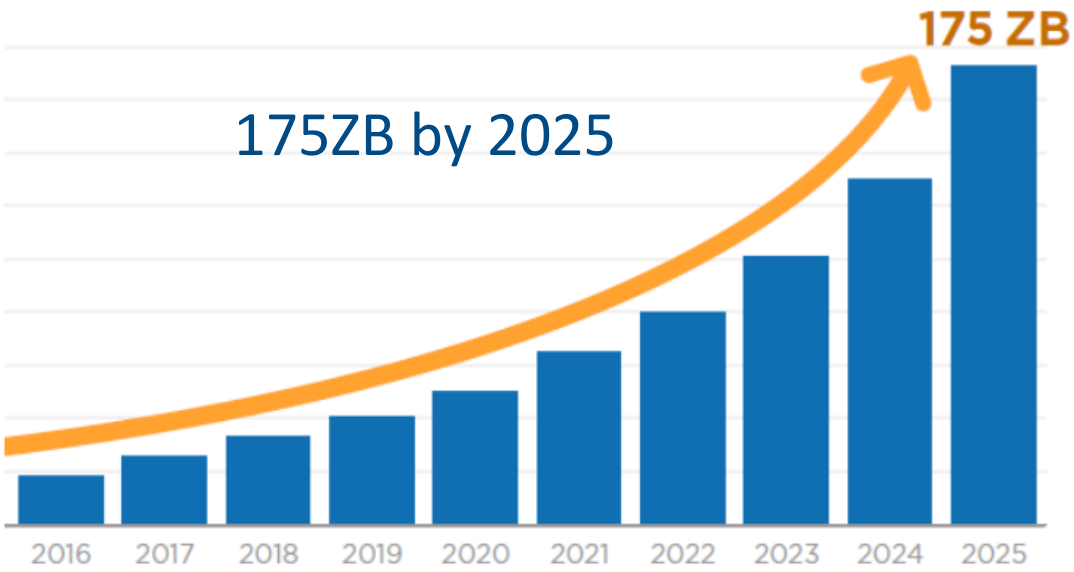
High Performance Chiplet Workshop @ISCA 2022

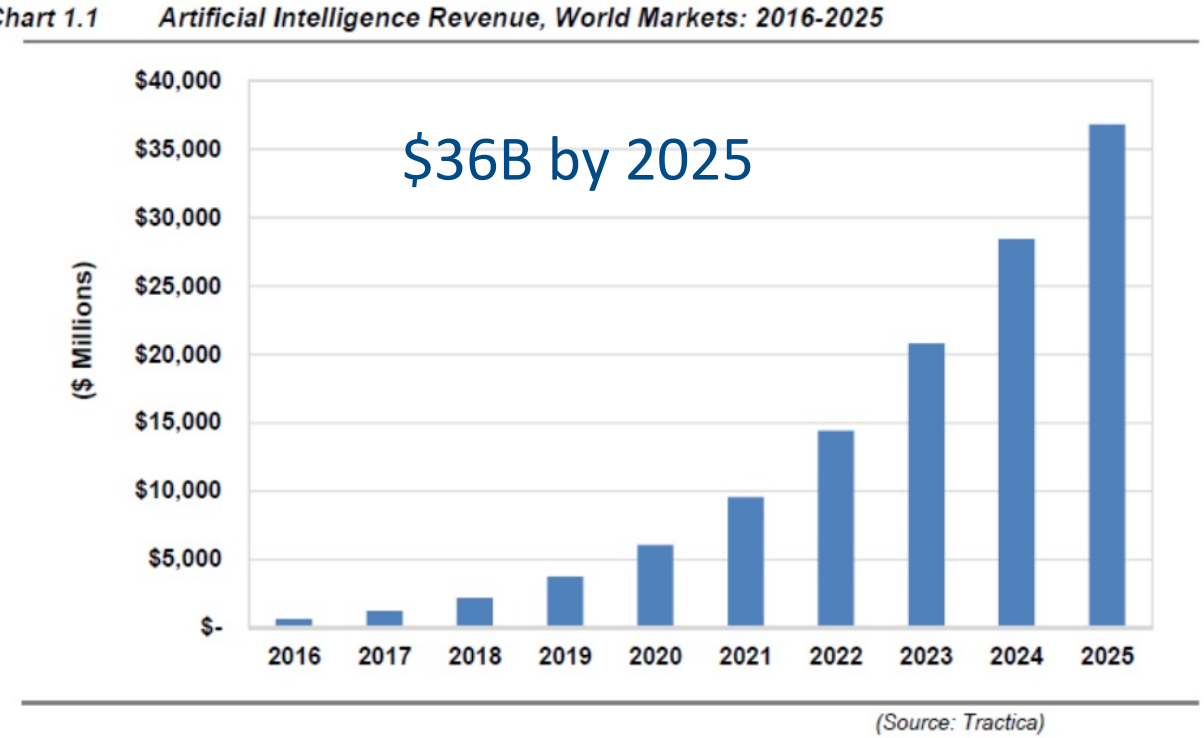
AI & HPC system opportunity with integrated photonics chiplets

Edi Roytman, Ajaya Durg, Thomas Liljeberg, Ling Liao, Robert Munoz – all from Intel Corporation.

Explosion of Data. AI is essential to analyze it.



IDC says 175 ZB will be created by 2025 (Image courtesy IDC)

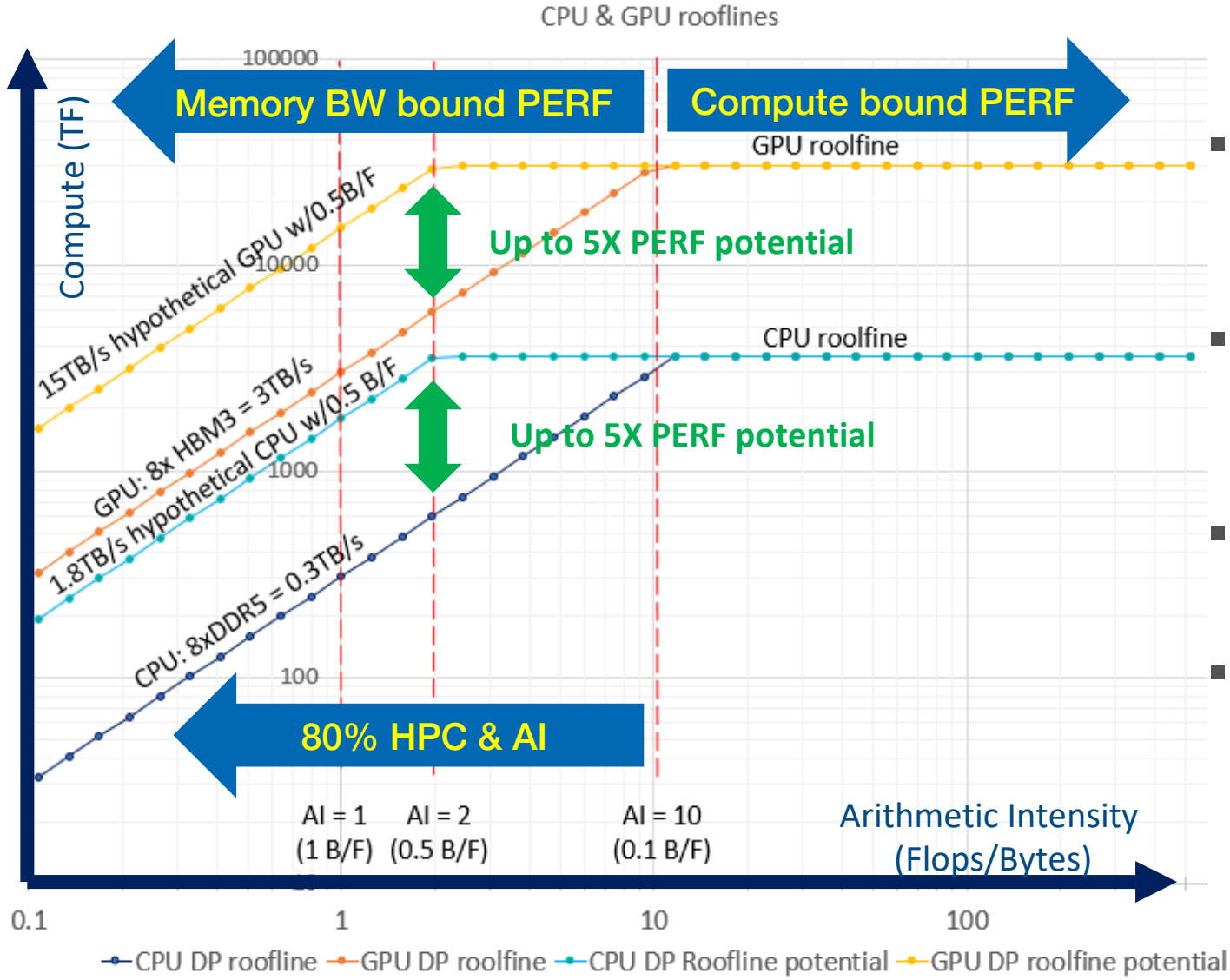


Strong correlation between Data and AI is no coincidence

AI is an essential technology to filter, analyze and translate data into insights

https://www.datanami.com/2022/01/11/big-growth-forecasted-for-big-data/
https://www.top500.org/news/market-for-artificial-intelligence-projected-to-hit-36-billion-by-2025/

AI & HPC Workloads: Memory BW perspective



- 80% of HPC and AI algorithms require *many* bytes of data to retire *one* instruction, aka low arithmetic intensity.
- CPUs & GPUs differ in Compute and Memory BW, yet exhibit similar rooflines with only ~0.1 Bytes/Flop (AI=10)
- PERF of HPC and AI is left on the table if Memory BW is not balanced vs. Compute.
- Estimated 5X PERF potential in AI & HPC if Memory BW is improved to 0.5B/F (AI=2) or higher.

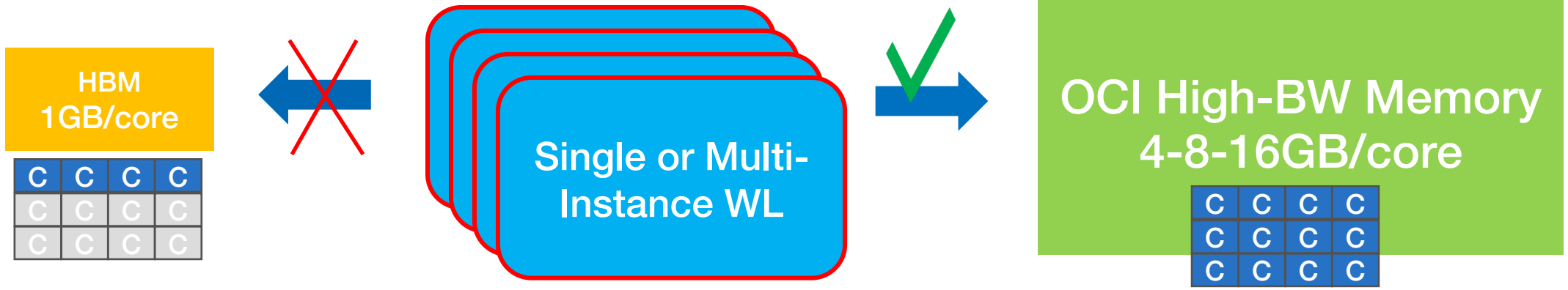
Can HBM solve a BW problem? It depends…

- **CPU perspective**
  - Moving CPU from DDR to HBM can bring CPU close to 0.5B/F point.
- **GPU perspective**
  - GPU with HBM is a baseline to get 0.1B/F…. no practical B/F upside potential.
- **Common HBM challenges:**
  - 6-8 HBM3 stacks take >600-800mm^2 of SoC area and ~120-160W of SoC power.
  - Limited HBM capacity can impact PERF if it results in excessive scale-out.
  - High cost per GB: HBM is 2X more expensive than DDR5

Why PERF moves w.BW AND Capacity?

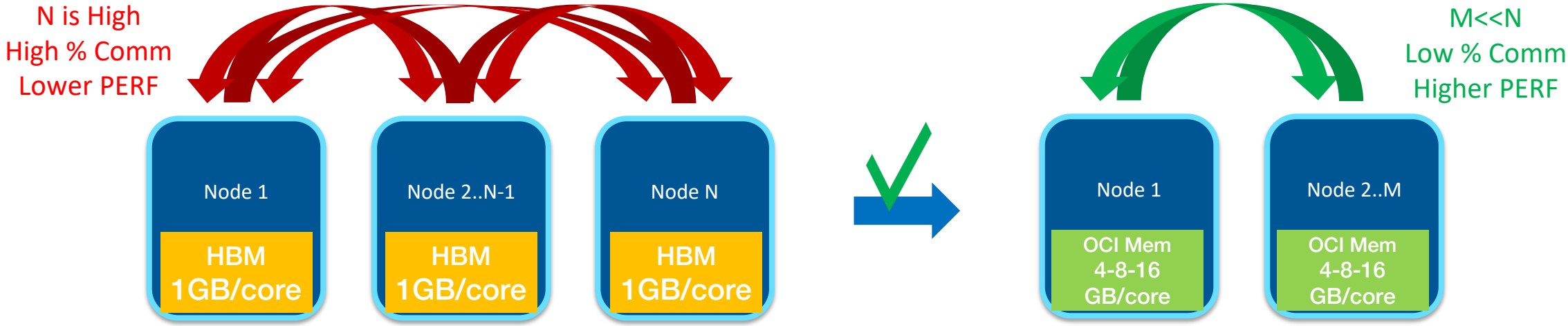
Node Throughput

1 BW-bound Single or Multi-Instance WL **fits in OCI Mem capacity** full or tiered, all CC in use



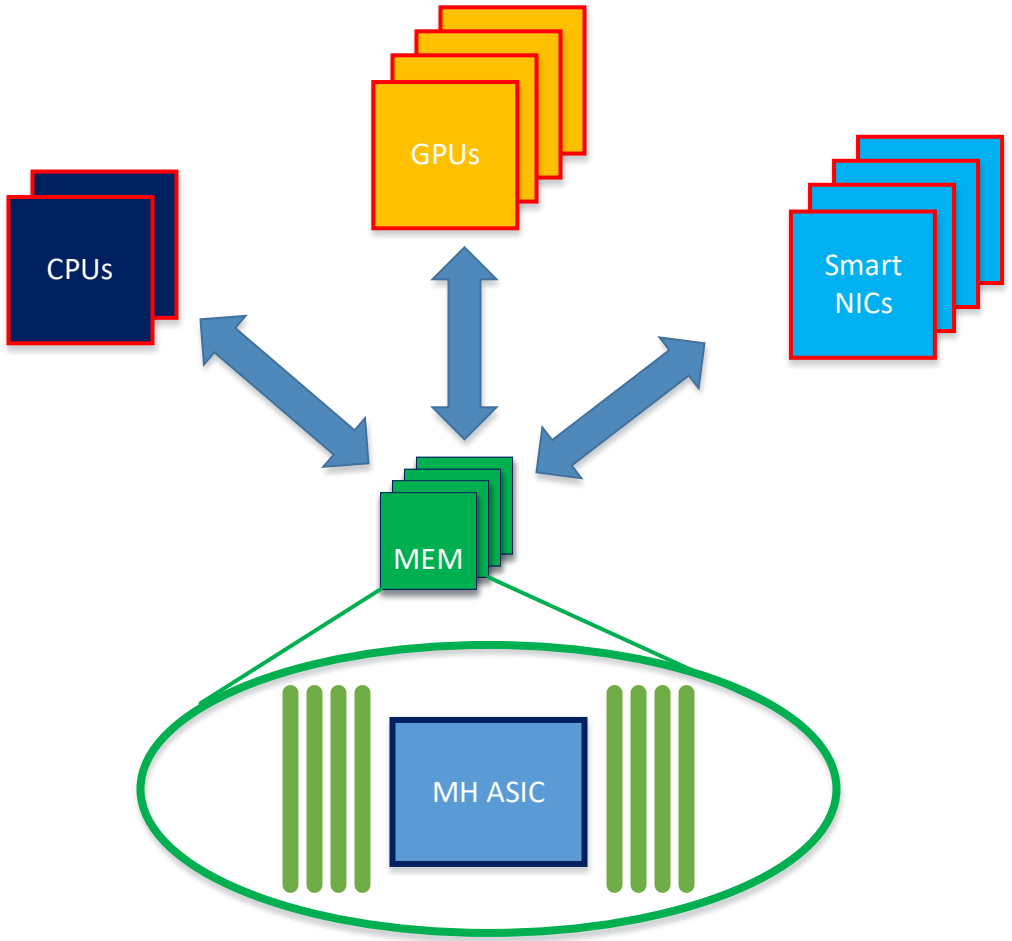
System Throughput

2 BW-bound All-to-All Comm, Sparse, Big problems **PERF-benefit from fat node**



An ideal system memory for AI/HPC node….

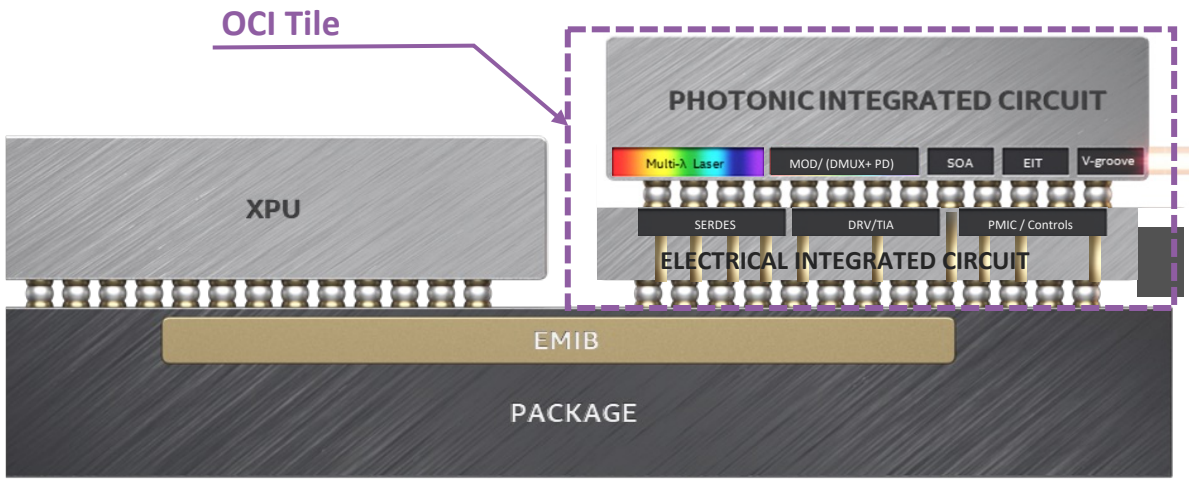
- ✓Directly accessible by all compute and communication types
- ✓Modular, Composable and Scalable
- ✓Shareable and pool(able).
- ✓HBM-like BW
- ✓DDR-like Capacity, Latency and ECC
- ✓LPDDR-like energy efficiency

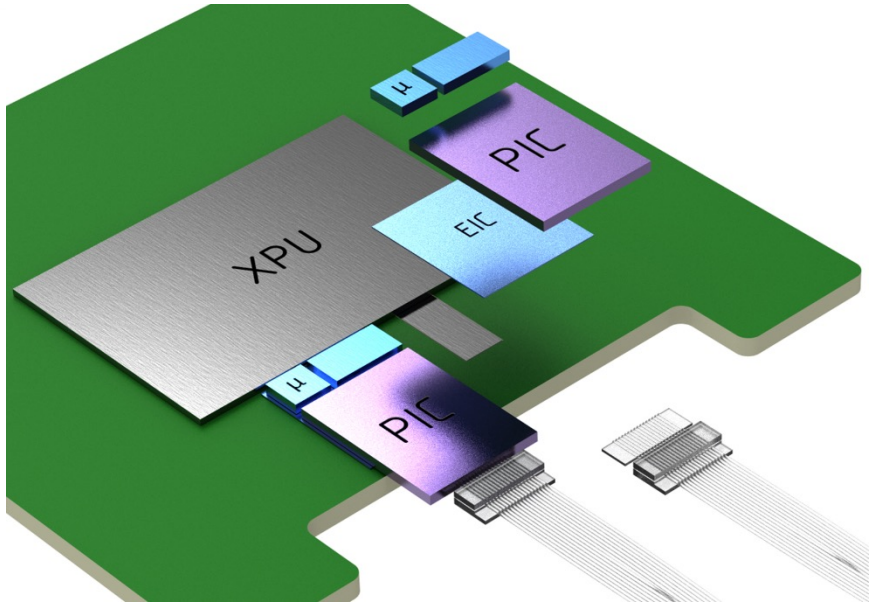


+ Disaggregation unlocks many degrees of innovation freedom

– Disaggregation causes higher latency vs. integrated memory

Intel Optical Compute Interconnect Vision





Ultra-high bandwidth

~1Tbps per fiber (equivalent to 1 UXI Link)

Reach

>100m, orders of magnitude better than electrical

Shoreline Density

>4x improvement over PCIe6

Energy Efficiency

3pJ/b (>30% better than PCIe6)

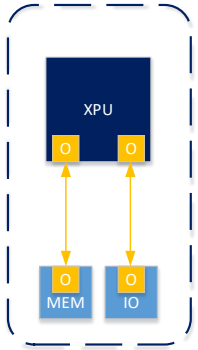
Latency

<10ns + TOF, comparable to electrical

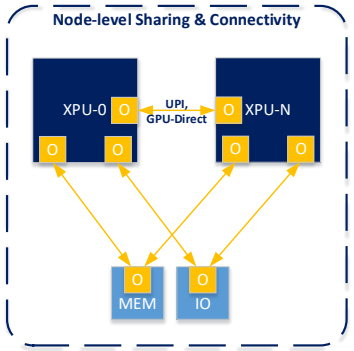
Use Cases

compute scale-up, resource pooling

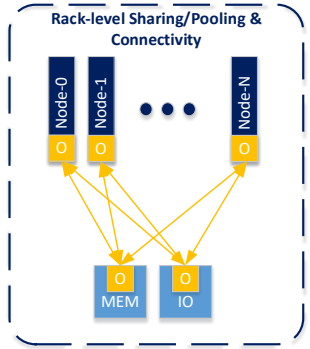
OCI Opportunities for System Architecture



Dedicated Memory

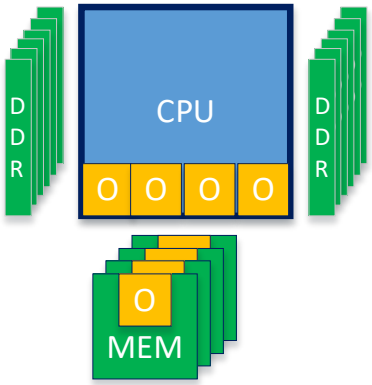


Common Node Memory & XPU IO

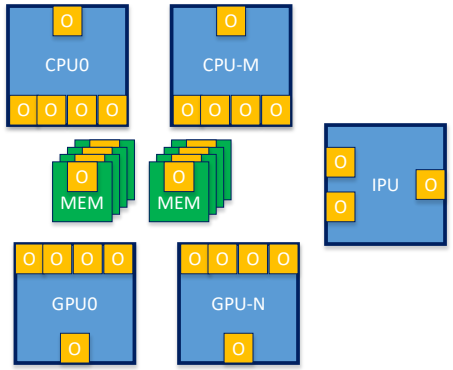


Shared/Pooled Memory + IO

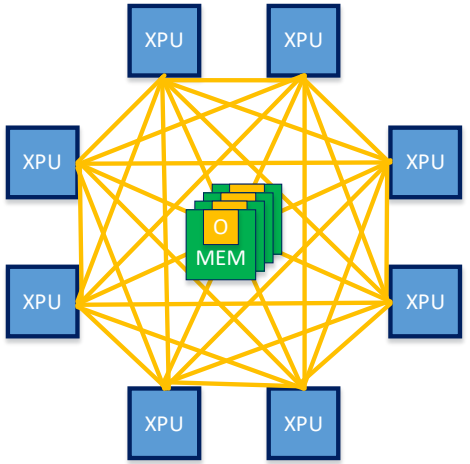
DDR augmentation
HBM replacement



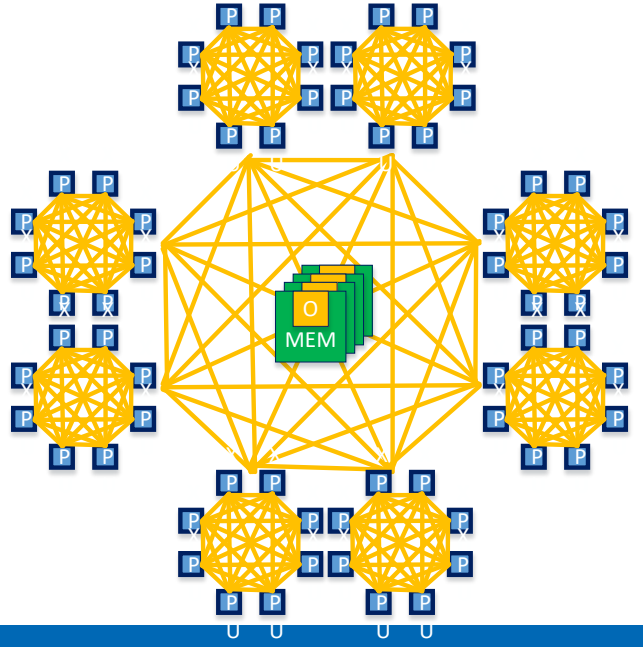
Central Memory for all types of Compute and Communication



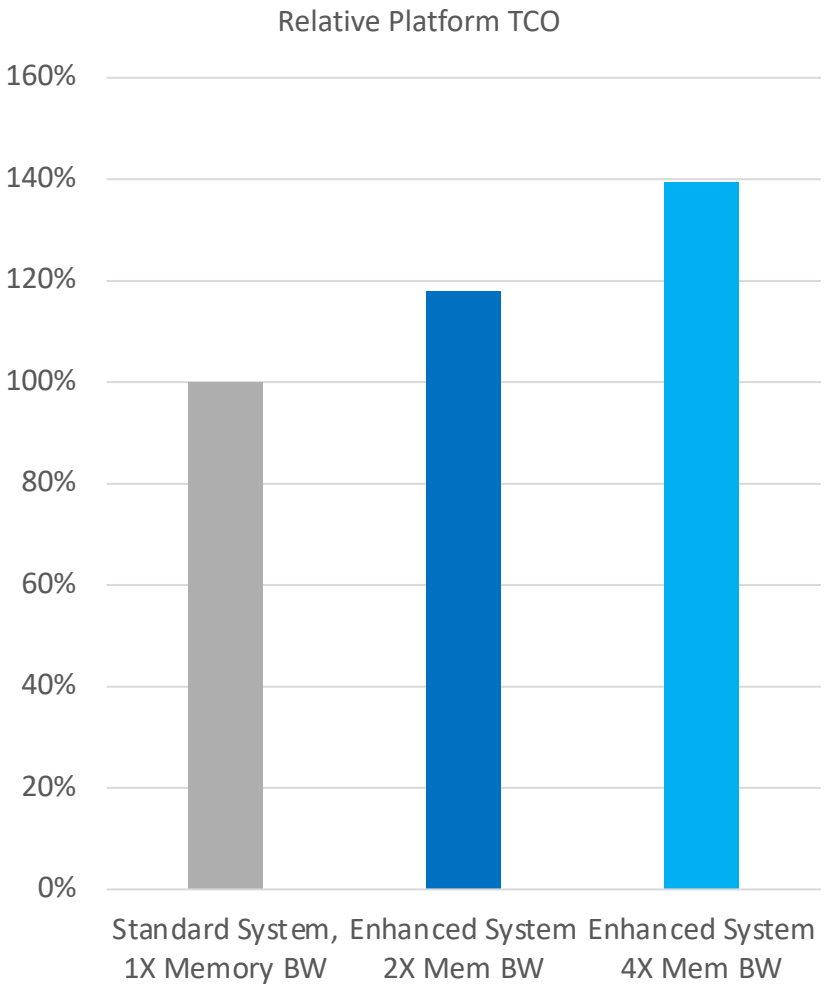
Scale-up Hetero node
Central Node Memory
Near-memory Compute?

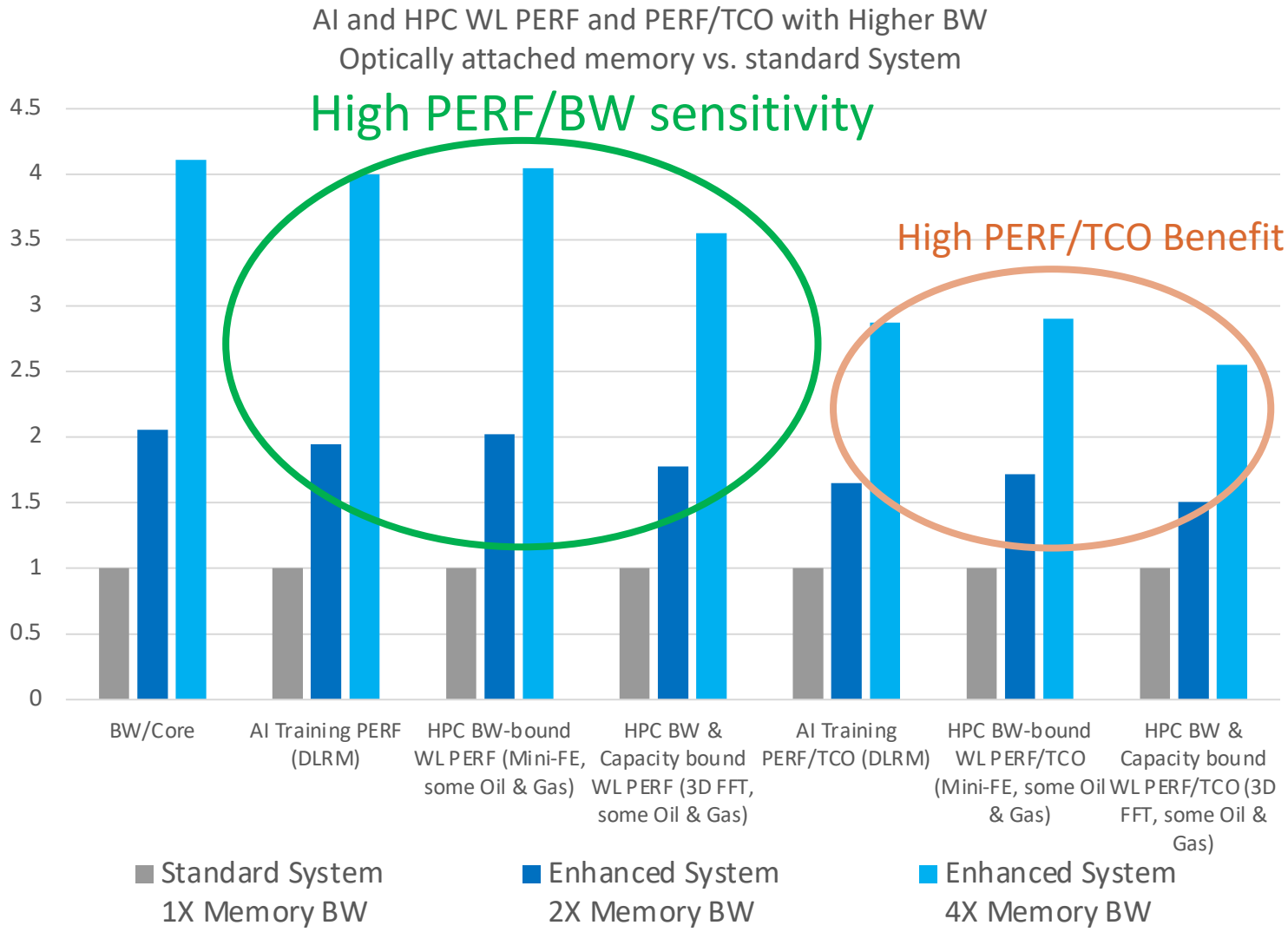


High Radix Switchless Cluster
Pooled/Shared memory
Near-Memory Compute?



AI & HPC workloads – Memory POV and PERF/TCO potential





Memory Performance of AI & HPC workloads increases >2.5X faster than system TCO

Q: Do I really need optics for this?
A: It depends…

CPU node	Gen6 Electrical	OCI Optical
System Config	8CPU node 8CH DDR5 4800	8CPU node 8CH DDR5 4800
Total Mem BW	2.4TB/s/dir	2.4TB/s/dir
Lanes required	300 Gen6 lanes	19 Fibers
SoC shoreline consumed	96mm	24mm
PHY power consumed	96W	48W
Retimes	Yes, 1 or 2	no
System TCO	1X	1X

Hetero node	Gen6 Electrical	OCI Optical
System Config	2CPU:8GPU node 8CH DDR5 4800 8xHBM3/GPU	8S scale-up node 8CH DDR5 4800 8xHBM3/GPU
Total Mem BW	**50TB/s/dir**	**50TB/s/dir**
Lanes required	6250 Gen6 lanes	390 Fibers
SoC shoreline consumed	1952mm	488mm
PHY power consumed	2000W	1000W
Retimes	Yes, 1 or 2	no
System TCO	1X	1X

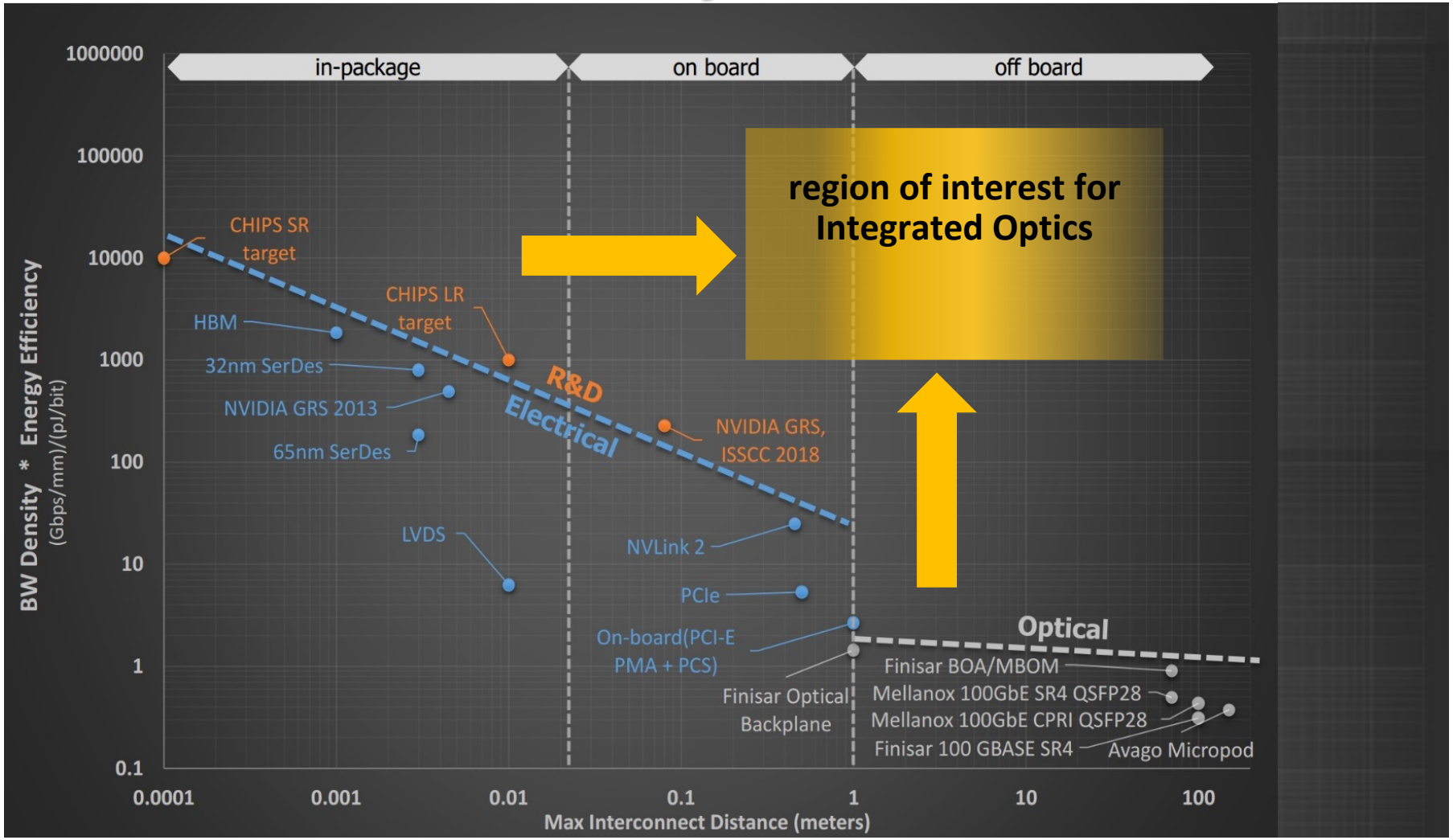
Copper IO is good enough

Integrated Optical IO is required

AI and HPC Compute, BW and Scale-up/out requirements create a pull for a feasible and affordable Integrated Optical IO technology.

BW Density * Energy Efficiency OR Reach?

Yes please…



Dr. Gordon Keeler, DARPA ERI Summit 2019

Three Value Vectors of Optical Compute Interconnect - a System Perspective

Performance	Scale	Power & Cost
As we scale disaggregated memory BW, Performance of AI/HPC WL increases 2.5X faster than system TCO does	Electrical implementation of AI/HPC scale-up & scale-out systems is challenged due to signal integrity limits, affecting scale achievable, modularity, latency.	Power saved by integrated optics is reapplied to boost compute.
		Package costs saved translate directly to product margins.
	Integrated optics unlocks SI degree of freedom allowing to solve all of the above with clear path to scale gen-gen.	System level TCO is similar between electrical and optical IO at ISO System size and BW.

Summary and Next Steps

- Growth in World Data is driving investment in AI technologies helping translate it to actionable insights.
- Optically connected High-BW and Right Capacity memory can help unleash 5X PERF potential and 2-3X PERF/TCO potential.
- Call to action – Optical Compute Interface Chiplets:
  - AI/HPC system architectures
    - Workload sensitivity studies
    - OCI-based reference designs
  - Open-standards based interoperability for XPU and Optical interfaces
    - Validation of cross-vendor interoperability, usability of enabling collateral
    - Input to Open-Standards consortia, UCIe for example.


intel