Tencent

AsteralNetwork: Efficient Large-scale Datacenter Network for LLM Training

Tencent Datacenter Network Architect

Baojia Li

Nov 07 2023

Network is Bottleneck of LLM Training

Tencent

AI Input

Text

Image

Videos

Graph

Initial model

AI pretraining

forward + All2All

loops

backward + AllReduce

Loss

Trained model

Fine-tuning

Task 1

Task 2

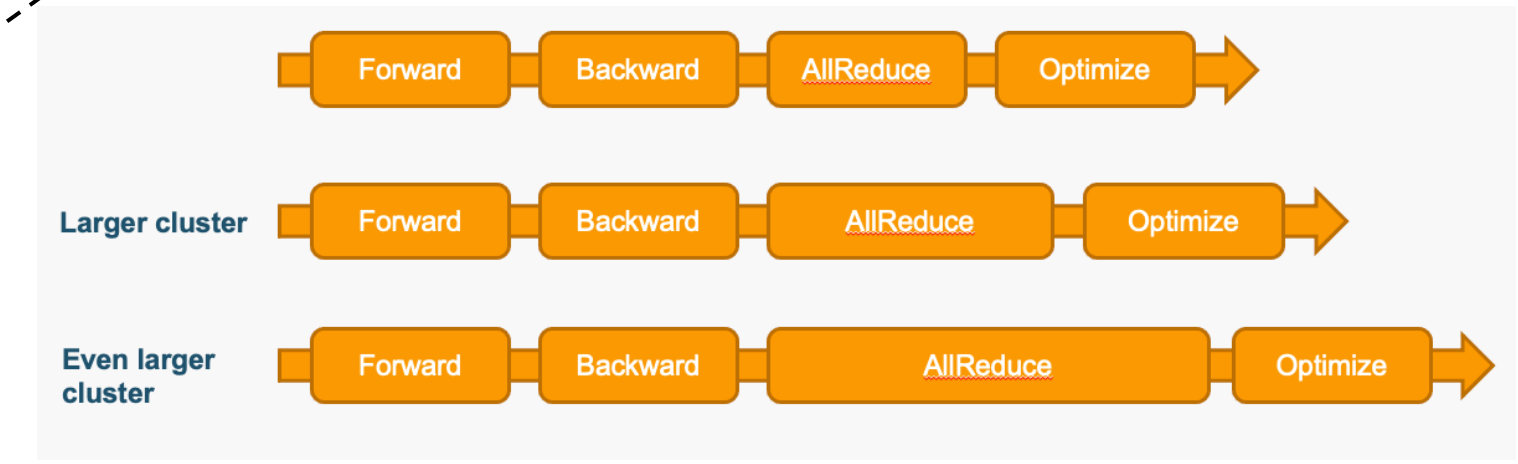
Task N

AI application

WechatChannel

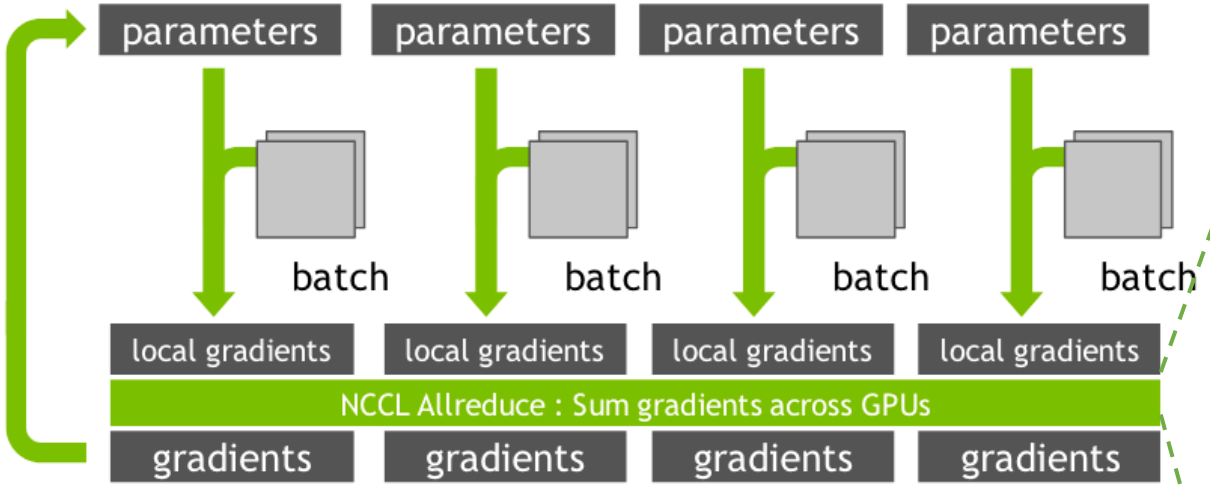
Tencent Ads

AIGC

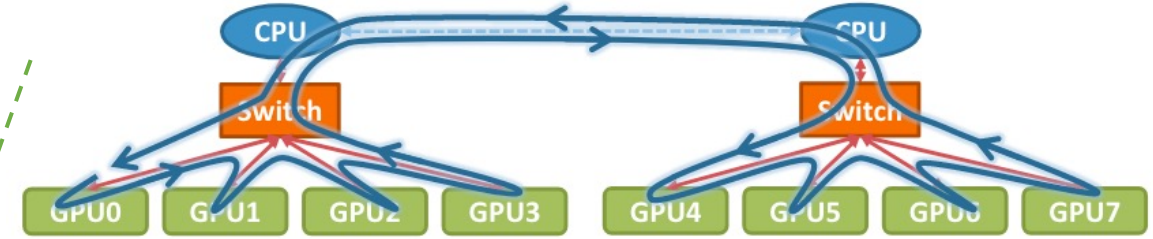


Collective Communication during LLM Training

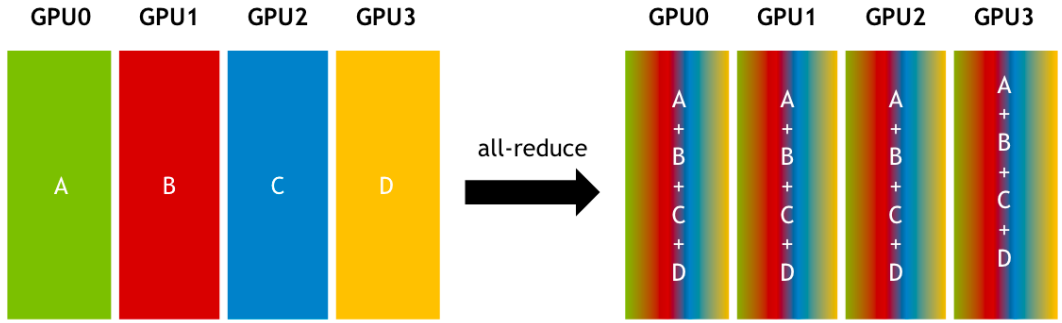
Tencent



Ring-Based Collective



AllReduce Algorithm

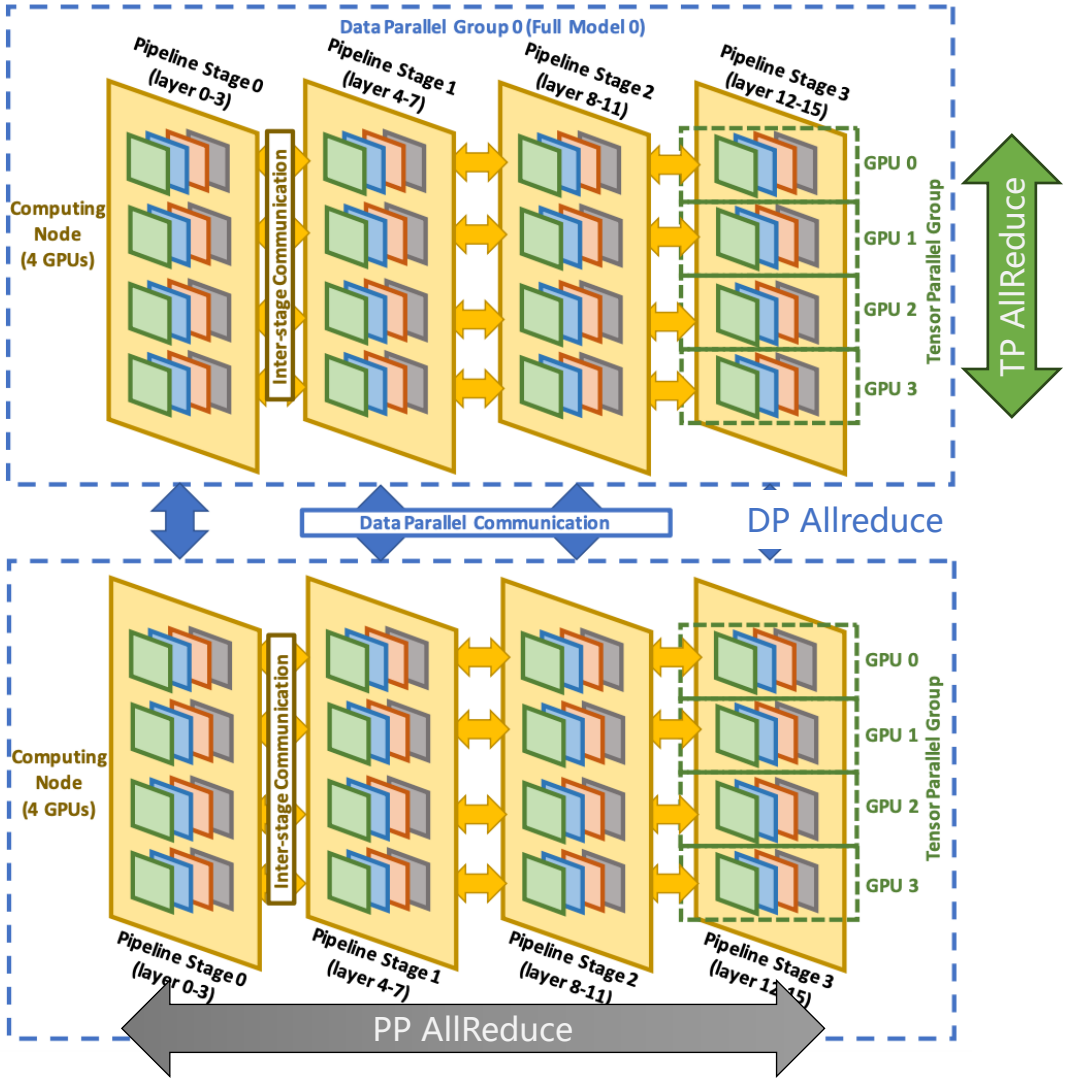


Combine data from all senders; deliver the result to all participants

Distributed LLM Training with 3D Parallelism

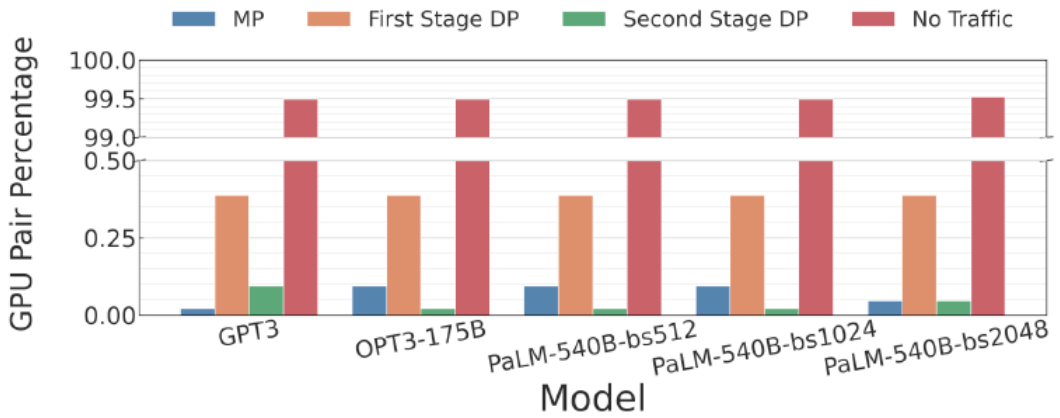
Tencent

3-Dimension Parallelism



Parallelism	Features	Requirements for Communication	
Tensor Parallelism	• 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

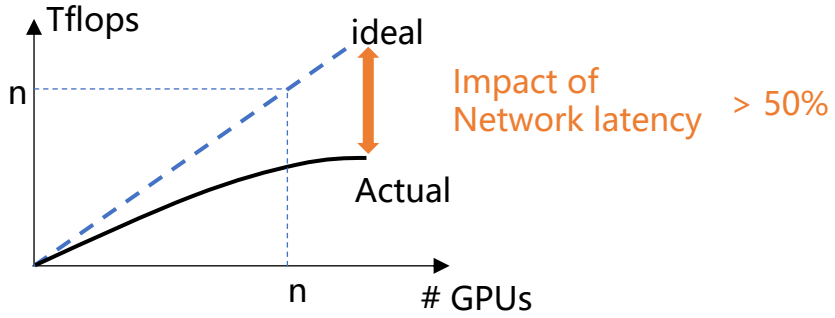
Image source: Optimus-CC: Efficient Large NLP Model Training with 3D Parallelism Aware Communication Compression

Network Requirements of Distributed LLM Training

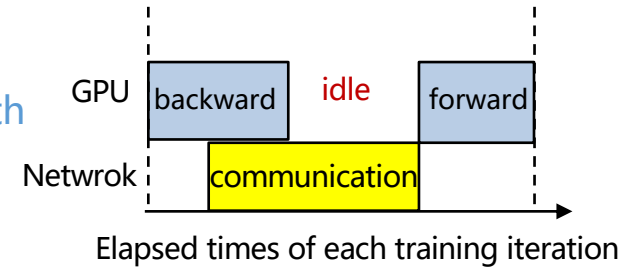
Tencent

Large Cluster Size ≠ Effective Flops

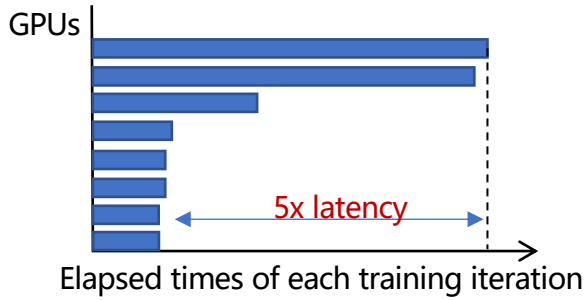
Network Performance constrains effective flops of GPU clusters



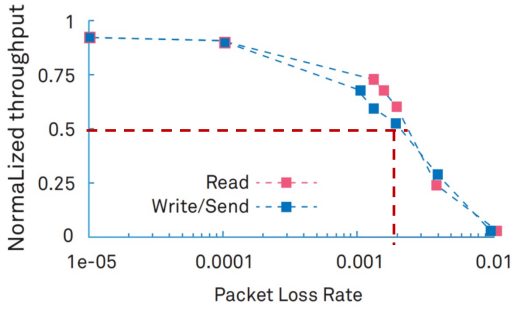
Impacts of Bandwidth (TB magnitude)



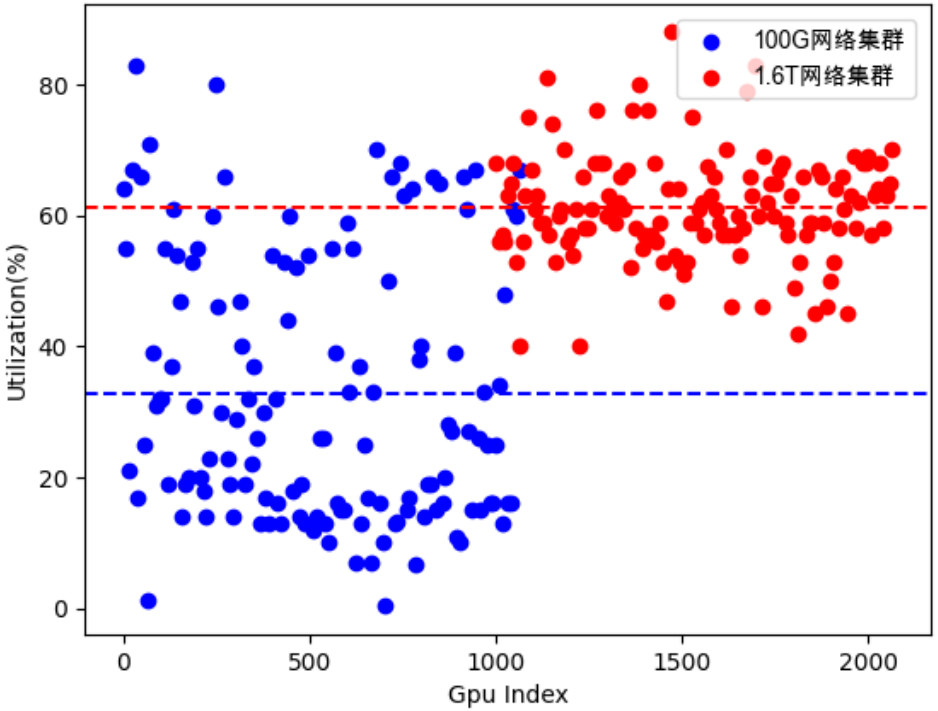
Impacts of Latency (5x)



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



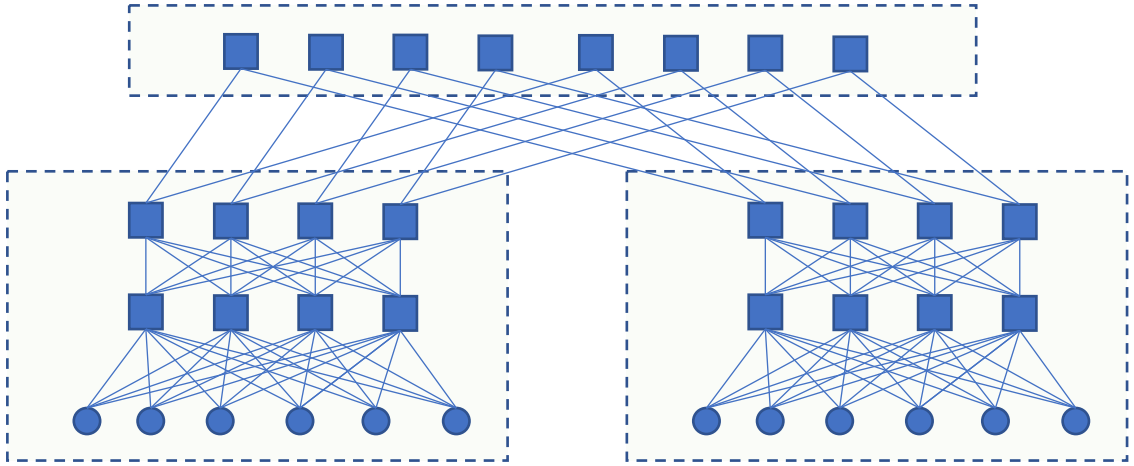
GPU usage **32%→61%**



Challenge1： Increase in Network Size

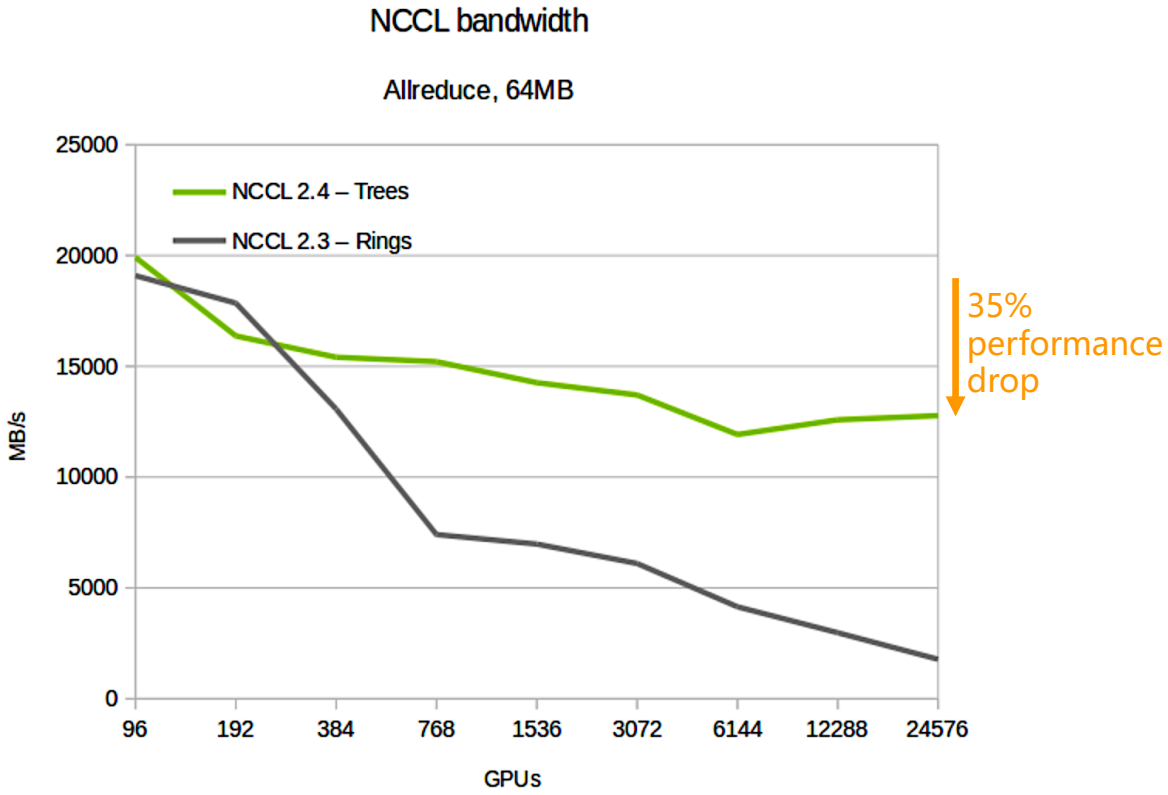
Tencent

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%**



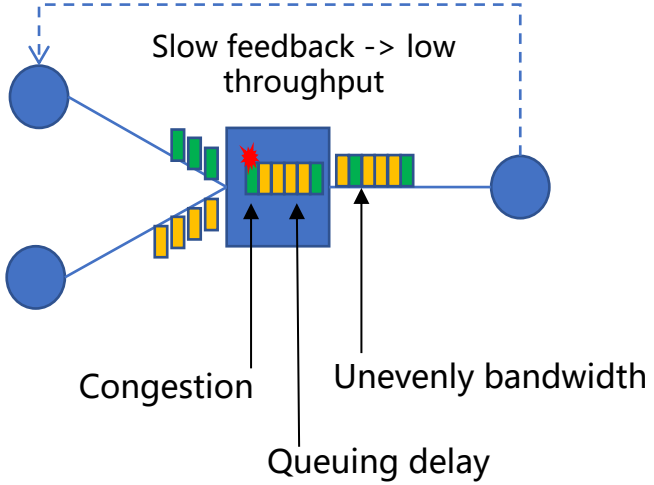
Challenge2: Tradition Protocols Restrict Efficiency

Tencent

Network Protocol

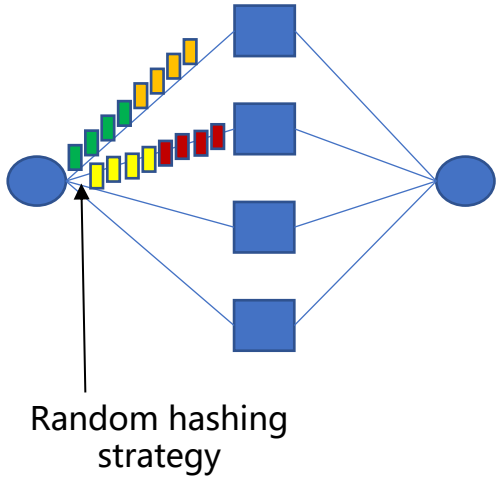
Congestion control

Congestion: low utilization, high delay, packet loss



Multi-path Load balancing

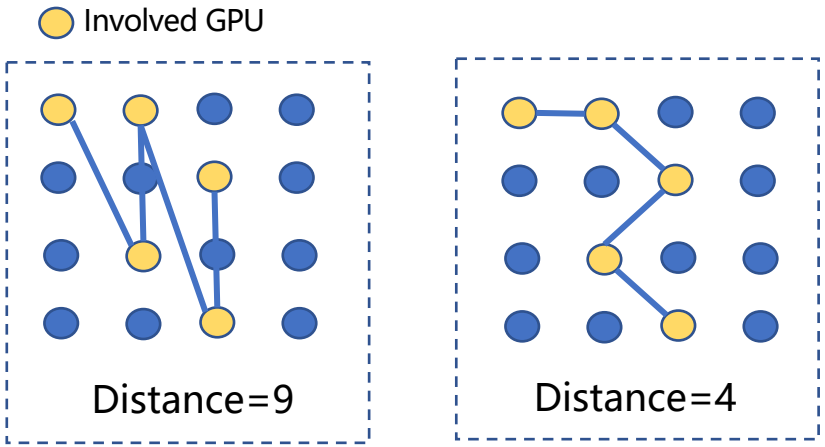
Random path selection: hash polarization, low utilization



Collective Communication Library

Collective Communication optimization

Topology-unaware: detours, sub-optimal paths

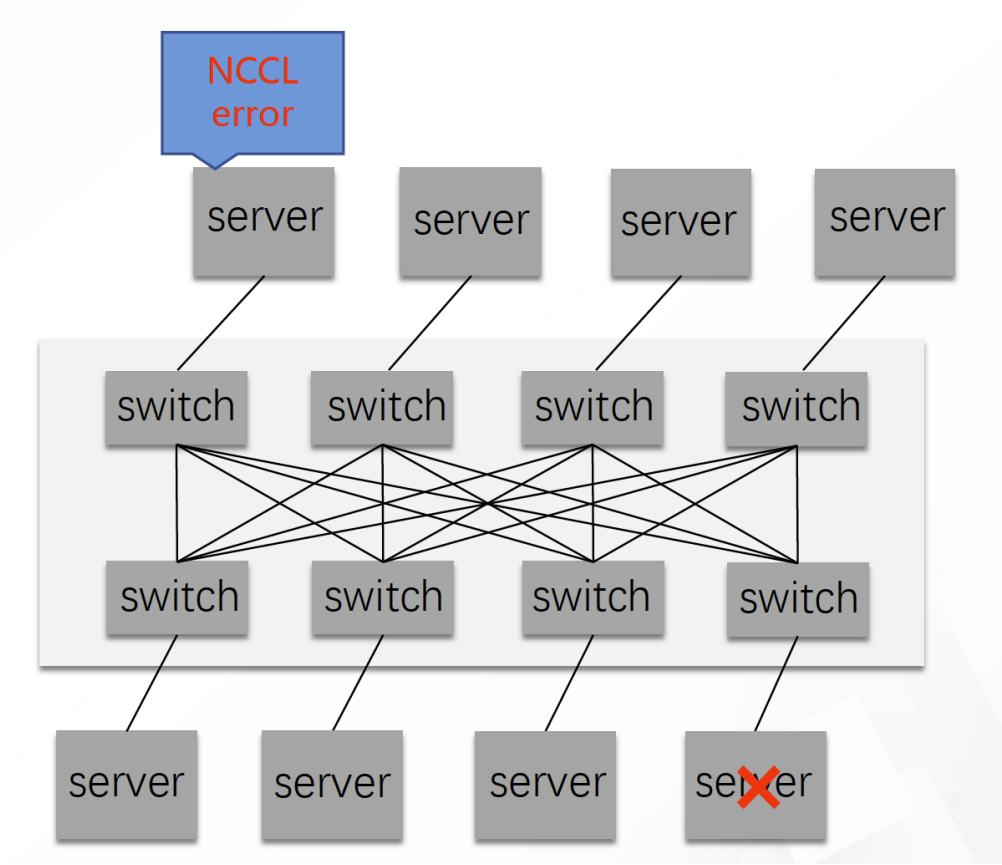


	Congestion Control	Load Balancing Bias	Packet Loss	Collective Communication
Tradition Protocol	**500us ~ 1ms**	**~ 40%**	**0.1% ~ 1%**	Topology-unaware
Self-developed Protocol	**10us ~ 40us**	**~ 5%**	**0**	Topology-aware

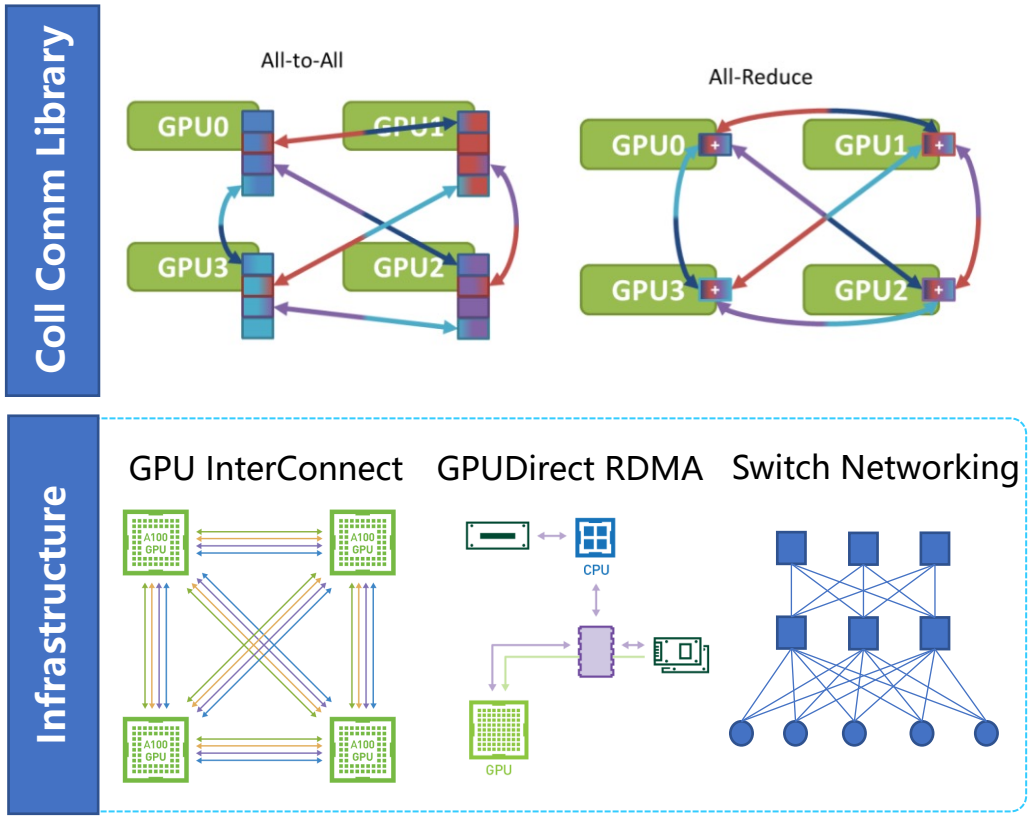
Challenge3: High Cluster Stability Requirements

Tencent

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.



AstralNetwork：Self-developed TiTa and TCCL

Tencent

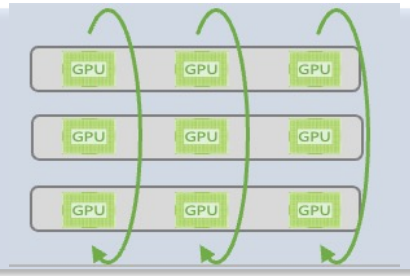
Extreme cluster performance and intelligent operation

Ads | Wechat Channel | Search | Dialog | AIGC

Training Framework (Pytorch, Tensorflow, Angel)

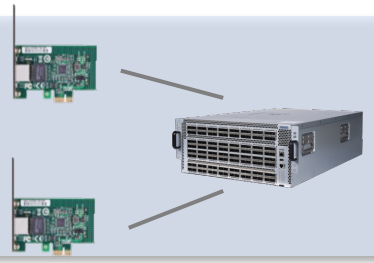
TCCL

- Heterogenous Communication optimization
- Topology-aware traffic scheduling

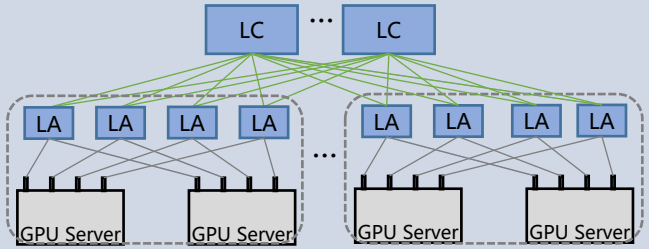


TiTa

- Congestion Control
- Global Hashing Routing



Self-developed switch



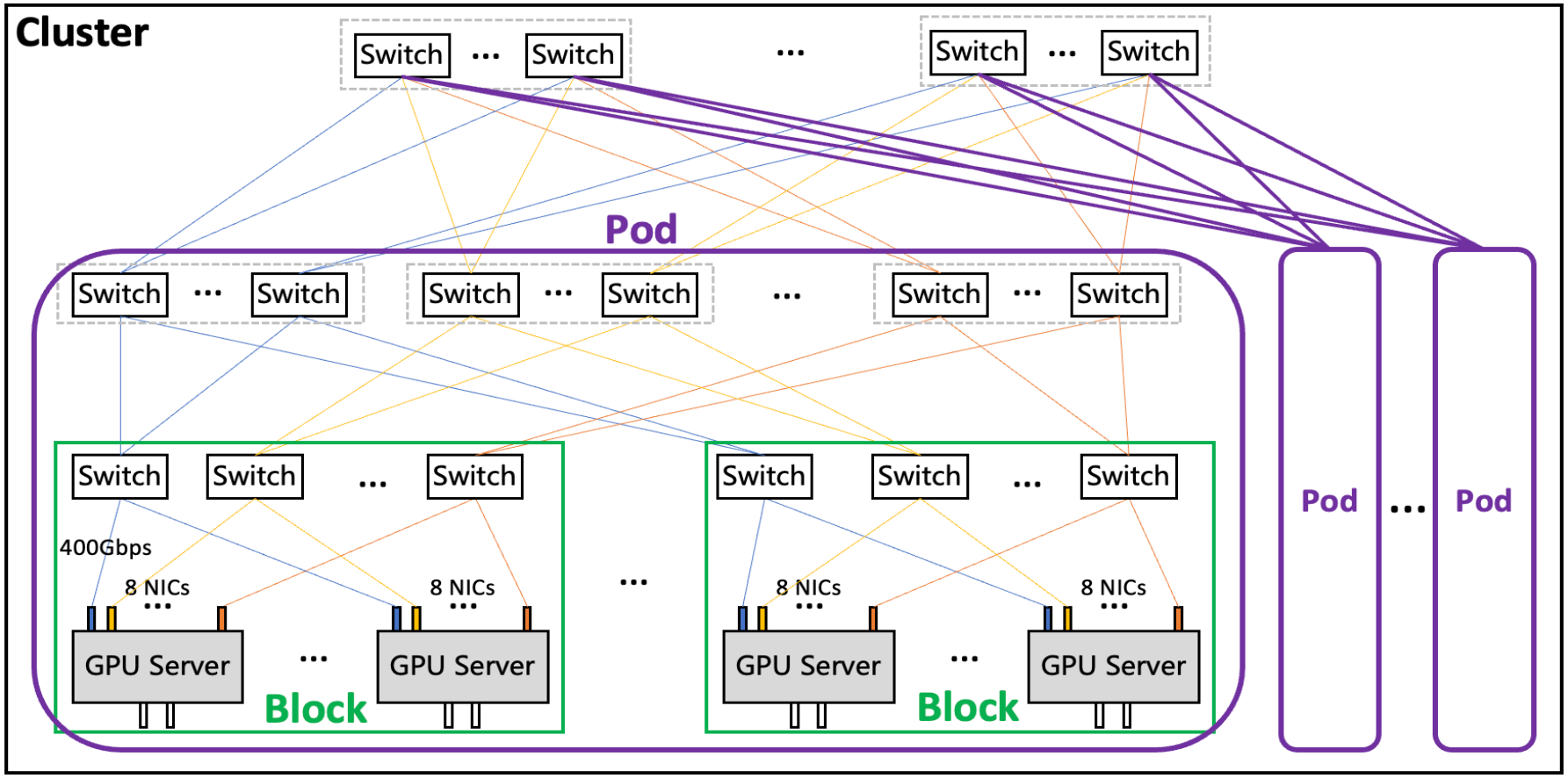
End-to-End Operating System

- Configures Checks
- High availability
- Fault inspection and location

Large-scale and Flexible Networking Architecture

Tencent

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

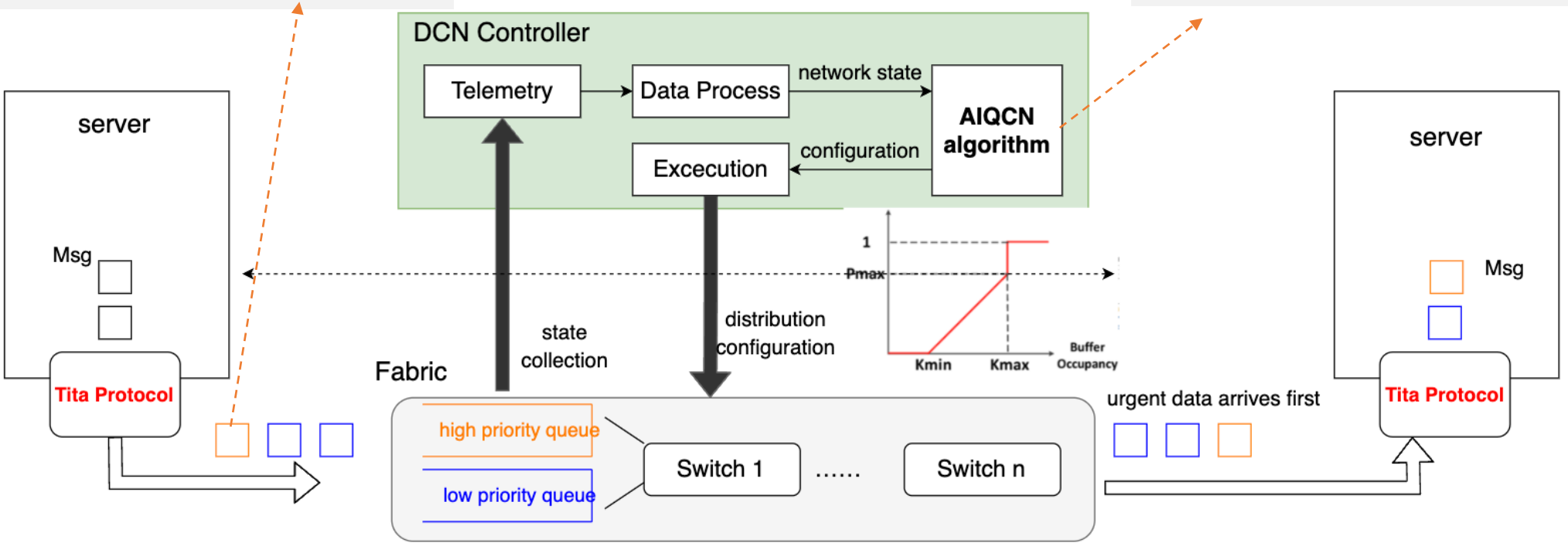
Congestion control based on end-network coordination

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



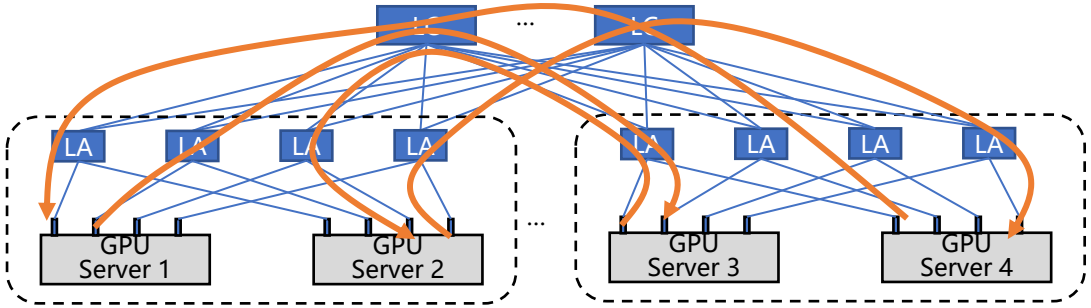
Effective load of the whole network reaches more than 90%

Tencent Collective Communication Library (TCCL)

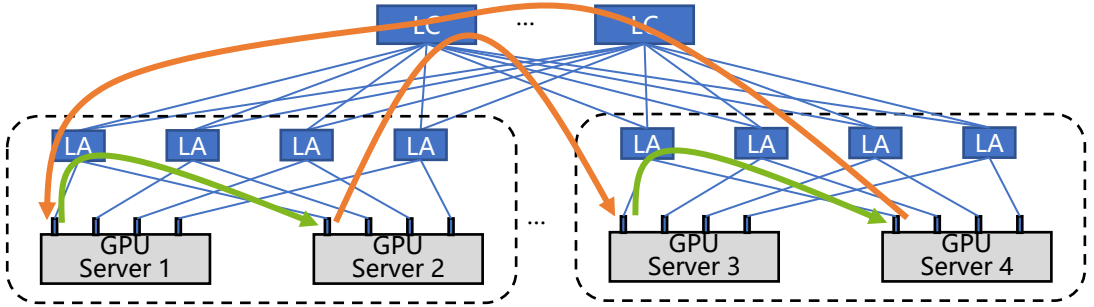
Tencent

Topology-aware affinity scheduling: minimizing traffic detours

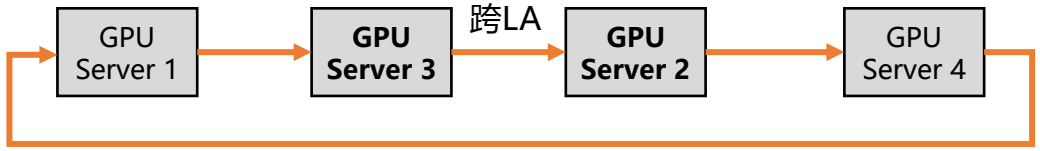
- Suboptimal traffic scheduling



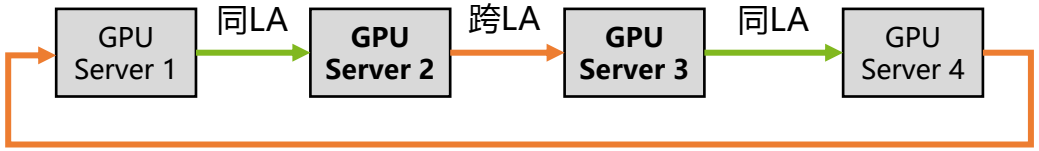
- Optimal traffic scheduling



- Suboptimal collective communication sequence



- Optimal collective communication sequence



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

Tencent

Thanks