PURDUE UNIVERSITY

STyGIANet

Introduction

CS59000: AI/DC Networking

Fall 2025

Vamsi Addanki
vaddank@purdue.edu

https://stygianet.github.io

Datacenter Networks

Image source: https://www.google.com/about/datacenters/gallery/

PURDUE UNIVERSITY

STyGIANet

Datacenter Networks

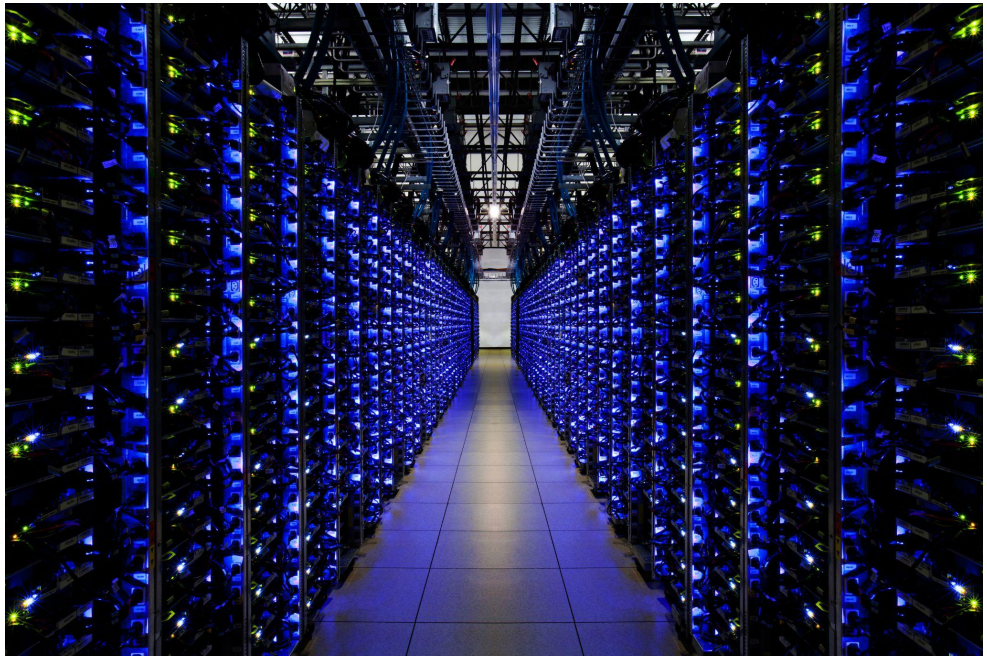


Image source: https://www.google.com/about/datacenters/gallery/

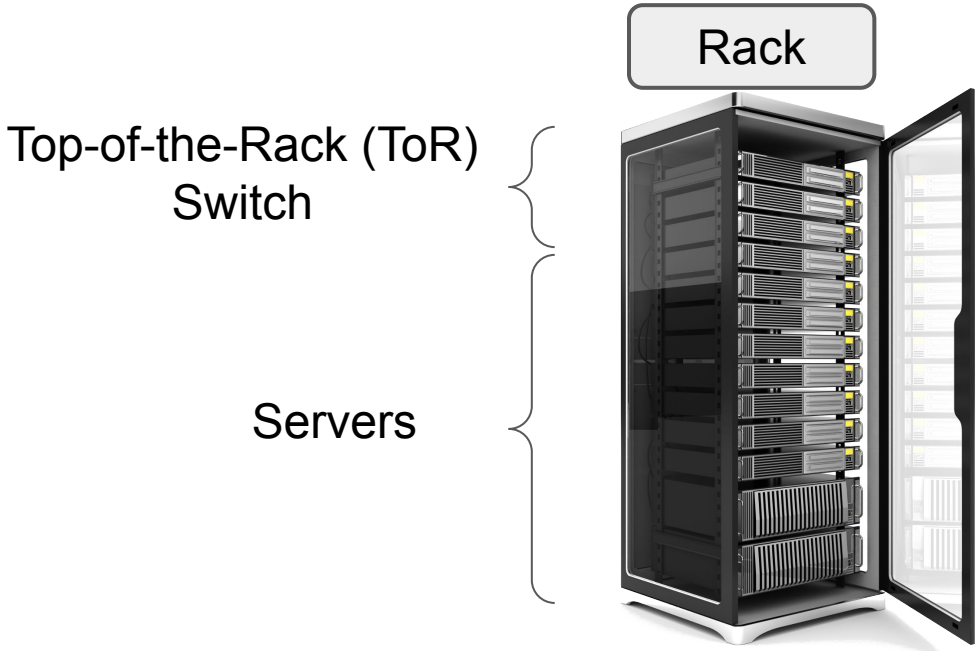
PURDUE UNIVERSITY

STyGIANet

Anatomy of Datacenter Networks

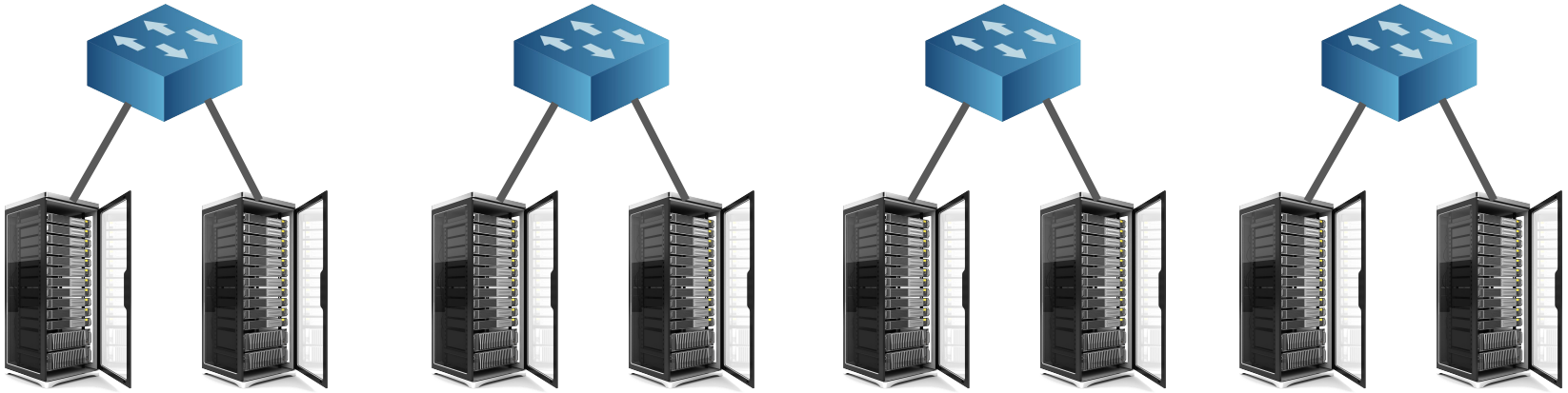
Server

Anatomy of Datacenter Networks



Anatomy of Datacenter Networks

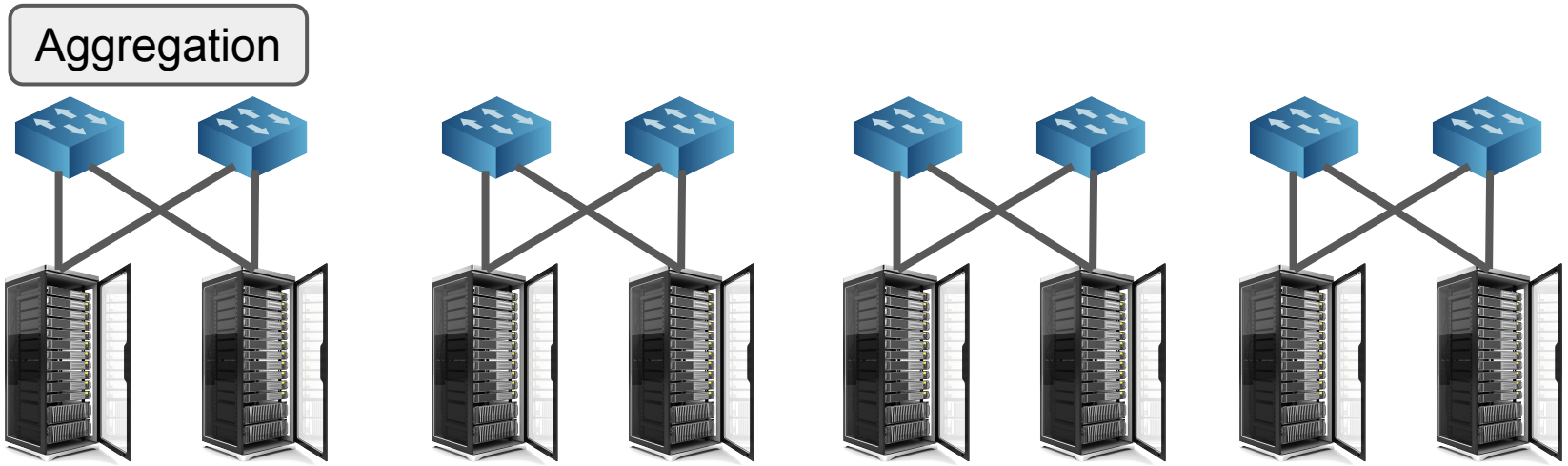
Anatomy of Datacenter Networks



PURDUE UNIVERSITY

STyGIANet

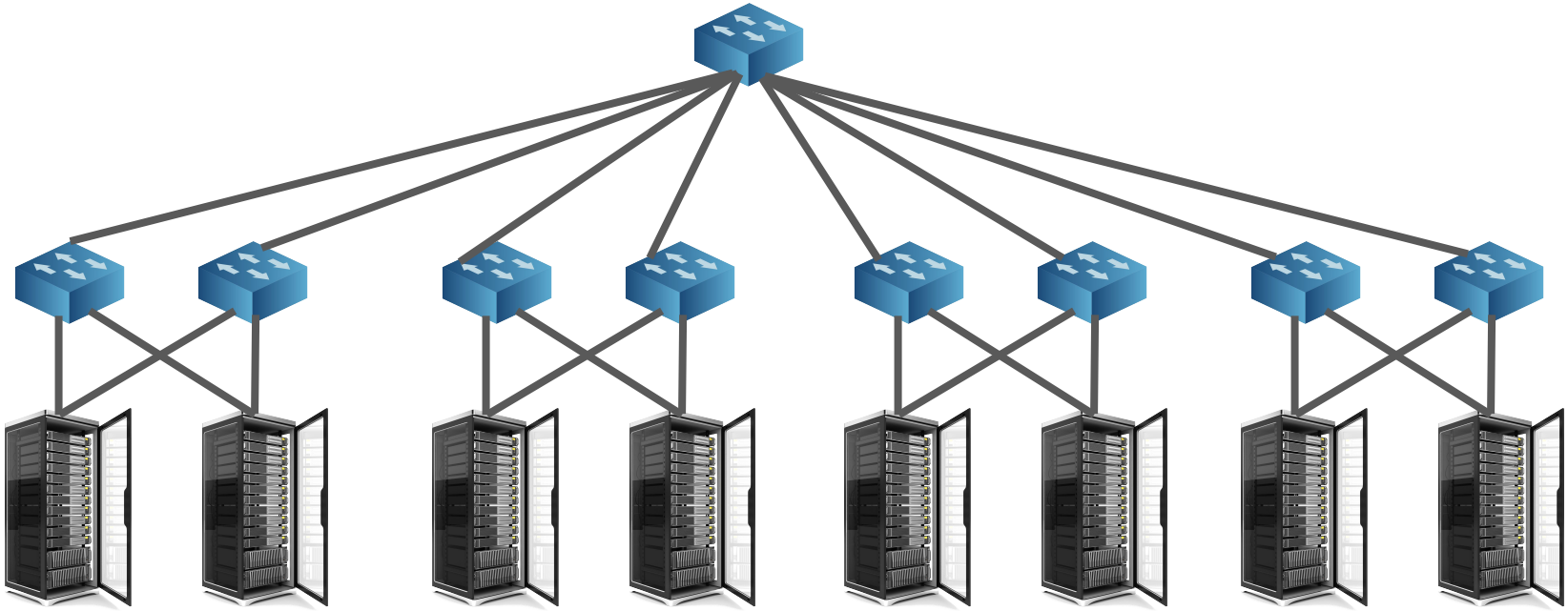
Anatomy of Datacenter Networks



PURDUE UNIVERSITY

STyGIANet

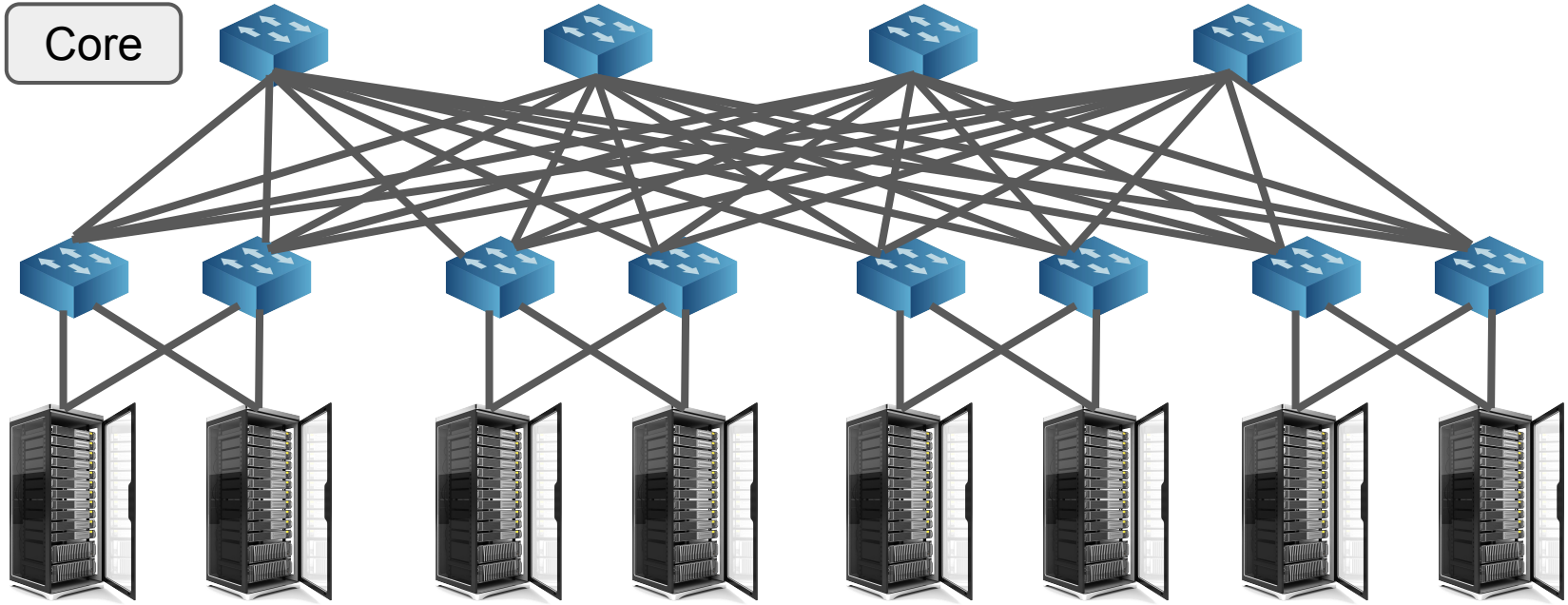
Anatomy of Datacenter Networks



PURDUE UNIVERSITY

STyGIANet

Anatomy of Datacenter Networks



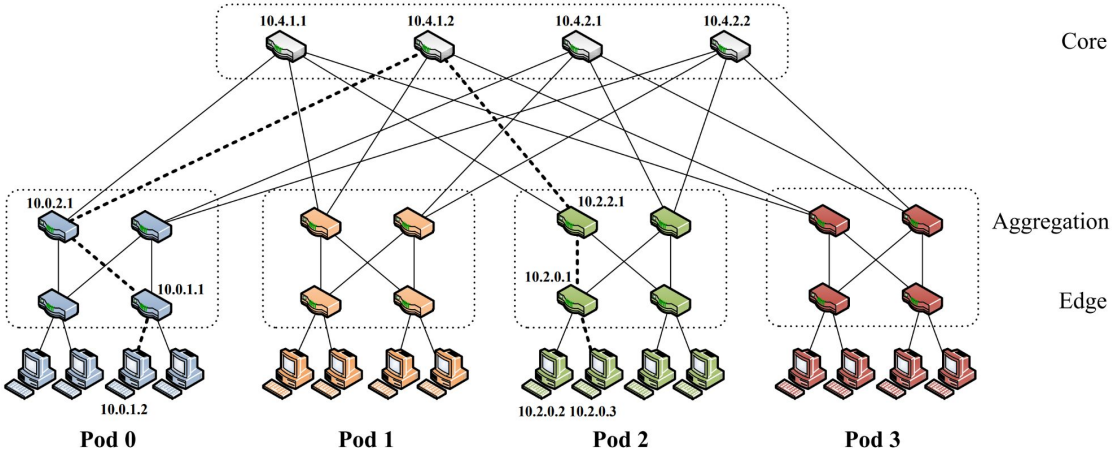
PURDUE UNIVERSITY

STyGIANet

Datacenter Topologies

- Predominantly Clos-based topologies are used in production datacenters
  - Google [1]
  - Meta [2]
  - Microsoft [3]
- Clos-based topologies offer:
  - Non-blocking network
  - Uniform and high bandwidth availability between servers

Fat Tree (Clos-based) Datacenter Topology



[4] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture. ACM SIGCOMM computer communication review. 2008 Aug 17;38(4):63-74.

PURDUE UNIVERSITY

STyGIANet

Further Reading: Alternative Topology Designs

- Topologies based on random graphs
  - JellyFish [5]
  - Xpander [6]
- Topologies optimized for Fault-tolerance
  - F10 [7]
- Topologies optimized for overall life cycle management of a datacenter
  - FatClique [8]
- Topologies optimized for specific traffic patterns (Self-adjusting Topologies)
  - **Upcoming seminar session**

Datacenters vs Internet

- Traffic Characteristics
- Autonomy
  - Large companies own datacenters and have partial or full control over their infrastructure.
  - Communication within Datacenter need not be standards compliant!
  - Internet is a collection of ASes, and controlled by many players.
  - Communication over the internet hence requires standards compliance.

Datacenter Applications

- Web search
- Datamining
- Web servers
- Cache Followers
- Hadoop
- DNN Training and various other ML workloads
- Netflix uses AWS, OpenAI uses Azure datacenters
- …

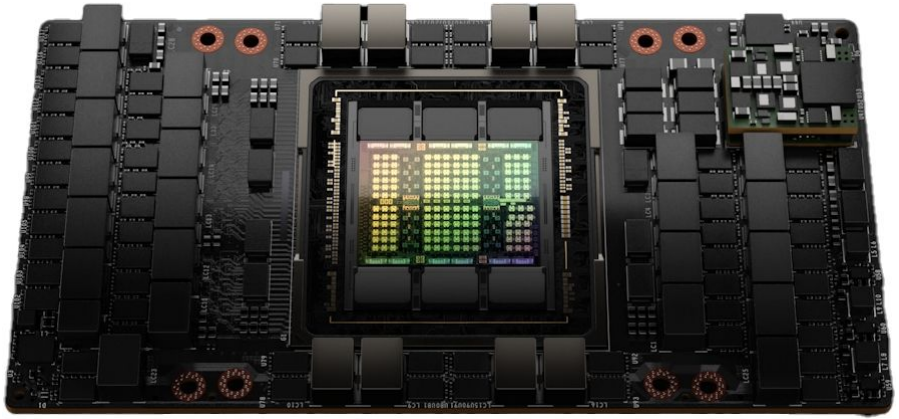
Datacenter Traffic Characteristics [Distributed Training]

- Traffic generated by GPU clusters training a large model in a distributed manner across multiple GPUs
- Traffic patterns are based on collective communication
  - All-to-All
  - All-reduce
  - …
- Flow sizes are mostly similar
- Synchronized communication patterns

More on Collective Communication in upcoming sessions!

Motivation: Parallel Computing

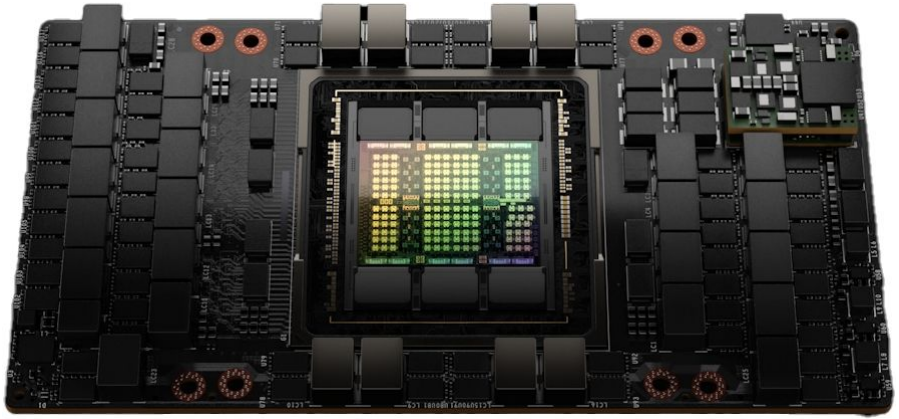
- Single GPU



Nvidia H100 GPU [9]

Motivation: Parallel Computing

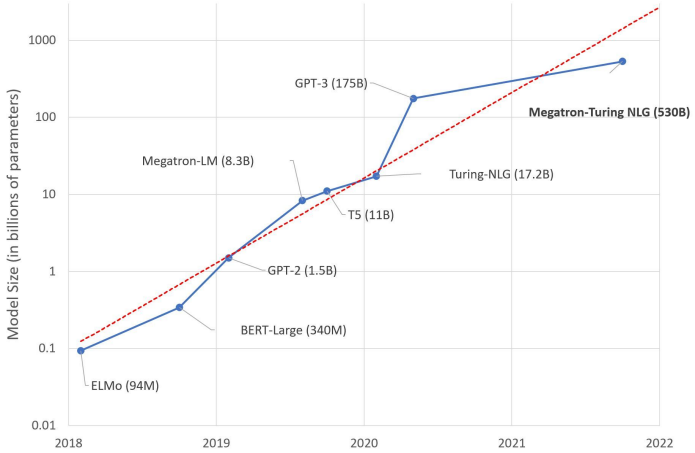
- Single GPU
  - **Limited compute resources and memory**



Nvidia H100 GPU [9]

Motivation: Parallel Computing

- Exponentially growing compute requirements of language models
  - *Can a single GPU support such large models?*



Trend of sizes of state-of-the-art NLP models over time [10]

PURDUE UNIVERSITY

STyGIANet

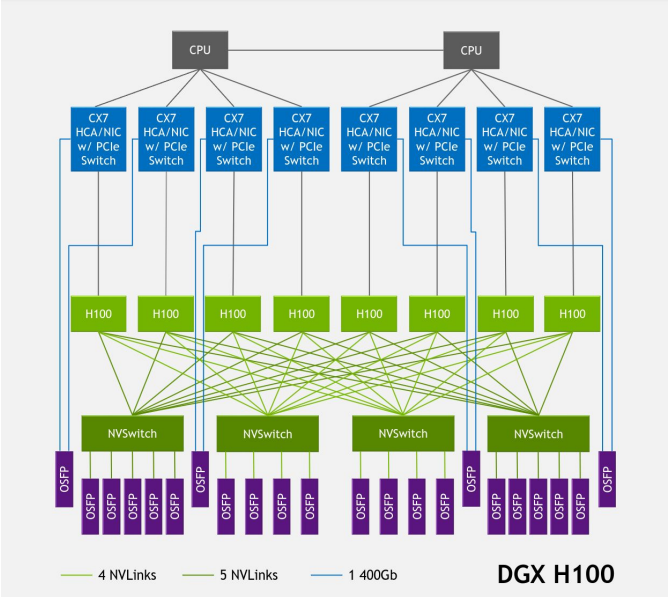
Motivation: Parallel Computing

- Enter parallelization

e.g., NVIDIA DGX server

Motivation: Parallel Computing

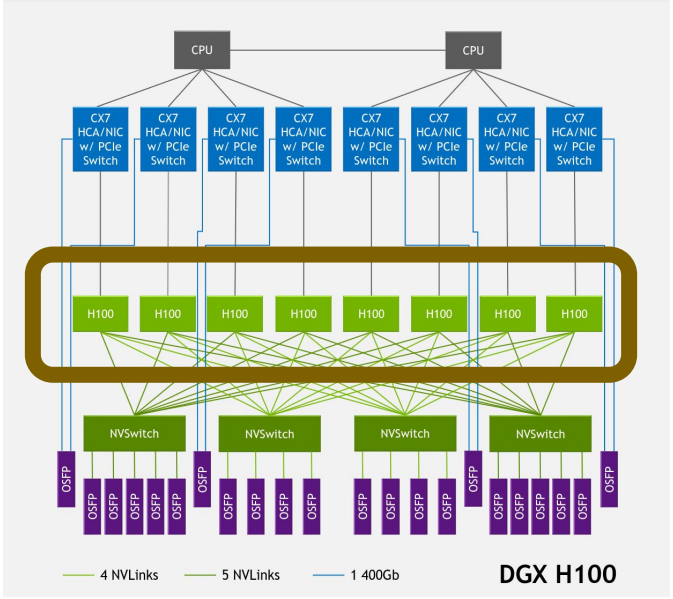
- Multi-GPU servers are on the rise



source [11]

Motivation: Parallel Computing

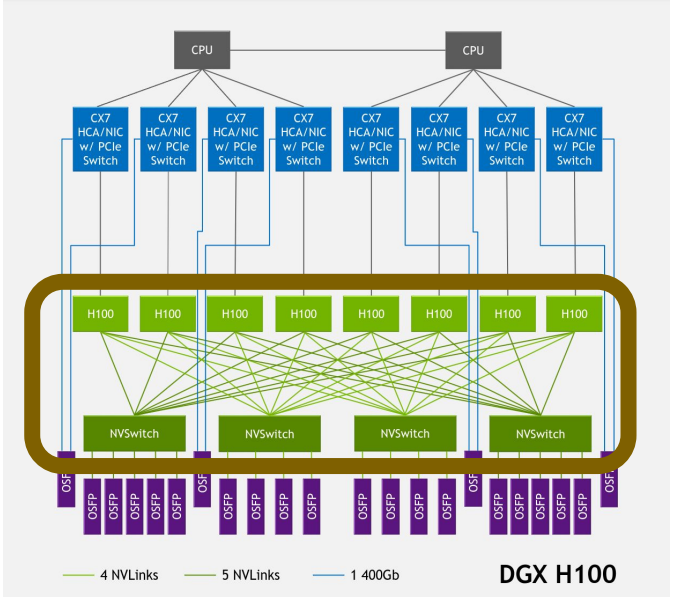
- Multi-GPU servers are on the rise
- e.g., NVIDIA DGX has 8 GPUs



source [11]

Motivation: Parallel Computing

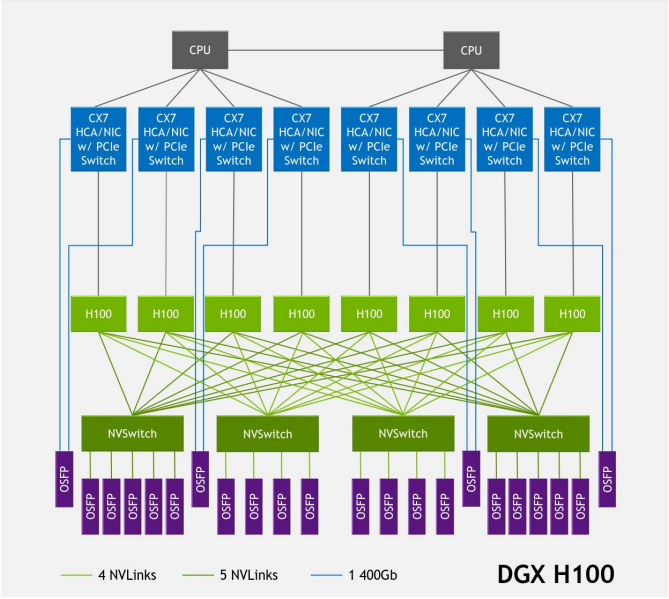
- Multi-GPU servers are on the rise
- e.g., NVIDIA DGX has 8 GPUs
- All 8 GPUs are interconnected to facilitate the *communication required for parallelization*



source [11]

Motivation: Parallel Computing

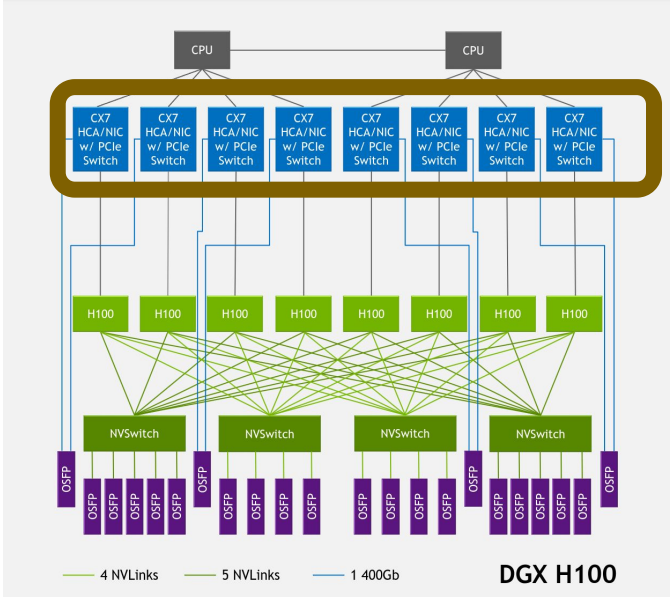
- Multi-GPU servers are on the rise
- e.g., NVIDIA DGX has 8 GPUs
- All 8 GPUs are interconnected to facilitate the *communication required for parallelization*
- A single task can now be parallelized across multiple compute resources



source [11]

Motivation: Parallel Computing

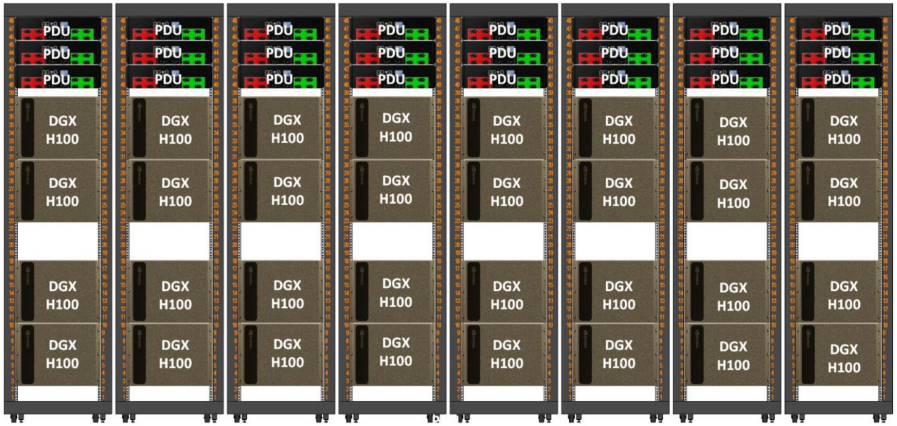
- Multi-GPU servers are on the rise
- e.g., NVIDIA DGX has 8 GPUs
- All 8 GPUs are interconnected to facilitate the *communication required for parallelization*
- A single task can now be parallelized across multiple compute resources
- Each GPU is also connected to a network interface card (NIC)
  - Allows connecting multiple DGX servers



source [11]

Motivation: Parallel Computing

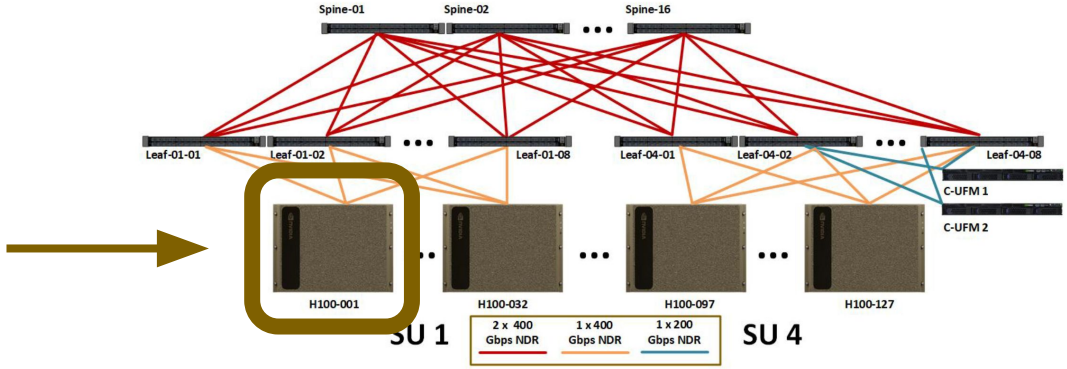
- Multiple multi-GPU servers can be further connected together to create a “Super Pod”



source [11]

Motivation: Parallel Computing

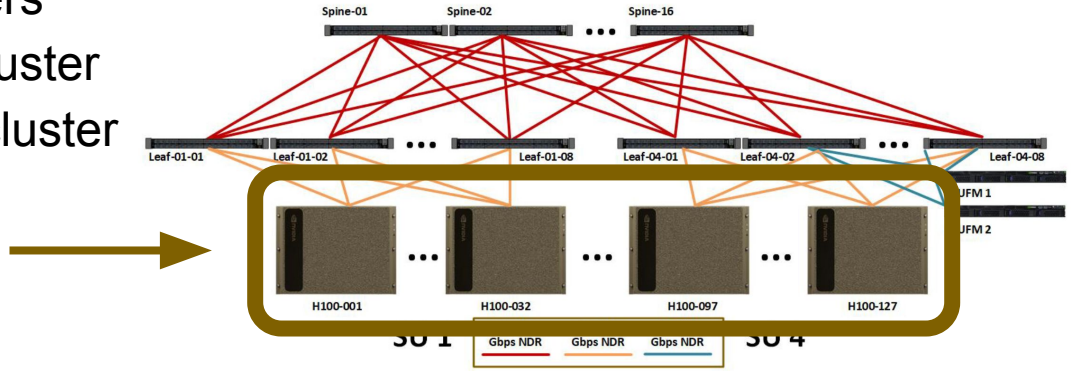
- Multiple multi-GPU servers can be further connected together to create a “Super Pod”
- Each DGX server provides 8 GPUs for parallelization



source [11]

Motivation: Parallel Computing

- Multiple multi-GPU servers can be further connected together to create a "Super Pod"
- Each DGX server provides 8 GPUs for parallelization
- Interconnecting multiple servers allows for scaling to a large cluster
- e.g, GPT-4 was trained on a cluster of ~25000 GPUs



source [11]

Motivation: Parallel Computing

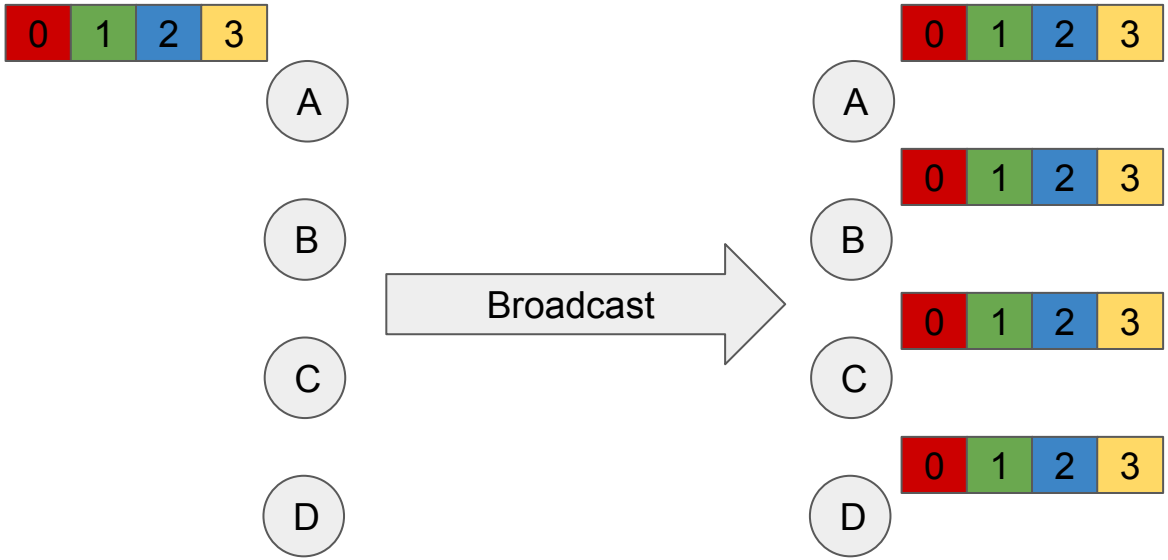
- GPUs ***communicate*** in order to aggregate the results globally
  - e.g., GPUs compute local gradients
  - GPUs in a cluster exchange gradients in order to aggregate results

Motivation: Parallel Computing

- Communication is an essential part of massively parallel distributed computing
- Later in this seminar:
  - Collective communication primitives
  - Cost model for analyzing collectives
  - Collective communication algorithms (Optimizing the comm. time)
- Fun fact: Collective communication is not new. GPU clusters are new. These algorithms have been widely studied for decades e.g., Message Passing Interface (MPI)
  - Nvidia's NCCL, AMD's RCCL,.... are all basically reincarnations of MPI
  - "CCL" stands for Collective Communication Library

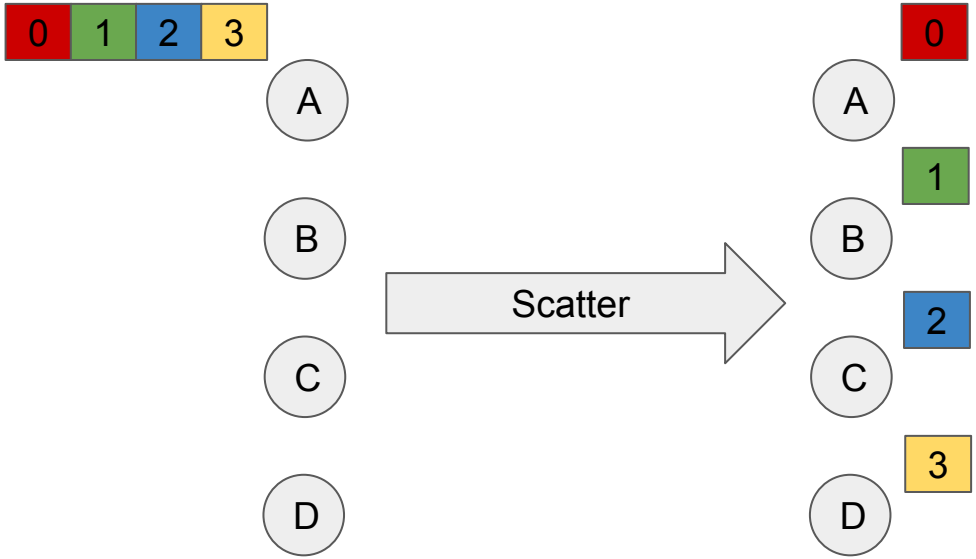
Collective Communication Primitives: Broadcast

- A single node transmits its entire data to all other nodes



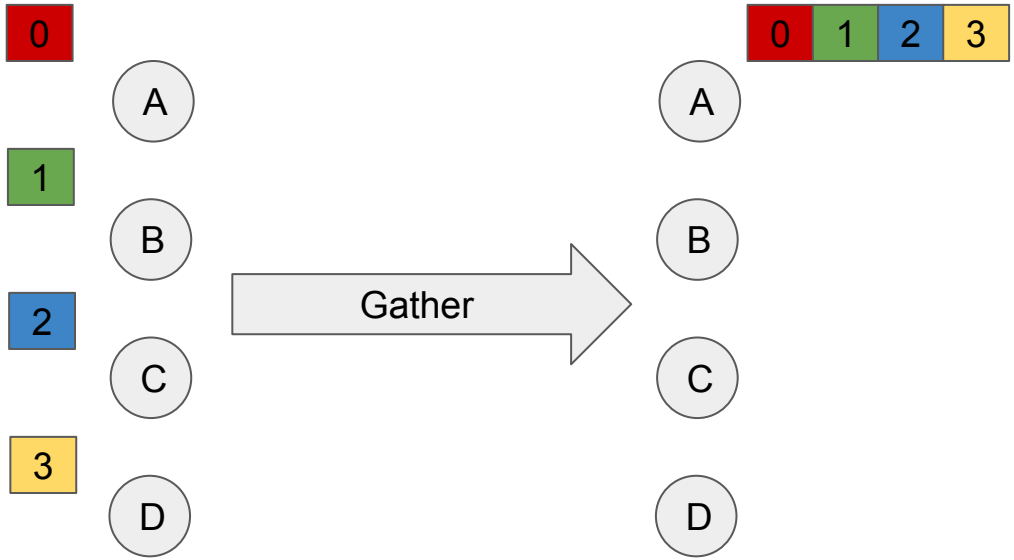
Collective Communication Primitives: Scatter

- A single node transmits *distinct* chunks of its data to all other nodes



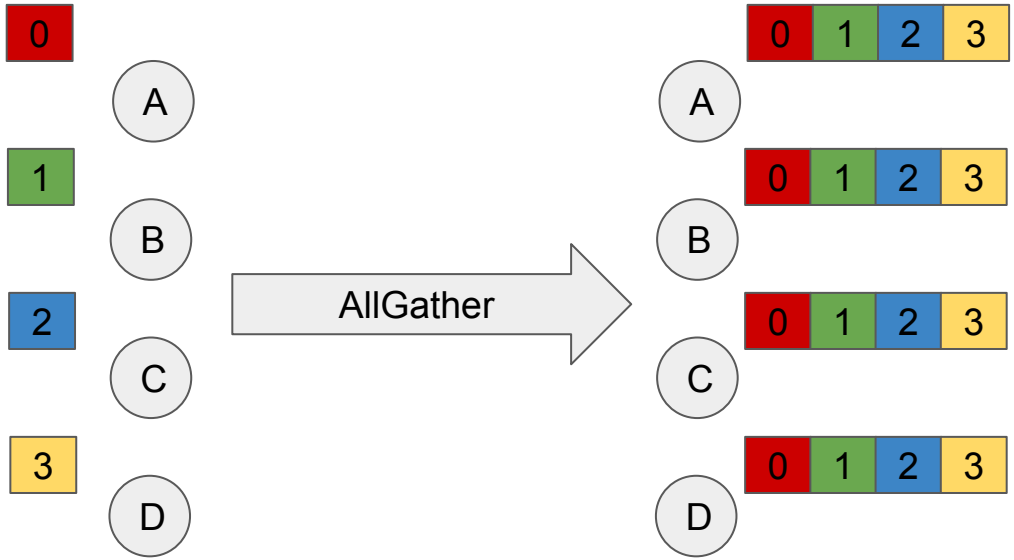
Collective Communication Primitives: Gather

- Every node transmits *distinct* chunks of its data to a single (root) node



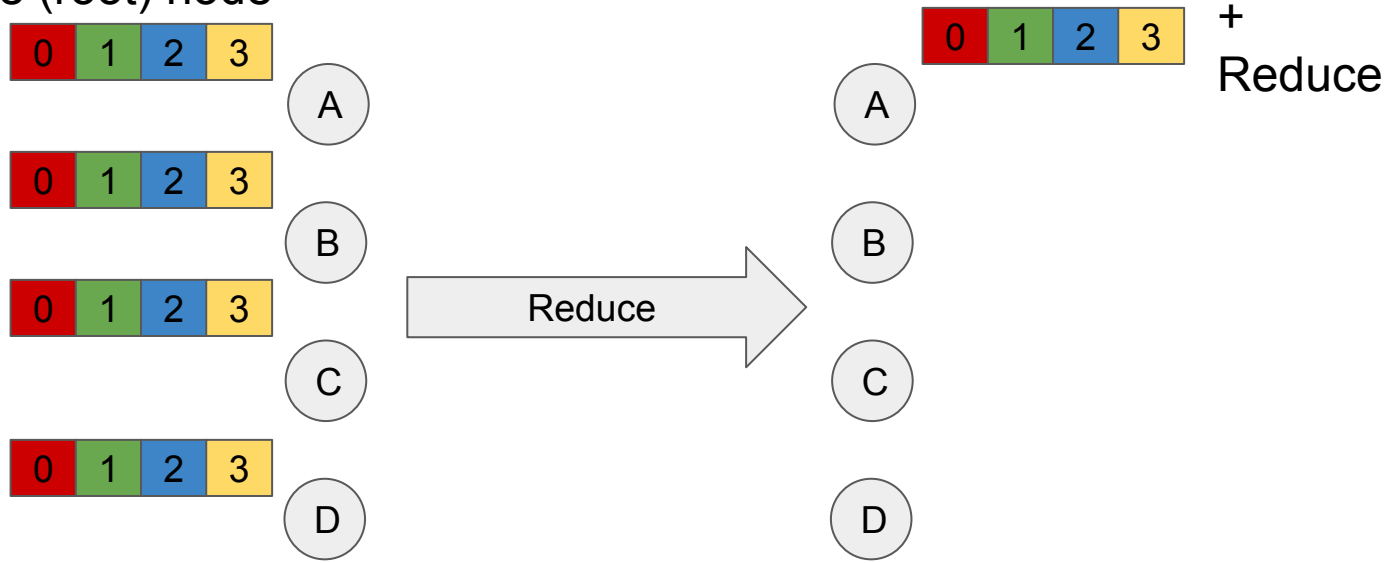
Collective Communication Primitives: AllGather

- Every node transmits *distinct* chunks of its data to all other nodes



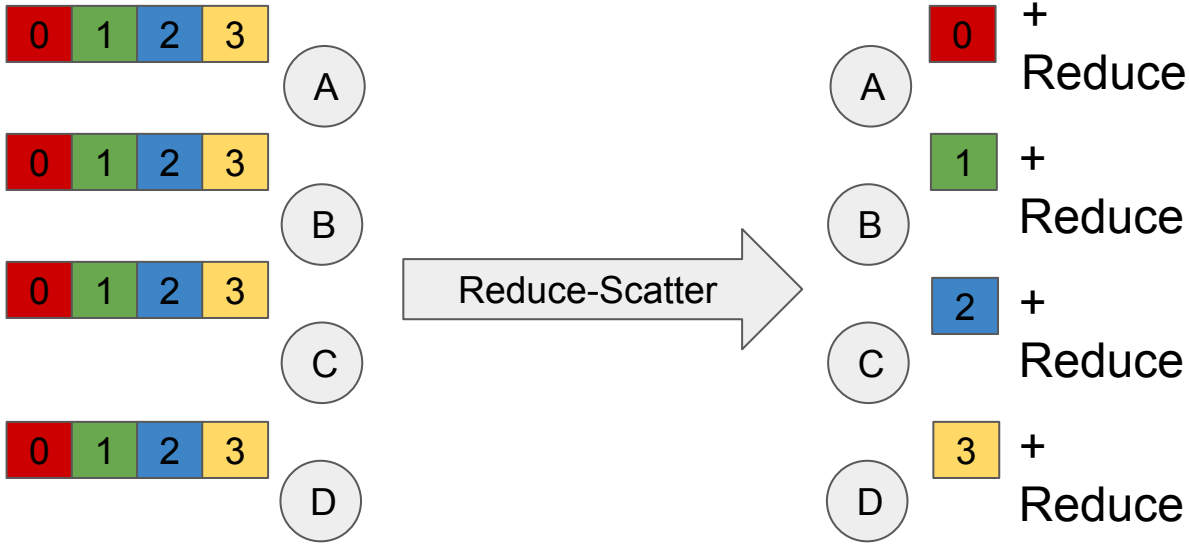
Collective Communication Primitives: Reduce

- Data across all the nodes is globally aggregated (reduced e.g., sum, max) at a single (root) node



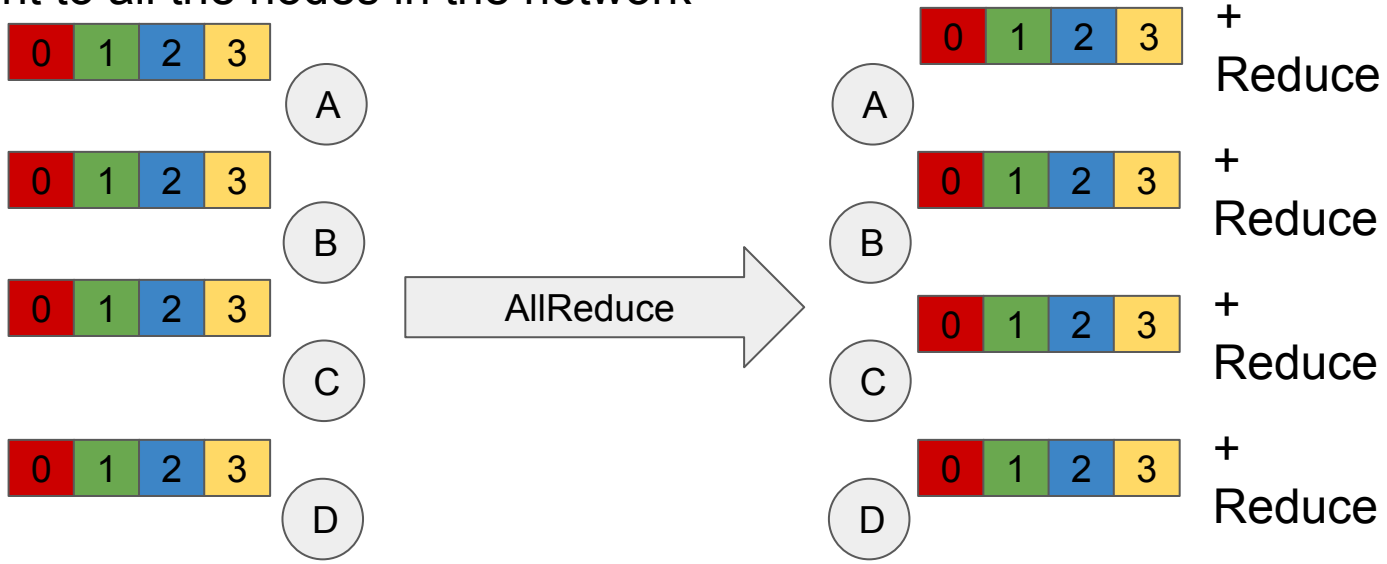
Collective Communication Primitives: Reduce-Scatter

- Data across all the nodes is globally aggregated (reduced e.g., sum, max) and distinct chunks of the reduced results are scattered across the nodes

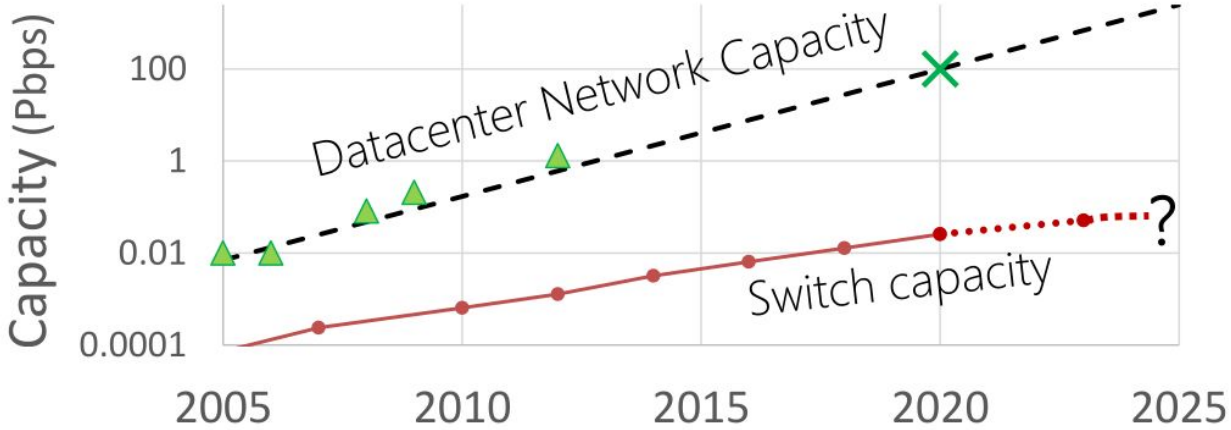


Collective Communication Primitives: AllReduce

- Data across all the nodes is globally aggregated (reduced e.g., sum, max) and sent to all the nodes in the network



Network Demand vs Capacity **Mismatch**



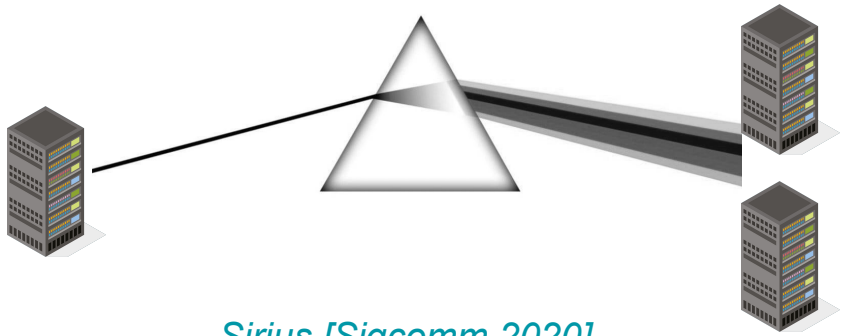
[12] A flat datacenter network with nanosecond optical switching (SIGCOMM 2020)

PURDUE UNIVERSITY

STyGIANet

Reconfigurable Datacenter Networks

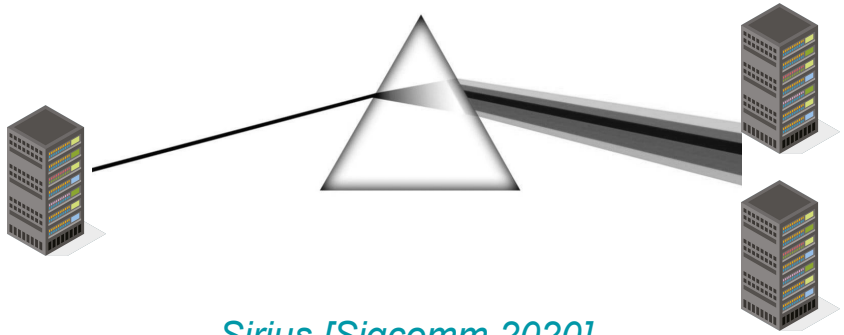
- Generalization of the design space: ***Topology can change over time***
- Static networks are a special case



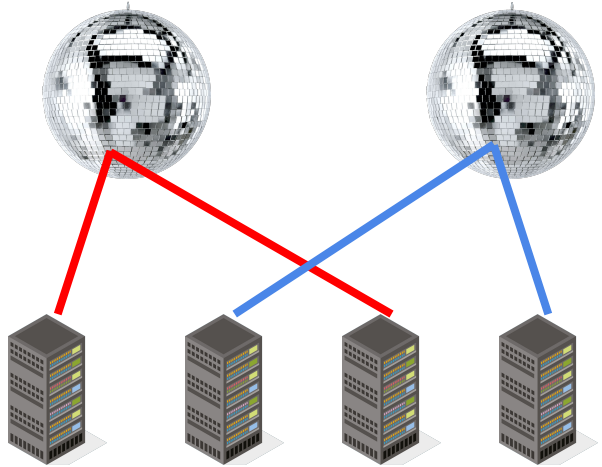
Sirius [Sigcomm 2020]

Reconfigurable Datacenter Networks

- Generalization of the design space: ***Topology can change over time***
- Static networks are a special case



Sirius [Sigcomm 2020]

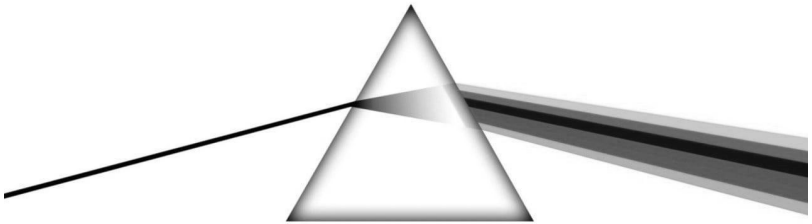


ProjecToR [Sigcomm 2016]

PURDUE UNIVERSITY

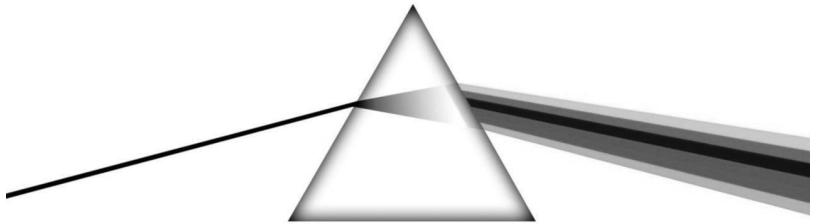
STyGIANet

Emerging Technologies - Photonics on the Rise

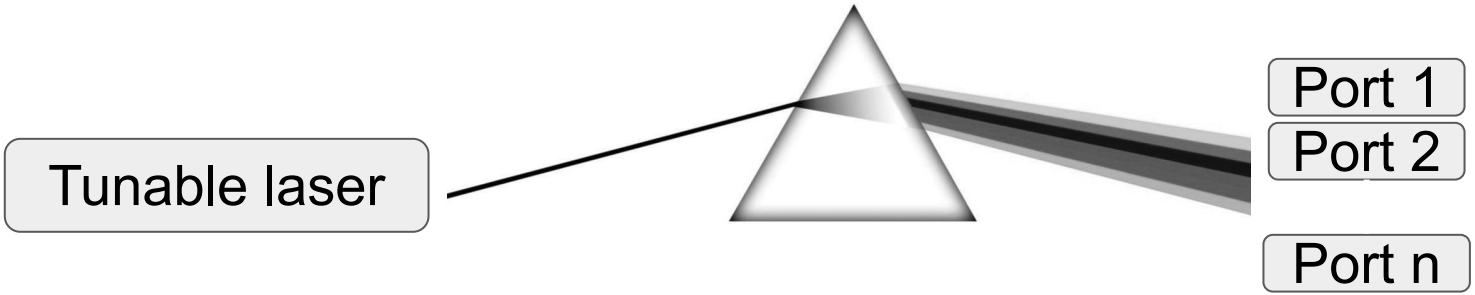


Emerging Technologies - Photonics on the Rise

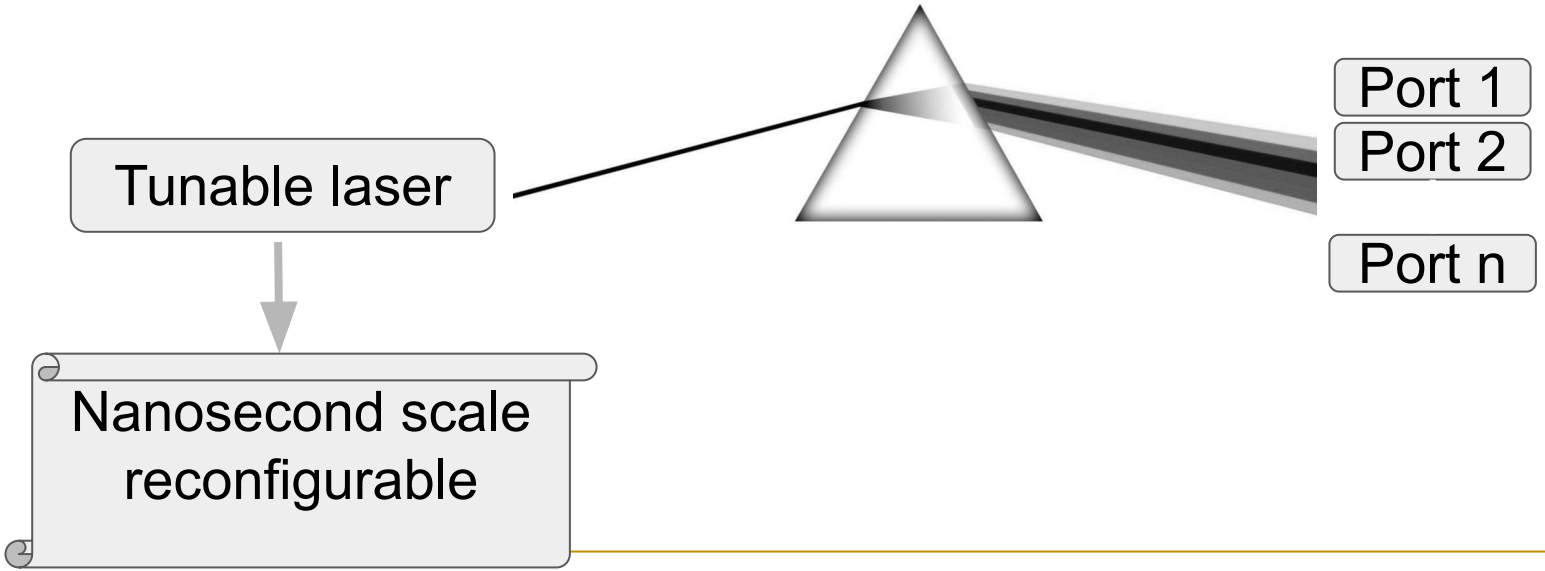
Tunable laser



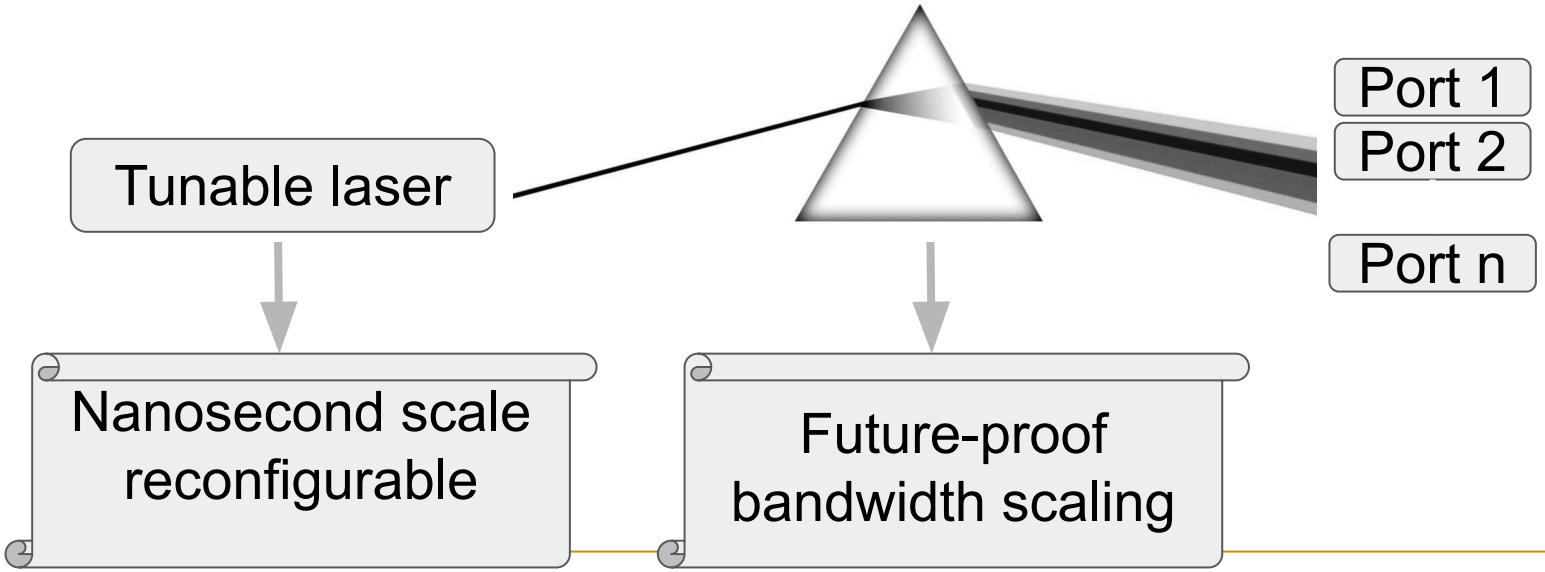
Emerging Technologies - Photonics on the Rise



Emerging Technologies - Photonics on the Rise



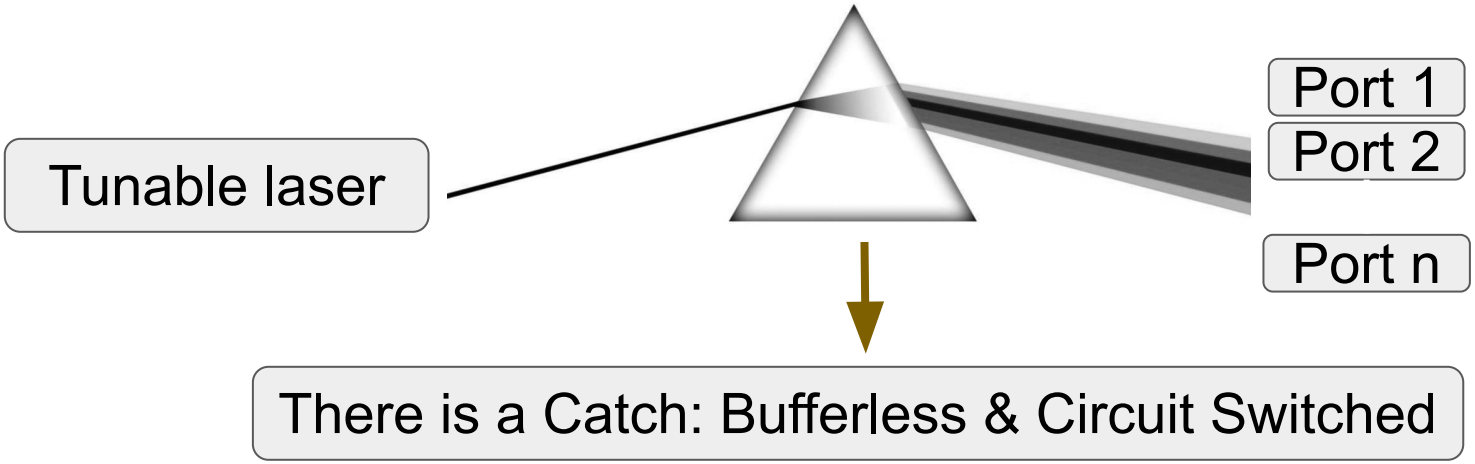
Emerging Technologies - Photonics on the Rise

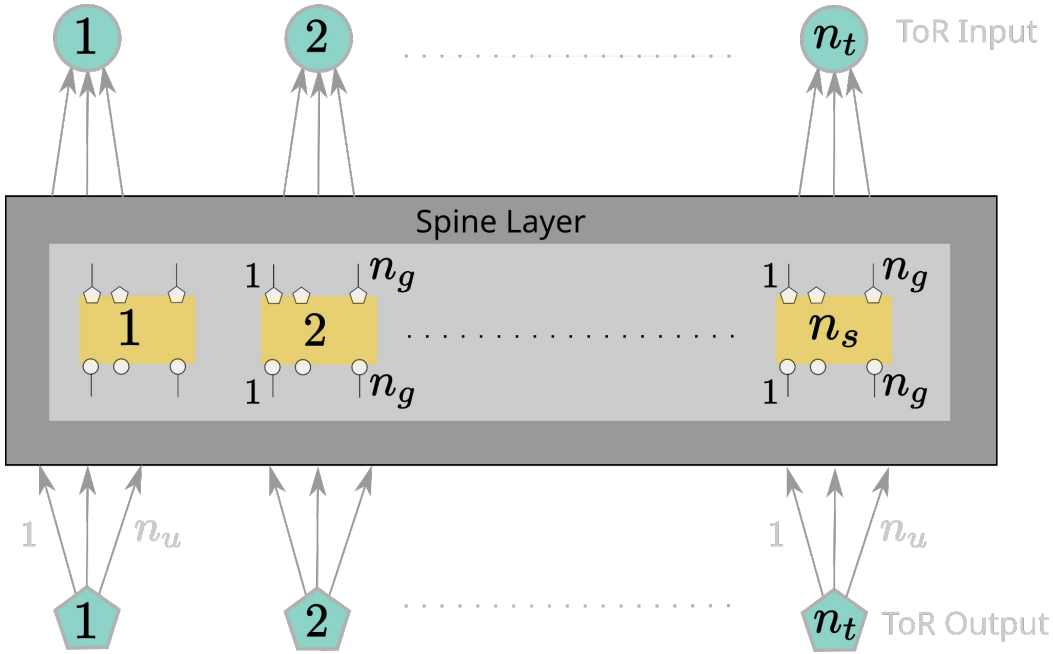


PURDUE UNIVERSITY

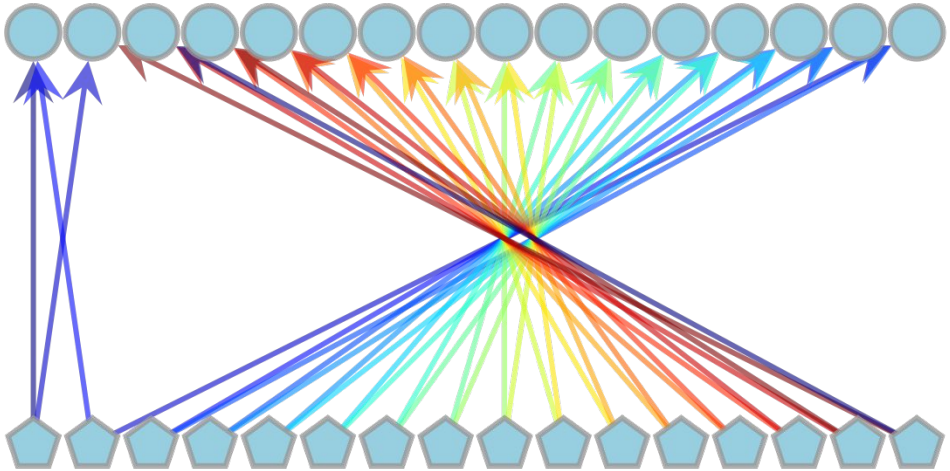
STyGIANet

Emerging Technologies - Photonics on the Rise




1
2
n_t
ToR Input
Spine Layer
1
1
n_g
2
1
n_g
1
n_g
n_s
1
n_g
1
n_u
1
n_u
1
2
n_t
ToR Output

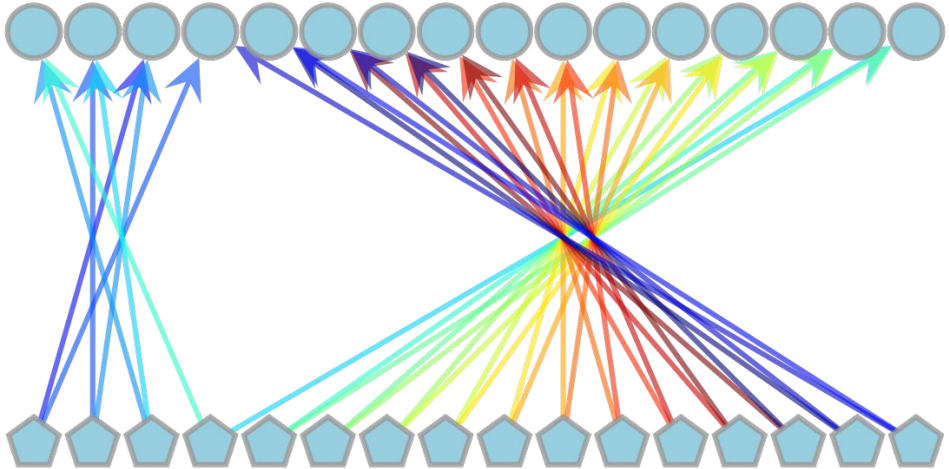
Timeslot 1



PURDUE UNIVERSITY

STyGIANet

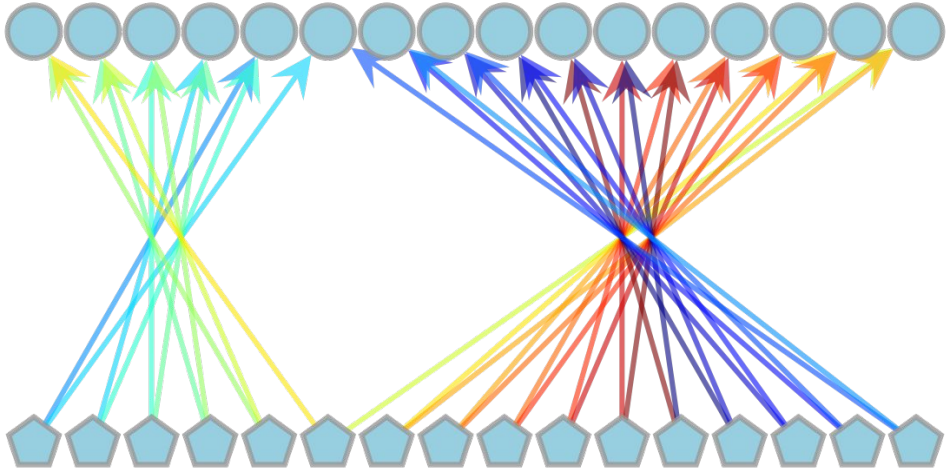
Timeslot 2



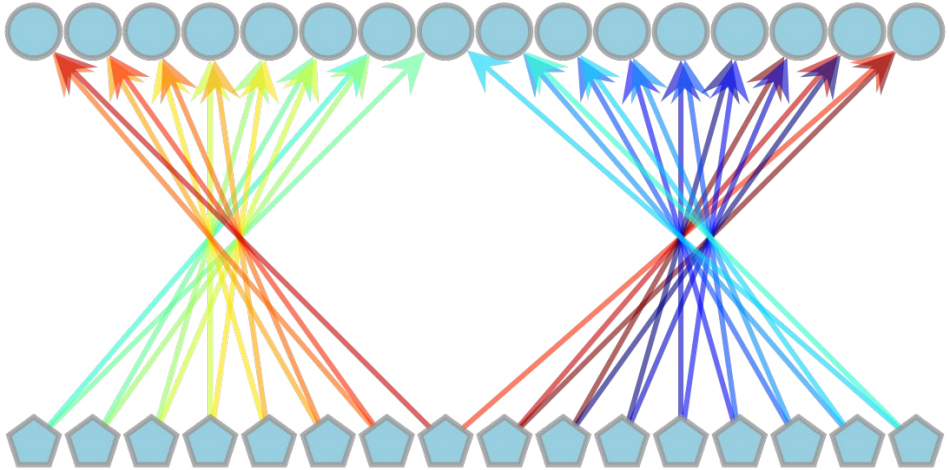
PURDUE UNIVERSITY

STyGIANet

Timeslot 3



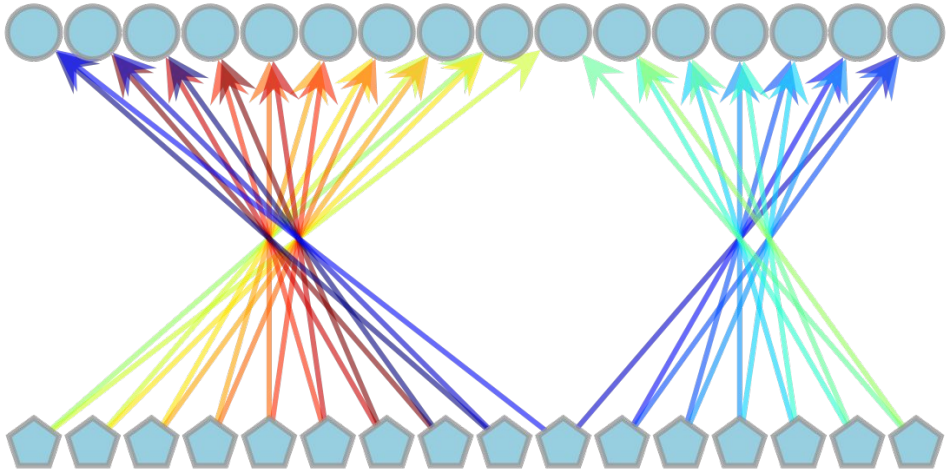
Timeslot 4



PURDUE UNIVERSITY

STyGIANet

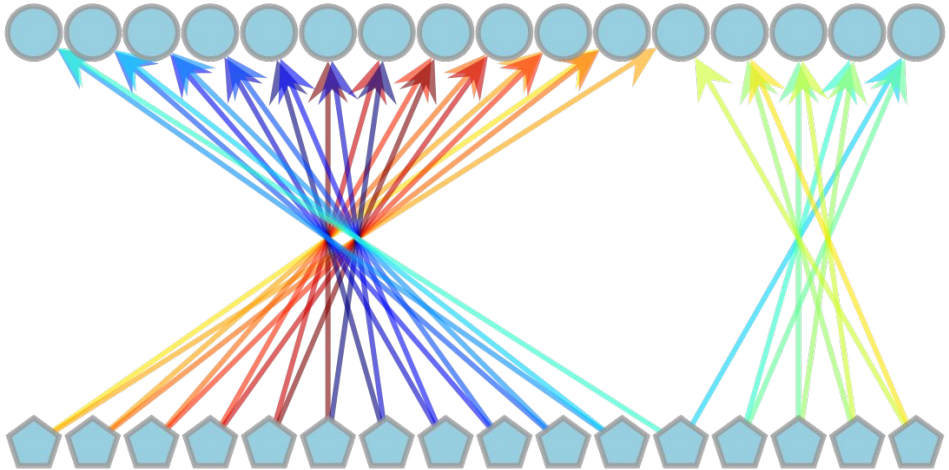
Timeslot 5



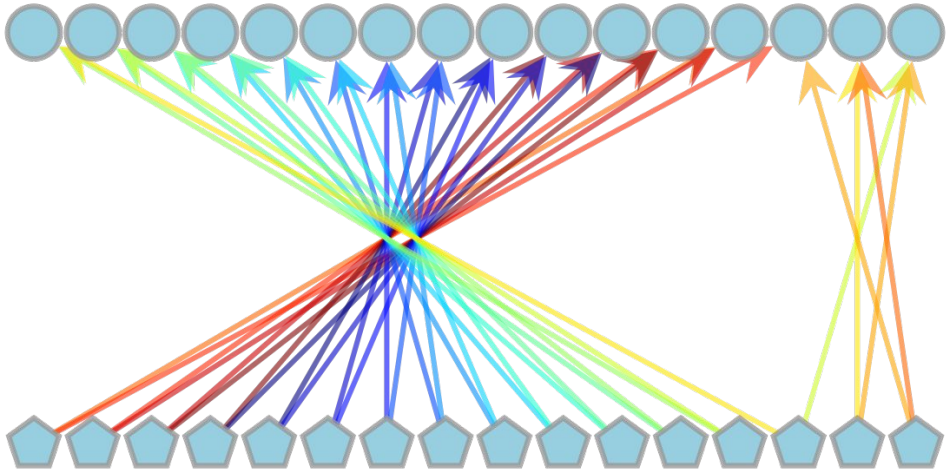
PURDUE UNIVERSITY

STyGIANet

Timeslot 6



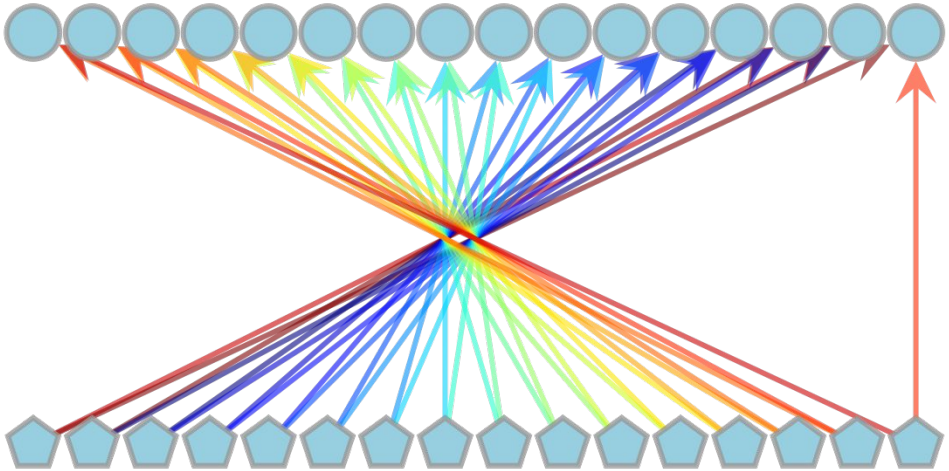
Timeslot 7



PURDUE UNIVERSITY

STyGIANet

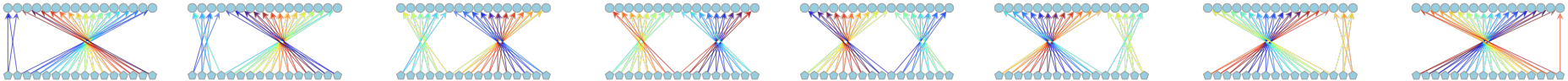
Timeslot 8



PURDUE UNIVERSITY

STyGIANet

Evolving Graph



Seminar Topics

- Large-scale LLM Training Network Architectures
- Collective Communication
- Photonic Interconnects
- AI for Networking

Course website: https://stygianet.github.io/courses/2025fallaidc

Schedule is fluid and much relaxed! The topics will be adjusted on-demand.

Grading

- 3 Assignments: 25%
- Mid-term exam (oral exam): 25%
- Final project/paper: 50%
  - Happy to help and work with you on finding a topic for research/project
  - Last two weeks of the schedule are dedicated for project feedback, final presentations, and submissions

Alright, enough formalities. Let's have some fun learning and building networks! 

The End