LIGHTMATTER

Accelerating Frontier MoE Training with 3D Integrated Optics

Mikhail Bernadskiy, Peter Carson, Thomas Graham, **Taylor Groves**, Ho John Lee, Eric Yeh

Hot Interconnects | Virtual Conference | August 20, 2025

HOT INTERCONNECTS

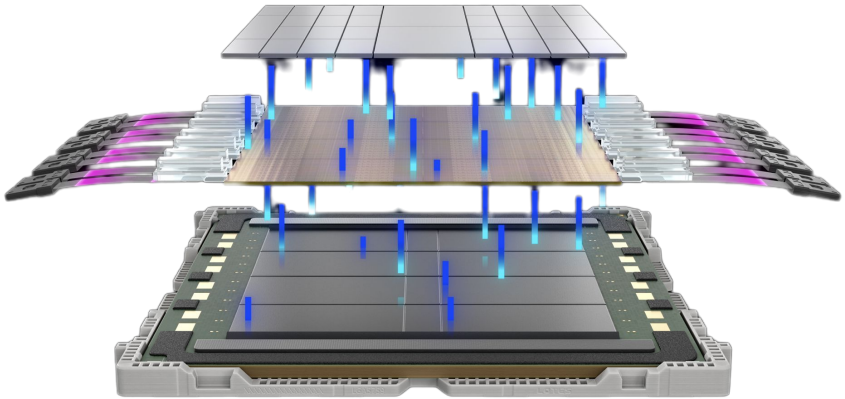
Passage: The Solution to AI Scaling

3D stack of Photonics and Electronics integrates massive bandwidth, low power, beyond copper reach.

8X increase to scale-up pod bandwidth using half the energy of conventional CPO.

6X reduction in package area expansion compared to CPO.

2.7X increase to MoE training throughput



Motivation

AI Scaling

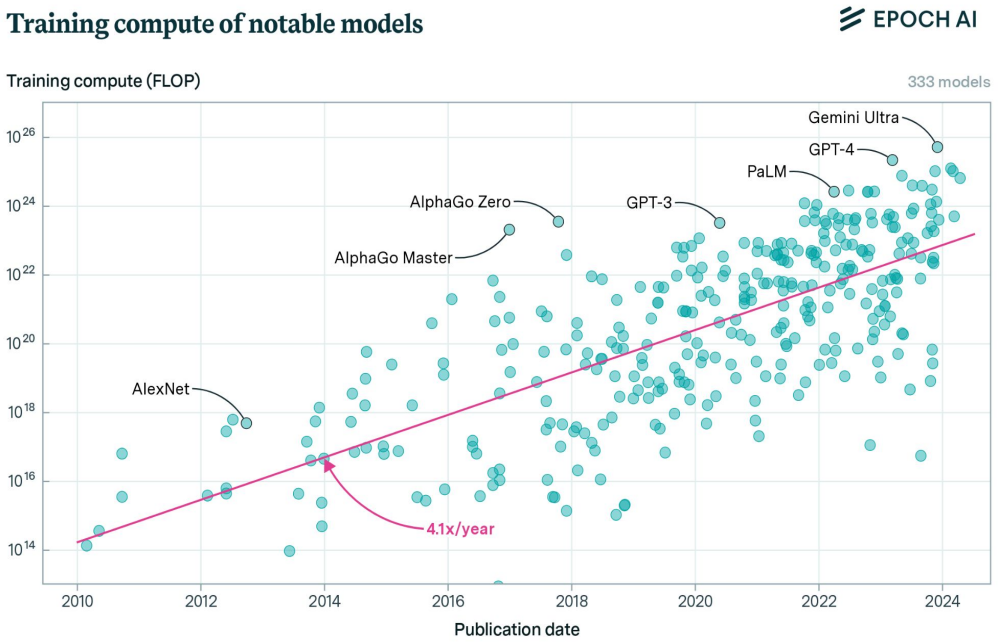
Exponential growth of model sizes and training workloads, faster than growth in device compute FLOPs, memory bandwidth, memory capacity

Requires extensive use of network:

GPU-to-GPU Parallelism: Tensor, Expert, Context, Pipeline, Data

GPU-to-I/O: Checkpointing, context caching, prefill

Chain of thought, reasoning models

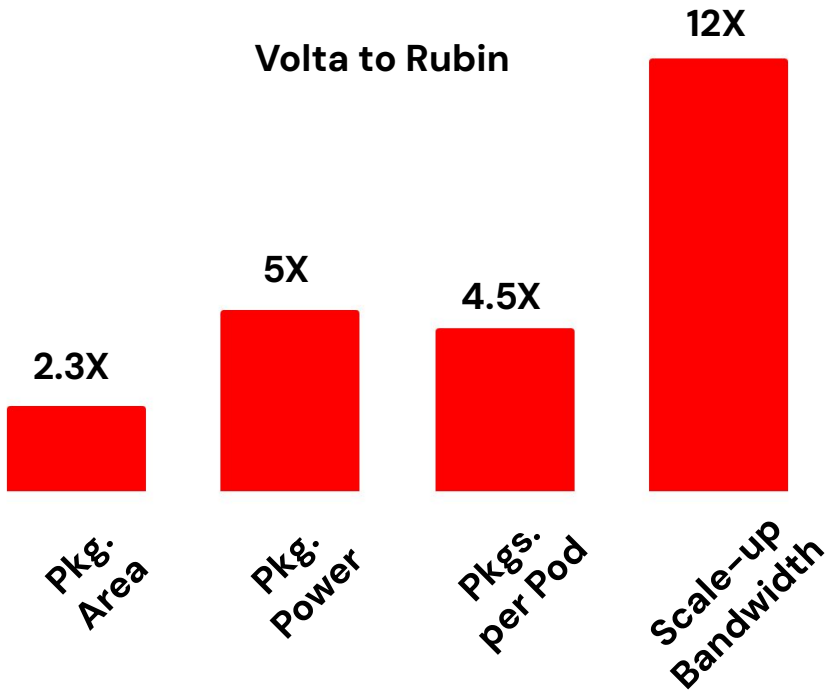


Scale-up Pod is the Unit of Compute

Process improvement not enough (15% per gen)

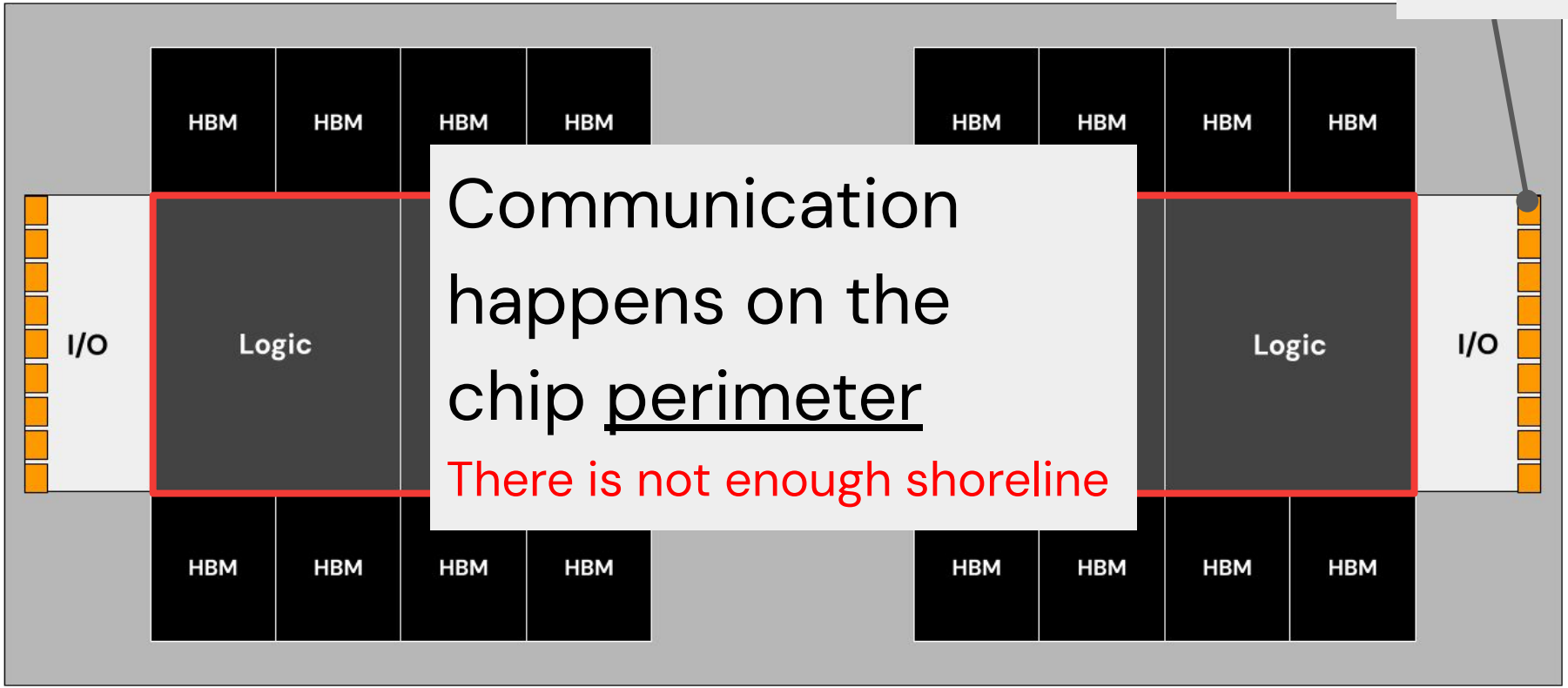
Continued Scaling Requires:

- **Increase package size**, number of logic and memory dies per package
- **Increasing Power:** at both the package and Pod
- **Increasing Pod Size and Bandwidth**: Tightly coupled Accelerator packages with high bandwidth and low latency



But, conventional technologies have hit a wall

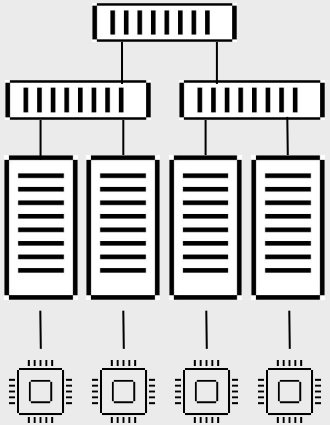
Challenge: Package Area and Shoreline



LIGHTMATTER

The Scale Up Challenge

Scale Out

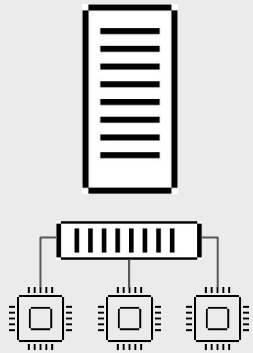


100k+ GPUs

How do we get the full GPU bandwidth to as many GPUs as possible in nanoseconds?

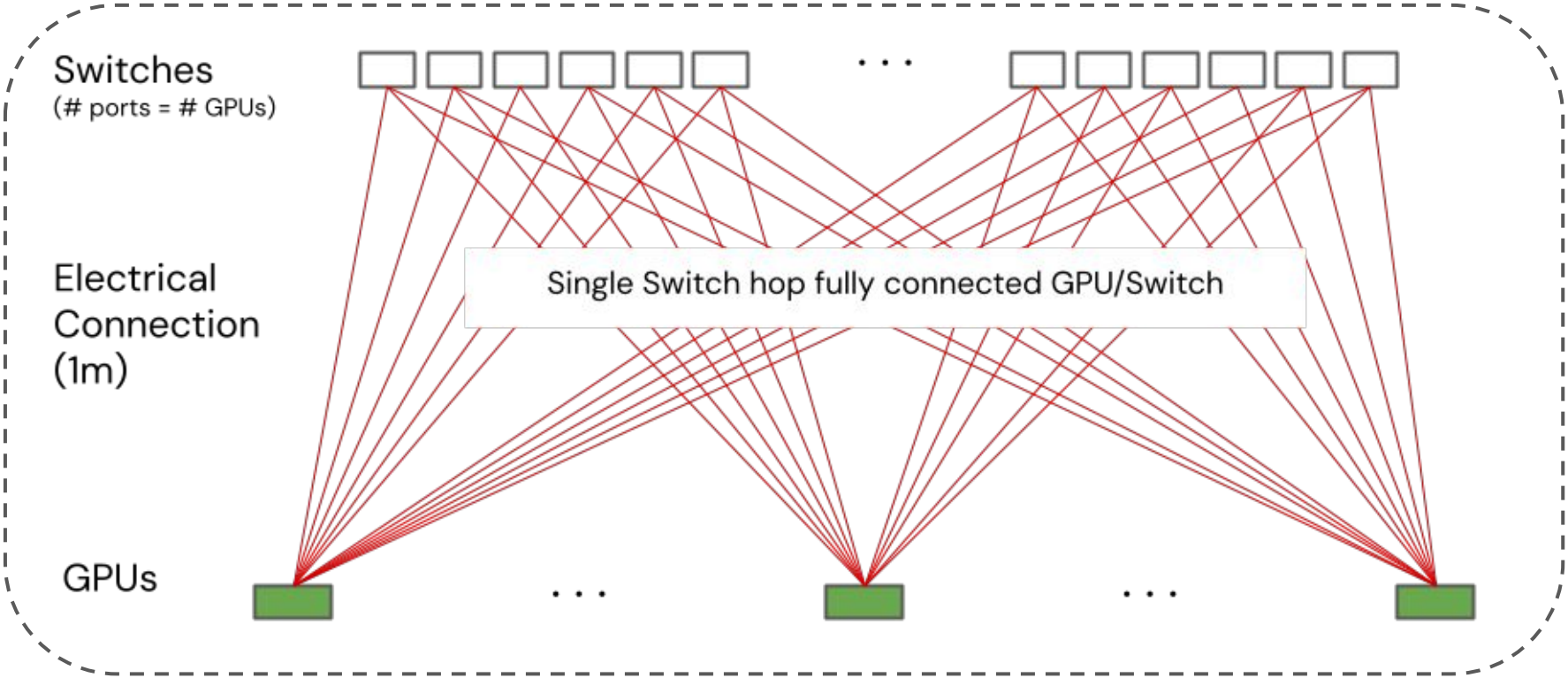
Network Type	No. GPUs	Latency	Bandwidth Per GPU	Energy
Scale-out	> 100k	multi-hop 2-10 us	→ 1.6 Tbps	16 pJ/bit
Scale-up	→ 1024	100-250 ns	> 12.8 Tbps	< 5 pJ/bit

Scale Up



→1024 GPUs

Multi-Rail Single Layer of Switching



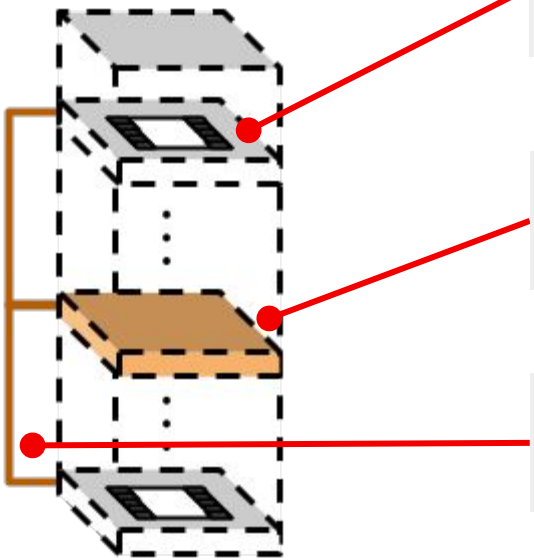
Scale-up Pod (512-1024 GPU Packages)

Challenge Copper Reach & Pod Density

Power is headed towards **1 MW Rack**

- More bandwidth achieved by faster data rate.
- Faster data rate **reduces reach.**
  - 224G PAM4 to 448G < 1 meter

The Copper Scale-up Pod is limited by how many GPUs that can fit within a meter of the switch

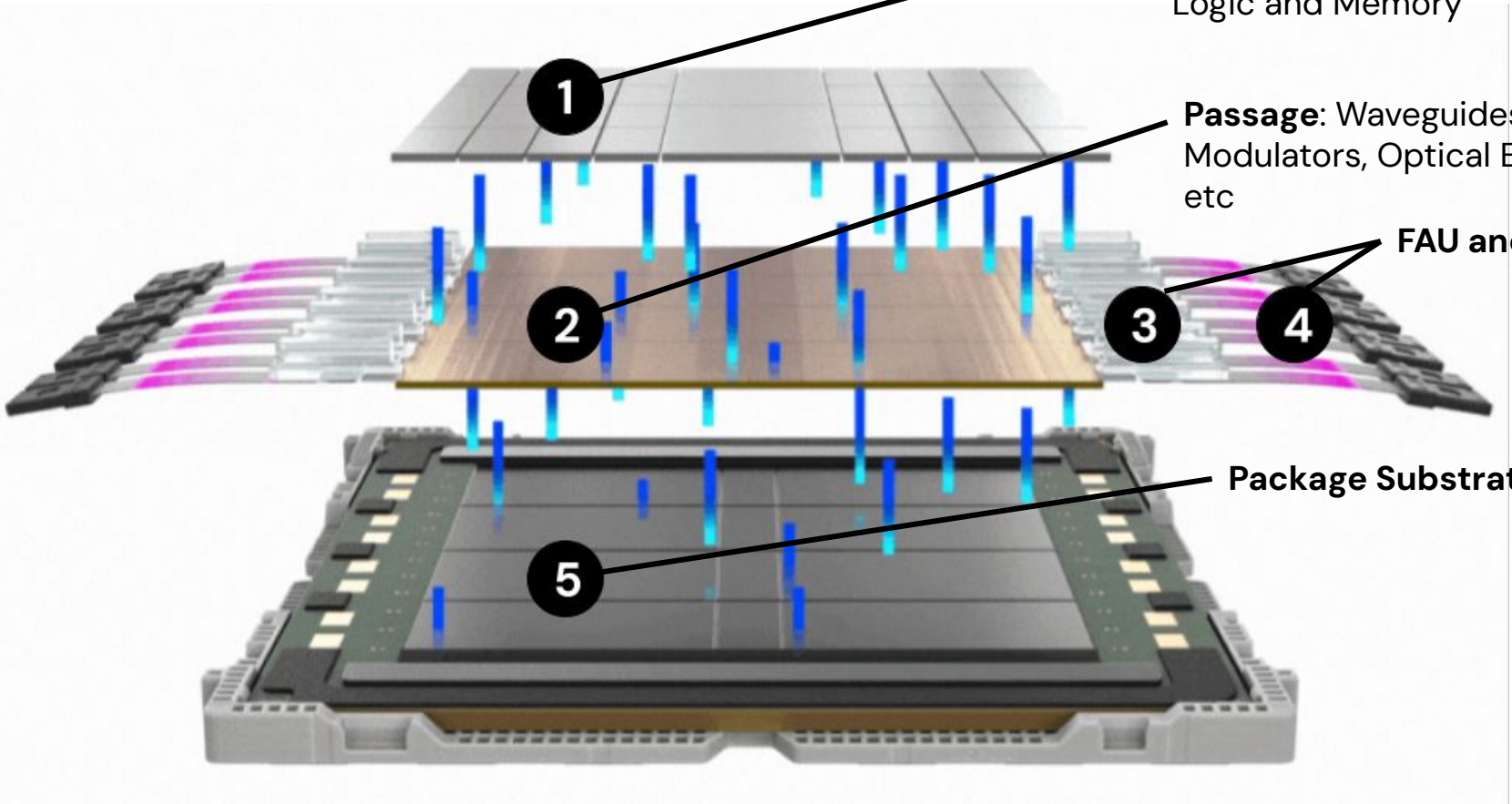


"NVIDIA is leading the transition to 800 VDC data center power infrastructure to support 1 MW IT racks and beyond"
https://developer.nvidia.com/blog/nvidia-800-v-hvdc-architecture-will-power-the-next-generation-of-ai-factories/

LIGHTMATTER

Passage Primer

Passage

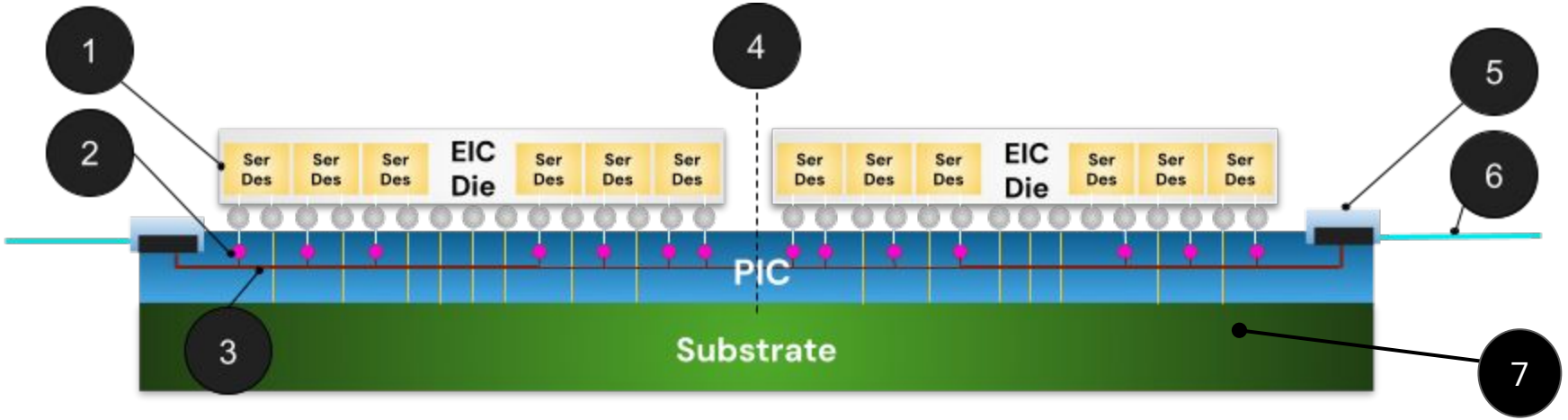


Electrical Integrated Circuit: SerDes (throughout area), Logic and Memory

Passage: Waveguides, TSVs, Modulators, Optical Elements, etc

FAU and Fibers

Package Substrate



1. Low Power SerDes (multiple rows deep) on top of (100 um between)
2. Driver/TIA to/from Microring Modulator (MRM)
   a. WDM multi TX and RX on same waveguide/fiber
3. Waveguide routing (4 um) and OCS throughout the PIC
4. Cross-reticle waveguide stitching enables waferscale designs
5. Pluggable Fiber Attach Unit enables HVM
6. Optical Fibers (127um) extend the reach of Scale-up to 100's meters
7. TSVs provide Power and Signal to EIC

LIGHTMATTER

AI Accelerated with SiPho Interconnect

Bandwidth Density + Radix

- > 896 Gbps per fiber 16λ
- TX + RX on the same fiber
- > 4 Fibers per mm
- 114Tbps demonstrated on package (see HotChips ‘25)

Datacenter reach

- 1000s of XPU in Scale-up
- Alleviate rack cooling and power challenges

3D Stacking and Power Efficiency

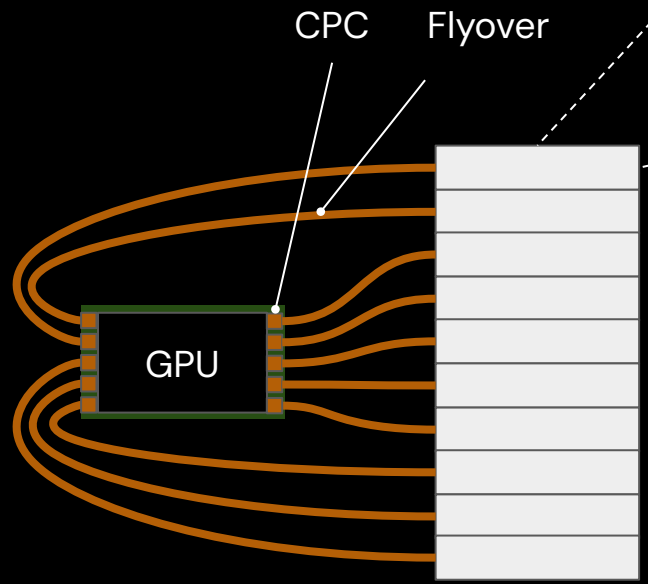
- 100 um Electrical ↕ Optical
- Power Delivery: >>1500W
- SiPho+Laser: 2.3 pJ/bit
- Minimal Package Growth

32T GPU
80 Ports @ 400Gbps

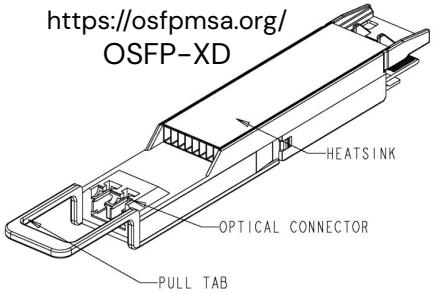
Conventional Optics vs Passage

Option 1. LPO

4576 sqmm GPU
13 pJ/b (includes host SerDes)
10X 3.2T modules, 2,389 sqmm per module
10 CPC (16 DP ea.)

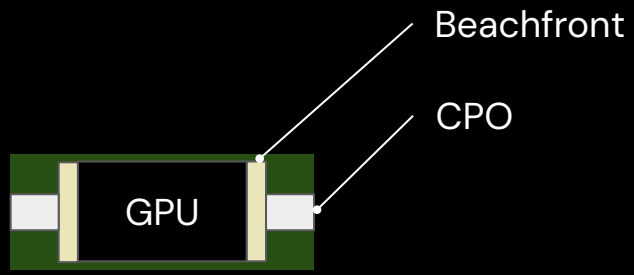


https://osfpmsa.org/
OSFP-XD



Option 2. 2D CPO

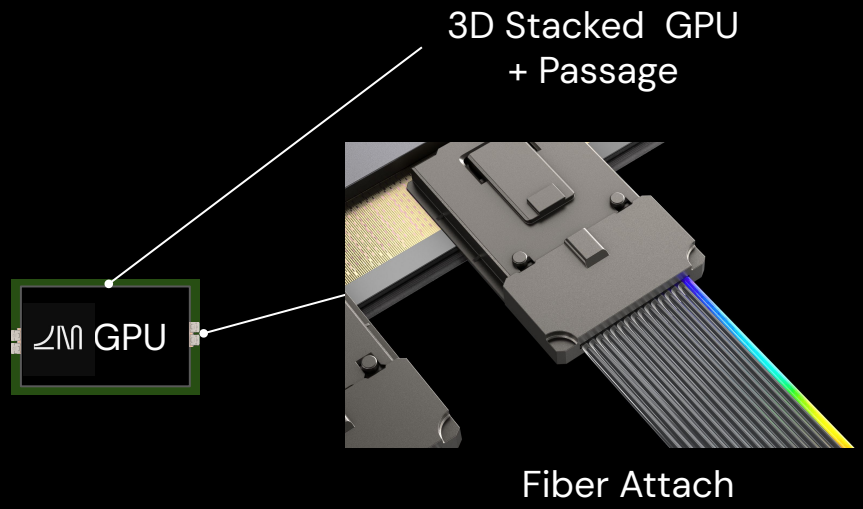
4576 sqmm GPU
12 pJ/b (includes host SerDes)
2 16T modules
1.7 Tbps/mm bidirectional
470 sqmm per Optical Engine + Beachfront and Substrate



32T Conventional Optics
Beachfront + Optical Engines
Doubles GPU Package Area

Option 3. Passage 3D Integration

4576 sqmm GPU
< 5 pJ/b (includes host SerDes)
400Gbps TX + 400Gbps RX per fiber
>2 Tbps/mm bidirectional
12.5 mm X 5 mm fiber attach per side (including laser)



32T Passage
1/6th the Package Growth of 2D CPO
224W of Power Savings per GPU

System and Application Impact

Increased Bandwidth and Scale-up Pod Size

Expert Parallelism and Relation to Scale-up

Multiple Experts with Independent FFNs

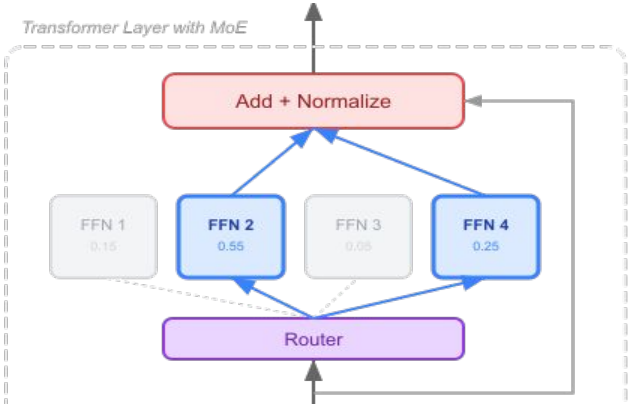
- Only subset of Experts trained per Token
- Significantly reduced computational requirements
- Specialization across experts

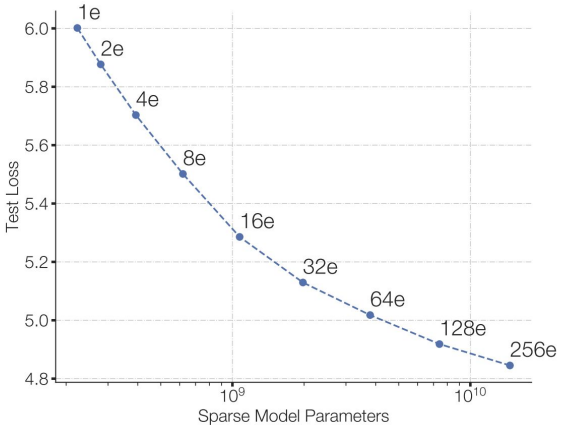
Adds communication:

- Router Activates ‘k’ Experts
- Combine Results of ‘k’ Experts via All-to-All

More Experts leads to better results.
How many Experts?

- Up to 256 used in this study





Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity

Model Parameters

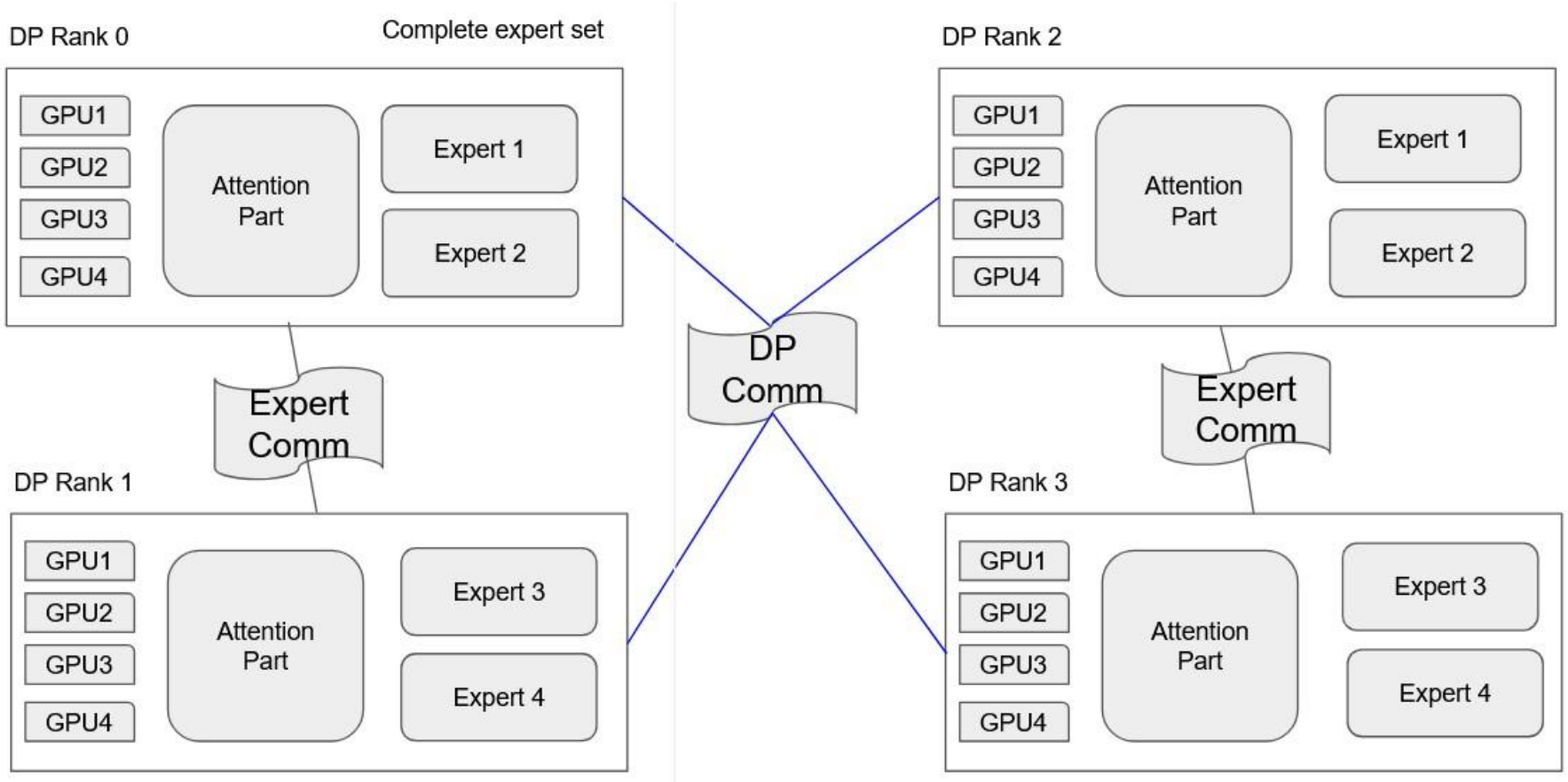
4.7T total parameters on 32,768 GPUs
Layers: 120
Model dimension: 12288
Attention heads: 128

Global batch size: 4096
Sequence length: 8192

Parallelization: TP 16, DP 256, PP 8
Four configurations of Fine Grained Expert Segmentation

Parameter	Config 1	Config 2	Config 3	Config 4
Active / total experts	1/32	2/64	4/128	8/256
Expert granularity (m)	1	2	4	8
Experts per DP rank	1	2	4	8

Model Parameters



Switch and GPU Comparison Points

Passage 14.4 Tbps

BF16 FLOPs	HBM	Switch Ports	Scale-up per GPU
8.5 PF	16 stacks, 209 Tbps	**512**	**14.4 Tbps TX + 14.4 Tbps RX**

Passage 32 Tbps

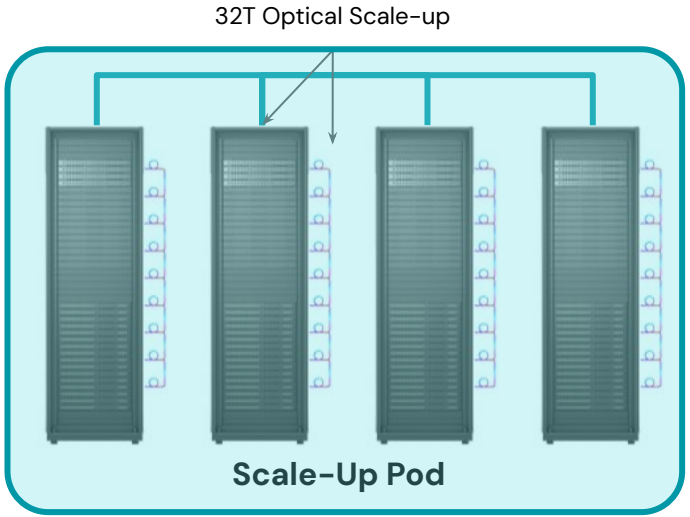
BF16 FLOPs	HBM	Switch Ports	Scale-up per GPU
8.5 PF	16 stacks, 209 Tbps	**512**	**32 Tbps TX + 32Tbps RX**

Electrical 14.4 Tbps

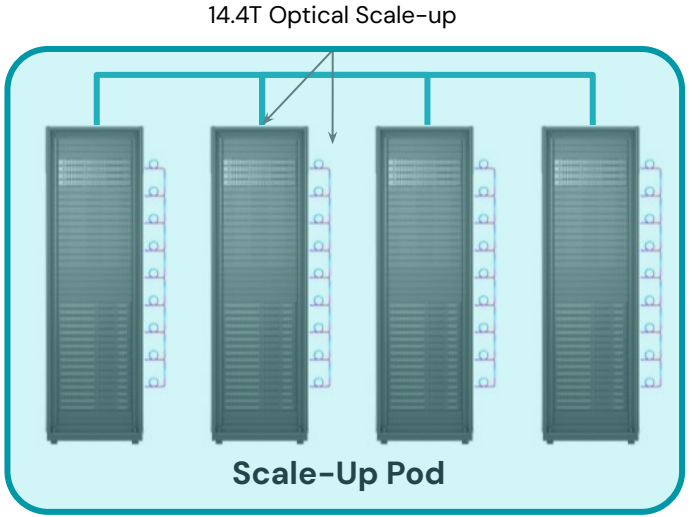
BF16 FLOPs	HBM	Switch Ports	Scale-up per GPU
8.5 PF	16 stacks, 209 Tbps	**144**	**14.4 Tbps TX + 14.4 Tbps RX**

Study 1. Impact of Bandwidth

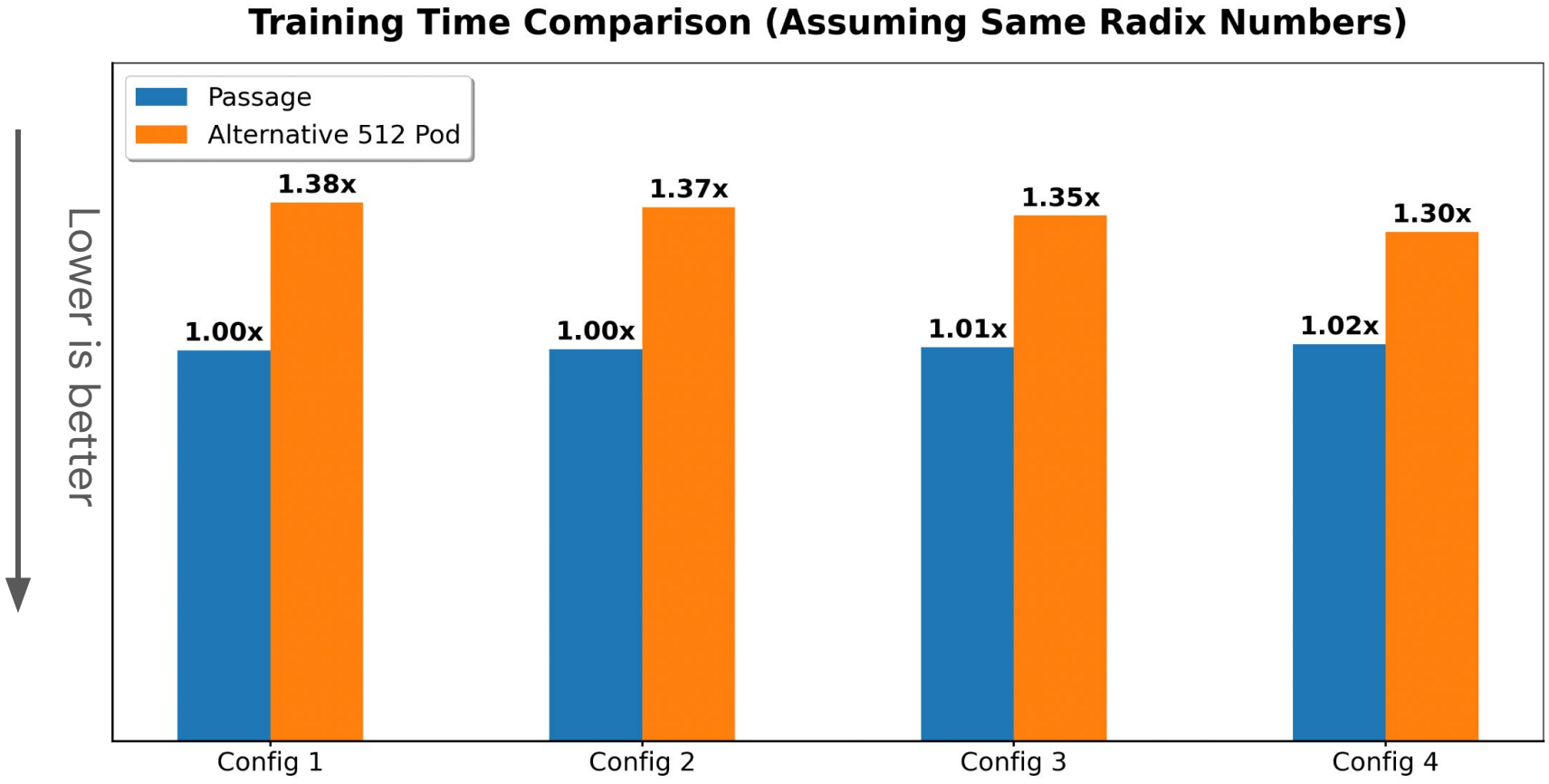
512-Pod Passage 3D CPO
512 active GPUs/Pod (2,048 GPU dies)
Optical 32T per GPU
Scale-out 1.6T per GPU



512-Pod Passage 3D CPO
512 active GPUs/Pod (2,048 GPU dies)
Optical 14.4T per GPU
Scale-out 1.6T per GPU

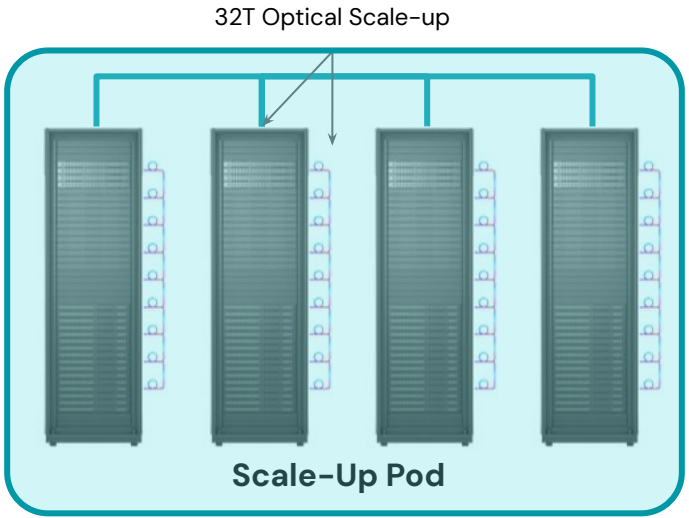


Passage Accelerates AI Model Training Time

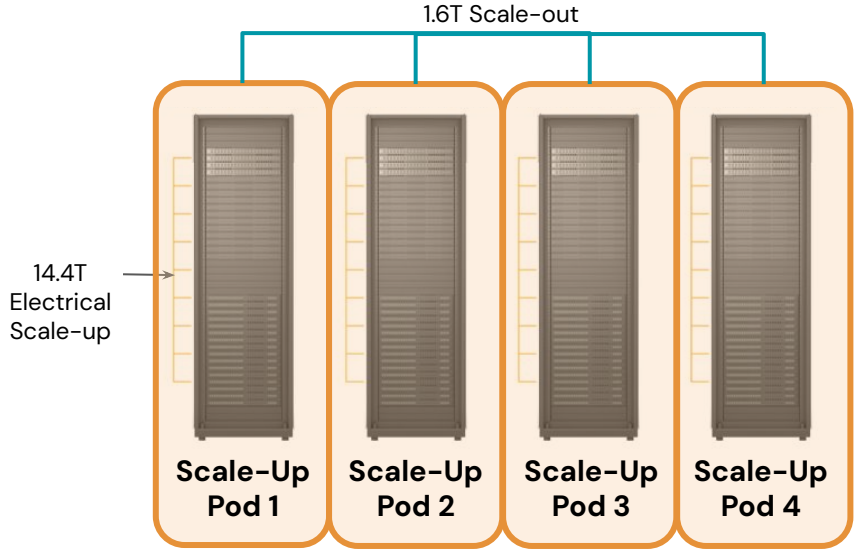


Study 2. Impact of Pod Size + Bandwidth

512-Pod Passage 3D CPO
512 active GPUs / Pod (2,048 dies)
Optical 32T per GPU
Scale-out 1.6T per GPU

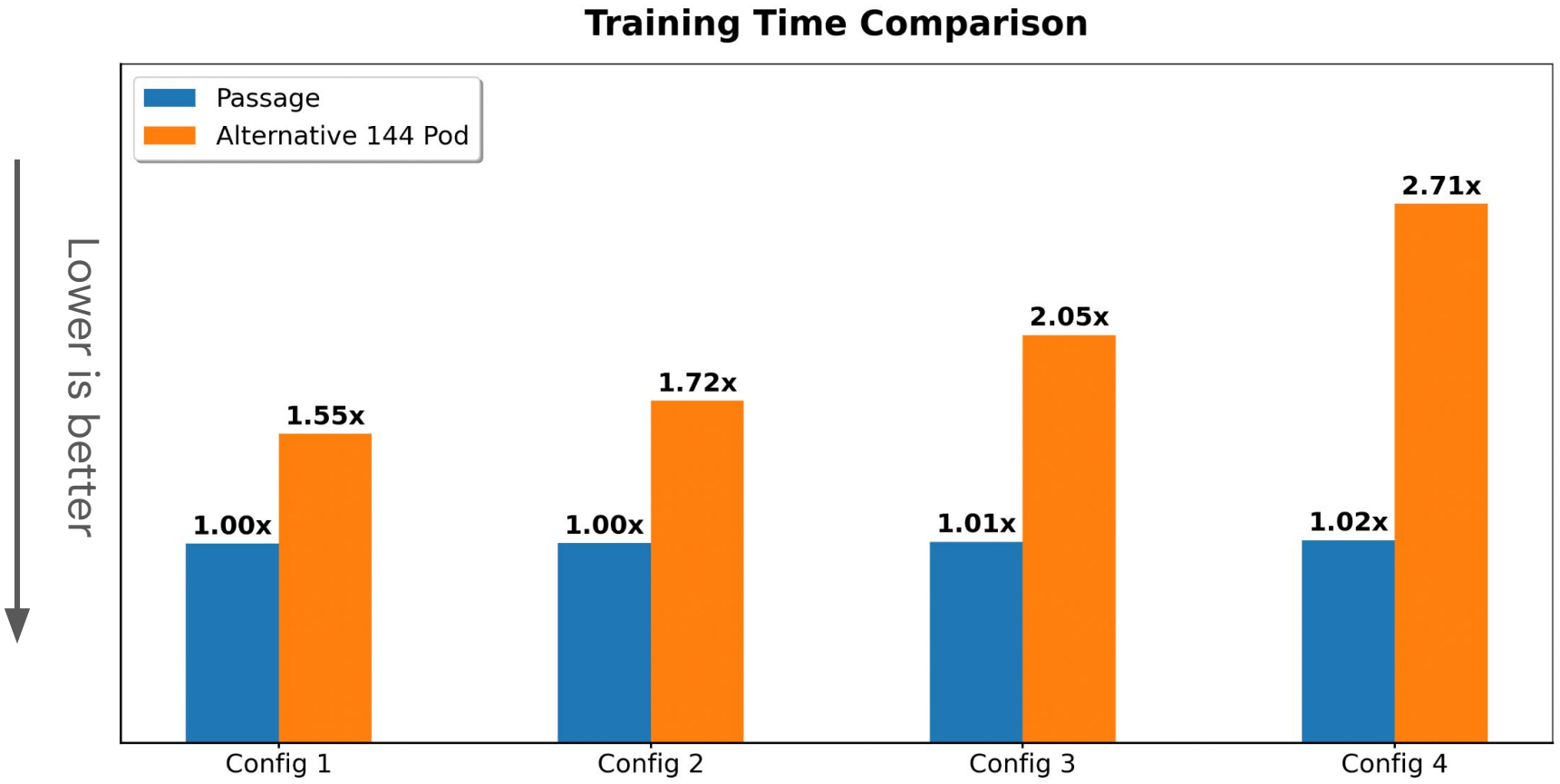


4X 144-Pods
512 active GPUs/ 4 Pods (2,048 dies)
Electrical 14.4T per GPU
Scale-out 1.6T per GPU



LIGHTMATTER

Passage Accelerates AI Model Training Time

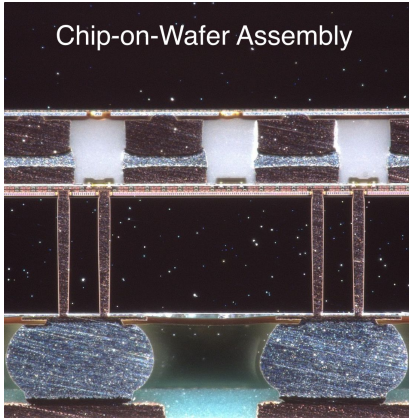


Conclusions

HOT CHIPS

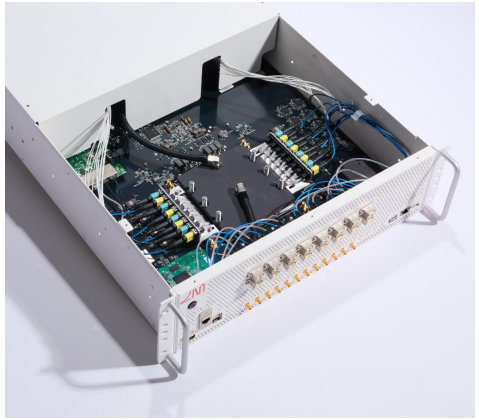
Darius Bunandar,
“Passage M1000: 3D photonic interposer for AI”,
HotChips ‘25, 10AM PST August 26, 2025

Multi-reticle 3D stacked CPO with fiber array connectivity (M1000)



Cross section for photonic interposer (M1000)

16λ per fiber, 200 GHz grid 16 fiber output



M1000 chassis with 256 fiber attach, laser and cooling

Increasing Scale-up Domain and Bandwidth are critical for continued training improvements

Passage Addresses Key Challenges for Next-gen Scale-up

- Bandwidth Density and Reduced Package Growth
- Power Efficiency
- Optical Reach for 1000 GPU Pods

LIGHTMATTER

THANK YOU!