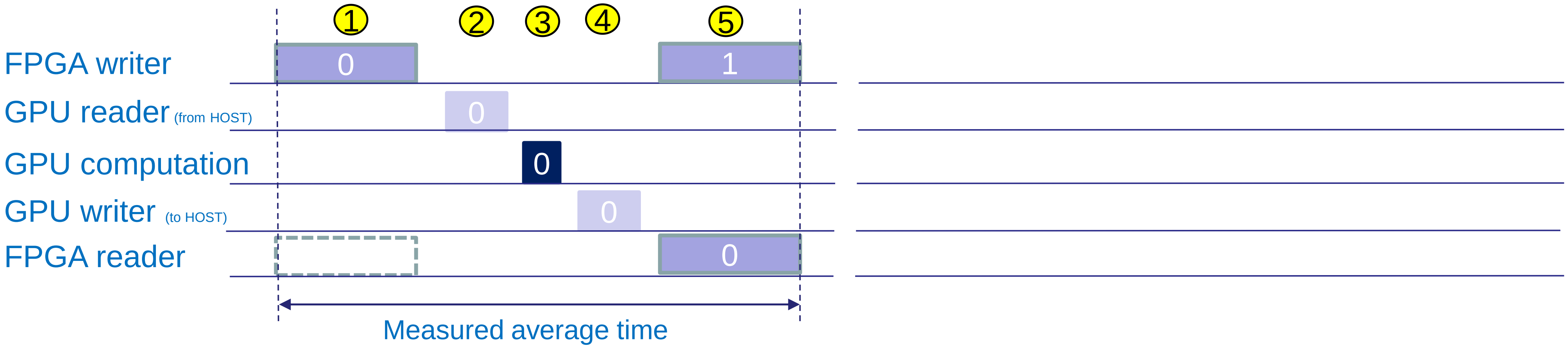
➔ FPGA-GPU combination

3 Host memory

Conventional system

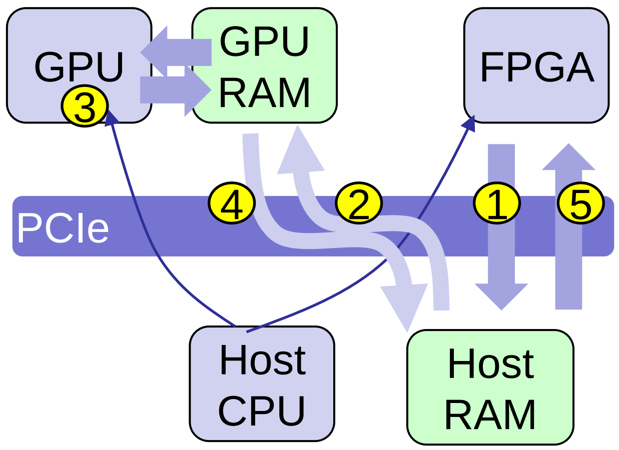


```
cudaDeviceGetAttribute(xxx);
cudaMemcpy(ibuff,bufferA, size, cudaMemcpyDeviceToHost);
// obuff = 2*ibuff
vector_add<<<4*numBlocks,numThreadsPerBlock>>>(ibuff,obuff,vector_size);
cudaMemcpy(bufferB, obuff, size, cudaMemcpyHostToDevice);
cudaDeviceSynchronize();
```



Conventional system

CAPI/OpenCAPI system

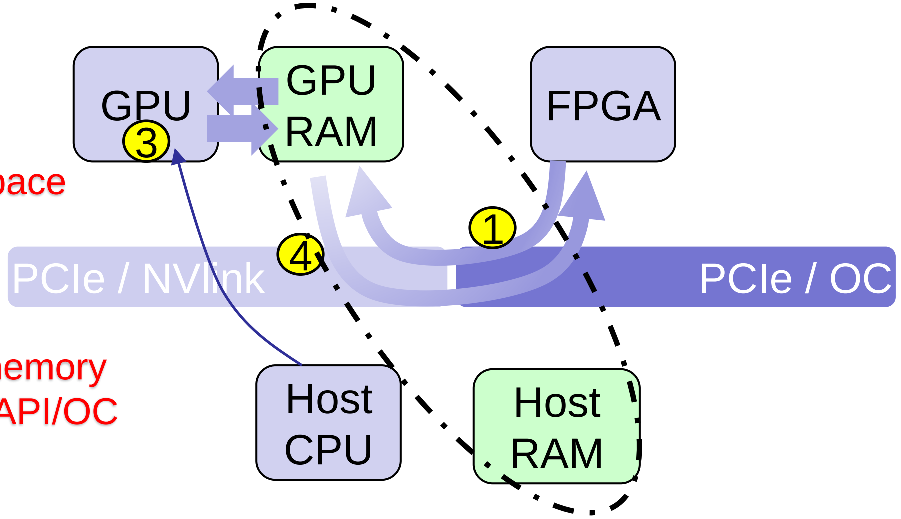


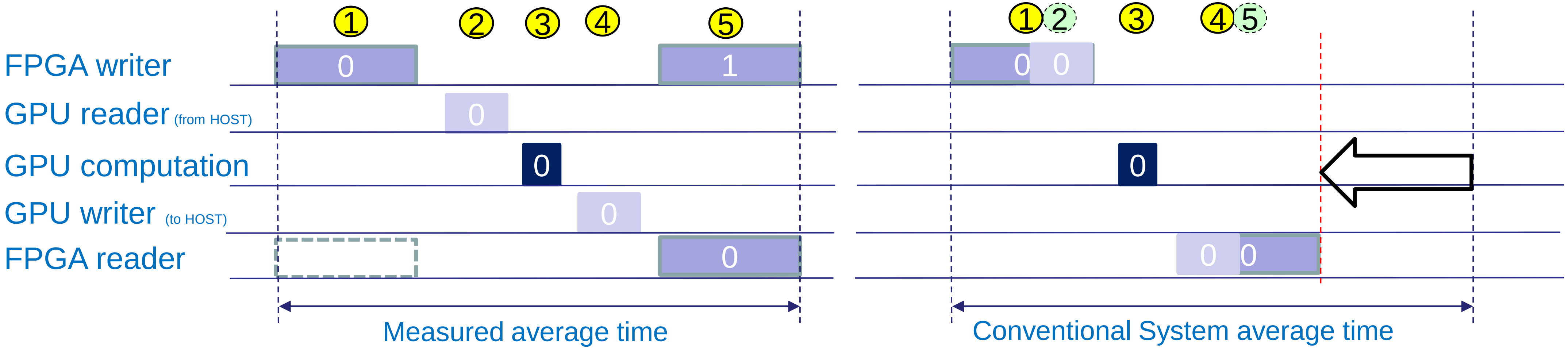
Memory Allocation:

- Unified memory mode
- Store within GPU memory space

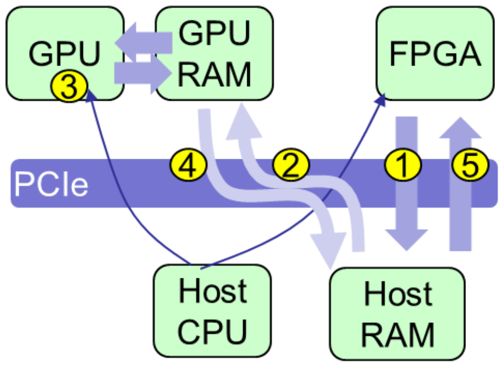
Advantage:

- No redundant copy at host memory
- Speedup with NVLink and CAPI/OC

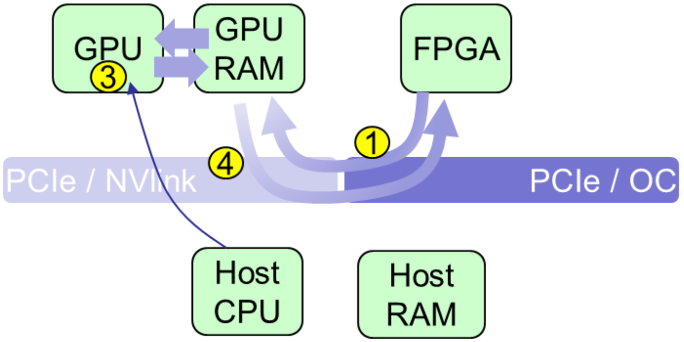




Conventional system



CAPI/OpenCAPI system



```
cudaDeviceGetAttribute(xxx);
cudaMemcpy(ibuff,bufferA, size, cudaMemcpyDeviceToHost);
// obuff = 2*ibuff
vector_add<<<4*numBlocks,numThreadsPerBlock>>>(ibuff,obuff,vector_size);
cudaMemcpy(bufferB, obuff, size, cudaMemcpyHostToDevice);
cudaDeviceSynchronize();
```

```
cudaDeviceGetAttribute(xxx);

// obuff = 2*ibuff
vector_add<<<4*numBlocks,numThreadsPerBlock>>>(ibuff,obuff,vector_size);

cudaDeviceSynchronize();
```

Using Host-GPU unified memory:

- **Double the bandwidth** and **cut by 2 the latency**
- Not depending on GPU interface used ➔ no reprogramming needed

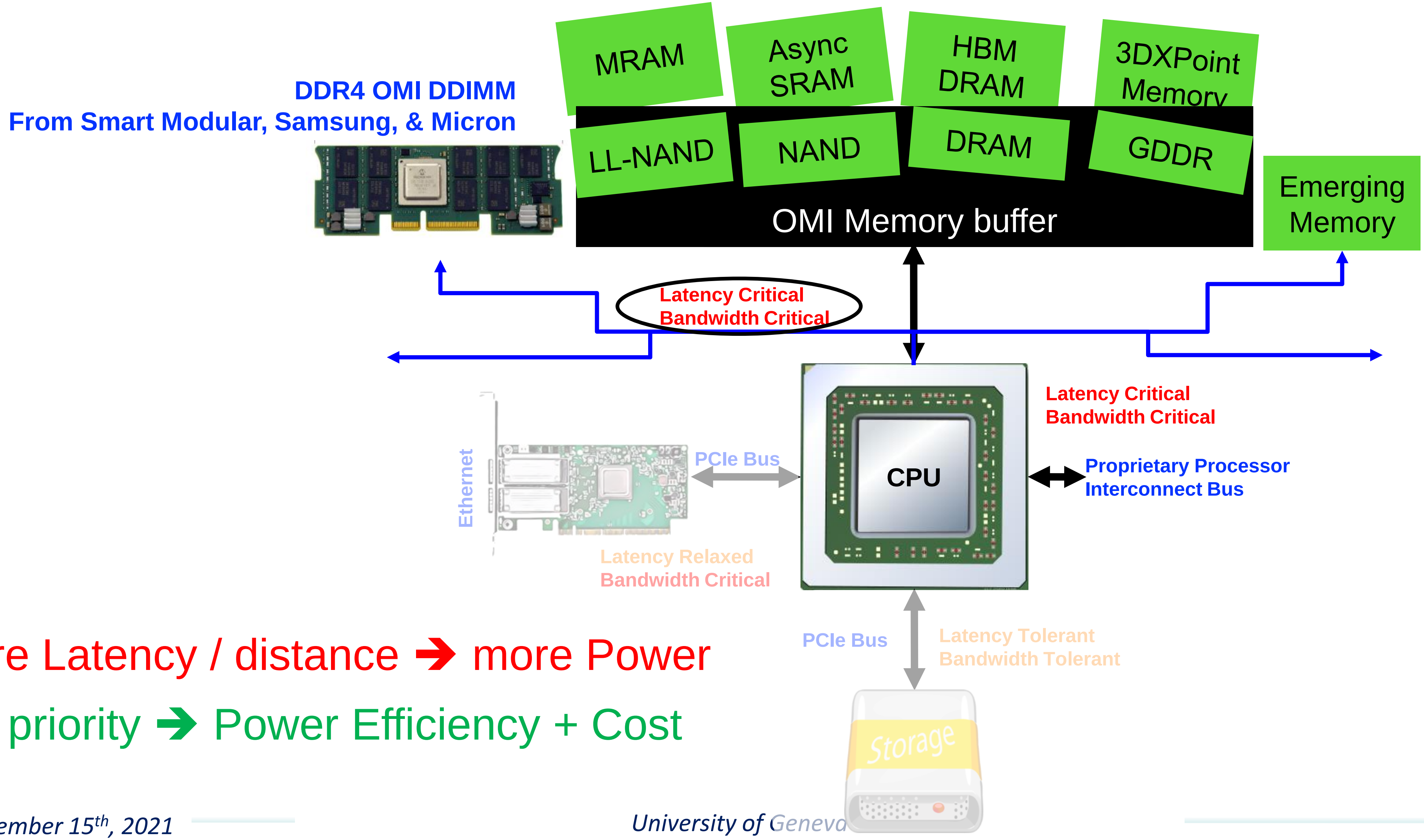
1. Adapters
2. Hardware Accelerators
   ➔ FPGA-GPU combination
3. Host memory

1 Adapters

2 Hardware Accelerators

3 Host memory

- → OMI: New memories around a universal bus
- → Work with pools of memories



DDR4 OMI DDIMM
From Smart Modular, Samsung, & Micron

MRAM

Async SRAM

HBM DRAM

3DXPoint Memory

LL-NAND

NAND

DRAM

GDDR

OMI Memory buffer

Emerging Memory

Latency Critical
Bandwidth Critical

Ethernet

PCIe Bus

Latency Relaxed
Bandwidth Critical

CPU

Latency Critical
Bandwidth Critical

Proprietary Processor
Interconnect Bus

PCIe Bus

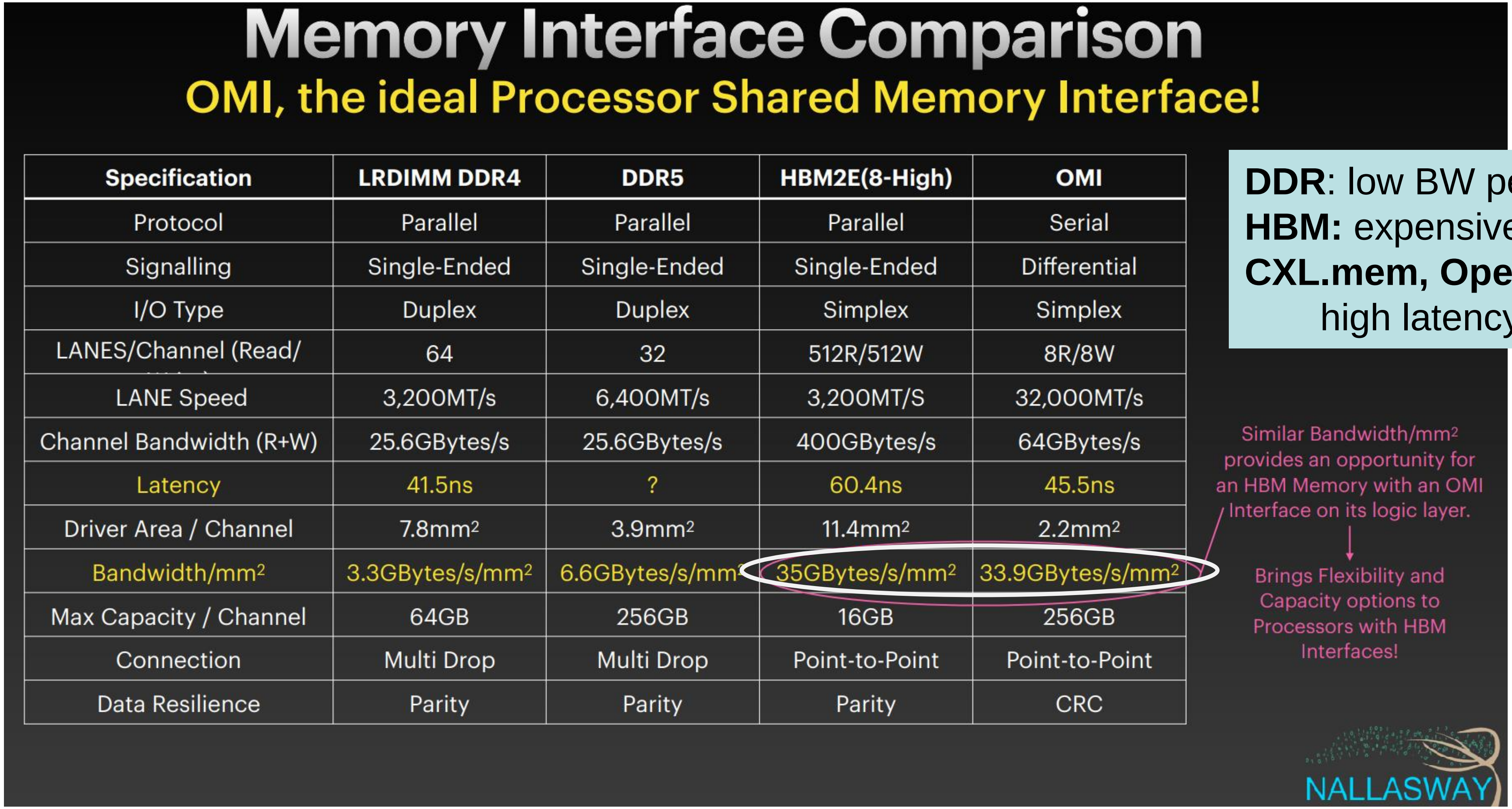
Latency Tolerant
Bandwidth Tolerant

Storage

More Latency / distance ➔ more Power

1st priority ➔ Power Efficiency + Cost

OMI = bandwidth of HBM at DDR latency, Capacity and Cost



Memory Interface Comparison

OMI, the ideal Processor Shared Memory Interface!

Specification	LRDIMM DDR4	DDR5	HBM2E(8-High)	OMI
Protocol	Parallel	Parallel	Parallel	Serial
Signalling	Single-Ended	Single-Ended	Single-Ended	Differential
I/O Type	Duplex	Duplex	Simplex	Simplex
LANES/Channel (Read/	64	32	512R/512W	8R/8W
LANE Speed	3,200MT/s	6,400MT/s	3,200MT/S	32,000MT/s
Channel Bandwidth (R+W)	25.6GBytes/s	25.6GBytes/s	400GBytes/s	64GBytes/s
Latency	41.5ns	?	60.4ns	45.5ns
Driver Area / Channel	7.8mm^2	3.9mm^2	11.4mm^2	2.2mm^2
Bandwidth/mm^2	3.3GBytes/s/mm^2	6.6GBytes/s/mm	35GBytes/s/mm^2	33.9GBytes/s/mm^2
Max Capacity / Channel	64GB	256GB	16GB	256GB
Connection	Multi Drop	Multi Drop	Point-to-Point	Point-to-Point
Data Resilience	Parity	Parity	Parity	CRC

Similar Bandwidth/mm^2 provides an opportunity for an HBM Memory with an OMI Interface on its logic layer.

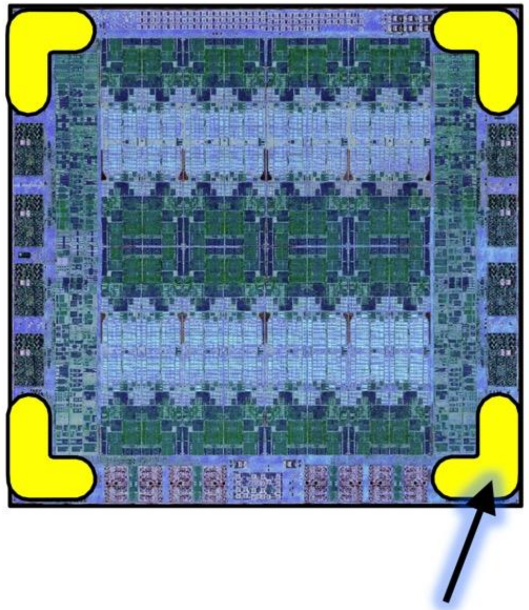
↓

Brings Flexibility and Capacity options to Processors with HBM Interfaces!

DDR: low BW per Die/Area
HBM: expensive + capacity limited
CXL.mem, OpenCAPI, CCIX, GenZ:
high latency, far memory

Allan Cantle's full presentation at https://youtu.be/c0DuGSwDpqY

System Composability: PowerAXON & Open Memory Interfaces



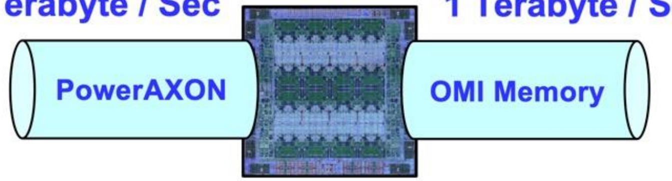
PowerAXON corner
4x8 @ 32 GT/s

Multi-protocol
“Swiss-army-knife”
Flexible / Modular Interfaces

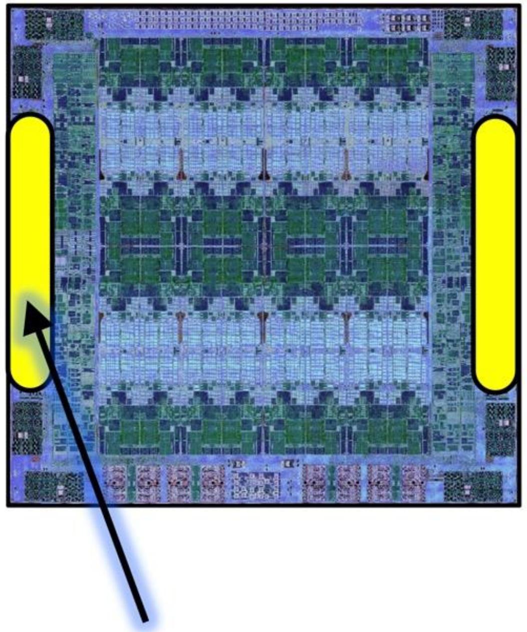
POWER10
Chip

1 Terabyte / Sec

1 Terabyte / Sec



Built on best-of-breed
Low Power, Low Latency,
High Bandwidth
Signaling Technology



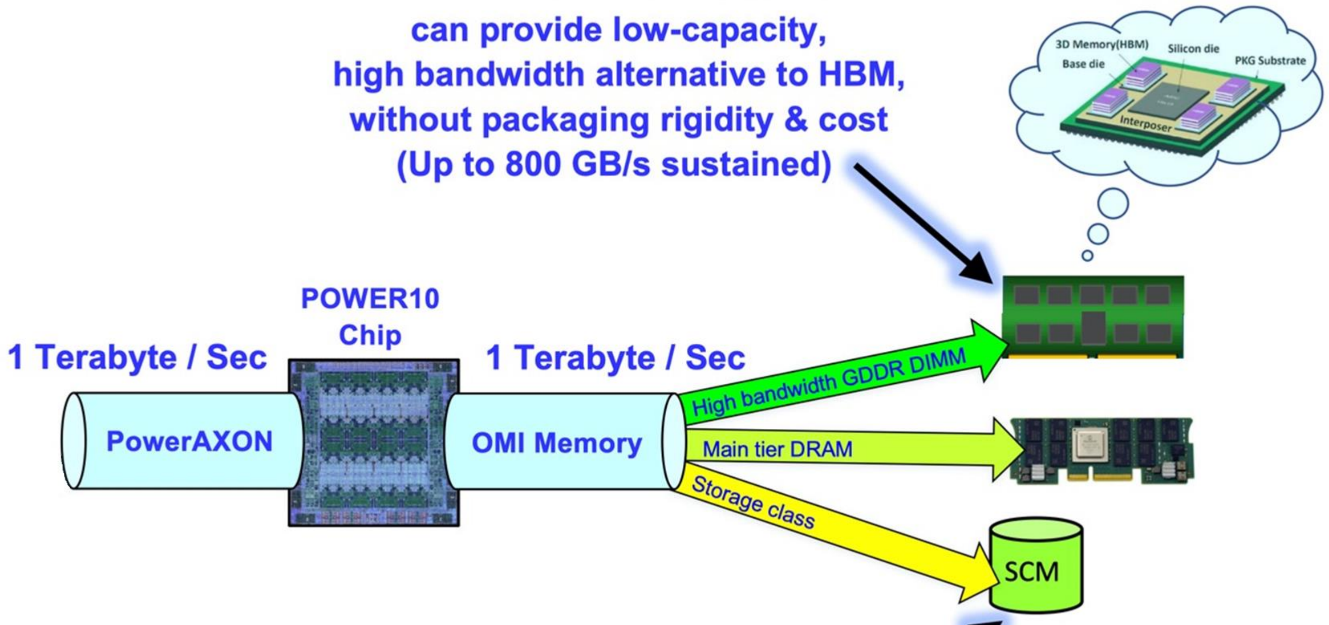
OMI edge
8x8 @ 32 GT/s

6x bandwidth / mm^2
compared to DDR4
signaling

IBM POWER10

Data Plane Bandwidth and Capacity: Open Memory Interfaces

OMI-attached GDDR DIMMs
can provide low-capacity,
high bandwidth alternative to HBM,
without packaging rigidity & cost
(Up to 800 GB/s sustained)



OMI-attached storage class memory
can provide high-capacity, encrypted,
persistent memory in a DIMM slot.
(POWER10 systems can support 2 petabytes
of addressable load/store memory)

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

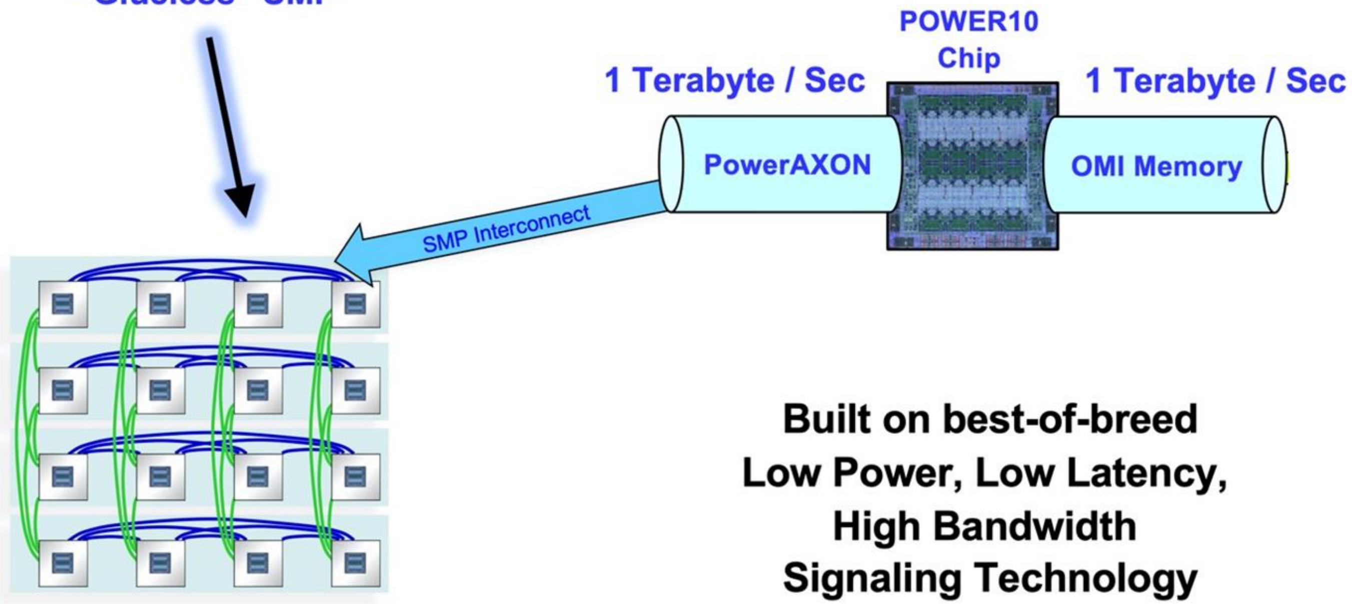
IBM POWER10

System Enterprise Scale and Bandwidth: SMP & Main Memory

Multi-protocol
“Swiss-army-knife”
Flexible / Modular Interfaces

Build up to 16 SCM socket
Robustly Scalable
High Bisection Bandwidth
“Glueless” SMP

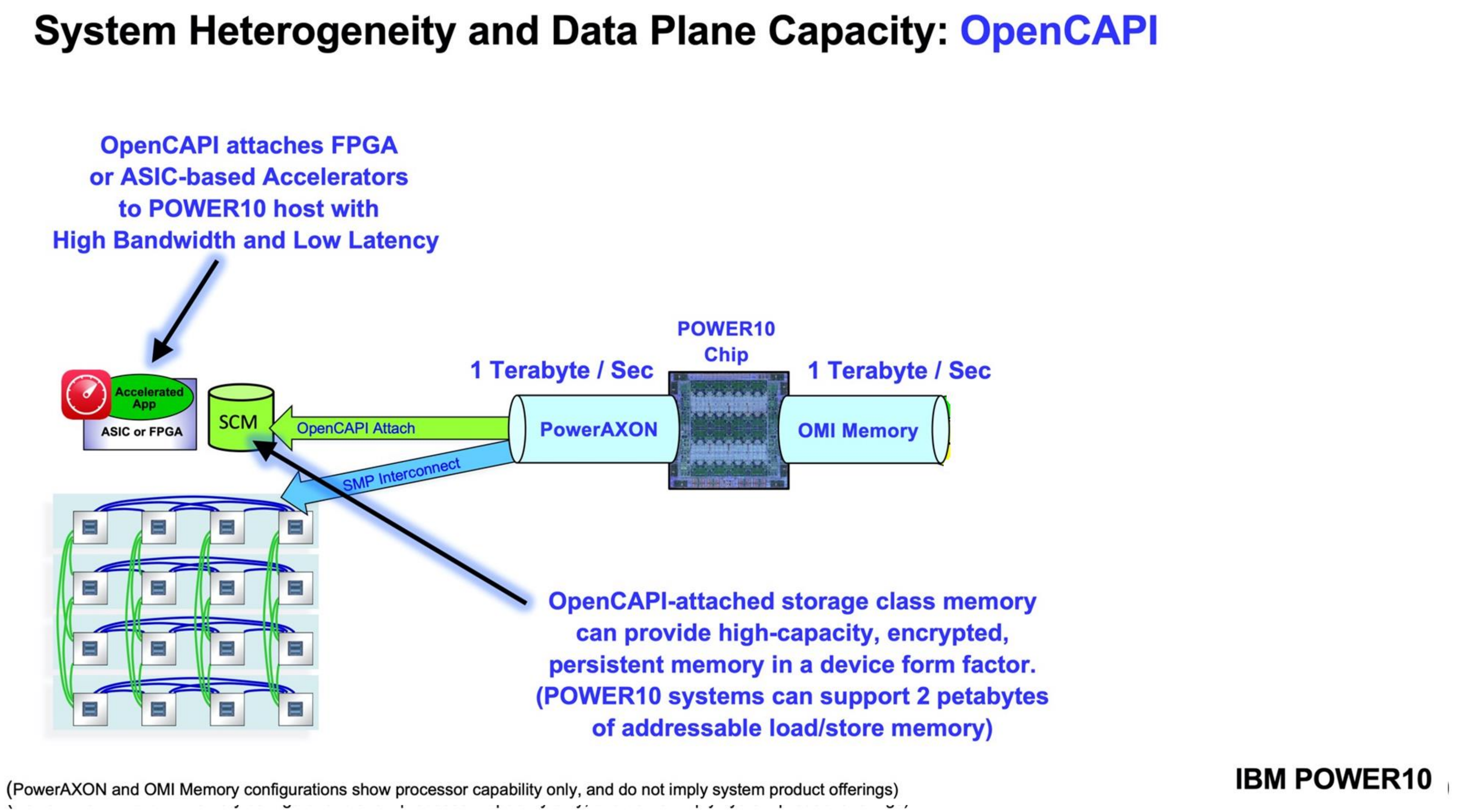
Allocate the bandwidth
however you need to use it



Built on best-of-breed
Low Power, Low Latency,
High Bandwidth
Signaling Technology

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

IBM POWER10



System Heterogeneity and Data Plane Capacity: OpenCAPI

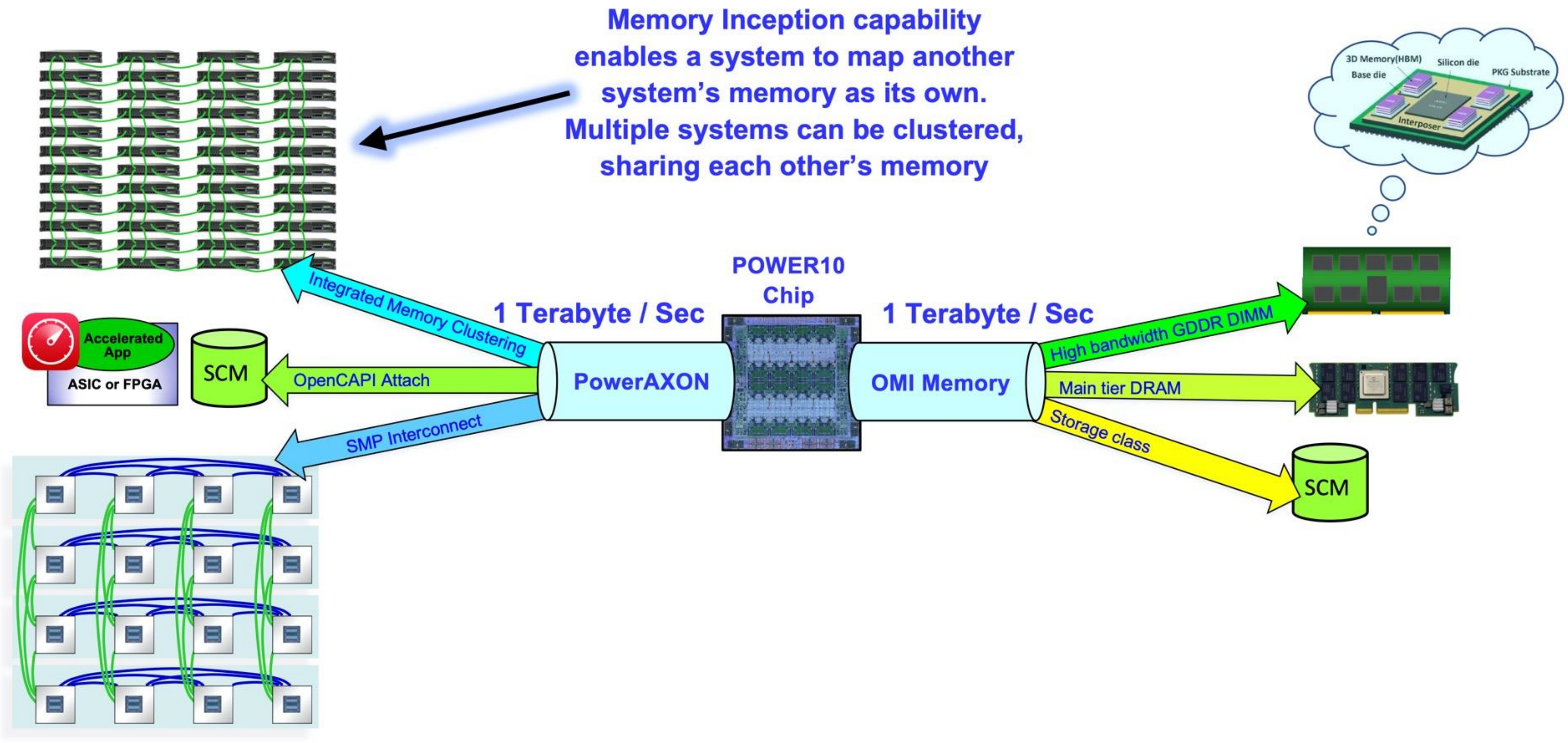
OpenCAPI attaches FPGA or ASIC-based Accelerators to POWER10 host with High Bandwidth and Low Latency

OpenCAPI-attached storage class memory can provide high-capacity, encrypted, persistent memory in a device form factor. (POWER10 systems can support 2 petabytes of addressable load/store memory)

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

IBM POWER10

Pod Composability: PowerAXON Memory Clustering



Memory Inception capability enables a system to map another system's memory as its own. Multiple systems can be clustered, sharing each other's memory

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

IBM POWER10

Memory Clustering: Distributed Memory Disaggregation and Sharing

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

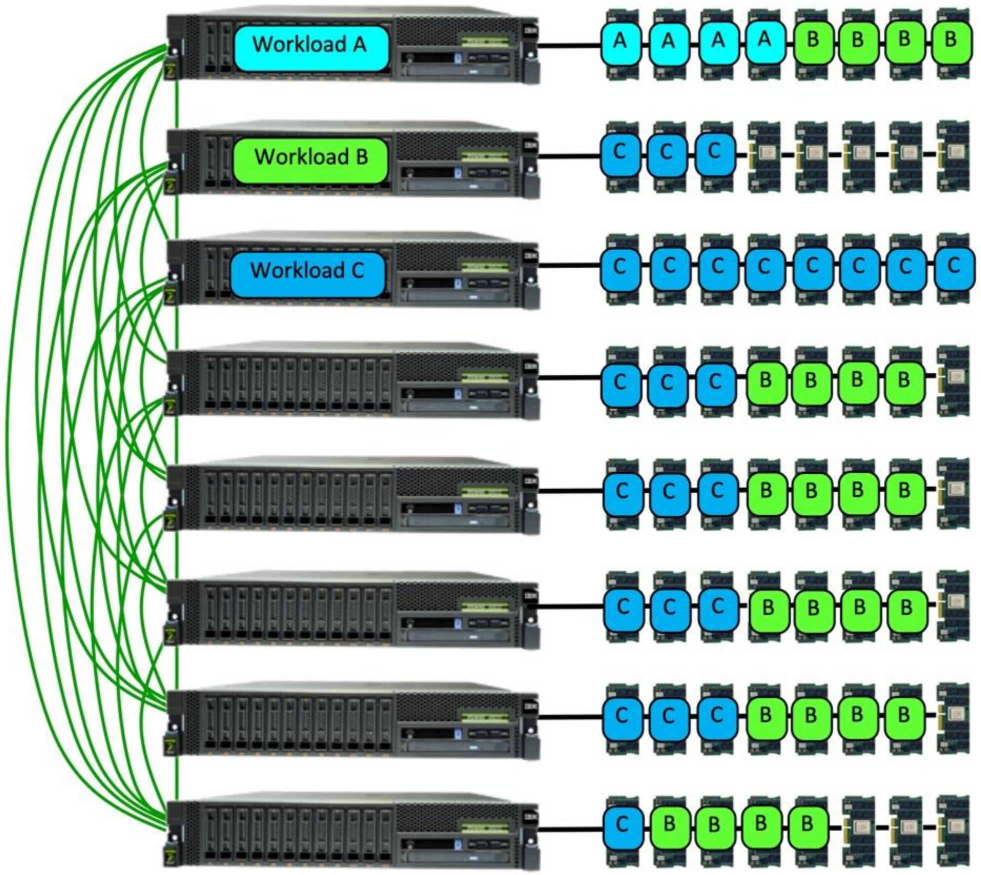
Workload A Rqmt: 4 TB low latency

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



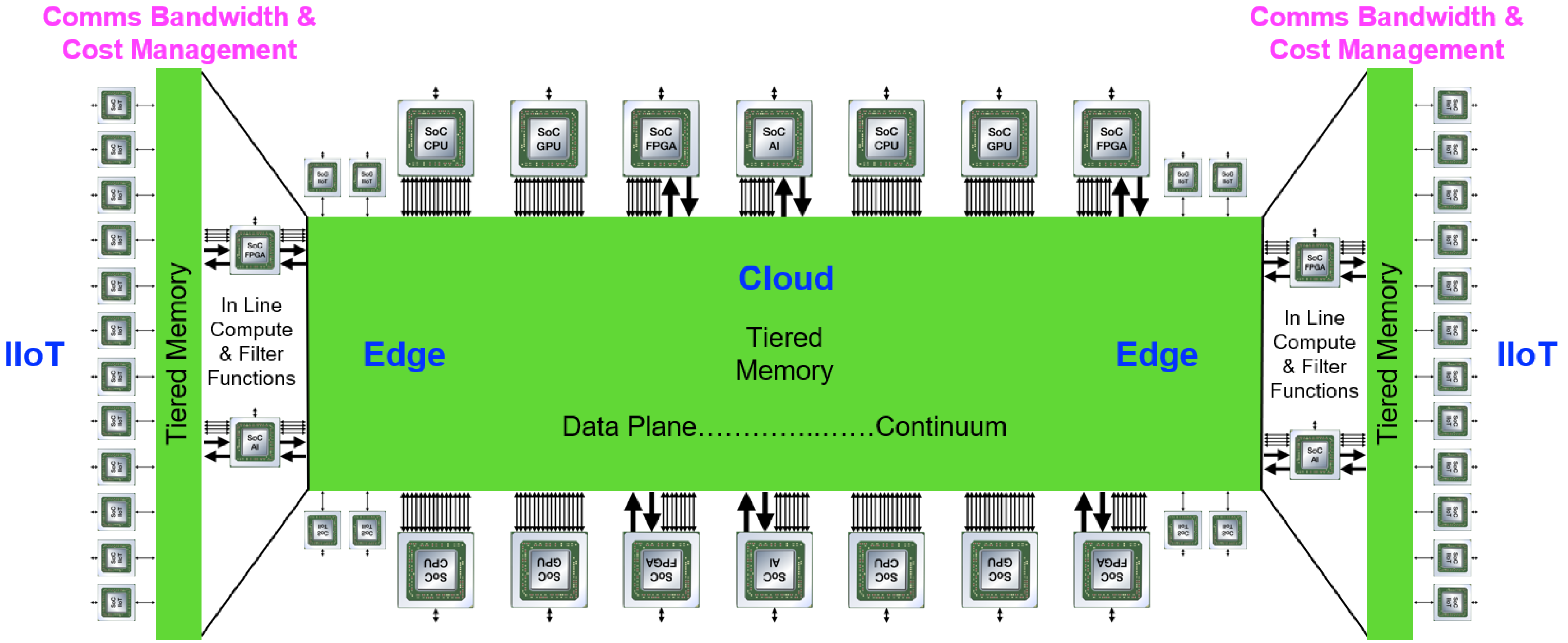
(Memory cluster configurations show processor capability only, and do not imply system product offerings)

IBM POWER10

Think big – memory at a system level architecture

Data Centric

Generic System Level "Domain Specific Architecture"



Comms Bandwidth & Cost Management

Comms Bandwidth & Cost Management

IIoT

Tiered Memory

In Line Compute & Filter Functions

Edge

Cloud

Tiered Memory

Data Plane……………..……Continuum

Edge

In Line Compute & Filter Functions

Tiered Memory

IIoT

I WANT MORE!

- *Know more about accelerators ?*
- *See a live demonstration?*
- *Access to a server?*
- *Do a benchmark ?*
- *Get answers to your questions?*

Paul Scherrer Institute : filip.leonarski@psi.ch

IBM local partner: **lclavien@inno-boost.com**

IBM OpenCAPI team :
alexandre.castellane@fr.ibm.com - bruno.mesnet@fr.ibm.com - fabrice_moyen@fr.ibm.com

OpenCAPI Repository: https://github.com/OpenCAPI/oc-accel

More about decoupling compute with OMI: https://youtu.be/c0DuGSwDpqY or https://openmemoryinterface.org/

OpenCAPI™

IBM