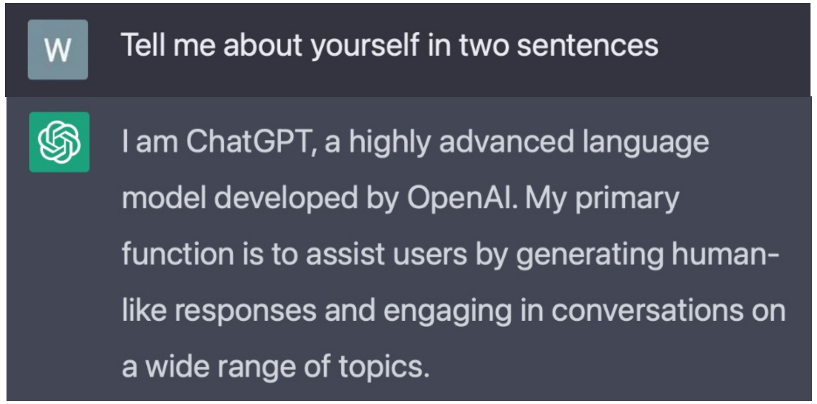
TopoOpt: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs

Weiyang (Frank) Wang, MIT CSAIL

Moein Khazraee, Zhizhen Zhong, Manya Ghobadi, Zhihao Jia, Dheevatsa Mudigere, Ying Zhang, Anthony Kewitsch

The era of large deep neural networks (DNNs)



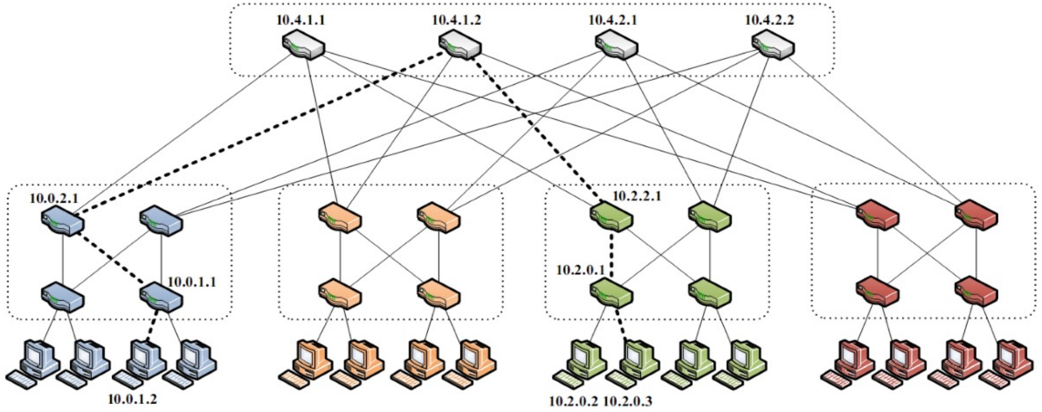
GPT
Large Language Model

Deep Learning Recommendation Model
Recommendation Model

DALL.E
Image Generation Model

- The growth of large DNN models creates demands efficient distributed DNN training systems

State-of-the-art training clusters



Fat-Tree network topology [1]

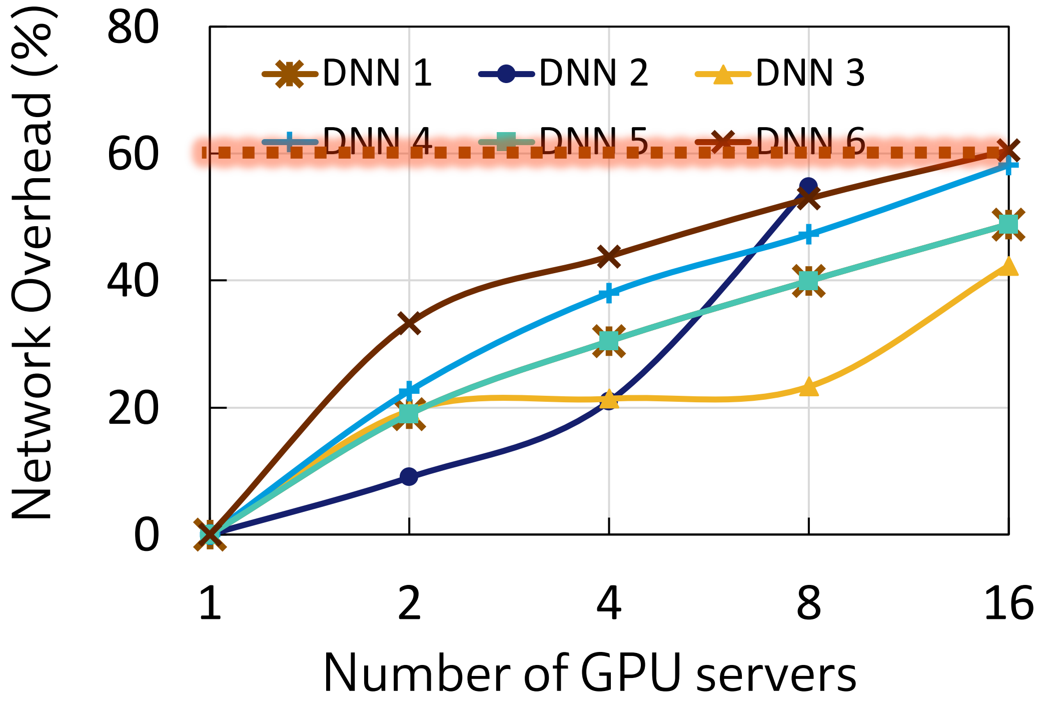
- The Fat-Tree network topology forms the basis of today's training cluster
- Traffic oblivious fabric the provides uniform, full-bisection bandwidth between server pairs
- Ideal when the workload is unpredictable and consists mostly of short transfers

- Full-bisection networks are not the best network topology for DNN training!

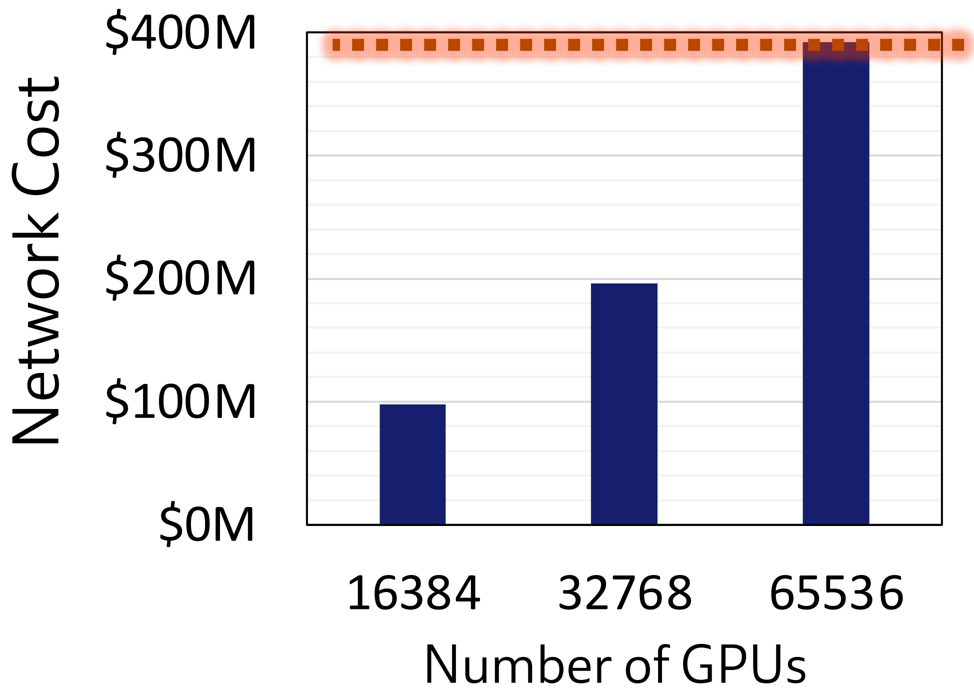
[1] *Al-Fares et al., A Scalable, Commodity Data Center Network Architecture, SIGCOMM '08*

Network is becoming a bottleneck and getting too expensive

- **Network Overhead**: the amount of time spent on communication only



- **Network Cost**: total cost of network switches and transceivers at 400Gbps



Previous work on distributed DNN training optimization does not consider physical topology

Compression and encoding
THC [NSDI '24]

Schedulers
Themis [NSDI '20]

Asynchronous transmit
DC-ASGD [PMLR '17]

Computation
+
Communication

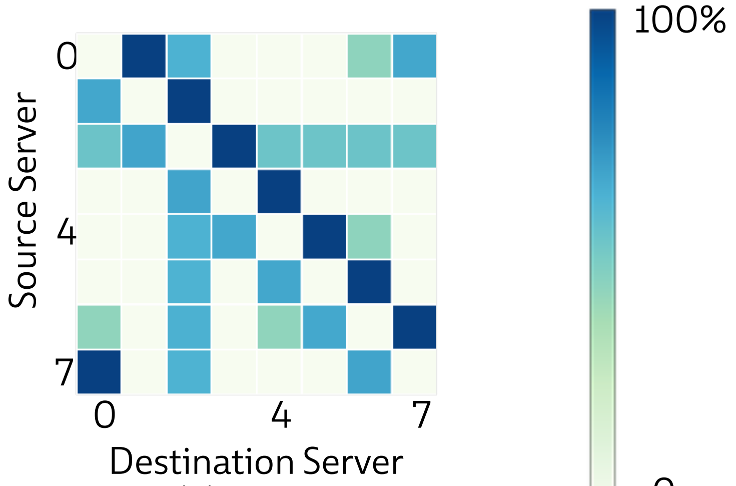
Parallelization strategy
Alpa [OSDI '22]

Collective communication
TE-CCL [SIGCOMM '25]

Hyper parameters
ASHA [MLSys '20]

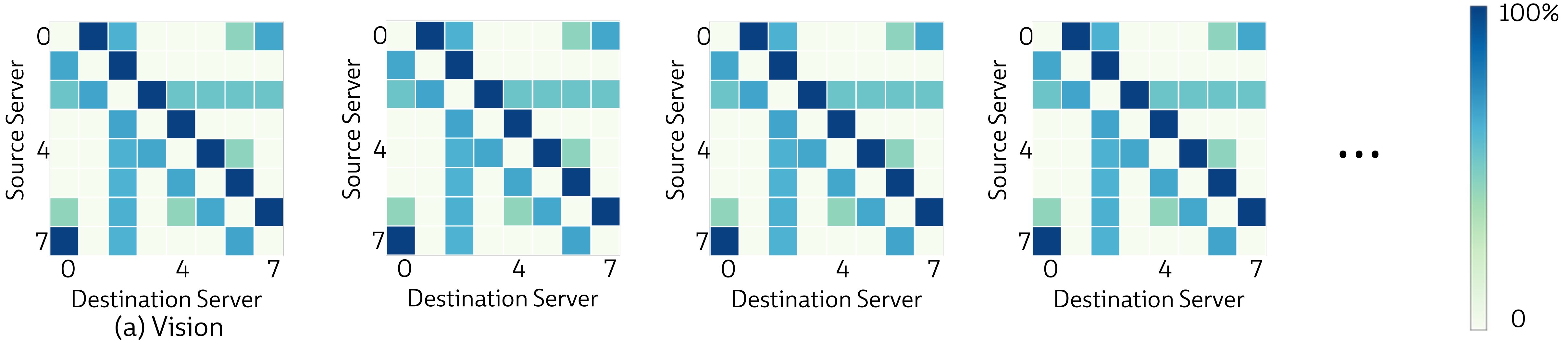
Network topology
?

DNNs training traffic has different properties



(a) Vision

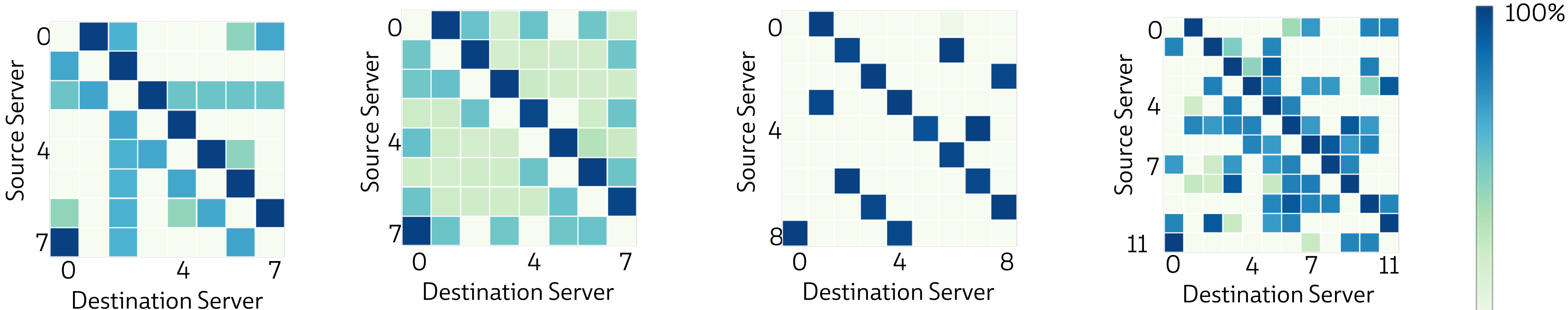
DNNs training traffic has different properties



Iteration 1 Iteration 2 Iteration 3 Iteration 4 …

- Key observations:
    1. Traffic patterns do not change across training iterations

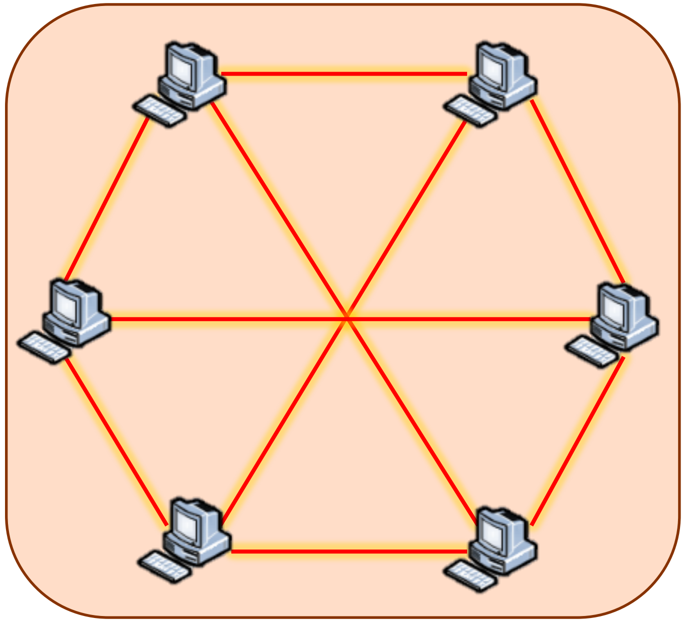
DNNs training traffic has different properties



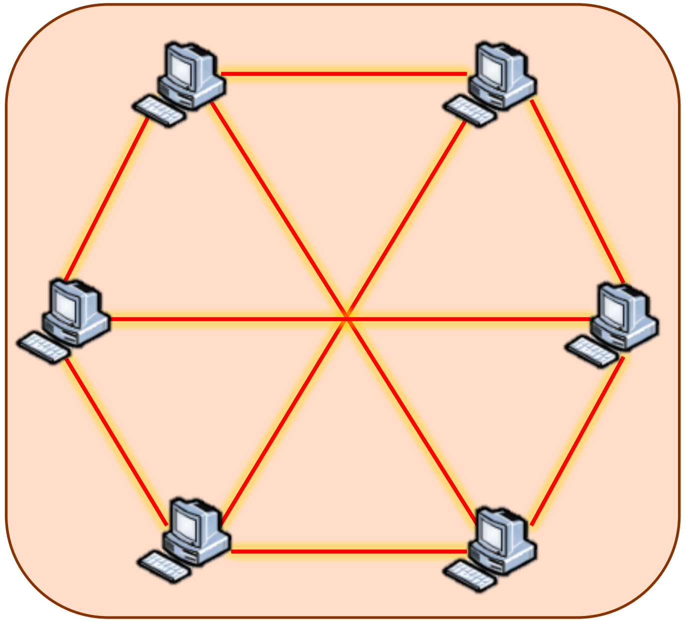
(a) Vision (b) Image processing (c) Object Tracking (d) Speech Recognition

- Key observations:
    1. Traffic patterns do not change across training iterations
    2. Traffic patterns are model-dependent

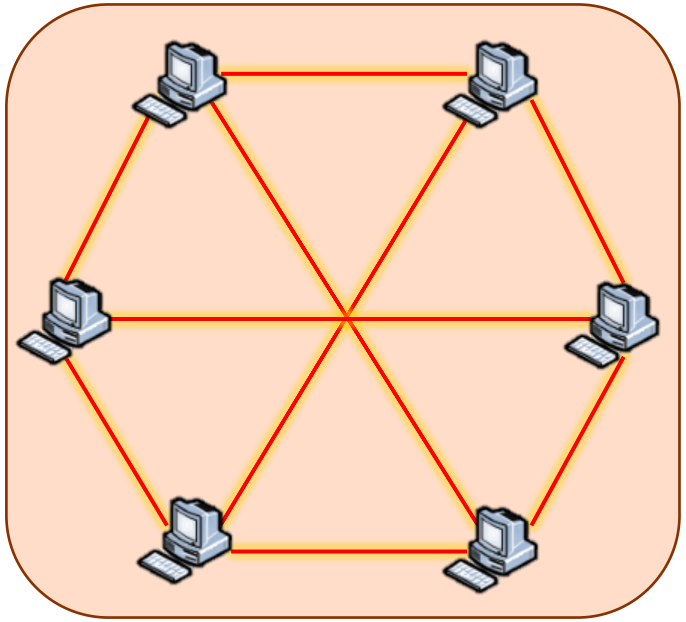
Reconfiguring physical network topology



Topology A

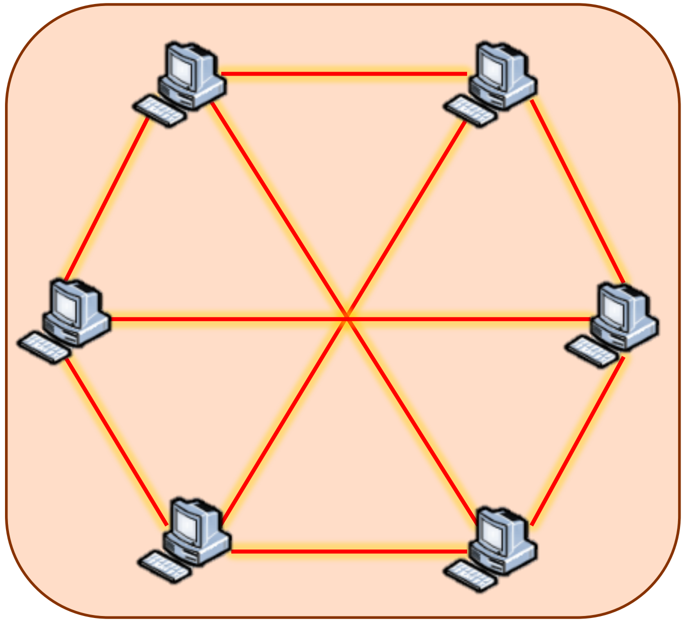


Topology A

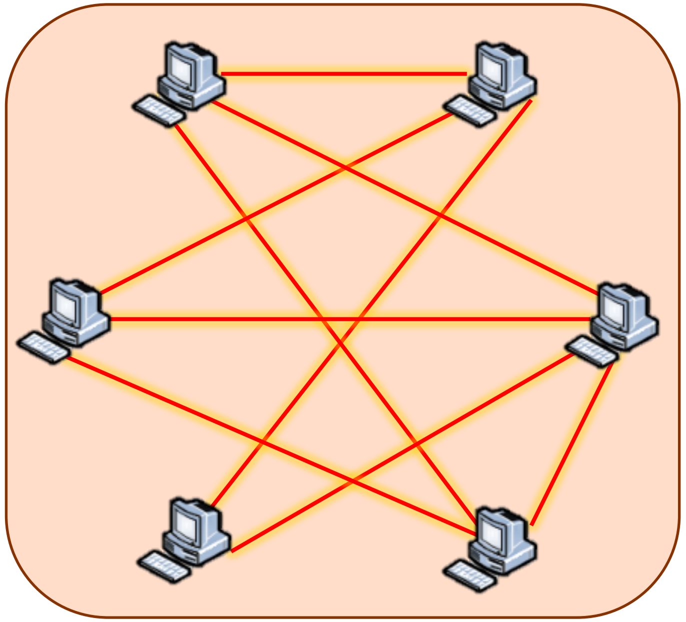


Topology A

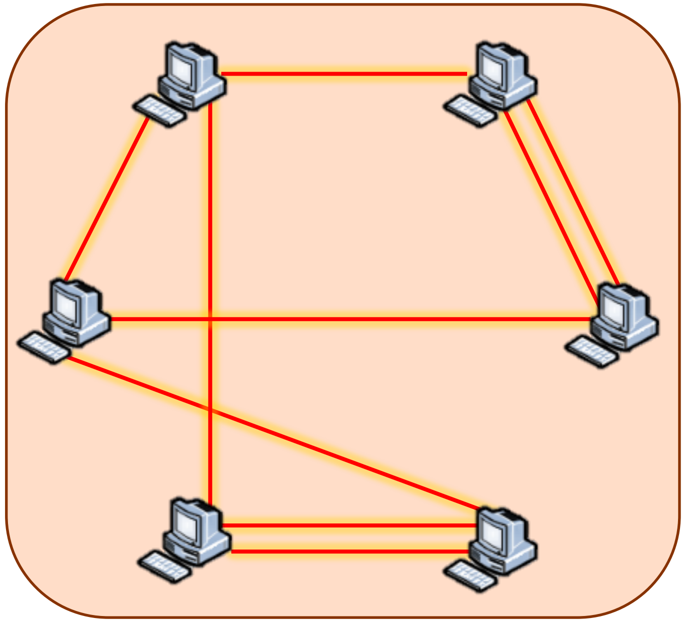
Reconfiguring physical network topology



Topology A



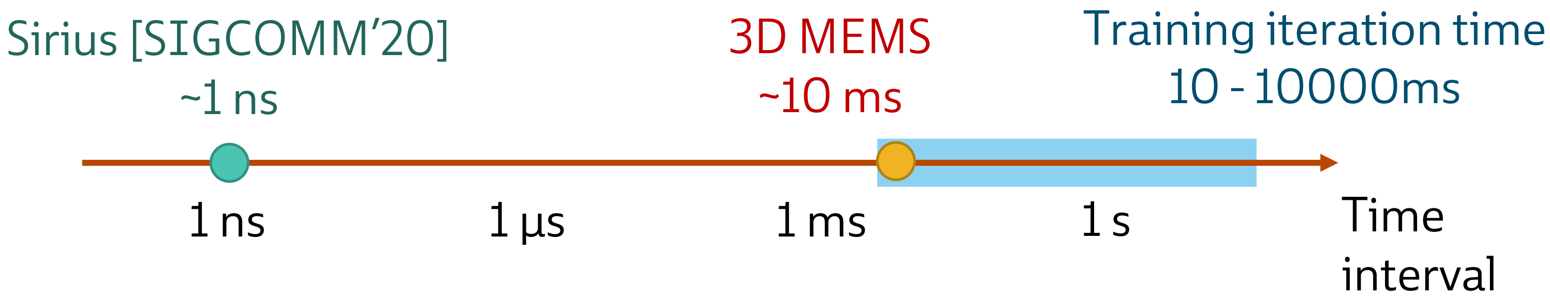
Topology B



Topology C

Reconfiguring physical network topology – how often?

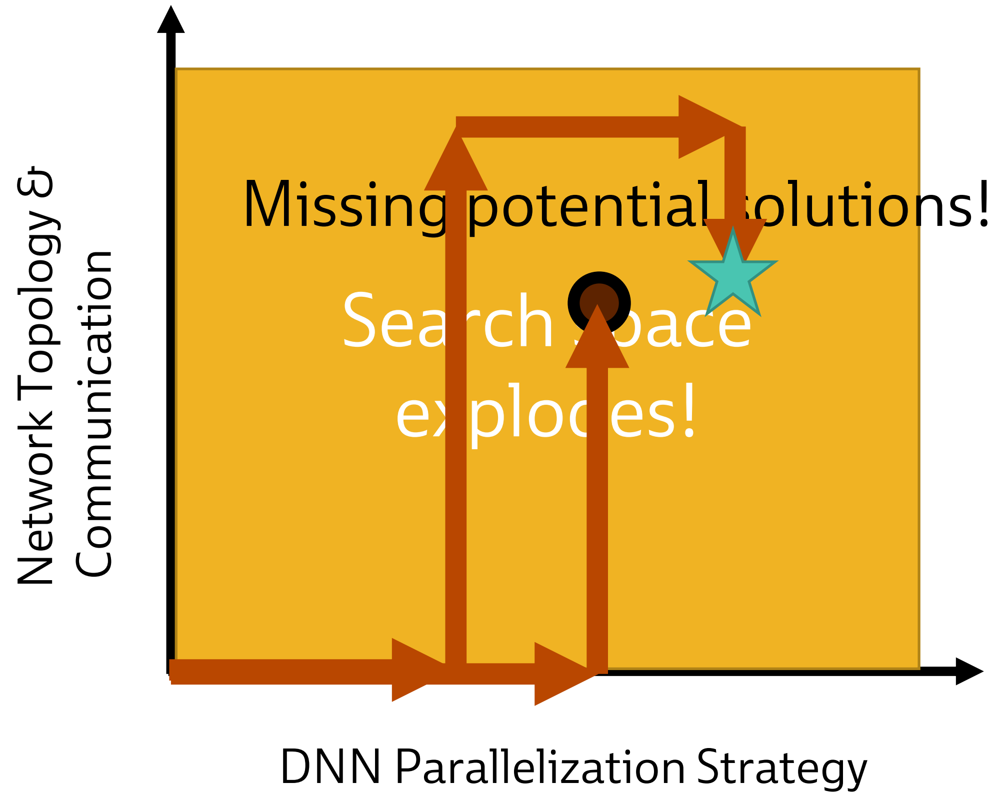
- Ideally, we could change the topology according to the instantaneous demand
- However, this is challenging with today's technology
  - Existing commercially available solutions that scales to thousands of ports are not fast enough for many DNN models



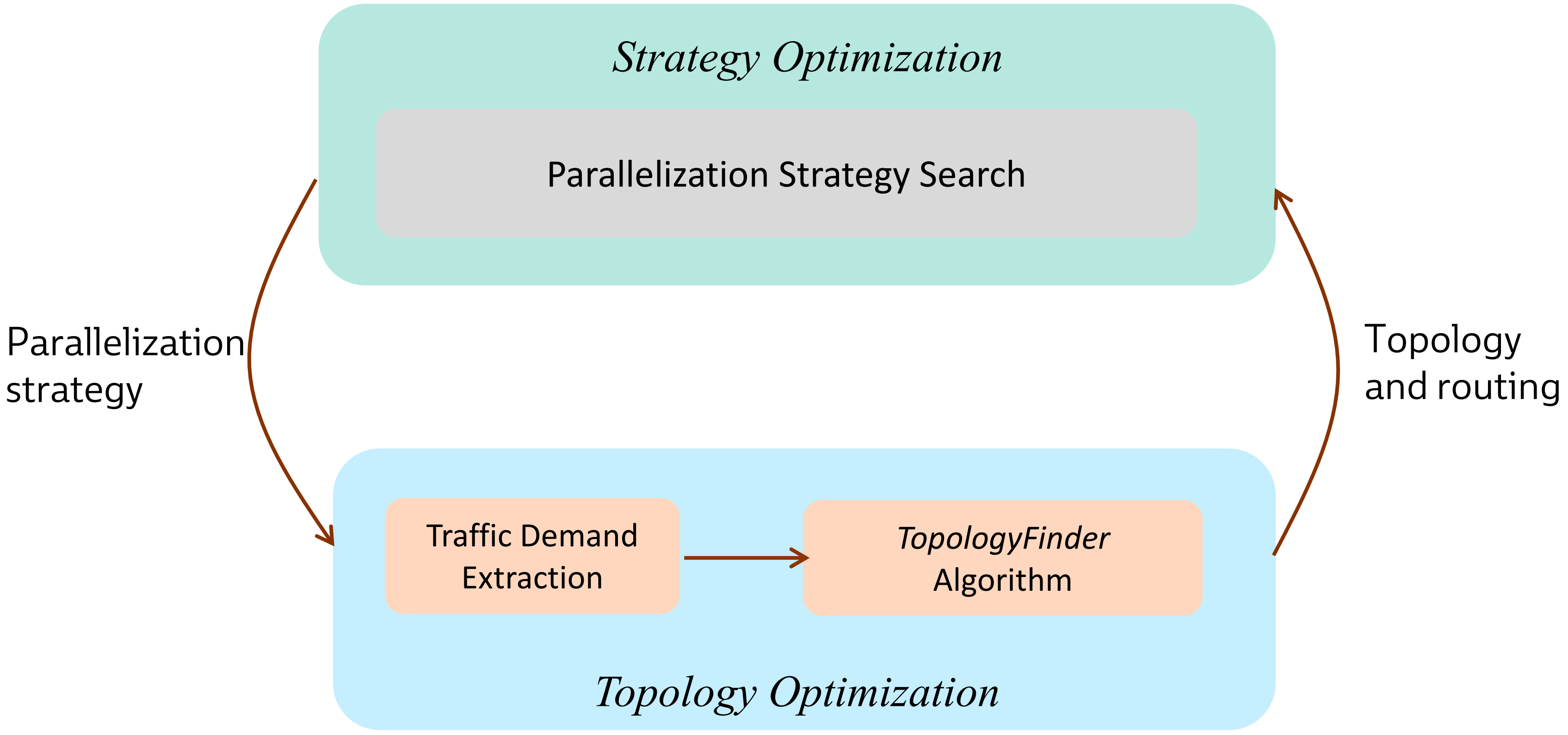
- In this presentation, we focus on a **one-shot reconfiguration** policy
  - Find one topology for the **entire duration** of each training job

Co-optimization challenge: Huge search space for optimal DNN training

- The configuration space is huge!



Alternating optimization framework to co-optimize DNN parallelization strategy and network topology



Alternating optimization framework to co-optimize DNN parallelization strategy and network topology

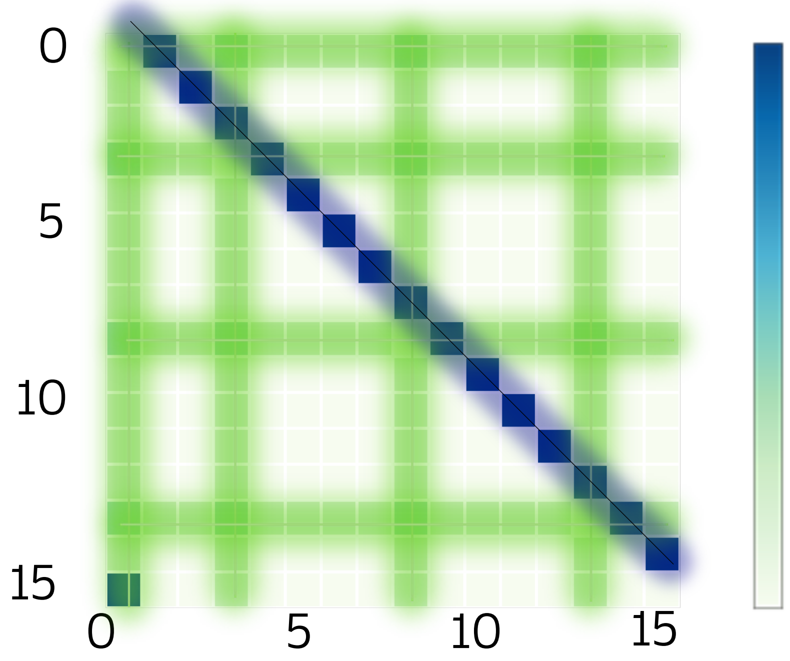
Traffic Demand Extraction

TopologyFinder Algorithm

Topology Optimization

What algorithm should we use to find the topology in this framework?

Characteristics of DNN training traffic for DLRM



Data Parallel AllReduce Transfers

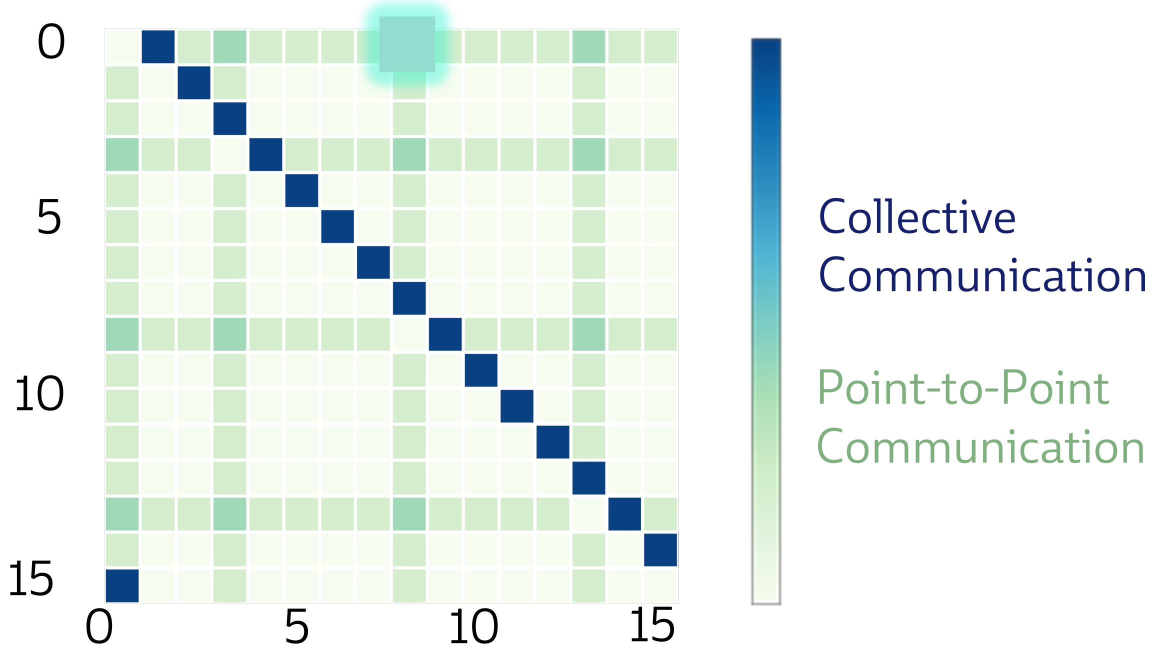
- Collective Communication (CC)
- Achieve some data distribution goals, in this case taking an average of the gradients located on all GPUs
- **Ring-AllReduce** generates a ring traffic pattern

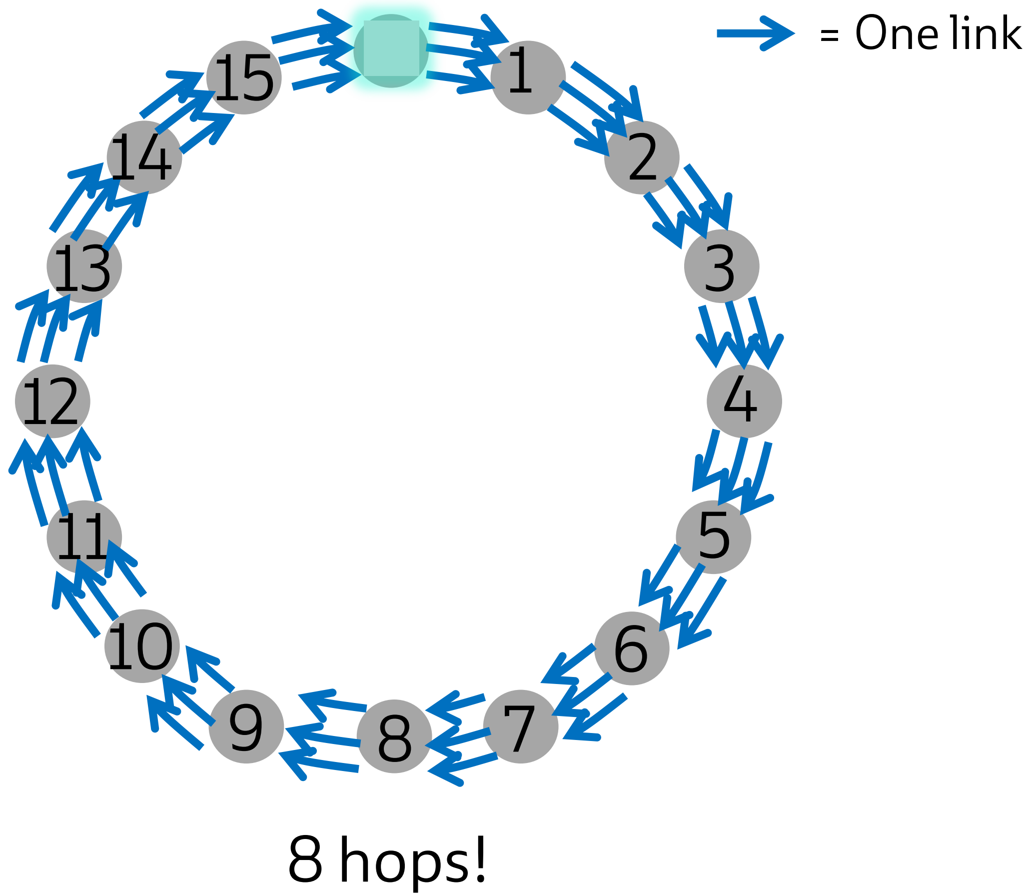
Model Parallel Transfers

- Point-to-Point Communication (P2P)
- An operator placed on one GPU communicating with another operator located on another GPU

Challenge: finding a good network topology for both Collective and Point-to-Point transfers

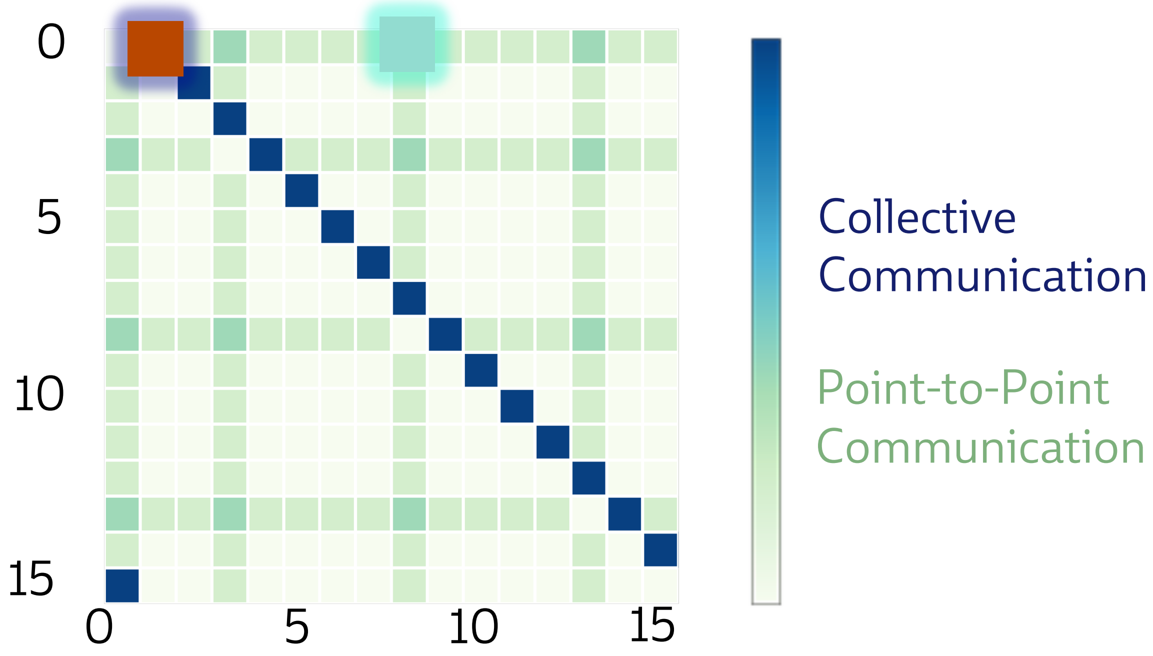
- Degree (d) = 3, unidirectional



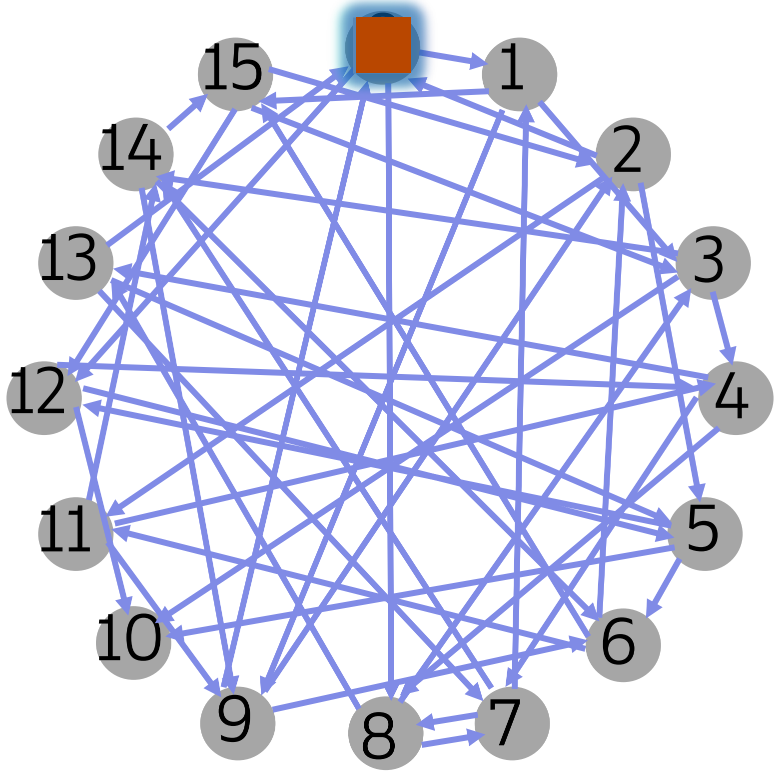


Challenge: finding a good network topology for both Collective and Point-to-Point transfers

- Degree (d) = 3, unidirectional

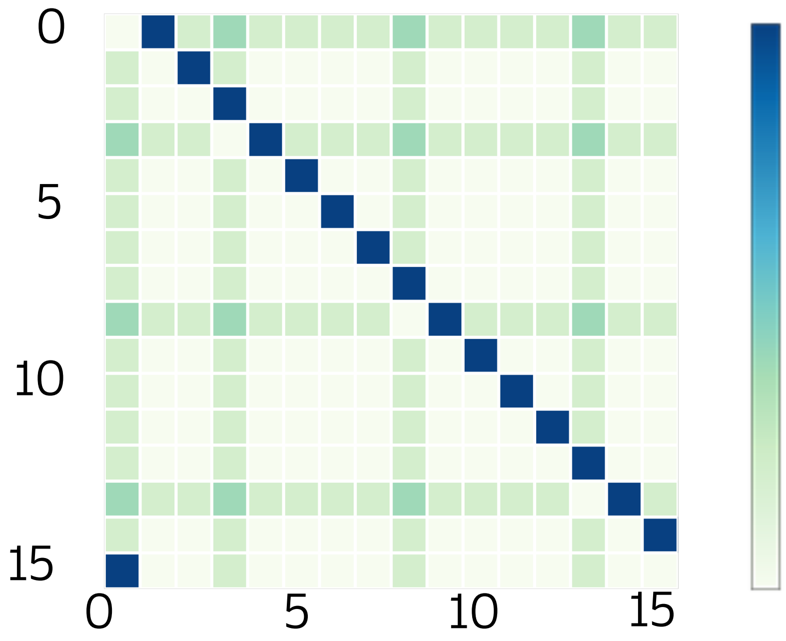


Low Bandwidth!



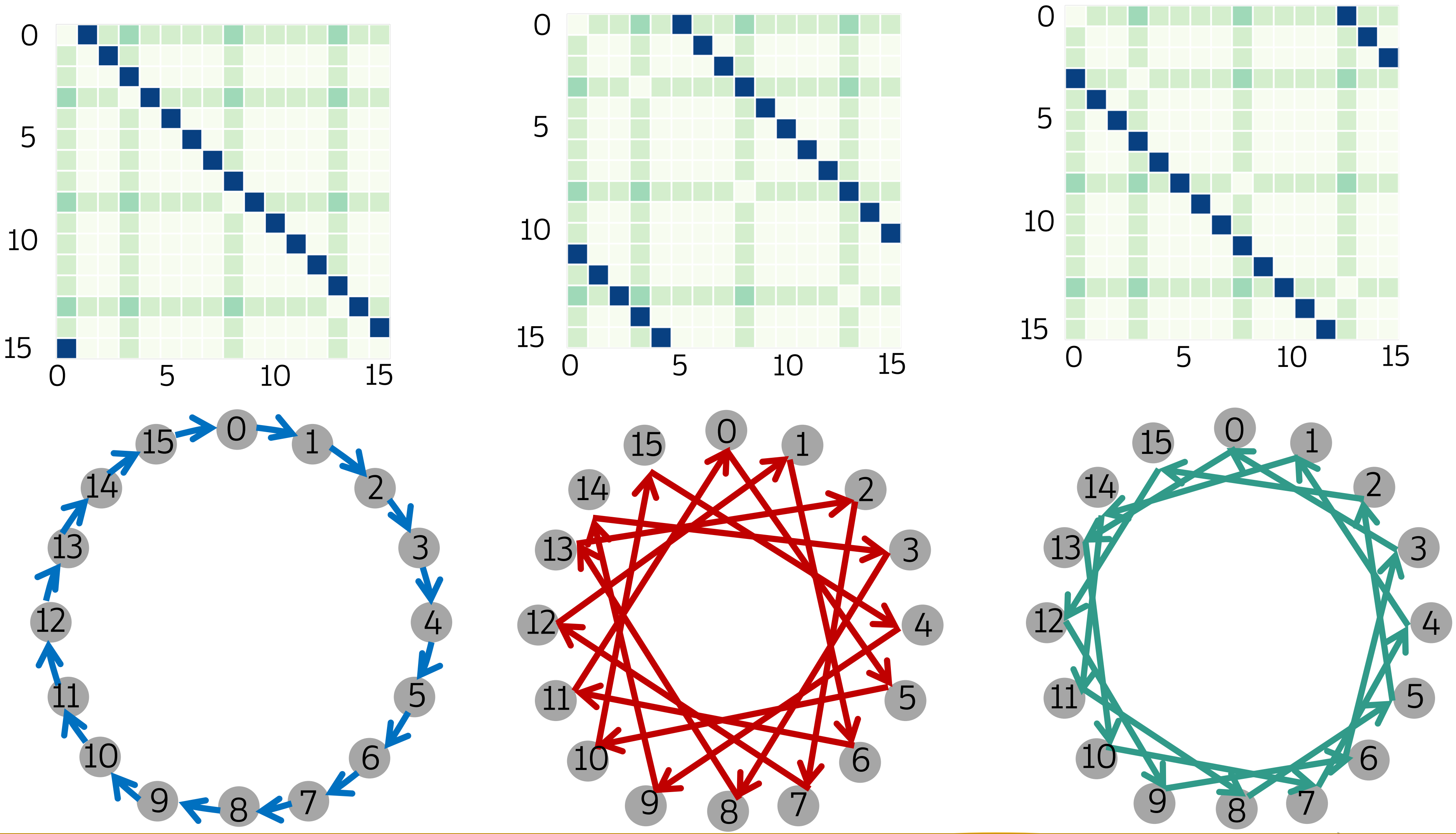
Meeting the requirements of both Point-to-Point and Collective transfers

- Degree (d) = 3, unidirectional



Transfer Type	Characteristics	Network Requirement
Collective Communication	Large, Sparse	Ample Bandwidth
Point-to-Point Communication	Small, Dense	Low hop-count

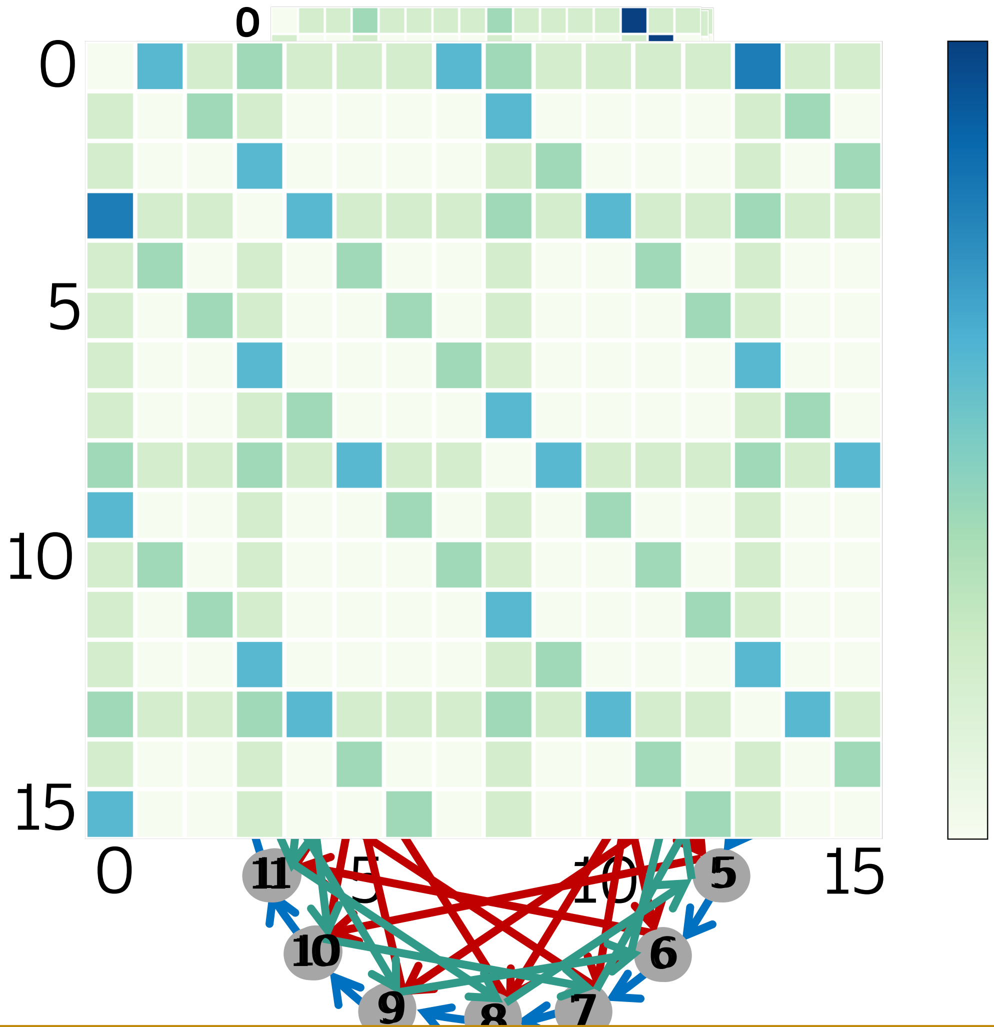
Key idea: **mutate the traffic matrix**

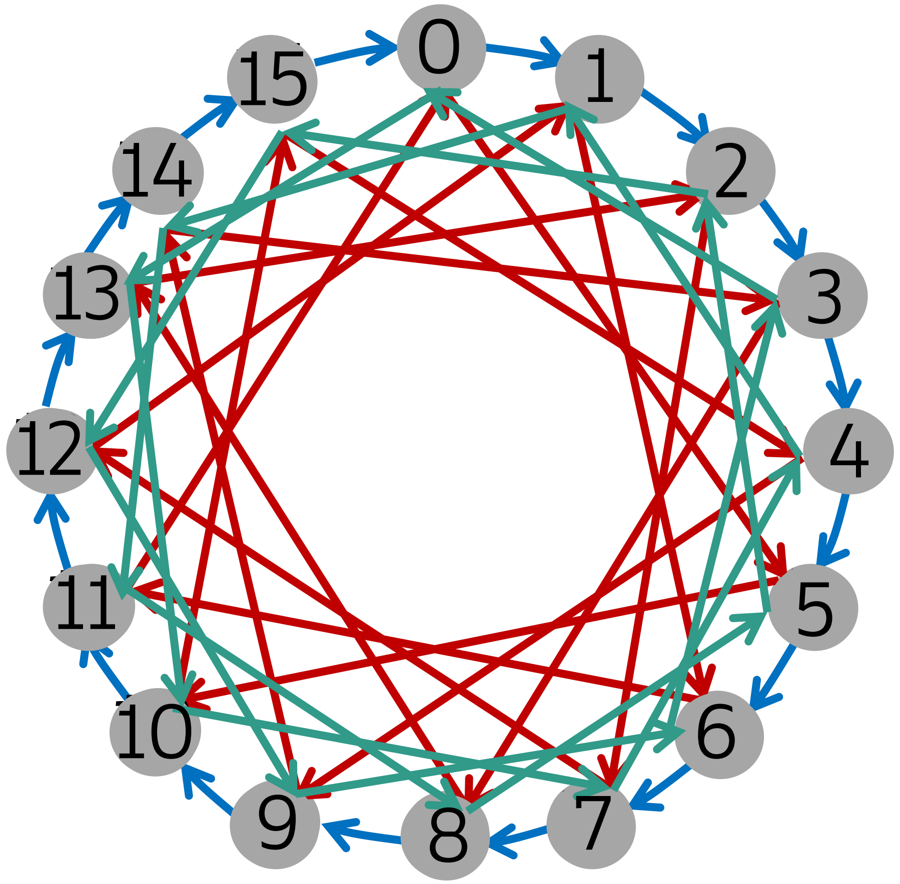


Collective Communications are **mutable.**

Point-to-Point transfers are not mutable.

Splitting AllReduce traffic

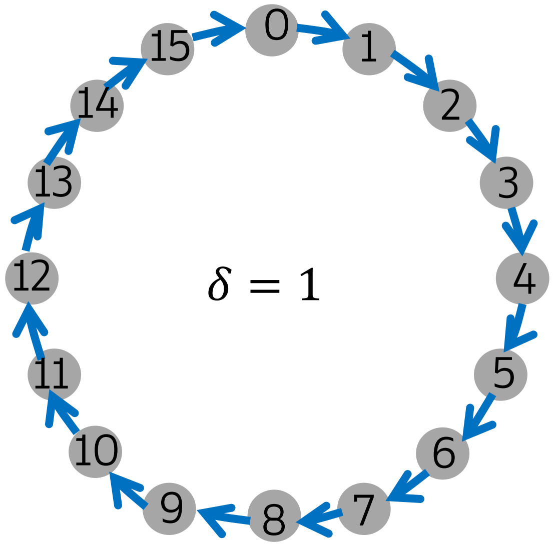


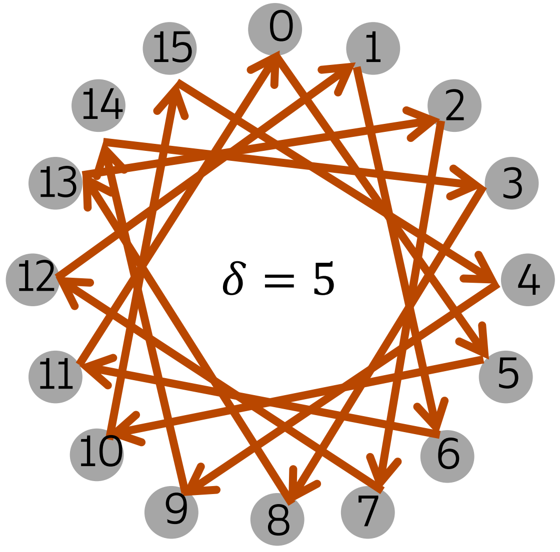


Leverage the mutability of Collective Communication to achieve high bandwidth for CC & low hop-count for Point-to-Point transfers!

Key technique: Regular permutations

- n total accelerator, each with degree d



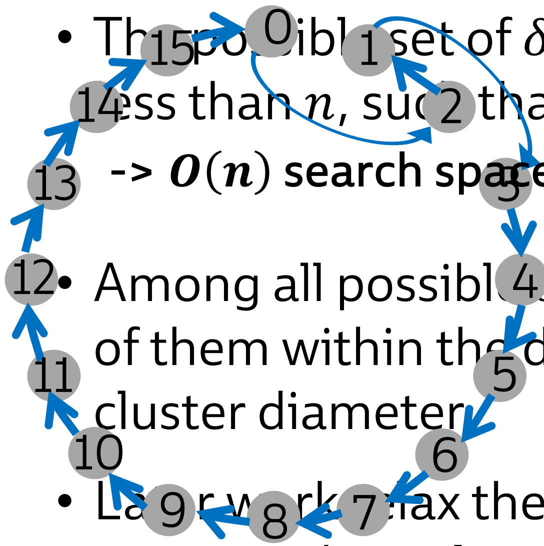


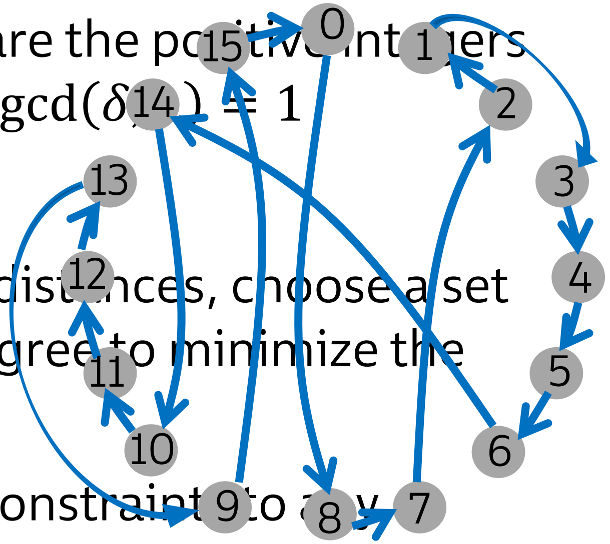
- The possible set of δ are the positive integers less than n, such that $\gcd(\delta, n) = 1$

  -> $O(n)$ search space!

- Among all possible δ distances, choose a set of them within the degree to minimize the cluster diameter

- Later work relax the constraint to any connected circulant graph

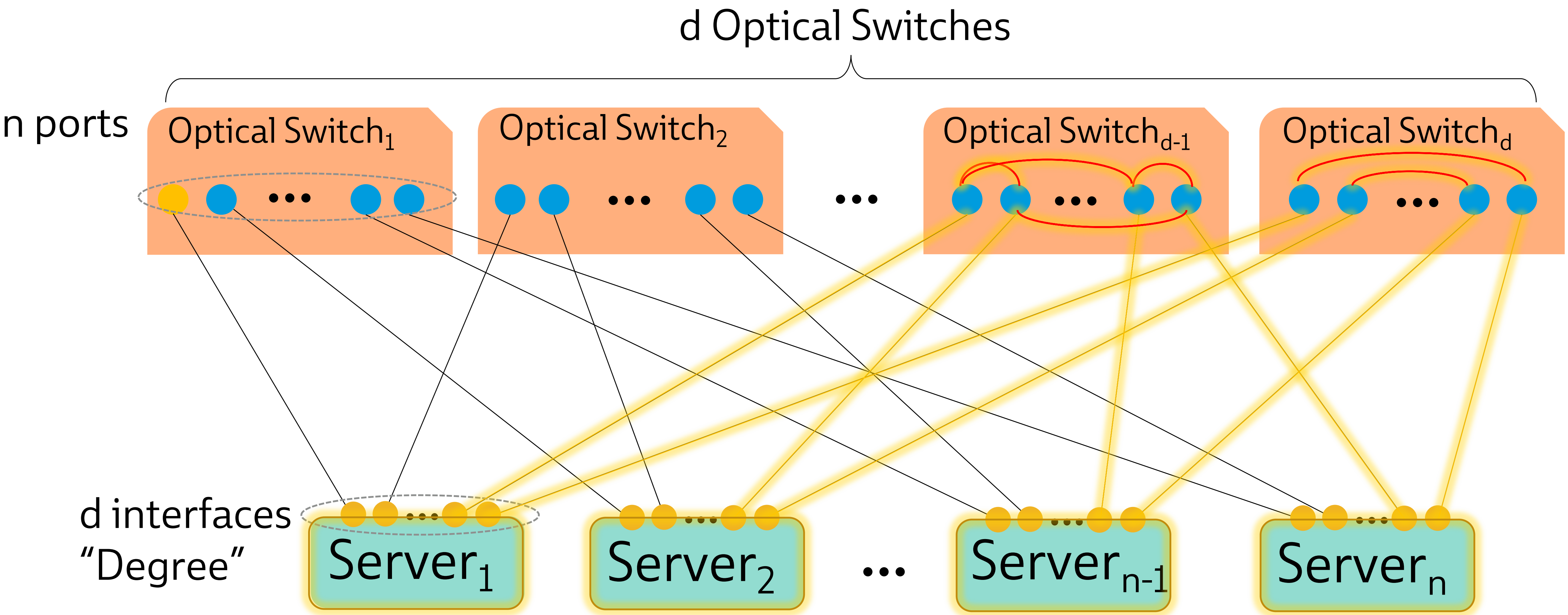




Irregular permutations

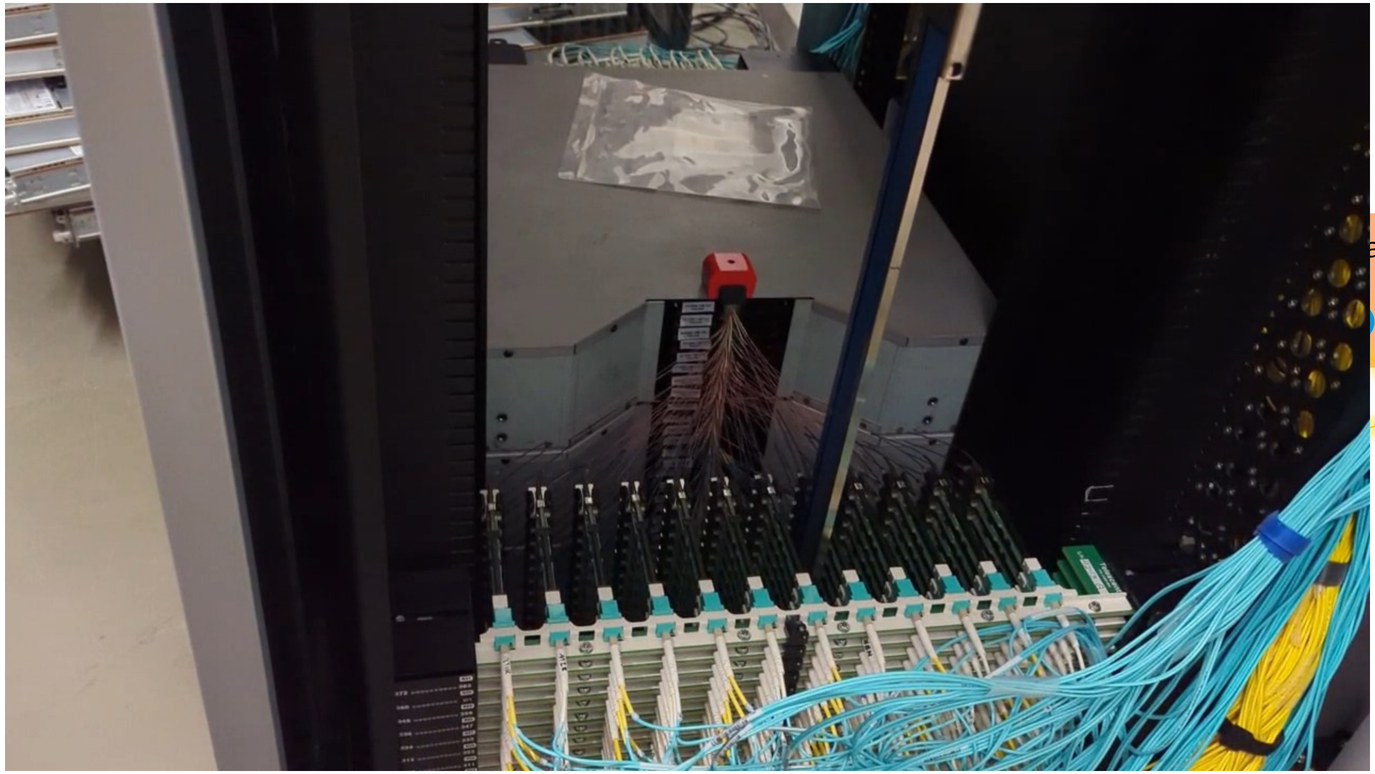
TopoOpt bounds the cluster diameter to $O(d \cdot \sqrt[d]{n})$

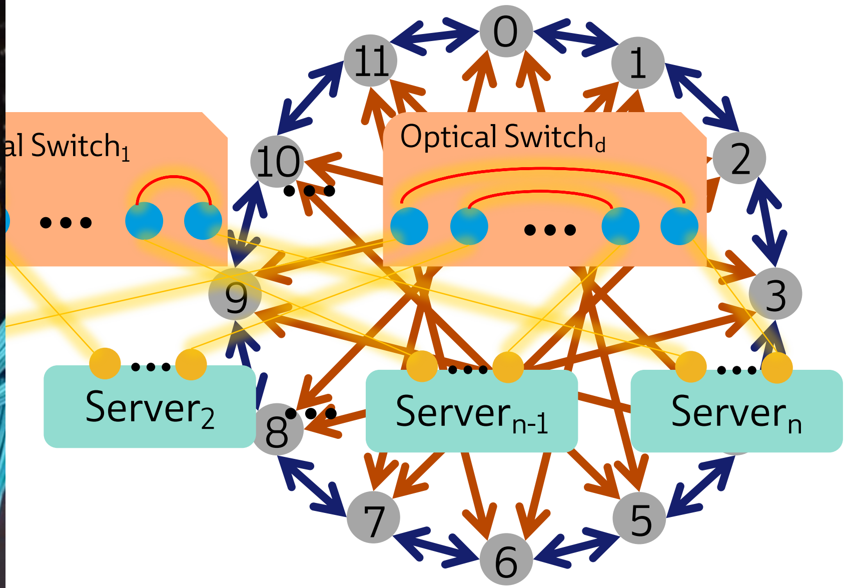
Physical interconnect of TopoOpt



TopoOpt uses optical switches

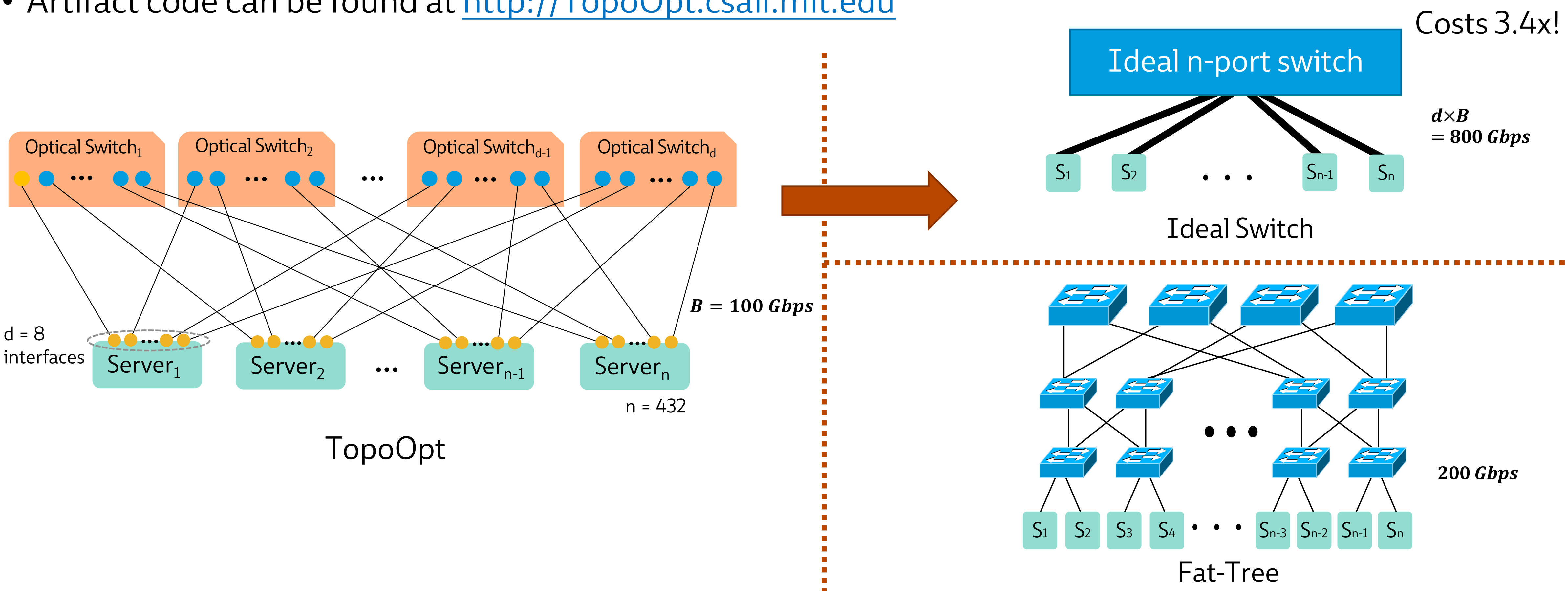
- Fully functional 12-node, degree 4 testbed integrated with NCCL





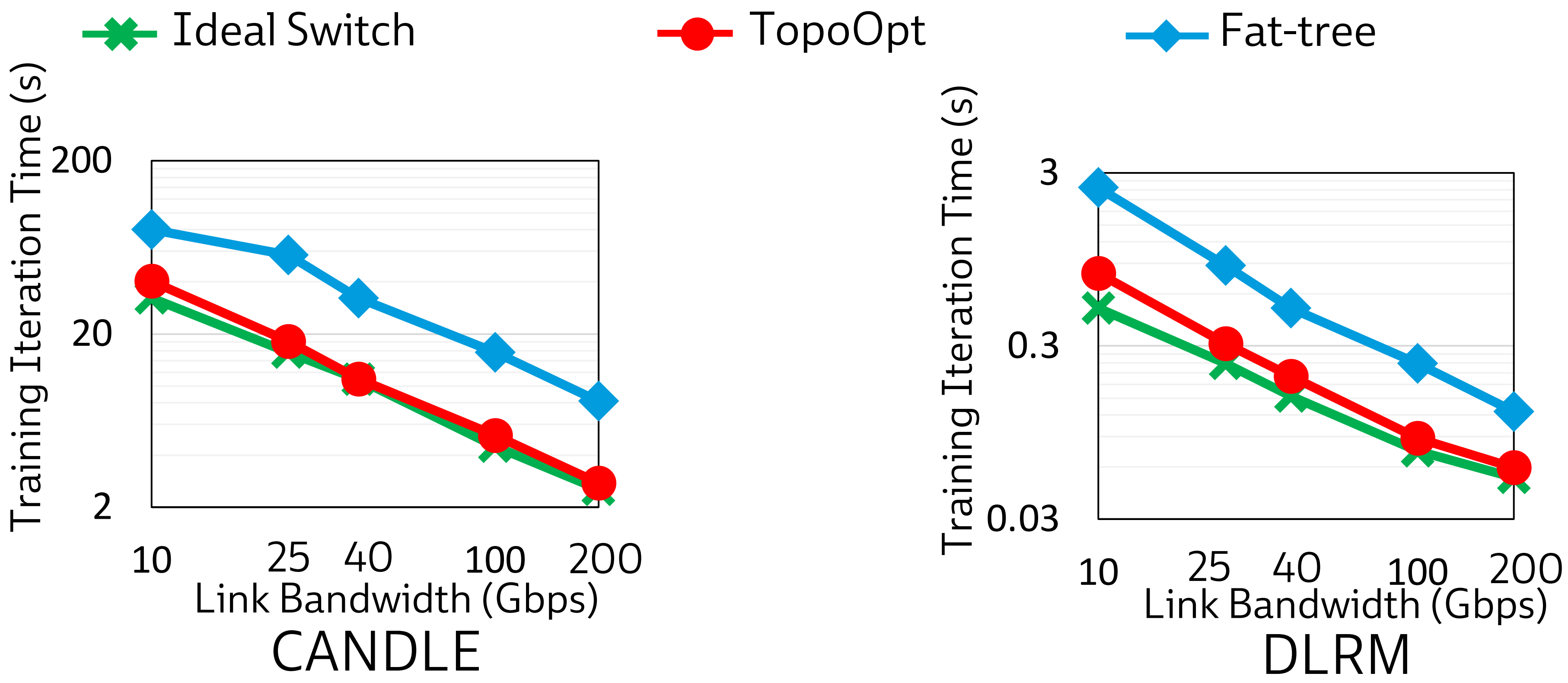
Evaluation

- We evaluate TopoOpt with large scale simulation and a small-scale prototype
- Artifact code can be found at http://TopoOpt.csail.mit.edu



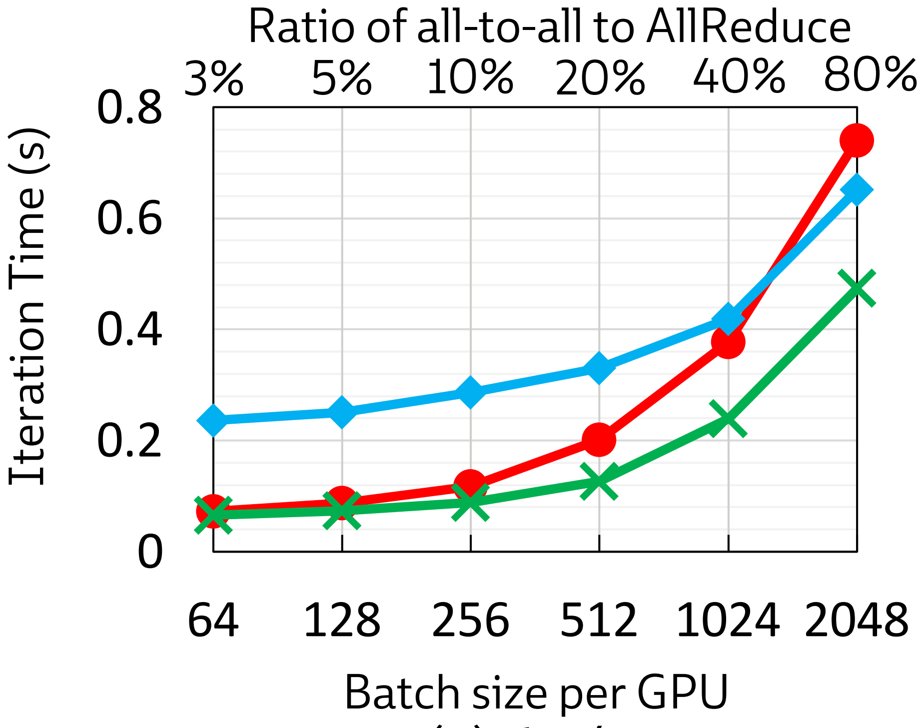
Simulation – iteration time

- Training DNN on a dedicated cluster of 128 nodes, d = 4, with different available bandwidth

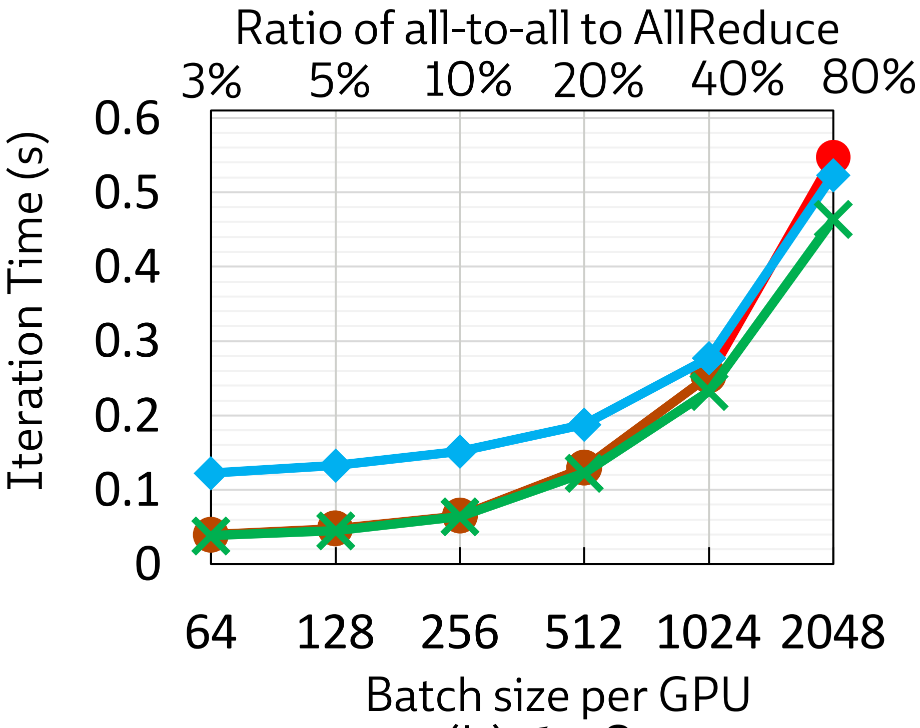


Simulation – Impact of All-to-All traffic

- Training DLRM model with different batch size



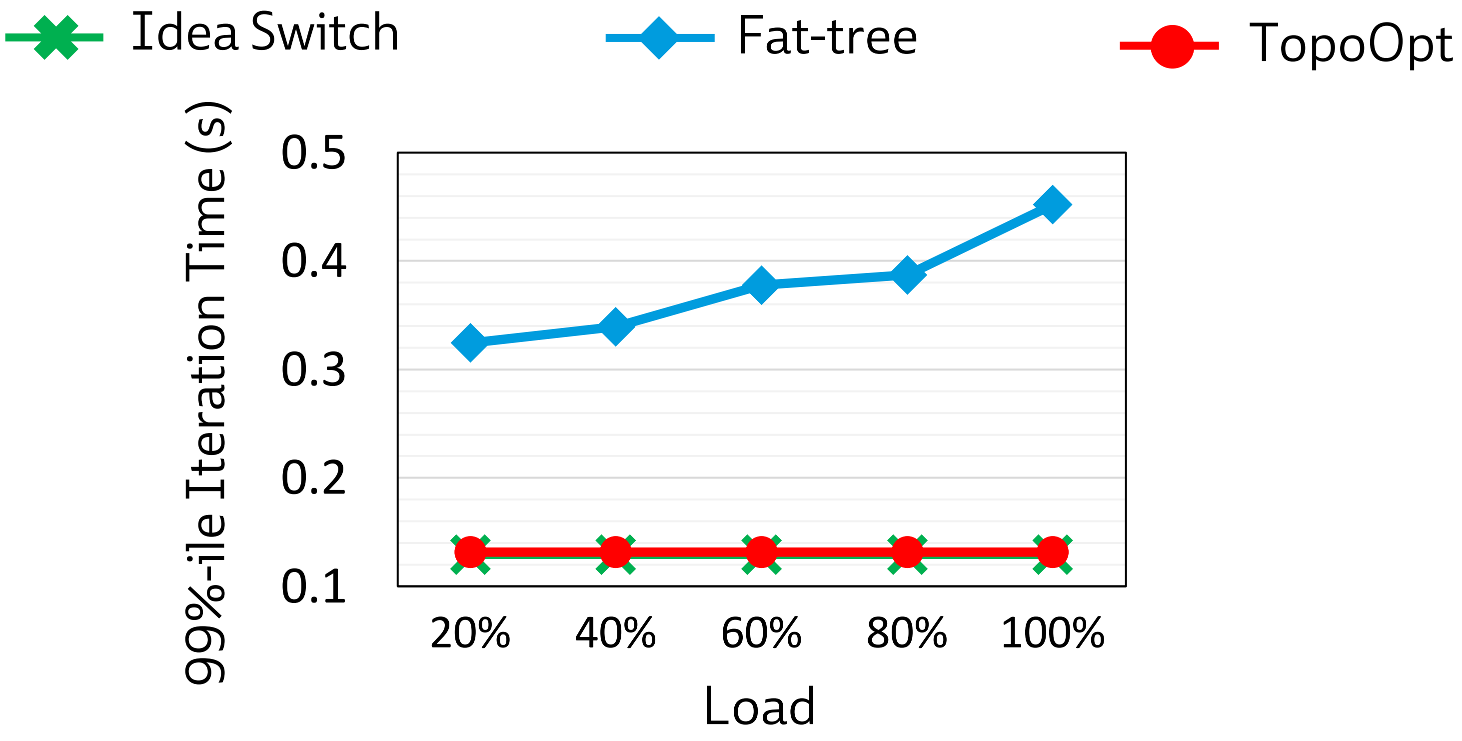
(a) d = 4



(b) d = 8

Simulation – tail completion time

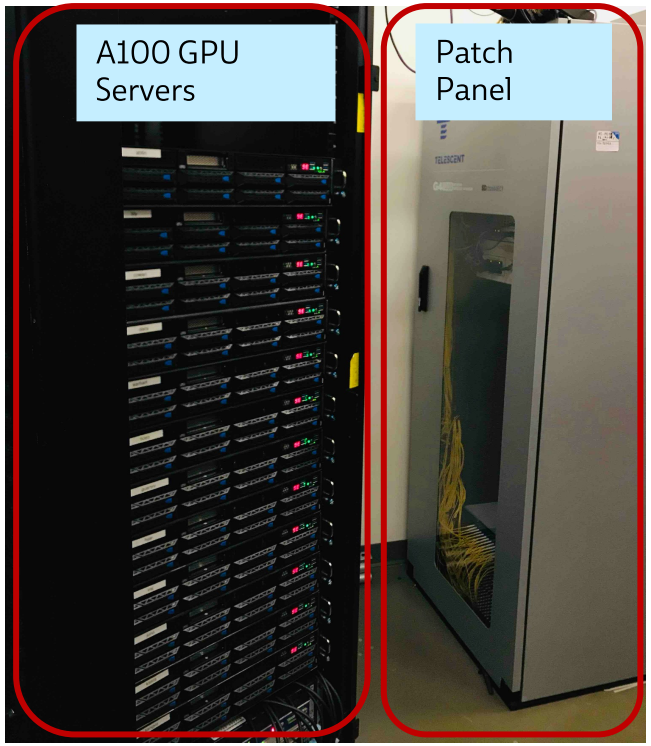
- Running several jobs together on a 432 node, d = 8, 100Gbps TopoOpt system, compared to several other options



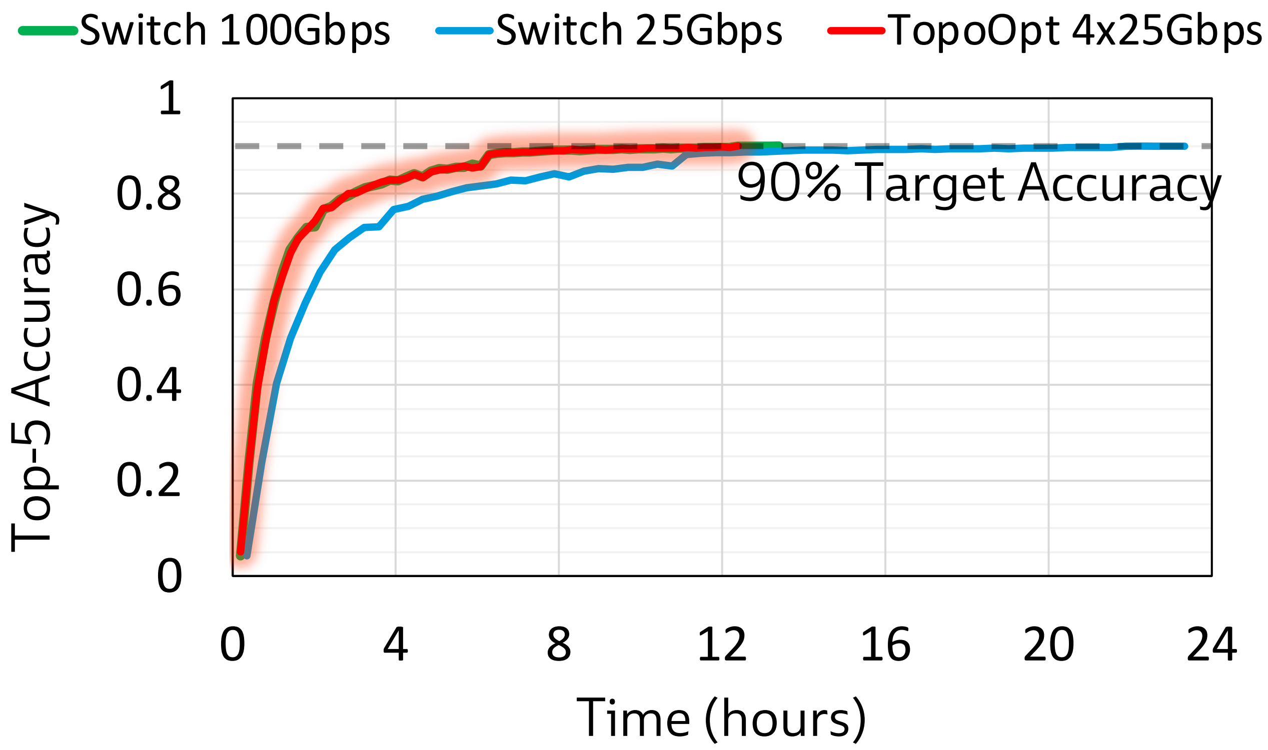
TopoOpt **isolates the jobs perfectly by design**, and achieves up to 3.4x faster 99%-tile latency compared to cost-equivalent Fat-trees

Testbed result

- We implemented a prototype for TopoOpt on a 12-node testbed, with Nvidia A100 GPUs and 4 x 25Gbps HPE NICs connected to an optical patch panel



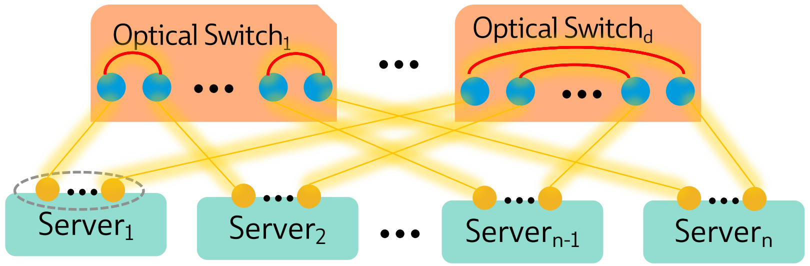
Testbed Photo



Time to Accuracy, VGG19

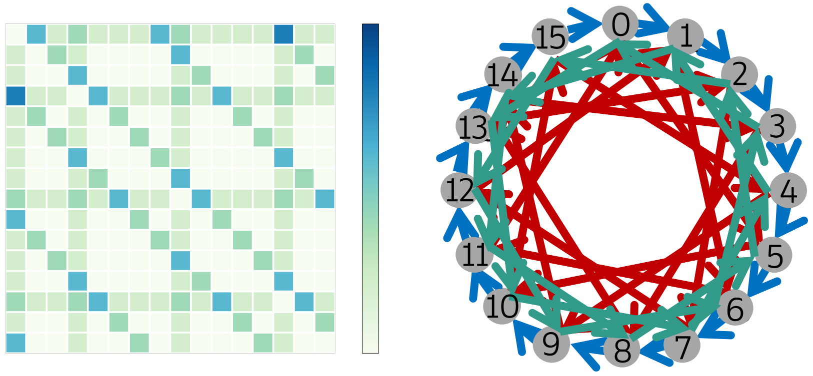
TopoOpt matches the performance of an ideal full-bisection bandwidth fabric

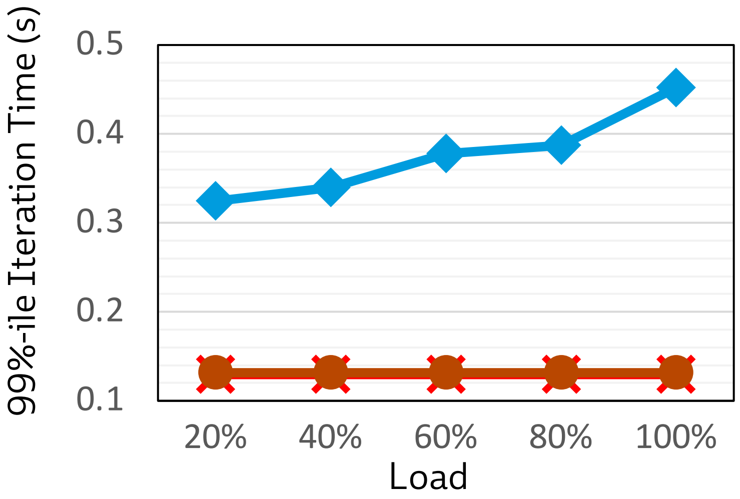
Summary



TopoOpt: the first system to co-optimize DNN training with demand-aware network topology

Leverages the mutability of DNN training traffic to search and construct the best topology





Achieves up to 3.4x faster 99%-ile training iteration time compared to cost equivalent Fat-trees

Future work and upcoming talks

- LLM with 3D parallelism and Mixture of Expert (MoE) layers:
  - Disjoint traffic across different parallelisms
  - Non-uniform, many-to-many dense communication

- Utilizing fast-reconfigurable optical switches to build efficient all-to-all communication primitive

- Network infrastructure for other popular ML workload – RLHF, RAG, fine-tuning and inferencing for LLMs and other DNNs