

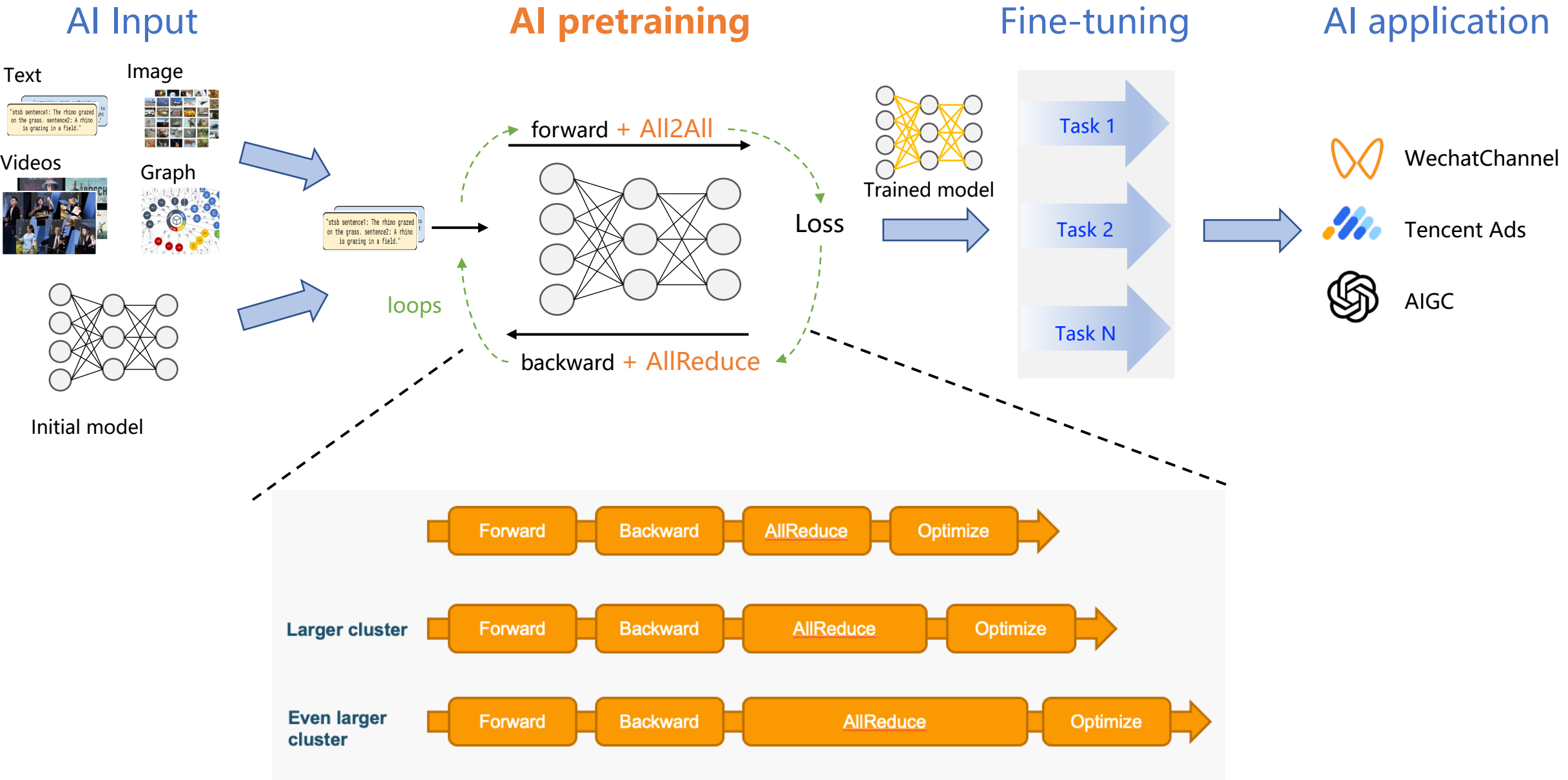
AsterNetwork: Efficient Large-scale Datacenter Network for LLM Training

Tencent Datacenter Network Architect

Baojia Li

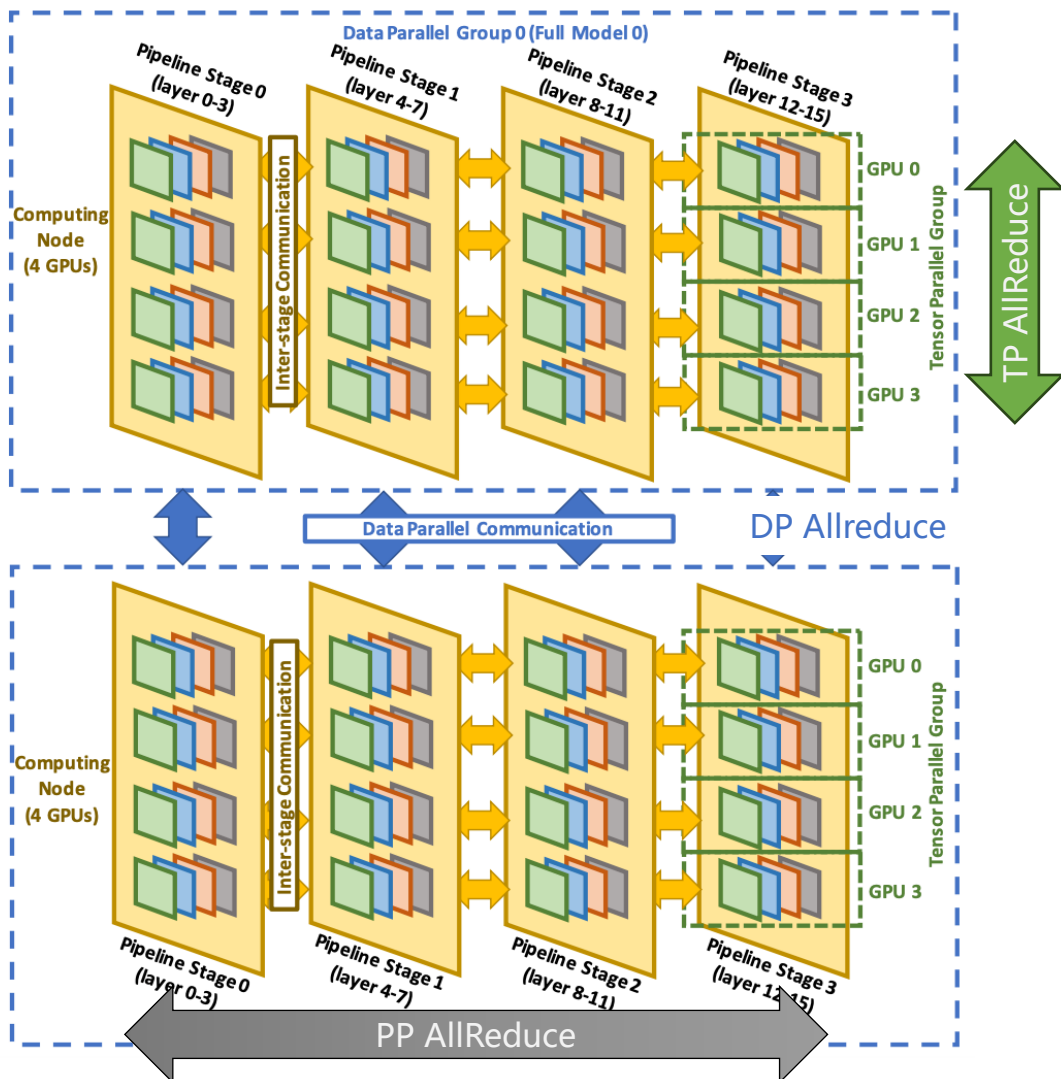
July 25 2023

Network is Bottleneck of LLM Training



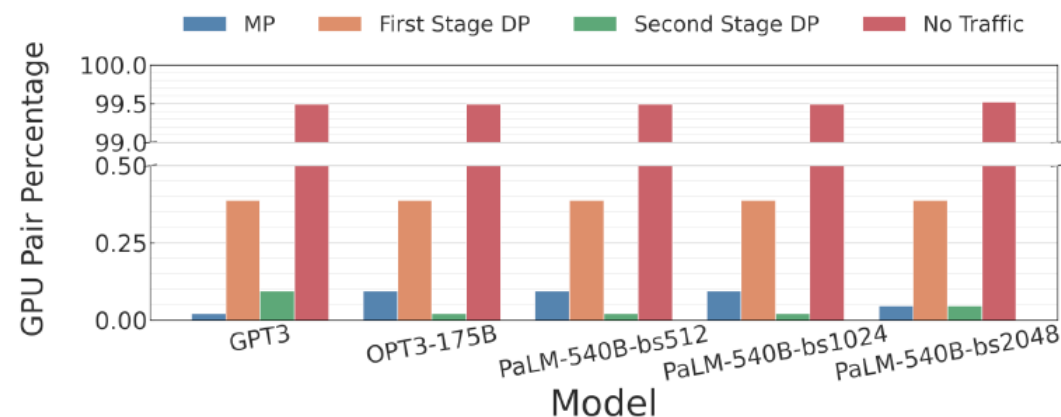
Distributed LLM Training with 3D Parallelism

3-Dimension Parallelism



Parallelism	Features	Requirements for Communication	
Tensor Parallelism	<ul style="list-style-type: none"> The highest proportion ~100GB Cannot be overlapped 	TP Allreduce	Intra-node over PCIe/Nvlink
Pipeline Parallelism	10M~10G Cannot be overlapped	PP allreduce	Inter-node over IB/Ethernet
Data Parallelism	1G~10G Can be overlapped	DP allreduce	Inter-node over IB/Ethernet
Expert Parallelism	~10G Cannot be overlapped	AlltoAll	Inter-node over IB/Ethernet

Traffic type distribution for all pairs of GPUs

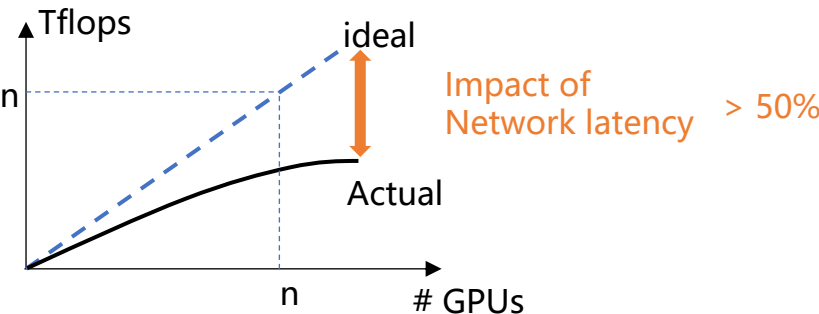


- Over 99% of GPU pairs carry no traffic
- 0.25% of GPU pairs carry MP and second stage DP traffic between them

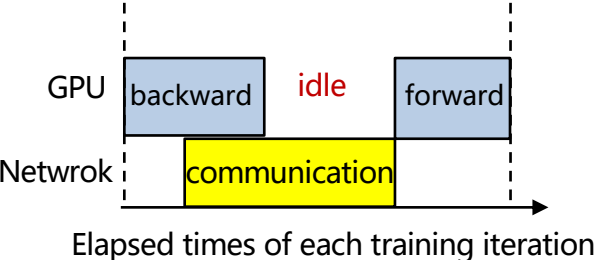
Network Requirements of Distributed LLM Training

Large Cluster Size \neq Effective Flops

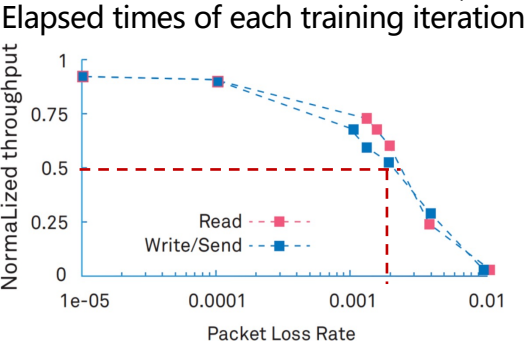
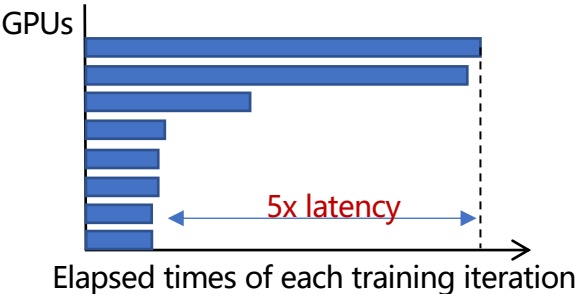
Network Performance constrains effective flops of GPU clusters



Impacts of Bandwidth (TB magnitude)

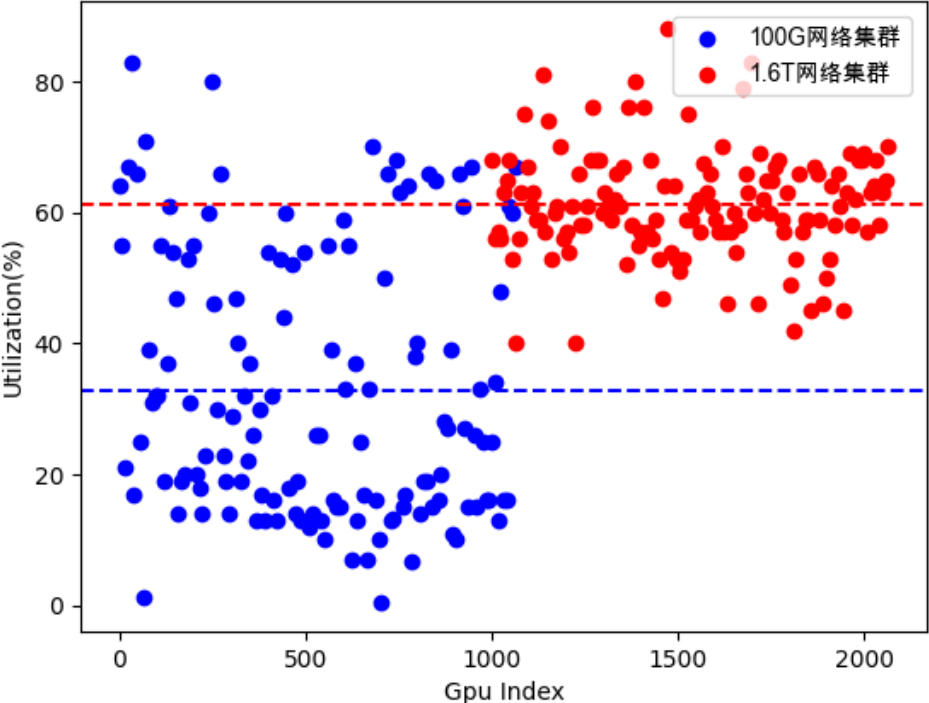


Impacts of Latency (5x)



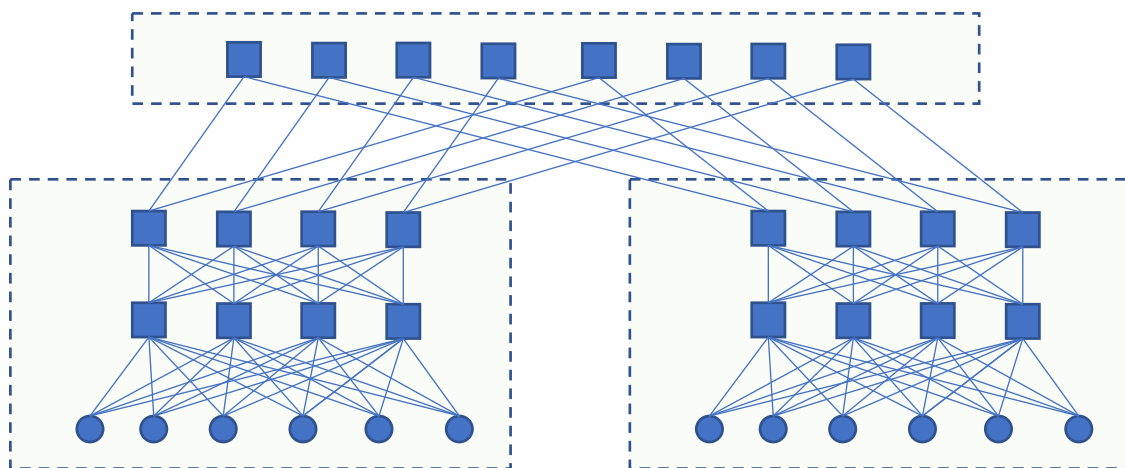
Impacts of packet loss (0.1% packet loss incur 50%)

GPU usage 32%→61%



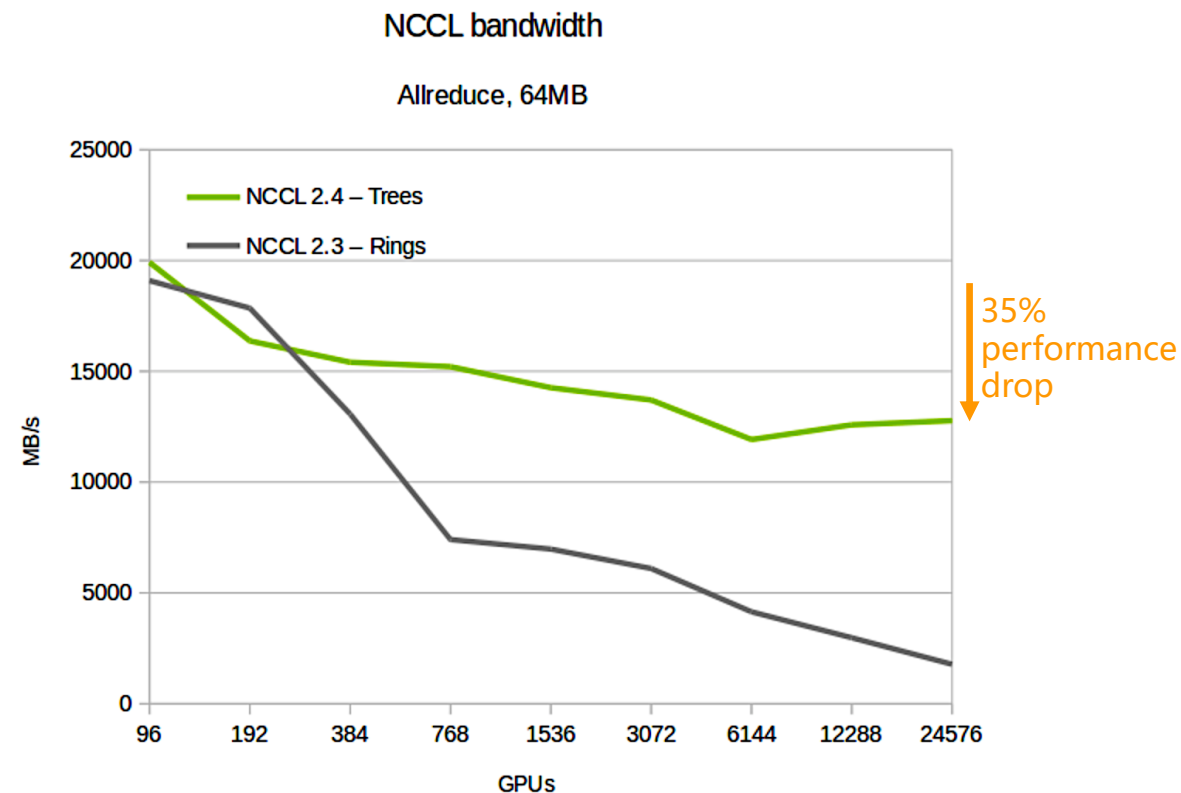
Challenge1: Increase in Network Size

Size of the **GPU cluster** is limited by the **throughput of the switching chip** and the **number of device ports**



- **Performance:** Affected by uneven multi-layer load balancing and multi-level network congestion, collective communication performance is degraded.
- **Cost:** the number of switches and optical modules increase.
- **Operation:** Affected by multi-layer PFC Pause, network operation risks increase.

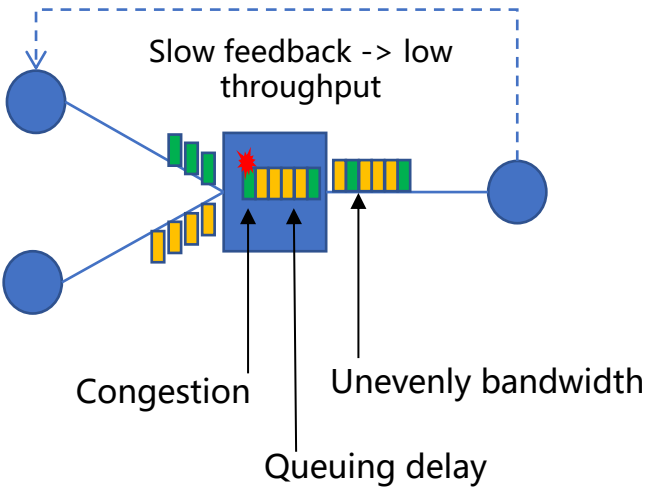
In large-scale scenarios, the collective communication performance **drops by 35%**



Network Protocol

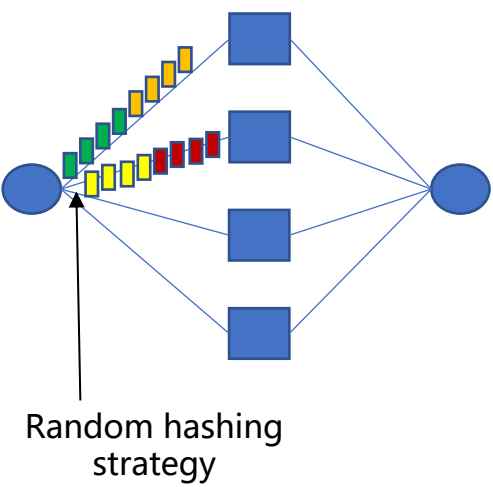
Congestion control

Congestion: low utilization, high delay, packet loss



Multi-path Load balancing

Random path selection: hash polarization, low utilization

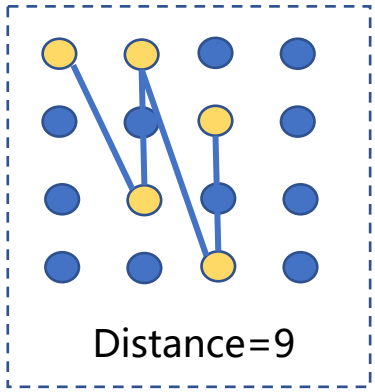


Collective Communication Library

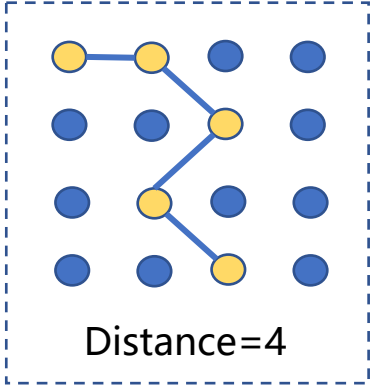
Collective Communication optimization

Topology-unaware: detours, sub-optimal paths

● Involved GPU



Topology-unaware



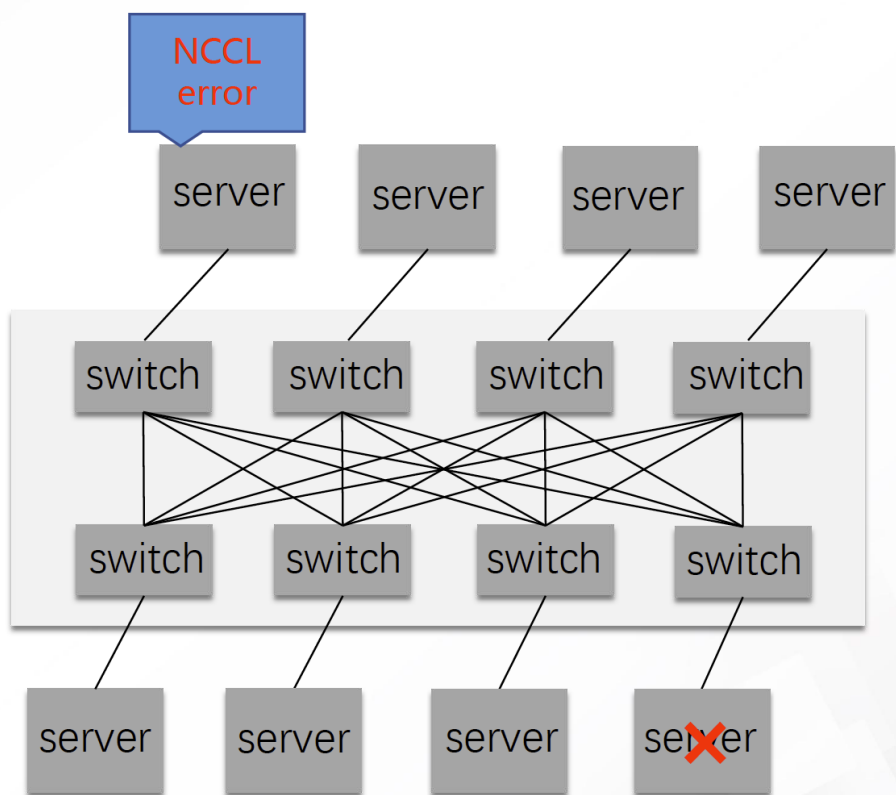
Topology-aware

	Congestion Control	Load Balancing Bias	Packet Loss	Collective Communication
Tradition Protocol	500us ~ 1ms	~ 40%	0.1% ~ 1%	Topology-unaware
Self-developed Protocol	10us ~ 40us	~ 5%	0	Topology-aware

Challenge3: High Cluster Stability Requirements

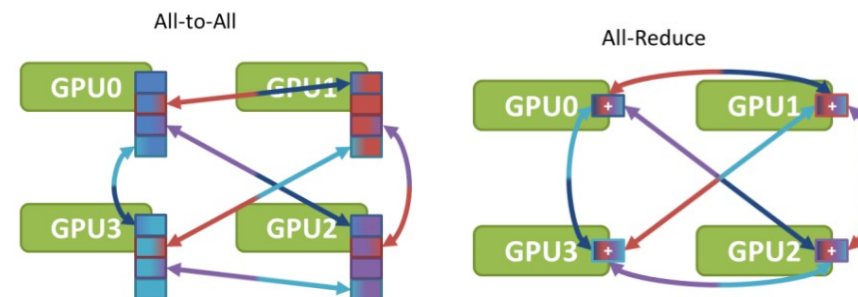
Tencent 腾讯

A single node/device failure can cause the entire training to be disrupted

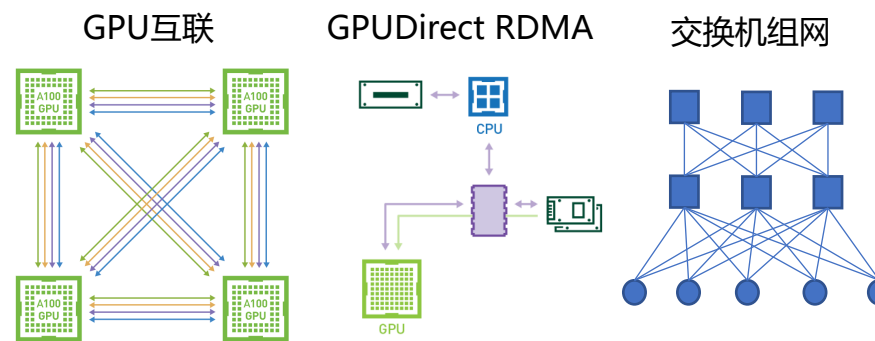


The technical system is complex and it is difficult to locate the fault point.

集合通信



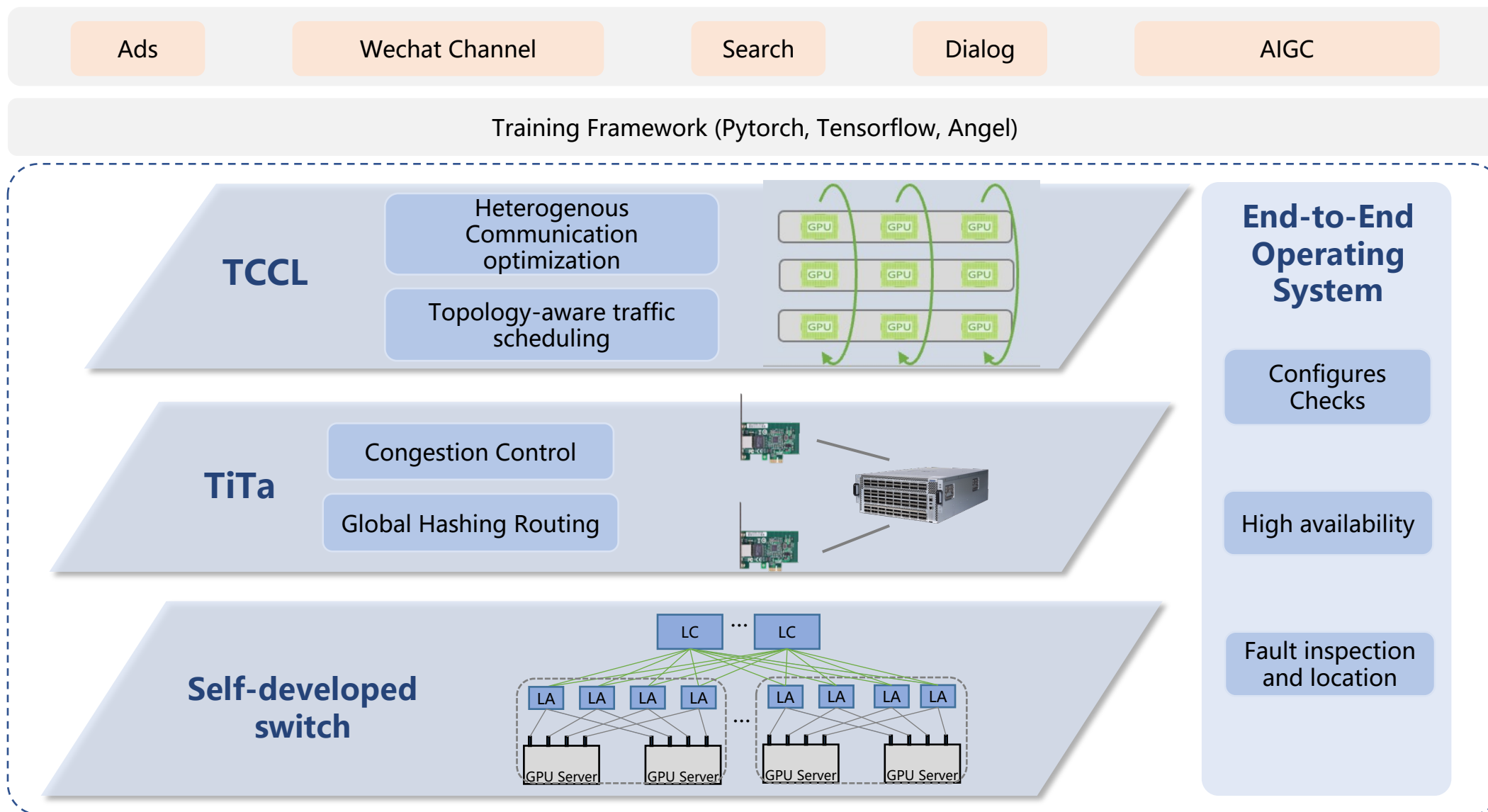
基础网络



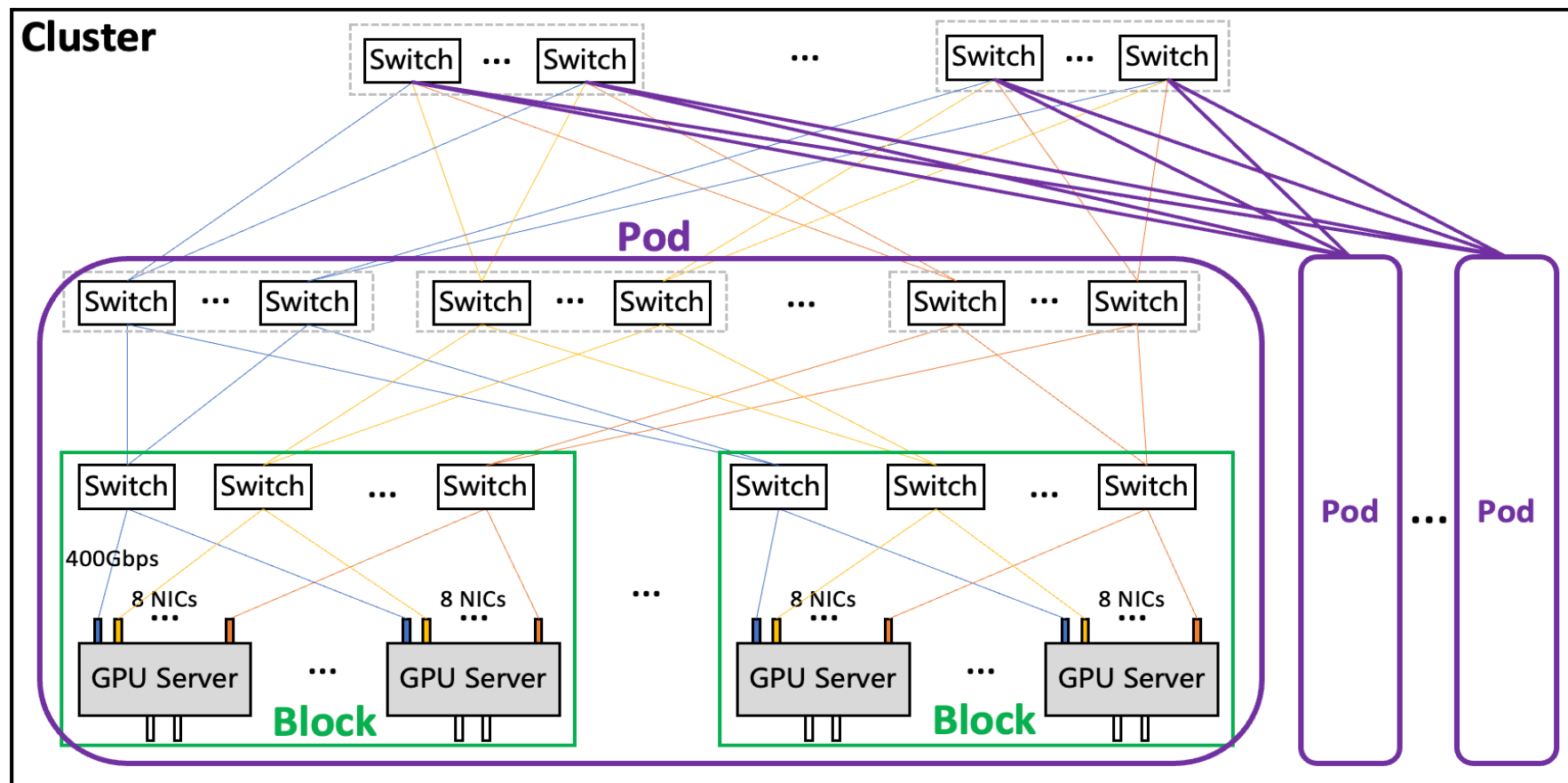
AstralNetwork: Self-developed TiTa and TCCL

Tencent 腾讯

Extreme cluster performance and intelligent operation



Network Scale of 100 thousands GPU in a single cluster



Block-Pod-Cluster

- **Block**: 256 GPU
- **Pod**: 16~64 Block (4K~16K GPU) **Cluster**: Up to 16 Pod (64K~256K GPU)

Customized interconnection of AI traffic to maximize performance

- Multi-plane: Distribute cross-server GPU communication on 8 planes
- Traffic affinity: ensure that the single-task 8K/16K GPU is accessed on the same plane

Tencent Intelligent Traffic Aware protocol (Tita)

Tencent 腾讯

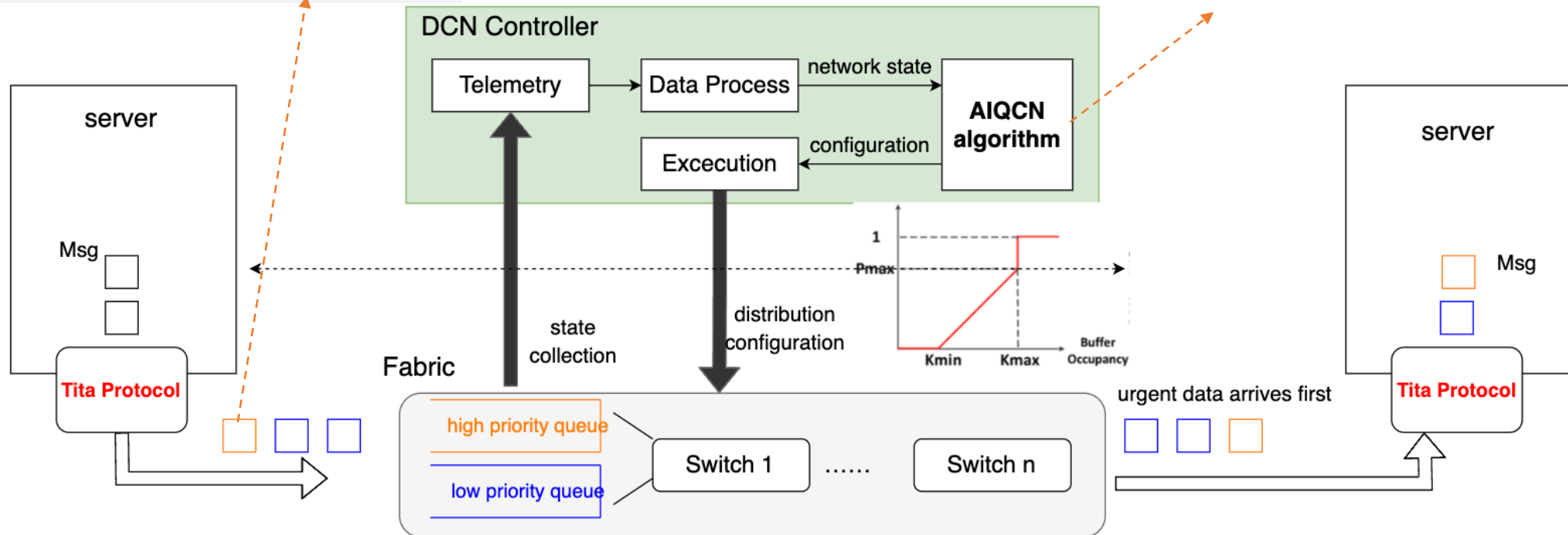
Congestion control based on end-network coordination

End

Set the message transmission priority according to the communication flow status

Network

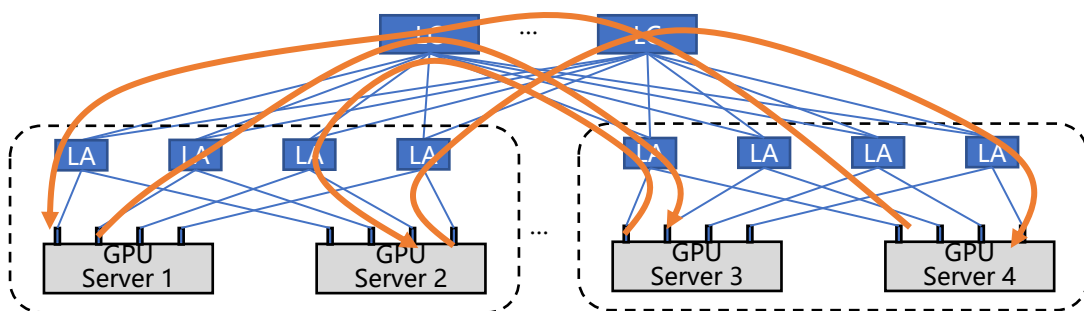
According to the degree of traffic conflict, control the sending rate of the source end



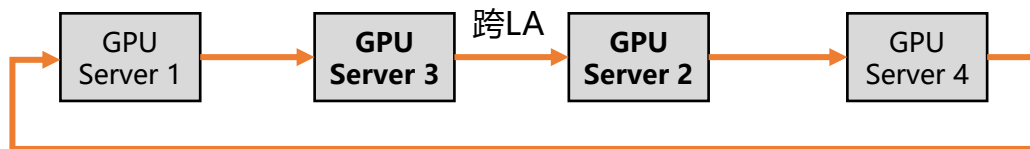
Effective load of the whole network reaches more than 90%

Topology-aware affinity scheduling: minimizing traffic detours

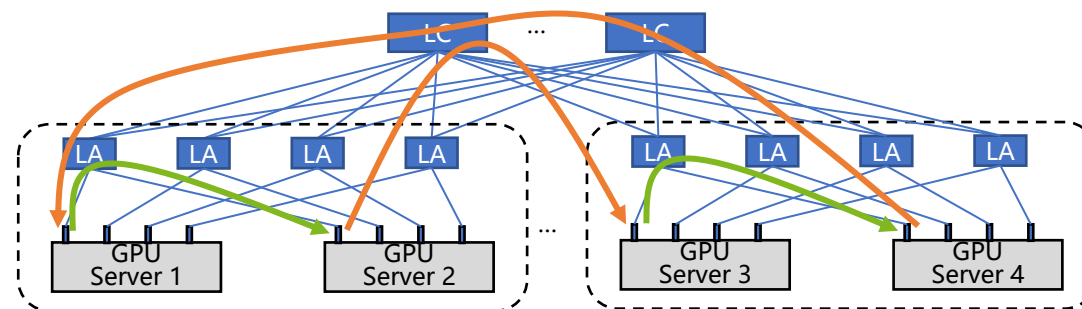
➤ Suboptimal traffic scheduling



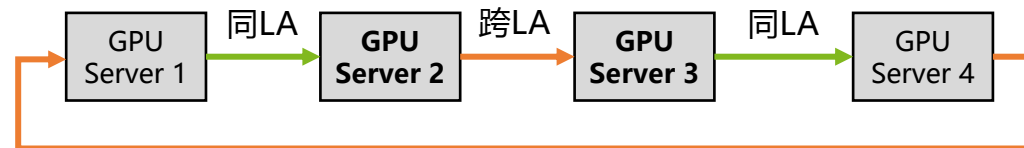
➤ Suboptimal collective communication sequence



➤ Optimal traffic scheduling



➤ Optimal collective communication sequence



Traffic cross LA groups was reduced by 50%~80%

Thanks