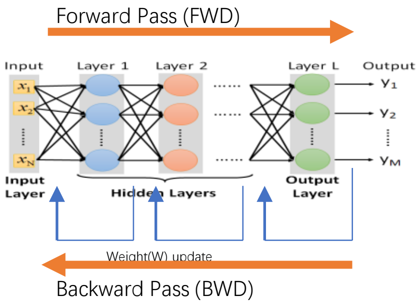
AICN Clarification

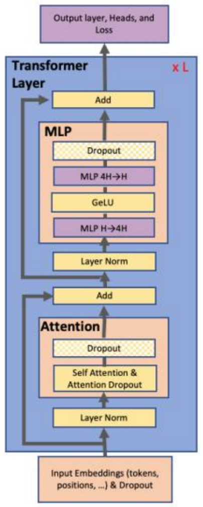
Lily Lyu

November Plenary 2024

AICN：Connecting Accelerators for AI Training

Neural Network

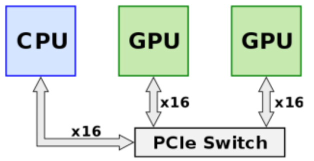



Output layer, Heads, and Loss
Transformer Layer
x L
Add
MLP
Dropout
MLP 4H→H
GeLU
MLP H→4H
Layer Norm
Add
Attention
Dropout
Self Attention & Attention Dropout
Layer Norm
Input Embeddings (tokens, positions, …) & Dropout

This is AI model, not the 'network' we are talking about. But it impacts the network development.

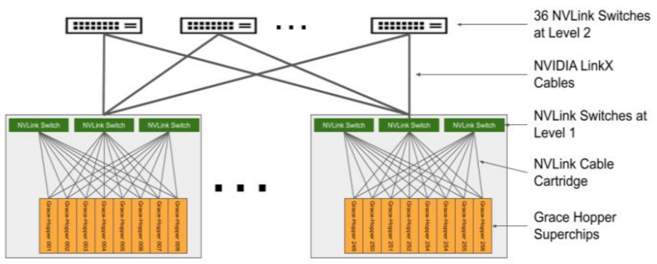
Scale-up

Bus technology evolves, trying to connect more GPUs in bus domain.



https://en.wikichip.org/wiki/nvidia/nvlink

Fully Connected NVLink across 256 GPUs



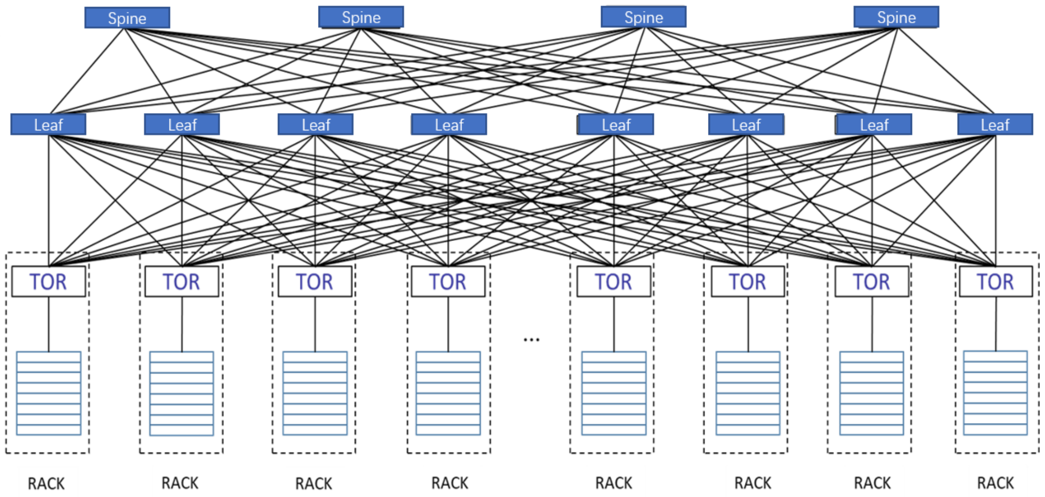
https://developer.nvidia.com/blog/announcing-nvidia-dgx-gh200-first-100-terabyte-gpu-memory-system/

- PCIe, NVLink, CXL …
  - Ultra high bandwidth: e.g NVLink5.0 is a 1.8TB/s bidirectional, direct GPU-to-GPU interconnect
- Server-scale -> rack-scale ->pod-scale
  - 10 -> 1000 GPUs

Scale-out **AICN focus**

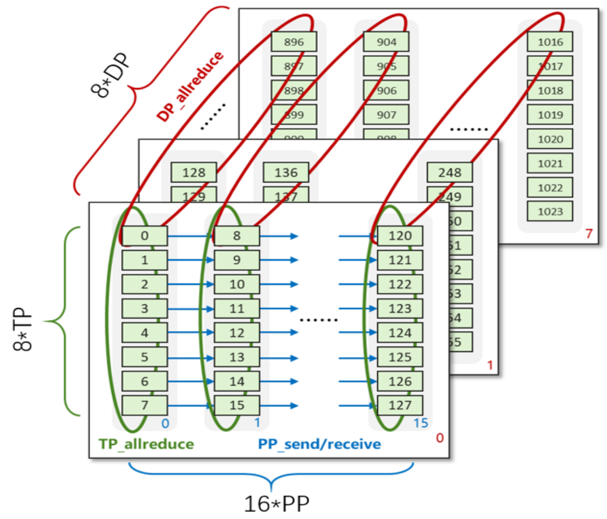
Network technology evolves, trying to improve performance (reliability, latency, throughput)

- Infiniband, Ethernet (RoCEv2) , …
  - High bandwidth: 800GE->1.6TGE
- Pod-scale -> across DC scale
  - Towards 10K+ GPUs



3D Parallelism Deployment

3D Parallelism

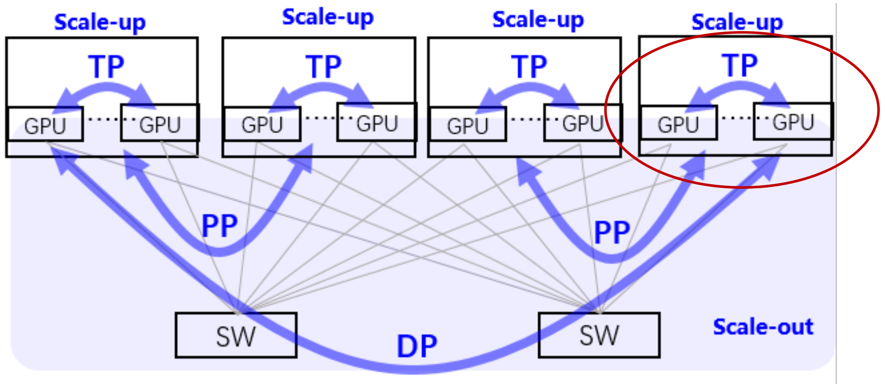


GPT-3 example:

- L=96(layer number), h=12288(hidden dimension), b=1536(global batch size)，s=2048(sequence length)
- T=8, P=8, D=16 (totally 8*8*16=1024 GPUs)
- AllReduce = reduce scatter + all gather, introduces 2 times traffic amount
- 2 bytes for each parameter

	Collective communication		GPU Traffic amount/time	Times/iteration	GPU traffic amount/iteration
DP	AllReduce	MLP	4h*h*2/T/D * 2(D-1) * 2byte = 540MB	L/P=12	12*(540+270) = **9.49GB**
		Attention	4h*h/T/D *2(D-1) * 2byte = 270MB	L/P=12	
PP	Send/Receive	Transformer	b/D * s * h * 2byte = 4.5GB	2	4.5*2 = **9GB**
TP	AllReduce	MLP	b/D*s*h/T * 2(T-1) * 2byte = 7.875GB	2 * L/P =24	7.875*24*2= **378GB**
		Attention	b/D*s*h/T * 2(T-1) * 2byte = 7.875GB	2 * L/P =24	

High data volume, deployed in scale-up domain



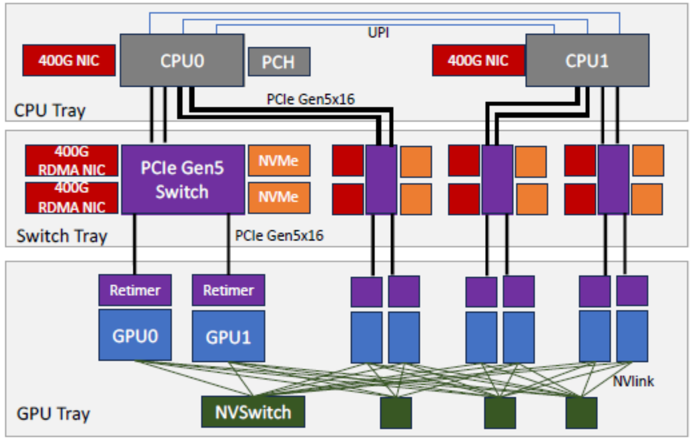
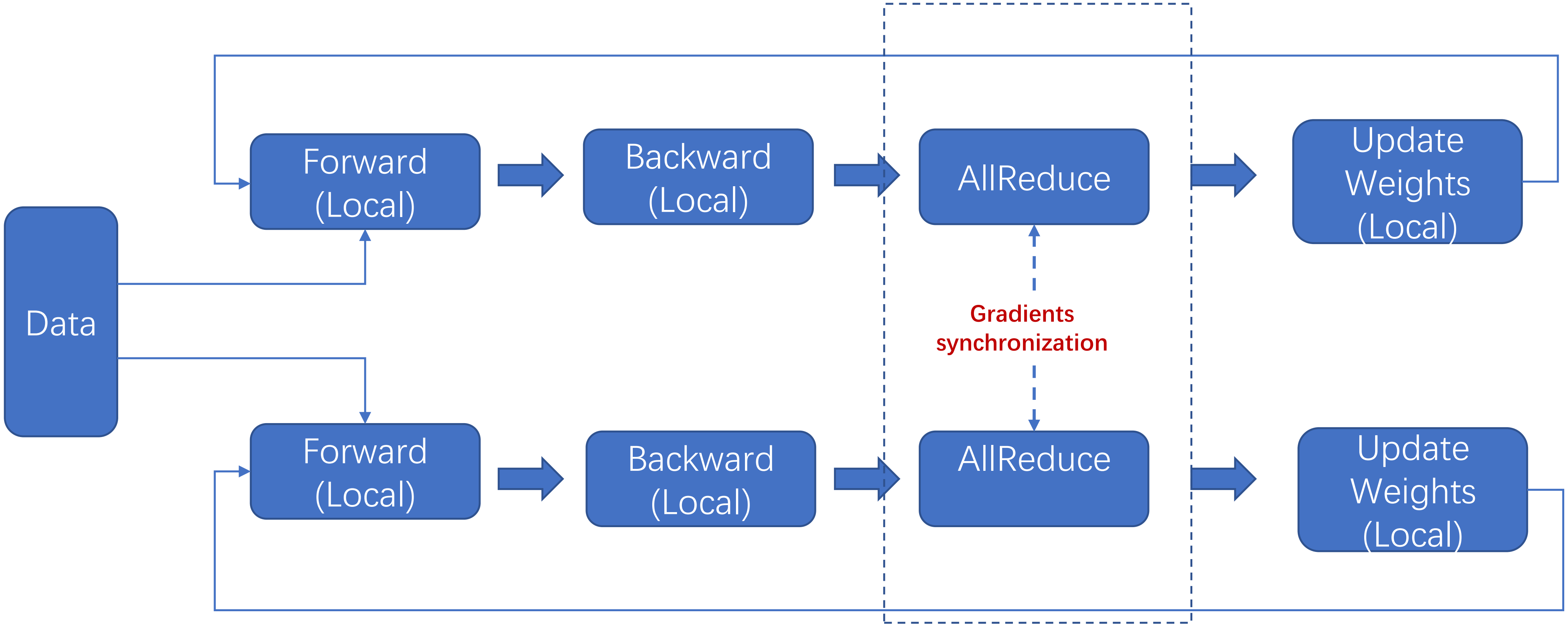


Figure 4: Grand Teton platform

1:1 mapping between GPUs and NICs.

Sigcomm2024: RDMA over Ethernet for Distributed AI Training at Meta Scale

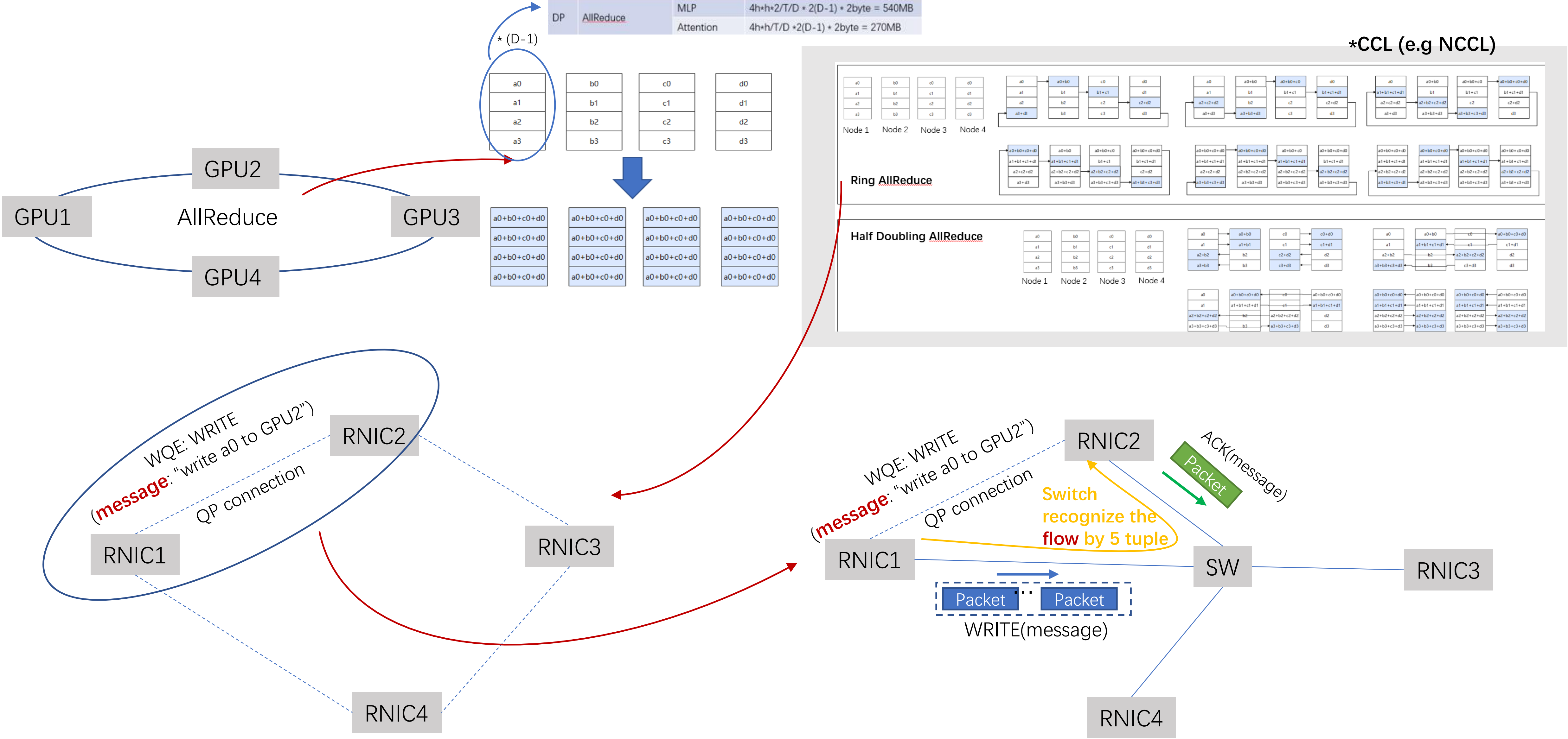
DP Traffic



The forward pass completes on each of the ranks followed by the backward pass.
During the backward pass, gradients are synchronized using collective communication AllReduce.

DP Traffic

	Collective communication		GPU Traffic amount/time
DP	AllReduce	MLP	4h*h*2/T/D * 2(D-1) * 2byte = 540MB
		Attention	4h*h/T/D *2(D-1) * 2byte = 270MB



a0	b0	c0	d0
a1	b1	c1	d1
a2	b2	c2	d2
a3	b3	c3	d3

a0+b0+c0+d0	a0+b0+c0+d0	a0+b0+c0+d0	a0+b0+c0+d0
a0+b0+c0+d0	a0+b0+c0+d0	a0+b0+c0+d0	a0+b0+c0+d0
a0+b0+c0+d0	a0+b0+c0+d0	a0+b0+c0+d0	a0+b0+c0+d0
a0+b0+c0+d0	a0+b0+c0+d0	a0+b0+c0+d0	a0+b0+c0+d0

DP Traffic Pattern

Burstiness

Gradients synchronization happens just during backward pass. It repeats iteration by iteration, involving multiple GPUs' communication at the same time.

ON-OFF pattern

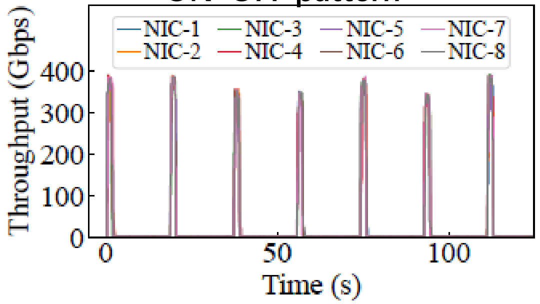


Figure 2: NIC egress traffic pattern during production model training.

Sigcomm2024: Alibaba HPN: A Data Center Network for Large Language Model Training

"On the time dimension, the flows usually exhibit the "on and off" nature in the time granularity of milliseconds"

Sigcomm2024: RDMA over Ethernet for Distributed AI Training at Meta Scale

Low Entropy

Regular communication between predictable communication pairs.

"a general cloud computing instance typically generates hundreds of thousands of connections; on the contrary, each node in the LLM training generates very few connections."

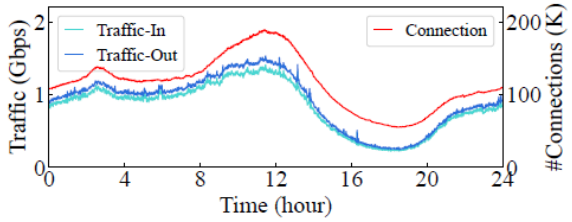


Figure 1: Traditional cloud computing traffic pattern.

VS.

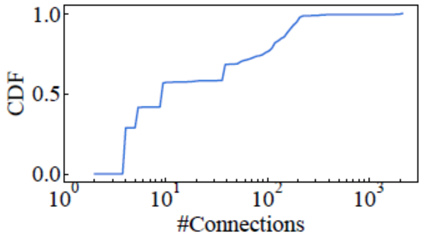


Figure 3: Number of connections per host.

Sigcomm2024: Alibaba HPN: A Data Center Network for Large Language Model Training

"Flow entropy per NIC is log(M). M is number of channels used in NCCL."

Sigcomm2024: RDMA over Ethernet for Distributed AI Training at Meta Scale

Elephant Flows

Even no clear definition, >1MB is usually considered as elephant flow.

"flows receiving more than 1048555 bytes (approximately 1 MB) will be identified as elephant flows. An elephant flow will time out from the ETRAP flow table if it has less than 500 bytes data over a period of 500 microseconds."

https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-738488.html

Few connection & huge data volume → >1MB flow is common

"For each burst, the intensity of each flow could reach up to the line rate of NICs."

Sigcomm2024: RDMA over Ethernet for Distributed AI Training at Meta Scale

Thanks!