



**THE CHIPS
TO SYSTEMS
CONFERENCE**

SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

The Case for Chiplets for Photonic Resource Disaggregation

John Shalf

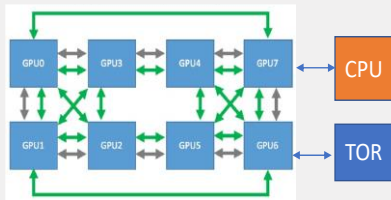
Department Head for Computer Science,
Lawrence Berkeley National Laboratory



Diverse Node Configurations for Diverse Workload Resource Requirements

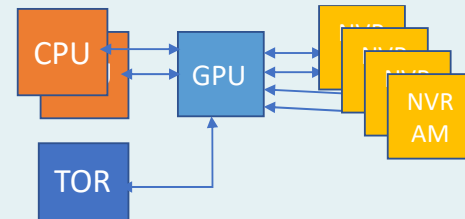
Training

- 8 connections: GPU
- 8 links to HBM (weights)
- 8 links: to NVRAM
- 1 links: to CPU (control)



Data Mining

- 6-links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)



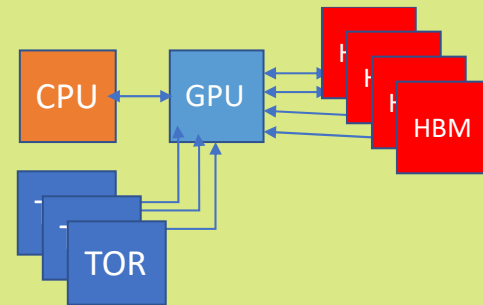
Inference

- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU



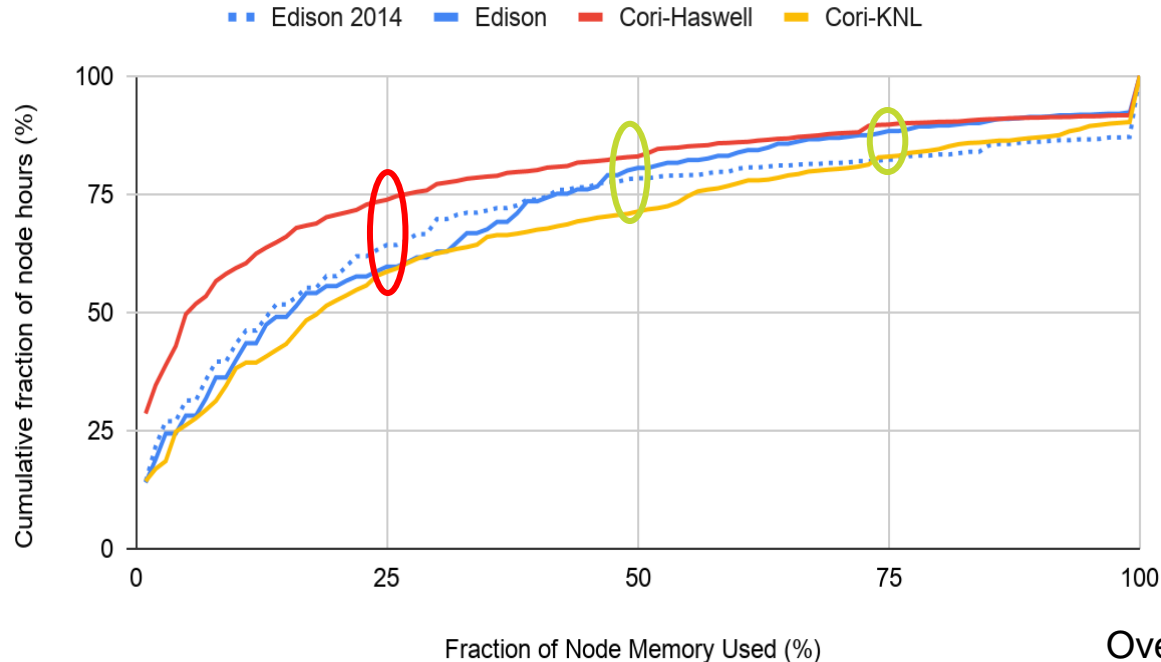
Graph Analytics

- 16 links HBM
- 8 links TOR
- 1 Link CPU



Diverse Capacity Requirements Drive Need for Memory Disaggregation

Memory pressure at NERSC, 2018



About 15% of NERSC workload uses more than 75% of the available memory per node.

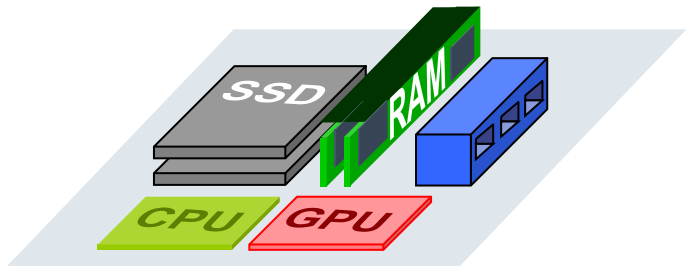
And ~25% uses more than 50% of available memory.

But 75% of Haswell job hours (60% of KNL) use < 25% memory

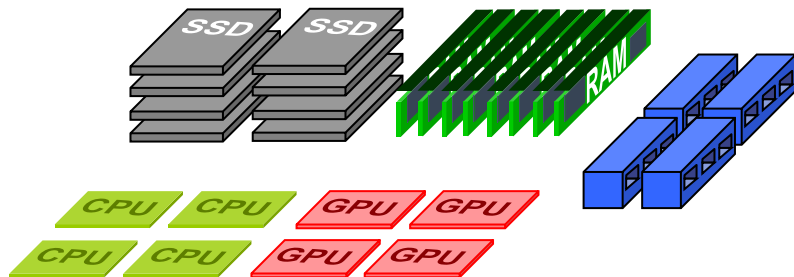
Overestimate: $\text{maxrss} \times \text{ranks_per_node}$
Assumes memory balance across MPI ranks.

Disaggregated Node/Rack Architecture

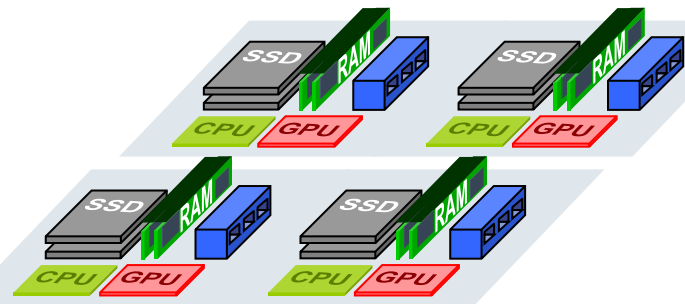
Current server



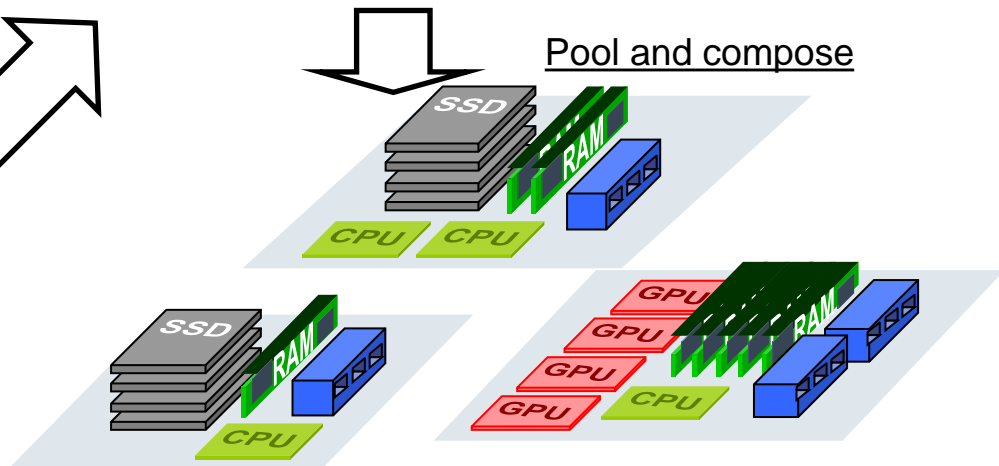
Disaggregated rack



Current rack

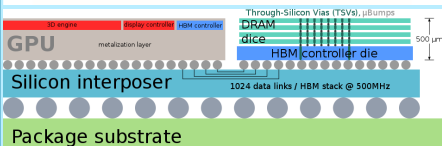
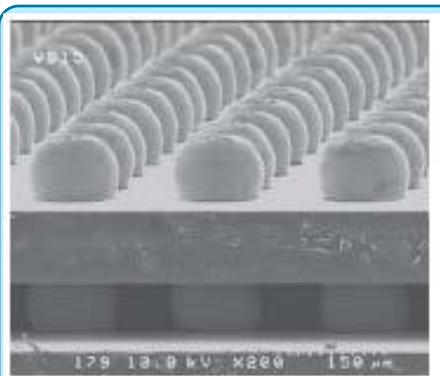
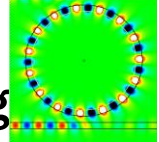


Pool and compose



Most solutions current disaggregation solutions use Interconnect bandwidth (1 – 10 GB/s)
But this is significantly inferior to RAM bandwidth (100 GB/s – 1 TB/s)

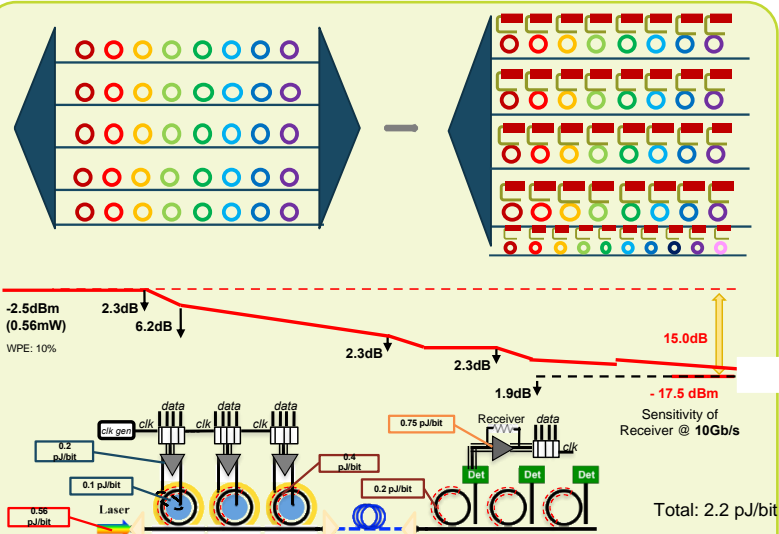
Co-Packaged Optics: Chiplets Impedance Matching to Packaging Technology



In-package integration

Solder Microbumps
& Copper Pillars @ ~10Gbps

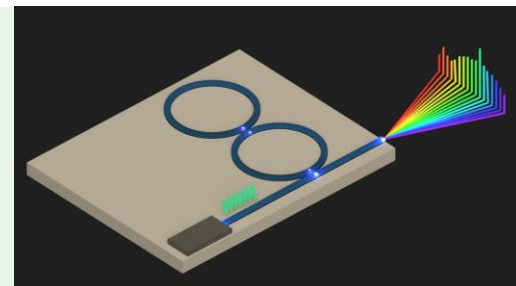
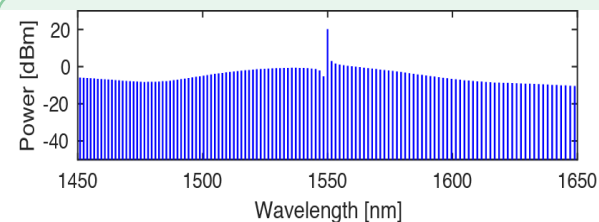
Wide and Slow!



DWDM Using Silicon Photonics

Ring Resonators @ ~10-25 Gb/sec per chan
Many channels to get bandwidth density

Wide and Slow!

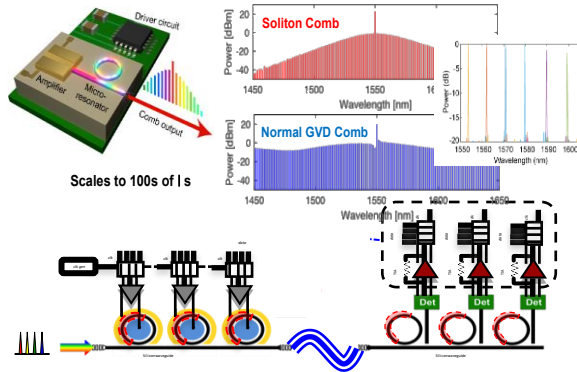


Comb Laser Sources

Single laser to efficiently
generate 100s of frequencies

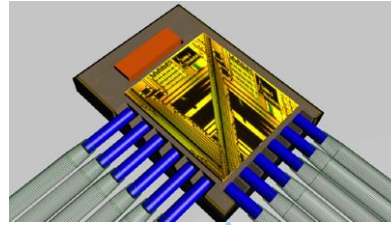
Wide and Slow!

Photonic MCM (Co-Packaged Optics)

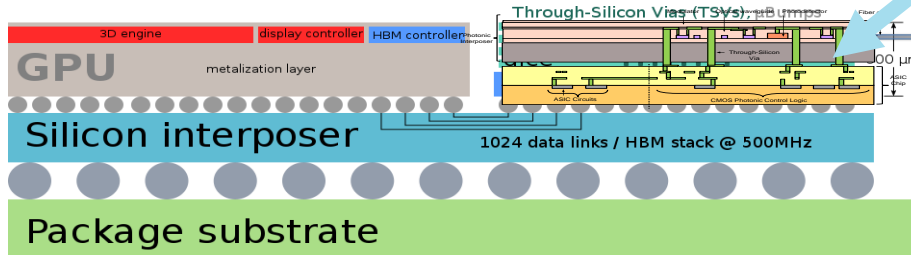
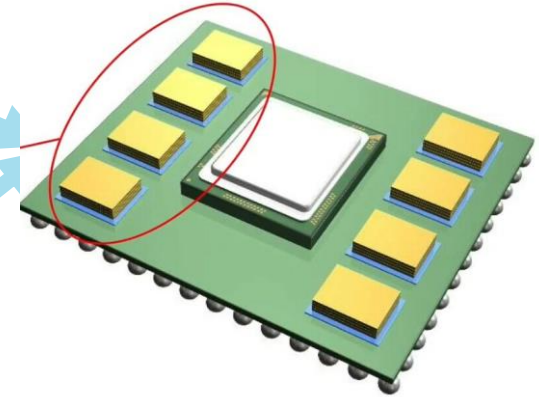


Comb Laser Source with
DWDM Silicon Photonics

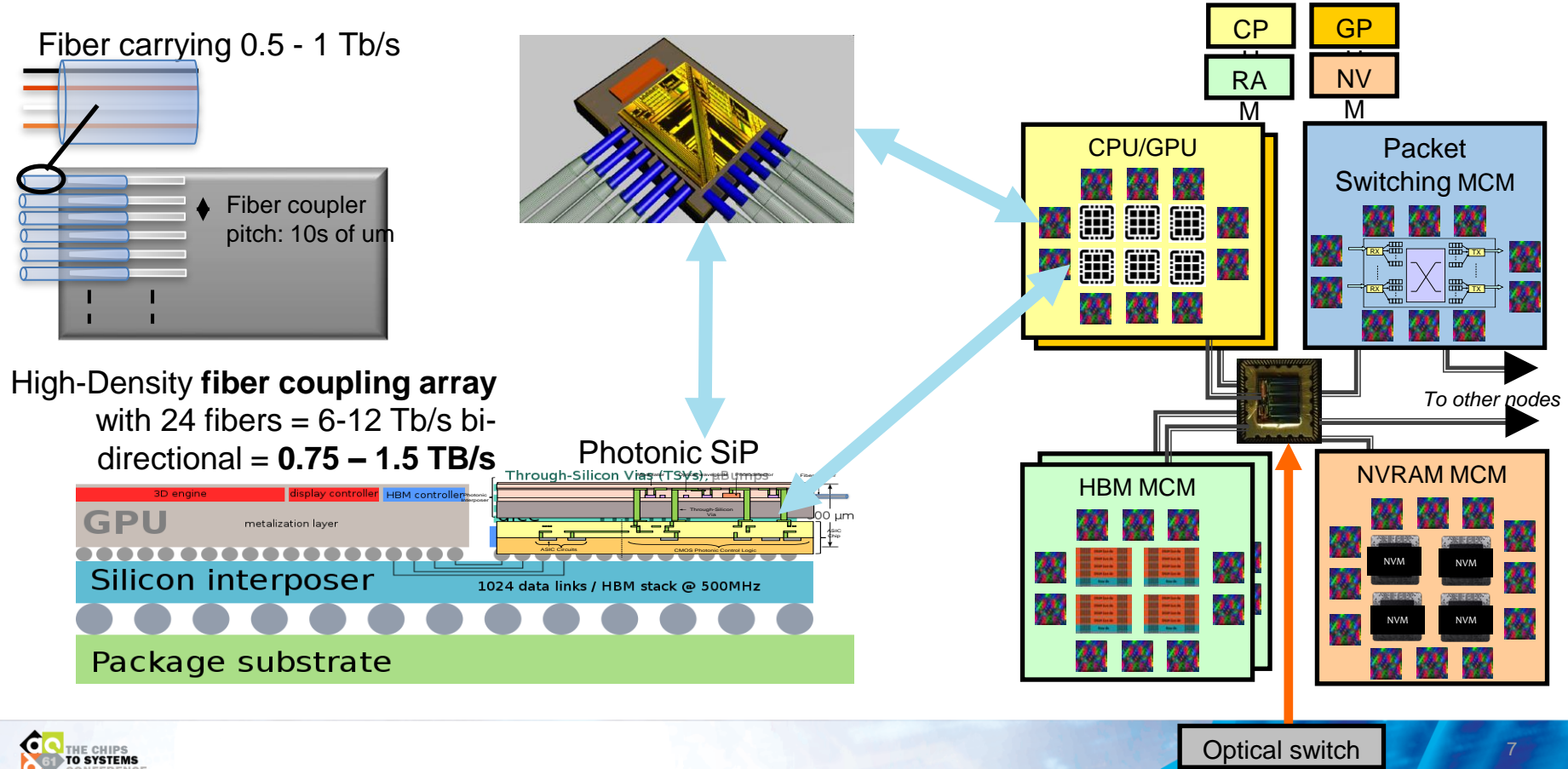
Wide-and Slow for high speed links



Photonic SiP

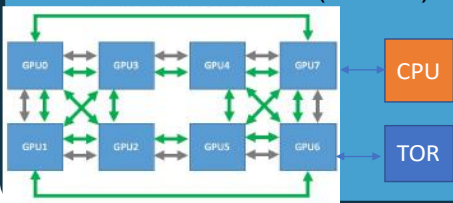


Photonic MCM (Co-Packaged Optics)



Training

- 8 connections: Peer GPU
- 8 links to HBM (weights)
- 8 links: to NVRAM
- 1 links: to CPU (control)



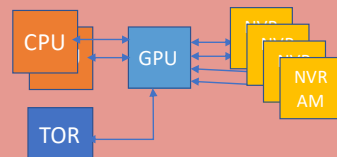
Inference

- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU



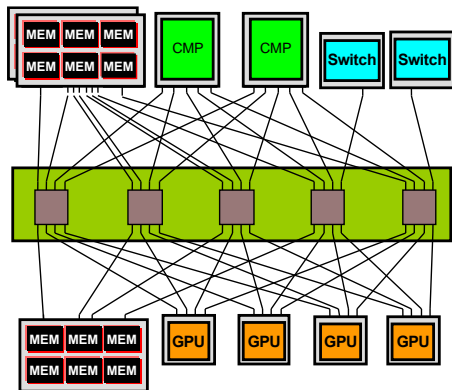
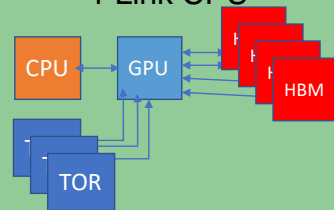
Data Mining

- 6-links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)



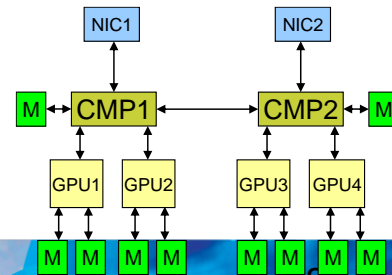
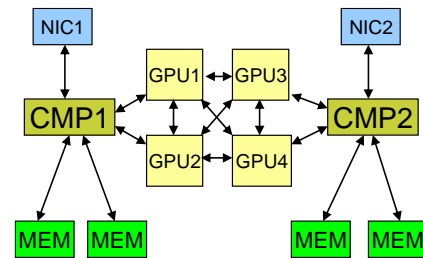
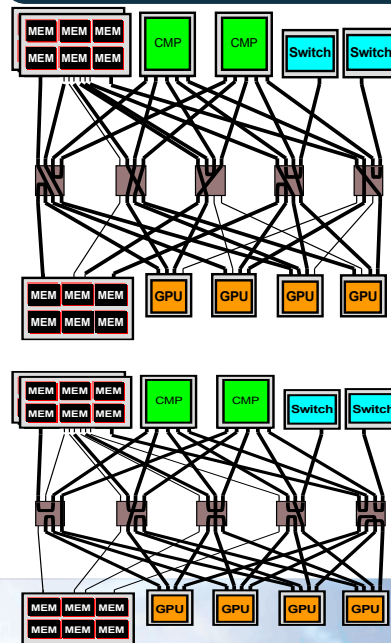
Graph Analytics

- 16 links HBM
- 8 links TOR
- 1 Link CPU



Configure for Training

Configure for Inference





THE CHIPS TO SYSTEMS CONFERENCE

SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

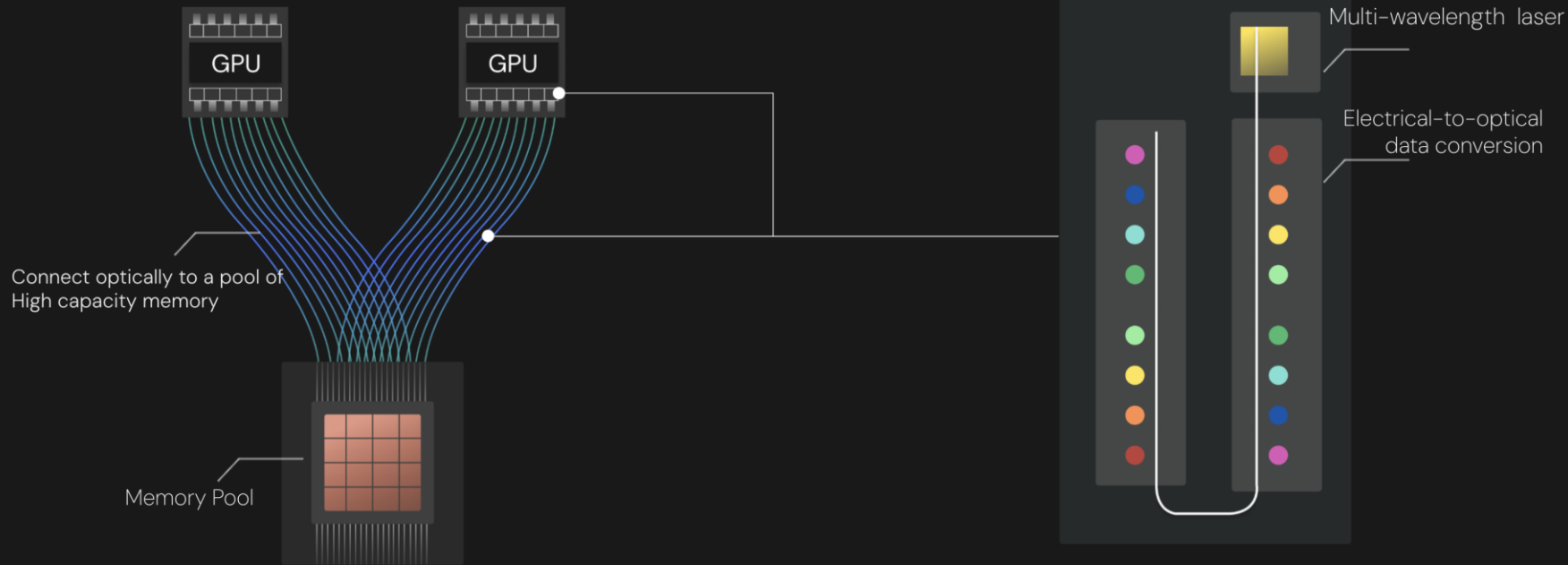
MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA



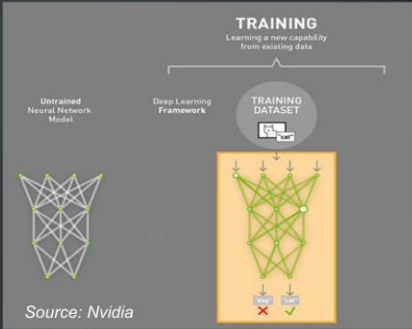
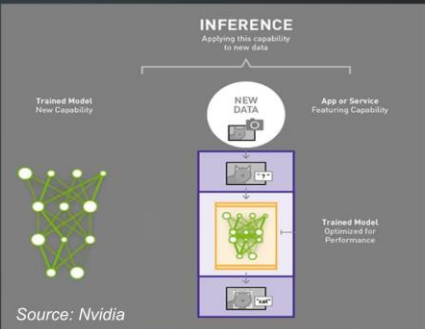

Solution : Leverage Photonics Chiplets to scale AI fabric

Wavelength division multiplexing (WDM) photonics chiplets in package offers the highest BW density

Maximize bandwidth out of GPU by converting electrical signals to optical signals in the same package (co-packaged optics)



Different WDM requirements for different AI workloads

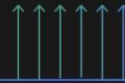
	AI TRAINING	AI INFERENCE	EDGE AI/5G/6G
	 <p>Source: Nvidia</p>	 <p>Source: Nvidia</p>	 <p>https://lionbridge.ai/articles/what-is-edge-ai-computing/</p>
BW	0.4 – 0.8 Tbps	> 8 Tbps	0.1 – 0.4 Tbps
Latency	μ s – ms	10 – 100 ns	~ ms
Reach	Up to 3 km	< 100 m	Up to 40 km
WDM	1 – 4 λ 20-nm Grid	8 – 32 λ 1.1-nm Grid	4 – 8 λ , Tunable 4.5-nm Grid

CombX Technology

Industry's first programmable laser for different AI Workloads



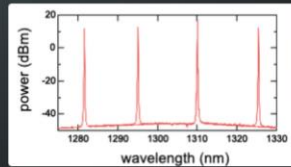
Actual CombX Module



Programmable number of
wavelengths and spacing

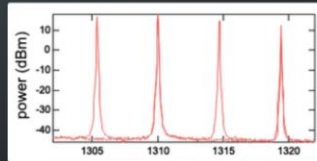


AI TRAINING



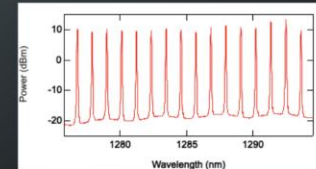
20-nm CWDM4 Grid

EDGE AI/5G/6G



4.5-nm LR4 Grid

AI INFERENCE



1-nm CW-WDM Grid

Thank You

