Alibaba HPN Introduction

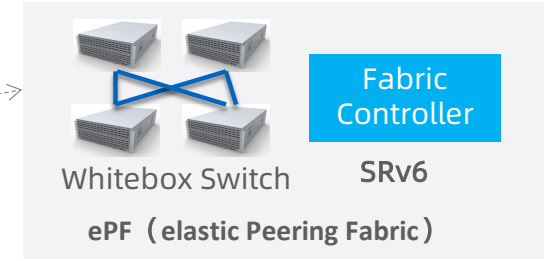
Eddie Ruan
07/2024

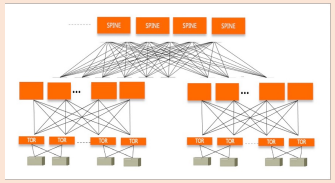
The big picture – AliCloud Network

Global Service Center

NetO Network Orchestrator

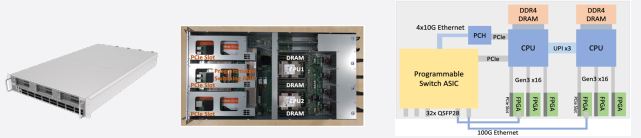
Global DCI Network





SuperScale DCN Network

HPN in AI Infra.



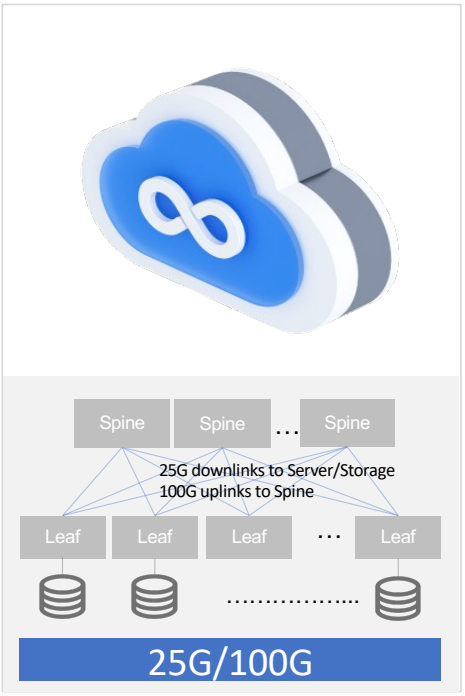
Distributed Edge Cloud

SONiC P4 NPL

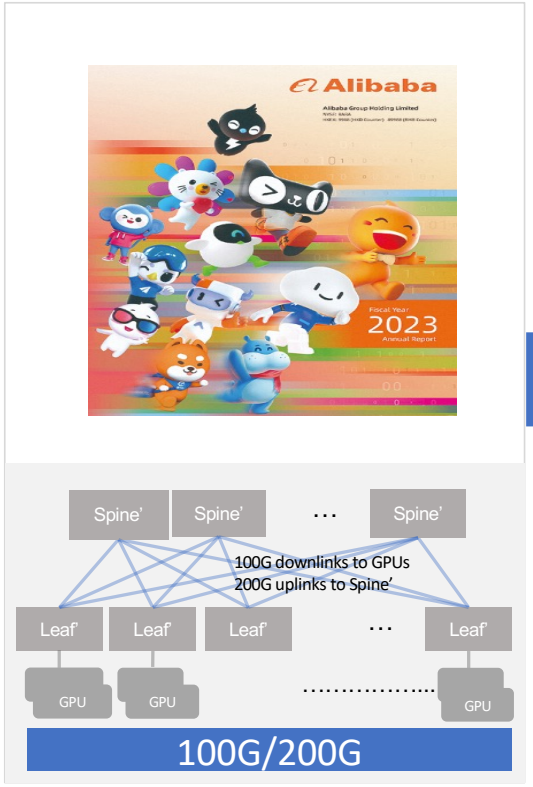
IP/SRv6 + Single-chip white box+ Programmability

The hotspot: AI infrastructure within AliCloud

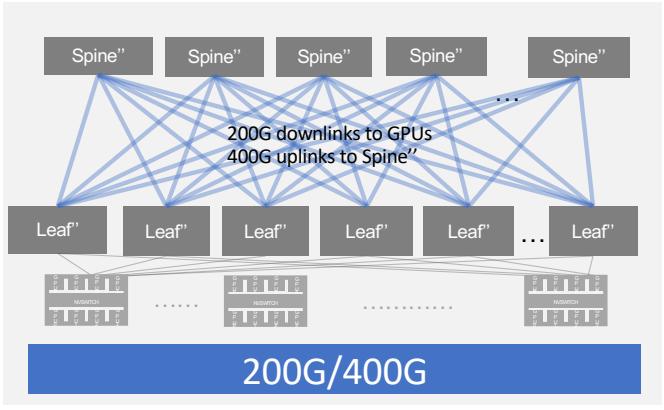
2017 - DC 5.0, 5.2



2021 – DC 6.0



2023- HPN 7.0/ DC 7.0



Alibaba HPN: A Data Center Network for Large Language Model Training

Kun Qian, Yongqing Xi, Jiamin Cao, Jiaqi Gao, Yichi Xu, Yu Guan, Binzhang Fu, Xuemei Shi
Fangbo Zhu, Rui Miao, Chao Wang, Peng Wang, Pengcheng Zhang, Xianlong Zeng
Eddie Ruan, Zhiping Yao, Ennan Zhai, Dennis Cai
Alibaba Cloud

Abstract

This paper presents HPN, Alibaba Cloud's data center network for large language model (LLM) training. Due to the differences between LLMs and general cloud computing (*e.g.*, in terms of traffic patterns and fault tolerance), traditional data center networks are not well-suited for LLM training. LLM training produces a small number of periodic, bursty flows (*e.g.*, 400Gbps) on each host. This characteristic of LLM training predisposes Equal-Cost Multi-Path (ECMP) to hash polarization, causing issues such as uneven traffic distribution. HPN introduces a 2-tier, dual-plane architecture capable of interconnecting 15K GPUs within one Pod, typically accommodated by the traditional 3-tier Clos architecture. Such a new architecture design not only avoids hash polarization but also greatly reduces the search space for path selection. Another challenge in LLM training is that its requirement for GPUs to complete iterations in synchronization makes it more sensitive to single-point failure (typically occurring on ToR). HPN proposes a new dual-ToR design to replace the single-ToR in traditional data center networks. HPN has been deployed in our production for more than eight months. We share our experience in designing, and building HPN, as well as the operational lessons of HPN in production.

'24), August 4–8, 2024, Sydney, NSW, Australia. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3651890.3672265

1 Introduction

The large language model (or LLM) [13, 15–17, 24] has brought about tremendous revolutions to today's AI and cloud services. The training of an LLM, which has hundreds of billions of parameters, relies on a large-scale distributed training cluster, typically equipped with tens of millions of GPUs. Due to its unique characteristics, LLM training presents new challenges to the design of data center networks.

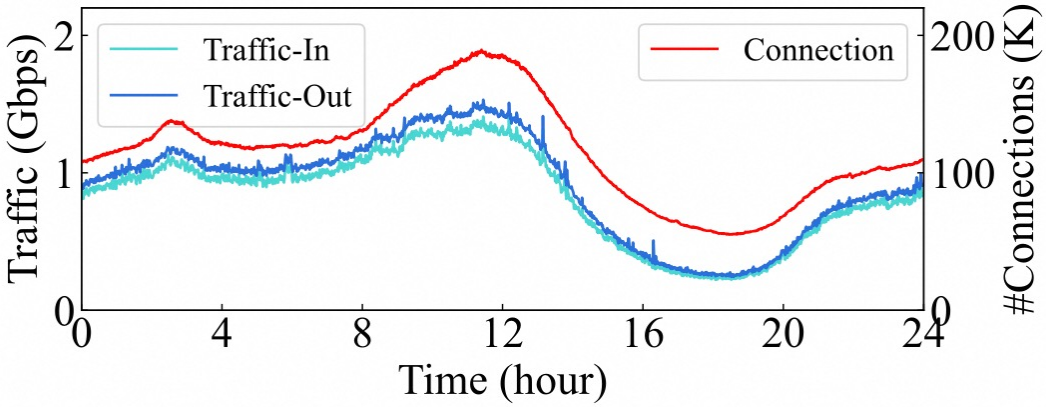
Problem 1: Traffic patterns. The traffic patterns of LLM training are different from those of general cloud computing in terms of (1) low entropy [19, 22] and (2) bursty traffic. Specifically, general cloud computing generates millions of flows, which gives the network high entropy. Each flow is continuous and low-utilization (*e.g.*, typically below 20% of the NIC capacity), as shown in Figure 1. On the contrary, LLM training generates very few but periodically bursty flows, resulting in low entropy and high utilization for the network. The burst can directly reach the NIC capacity, which is

https://ennanzhai.github.io/pub/sigcomm24-hpn.pdf

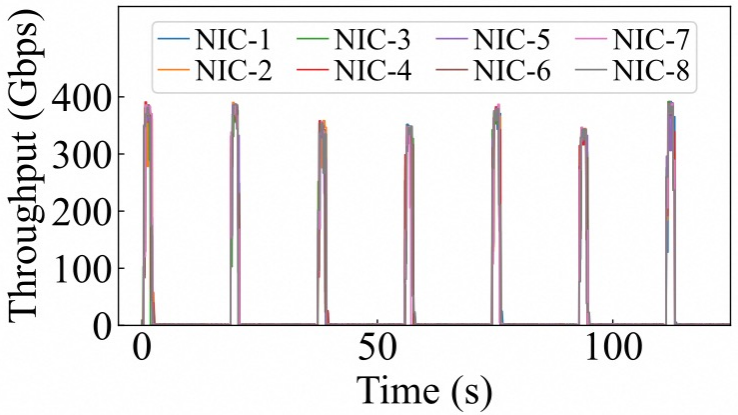
LLM Traffic Pattern

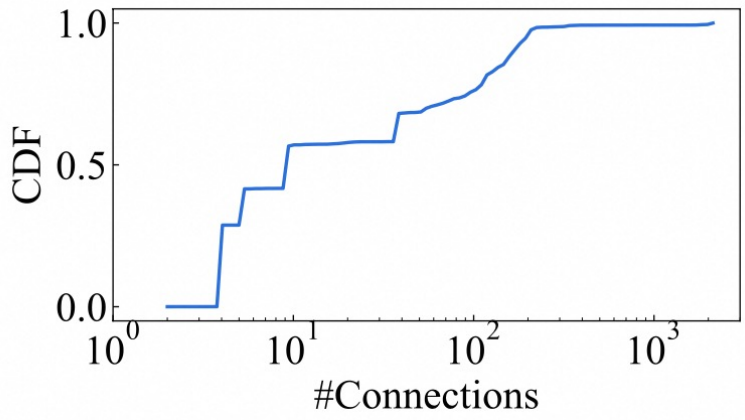
Periodic burst

Small number of flows



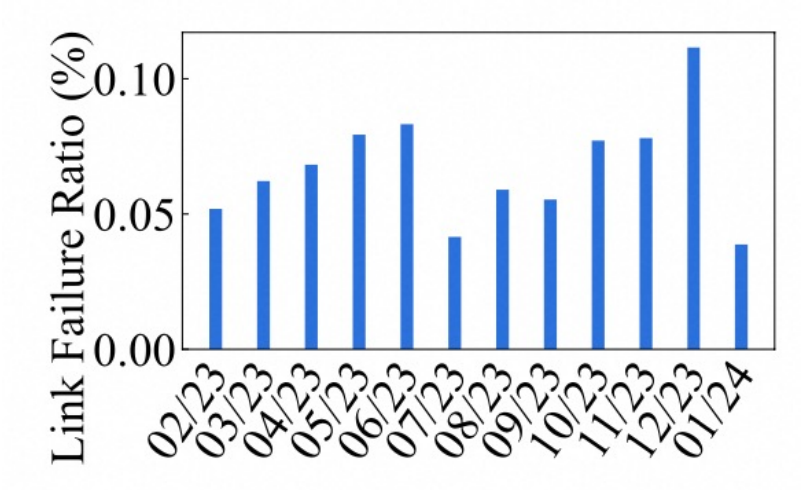
General cloud computing





LLM training

LLM Training is Sensitive to Faults



In each month, 0.057% of links fail.
In each month, 0.051% of switches encounter critical errors and crashes.
Lots of link issues happen every day due to optics failures.

Single-point failures would cause the entire LLM training crash.

Limitations of Traditional Clos Topology

Performance:

Hash polarization.

Reliability:

Single-point failure.

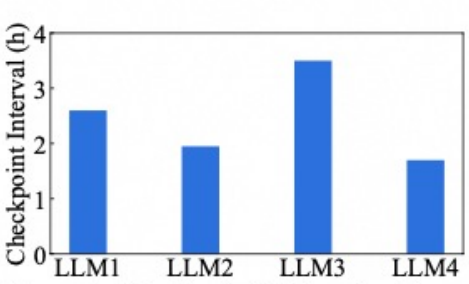
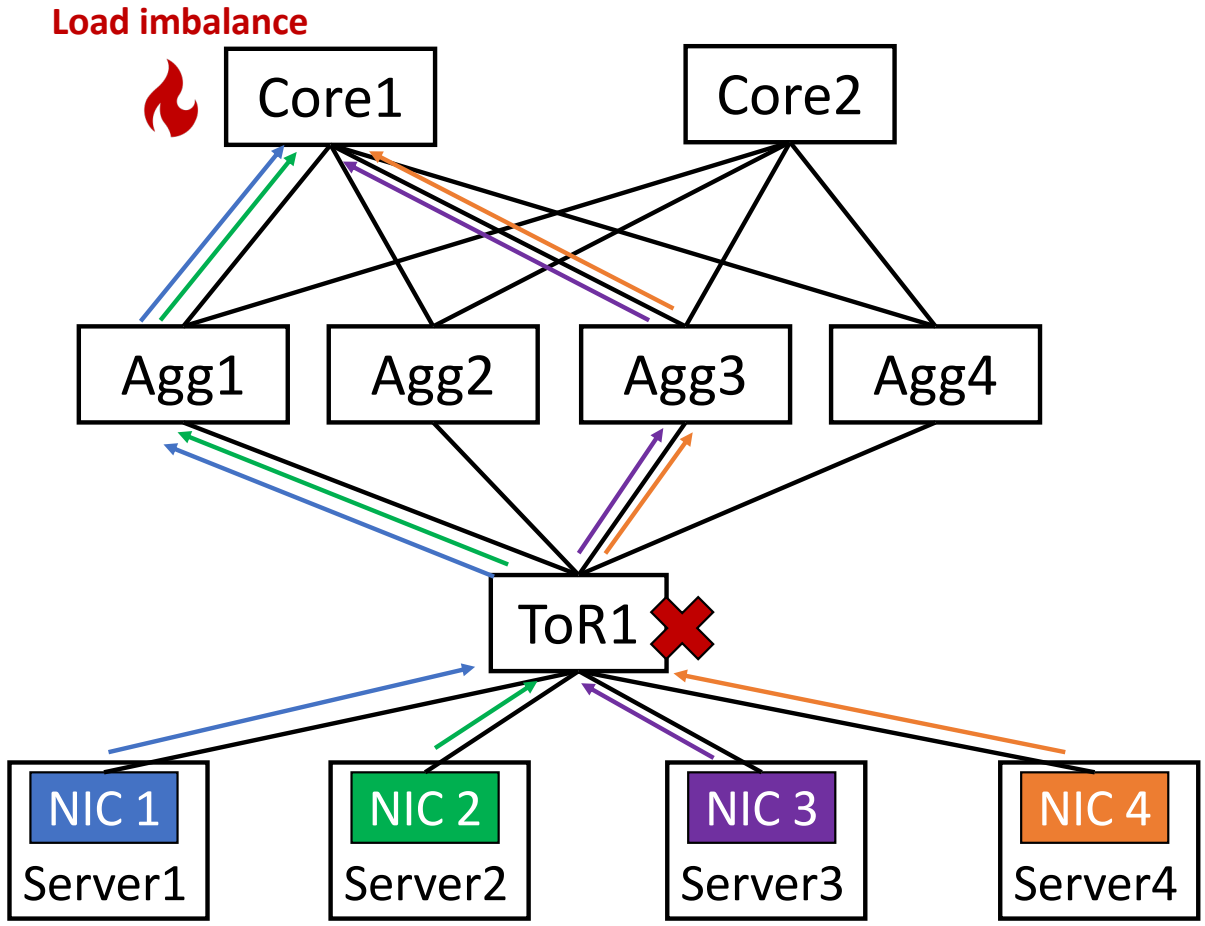


Figure 4: Checkpoint intervals of representative LLM jobs.



Our Goals

Scalability:

Primary capacity goal: interconnecting 15K GPUs.

Advanced capacity goal: interconnecting 100K GPUs.

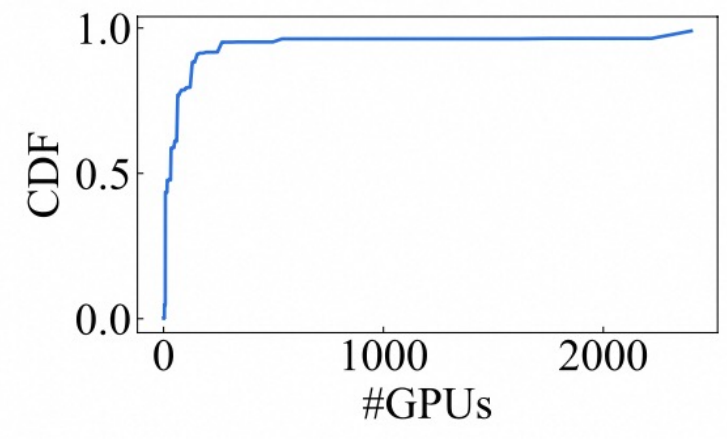
Performance:

Low latency.

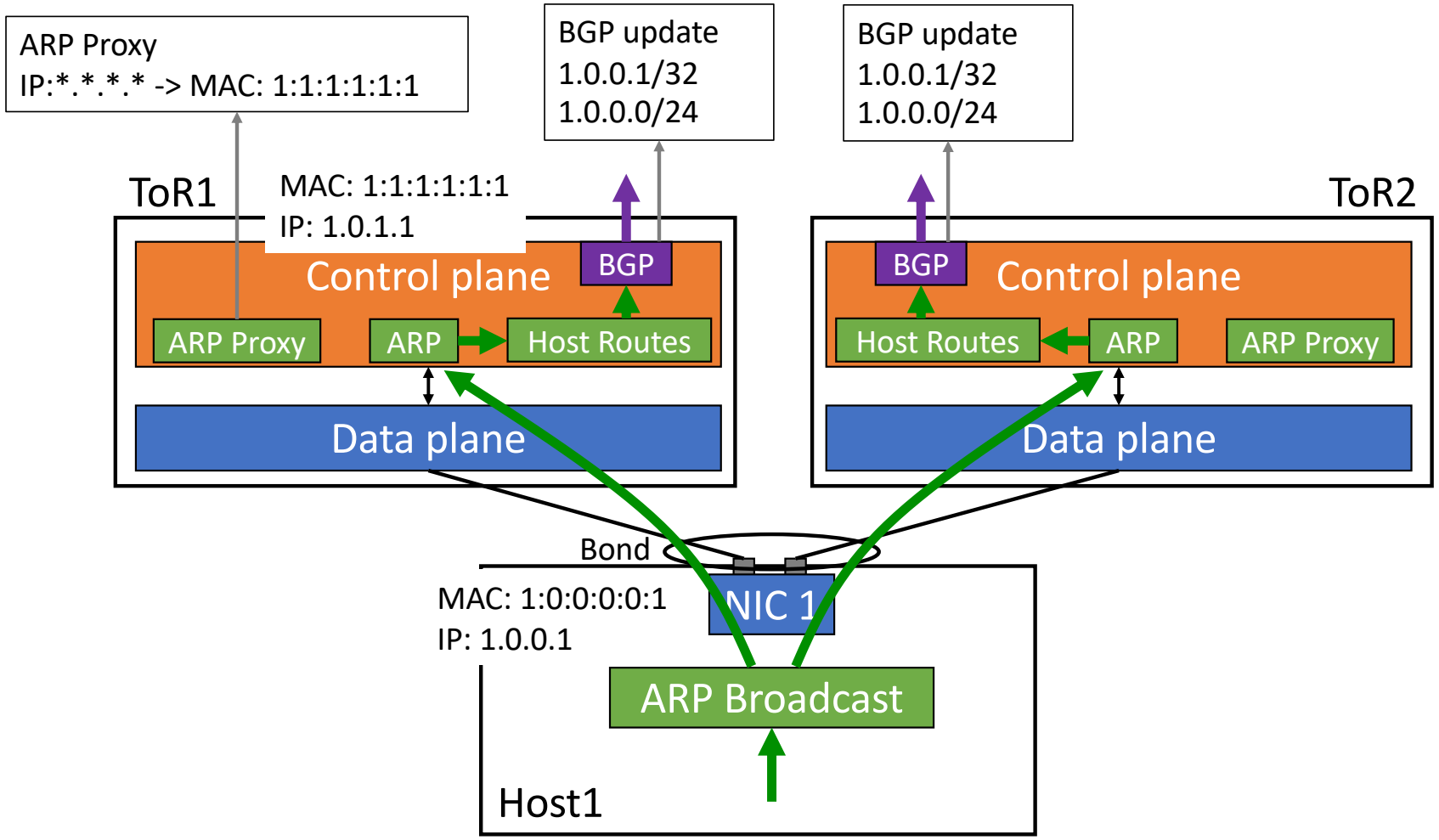
Maximizing network utilization.

Reliability:

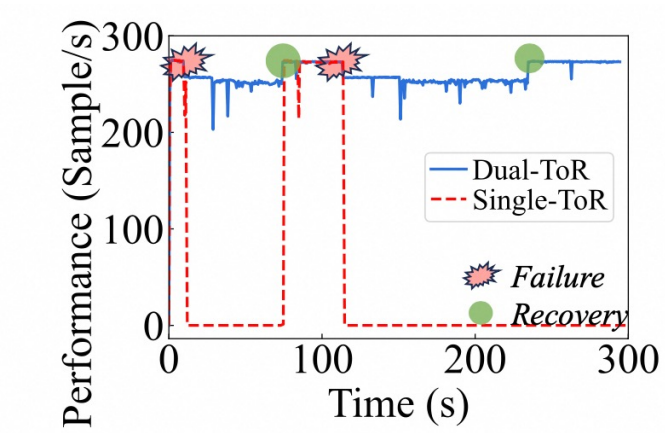
Eliminating single-point failure risk in the network.



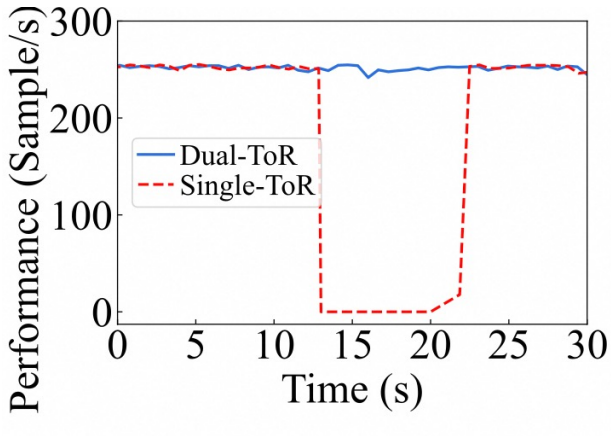
Access: Non-Stacked Dual-ToR



Handling Single-ToR Failure



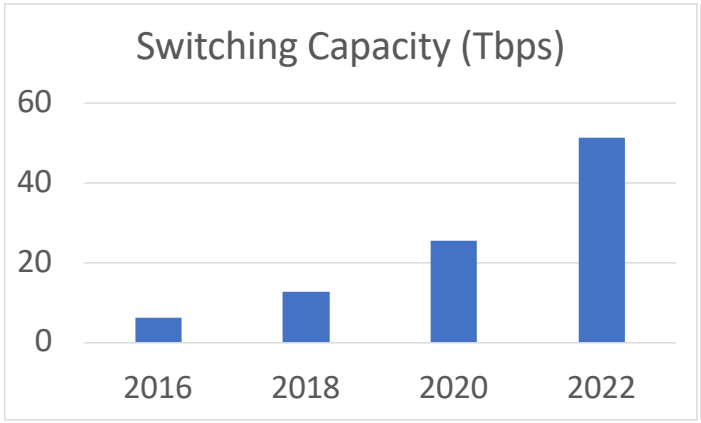
Link failure

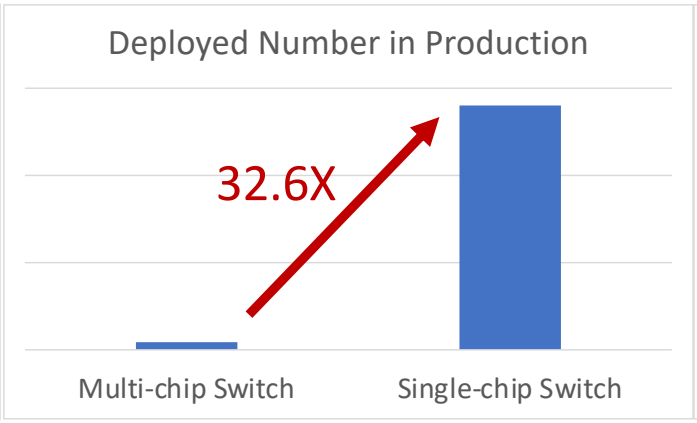


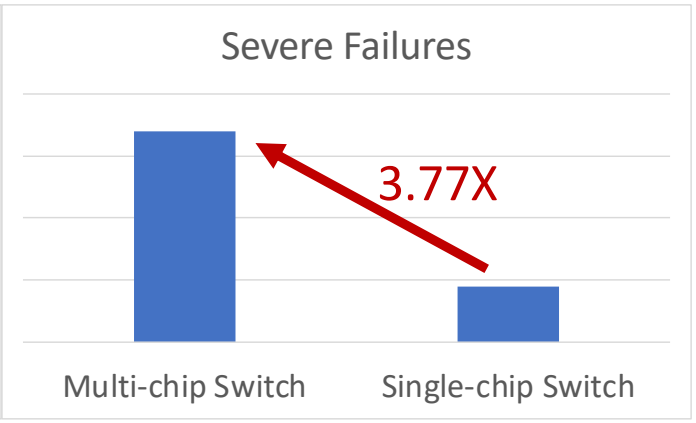
Link flapping

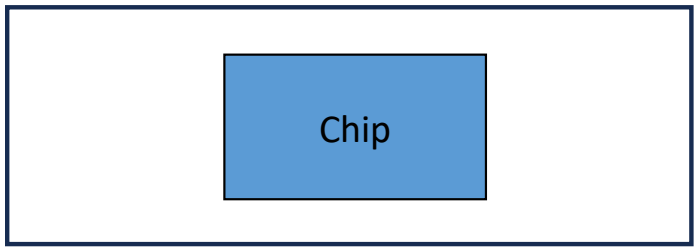
Tier1: 1K GPUs in a Segment

Employing latest single-chip switch

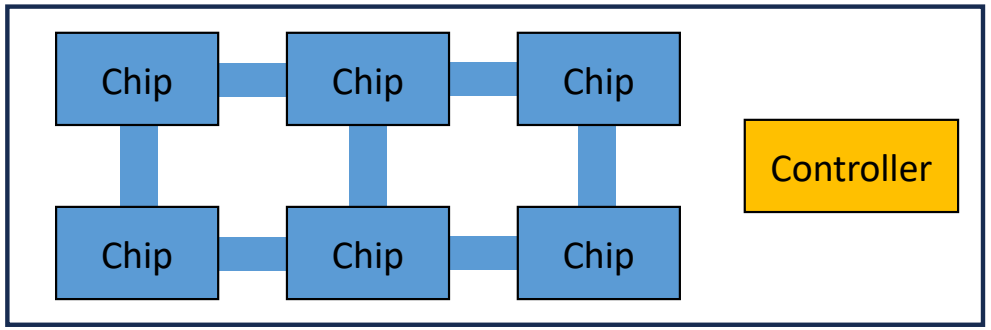








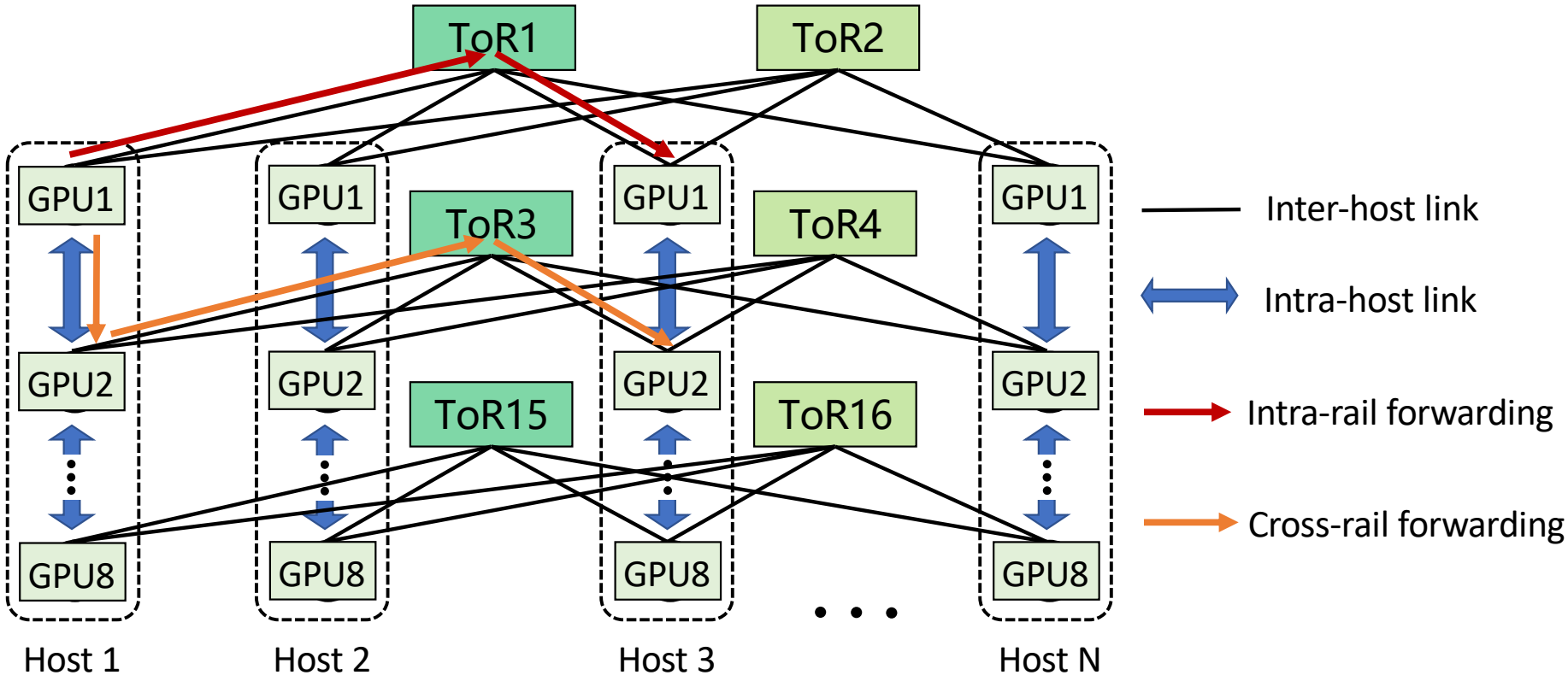
Single-chip switch



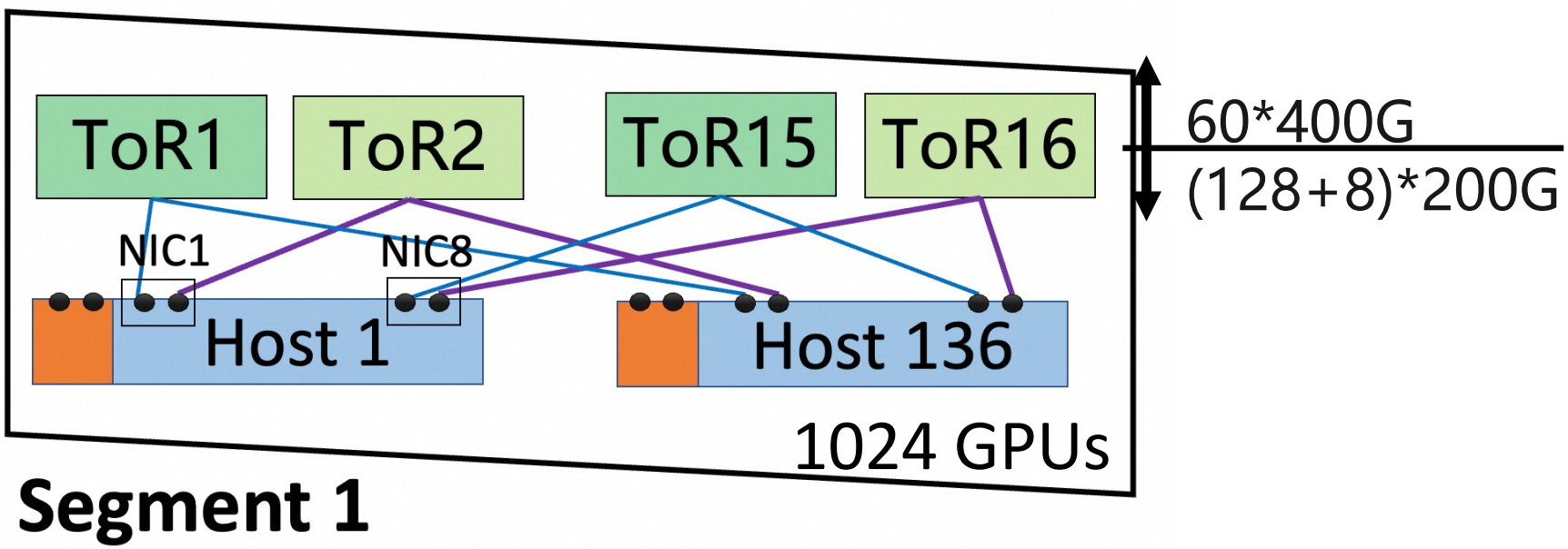
Multi-chip switch

Tier1: 1K GPUs in a Segment

Rail-optimized topology

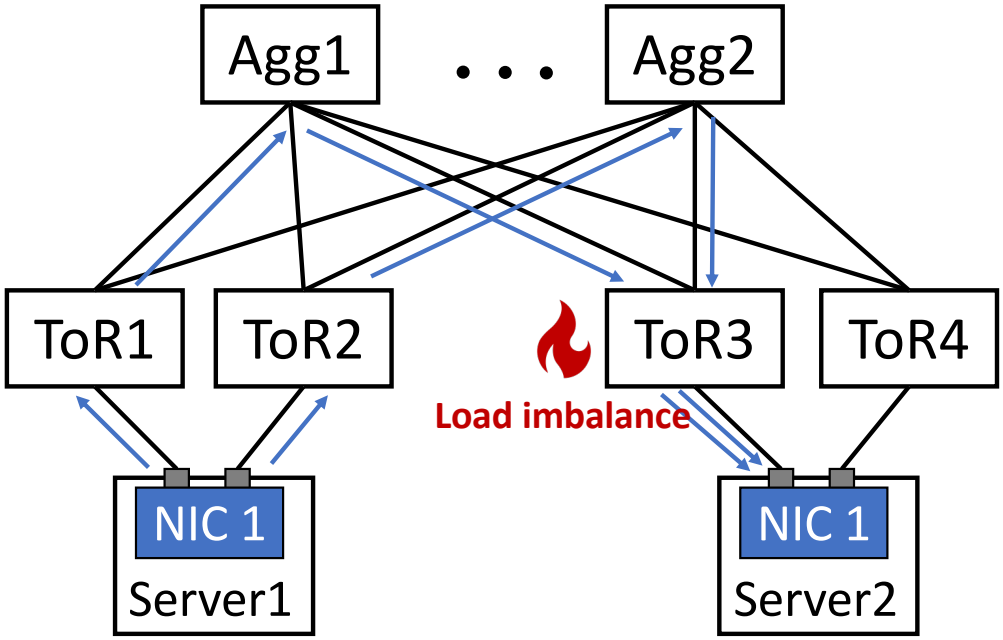


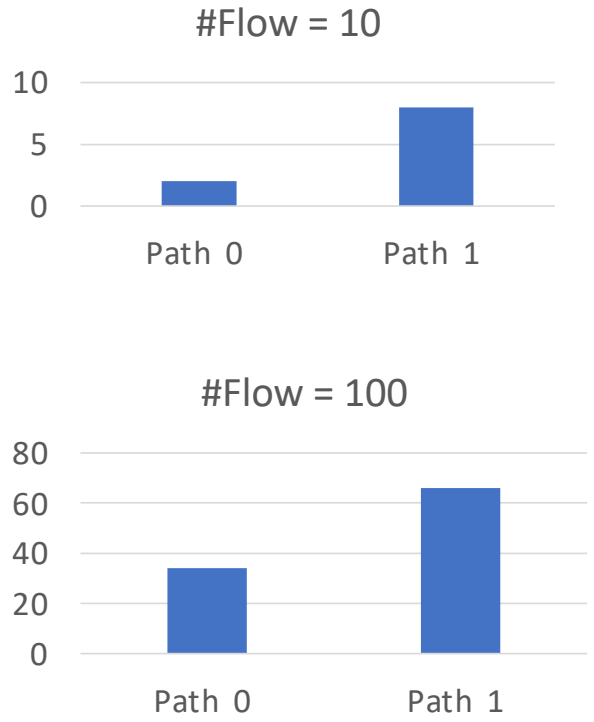
Tier1: 1K GPUs in a Segment



Tier2: Eliminating Load Imbalance

Load imbalance in 2-layer network



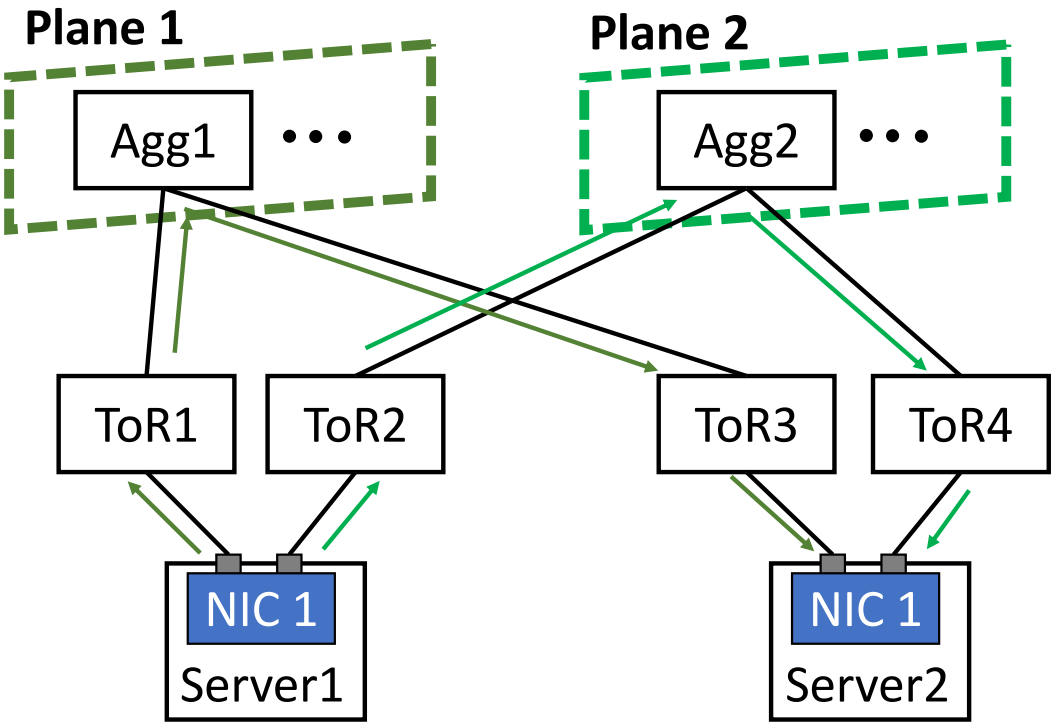


60

2

Tier2: Eliminating Load Imbalance

Dual plane



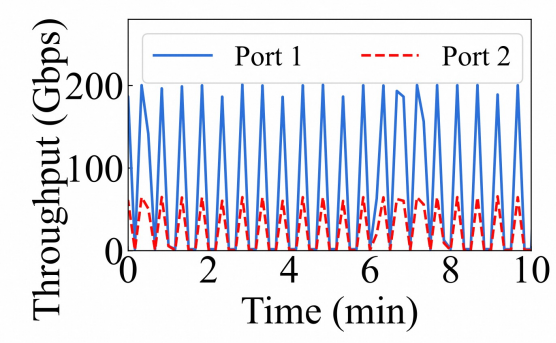
Tier2: Eliminating Load Imbalance

Eliminating hash-polarization with Dual-plane

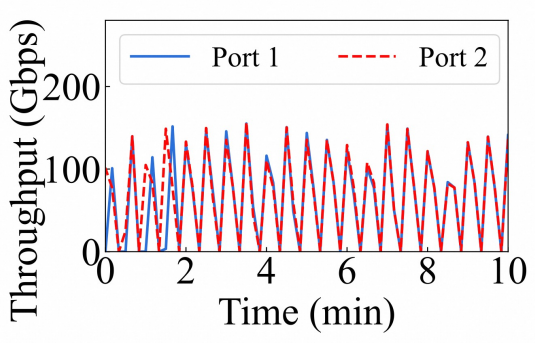
	Supported #GPUs	#Tiers	Switches participating load balance	Path selection complexity
Pod in HPN	15360	2	ToR	O(60)
SuperPod [21]†	16384	3	ToR+Aggregation+Core	O(32×32×4)=O(4096)
Jupiter [63]	26000 [12]	3	ToR+Aggregation	O(8×256)=O(2048)
Fat tree (k=48) [53]	27648	3	ToR+Aggregation	O(48×48)=O(2304)

Tier2: Eliminating Load Imbalance

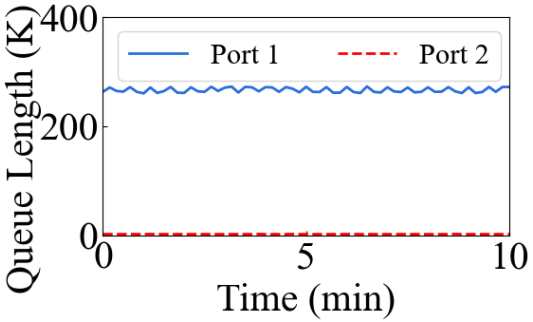
Eliminating hash-polarization with Dual-plane



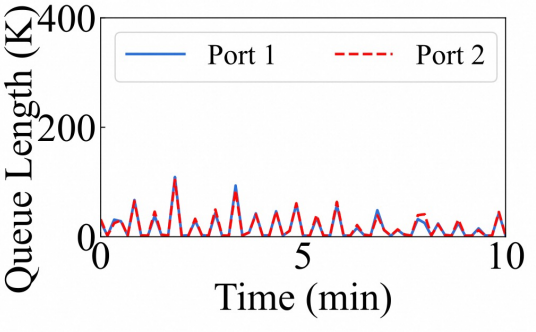
Typical Clos



Dual plane

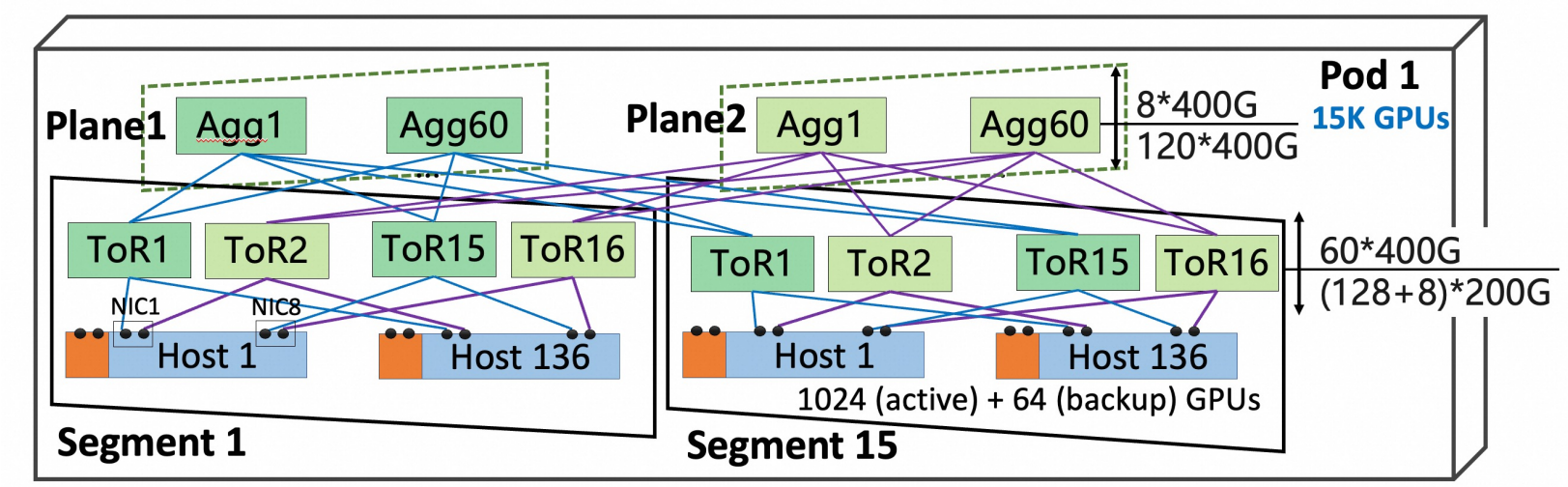


Typical Clos



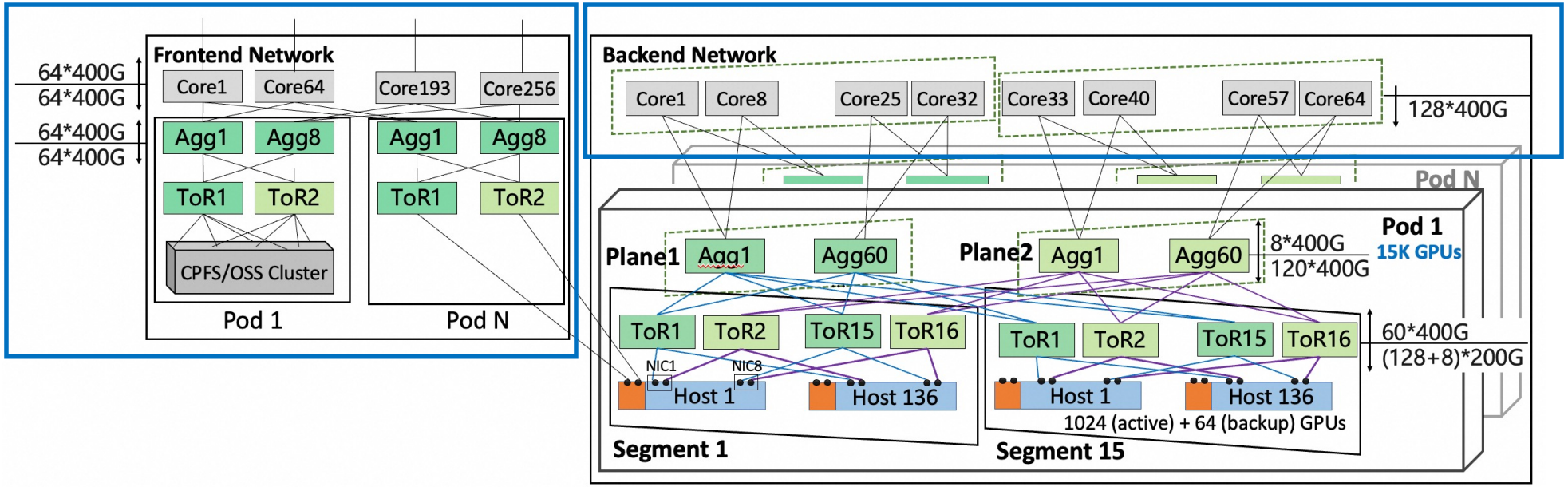
Dual plane

Tier2: 15K GPUs in a Pod

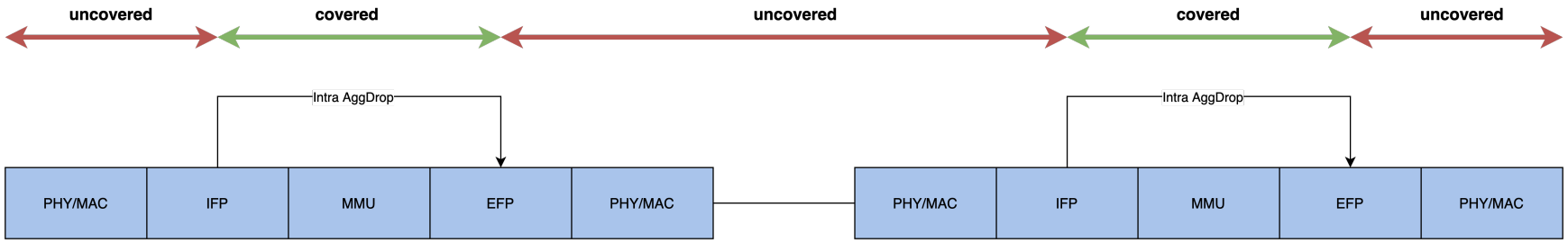


Alibaba HPN 7.0

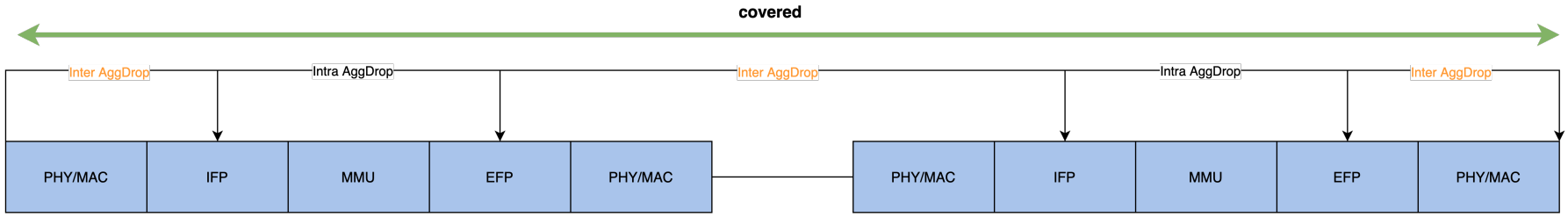
Tier3 connects multiple Pods together to interconnecting more GPUs.
Independent frontend network isolate storage traffic from training.



Packet Drop Detection

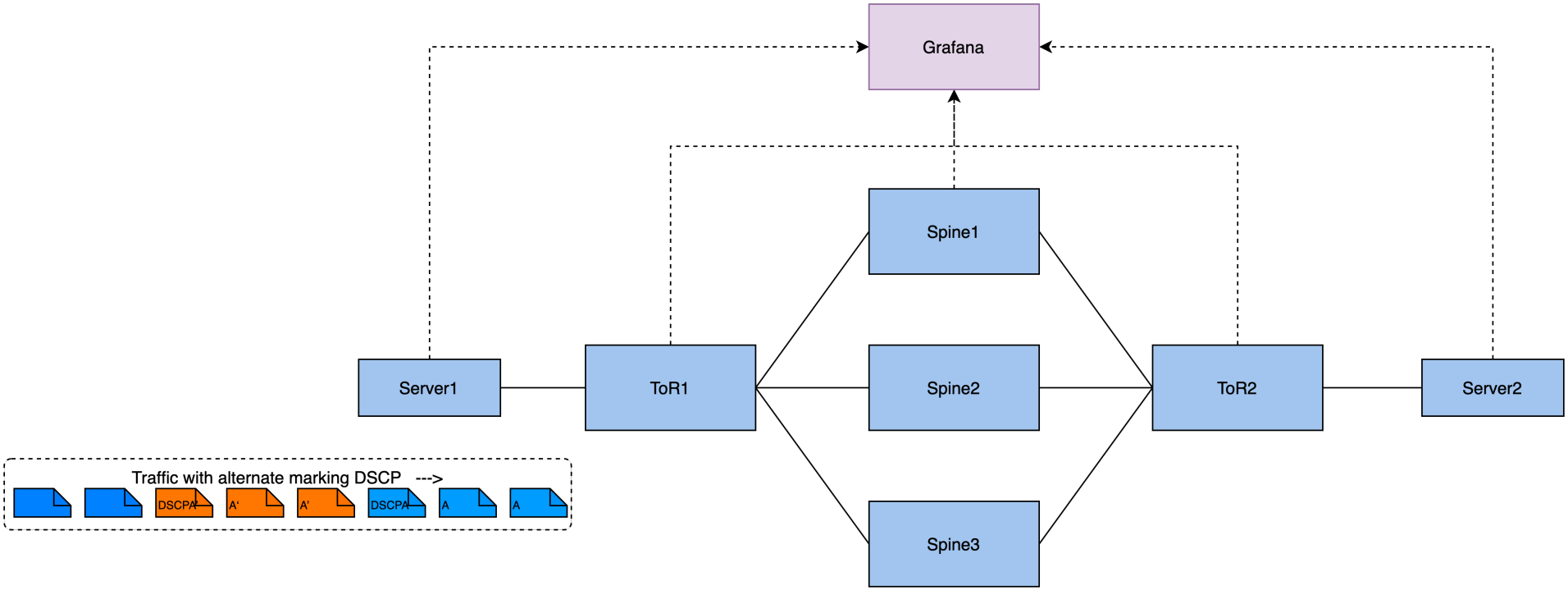


Normal Drop Detection’s limitation

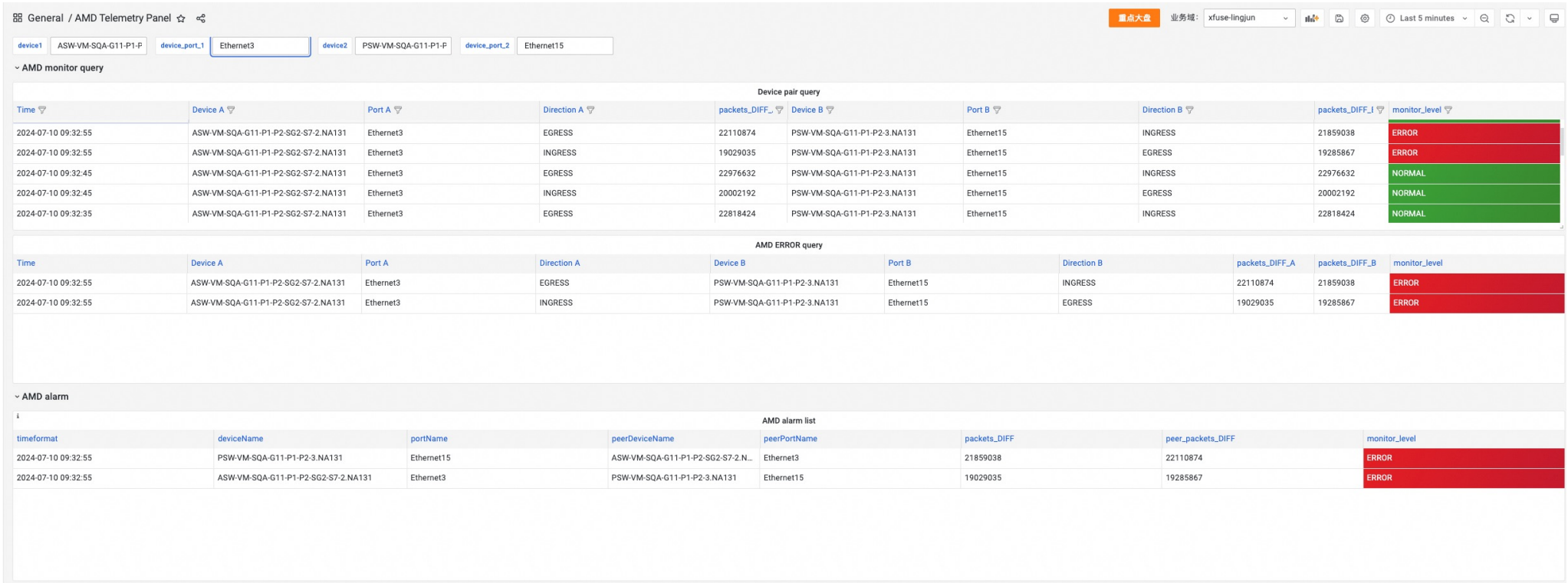


End to End Packet Drop Detection

Alternative Marking DSCP （A.M.D)



A.M.D Result


General / AMD Telemetry Panel
重点大盘
业务域: xfuse-lingjun
Last 5 minutes
device1 ASW-VM-SQA-G11-P1-P
device_port_1 Ethernet3
device2 PSW-VM-SQA-G11-P1-P
device_port_2 Ethernet15
AMD monitor query
Device pair query
Time | Device A | Port A | Direction A | packets_DIFF_... | Device B | Port B | Direction B | packets_DIFF_I | monitor_level
2024-07-10 09:32:55 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131 | Ethernet3 | EGRESS | 22110874 | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | INGRESS | 21859038 | ERROR
2024-07-10 09:32:55 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131 | Ethernet3 | INGRESS | 19029035 | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | EGRESS | 19285867 | ERROR
2024-07-10 09:32:45 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131 | Ethernet3 | EGRESS | 22976632 | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | INGRESS | 22976632 | NORMAL
2024-07-10 09:32:45 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131 | Ethernet3 | INGRESS | 20002192 | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | EGRESS | 20002192 | NORMAL
2024-07-10 09:32:35 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131 | Ethernet3 | EGRESS | 22818424 | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | INGRESS | 22818424 | NORMAL
AMD ERROR query
Time | Device A | Port A | Direction A | Device B | Port B | Direction B | packets_DIFF_A | packets_DIFF_B | monitor_level
2024-07-10 09:32:55 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131 | Ethernet3 | EGRESS | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | INGRESS | 22110874 | 21859038 | ERROR
2024-07-10 09:32:55 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131 | Ethernet3 | INGRESS | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | EGRESS | 19029035 | 19285867 | ERROR
AMD alarm
AMD alarm list
timeformat | deviceName | portName | peerDeviceName | peerPortName | packets_DIFF | peer_packets_DIFF | monitor_level
2024-07-10 09:32:55 | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.N... | Ethernet3 | 21859038 | 22110874 | ERROR
2024-07-10 09:32:55 | ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131 | Ethernet3 | PSW-VM-SQA-G11-P1-P2-3.NA131 | Ethernet15 | 19029035 | 19285867 | ERROR

Summary

More detail information could be found at

https://ennanzhai.github.io/pub/sigcomm24-hpn.pdf

Alibaba HPN: A Data Center Network for Large Language Model Training

Kun Qian, Yongqing Xi, Jiamin Cao, Jiaqi Gao, Yichi Xu, Yu Guan, Binzhang Fu, Xuemei Shi
Fangbo Zhu, Rui Miao, Chao Wang, Peng Wang, Pengcheng Zhang, Xianlong Zeng
Eddie Ruan, Zhiping Yao, Ennan Zhai, Dennis Cai
Alibaba Cloud

Abstract

This paper presents HPN, Alibaba Cloud's data center network for large language model (LLM) training. Due to the differences between LLMs and general cloud computing (*e.g.*, in terms of traffic patterns and fault tolerance), traditional data center networks are not well-suited for LLM training. LLM training produces a small number of periodic, bursty flows (*e.g.*, 400Gbps) on each host. This characteristic of LLM training predisposes Equal-Cost Multi-Path (ECMP) to hash polarization, causing issues such as uneven traffic distribution. HPN introduces a 2-tier, dual-plane architecture capable of interconnecting 15K GPUs within one Pod, typically accommodated by the traditional 3-tier Clos architecture. Such a new architecture design not only avoids hash polarization but also greatly reduces the search space for path selection. Another challenge in LLM training is that its requirement for GPUs to complete iterations in synchronization makes it more sensitive to single-point failure (typically occurring on ToR). HPN proposes a new dual-ToR design to replace the single-ToR in traditional data center networks. HPN has been deployed in our production for more than eight months. We share our experience in designing, and building HPN, as well as the operational lessons of HPN in production.

'24), August 4–8, 2024, Sydney, NSW, Australia. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3651890.3672265

1 Introduction

The large language model (or LLM) [13, 15–17, 24] has brought about tremendous revolutions to today's AI and cloud services. The training of an LLM, which has hundreds of billions of parameters, relies on a large-scale distributed training cluster, typically equipped with tens of millions of GPUs. Due to its unique characteristics, LLM training presents new challenges to the design of data center networks.

Problem 1: Traffic patterns. The traffic patterns of LLM training are different from those of general cloud computing in terms of (1) low entropy [19, 22] and (2) bursty traffic. Specifically, general cloud computing generates millions of flows, which gives the network high entropy. Each flow is continuous and low-utilization (*e.g.*, typically below 20% of the NIC capacity), as shown in Figure 1. On the contrary, LLM training generates very few but periodically bursty flows, resulting in low entropy and high utilization for the network. The burst can directly reach the NIC capacity, which is

阿里云

Thank you