Optical Interconnect Challenges for Scaling Future AI Infrastructure

Liron Gantz, Nvidia Electro-Optics | EPIC | Berlin June 2024

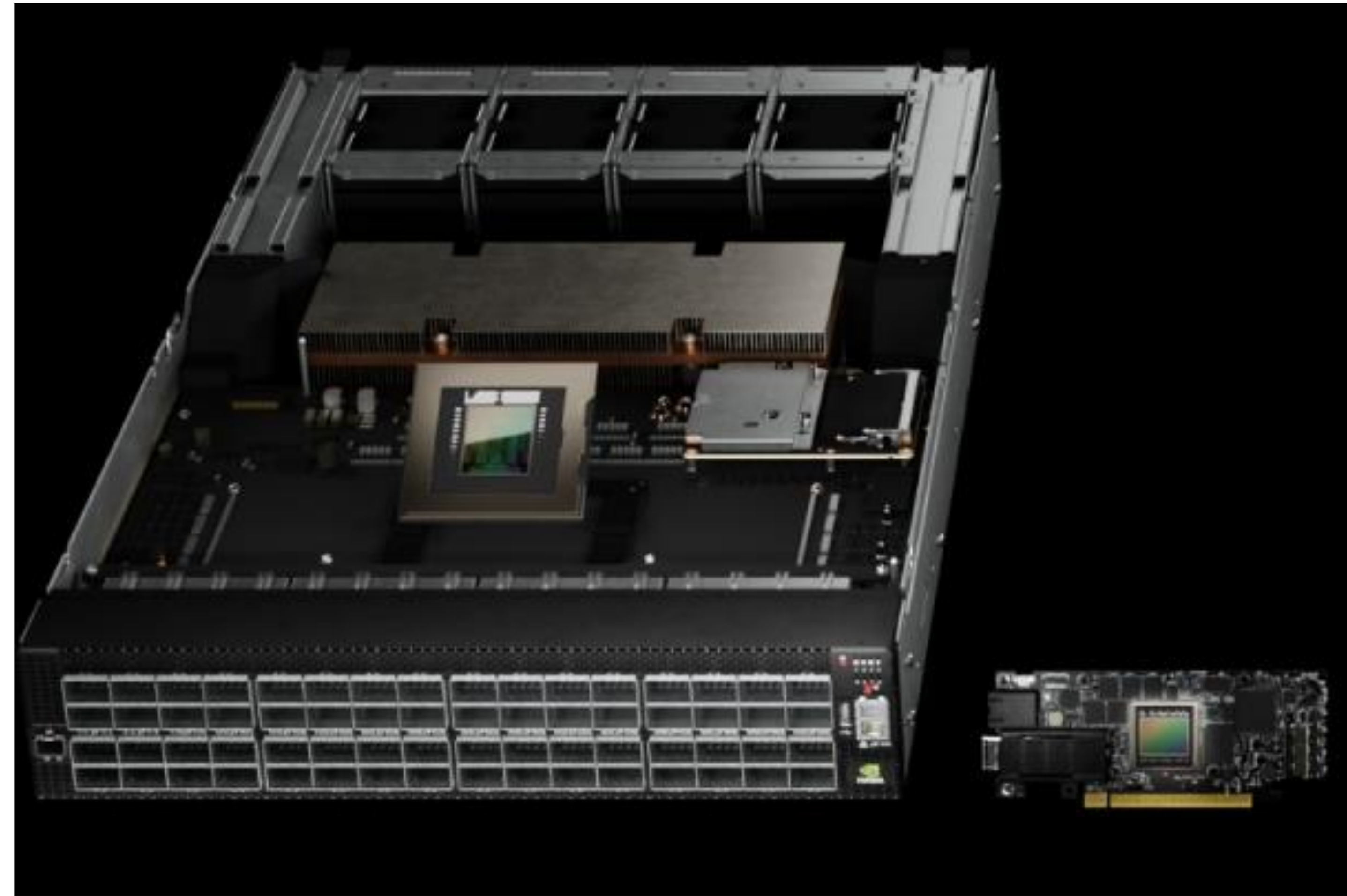
Thanks to Shai Cohen, Ben Lee and Peng Sun for their contributions

Agenda

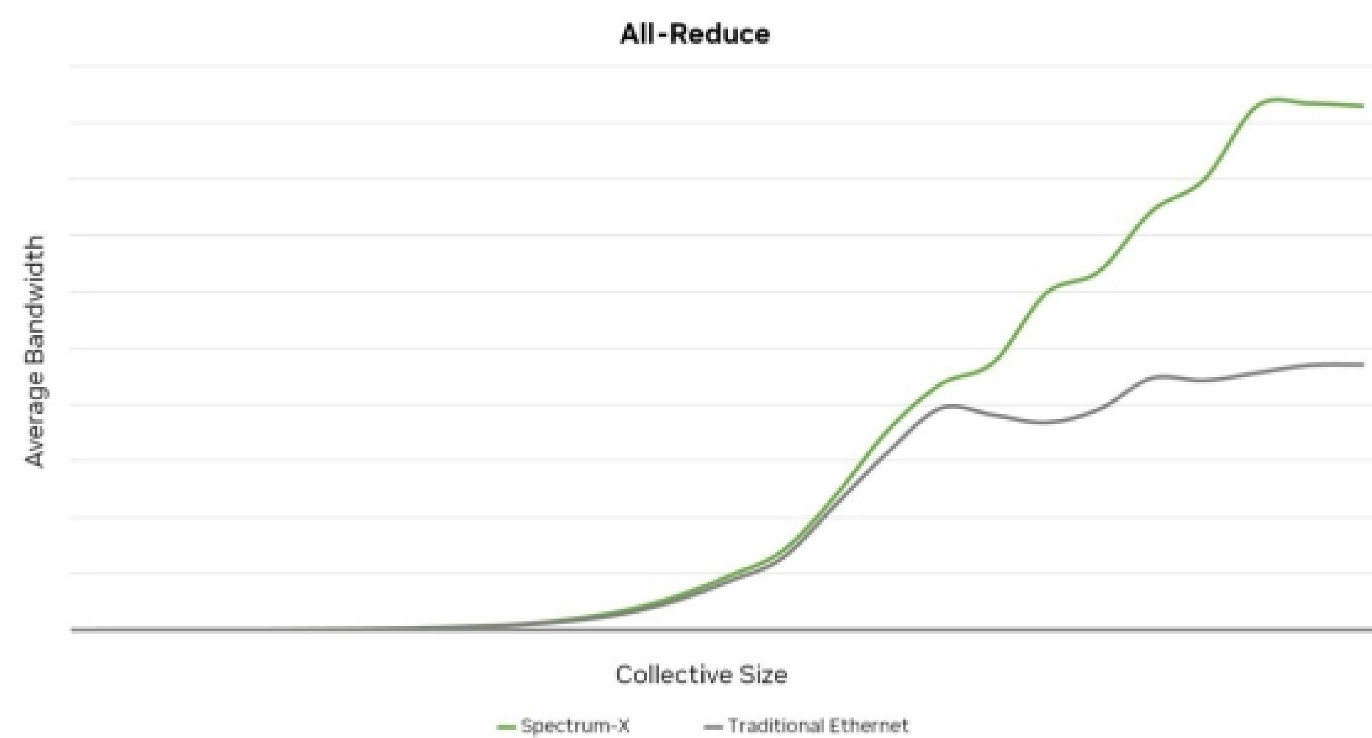
- **Trends in AI and the need for interconnect**
- What can photonic integration gain?
- The challenges of photonic integration
- Accelerating photonic design with GPU and inverse design

NVIDIA Networking

End-to-End Networking solutions



SWITCH
InfiniBand (Quantum)
Ethernet (Spectrum)



Network Interface card (NIC)

NVIDIA BlueField-3 SuperNIC

The BlueField-3 SuperNIC is a novel class of network accelerators based on the BlueField-3 networking platform, purpose-built for supercharging hyperscale AI workloads. Designed for network-intensive, massively parallel computing, the BlueField-3 SuperNIC provides up to 400Gb/s of remote direct-memory access (RDMA) over Converged Ethernet (RoCE) network connectivity between GPU servers, optimizing peak AI workload efficiency. Creating a new era of AI cloud computing, the BlueField-3 SuperNIC enables secure, multi-tenant data center environments with deterministic and isolated

Data Process Unit (DPU)

NVIDIA BlueField-3 DPU

The NVIDIA BlueField-3 DPU is a 400 gigabits per second (Gb/s) infrastructure compute platform with line-rate processing of software-defined networking, storage, and cybersecurity. BlueField-3 combines powerful computing, high-speed networking, and extensive programmability to deliver software-defined, hardware-accelerated solutions for the most demanding workloads. From accelerated AI to hybrid cloud, high-performance computing to 5G wireless networks, BlueField-3 redefines the art of the possible.

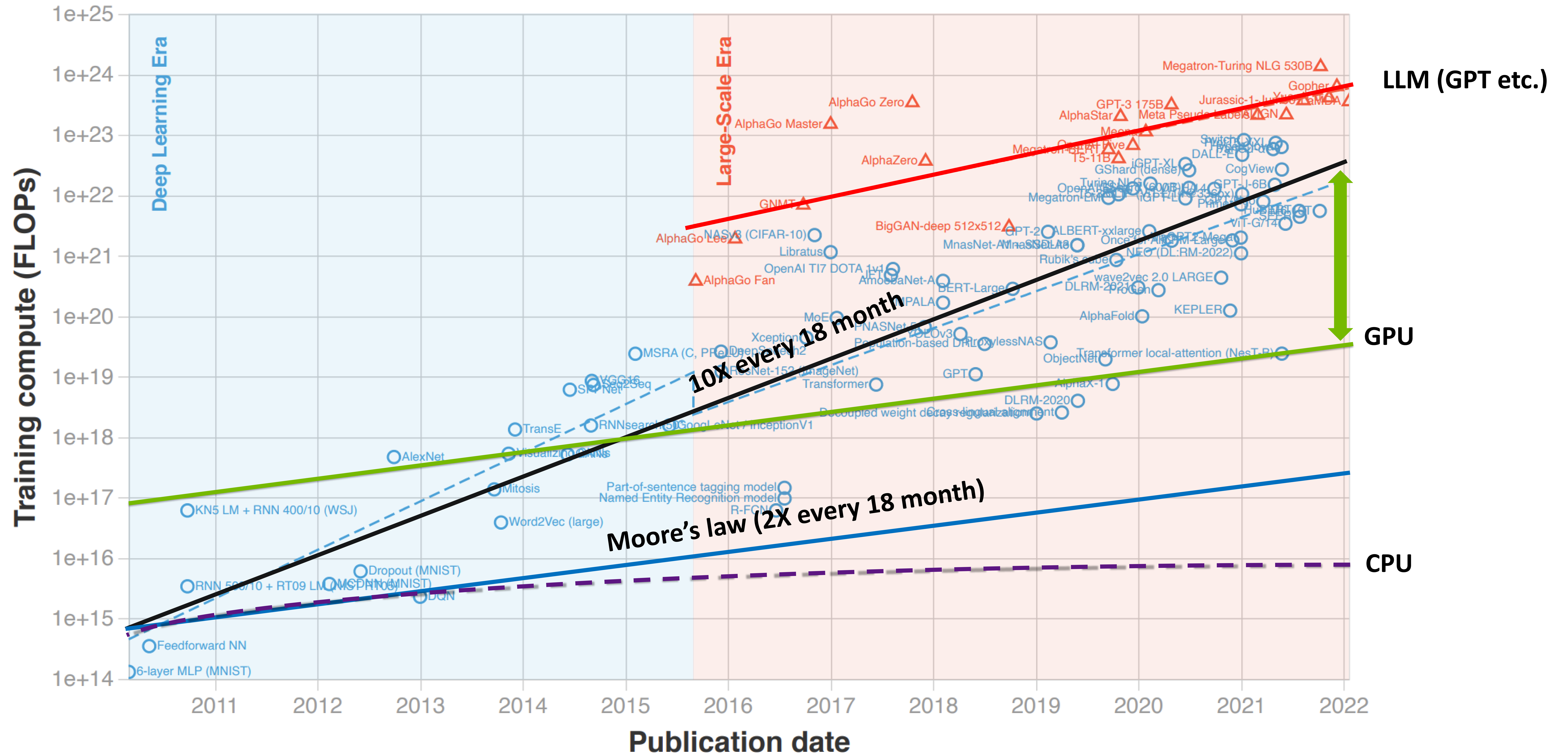
Cable, Transceivers

https://www.nvidia.com/en-us/networking/interconnect/

https://www.nvidia.com/en-us/networking/products/data-processing-unit

Compute Supply and Demand

What LLM needs and what HW can deliver



FLOPs = flouting point operations per second

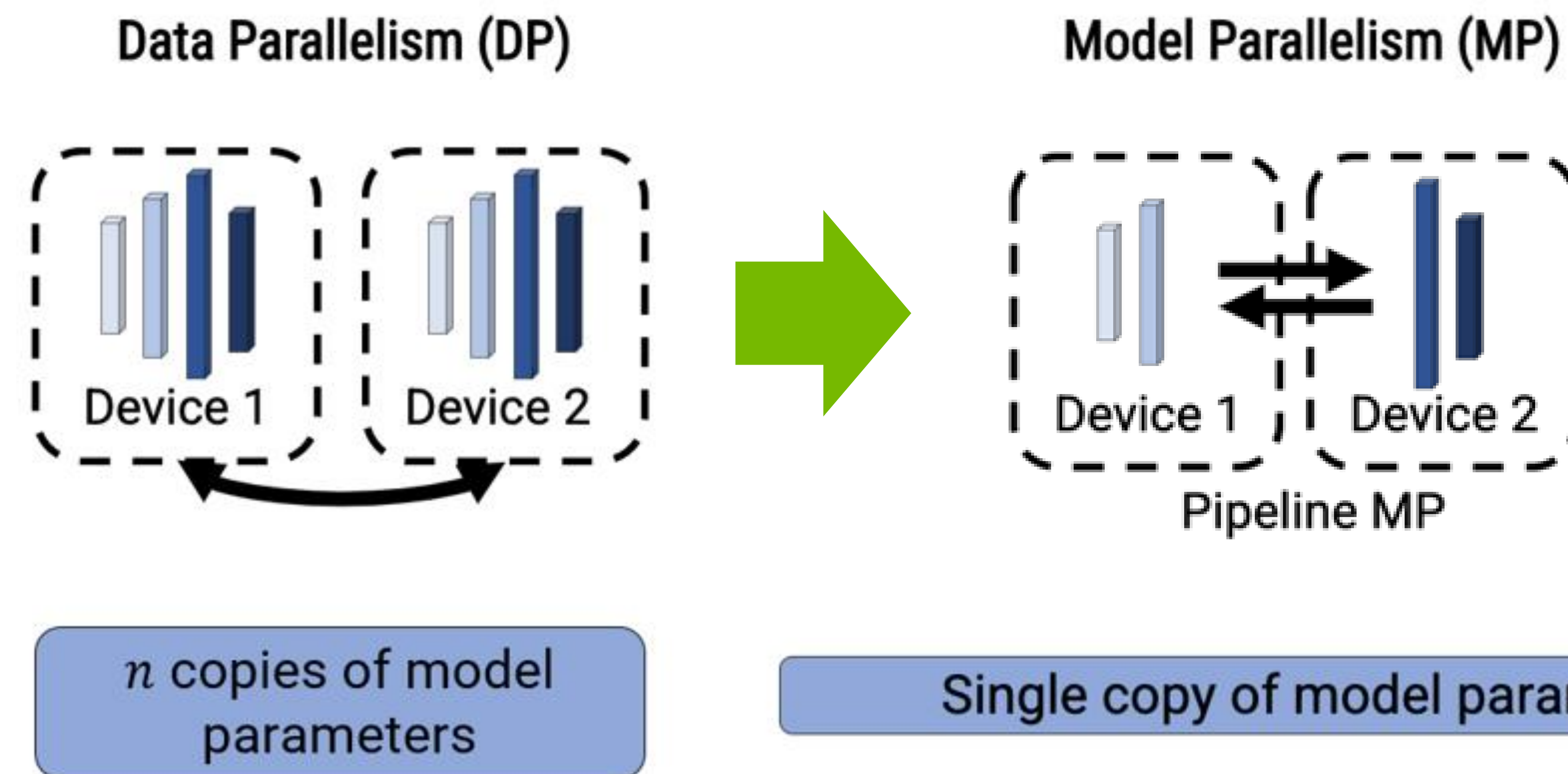
"Compute trends across three eras of machine learning", J. Sevilla, https://ar5iv.labs.arxiv.org/html/2202.05924

NVIDIA

The Building Blocks of Scalable Computing

A modular way to scale both up (compute) and out (networking)

Data and Model Parallelism



Sharing the load requires high BW low latency connectivity

Computing

DGX H100
32 petaFLOPS server

Networking

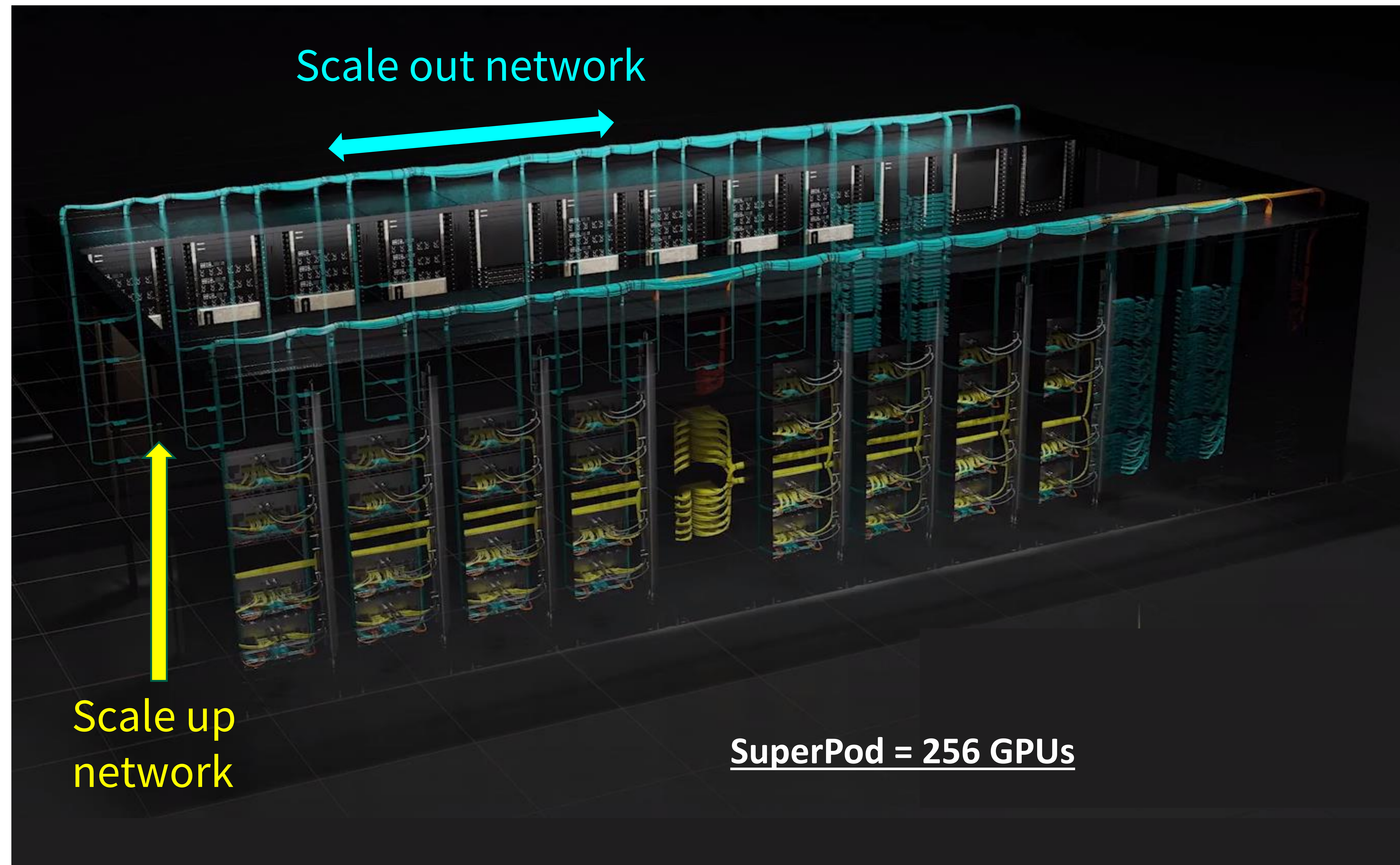
Quantum-2
64 x 400 Gbps switch

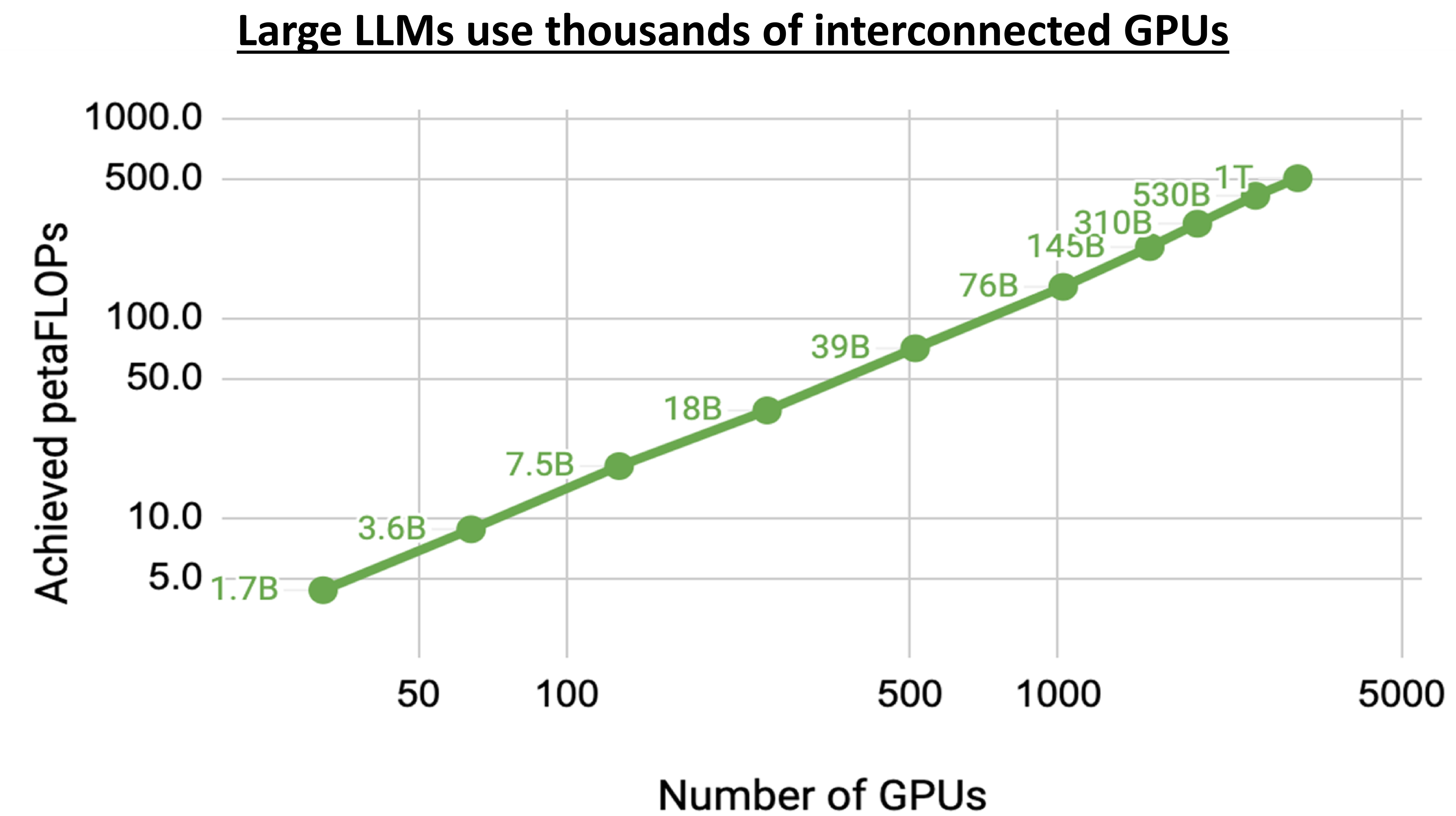
800 Gbps transceiver

https://www.nvidia.com/en-us/data-center

HW for Next-Gen AI/ML Clusters - The SuperPod

256 GPUs scales out to 10,000s GPUs



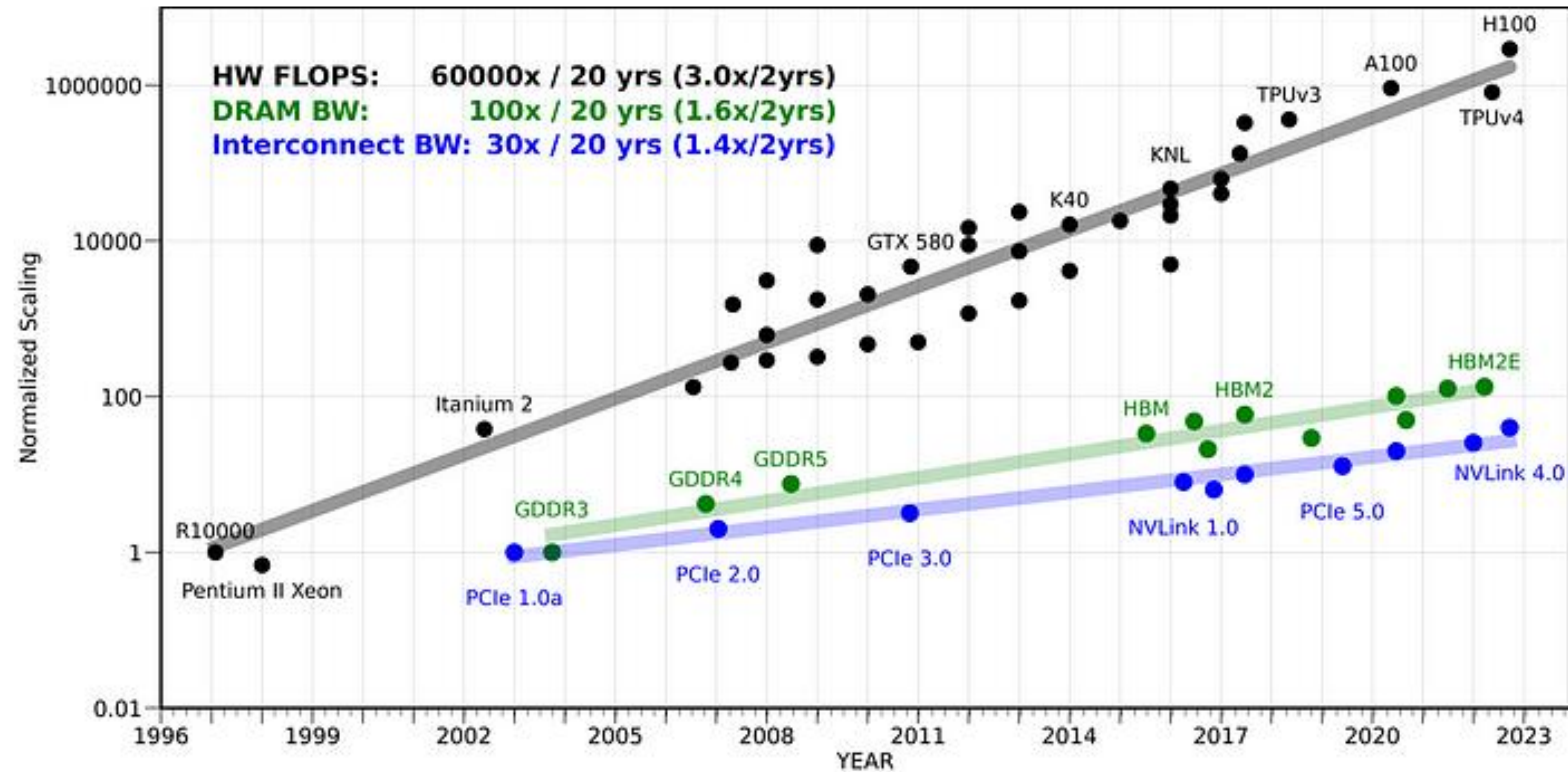


Linear FLOPs scaling - Bill D. Hot interconnects 2023

NVIDIA.

The Challenges With Current Interconnect

Power and cost becomes prohibitive



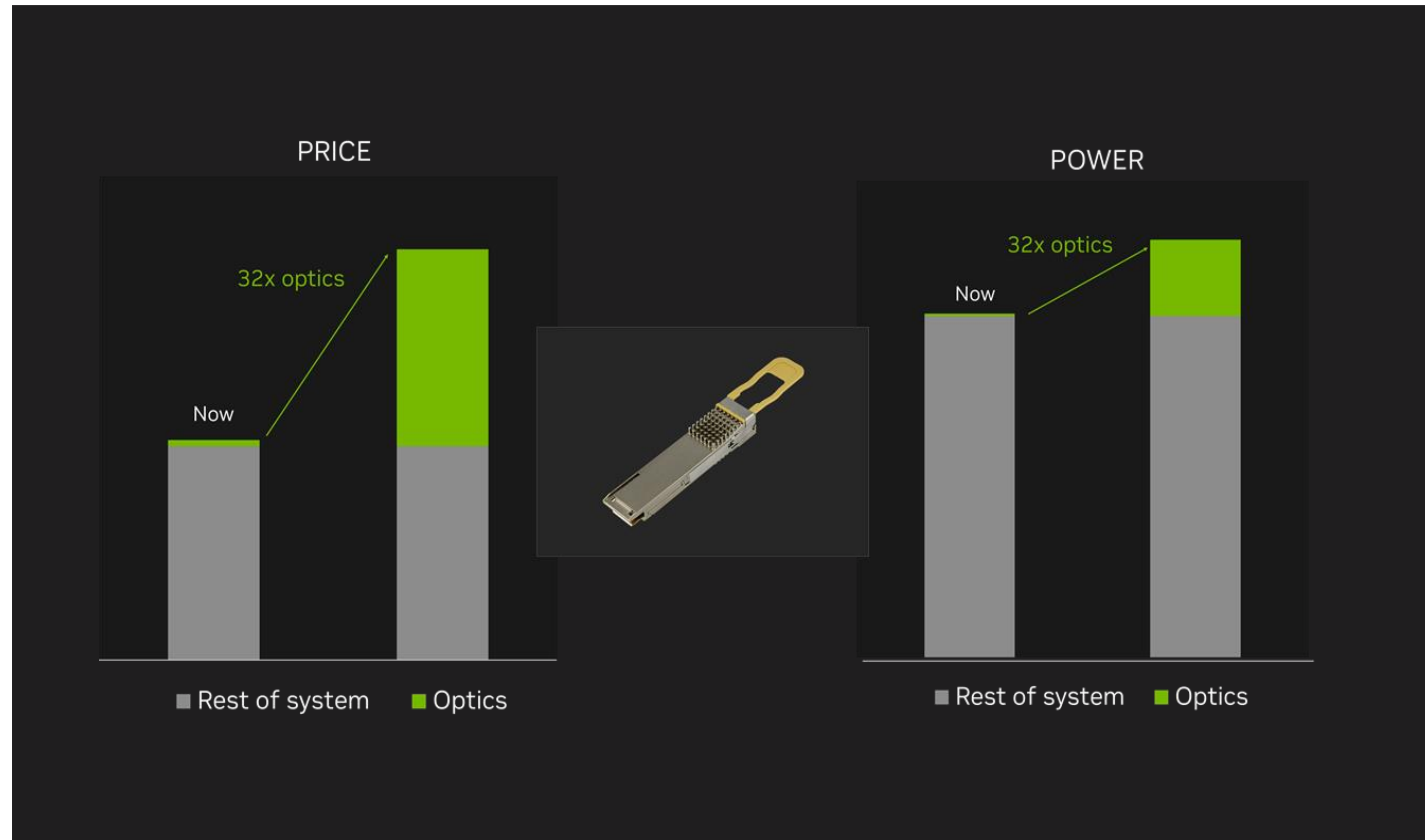
https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

ever-growing gap between Compute-BW and interconnect-BW

an exponential increase in number of communication lanes is needed

The Challenges With Current Interconnect

Power and cost becomes prohibitive



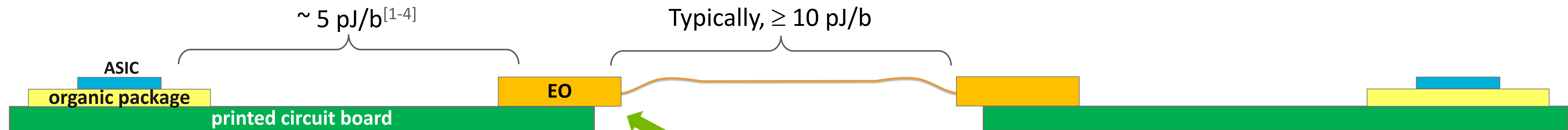
- State of the art optical communication modules already introduce a significant cost and power challenges
  - For example: **10pJ/b** modules serving a **100Tbps** Switch will consume about **1KW** of power

NVIDIA

- Trends in AI and the need for interconnect
- **What can photonic integration gain?**
- The challenges of photonic integration
- Accelerating photonic design with GPU and inverse design

Electrical Interfaces

LR interfaces to edge-of-card optics over printed circuit board



- Limited Pin density requires 100+ Gb/s per trace.
- ~ 500 W of in-package I/O power for a 100Tb/s ASIC

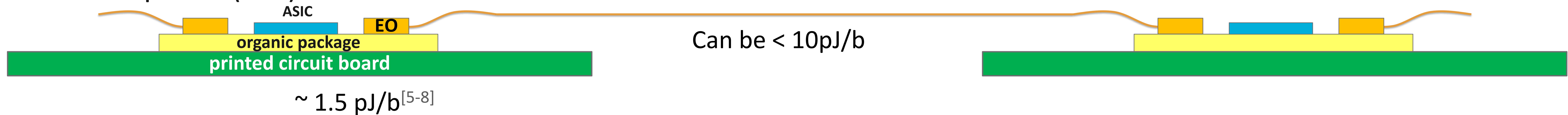
[1] P. Mishra, "A 112Gb/s ADC-DSP-based PAM-4 transceiver for long-reach applications with >40dB channel loss in 7nm FinFET," *ISSCC 2021*, pp. 138-140.
[2] Z. Guo, "A 112.5Gb/s ADC-DSP-based PAM-4 long-reach transceiver with >50dB channel loss in 5nm FinFET," *ISSCC 2022*, pp. 116-118.
[3] A. Varzaghani, "A 1-to-112Gb/s DSP-based wireline transceiver with a flexible clocking scheme in 5nm FinFET," *VLSI 2022*, paper C03-1.
[4] H. Park, "A 4.63pJ/b 112Gb/s DSP-Based PAM-4 Transceiver for a Large-Scale Switch in 5nm FinFET," *ISSCC 2023*, pp. 5-7.

Electrical Interfaces

XSR interfaces to co-packaged optics (CPO) on organic multi-chip module (MCM)

On-board or edge-of-card optics

Co-packaged optics on multi-chip module (MCM)



- Overall power consumption is reduced (to ~60%)

- However, this configuration raises two challenges:
  - Power density in the MCM → similar power inside the MCM over smaller area.
  - MCM cost and trace loss present a strong incentive for area efficient optics → **edge and area BW density become a nontrivial requirement**

[5] R. Shivnaraine, "A 26.5625-to-106.25Gb/s XSR SerDes with 1.55pJ/b efficiency in 7nm CMOS," *ISSCC 2021*, p. 181.
[6] G. Gangasani, "A 1.6Tb/s chiplet over XSR-MCM channels using 113Gb/s PAM-4 transceiver …," *ISSCC 2022*, pp. 122-124.
[7] C. F. Poon, "A 1.24-pJ/b 112-Gb/s (870 Gb/s/mm) transceiver for in-package links in 7-nm FinFET," *JSSC 2022*, vol. 57, no. 4, pp. 1199-1210.
[8] R. Yousry, "A 1.7pJ/b 112Gb/s XSR transceiver for intra-package communication in 7nm FinFET technology," *ISSCC 2021*, pp. 180-182.

Electrical Interfaces

2.5D co-packaged optics (CPO) on silicon interposer

On-board or edge-of-card optics

Co-packaged optics on MCM

2.5D co-packaged optics on interposer

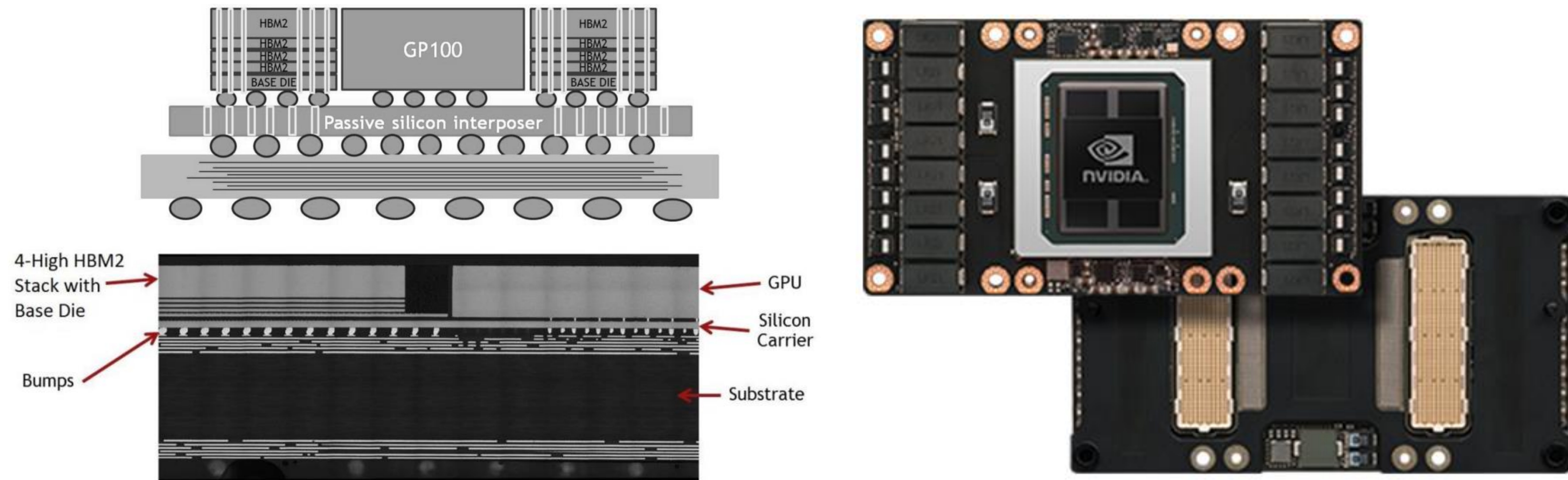
- **Why 2.5D integration?**
  Higher wire density → Lower serial rates → Better energy efficiency → edge and area BW density become critical
- **How much better can it be?**
  [9] demonstrated 50-Gb/s per lane:
  - Consuming 0.3 pJ/b
  - Edge bandwidth density of > 2 Tb/s/mm, scalable up to > 10 Tb/s/mm
  - Area BW density of 2.225 Tb/s/mm^2

[9] Y. Nishi, "A 0.297-pJ/bit 50.4-Gb/s/wire inverter-based short-reach simultaneous bidirectional transceiver for die-to-die interface in 5nm CMOS," *VLSI 2022*, pp. 154-155.

2.5D Integration Used in GPU Products

NVIDIA P100 Tesla GPU (2016)

Tesla P100 tightly integrates compute and data on the same package by adding chip-on-wafer-on-substrate (CoWoS) with HBM2 technology.



[A. Khan, "GPU-HBM SiP Interconnect Link Test and Repair," *ITC India 2019*, pp. 1-6.]
[https://www.nvidia.com/en-us/data-center/tesla-p100/]

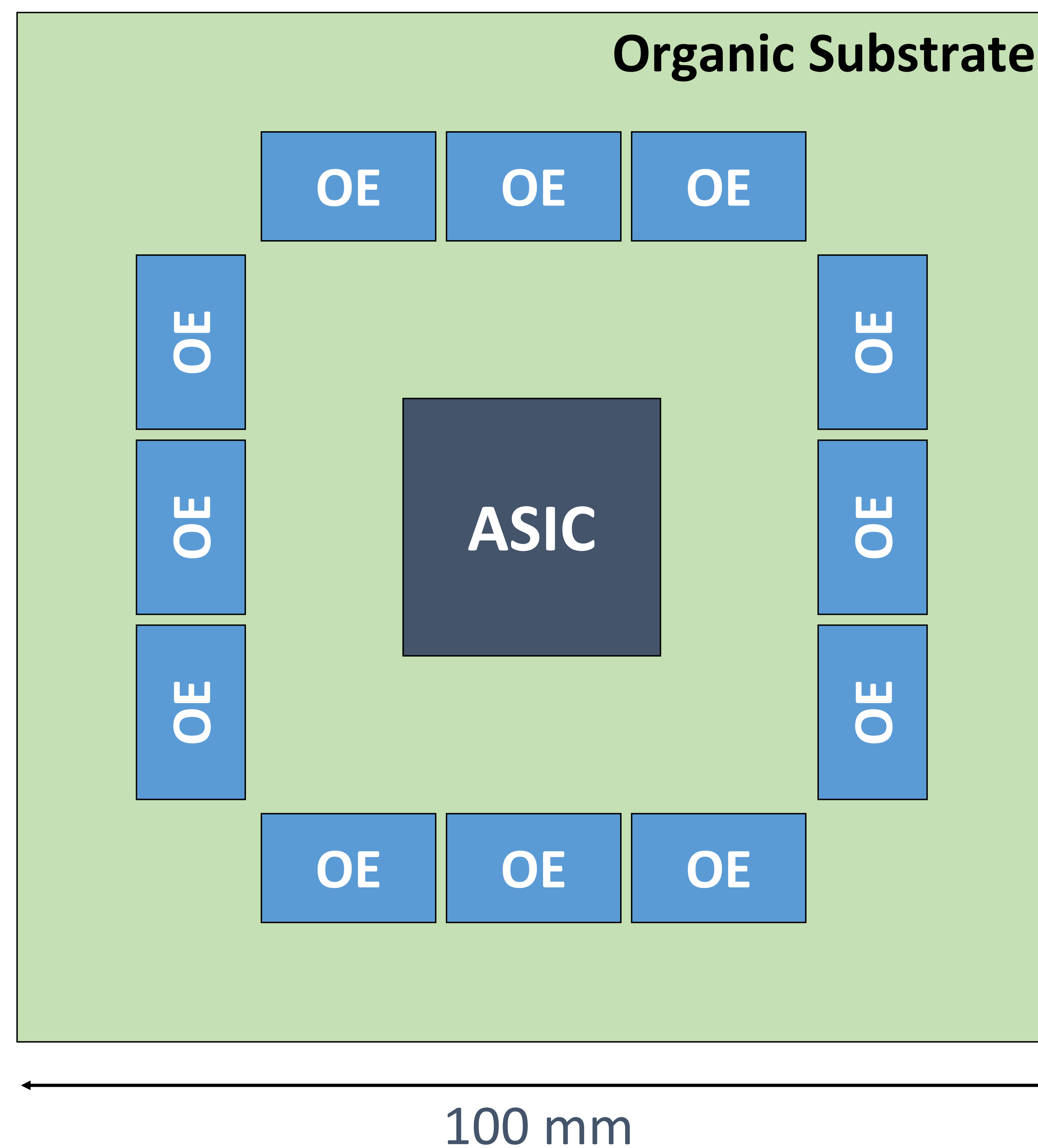
So, what are the challenges for 2.5D integrated optics?
(1) Photonics (2) Lasers (3) Packaging

- Trends in AI and the need for interconnect
- What can photonic integration gain?
- **The challenges of photonic integration**
- Accelerating photonic design with GPU and inverse design

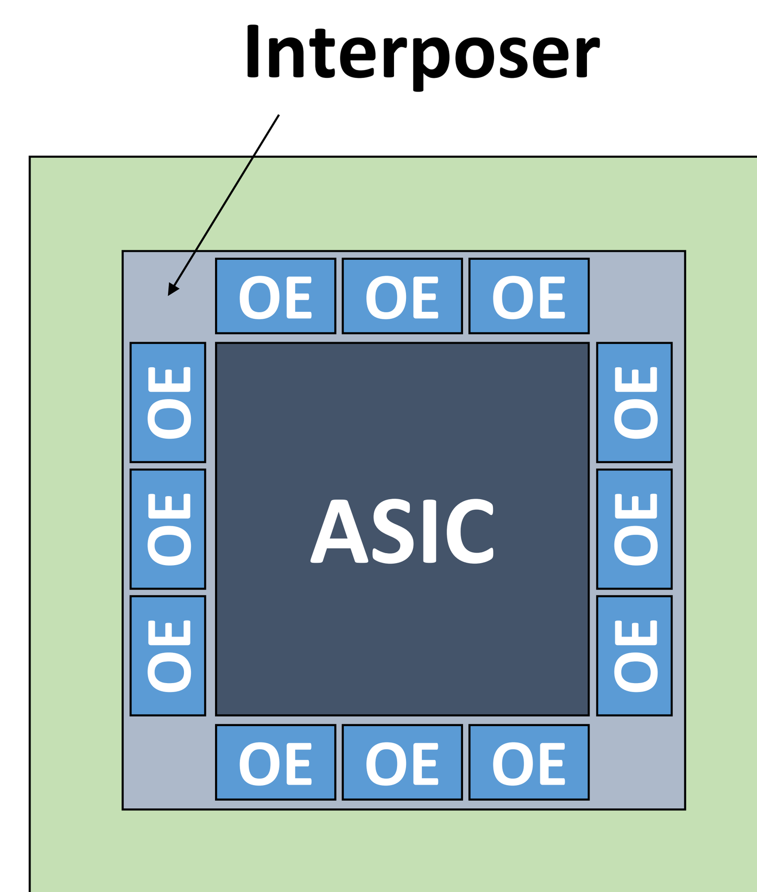
Photonics for 2.5D-Integrated Optics

Bandwidth density and energy efficiency

CPO on MCM



2.5D CPO

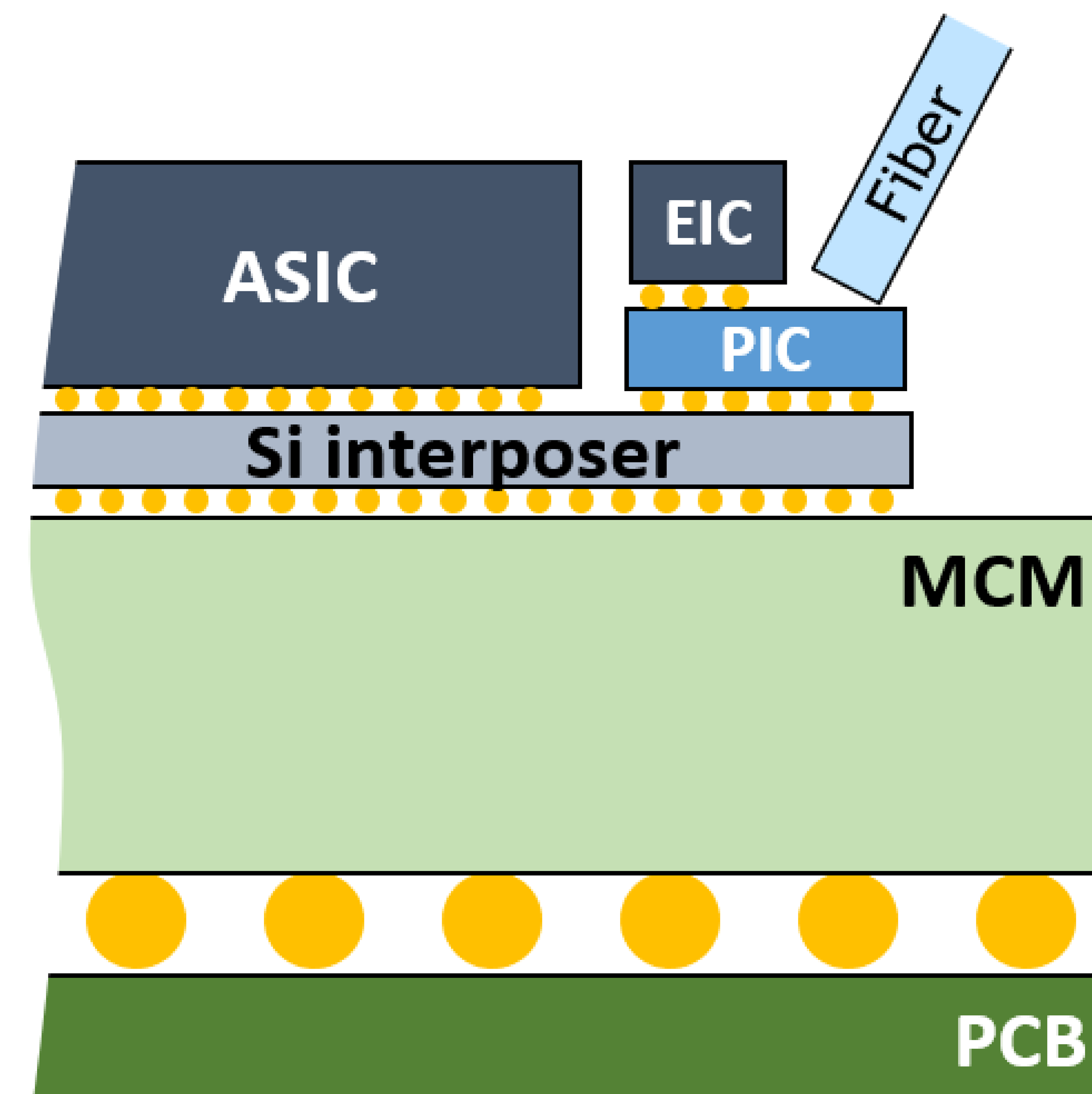


- Requires high optical edge bandwidth density
  Target: 2 + 2 Tb/s/mm
  - No optical beachfront expansion
  - At 127-µm pitch with 1:1:1 TX:RX:supply fiber ratio, need ~ 0.8 Tb/s per fiber
- Requires low-energy optics
  Target: 1.5 pJ/b
  - Package I/O power = 300 W @ 200T
  - Need efficient modulator
- Requires high areal bandwidth density
  Target: 0.5 Tb/s/mm^2
  - Need compact modulator

Packaging for 2.5D-Integrated Optics

Technology requirements

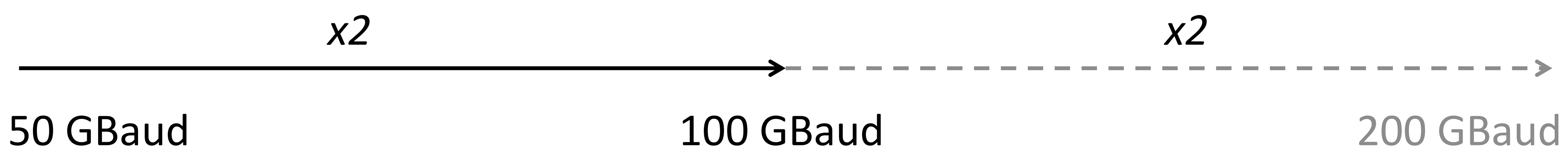
- No sockets; need very high reliability optics
- PIC with TSVs
- Edge or surface coupling, but need mechanically robust features
- Detachable optical connector
- Thermal environment

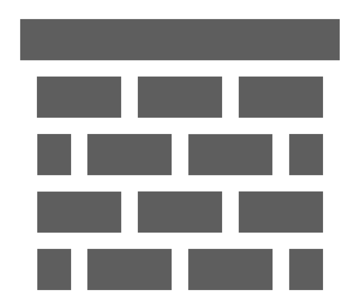


Throughput Scaling Dimensions

How to mux data?

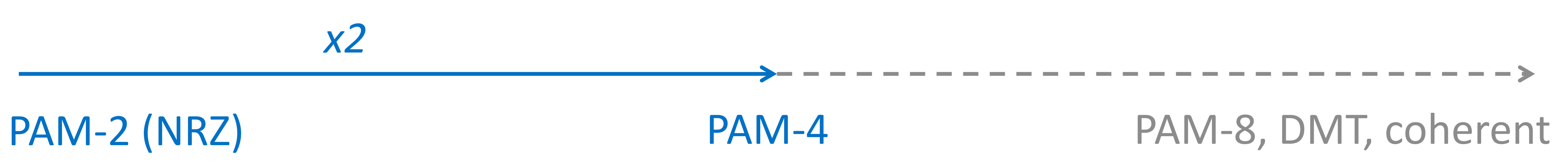
Time domain

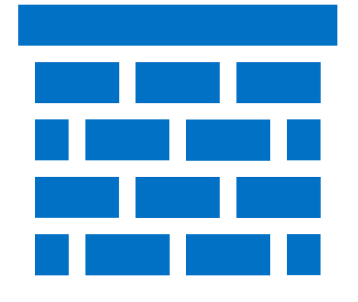




Diminishing energy efficiencies

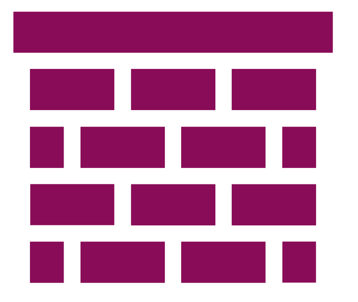
Amplitude & Phase domain





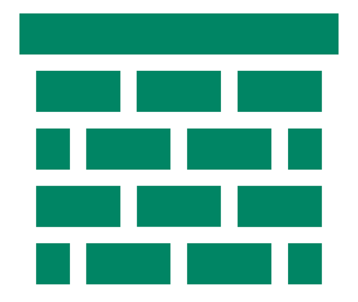
Diminishing signal to noise requires more DSP

Polarization domain



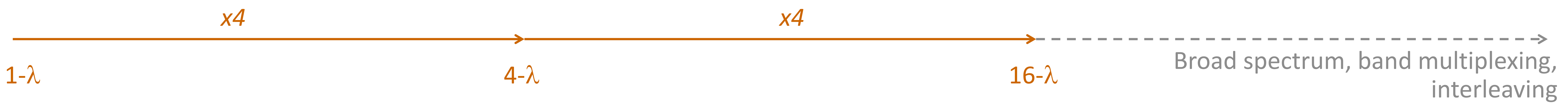
Physics

Space domain



Increasing development costs, increasing operational costs

Wavelength domain



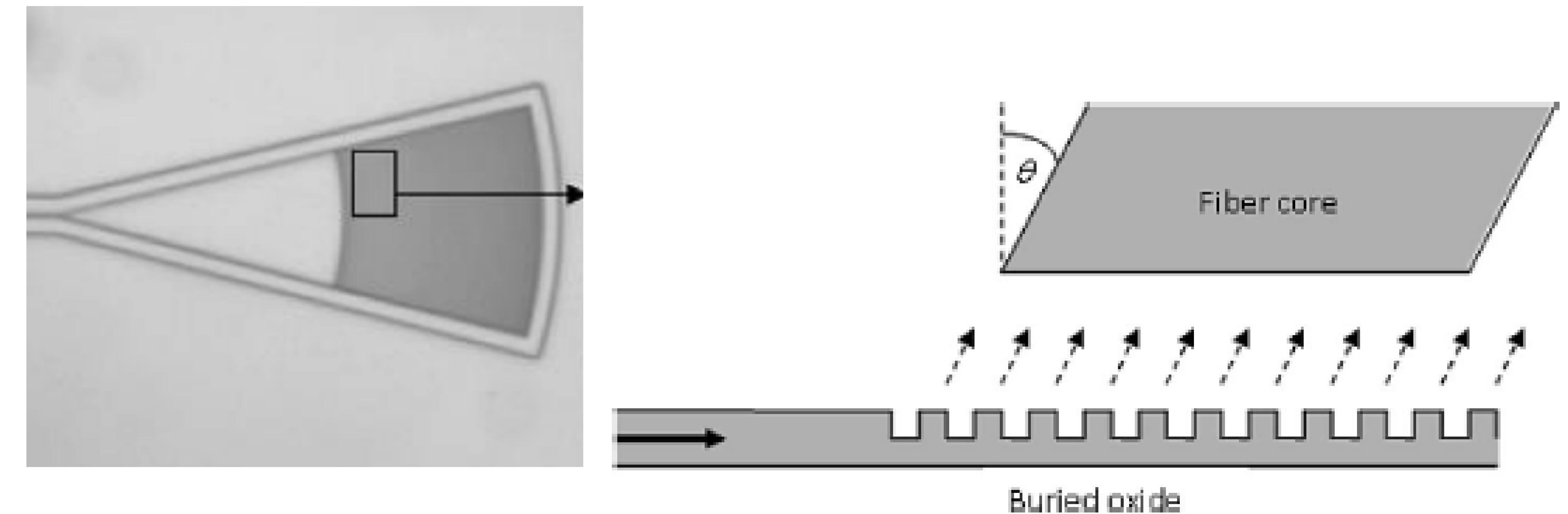
Couplers and High-Density Fiber Attachment

Optical beachfront utilization

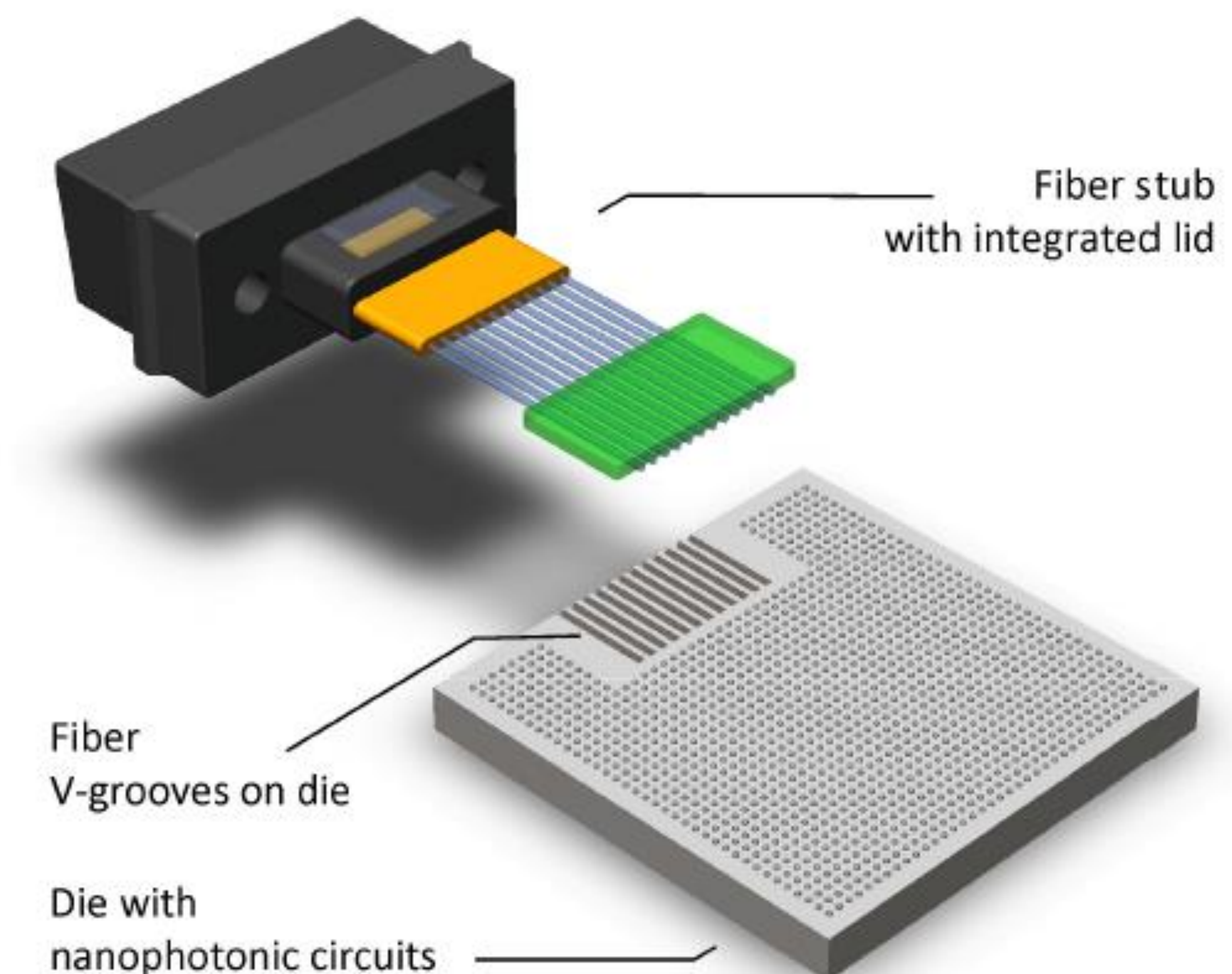
- **Grating couplers**
  - Limited spectral BW (supports 10-20 λ's),
  - Simple to fabricate
  - Allows 2D arrays of couplers → many fibers connectivity

- **Edge couplers**
  - Very wide spectral BW
  - Complicated fabrication (V-grooves, metamaterials),
  - Incompatible with die stacking (TSVs needs thinning)

- **Main Challenges & open research areas:**
  - Typically, <1-2dB loss per coupler
  - Both photonic connector solutions introduce a significant mechanical dead area on chip, limiting edge BW density



[A. Mekis et al., JSTQE2011]

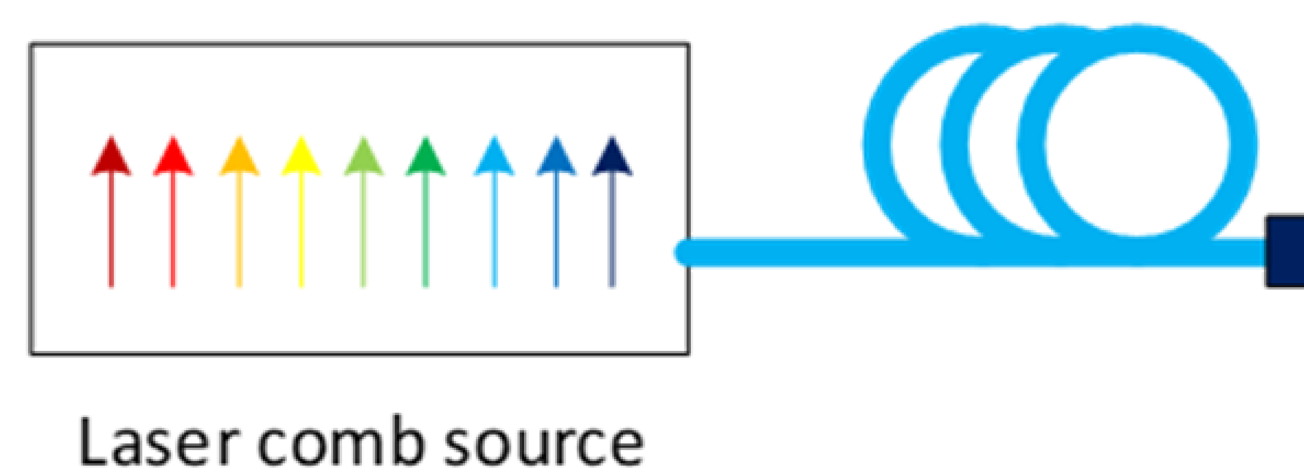


T. Barwicz et al., JSTQE2016

Photonics for 2.5D-Integrated Optics

Example of a Photonic link architecture

- Dense wavelength-division multiplexing (DWDM)
- Energy optimized lane rates ~ 25 Gb/s to 50 Gb/s
- Micro-ring resonator-based link architecture
  - Bandwidth scalable
  - Energy efficient
  - Area efficient



[B. Dally, Hot interconnect 2023]

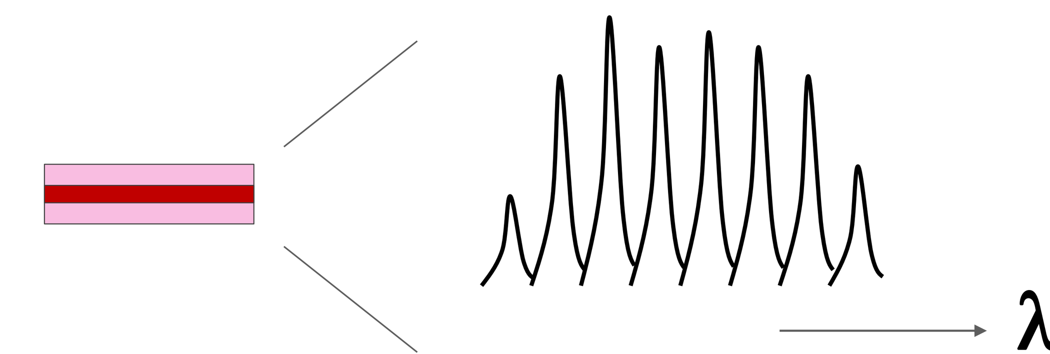
Lasers for DWDM Photonics

Working with multiple partners on various options

Singlet DFBs

Requires many precision active alignment

QD Mode-Locked Combs

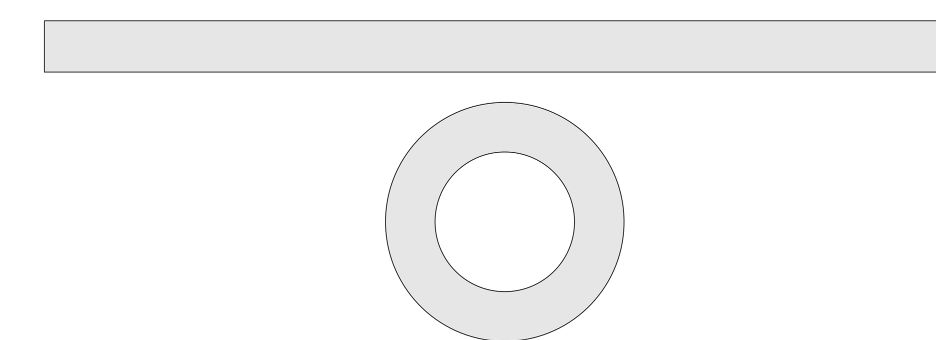


Balancing cavity gain and mode spacing.

DFB Arrays

Array yield present a significant challenge

Pumped Nonlinear Resonant Combs

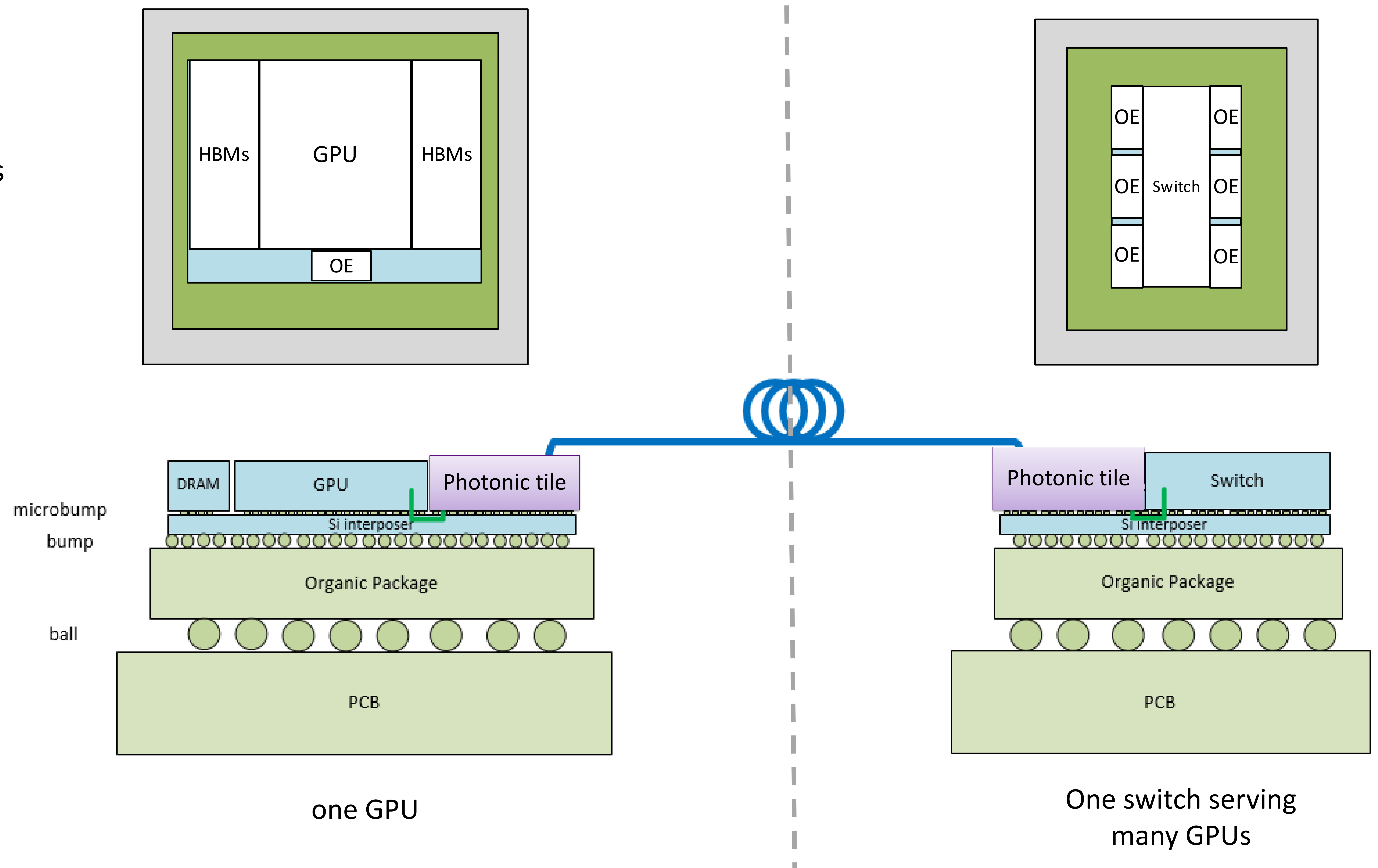


Achieving sufficient efficiency from the nonlinear process.

Photonically Connected ASICs - Challenges

EIC/PIC Optical Engine on Silicon Interposer - summary

- Interposer integration removes bandwidth bottleneck
- Die stacking minimizes interposer area, losses and Xtalk, enables process mixing
- **Challenges:**
  - **Energy and area efficient components**
  - **Comb lasers**
  - **Loss management**
  - **Polarization handling**
  - **Thermal management**



NVIDIA

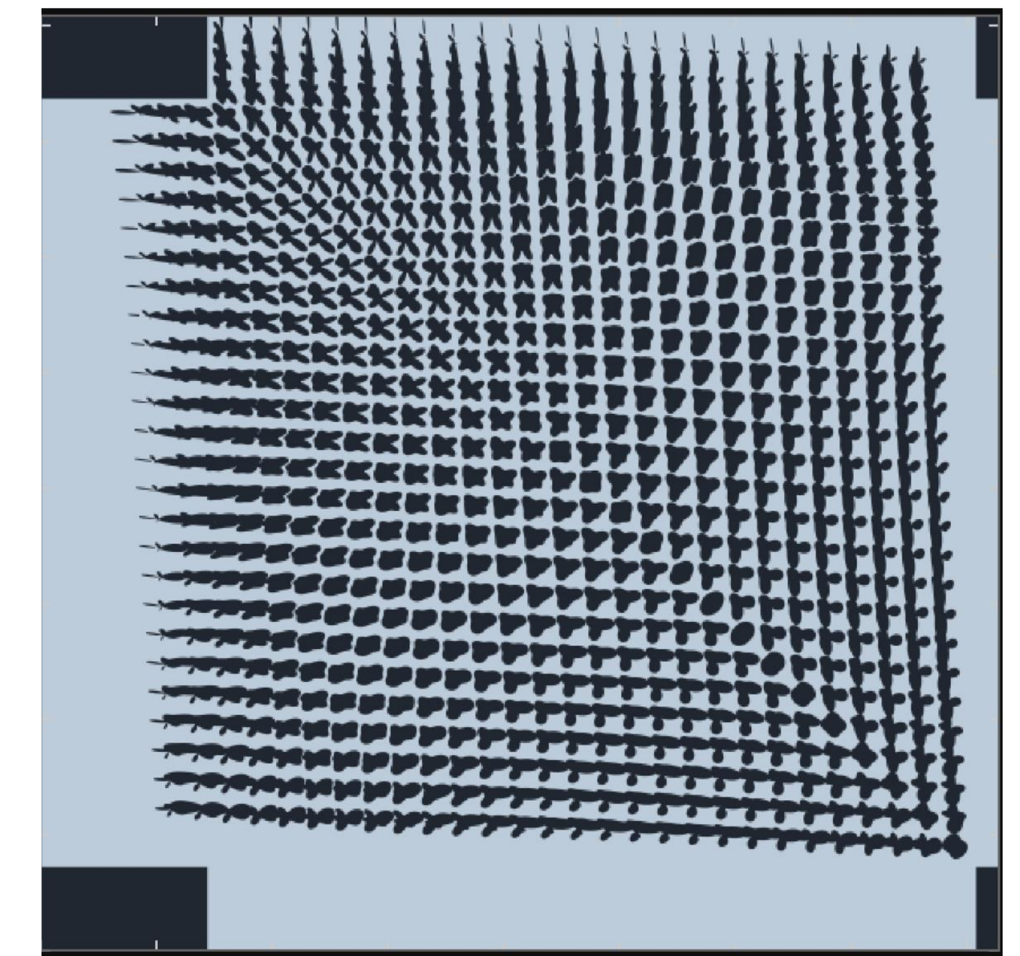
- Trends in AI and the need for interconnect
- What can photonic integration gain?
- The architecture and challenges of photonic integration
- **Accelerating photonic design with GPU and inverse design**

The Playground

Using inverse design to design 2D GC

- **EMopt** is an open-source EDA tool for inverse design of photonic devices
- Problem: slow to optimize large, complicated devices
- Solution:
  - Ported EMopt to GPUs for hardware acceleration
  - Developed Autograd-based algorithms for software acceleration
- **Case study for benchmark:**
  - Optical I/O device: polarization-splitting grating coupler (PSGC)
  - ~800 polygons, 325 design parameters
  - ~29 million FDTD grids

Topview Layout

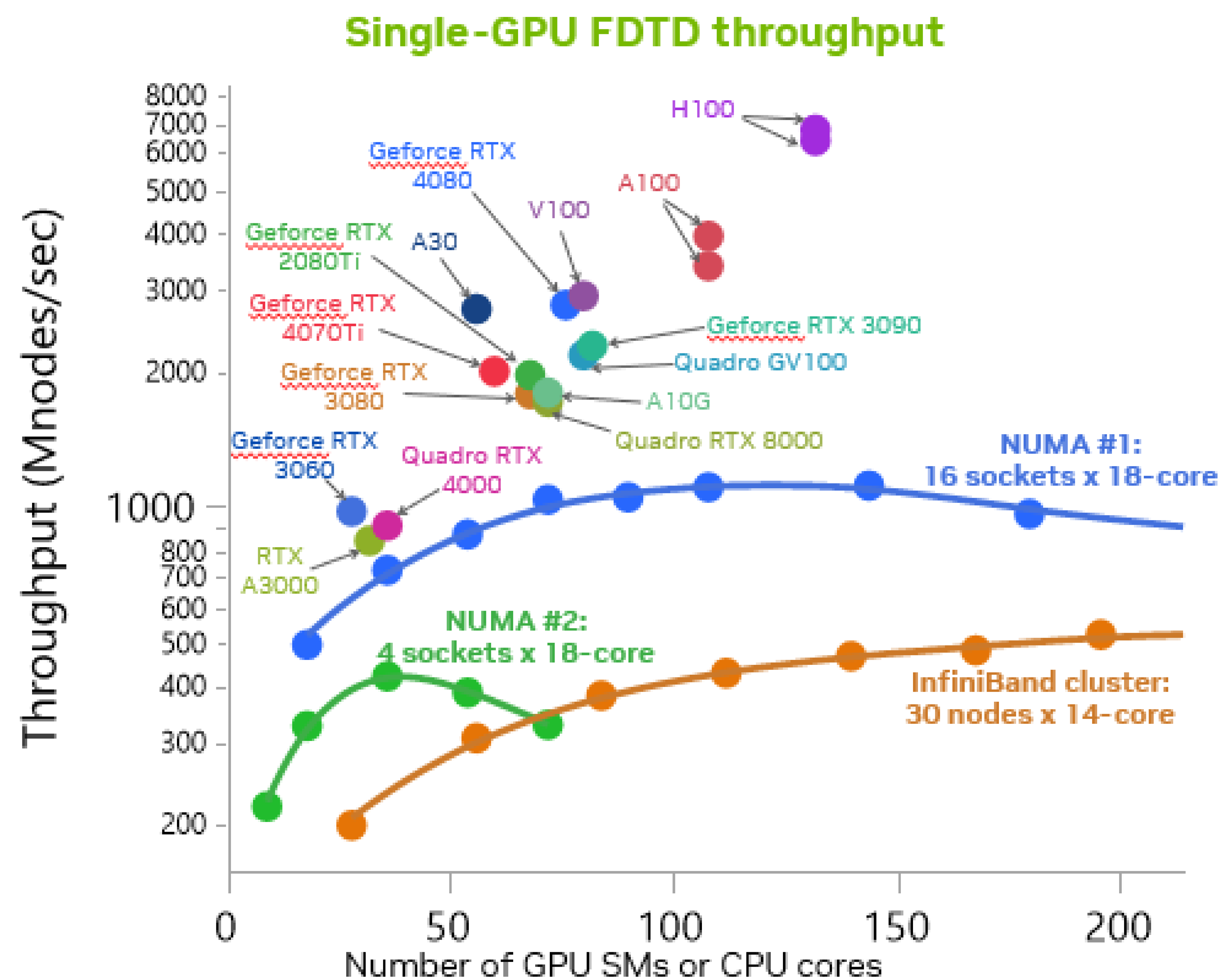


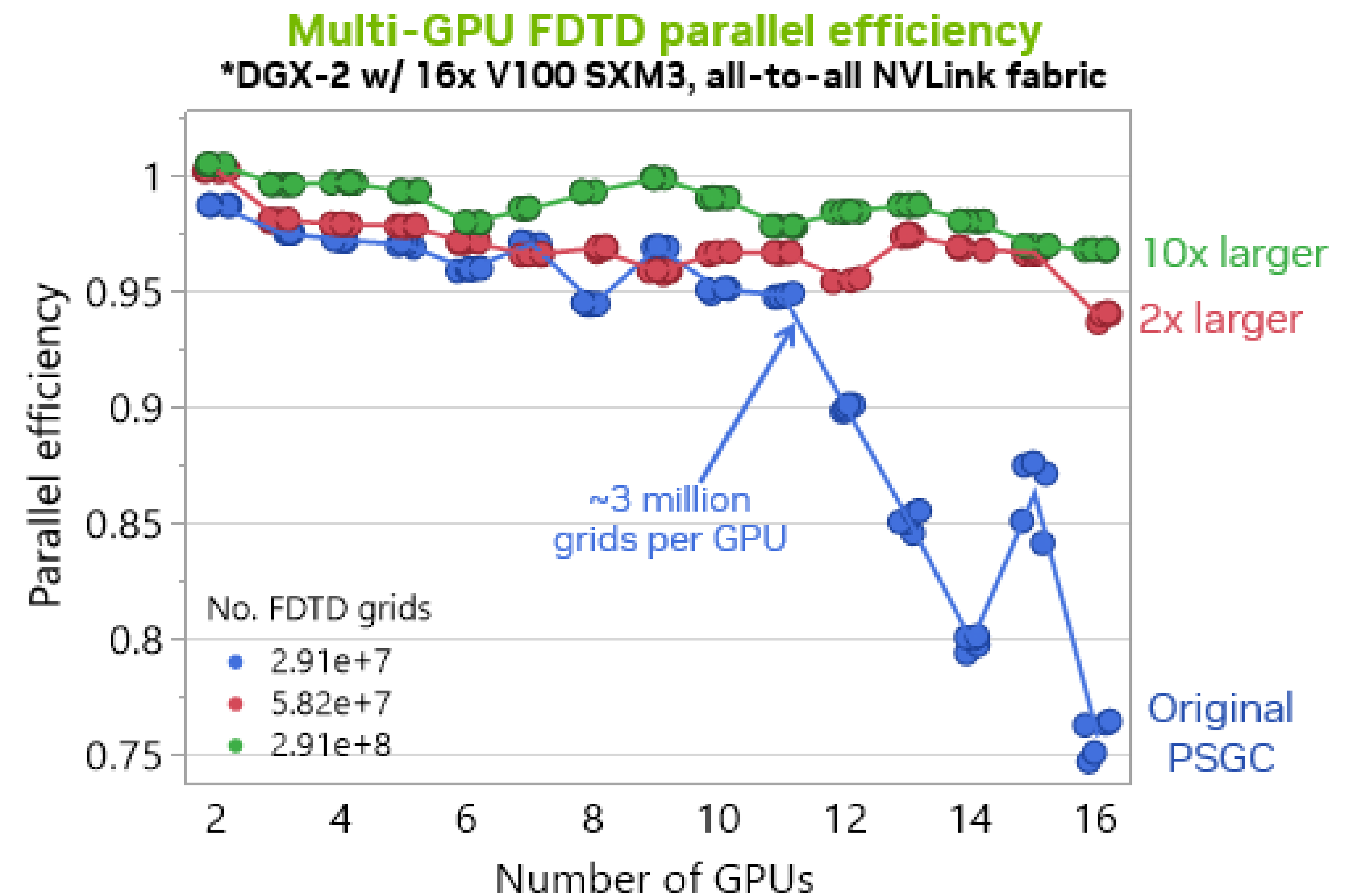
* Opt. Express 31, 31317 (2023)

GPU-Accelerated FDTD

Performance

- A single GPU can achieve higher FDTD throughput than a multi-socket CPU server
  - *Nsight Compute* profiling shows >85% peak GPU performance achieved
- Parallel efficiency ~99% on 2x GPUs, and drops to ~75% on 16x GPUs
  - *Nsight Systems* profiling shows utilization gaps when the simulation is distributed over too many GPUs





NVIDIA.

Accelerate Gradient Calculation With Autograd

Performance

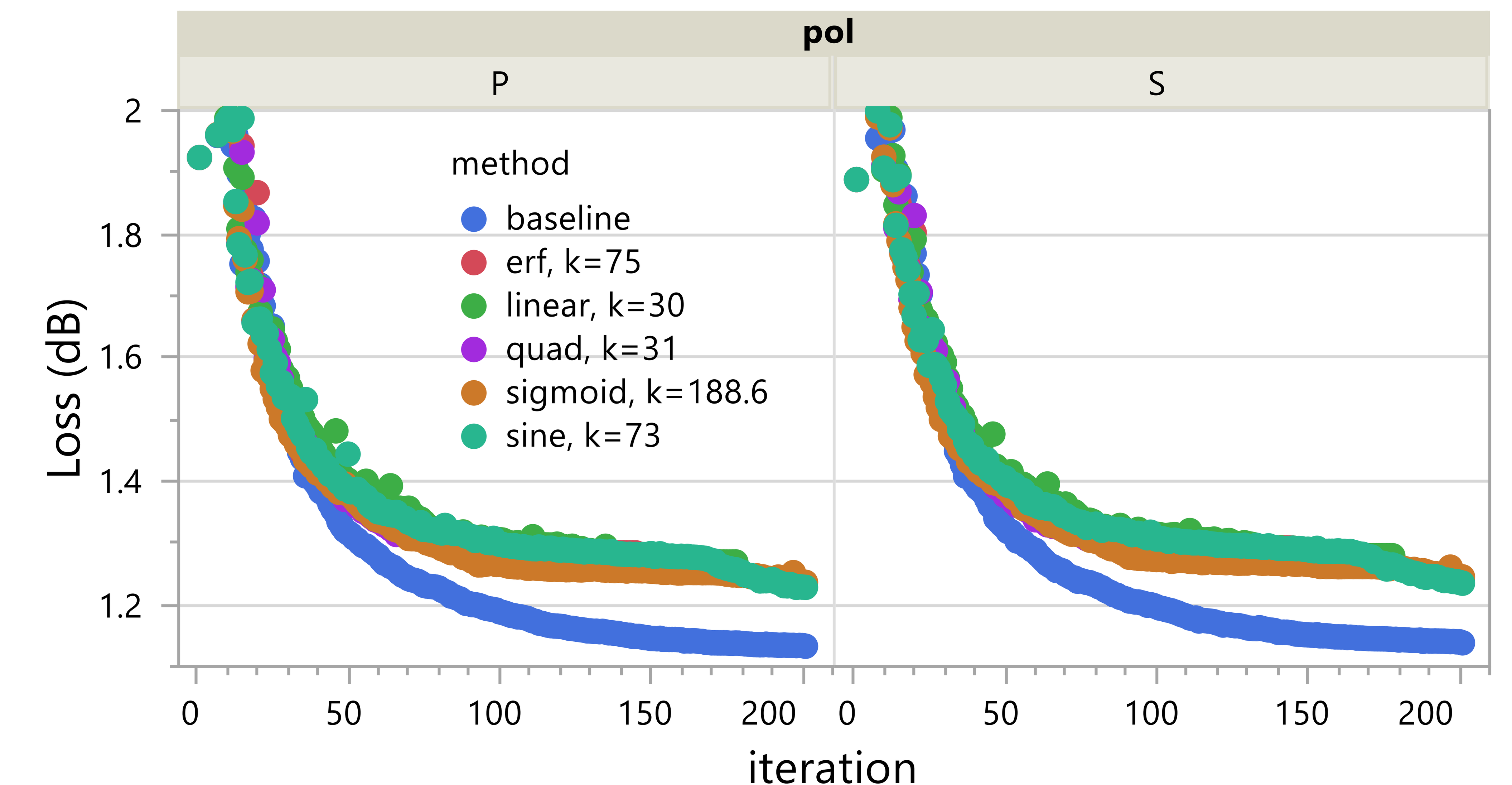
- Collaborating with Hewlett-Packard Labs: arXiv:2311.05646
- Represent the PSGC geometry as an ensemble of differentiable functions
- Use PyTorch's Autograd for automatic differentiation
    - 60x faster gradient calculation, and 10x overall speed up
- Caveat: gradient errors

Runtime of the PSGC problem

	CPU-EMopt	GPU-EMopt	GPU-EMopt w/ Autograd
FDTD	~14 min	~4 min	~4 min
Gradient	~38 min	~17 min	~0.6 min
Total	~ 52 min	~ 21 min	~5 min

FDTD becomes the bottleneck -> buy more GPUs

Convergence paths of the PSGC, various AutoDiffGeo functions



NVIDIA

lironga@nvidia.com