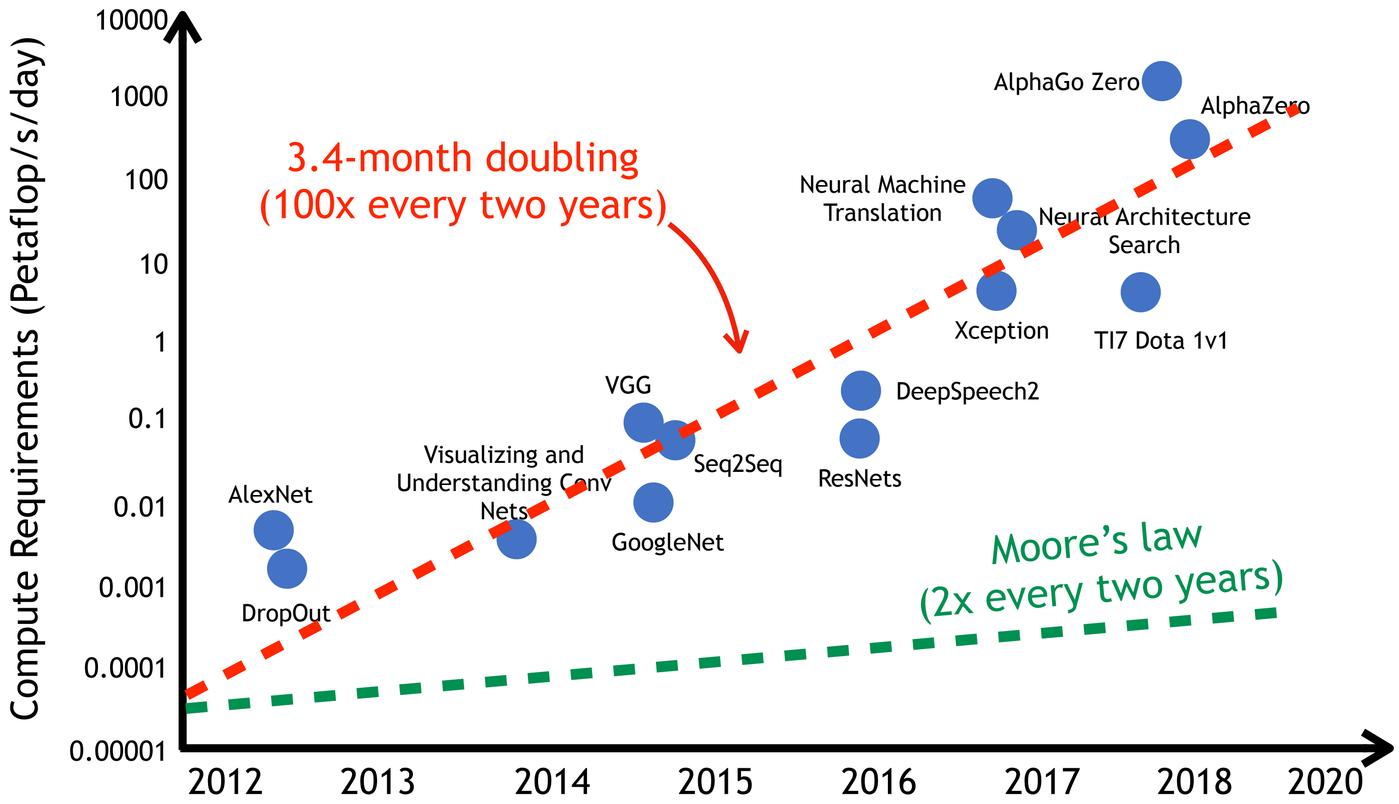
Optimizing AI Systems with Optical Technologies

Manya Ghobadi
MIT CSAIL
ghobadi@csail.mit.edu

Massachusetts
Institute of
Technology

Steady Growth of Machine Learning Models

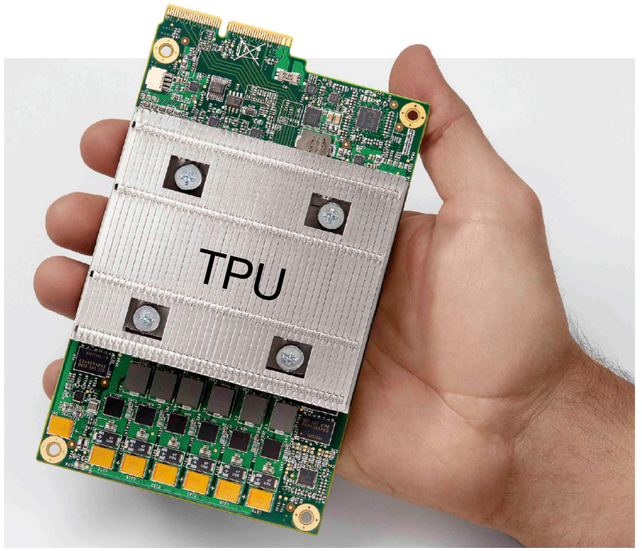


From AlexNet to AlphaGo Zero: 300,000x increase in compute requirements

Figure adopted from OpenAI
https://openai.com/blog/ai-and-compute/

The Need for Distributed ML Training

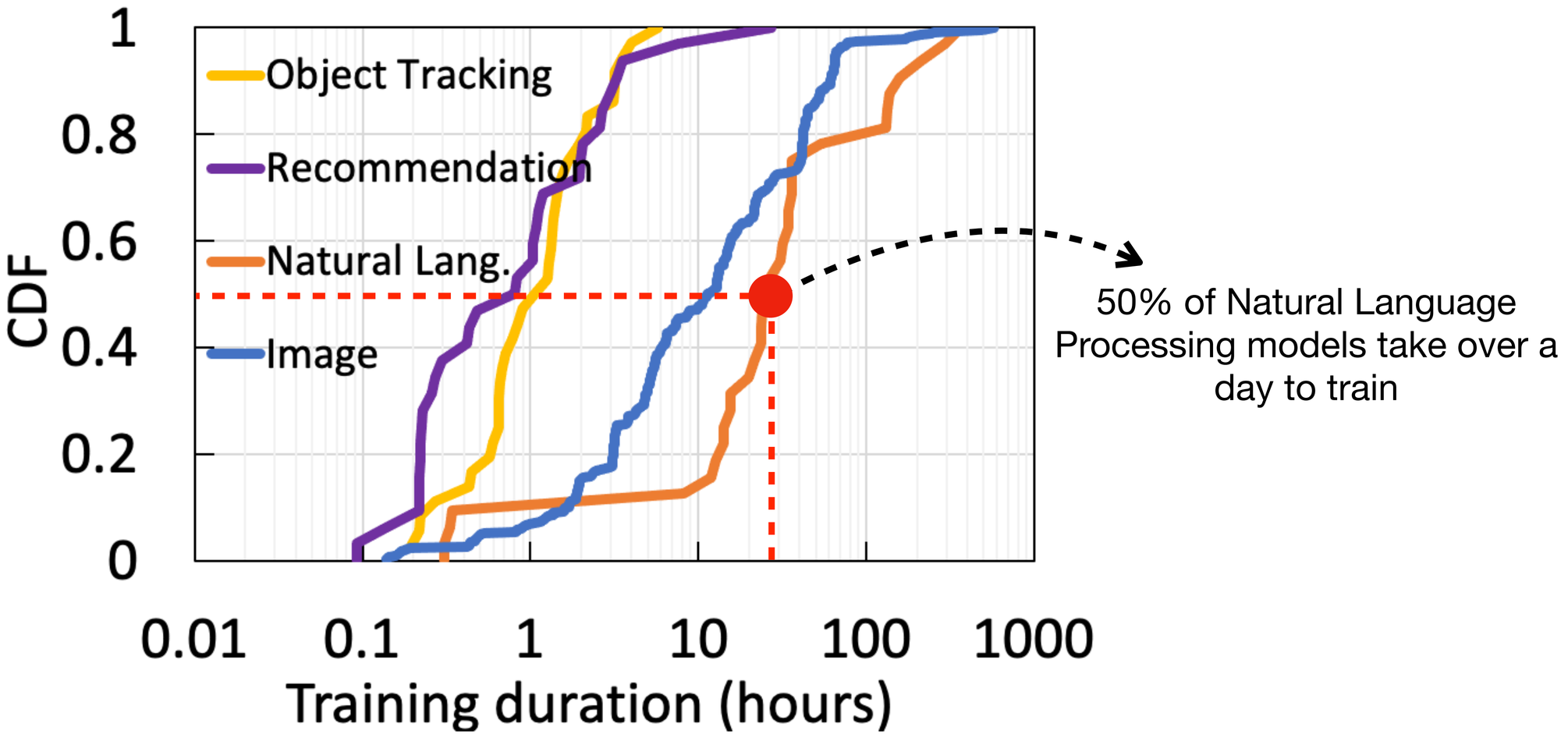
- Rapid development of hardware accelerators and software stacks.
- New models are invented daily, increasing the memory and computation requirements for both inference and training.
- Future advancements to deep learning are significantly limited by the *amount of computation and memory that can fit on a single chip package*.
- Distributed training is the key enabler for wide-adoption of ML.



Problem: Today's ML tasks still take days and even weeks to train.

Figure adopted from ACM
https://cacm.acm.org/magazines/2018/4/226374-chips-for-artificial-intelligence/fulltext

Measurements at Facebook



Where is the Bottleneck?

- Why don’t we throw more GPUs at these DNN jobs?
  - We have been! Each job is running with hundreds, sometimes thousands, GPUs.
- As the number of ML workers scale, the network bandwidth becomes a bottleneck [SIGCOMM’21].

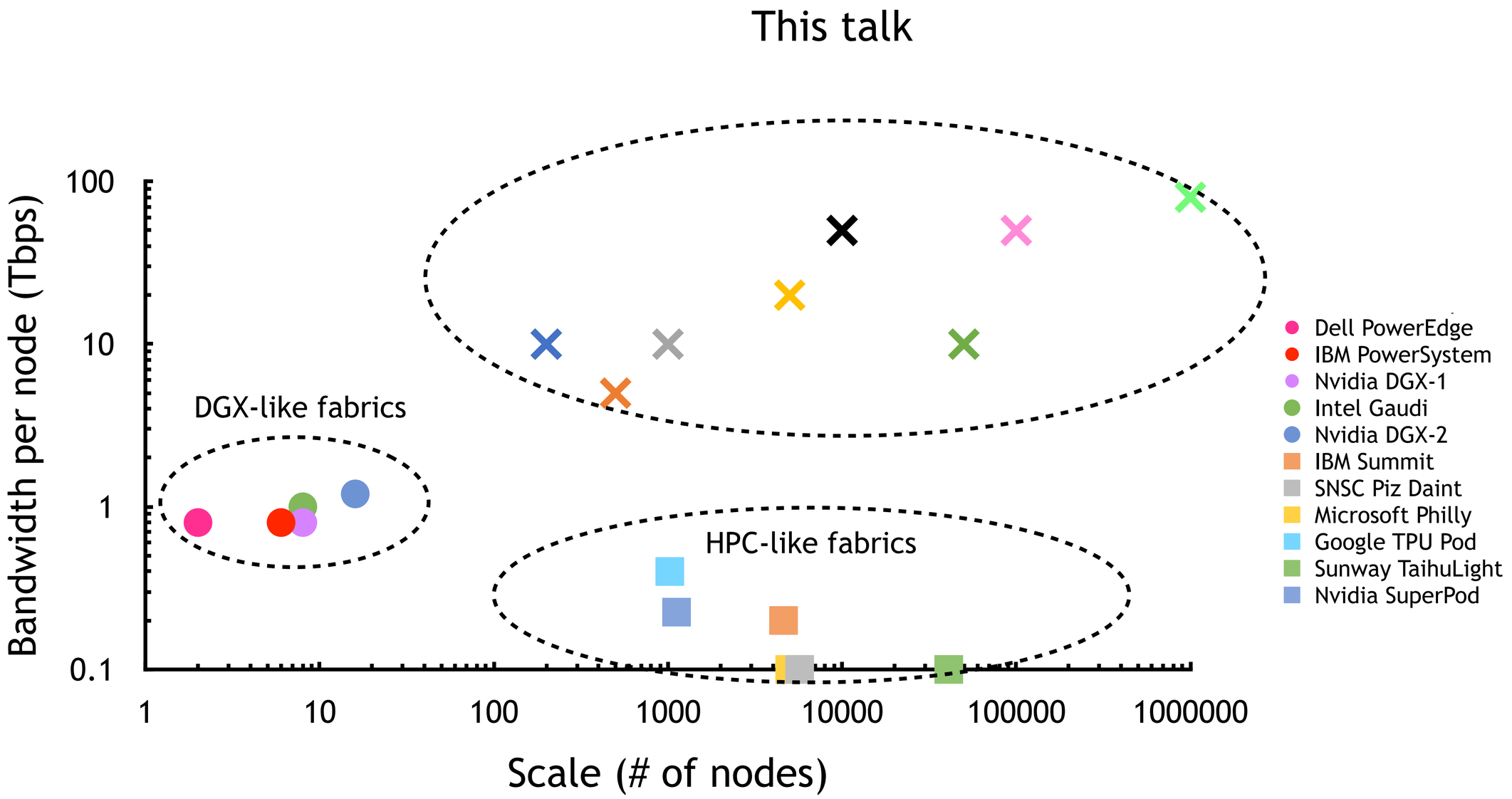
Car parts are delivered slowly

1 per minute

Can produce 1 car per second

Image source: https://www.investors.com/wp-content/uploads/2021/04/Stock-manufacturingcar-02-adobe.jpg

Vision: Next-generation DNN Training Clusters



Outline

- Key Concepts for Designing Scalable ML Training Interconnects
  - Parallelization strategies
  - Weak and strong scaling
- Network bandwidth requirements for strong scaling
- Silicon photonics
- Optical network designs, their advantages and challenges to build high-performance ML training interconnects

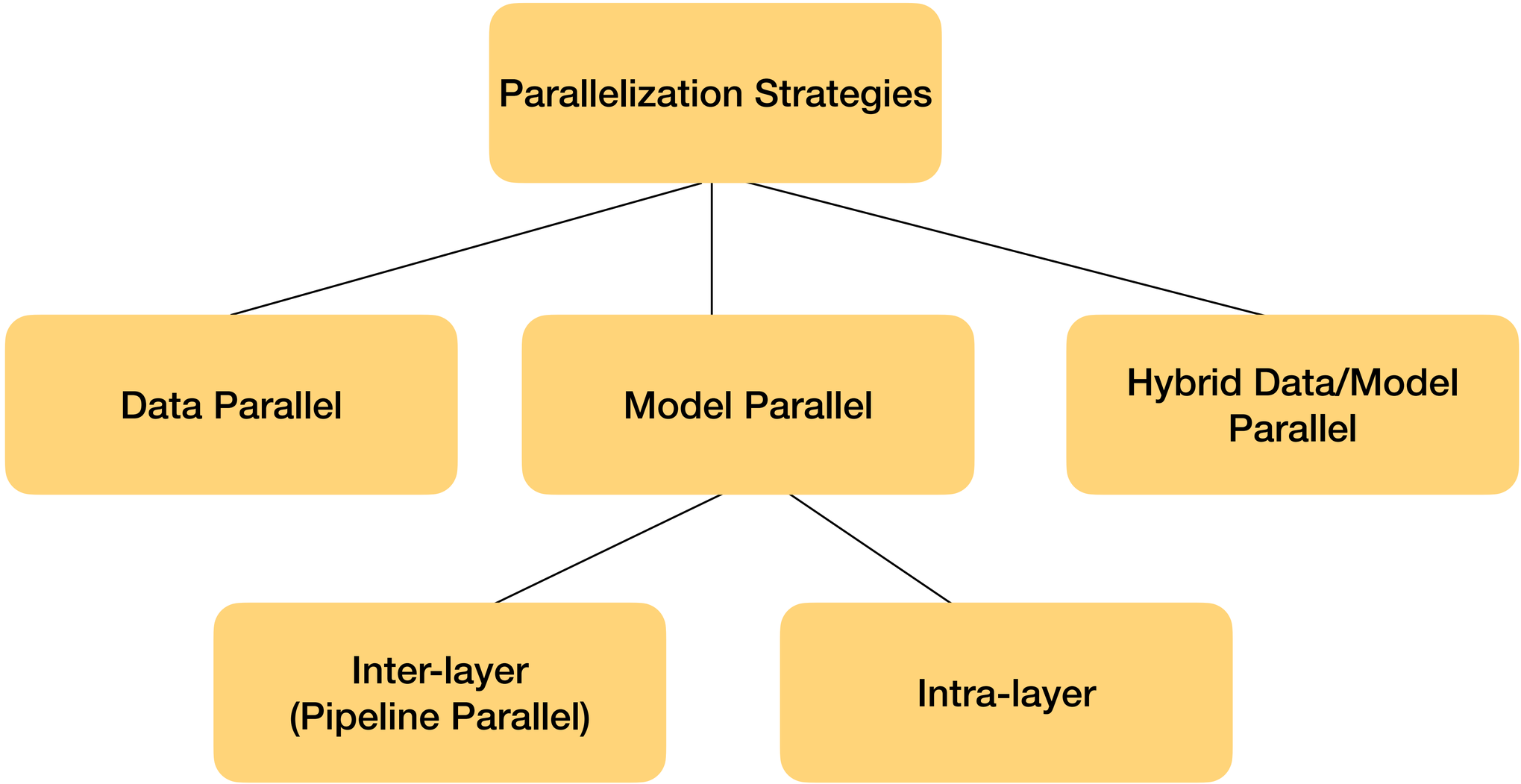
Background on Distributed Training

DNN Training

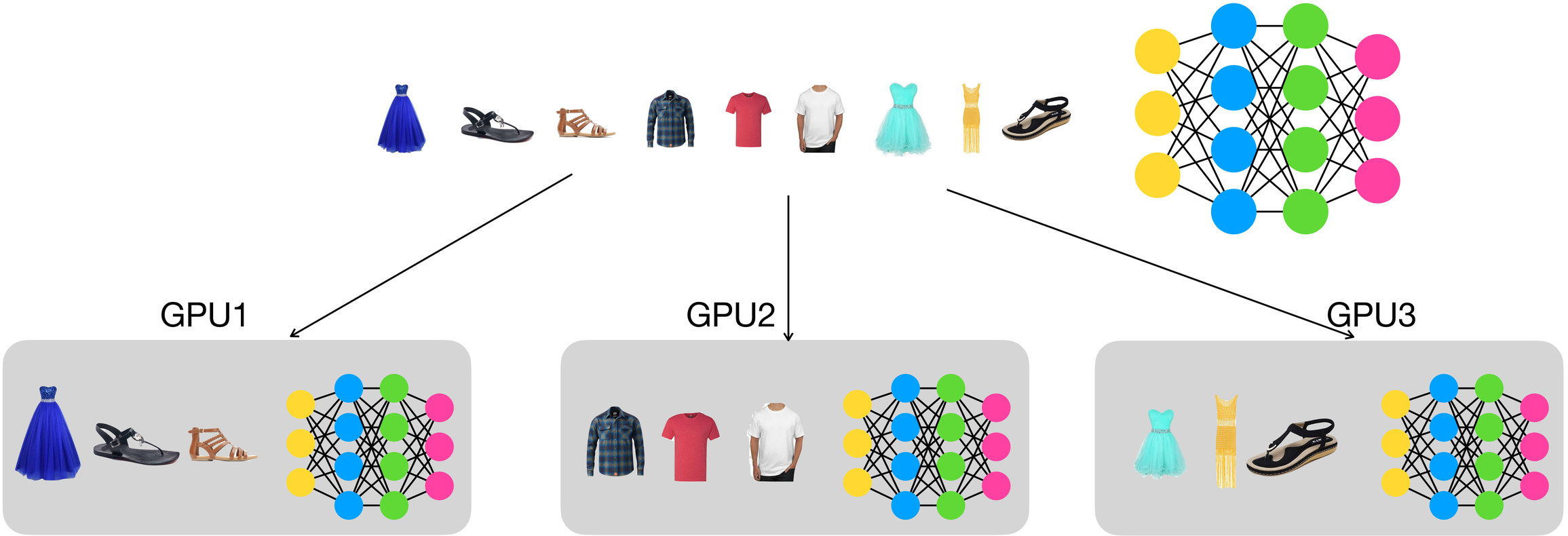
- Stochastic Gradient Descent (SGD)
- Training starts with randomly initialized weights
  - Iterate through a batch of training data samples at a time:
    - Forward pass, Backward pass, Weight update
- Three important metrics:
  - Throughput
  - Iteration time
  - Time-to-accuracy

The main goal of systems for training is to reduce the time-to-accuracy

Distributed Training

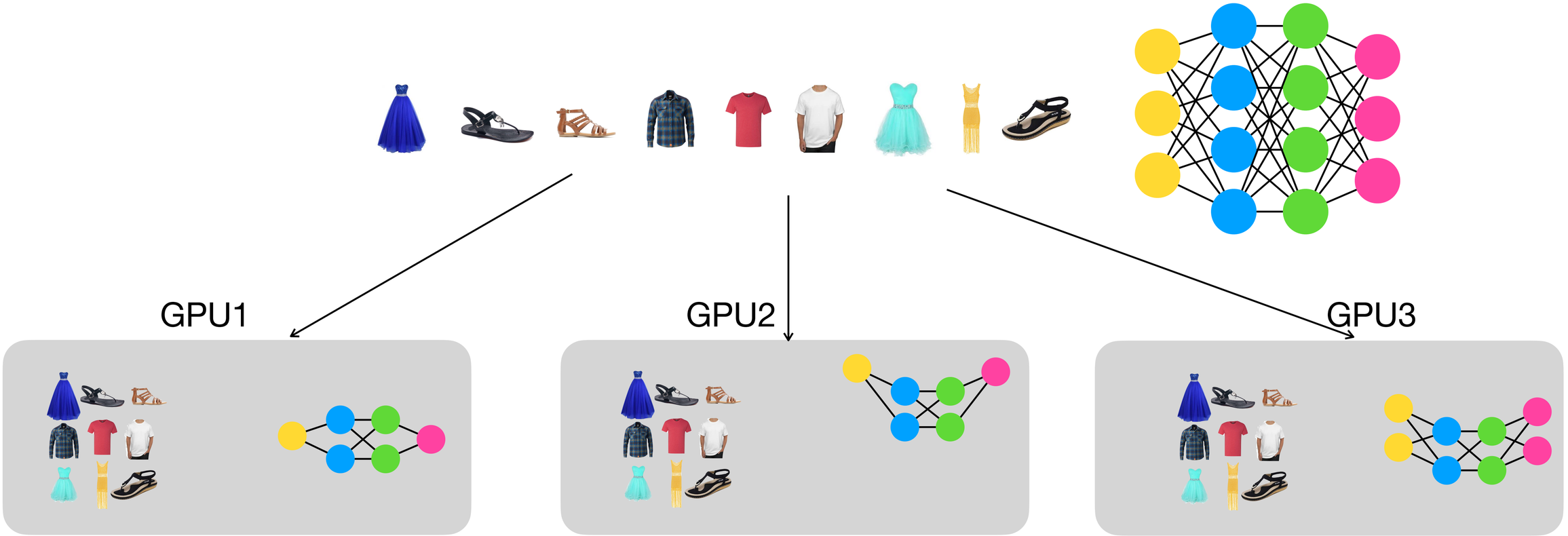


Data Parallelism: A Popular Parallelization Strategy



- At the end of every iteration, GPUs need to exchange gradient updates with each other
  - Parameter server, ring-allreduce, tree-allreduce, …
- The amount of data per GPU is proportional to the size of the DNN model

Model Parallelism: Effective but Challenging



- Within each iteration, GPUs need to exchange activation updates with each other
  - The amount of data per GPU depends on the batch size and where the model was cut.

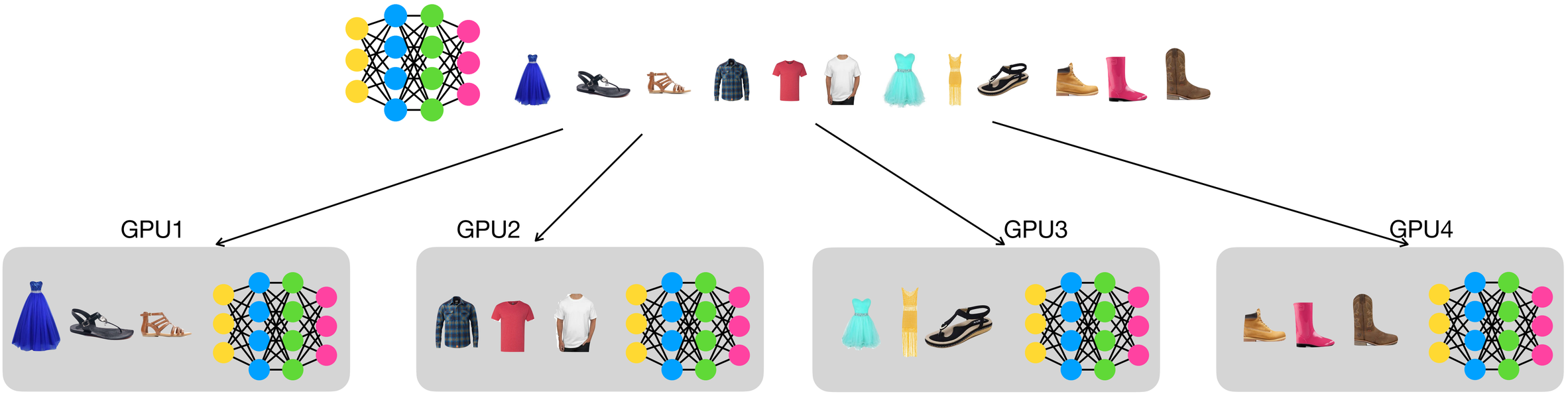
Weak and Strong Scaling of ML Jobs

Main Goal of Distributed ML Training

- Reduce the time-to-accuracy as the number of workers is scaled.
- Two dominating scaling approaches:
  - Weak scaling
  - Strong scaling
- Key insight: strong scaling requires high bandwidth.

Weak Scaling (aka Throughput Scaling)

- Reduce the *number* of training iterations by increasing the throughput of data processing (number of processed data samples/sec) as the number of workers is increased.
- Principle technique for throughput scaling:
  - Keep the local batch size per worker fixed, and grow the global batch size as more worker are added to the training job.

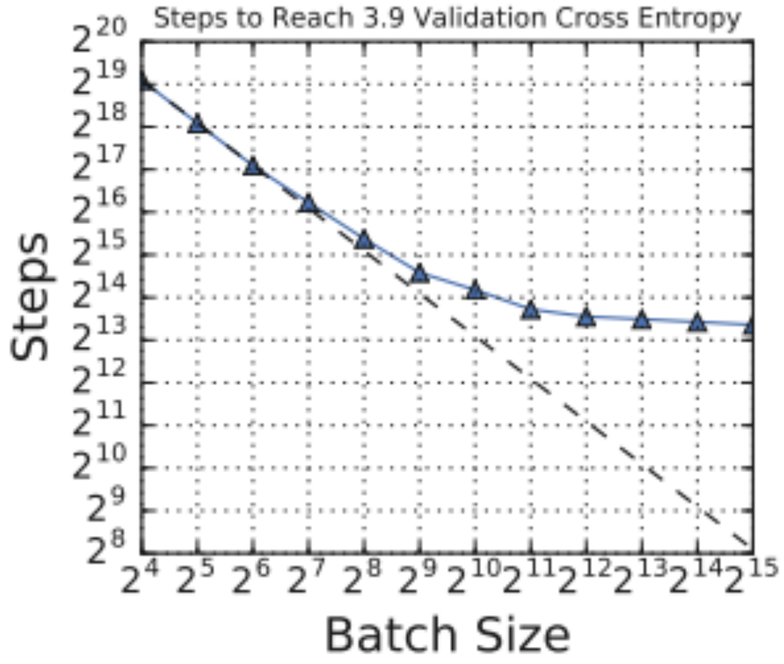


Weak Scaling (aka Throughput Scaling)

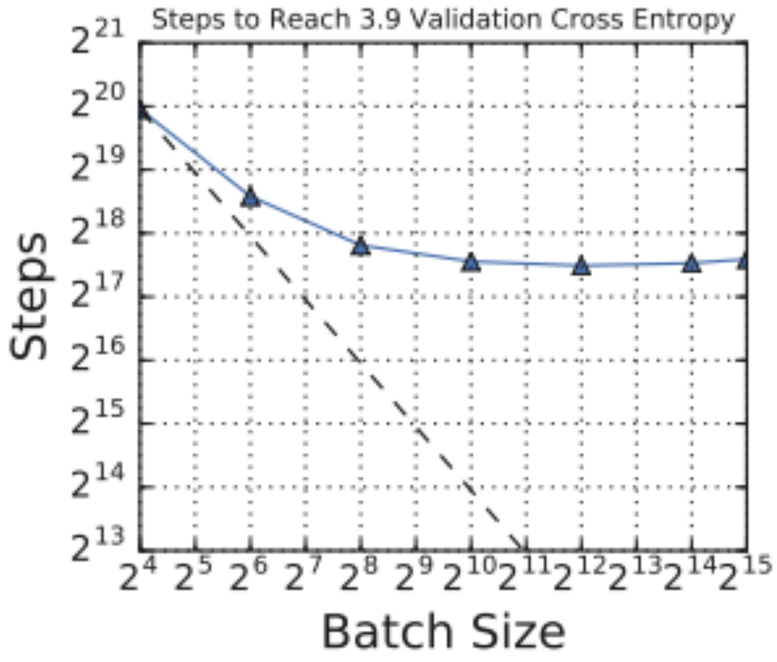
- Iteration time per worker is the same
- But the entire system is able to process a larger global batch
- It is widely believed that training with large batches reduces the time-to-accuracy
- Many systems today have been successful at demonstrating throughput scaling with thousands of worker nodes without requiring a high bandwidth interconnect
- What's the problem?

The Problem with Weak Scaling

Increasing the global batch size does not always translate to improving the number of iterations.



Transformer Model



LSTM Model

Source: Measuring the Effects of Data Parallelism on Neural Network Training
Christopher J. Shallue et al., Google
https://arxiv.org/pdf/1811.03600.pdf

Strong Scaling (aka Latency Scaling)

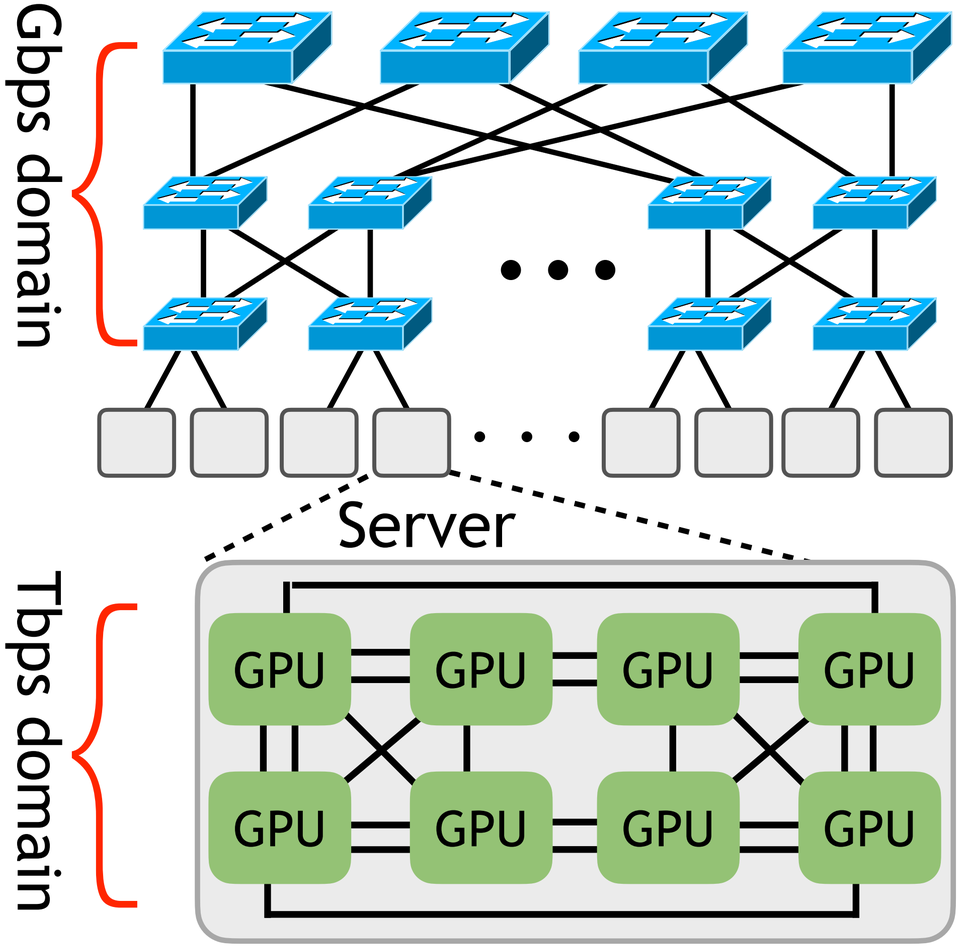
- Instead of reducing the number of iterations, reduce the iteration time as more workers are participating in the training job.
- Guaranteed to improve the time-to-accuracy.
- Strong scaling parallelizes the computation either by reducing the local batch size per worker or by partitioning the computation task across workers.
- Achieving strong scaling is challenging.

What is the Bandwidth Requirement for Strong Scaling?

- In strong scaling approaches, the bandwidth requirement increases as the system is scaled:
  - (i) Strong scaling leads into reduced computation time per worker hence the model needs more frequent updates.
  - (ii) The amount of data exchanged at each iteration depends on the model partitioning strategy.
- Today: the degree of MP has been limited to 8–32 workers within one box with Tbps communication bandwidth per GPU.

This talk will show how to enable model parallelism at 1,000-GPU scale.

The Problem with Today's Clusters



Today's interconnects are not optimal for scaling Model Parallel strategies.

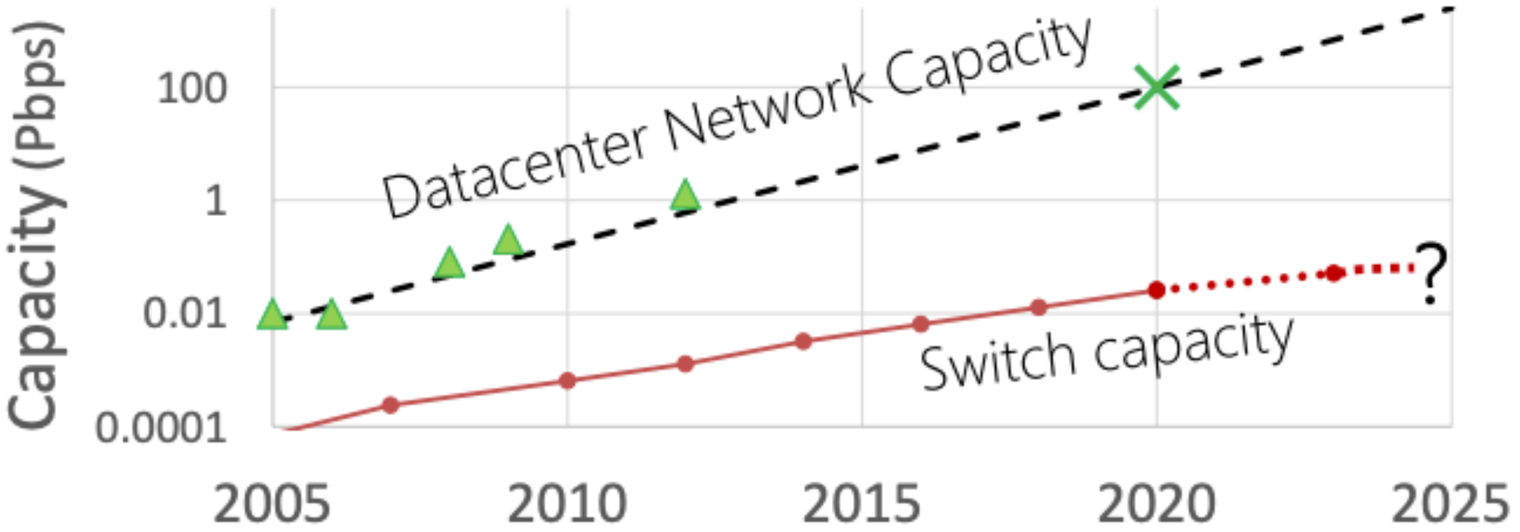
Silicon Photonics

For building the next generation of ML systems

Electrical Networks vs. Optical Networks?

- Straightforward design: electrical fabric

- Recent trends in SERDES/ packet switching technology suggest that we may hit a wall in capacity with standard electrical packet switching.



- At the same time: substantial progress with Silicon Photonics chiplets to bring optical interconnect very close (essentially on die) to the training ASICs.

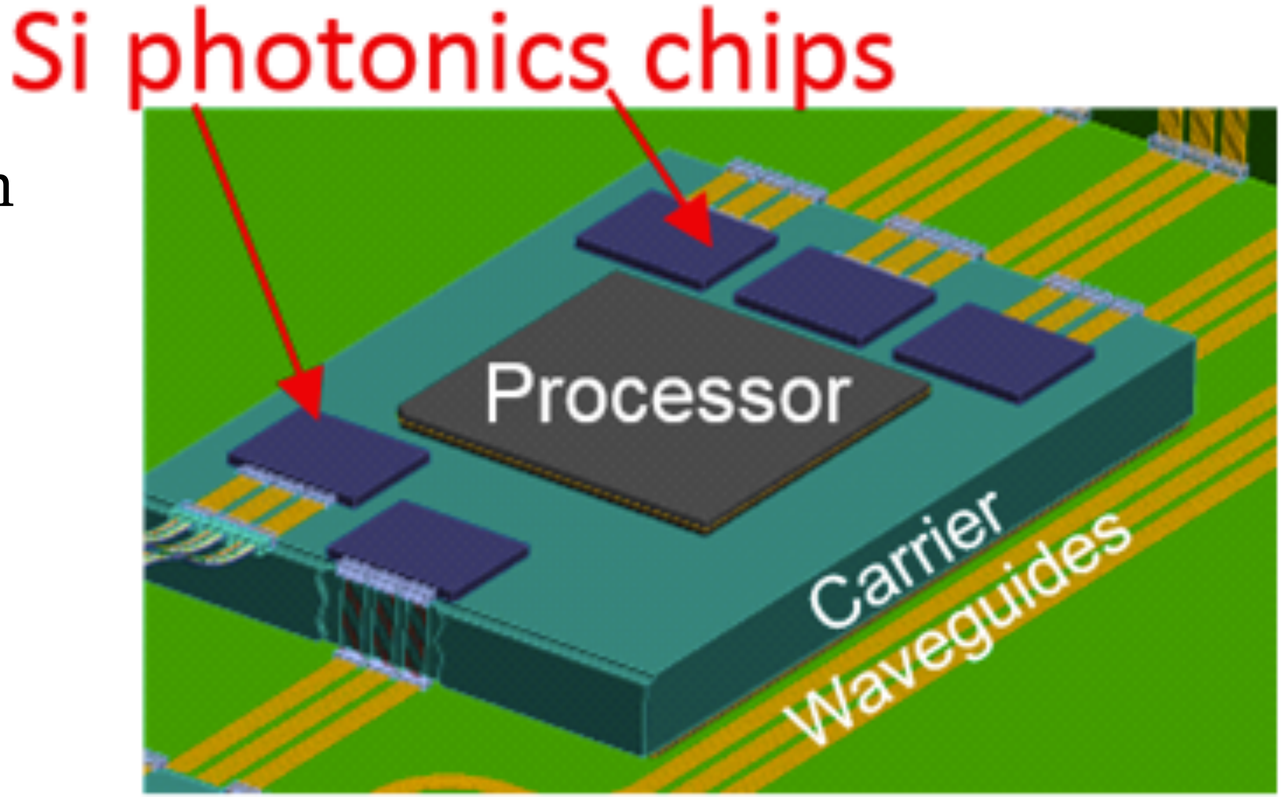
What would an optical future look like?

Figure source:
Ballani et al., Sirius: A Flat Datacenter Network with Nanosecond Optical Switching [SIGCOMM 2020]

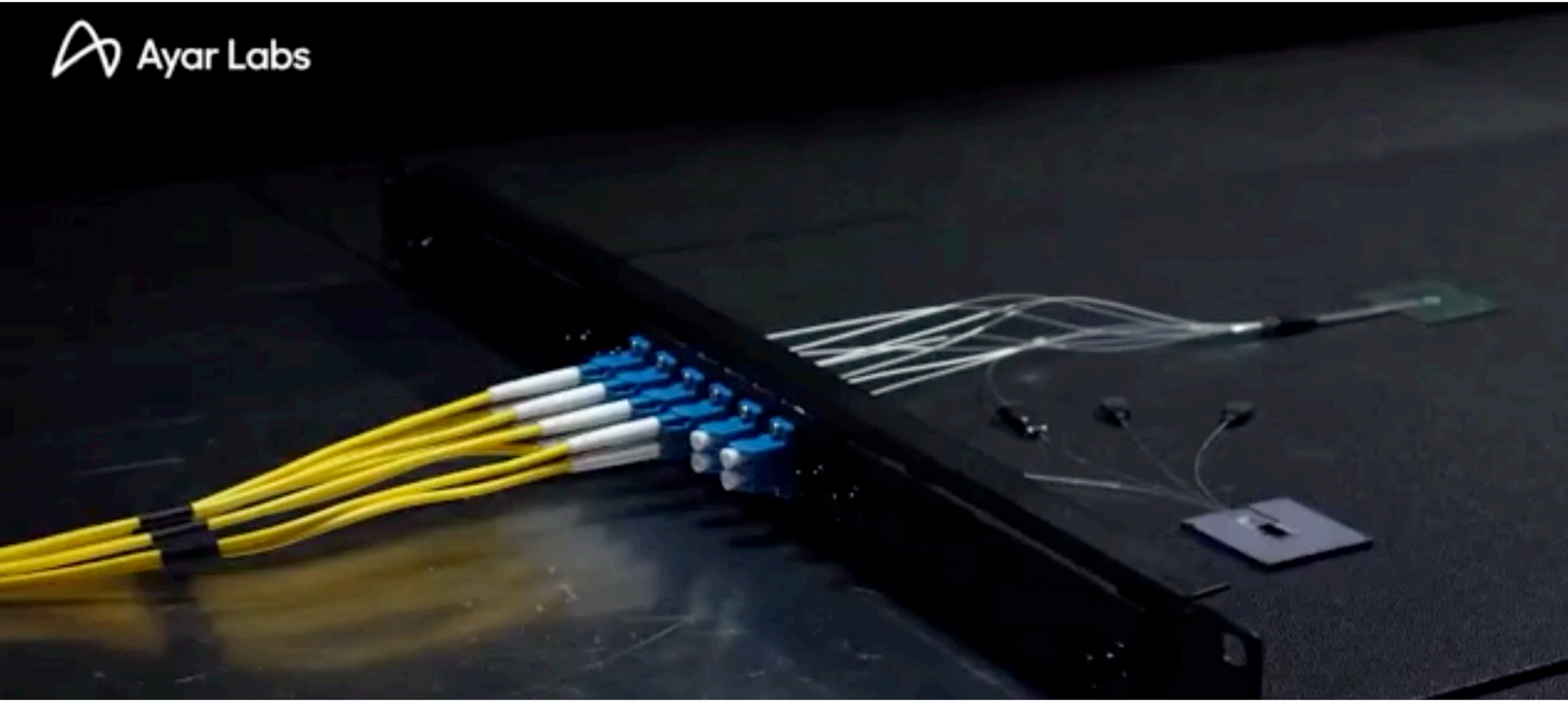
What is Silicon Photonics?

- Integrated CMOS-based silicon (Si) photonics monolithically combines electrical and optical functions on a single Si chip.

- The top candidate to provide bandwidth scalability.



Source: Adiabatic optical coupling
Bert Jan Offrein, IBM
https://www.zurich.ibm.com/st/photonics/adiabatic.html


Ayar Labs

intel

s@Intel CSR@Intel IOT@Intel Policy@Intel Technology@Intel Technology Provider Programmable Logic IT Peer Network Intel Developer Zone

PROGRAMMABLE LOGIC

Search Programmable Logic

AYAR LABS AND INTEL DEMO FPGA WITH OPTICAL TRANSCEIVERS IN DARPA PIPES PROJECT: 2 TBPS NOW, >100 TBPS IS THE GOAL

Written by Steven Leibson | March 27, 2020

- Intel's FPGA board with SiP interfaces capable of 2 Tbps I/O bandwidth
- Intel's projection is to achieve 100 Tbps I/O bandwidth integrated directly into CPU/GPU/FPGA/ASIC chiplets.

TECHNOLOGIES > ANALOG

CMOS Plus On-Chip Electro-Optical Interconnect Zooms Past 2 Tb/s

A highly advanced management of electronics, CMOS, and optical physics is pushing system-in-package designs past the terabit-per-second boundary in this DARPA-sponsored project with Ayar Labs and Intel.

Bill Schweber
AUG 19, 2020

Optical I/O interfaces

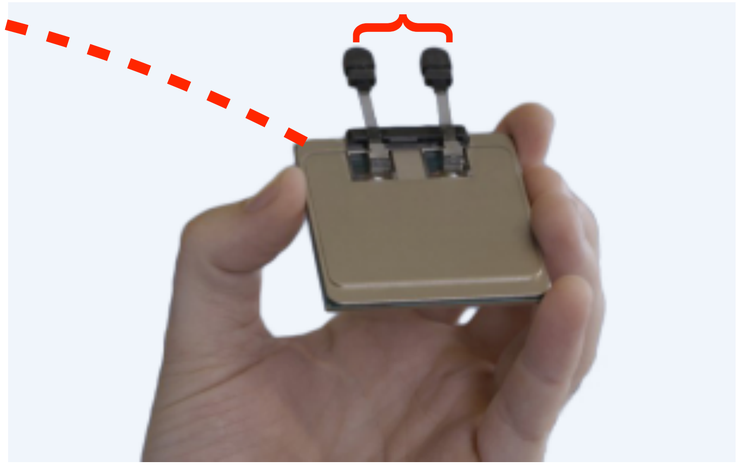
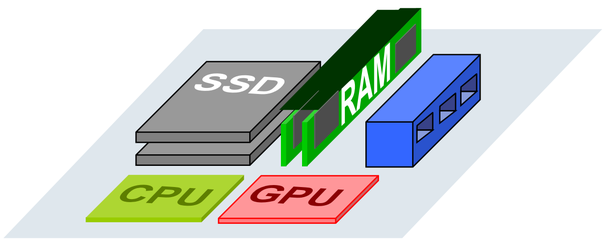


Image source: Intel
https://blogs.intel.com/psg/ayar-labs-and-intel-demo-fpga-with-optical-transceivers-in-darpa-pipes-project-2-tbps-now-100-tbps-is-the-goal/

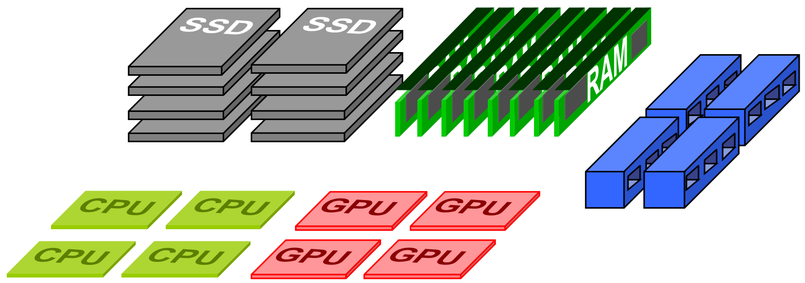
Reconfigurable Optical Links for Next-generation Clusters

- Tbps SiP I/O integration enables building next-generation computer architectures that are fundamentally impossible with today's technologies.
- Significant benefits from reconfigurable optical links.

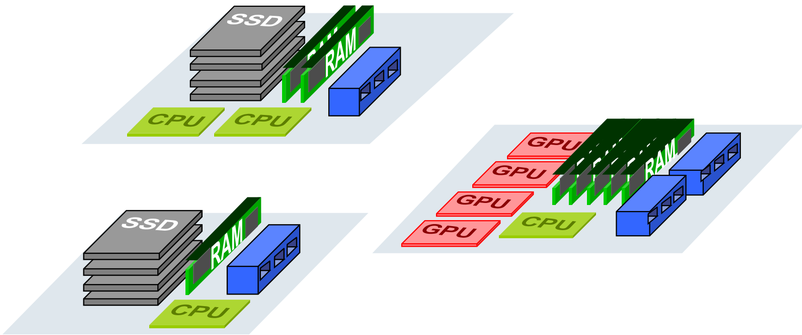
Current server



Disaggregated rack



Pool and compose



[IEEE Optical Interconnects'18]

Reconfigurable Optical Links and Hogwarts Grand Staircase

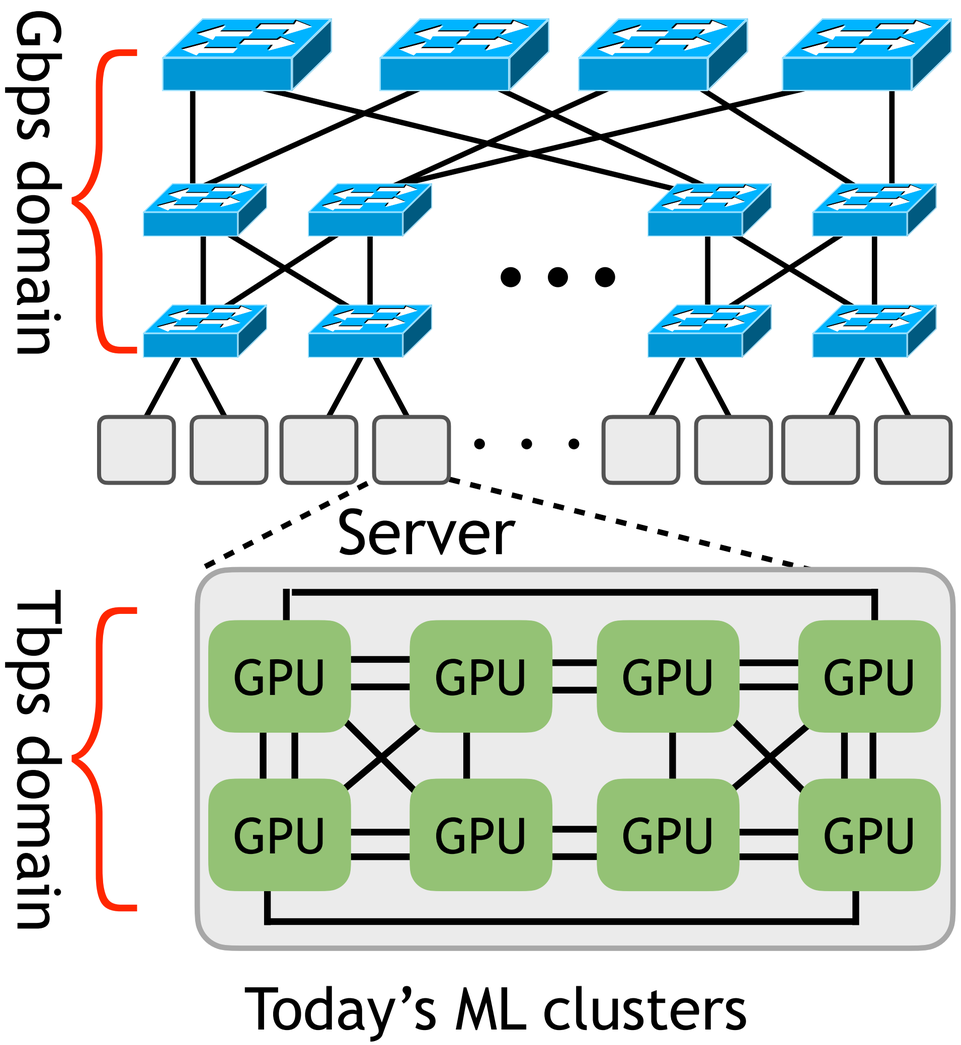
Movie Scene: Harry Potter and the Sorcerer's Stone
Credit: Fred Douglis, https://spectrum.ieee.org/tech-talk/computing/hardware/darpa-supercomputer-network-interface

Match Made in Heaven: ML workloads and Optical Interconnects

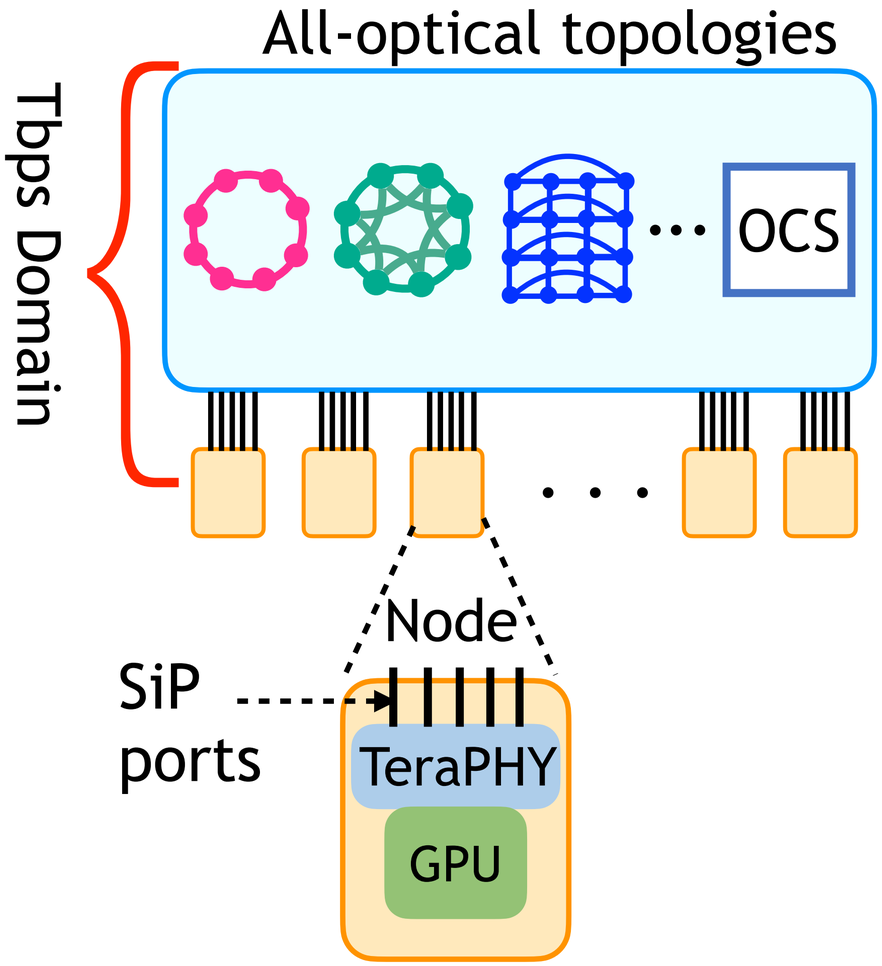
- Several optical proposals in the past decade to address the bandwidth growth of data center workloads.
- ML workloads open up new possibilities to build specialized circuit-based interconnects.
- Conventional datacenter workloads:
  - Unpredictable, mostly short flows.
- ML workloads:
  - Predictable, mostly large transfers.

The parallelization algorithm determines the circuit schedules and the entire training repeats the same communication pattern.

Novel Data Center Architecture

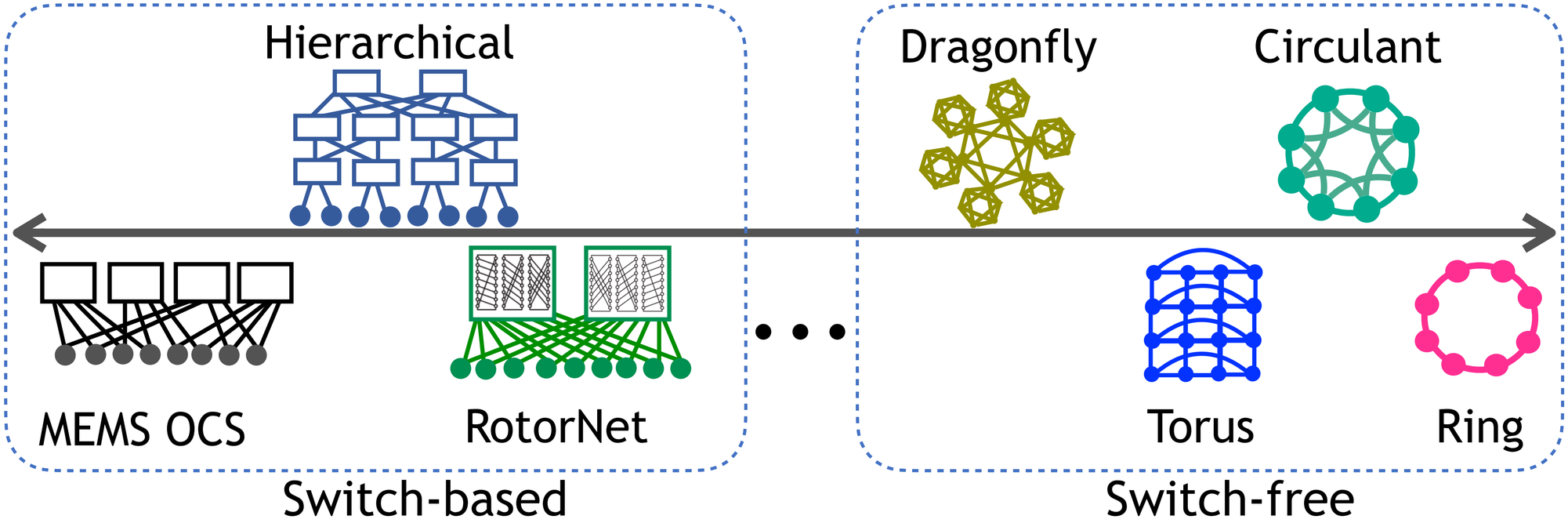


Today's ML clusters

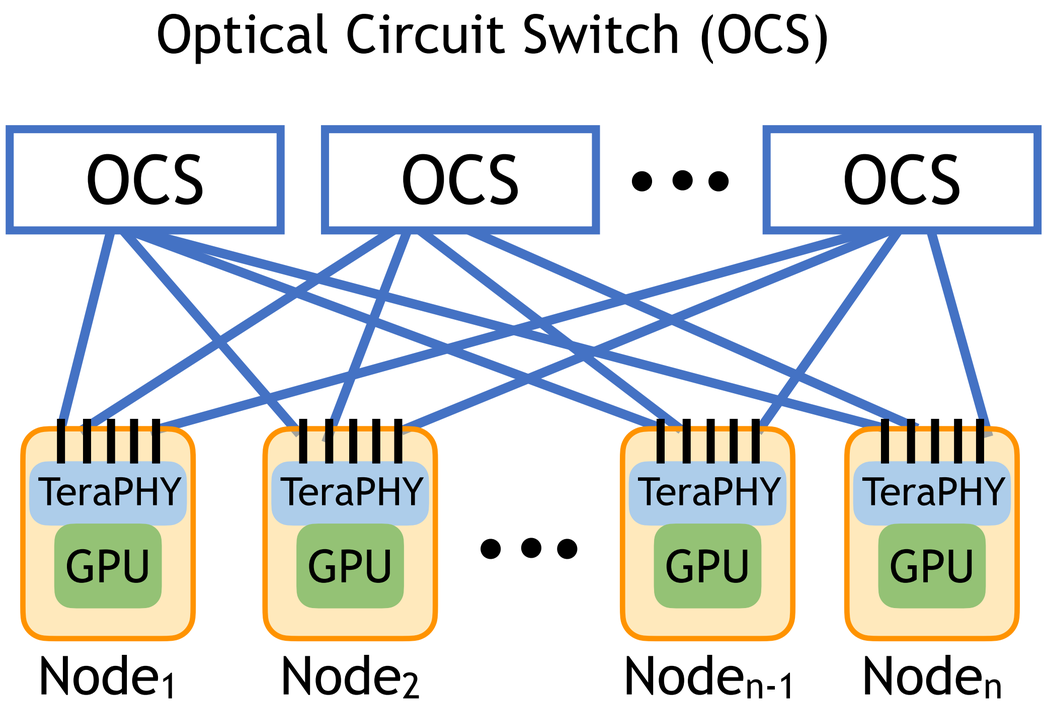


SiP-based cluster

Spectrum of Possible Optical Topologies

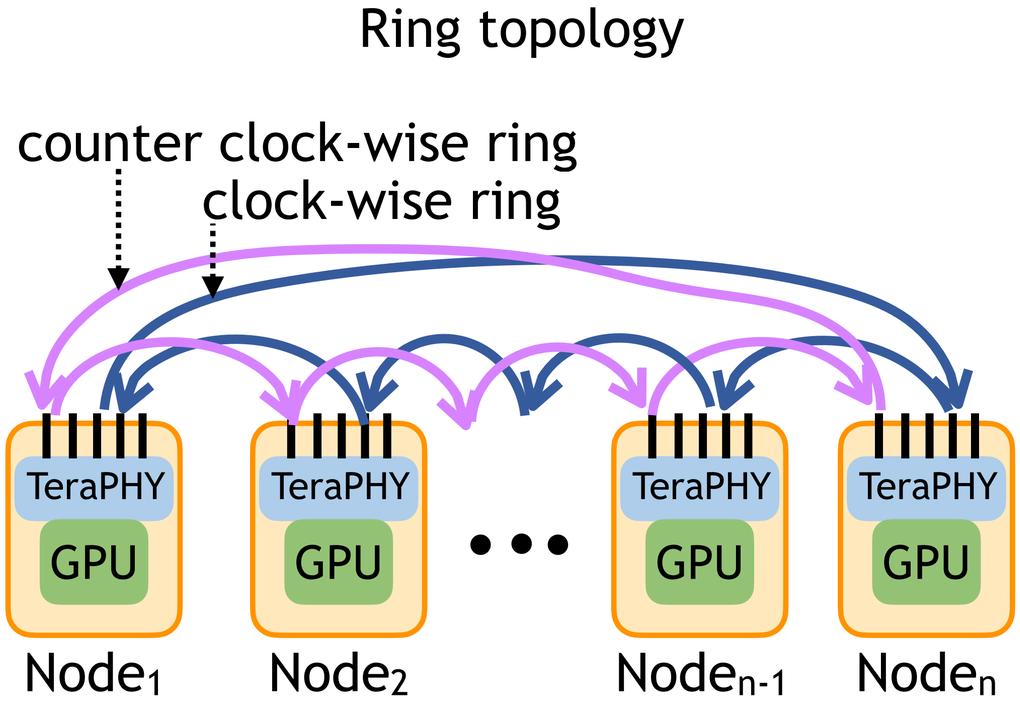


Switch-based Optical Topology



- Commercially available today
- Long reconfiguration delay (~30 ms)
- Suitable only for circuits that can last for several hundreds of millisecond
- Reconfigure the fabric once before the training job starts.

Switch-free Optical Topology

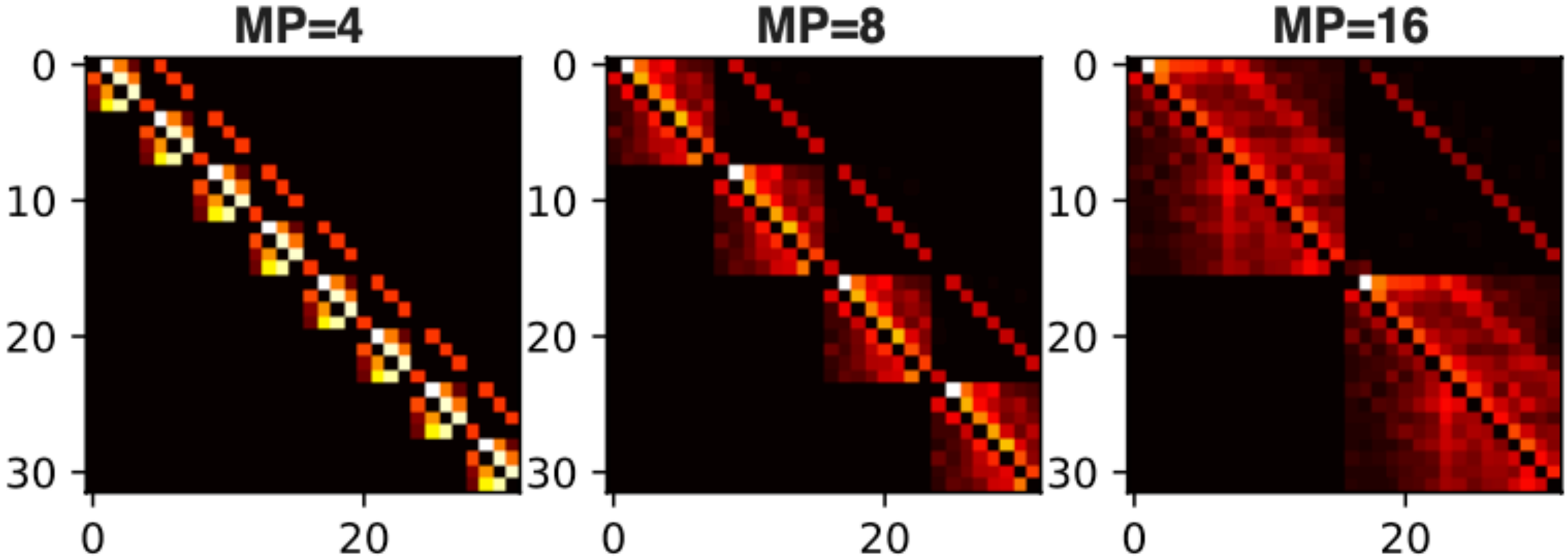


- Extreme design point
- Fast reconfiguration (~20 us)
- Less expensive than switch-based design
- Not palatable for general-purpose data center workloads, but we identify a unique opportunity to build switch-free optical interconnects for dedicated ML clusters.

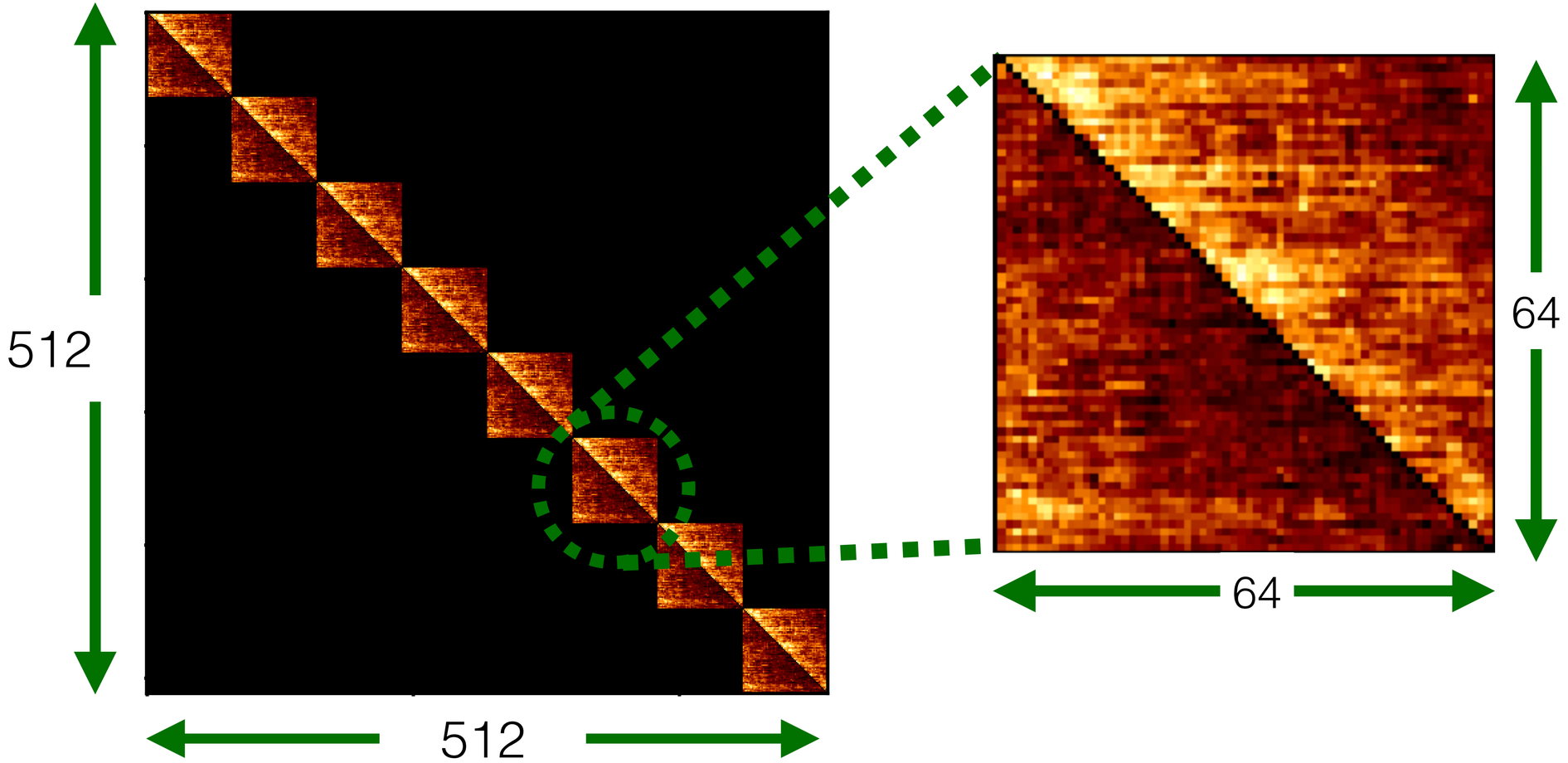
Task Placement Algorithm

- The choice of topology influences the parallelization strategy.
- Switch-based topologies:
  - General-purpose interconnects that can support all-to-all traffic patterns.
  - But high reconfiguration latency enforces a one-shot circuit establishment requirement.
- Switch-free topologies:
  - The parallelization strategy favors short path lengths (i.e., most communication should occur between nearby nodes) to allow wavelengths to be reused around the ring.

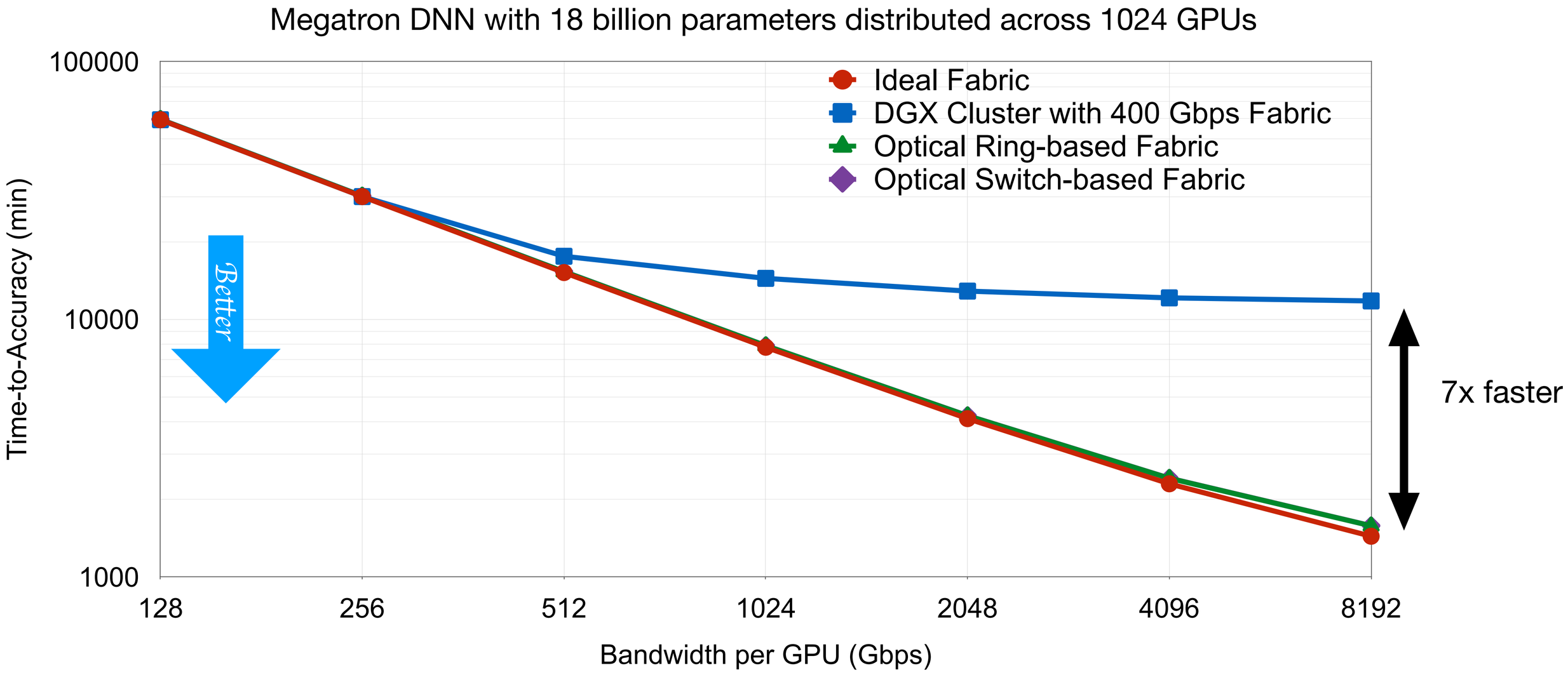
Task Placement Finds the Best MP based on per-GPU Bandwidth



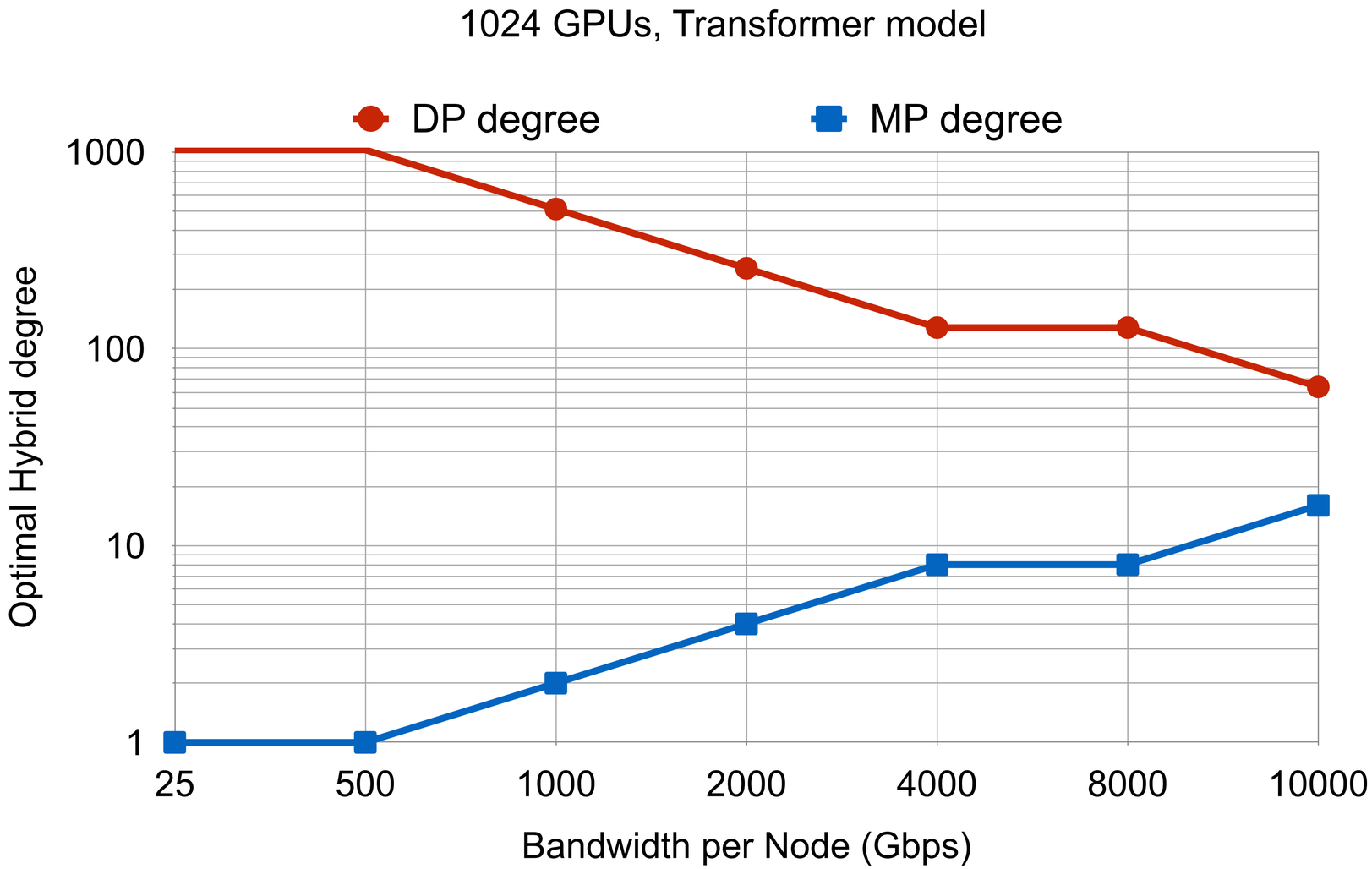
512 GPUs with 8-way DP and 64-way MP



Impact of Network Bandwidth on Training Time

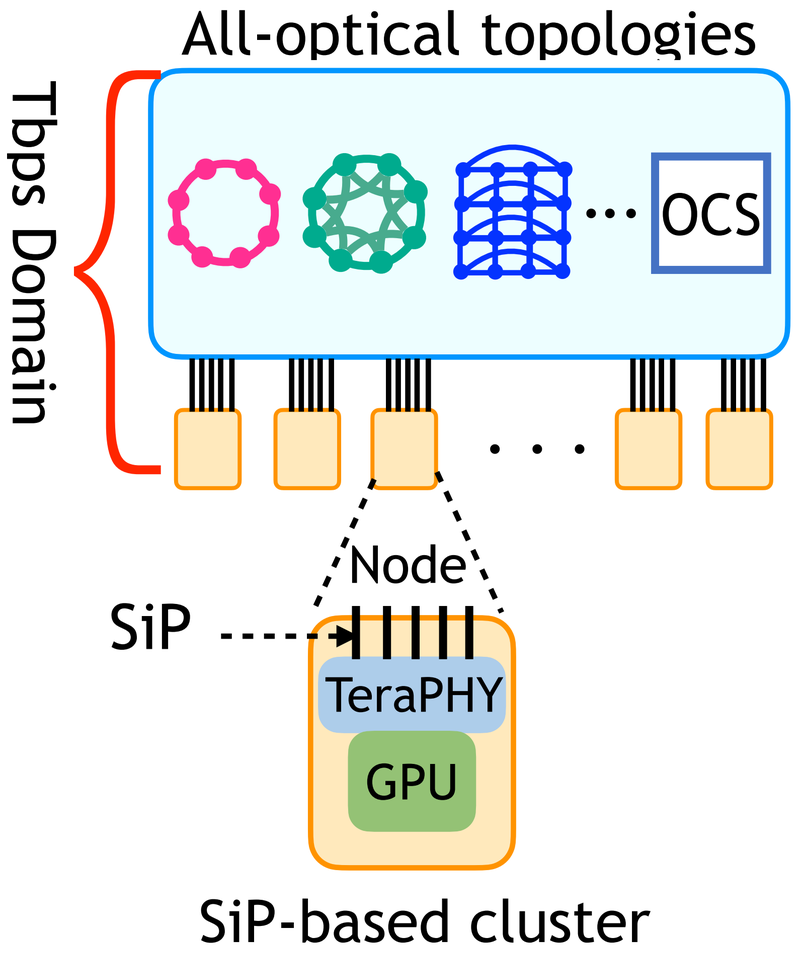


Balance Between Model and Data Parallel



Summary

- SiP-based optical I/O-enabled GPUs for building high bandwidth DNN training clusters.
- The interplay between topology and parallelization strategy provides a powerful tool to design ML networks.
- Despite the seemingly limited connectivity in switch-free topologies, they can support a logically rich communicate pattern by reconfiguring wavelengths in SiP ports.



Paper: M. Khani et al., SiP-ML: High-Bandwidth Optical Network Interconnects for Machine Learning Training [SIGCOMM'21]
Code: https://github.com/MLNetwork/rostam
Email: ghobadi@mit.edu