sigcomm 2025 COIMBRA

iSingLab HKUST

MixNet: A Runtime Reconfigurable Optical-Electrical Fabric for Distributed Mixture-of-Experts Training

Xudong Liao, Yijun Sun, Han Tian, Xinchen Wan, Yilun Jin,
Zilong Wang, Zhenghang Ren, Xinyang Huang, Wenxue Li, Kin Fai Tse,
Zhizhen Zhong, Guyue Liu, Ying Zhang, Xiaofeng Ye, Yiming Zhang, Kai Chen

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

MIT

PEKING UNIVERSITY 1898

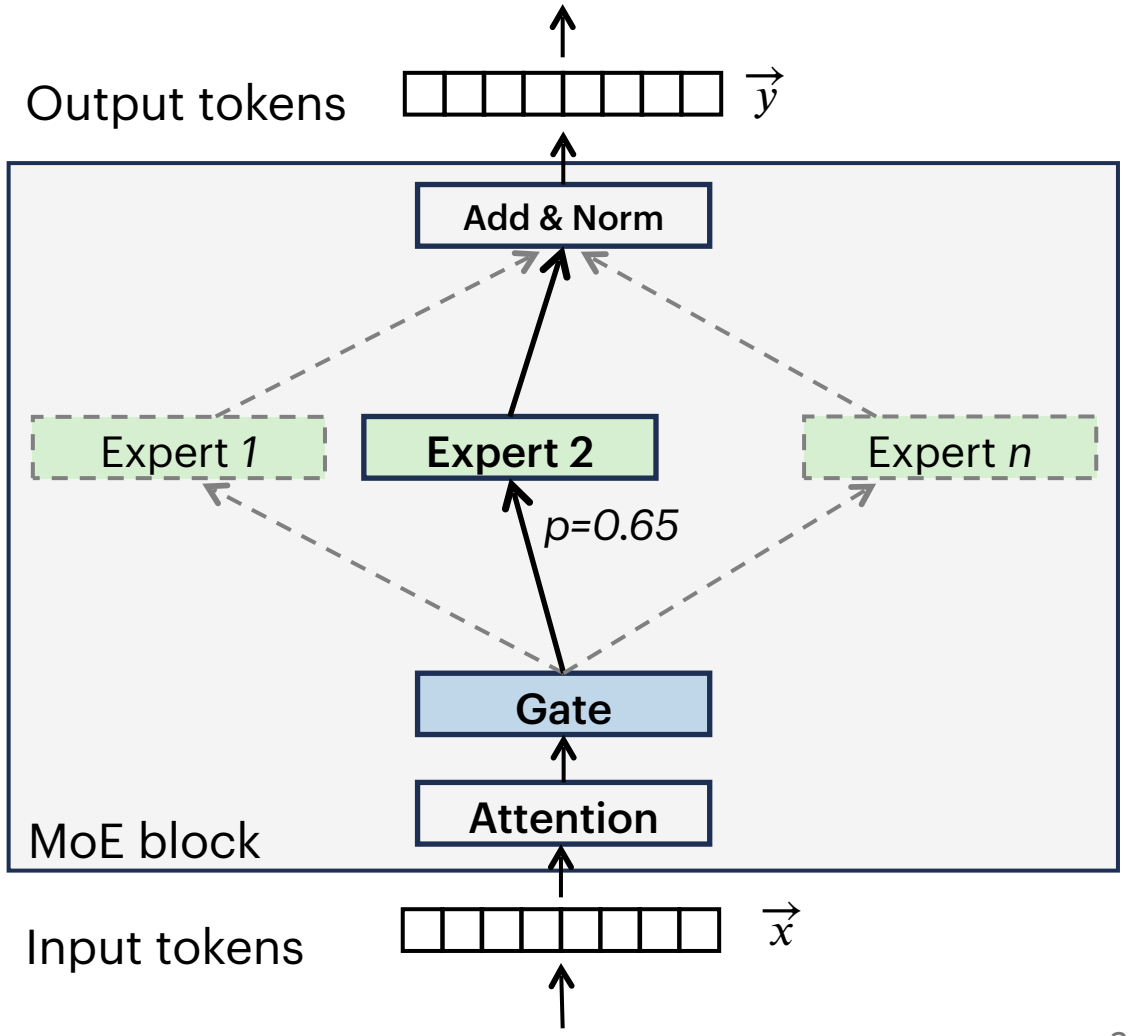
Meta

恒为科技 EmbedWay

Mixture-of-Experts Training

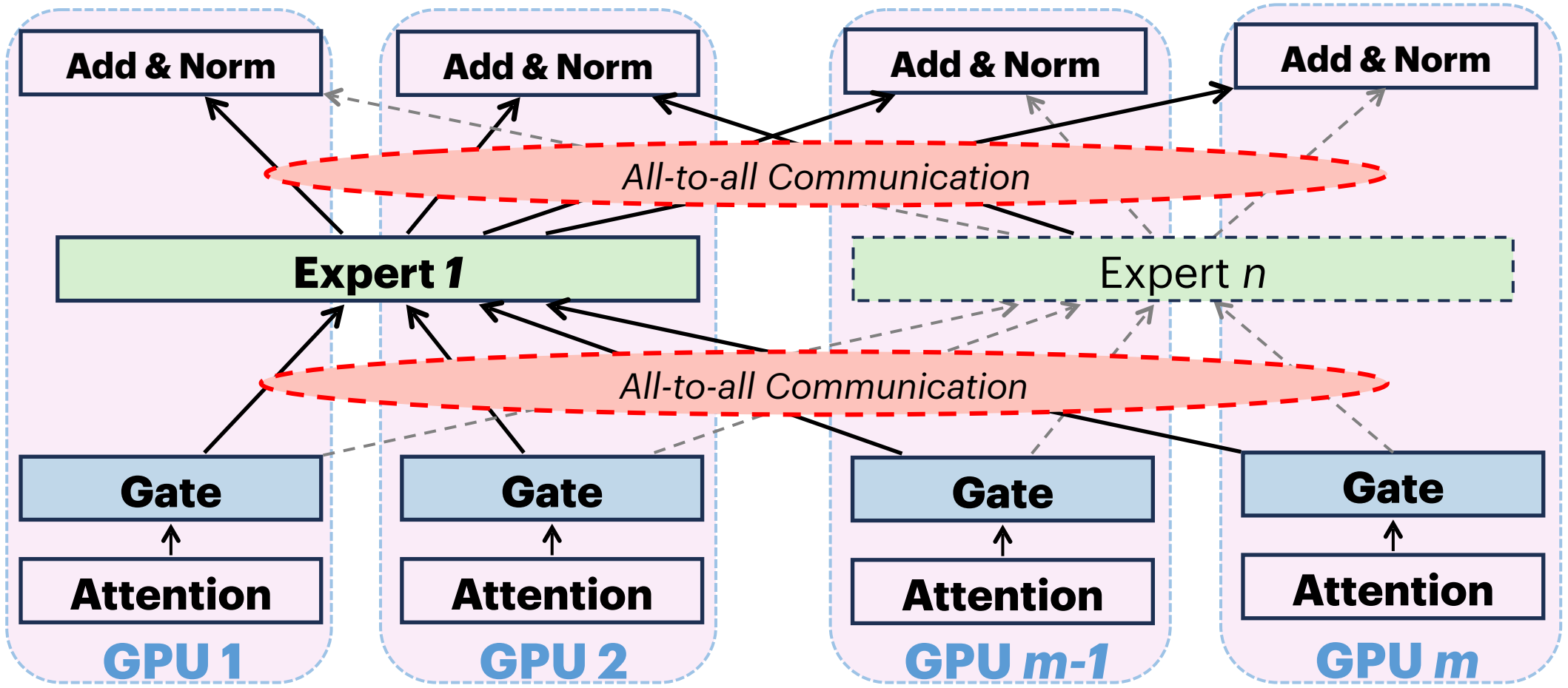
Mixture-of-Experts (MoE) models:

- ➢Sparse architecture with partially activated ***experts***
  - ➢Non-linearly increasing computation cost with increasing model size
- ➢Computation-based routing on each token
  - ➢Dynamic token trajectories at runtime



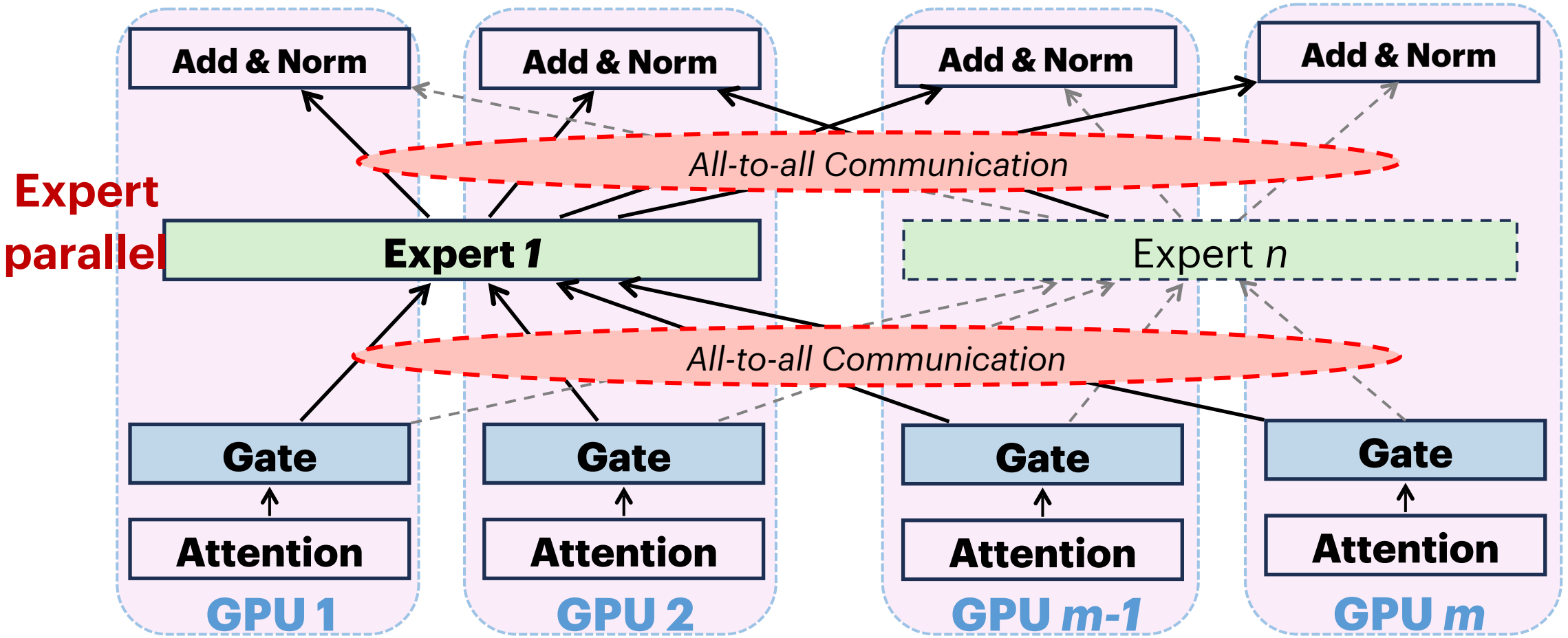
Parallelisms in MoE Training

Expert Parallel (EP) in ***Mixture-of-Experts*** (MoE) training



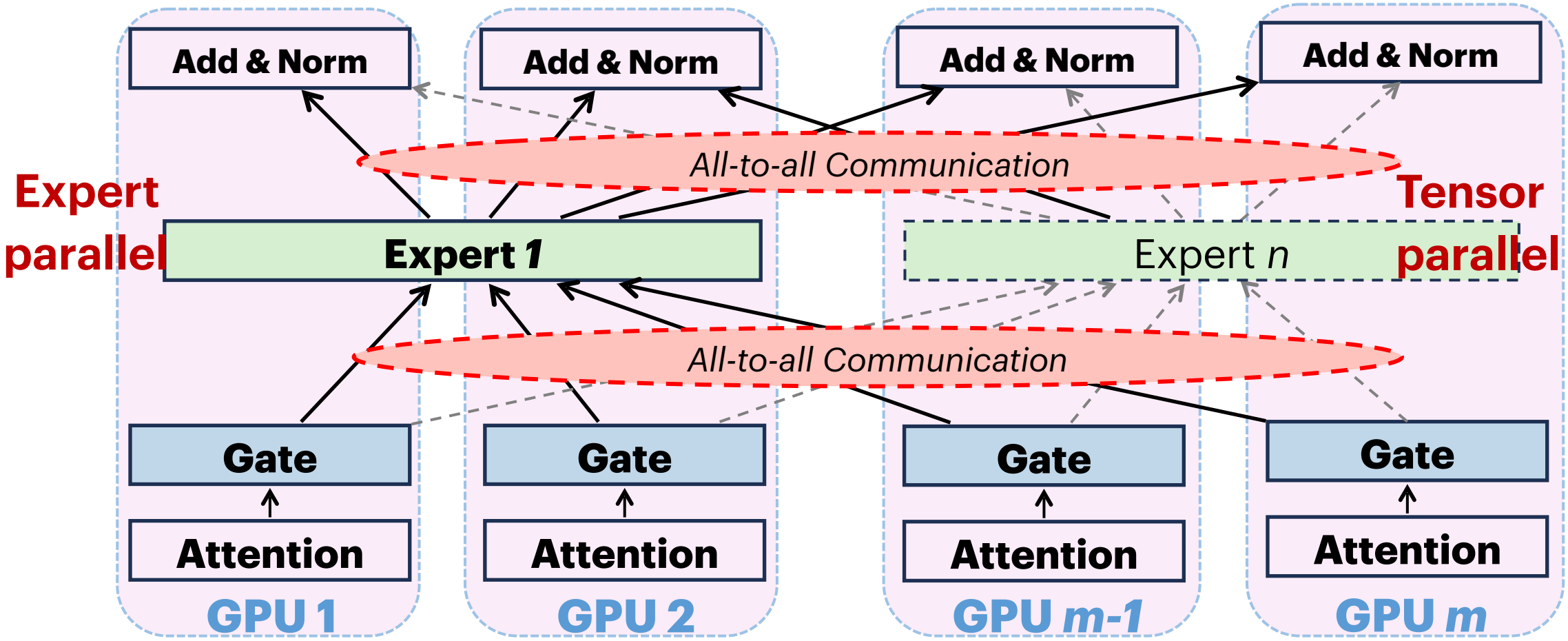
Parallelisms in MoE Training

Expert Parallel (EP) in ***Mixture-of-Experts*** (MoE) training



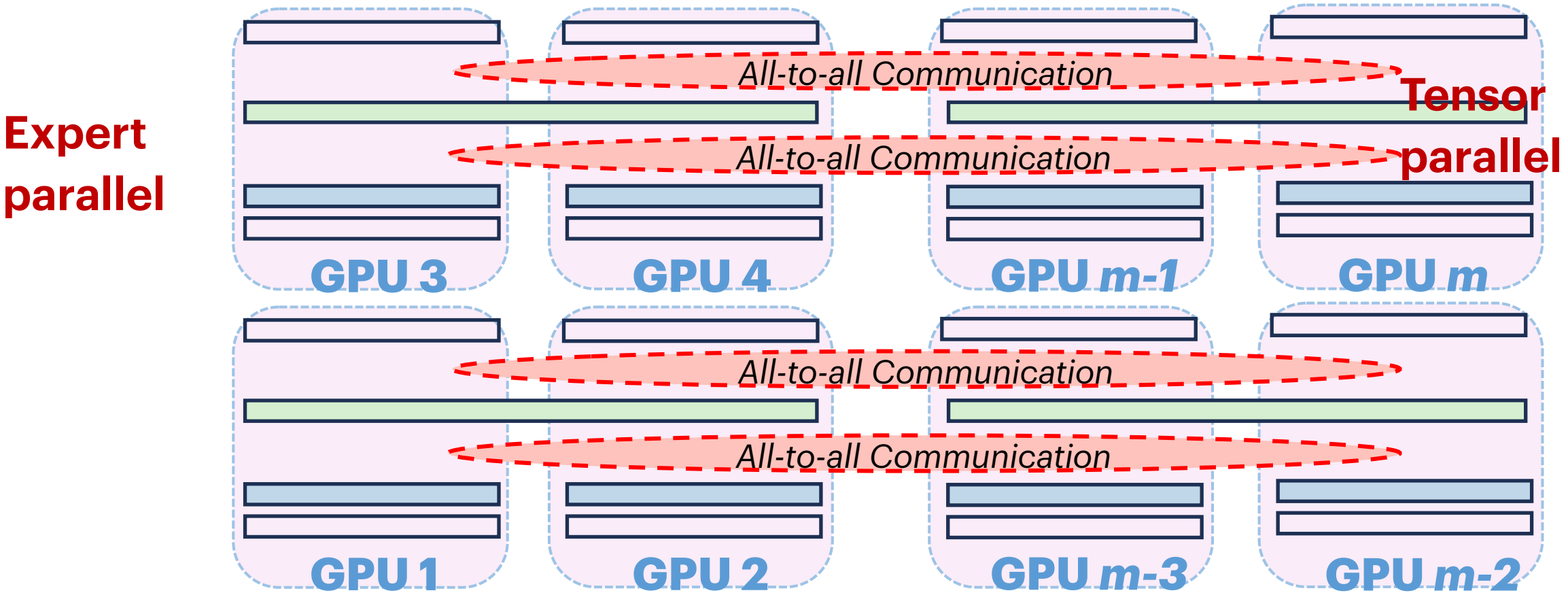
Parallelisms in MoE Training

Expert Parallel (EP) in ***Mixture-of-Experts*** (MoE) training



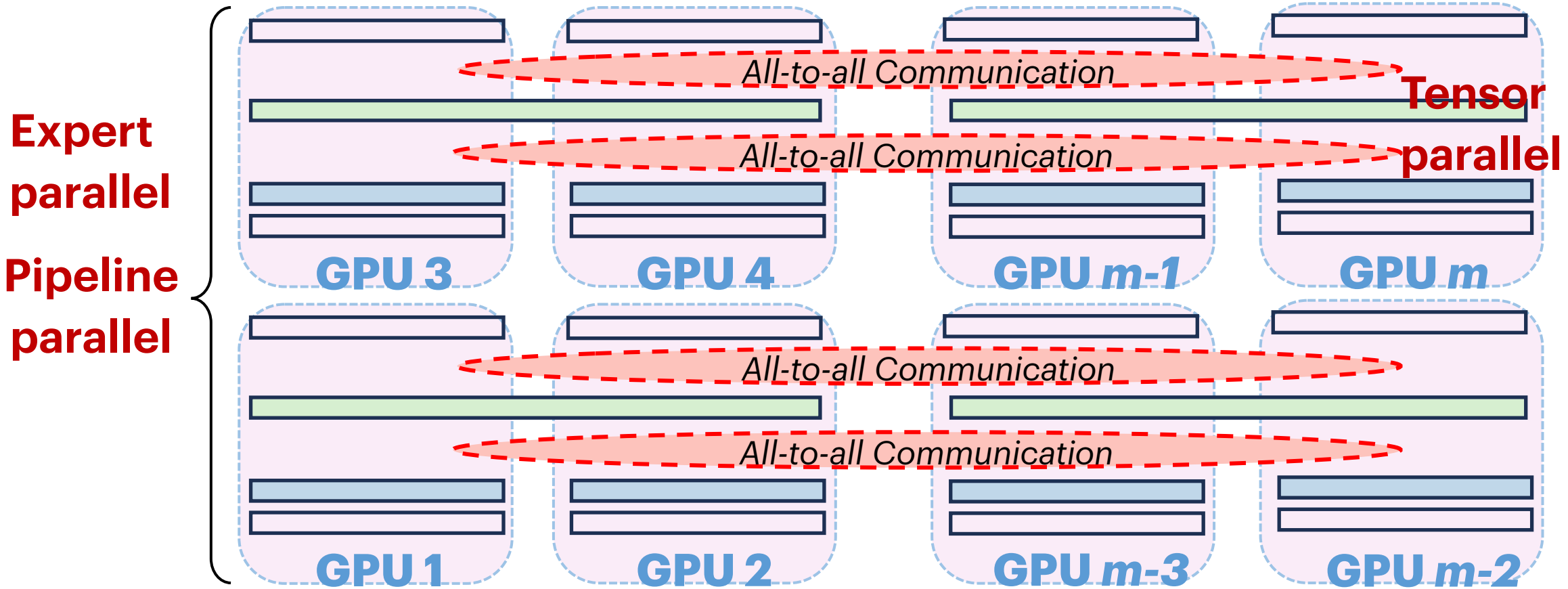
Parallelisms in MoE Training

State-of-the-art MoE training utilizes the ***hybrid parallelisms***: EP + TP + PP + DP



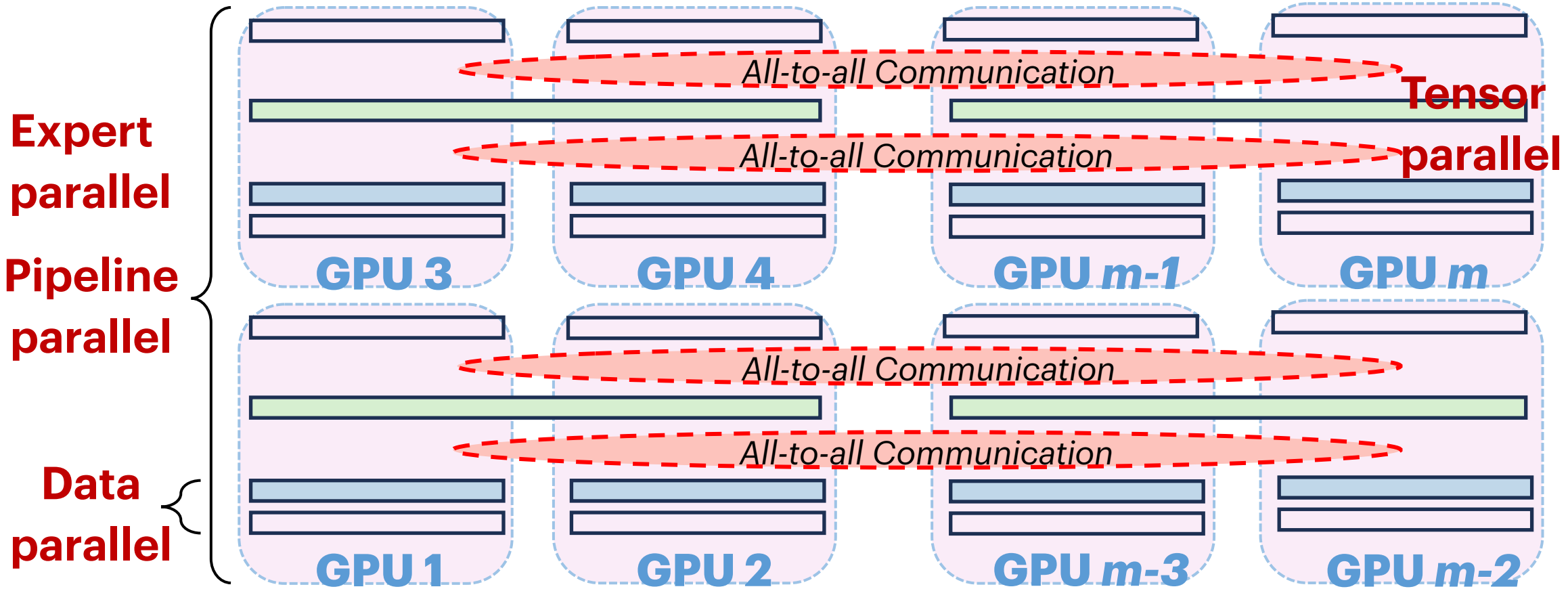
Parallelisms in MoE Training

State-of-the-art MoE training utilizes the ***hybrid parallelisms***: EP + TP + PP + DP



Parallelisms in MoE Training

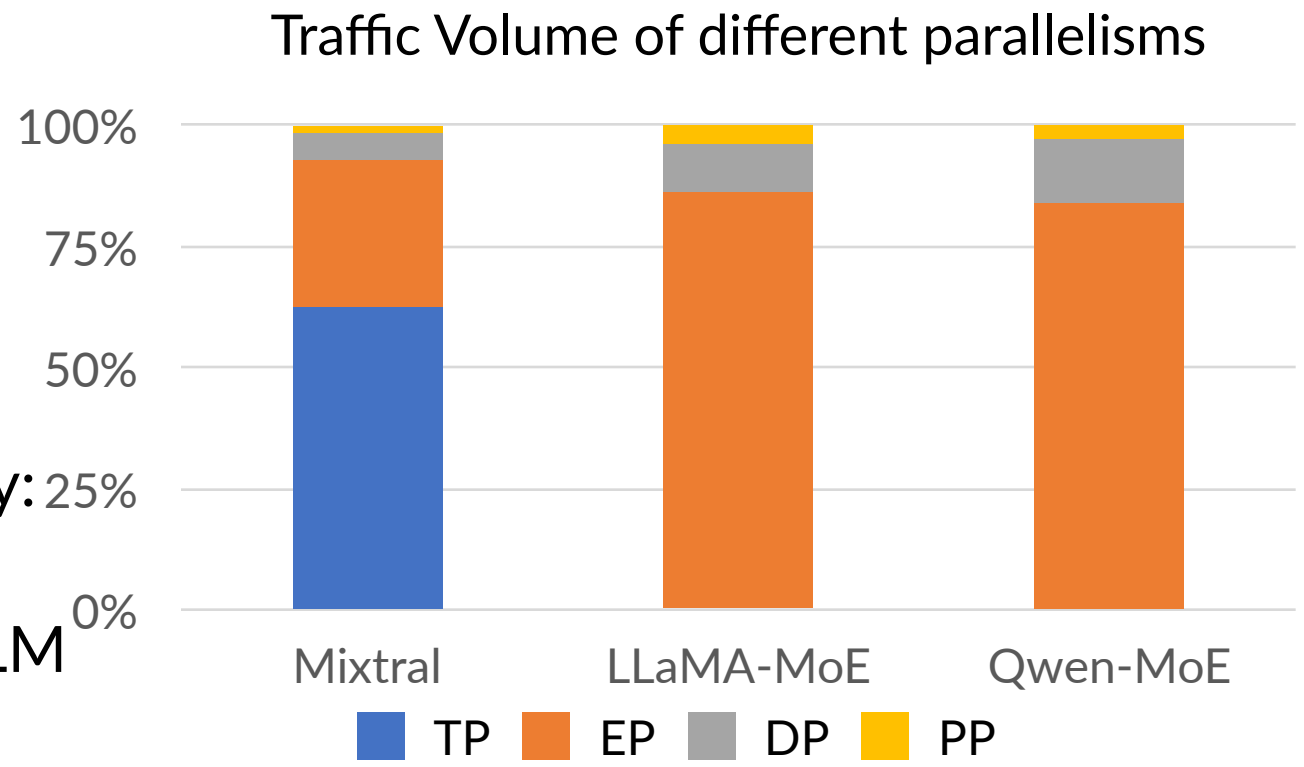
State-of-the-art MoE training utilizes the ***hybrid parallelisms***: EP + TP + PP + DP



MoE Measurements in Production Cluster

Cluster setup:

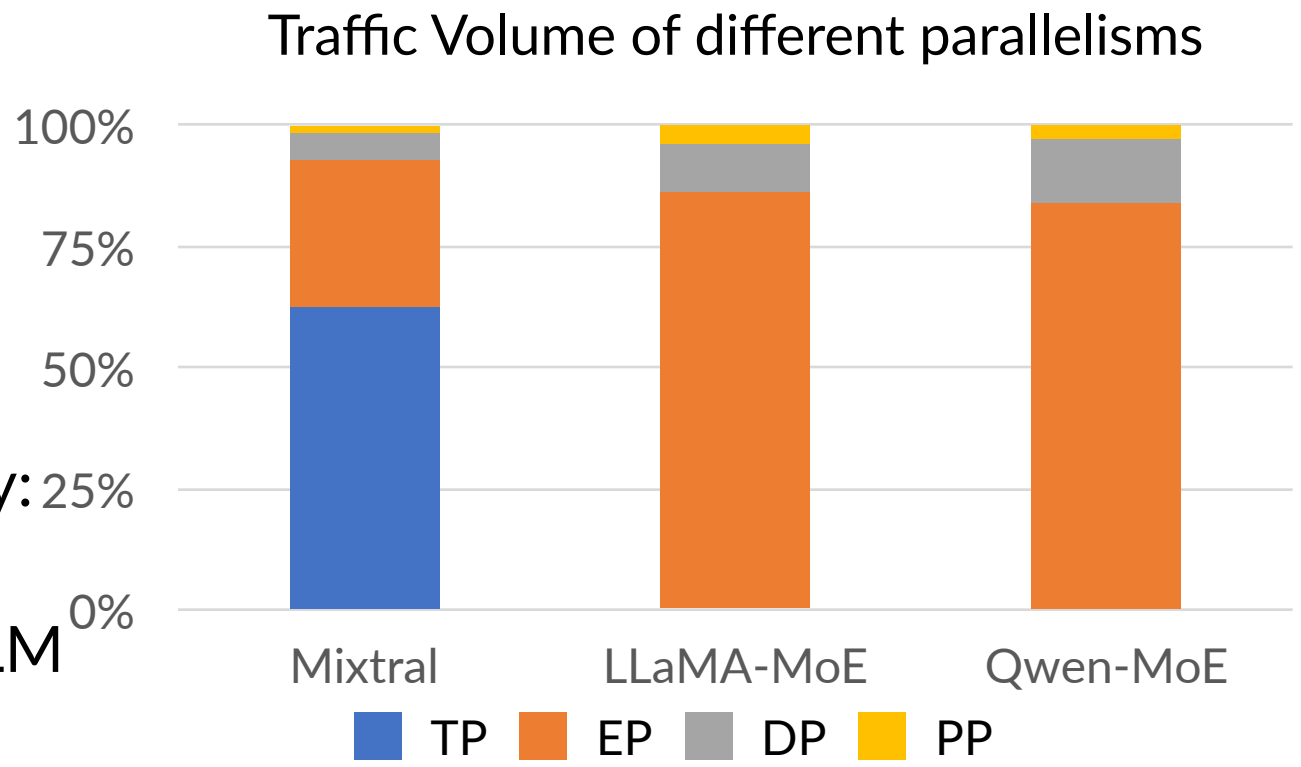
- Hardware: 128 H800 GPU and 128 400 Gbps ConnectX-7 with Infiniband network
- Topology: rail-optimized
- Collective communication library: NCCL
- Training framework: Megatron-LM



MoE Measurements in Production Cluster

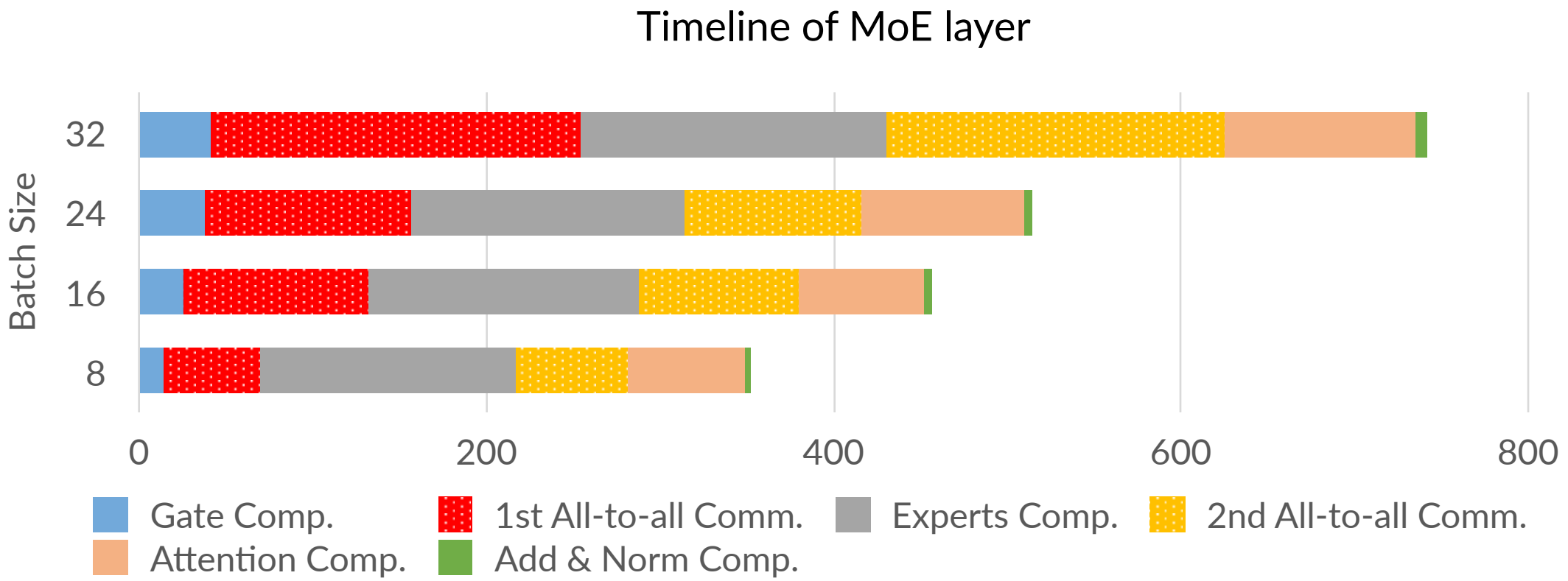
Cluster setup:

- Hardware: 128 H800 GPU and 128 400 Gbps ConnectX-7 with Infiniband network
- Topology: rail-optimized
- Collective communication library: NCCL
- Training framework: Megatron-LM



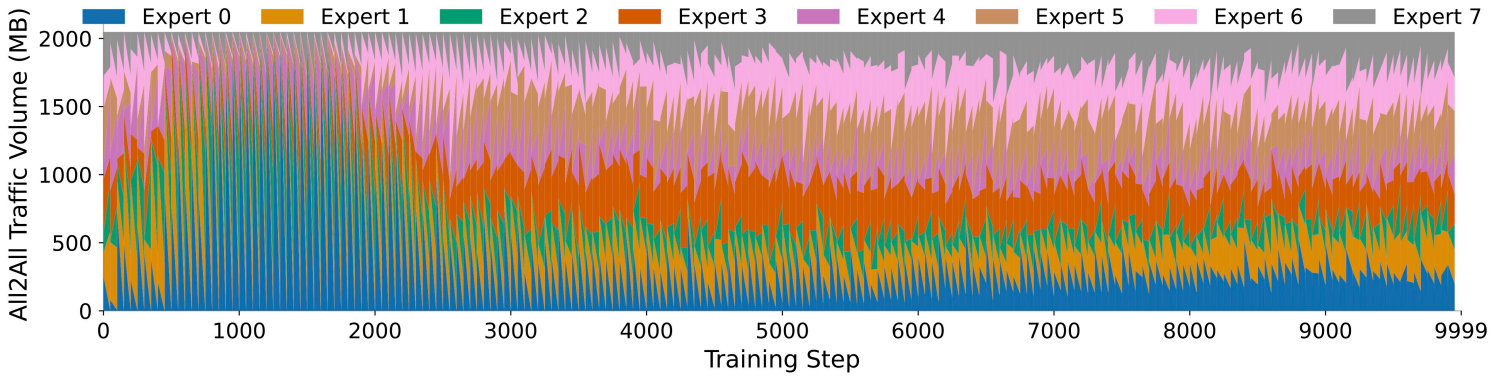
EP and TP are most ***communication-intensive.***

Communication Matters in MoE Training!

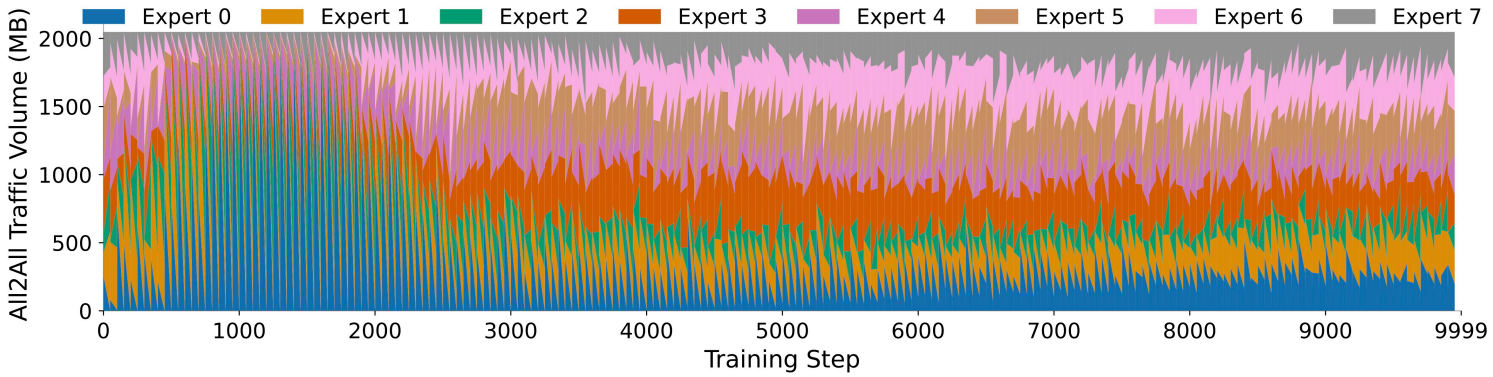


All-to-all communications account for ***35% - 55%*** training time in each MoE layer under a H800 SuperPod cluster with 400 Gbps Infiniband network.

Measurements: Temporal & Spatial Patterns

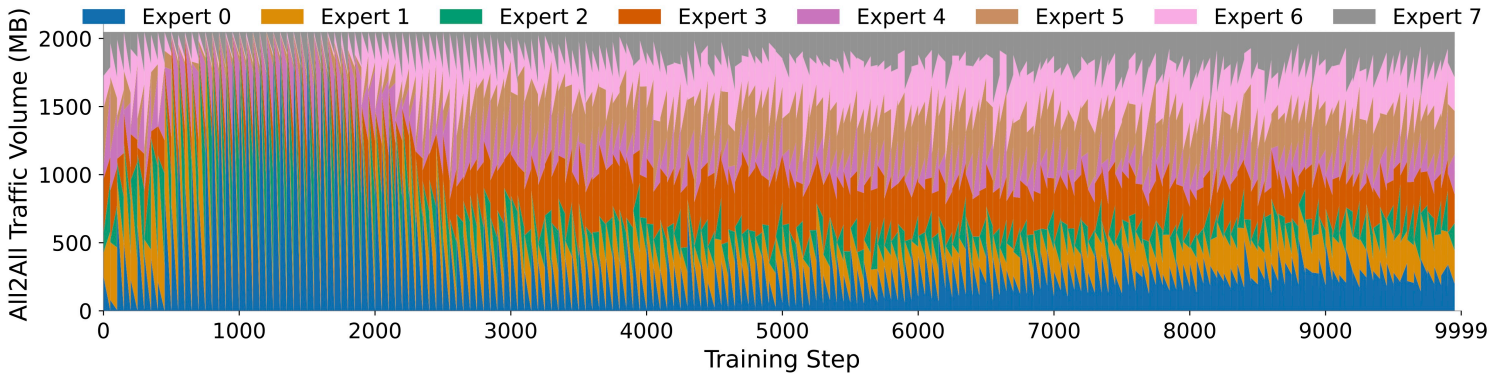


Measurements: Temporal & Spatial Patterns

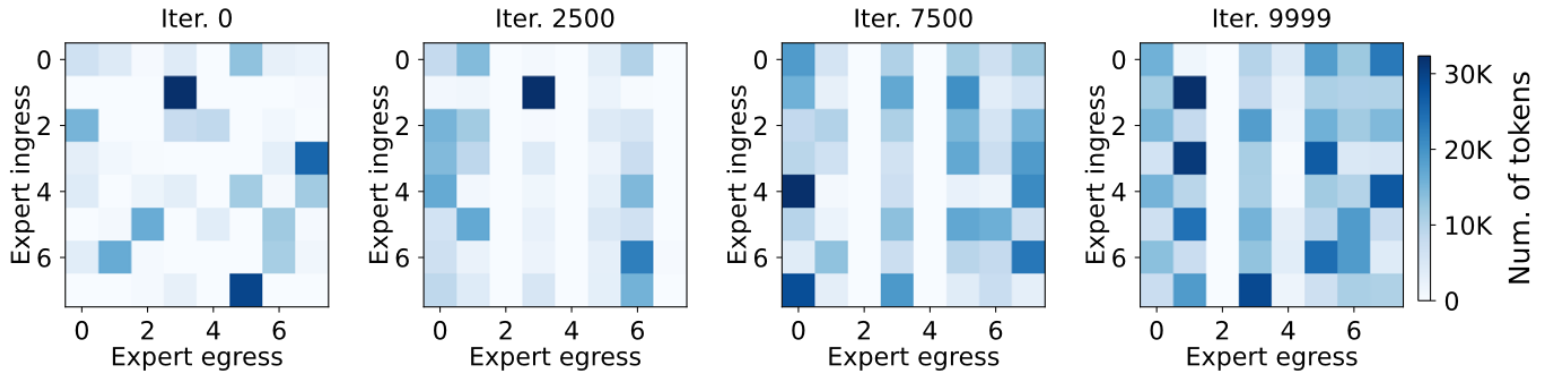


Non-deterministic: activation intensities of each expert vary significantly across different iterations.

Measurements: Temporal & Spatial Patterns

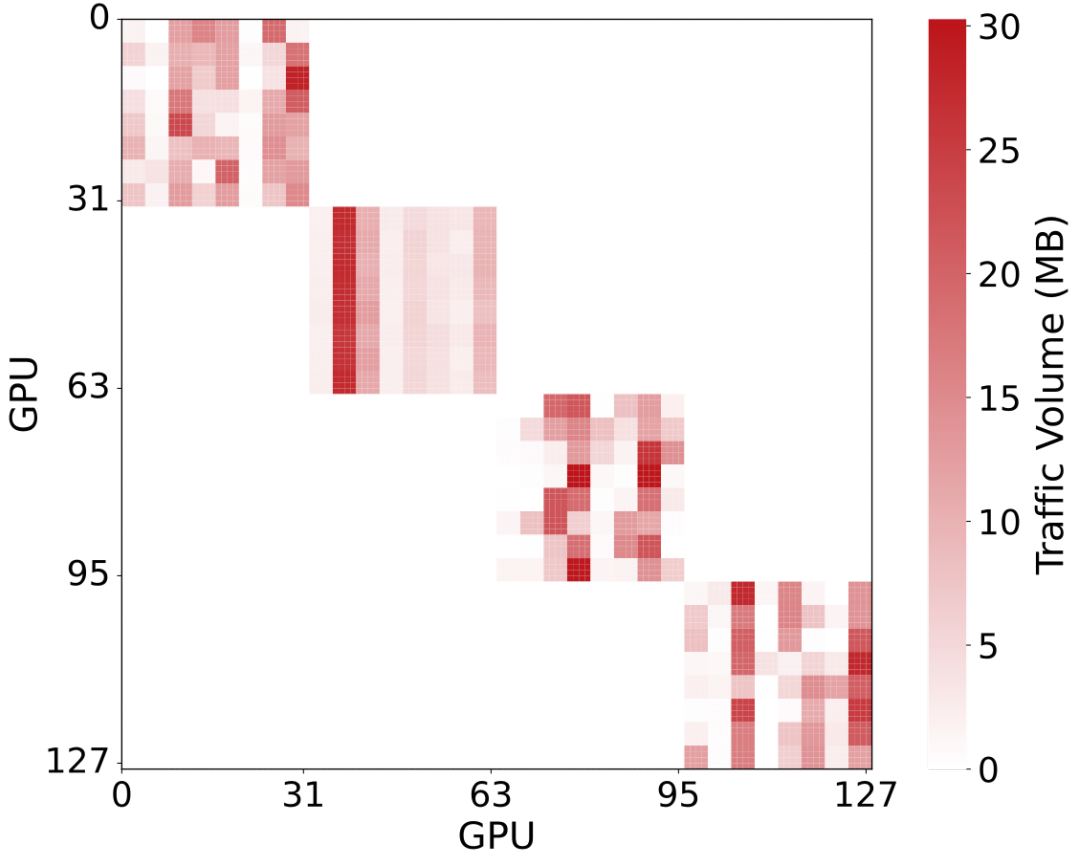


Non-deterministic: activation intensities of each expert vary significantly across different iterations.

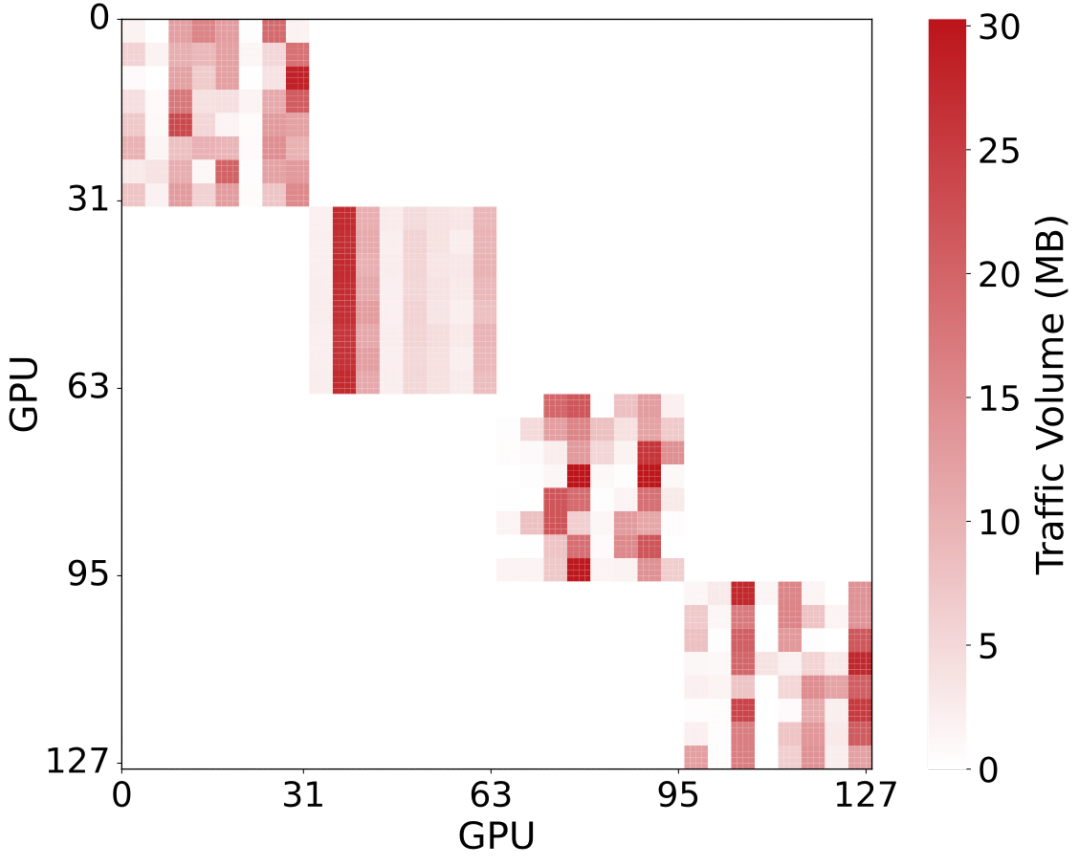


Non-uniform: heavy communications only occur between limited number of GPU pairs.

Measurements: Spatial Patterns

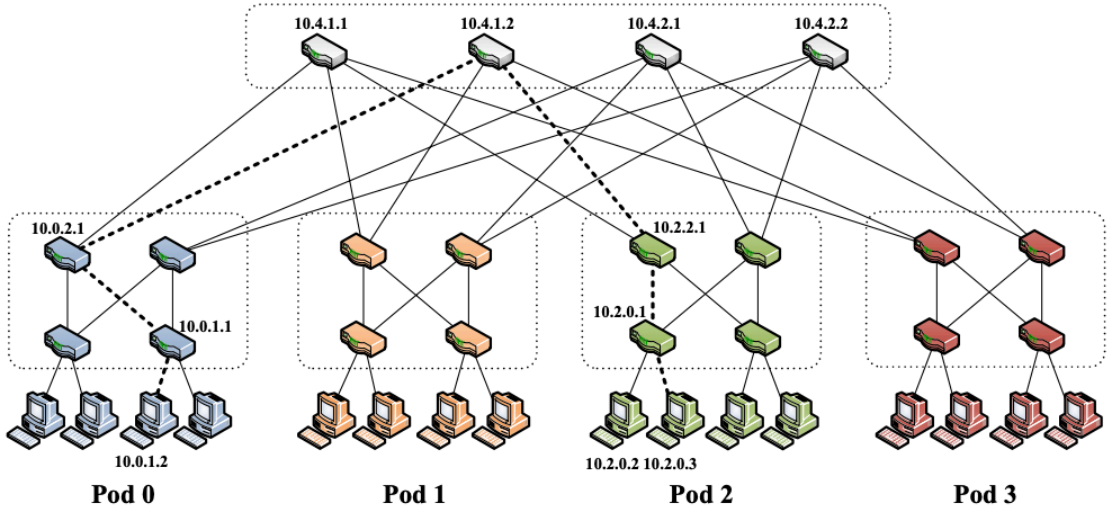


Measurements: Spatial Patterns



EP traffic has ***strong locality*** and is ***regional***: only the expert layers within the same MoE block need all-to-all communications

Motivation for a Cost-efficient Interconnect

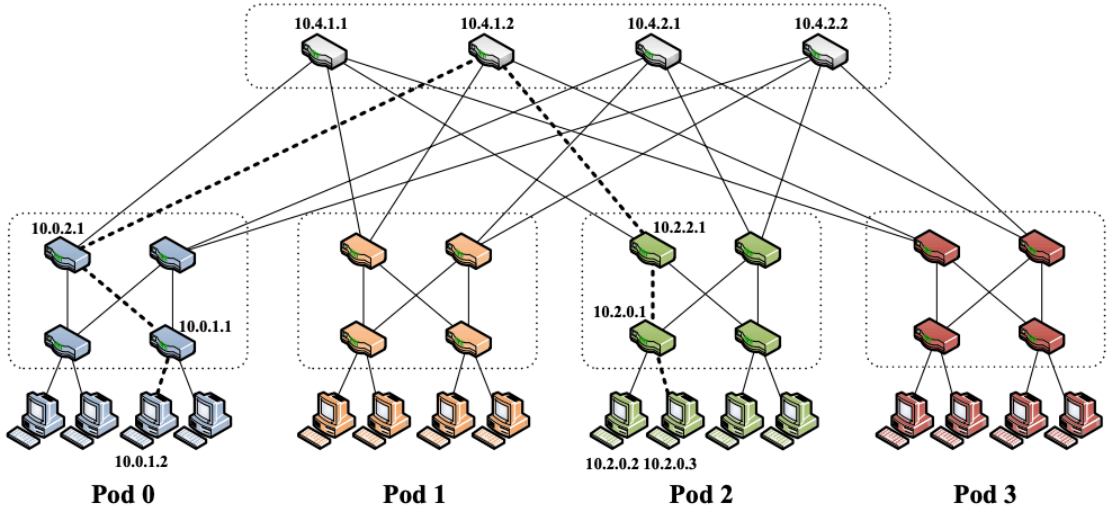


Fat-tree

[1] A scalable, commodity data center network architecture. SIGCOMM 2008
[2] TopoOpt: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs. NSDI 2023

Motivation for a Cost-efficient Interconnect

- Electrical Fat-tree [1] interconnect is an ***expensive overkill*** (20-30% networking cost);

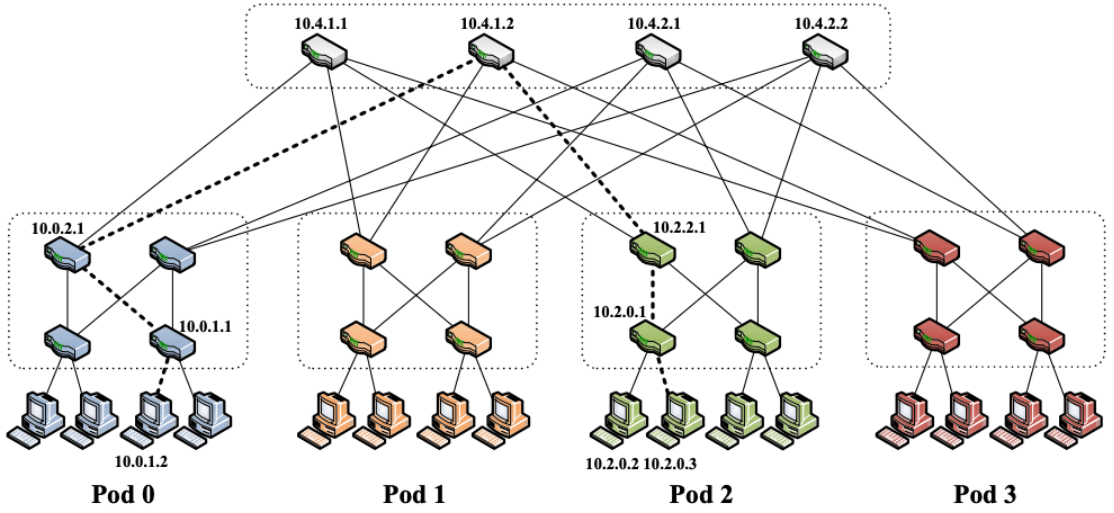


Fat-tree

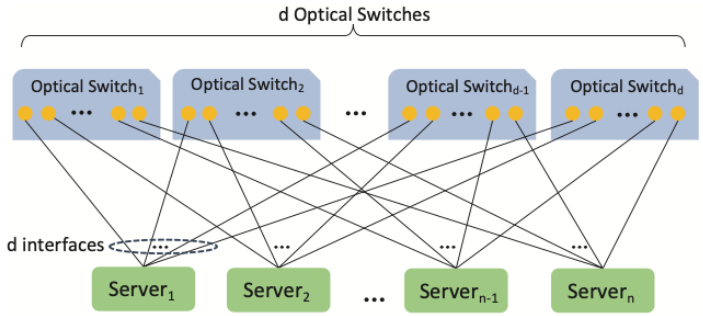
[1] A scalable, commodity data center network architecture. SIGCOMM 2008
[2] TopoOpt: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs. NSDI 2023

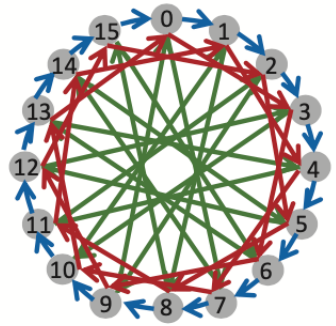
Motivation for a Cost-efficient Interconnect

- Electrical Fat-tree [1] interconnect is an ***expensive overkill*** (20-30% networking cost);
- Optical interconnect TopoOpt [2] fails to adapt to ***runtime dynamics***.



Fat-tree



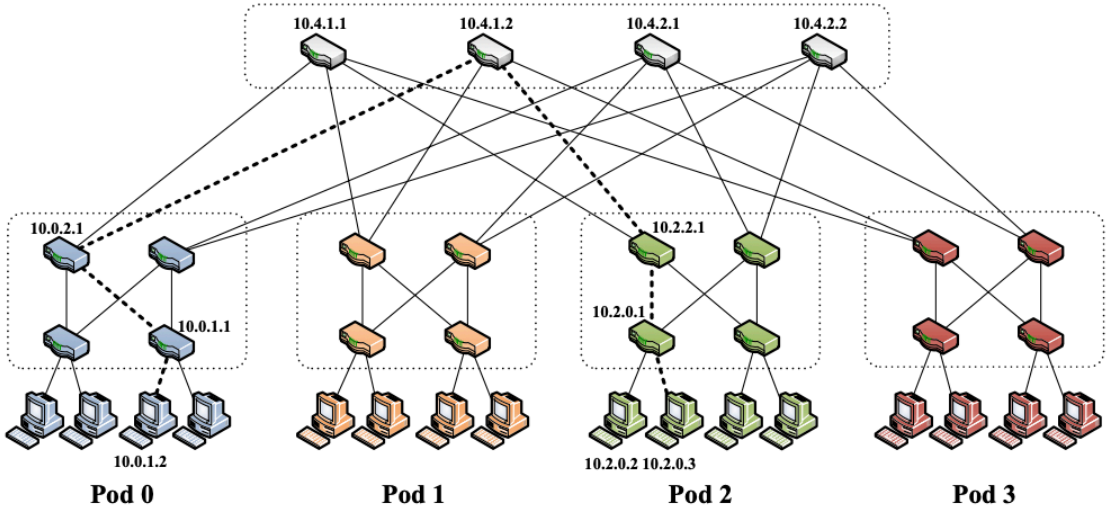


TopoOpt: model-wise one-off reconfiguration

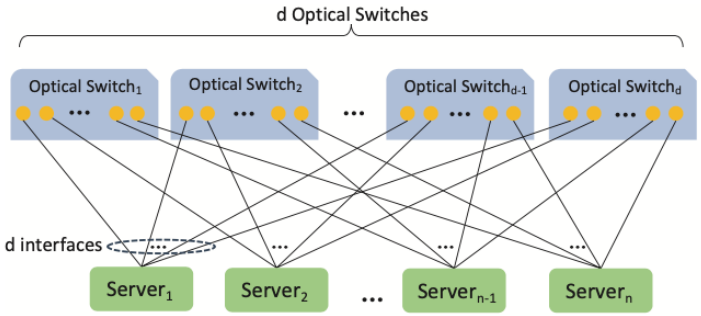
[1] A scalable, commodity data center network architecture. SIGCOMM 2008
[2] TopoOpt: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs. NSDI 2023

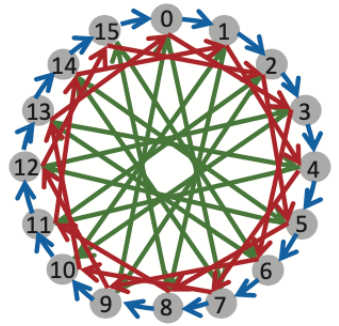
Motivation for a Cost-efficient Interconnect

- Electrical Fat-tree [1] interconnect is an ***expensive overkill*** (20-30% networking cost);
- Optical interconnect TopoOpt [2] fails to adapt to ***runtime dynamics***.



Fat-tree





TopoOpt: model-wise one-off reconfiguration

How to ***architect*** GPU interconnects for large-scale MoE training?

[1] A scalable, commodity data center network architecture. SIGCOMM 2008
[2] TopoOpt: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs. NSDI 2023

First-principle Analysis for the Interconnect

Quest for best fit between interconnect fabric and MoE training strategies:

First-principle Analysis for the Interconnect

Quest for best fit between interconnect fabric and MoE training strategies:

	Bandwidth	Predictability	Locality	Reconfigurability	Desired Fabric
TP	Highest	Deterministic	Local & All-reduce	N	Crossbar Switch (NVSwitch)

First-principle Analysis for the Interconnect

Quest for best fit between interconnect fabric and MoE training strategies:

	Bandwidth	Predictability	Locality	Reconfigurability	Desired Fabric
TP	Highest	Deterministic	Local & All-reduce	N	Crossbar Switch (NVSwitch)
EP	High	Non-deterministic	Regional & All-to-all	Y	Circuit switching (Optical)

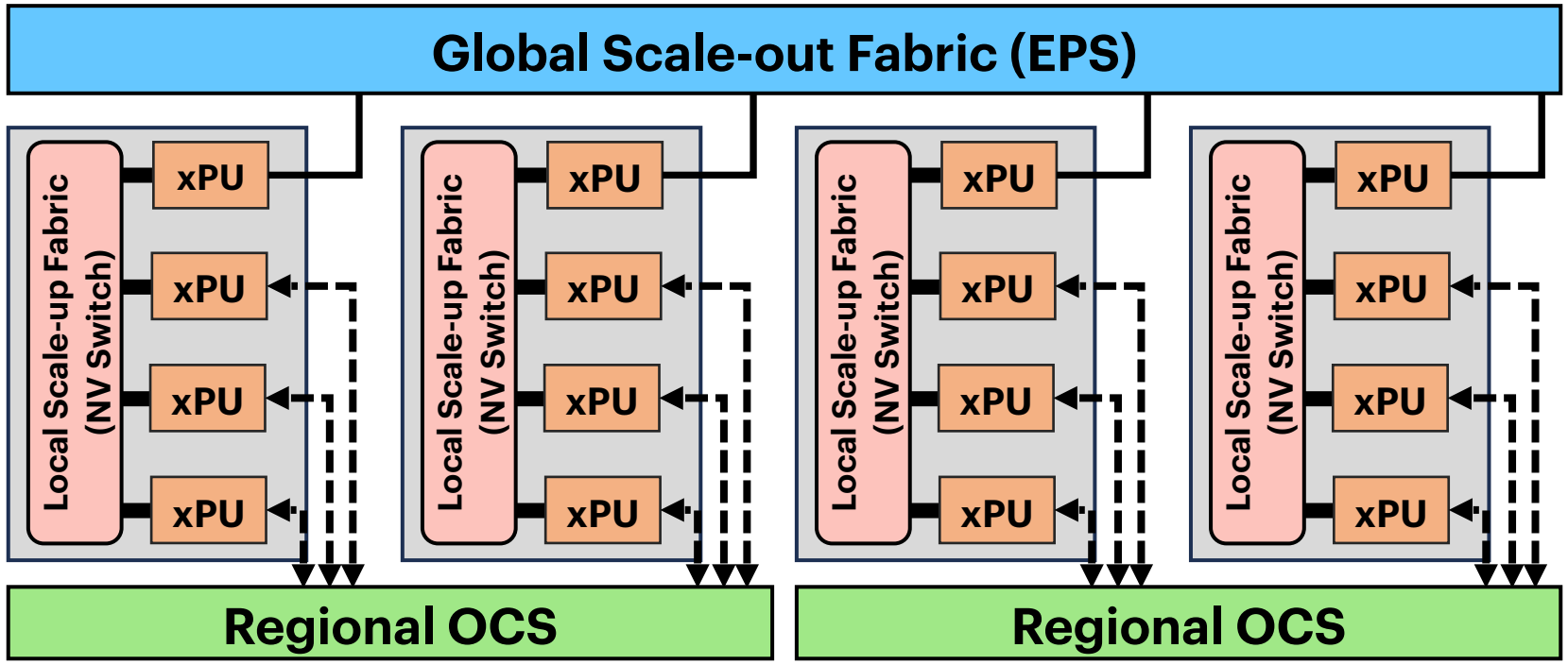
First-principle Analysis for the Interconnect

Quest for best fit between interconnect fabric and MoE training strategies:

	Bandwidth	Predictability	Locality	Reconfigurability	Desired Fabric
TP	Highest	Deterministic	Local & All-reduce	N	Crossbar Switch (NVSwitch)
EP	High	Non-deterministic	Regional & All-to-all	Y	Circuit switching (Optical)
DP	Low	Deterministic	Large & All-reduce	N	Electrical Packet Switching (Ethernet)
PP	Low	Deterministic	Large & Point-to-point	N	Electrical Packet Switching (Ethernet)

MixNet Architecture: An Ideal yet Practical Fabric

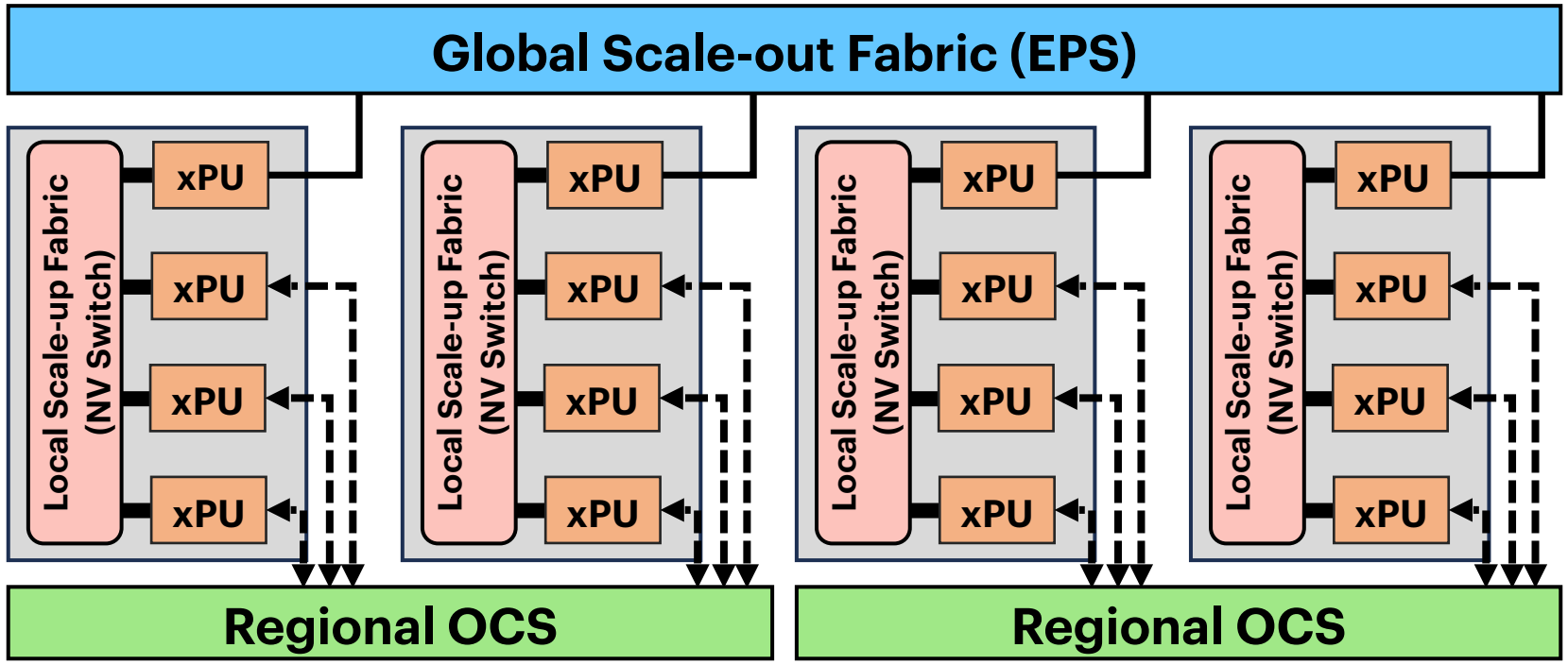
Design philosophy: Find an ***ideal*** fabric that balances the bandwidth and scale for on-demand networking.



MixNet Architecture: An Ideal yet Practical Fabric

Design philosophy: Find an ***ideal*** fabric that balances the bandwidth and scale for on-demand networking.

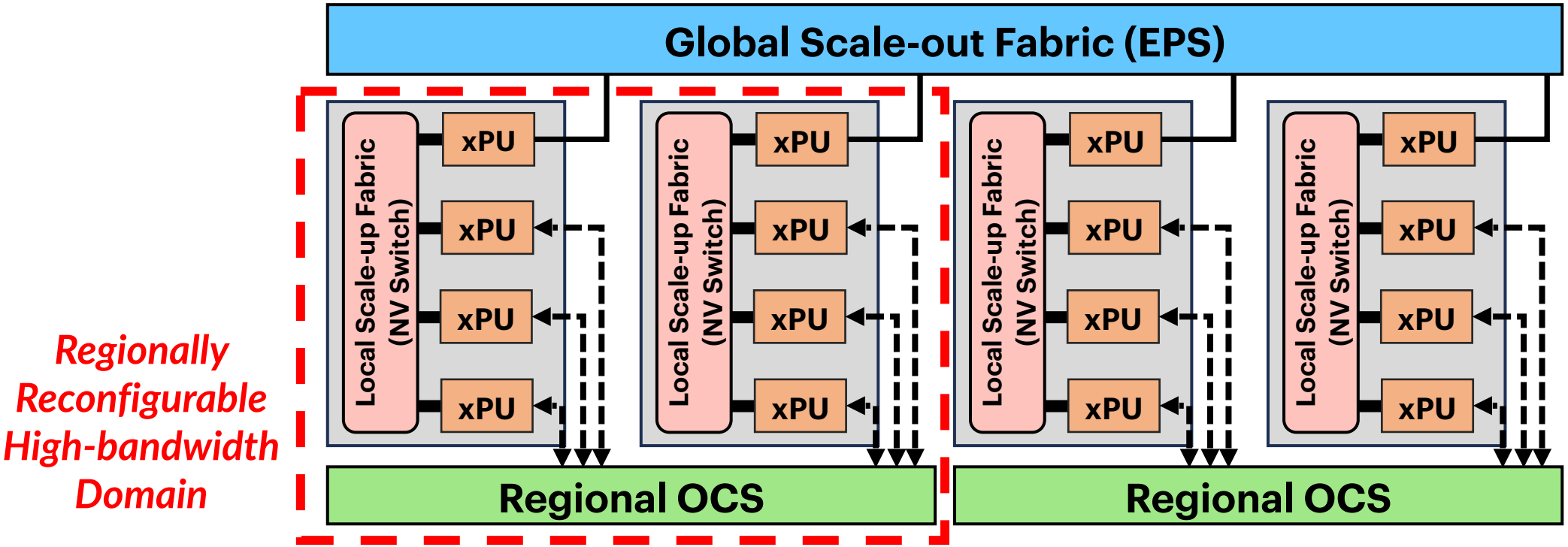
1. Expand the scale-up domain (NV Switch) efficiently with Optics.
2. Leverage the Electrical Ethernet to preserve the networking scale.



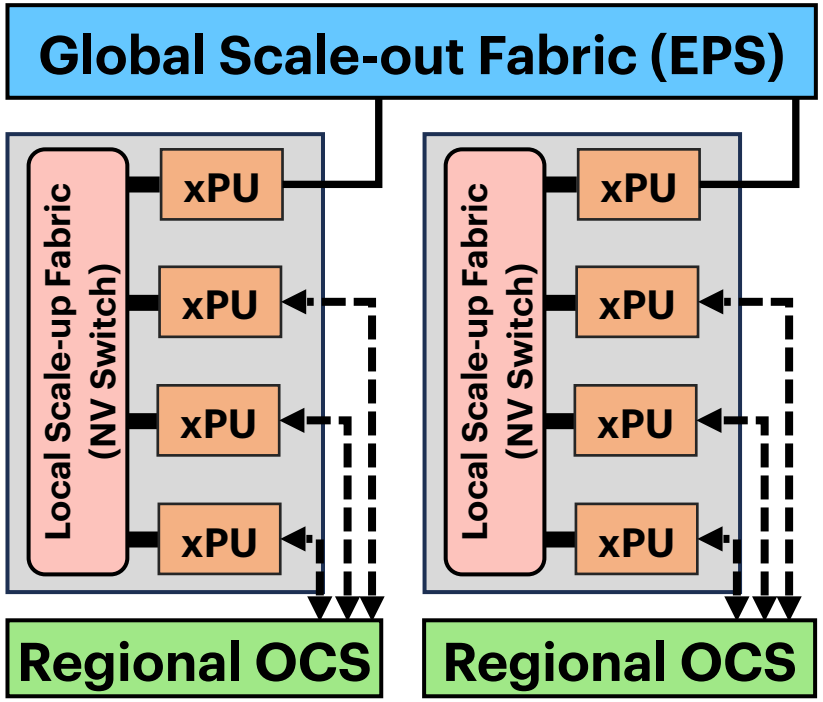
MixNet Architecture: An Ideal yet Practical Fabric

Design philosophy: Find an ***ideal*** fabric that balances the bandwidth and scale for on-demand networking.

1. Expand the scale-up domain (NV Switch) efficiently with Optics.
2. Leverage the Electrical Ethernet to preserve the networking scale.

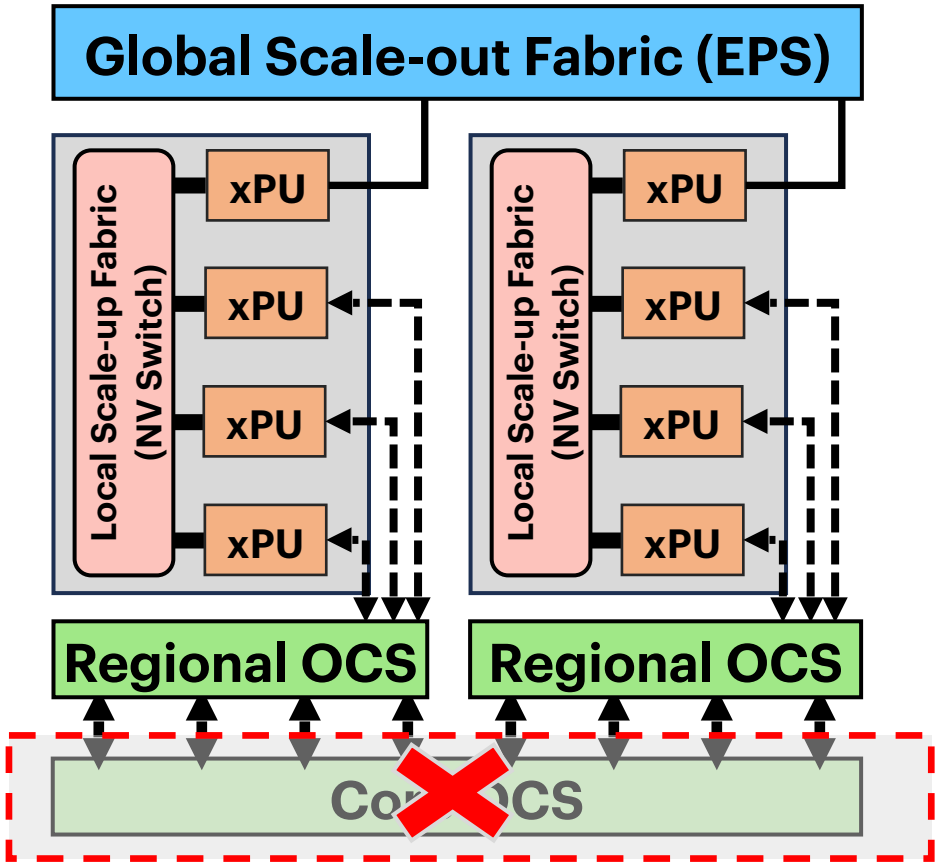


A Runtime Reconfigurable OCS-EPS Fabric



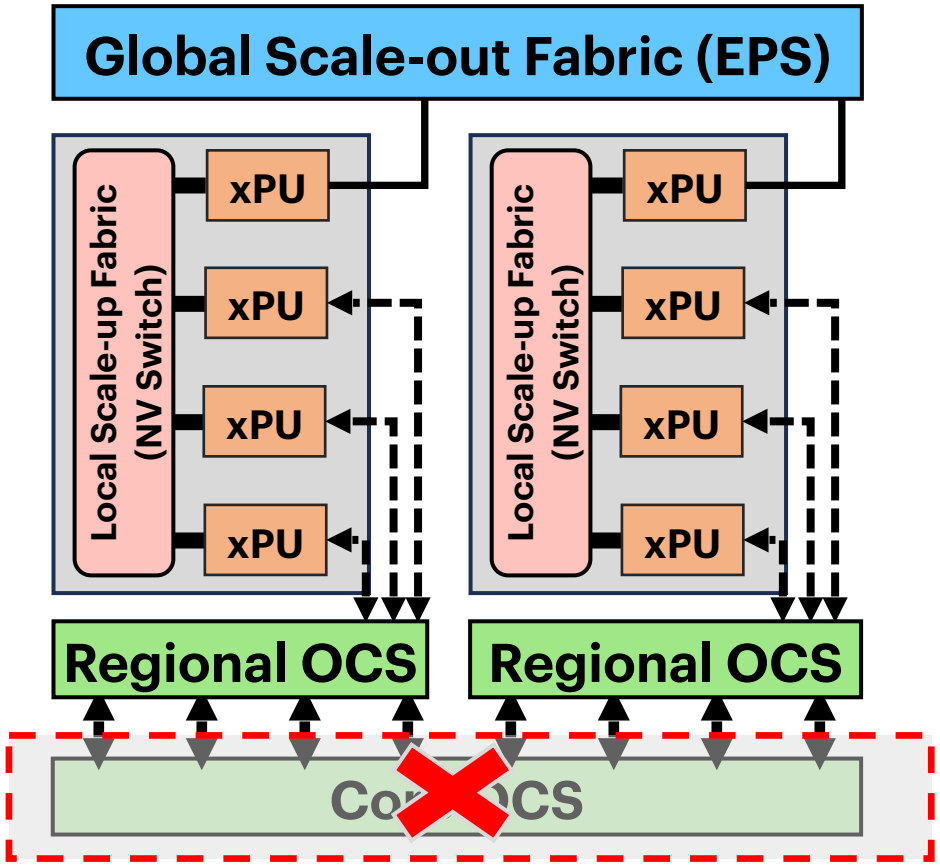
- NVSwitch handles the *intra-host TP* traffic
- Regional OCS handles *regionalized EP* traffic
  - MixNet reconfigures its topology based on runtime
- Global EPS handle both high-fanout *DP & PP* traffics

A Runtime Reconfigurable OCS-EPS Fabric



- NVSwitch handles the *intra-host TP* traffic
- Regional OCS handles *regionalized EP* traffic
  - MixNet reconfigures its topology based on runtime
- Global EPS handle both high-fanout *DP & PP* traffics

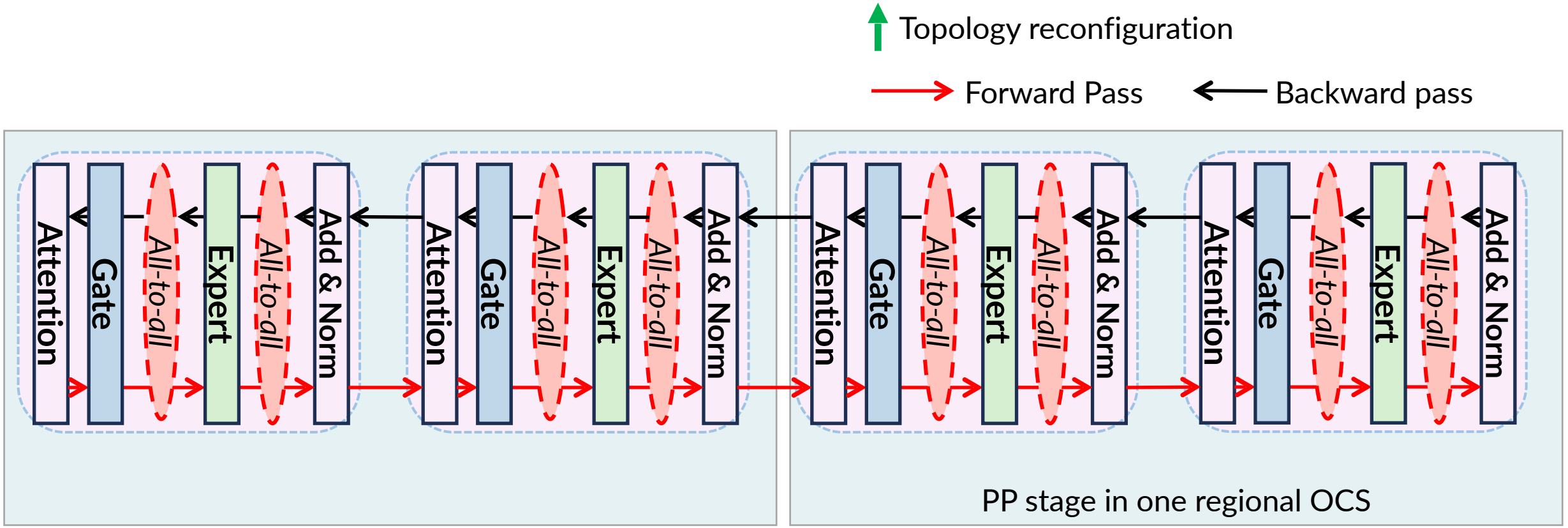
A Runtime Reconfigurable OCS-EPS Fabric



- NVSwitch handles the *intra-host TP* traffic
- Regional OCS handles *regionalized EP* traffic
  - MixNet reconfigures its topology based on runtime
- Global EPS handle both high-fanout *DP & PP* traffics

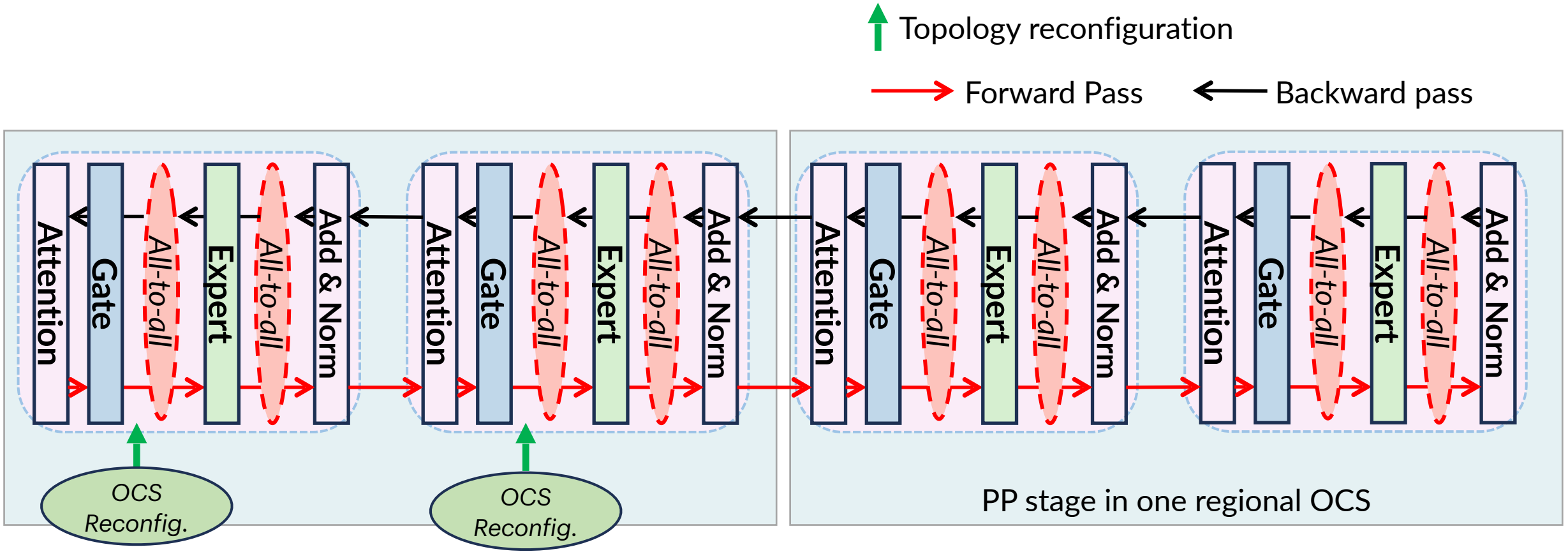
MixNet mitigates the tradeoff between ***OCS reconfigurable speed*** and ***port counts*** by designing the regional OCS.

MixNet Reconfiguration Timeline



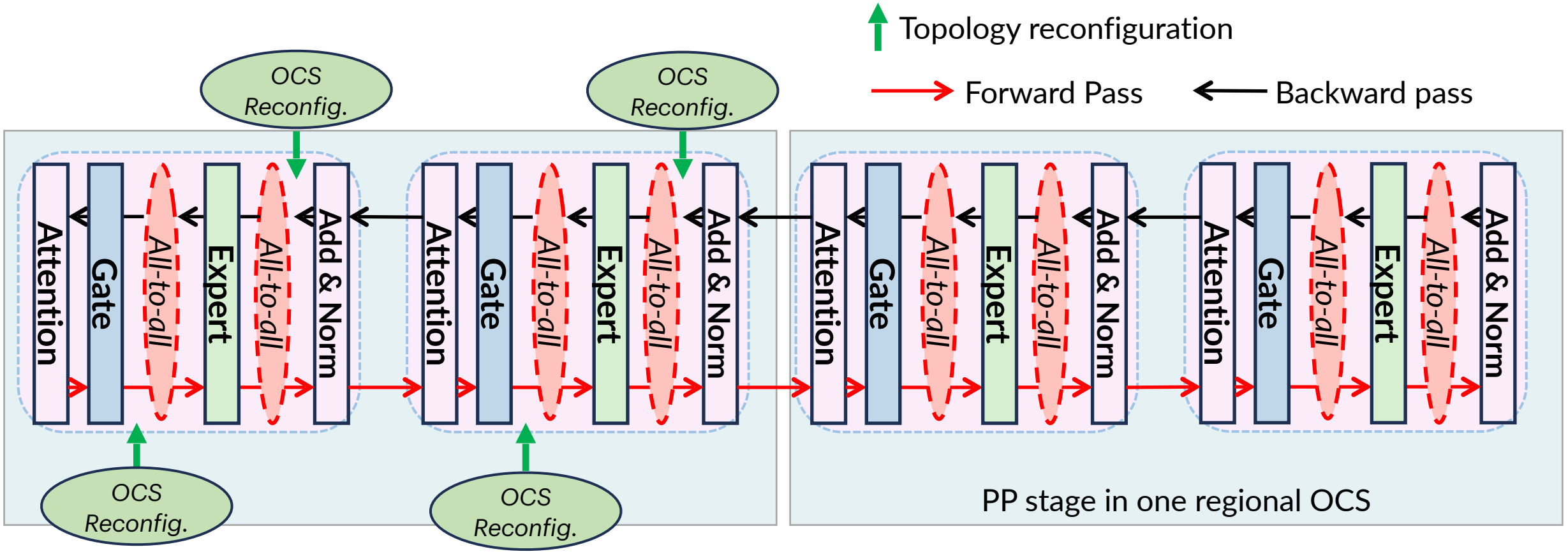
- Communication demand from `torch.dist.all_to_all_single()`
- MixNet hides the topology reconfiguration latency in the computation phase

MixNet Reconfiguration Timeline



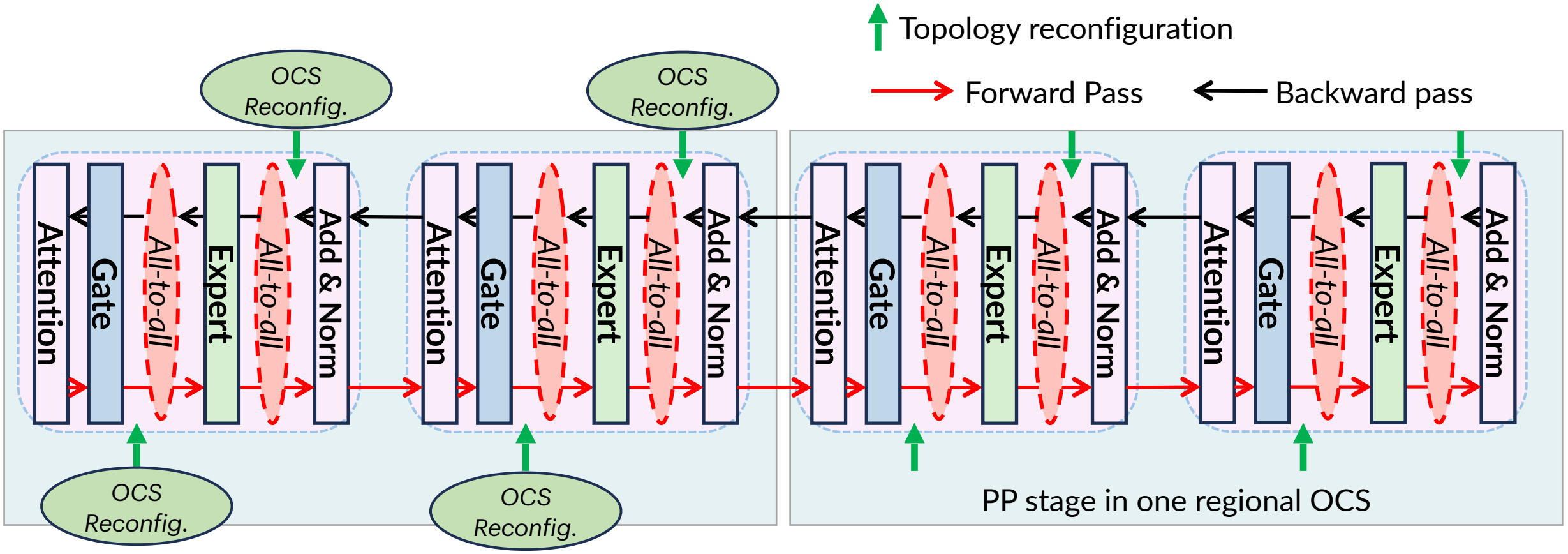
- Communication demand from `torch.dist.all_to_all_single()`
- MixNet hides the topology reconfiguration latency in the computation phase

MixNet Reconfiguration Timeline



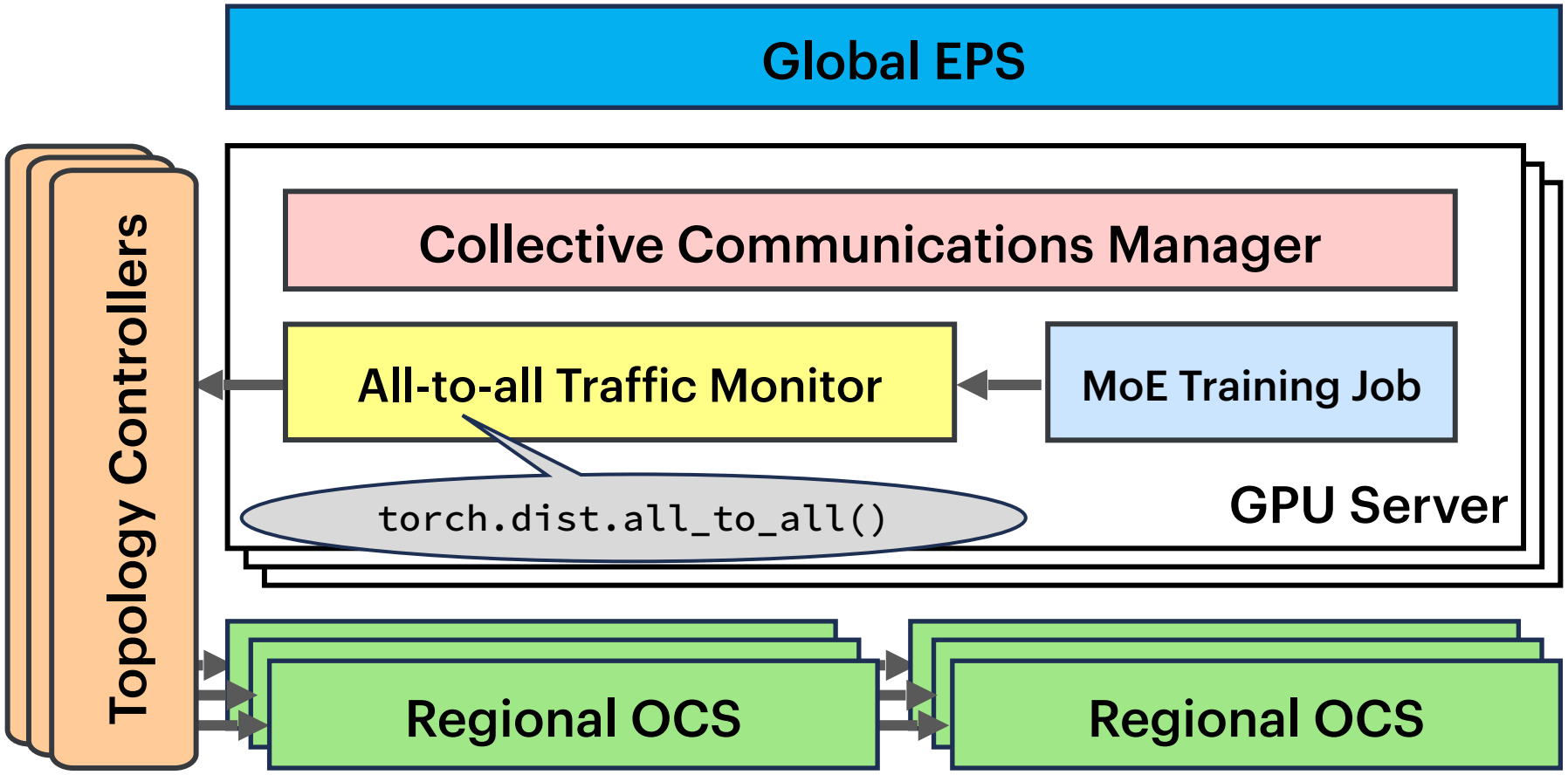
- Communication demand from `torch.dist.all_to_all_single()`
- MixNet hides the topology reconfiguration latency in the computation phase

MixNet Reconfiguration Timeline

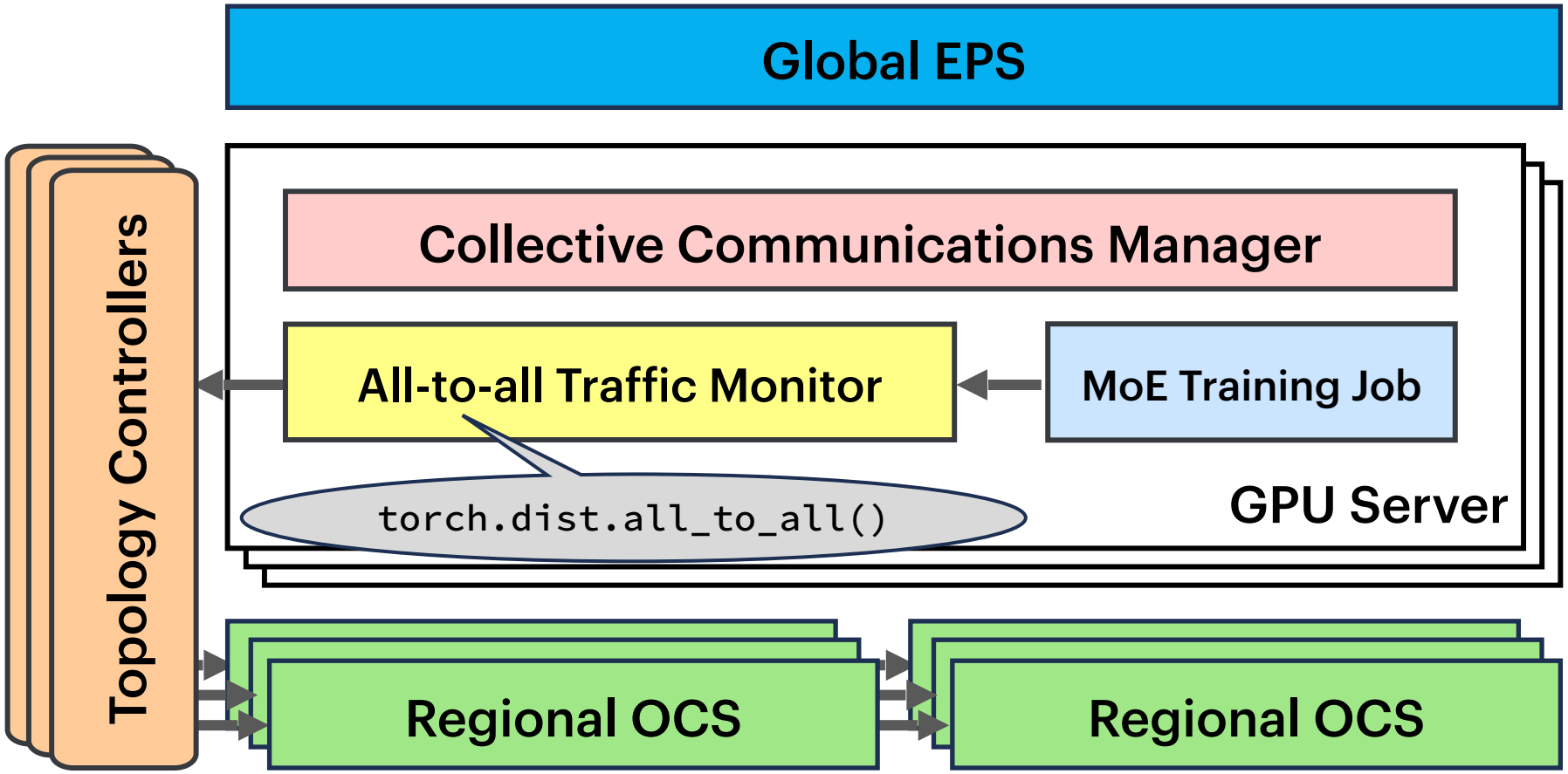


- Communication demand from `torch.dist.all_to_all_single()`
- MixNet hides the topology reconfiguration latency in the computation phase

MixNet System Design

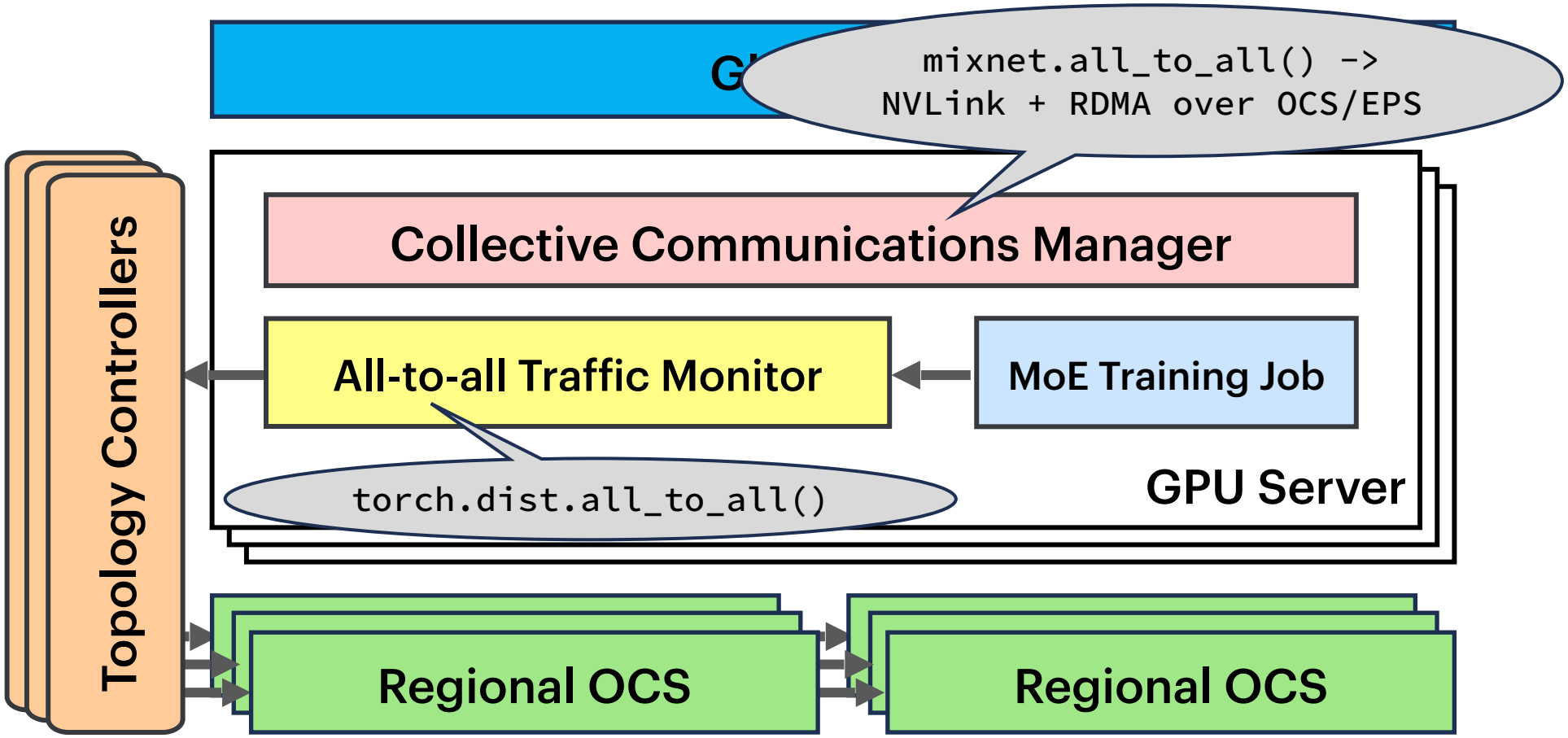


MixNet System Design



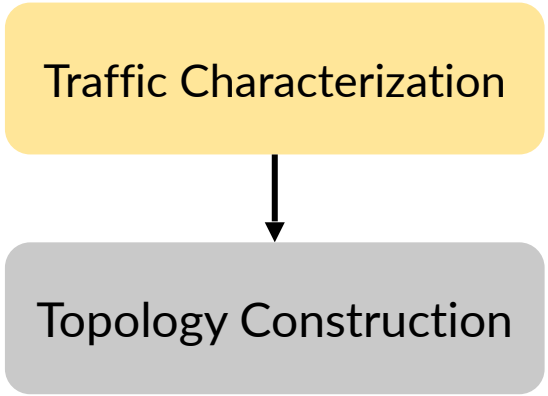
Fully localized control plane avoids the scalability limitation.

MixNet System Design



Fully localized control plane avoids the scalability limitation.

Collective Communication Schedule



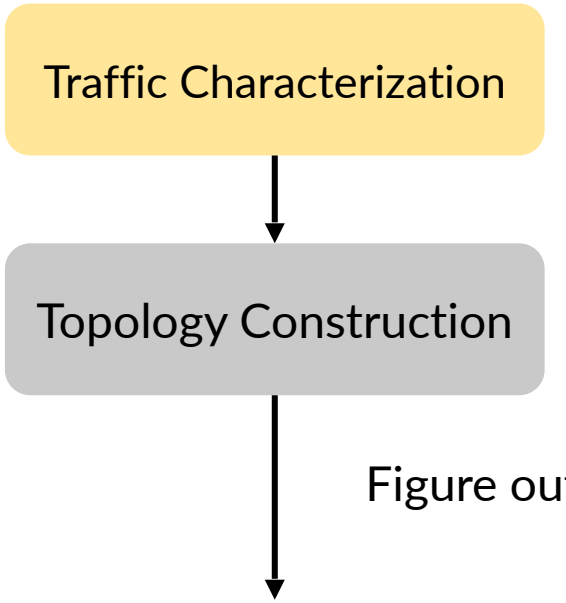
Infer runtime comm. demand from application-level collective comm. primitives (`mixnet.all_to_all`)

Compute topology and reconfigure OCS to enforce it

Inter-host

Intra-host

Collective Communication Schedule



Infer runtime comm. demand from application-level collective comm. primitives (`mixnet.all_to_all`)

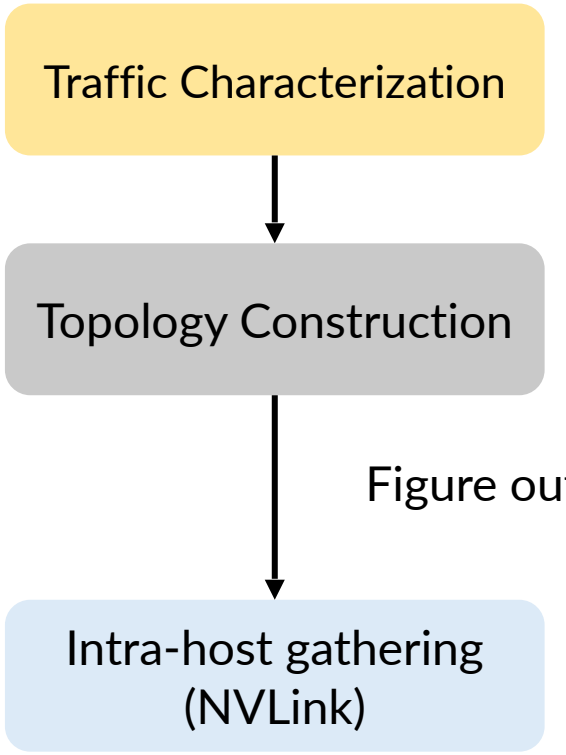
Compute topology and reconfigure OCS to enforce it

Figure out gateway GPUs and routing for inter-host connectivity

Inter-host

Intra-host

Collective Communication Schedule



Infer runtime comm. demand from application-level collective comm. primitives (`mixnet.all_to_all`)

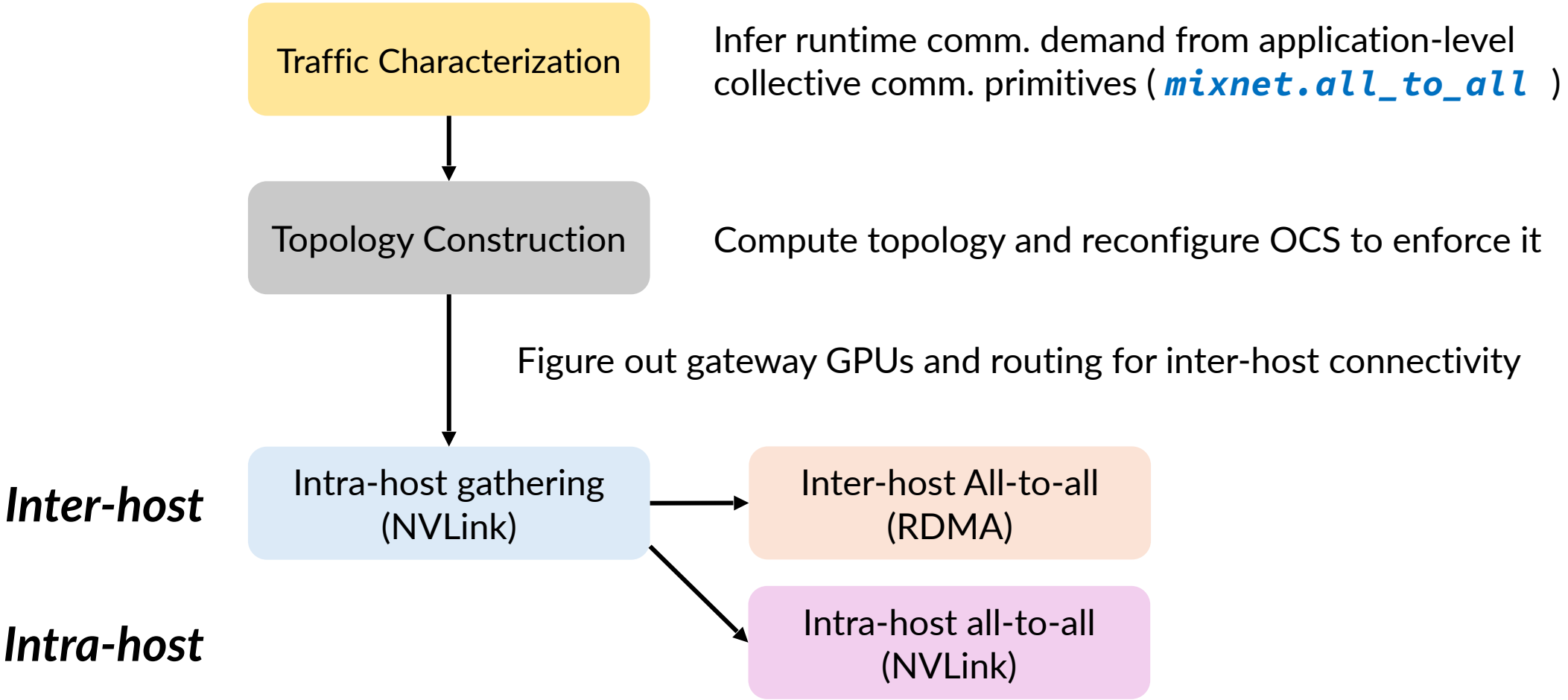
Compute topology and reconfigure OCS to enforce it

Figure out gateway GPUs and routing for inter-host connectivity

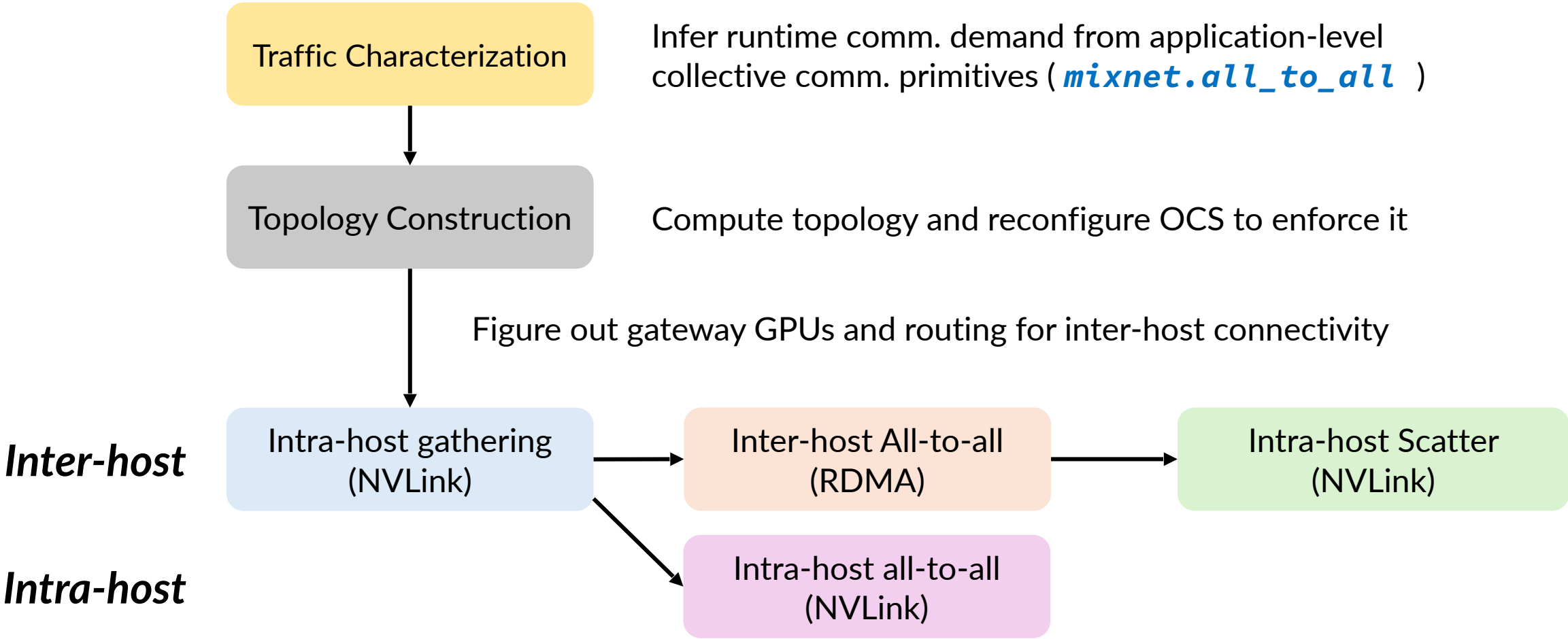
Inter-host

Intra-host

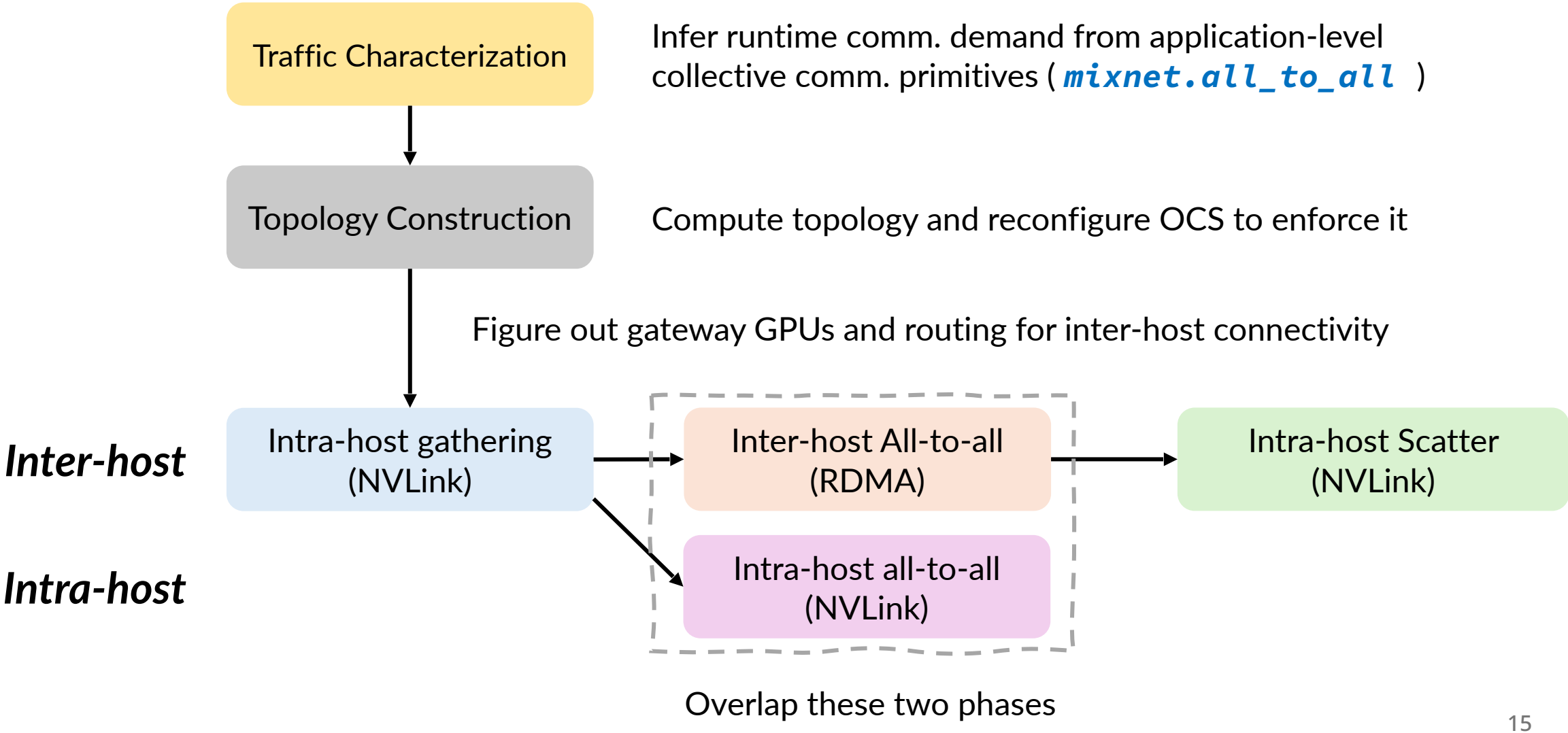
Collective Communication Schedule



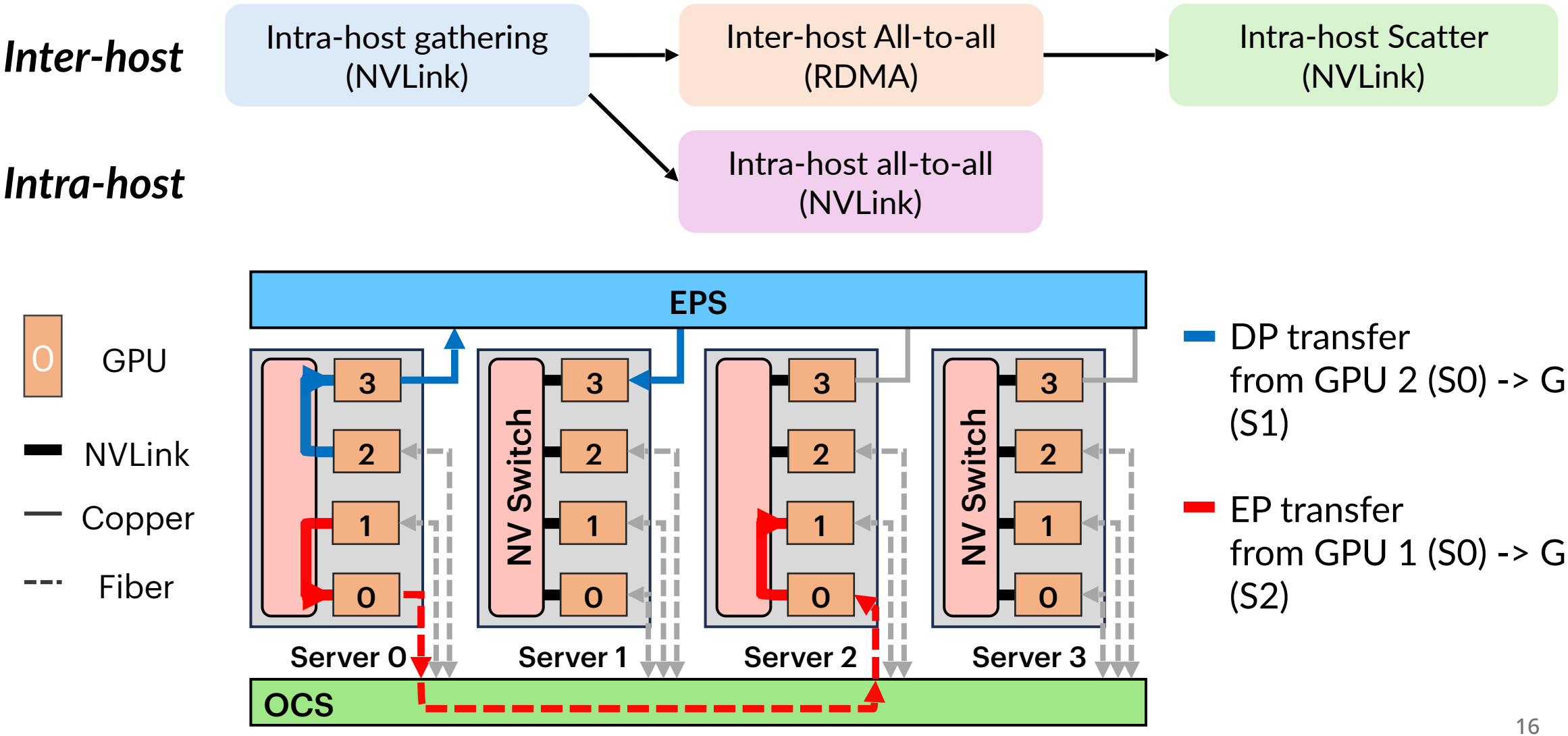
Collective Communication Schedule



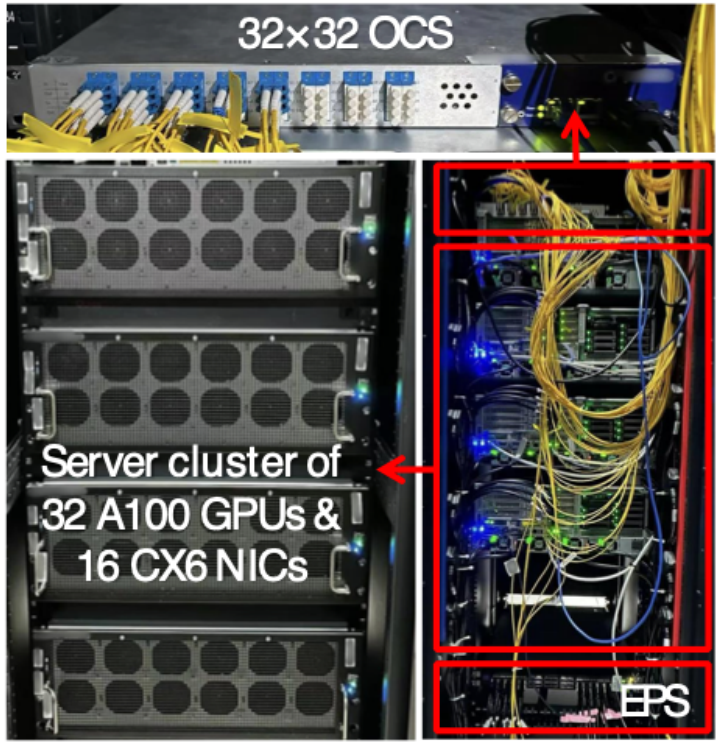
Collective Communication Schedule

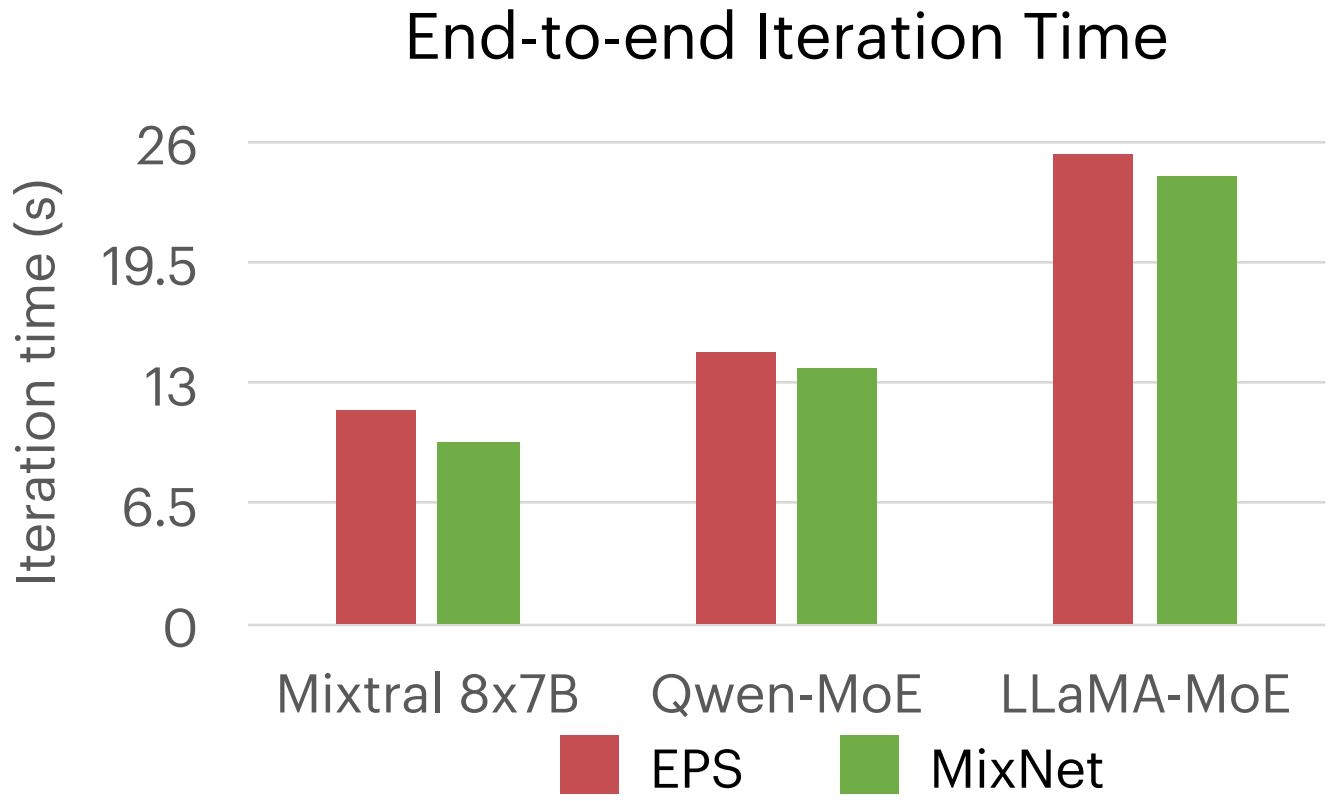


Collective Communication Schedule



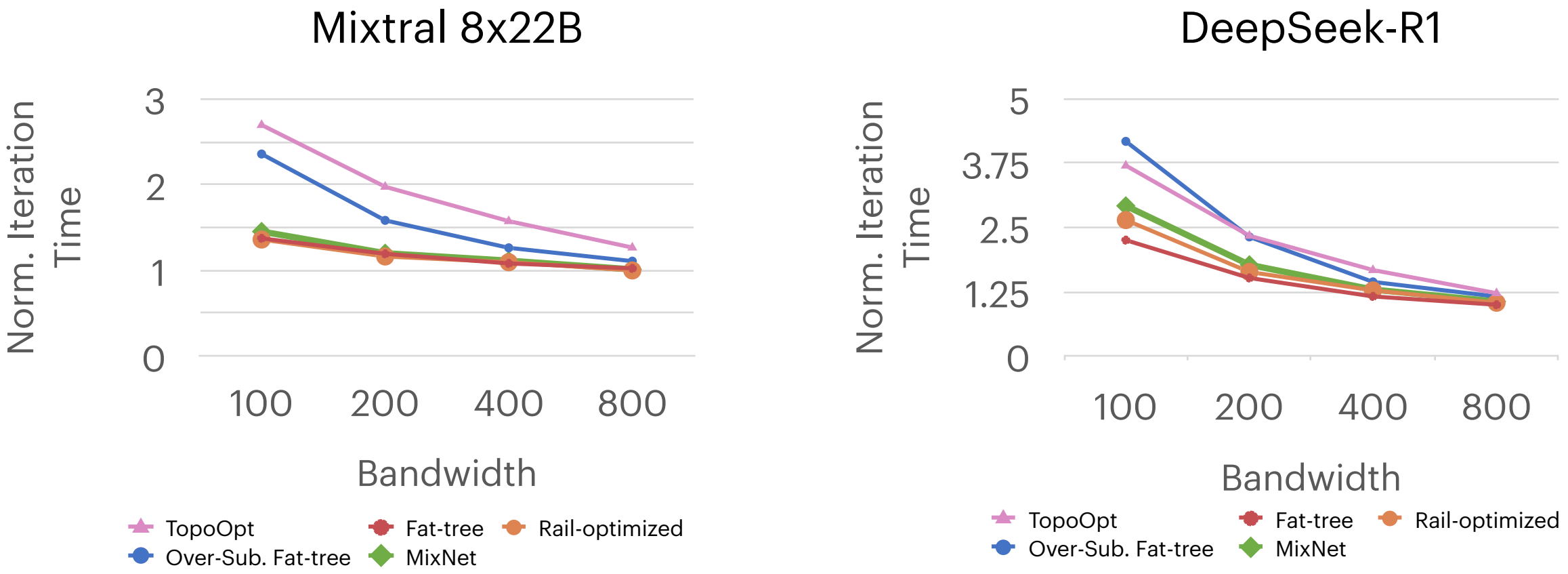
MixNet Prototype





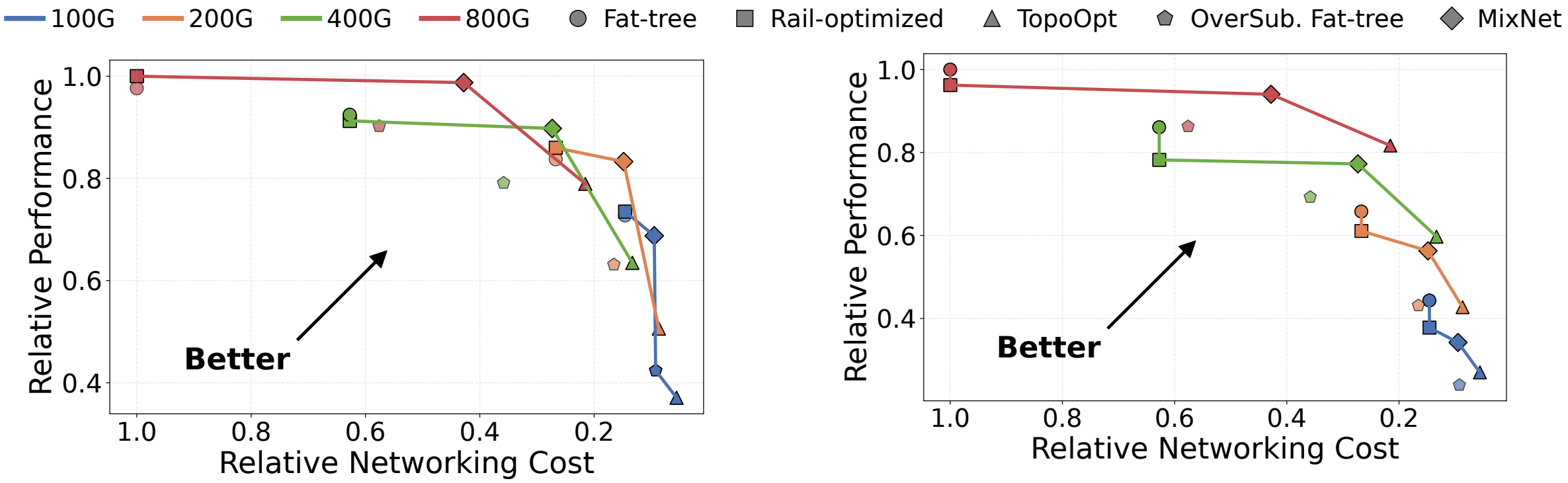
Feasibility: MixNet is able to train state-of-the-art MoE models and achieves comparable performance with bandwidth-equivalent EPS.

Large-scale Simulation



MixNet is comparable with 1:1 Fat-tree and outperforms TopoOpt by 2.5x.

Pareto Analysis on Performance vs. Cost

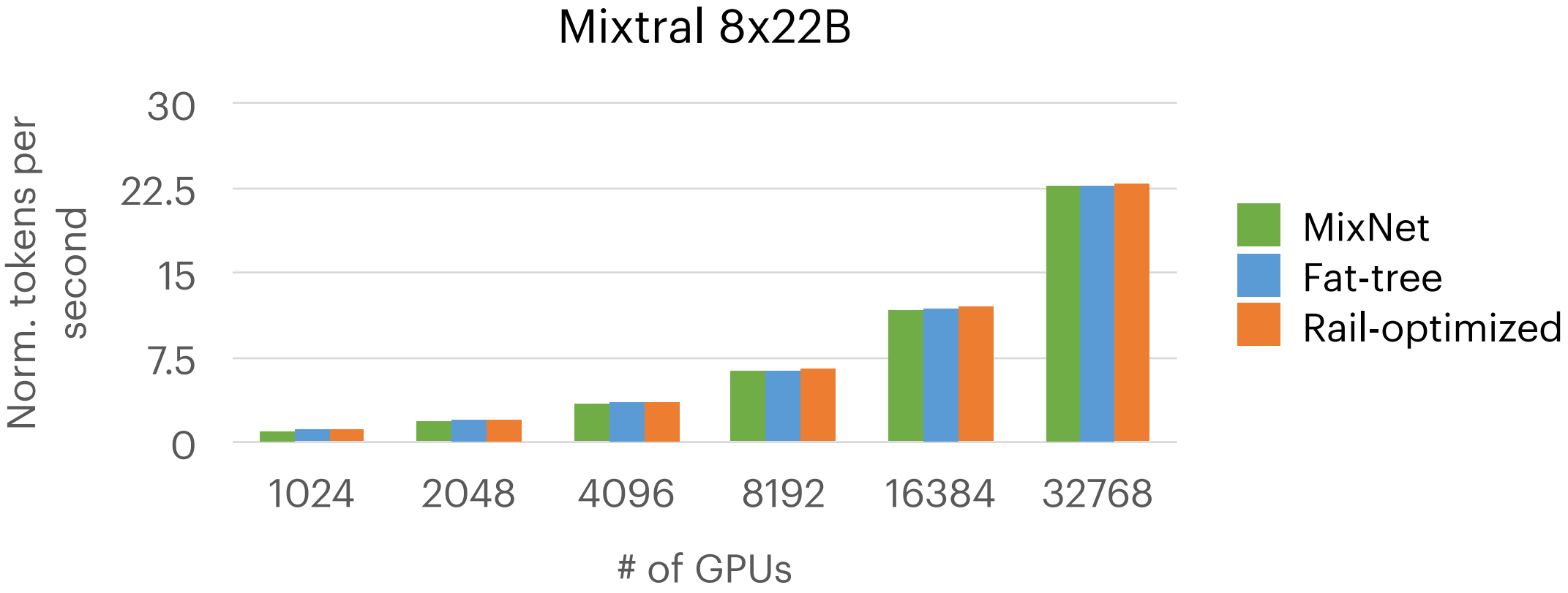


Mixtral 8*22B

DeepSeek-R1

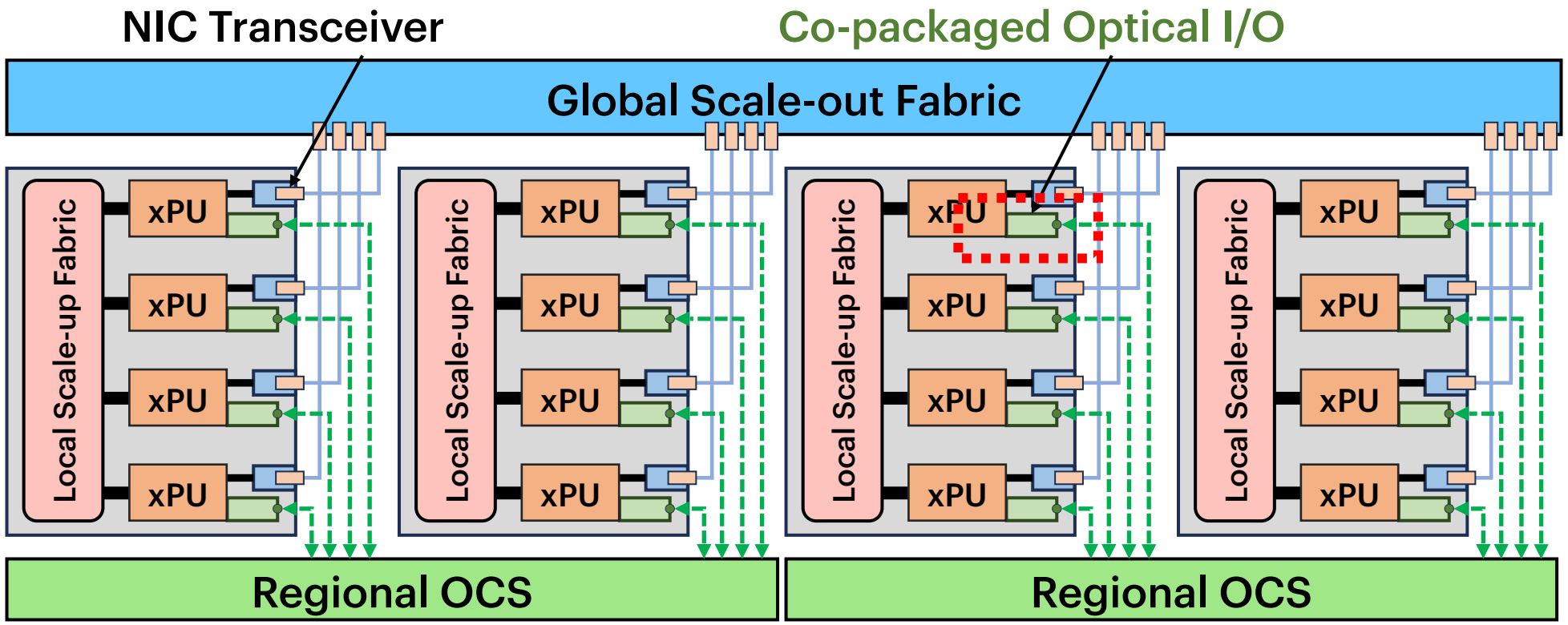
MixNet defines the ***Pareto Front*** among all interconnects and boosts the training cost efficiency by up to 2.3x at 400 Gbps network.

Scalability

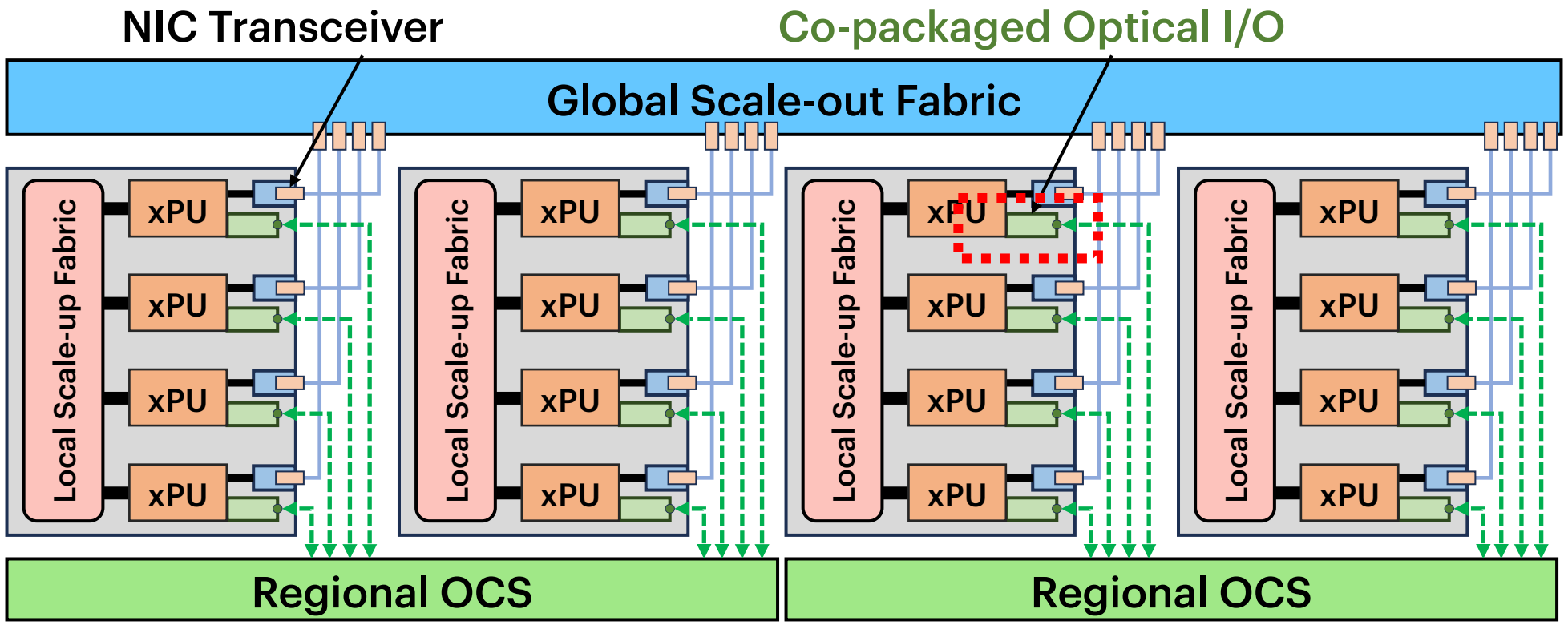


MixNet ***scales efficiently*** with an increasing number of GPUs, demonstrating strong scalability.

Look Forward: MixNet with Co-packaged Optical I/O

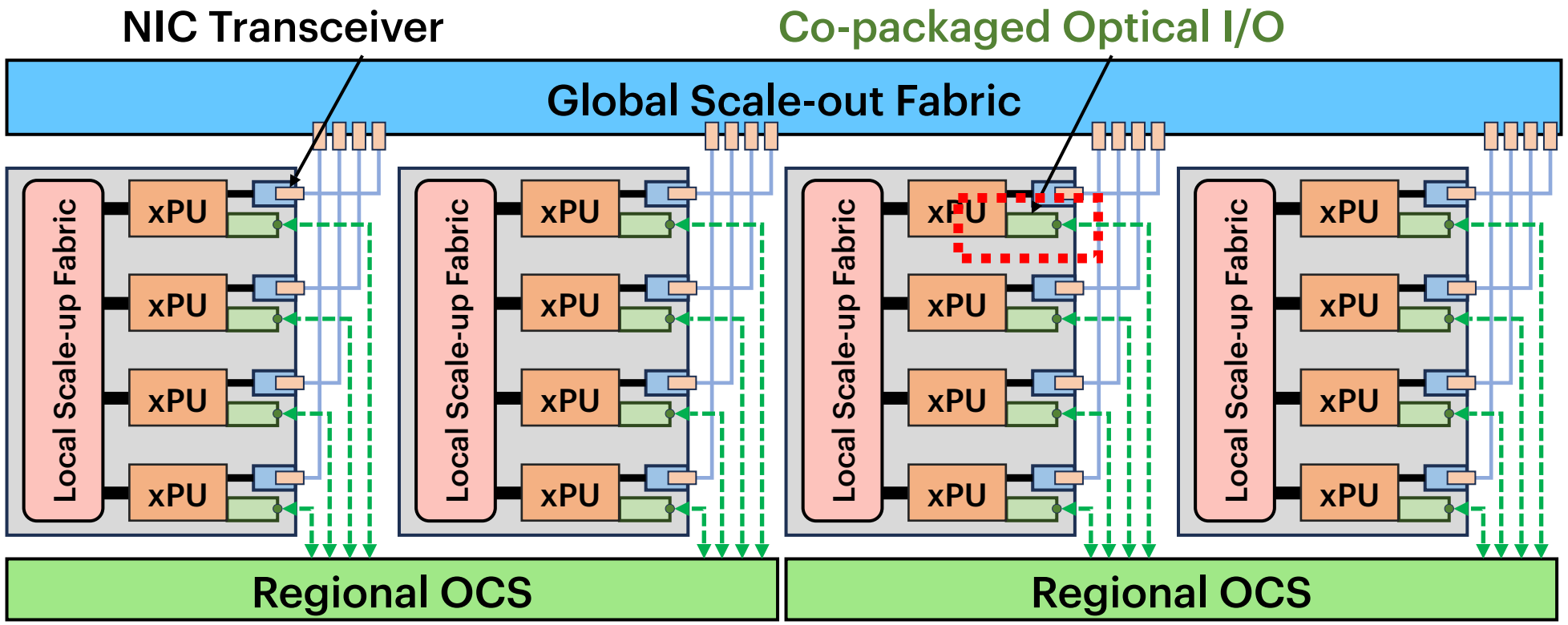


Look Forward: MixNet with Co-packaged Optical I/O



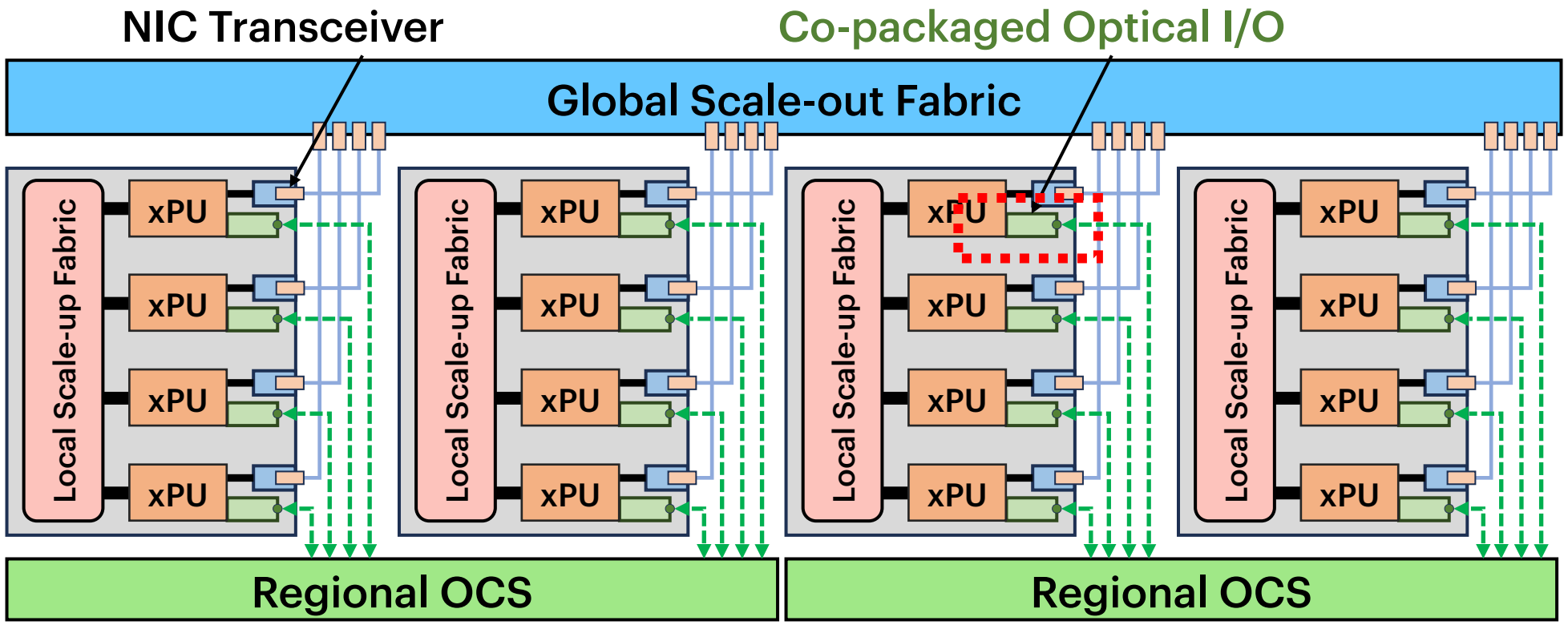
- ***Intra-node*** ultra-high speed scale-up electrical interconnect

Look Forward: MixNet with Co-packaged Optical I/O



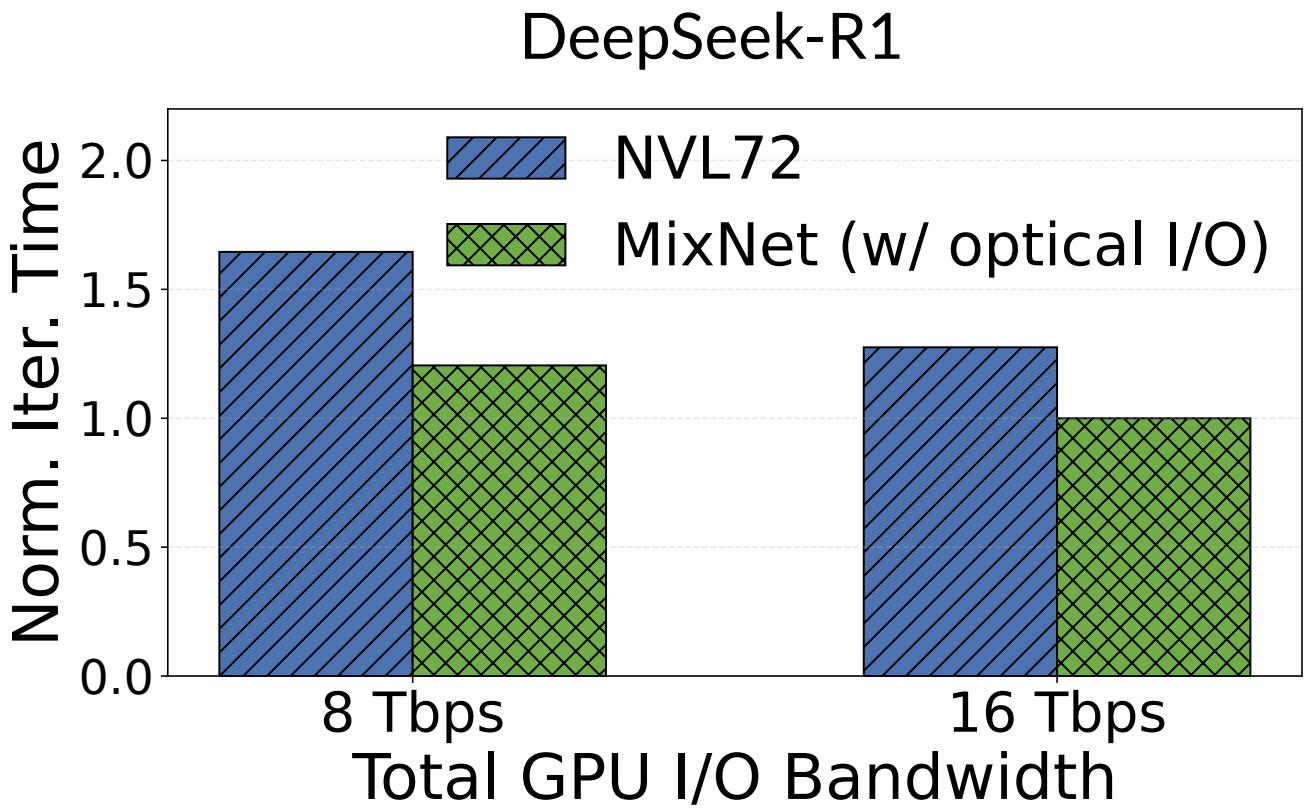
- ***Intra-node*** ultra-high speed scale-up electrical interconnect
- ***Regional*** medium-reach high-speed xPU-direct optical interconnect

Look Forward: MixNet with Co-packaged Optical I/O

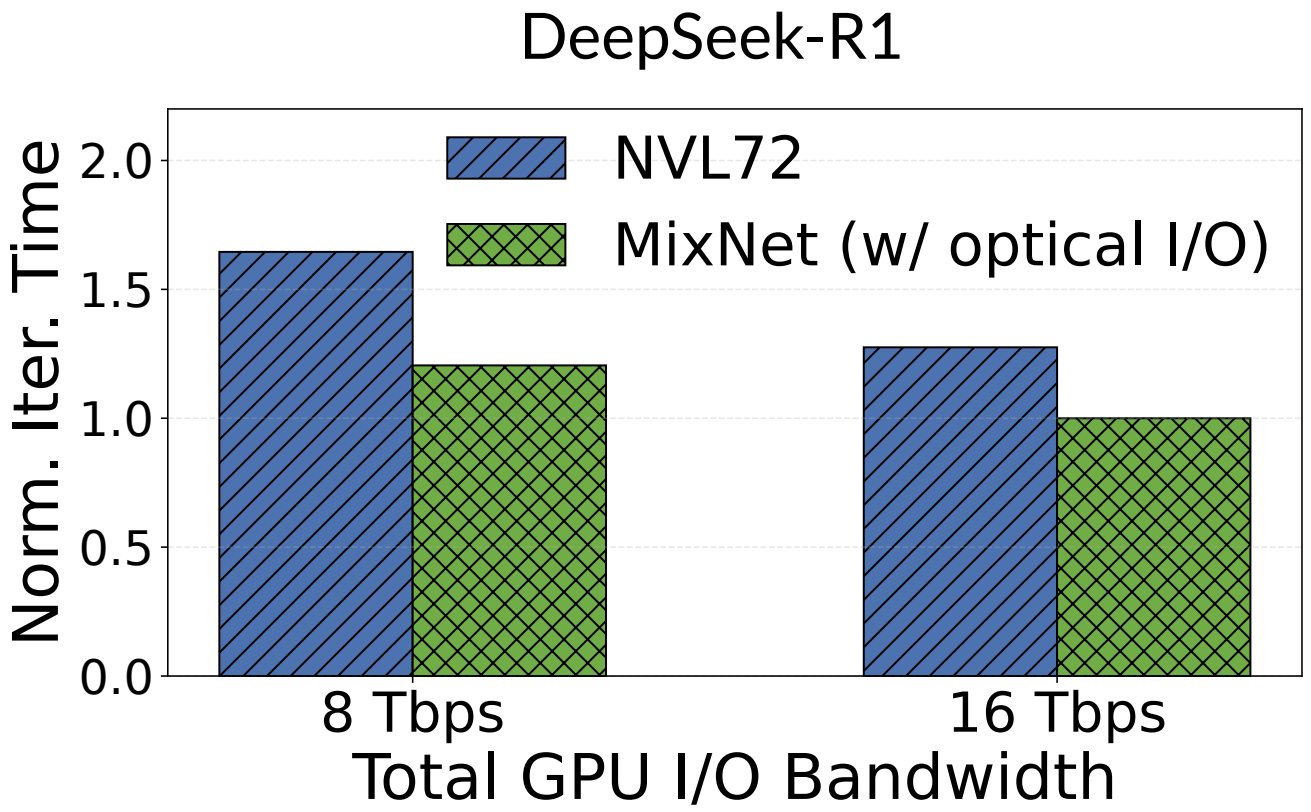


- ***Intra-node*** ultra-high speed scale-up electrical interconnect
- ***Regional*** medium-reach high-speed xPU-direct optical interconnect
- ***Cluster-scale*** electrical Ethernet interconnect

MixNet (w/ CPO) vs. NVL 72 System



MixNet (w/ CPO) vs. NVL 72 System



MixNet arguments the high-radix NVL72 system by 1.3x.

MixNet Recap

- MixNet is a ***first-of-its-kind*** system that designs mixed optical-electrical fabric for large-scale distributed MoE training, unlocking runtime topology reconfigurations.

- The core to MixNet is the design and implementation of ***regionally reconfigurable high-bandwidth domain***, making it flexible yet scalable.

- Forwarding-looking optical techniques can benefit from MixNet by arranging reconfigurable OCS within each EP group for providing massive bandwidth with intra-collective flexibility.

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Thank You!

Contact email: xudong.liao.cs@gmail.com
Personal homepage: https://xudongliao.github.io/