SkeletonHunter: Diagnosing and Localizing Network Failures in Containerized Large Model Training

Wei Liu, Kun Qian, Zhenhua Li, Tianyin Xu, Yunhao Liu, Weicheng Wang,
Yun Zhang, Jiakang Li, Shuhong Zhu, Xue Li, Hongfei Xu, Fei Feng, Ennan Zhai

清華大學 Tsinghua University

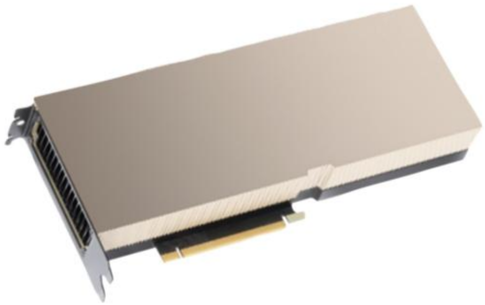
Alibaba Cloud

University of Illinois Urbana-Champaign

1. Background

Large model training is an important business for CSPs

- Large models are typically trained with
  - Significant infrastructure support
  - Hundreds of thousands of GPUs
  - Several weeks



O(1000)× High-end GPU

O(1000)×RDMA NIC (RNIC)

NVIDIA NVLINK

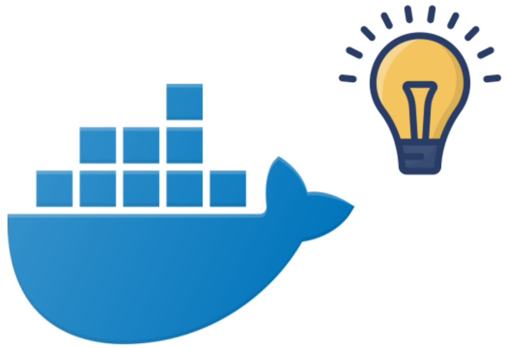
PCI EXPRESS

High-Speed, Reliable Interconnections

1. Background

Large model training is an important business for CSPs

- Large model training can mainly be launched by



Physical clusters

😄 **Highly customizable**, but…

🙁 Requires **professional experiences**

🙁 **Not flexible** enough

🙁 **High operational cost**

Container clusters

😄 **High flexibility**

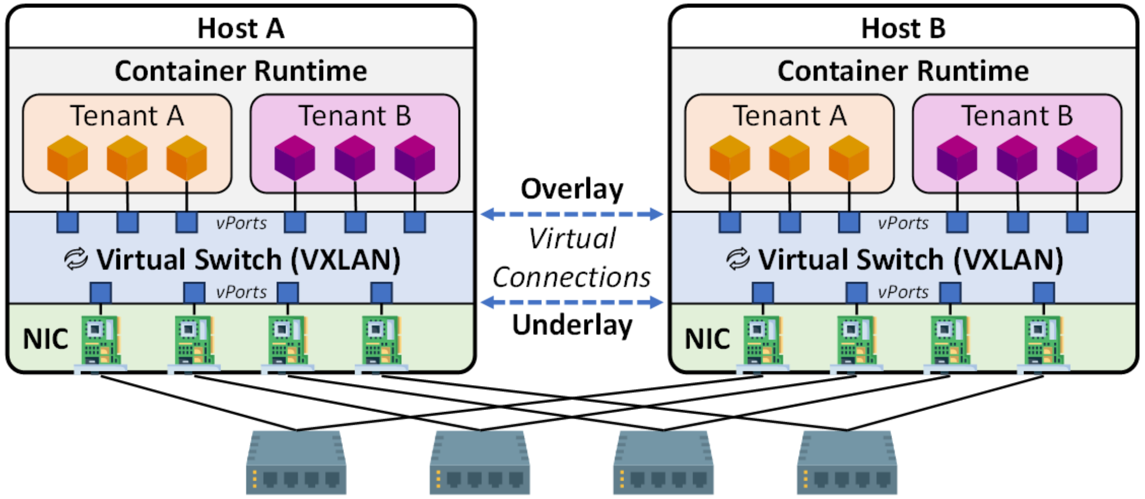
😄 **Easy to use**

😄 **Cost-efficient (on demand)**

1. Background

As a major CSP, we have…

- A large-scale, **multi-tenant** large model training cloud
- **40K+** RNICs, and **40K+** GPUs
- Operating over **3** years
- Serving **5M+** large model training tasks from users



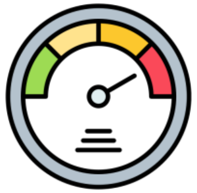
Data Center Topologies (Alibaba HPN 7.0)

2. Motivation

The reliability of containerized model training is crucial

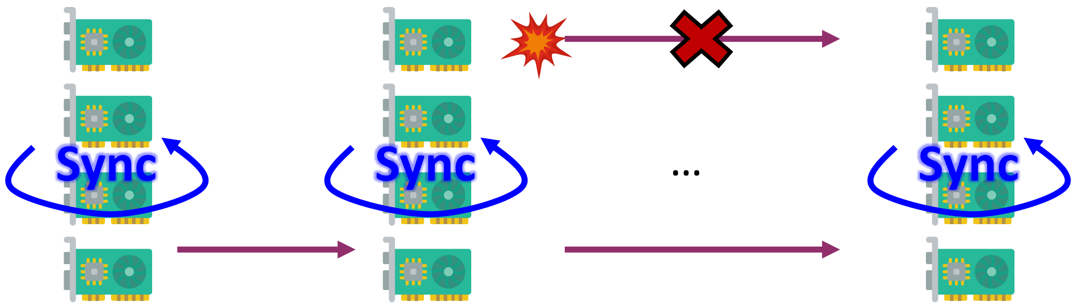
- Training nodes' GPUs are inter-connected
- Low-latency, high-bandwidth networks

RTT <20 us



Throughput >200 Gbps

- Training process is highly **synchronized**
- Sensitive to **single-point network failure**



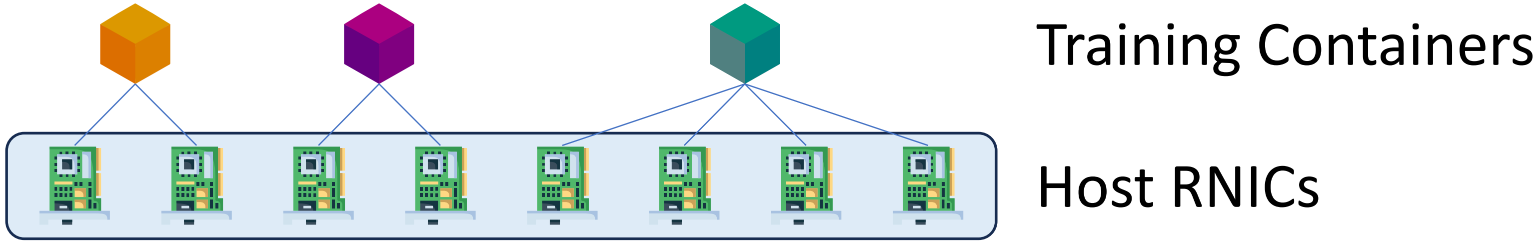
Financial loss to customers

2. Motivation

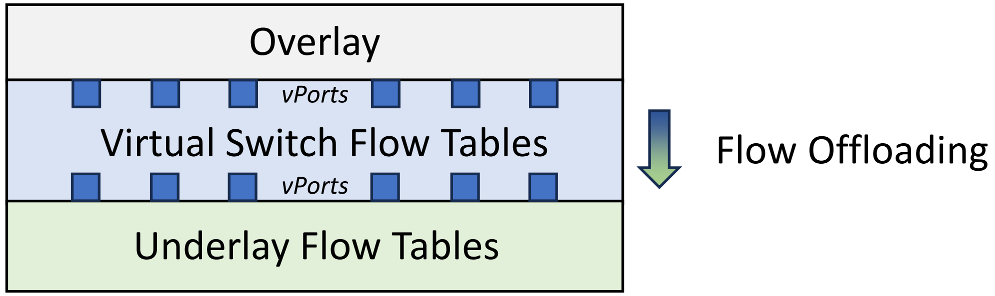
Pinpointing network connectivity issues is not easy

- **High dynamics of containers**

- **Endpoint-induced complexity**



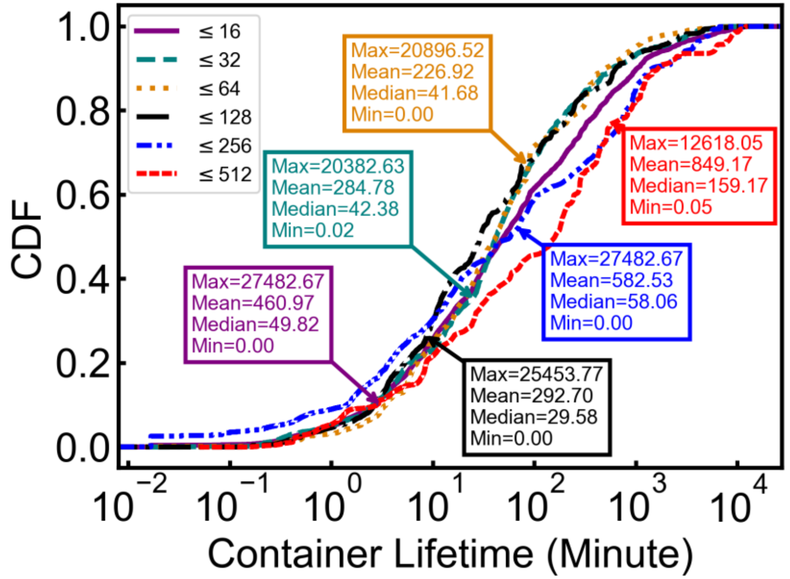
- **Interplay between overlay and underlay networks**

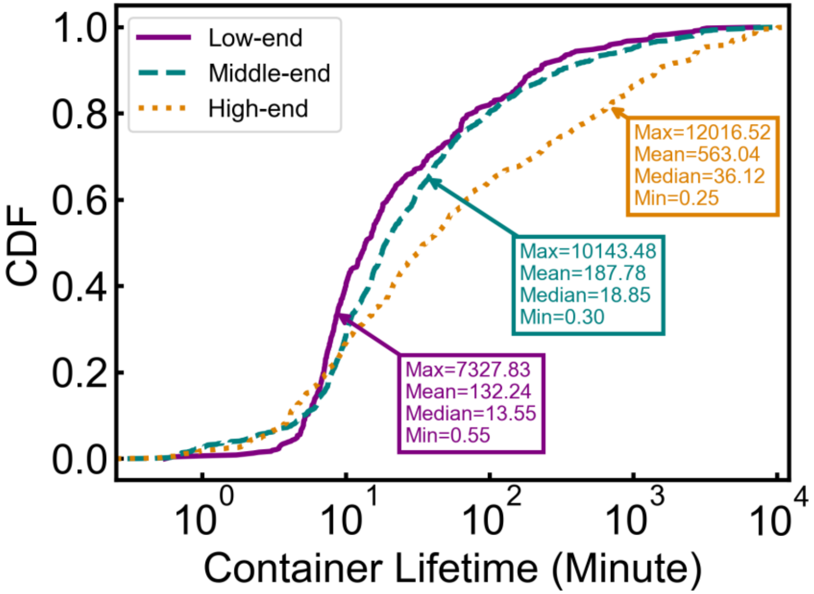


2. Motivation

High dynamics of containers

- Over **50%** of training containers have a lifetime of less than **60 minutes**
- Containers with **higher-end** configurations have a **longer lifetime**



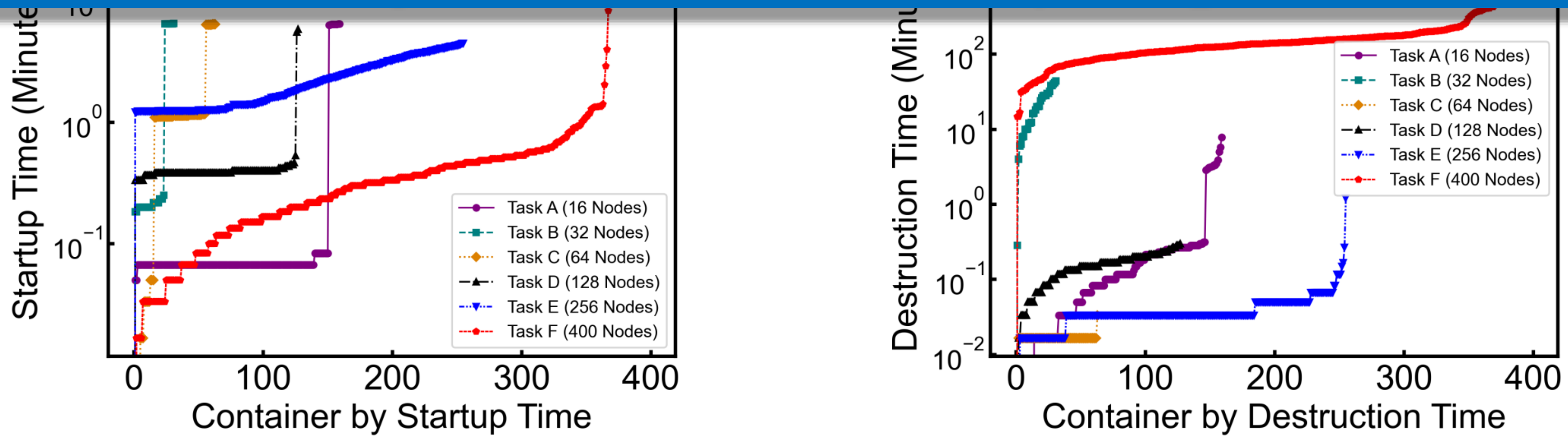


2. Motivation

- **High dynamics of containers**

Challenge 1

Requiring **fast** connectivity probing on the **highly dynamic** network topologies

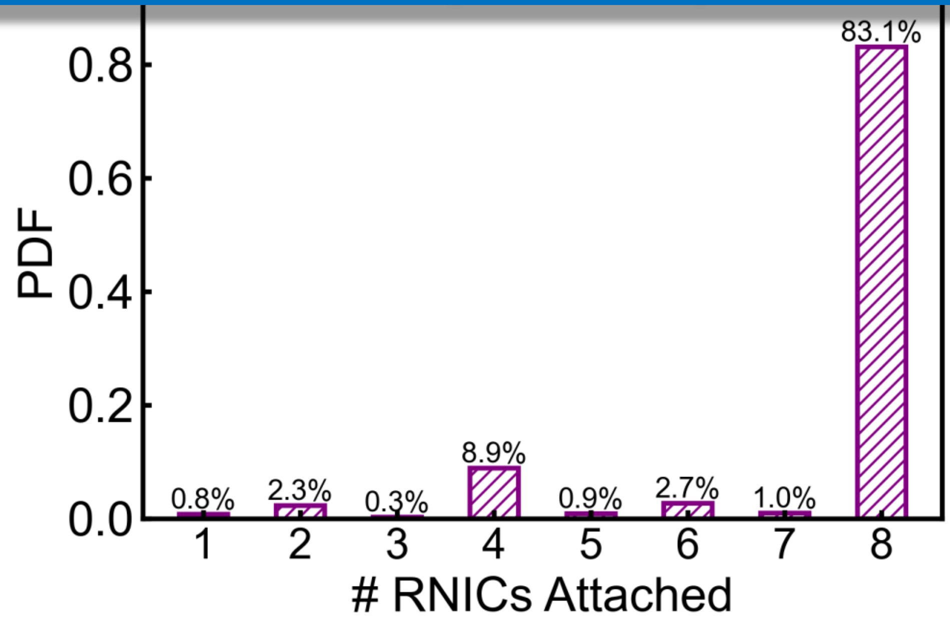


2. Motivation

Endpoint-induced complexity

Challenge 2

Requiring efficient **coverage** of the endpoint-induced complexity

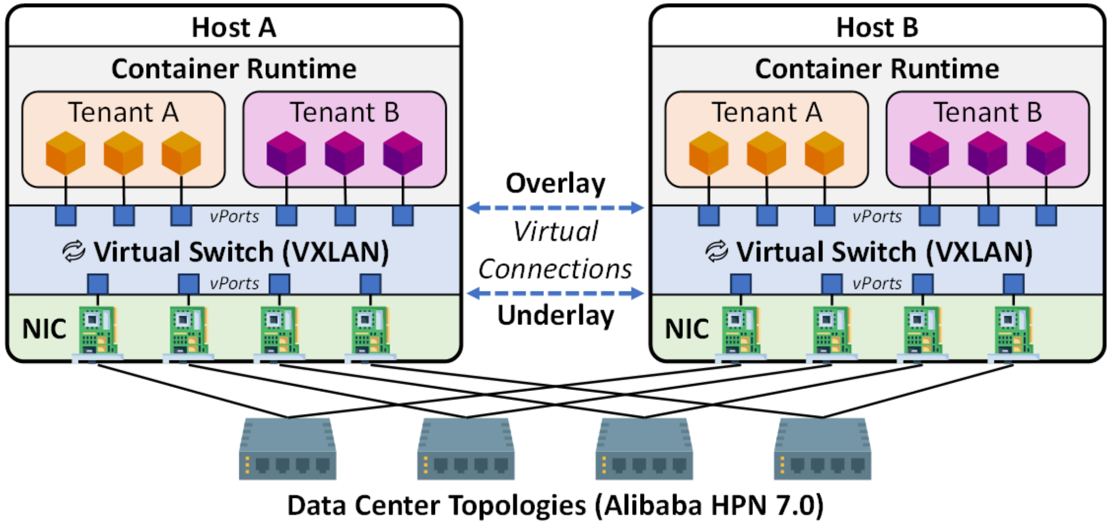


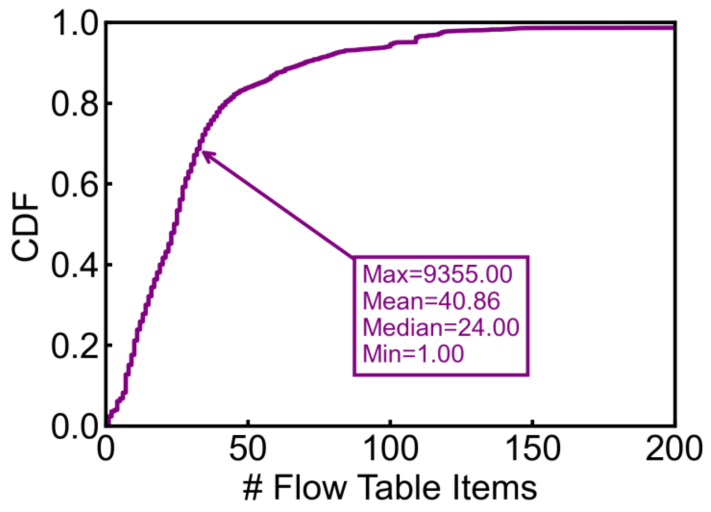
2. Motivation

Interplay between overlay and underlay networks

Challenge 3

Requiring **effective disentanglement** of the overlay-underlay **interplay**





2. Motivation

Multiplicative effect of the complexity

- X containers
- Y RNICs for each container
- Z virtual network components for each RNIC

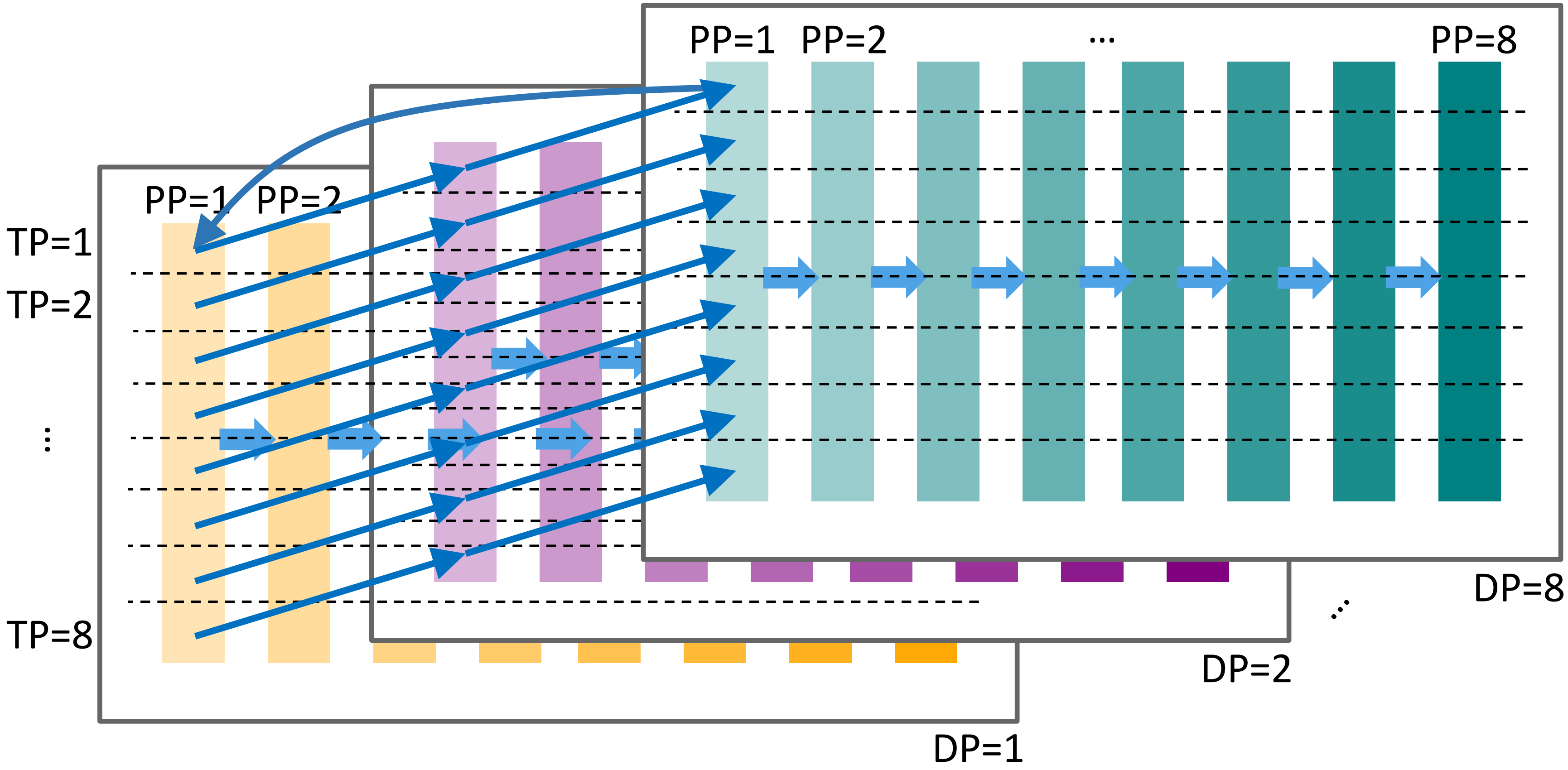
Examining **X×Y×Z** network components in each training round!

e.g., **1K×8 ×16=128K**

2. Motivation

Opportunity——Sparse spatial traffic distributions

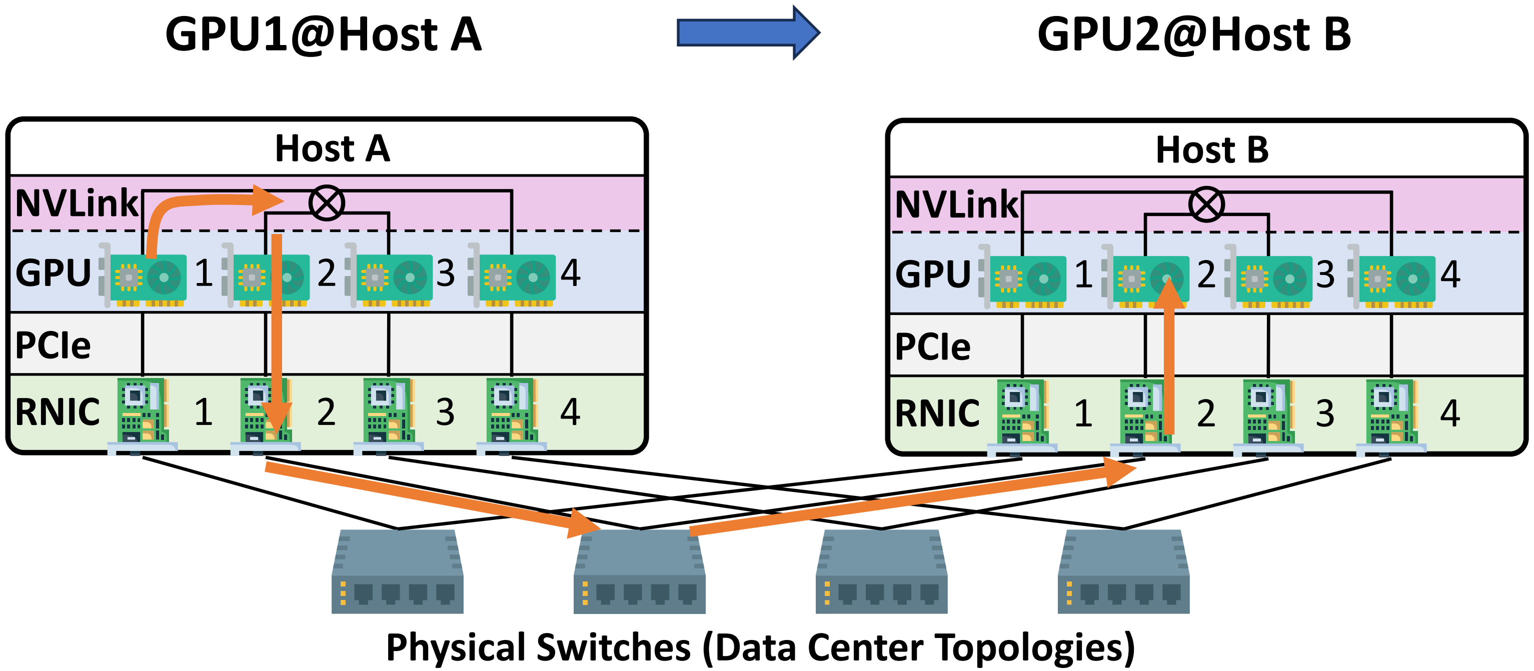
- Training data are only exchanged cross the GPUs with **dependencies**
- Derived from various parallelism strategies



2. Motivation

Opportunity——Sparse spatial traffic distributions

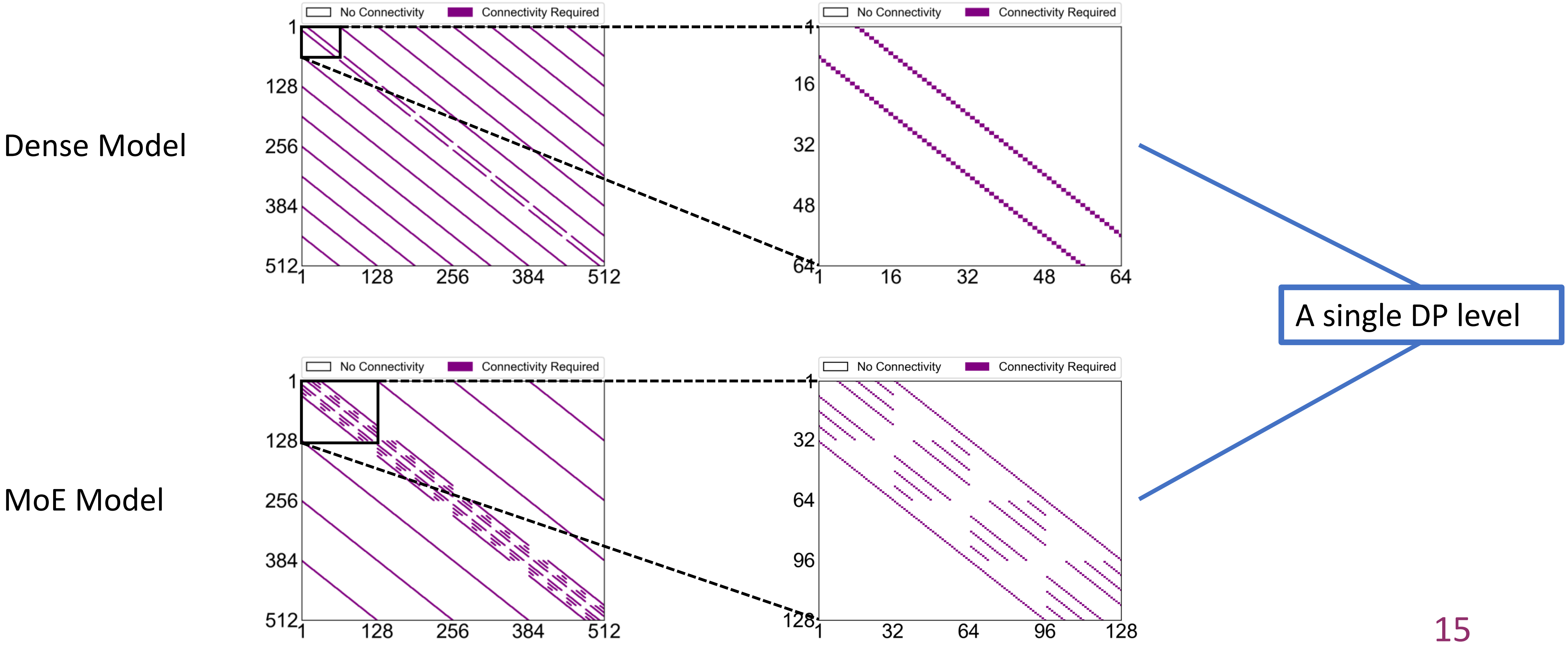
- Rail-optimized data center topologies
- Widely used collective communication libraries like NCCL and MPI



2. Motivation

Opportunity——Sparse spatial traffic distributions

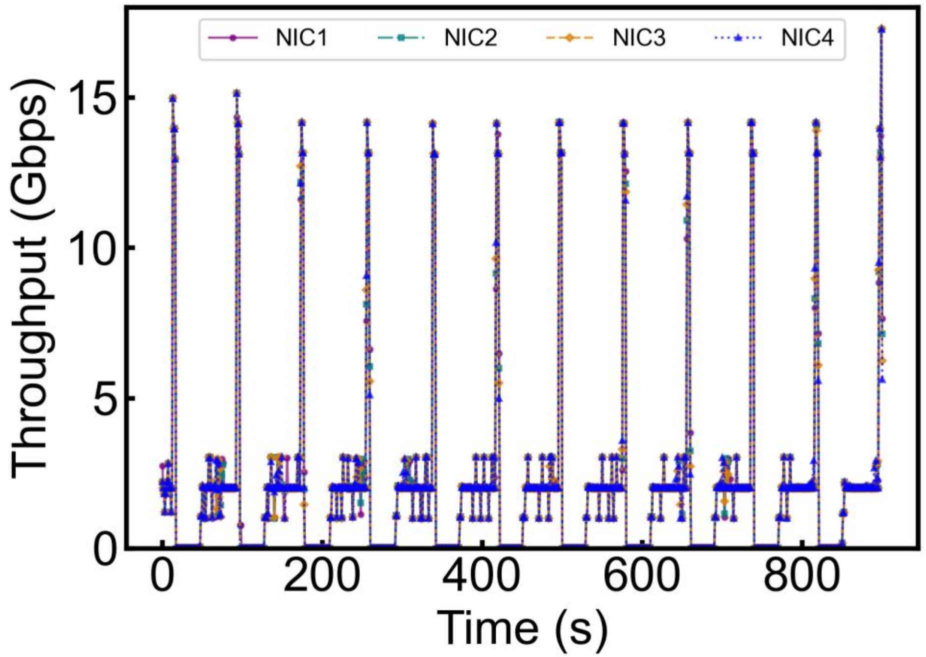
- Traffic matrix of model training



2. Motivation

□ Opportunity——Temporal burst cycles

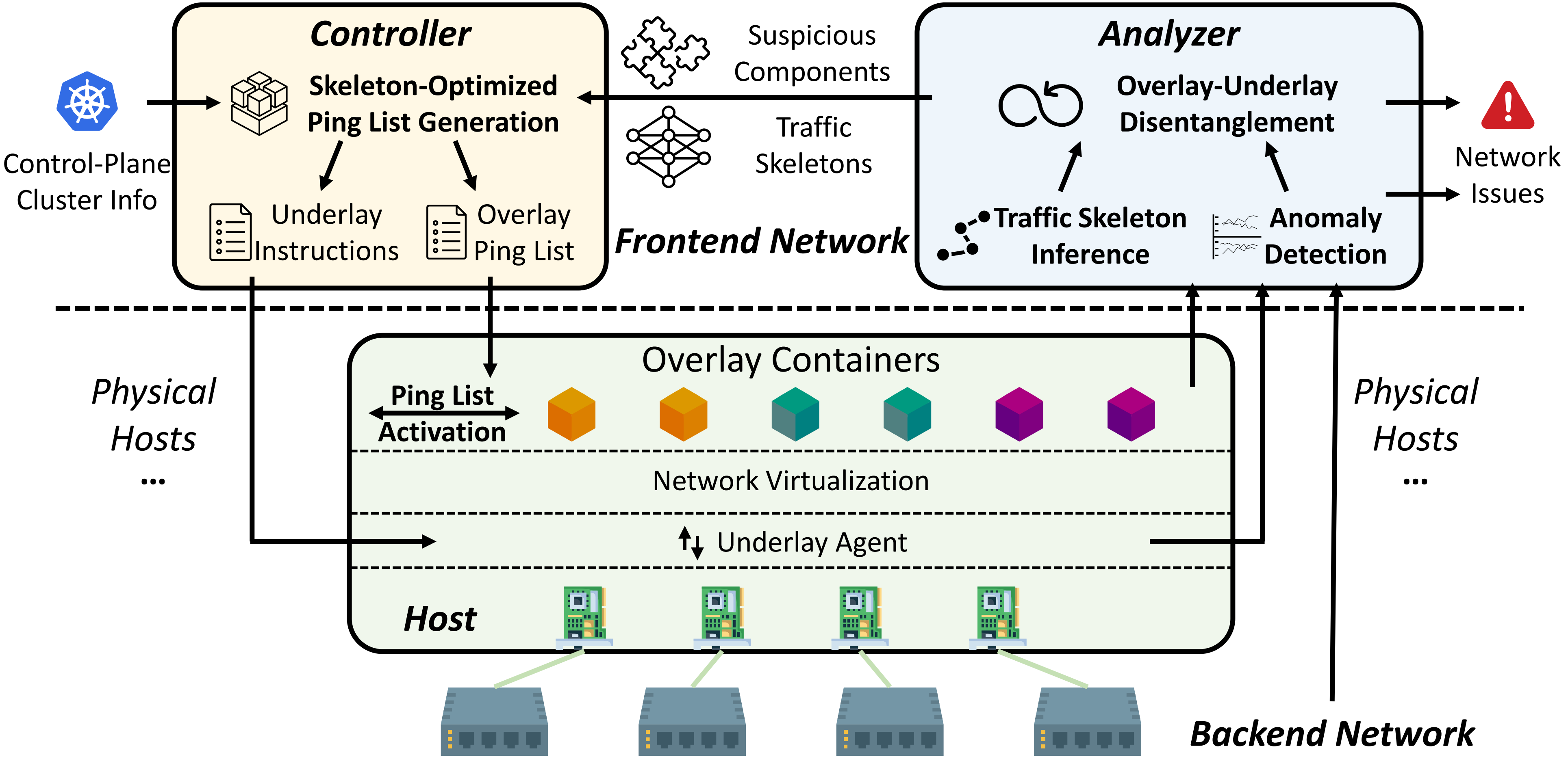
- Periodic and seasonal patterns
- Provide the opportunity to **distinguish the “role”** of each container in model parallelisms



3. Design of SkeletonHunter

Architectural overview

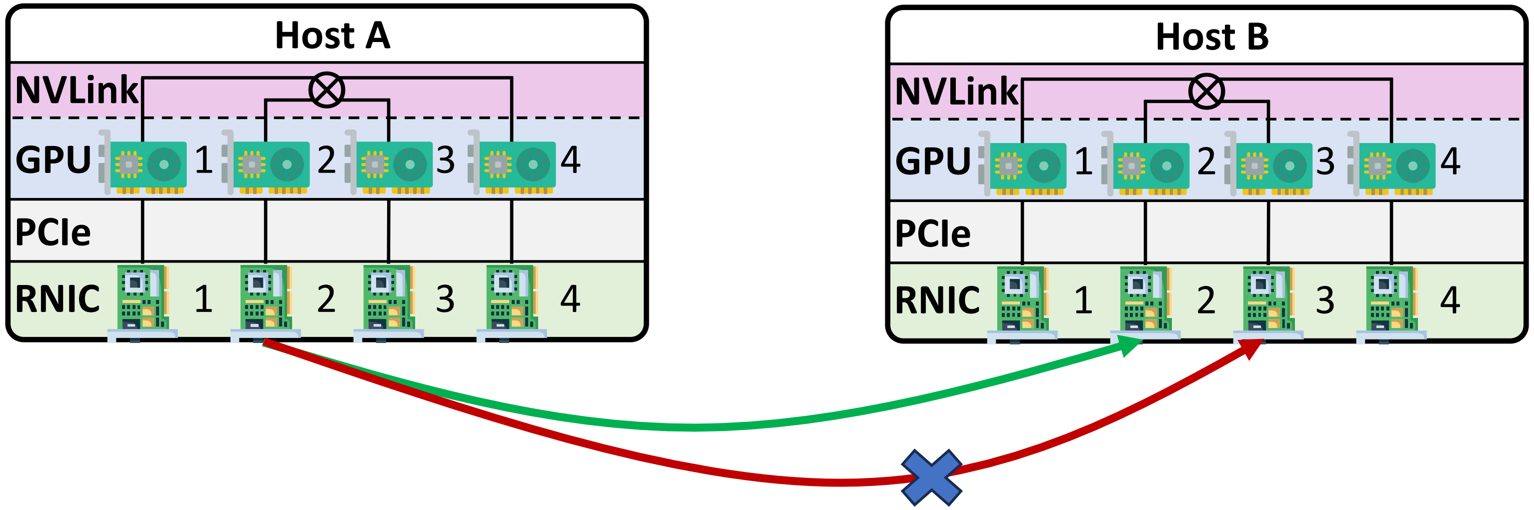
- Key idea——Infer **traffic skeletons** to reduce the monitoring complexity



3. Design of SkeletonHunter

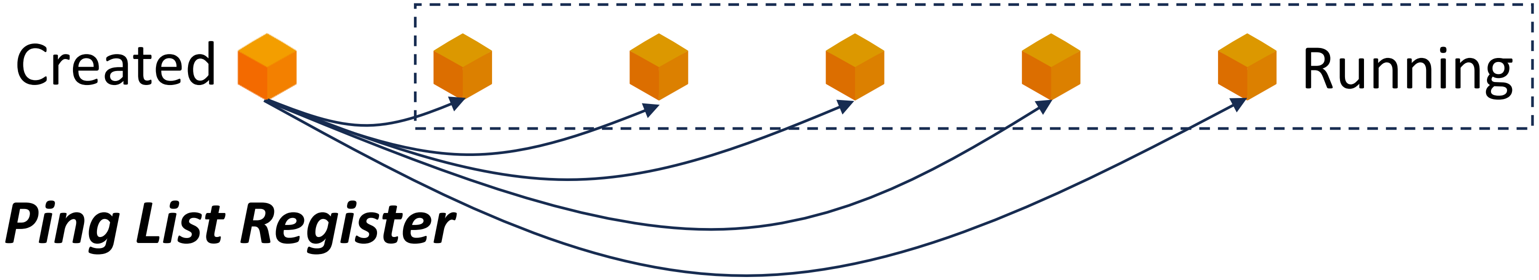
Traffic skeleton inference

- Preload: Remove ping list that are not in the same rail



7/8 reduction in ping list size

- Initialization: Incremental ping list activation

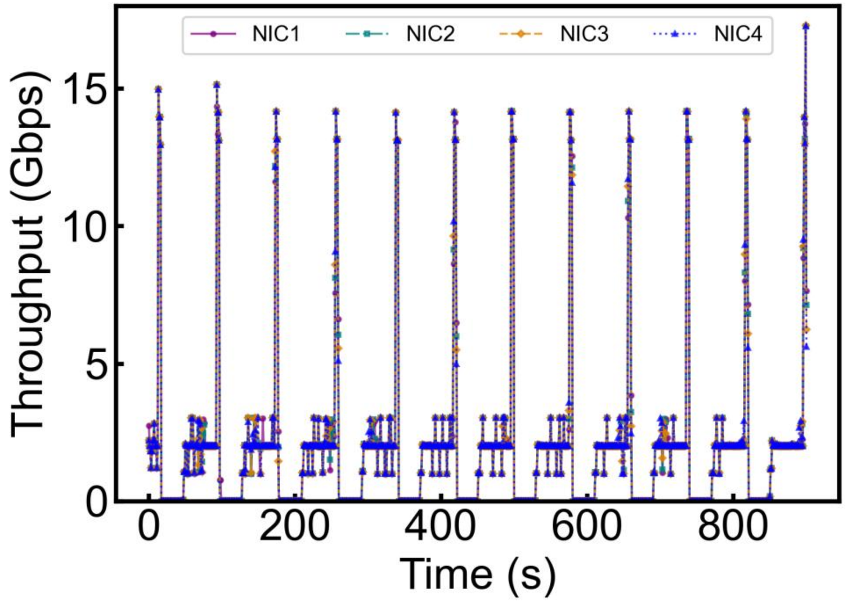


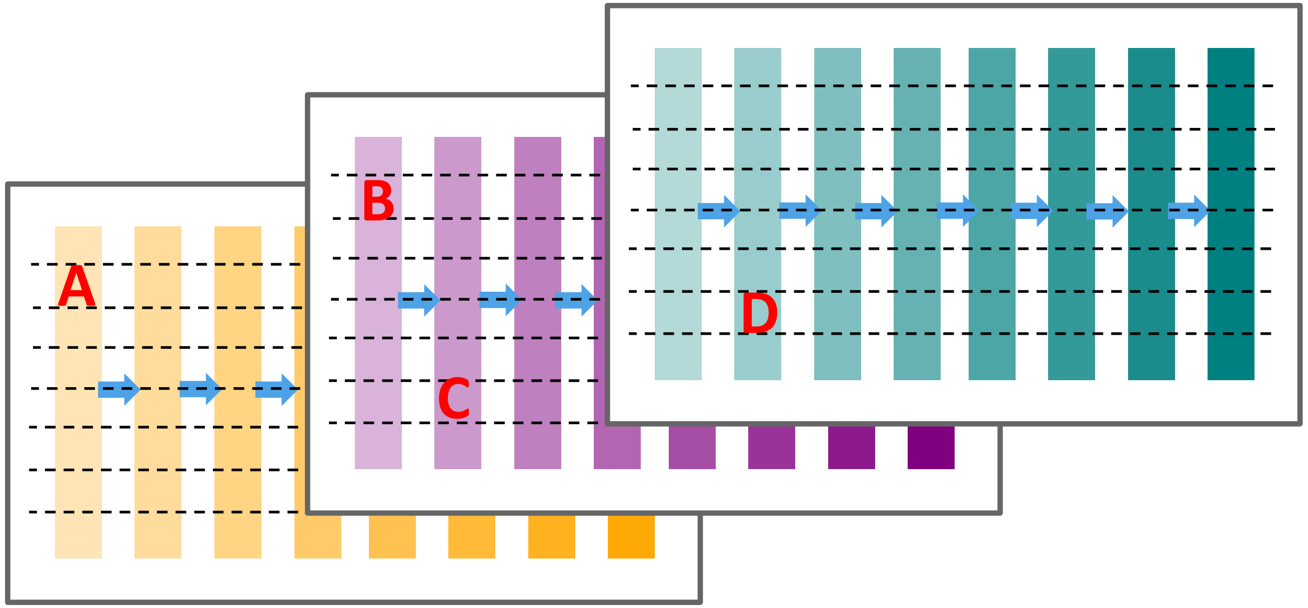
Ping List Register

3. Design of SkeletonHunter

Traffic skeleton inference

- Runtime: Optimization with inferred traffic skeletons



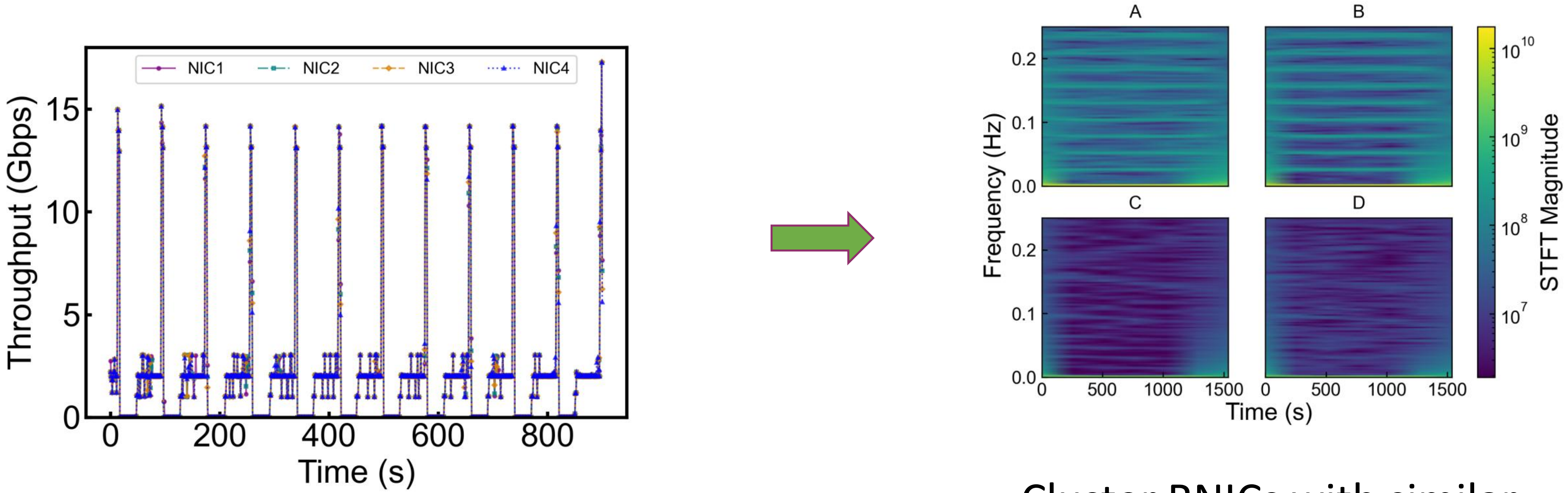


RNICs in the same **rank position** across different DPs have the same burst patterns in traffic

3. Design of SkeletonHunter

Traffic skeleton inference

- Runtime: Optimization with inferred traffic skeletons

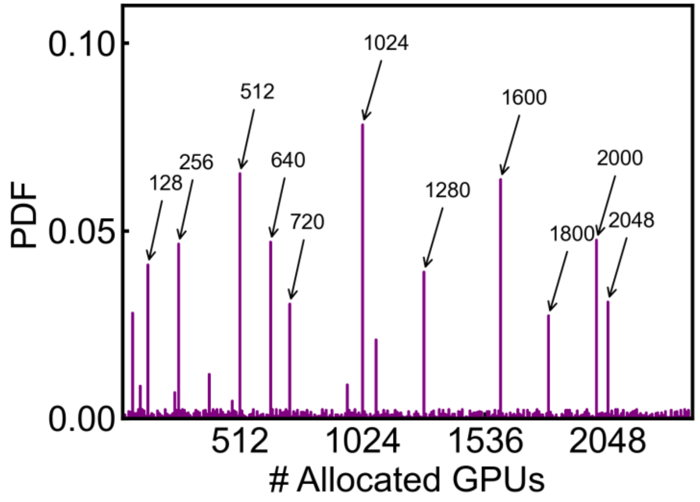


Cluster RNICs with similar STFT patterns

3. Design of SkeletonHunter

Traffic skeleton inference

- Runtime: Optimization with inferred traffic skeletons



Number of requested GPUs in a training job is confined to **limited set of values (e.g., 128, 512, and 1,024)**

Each DP group has the same number of RNICs

Degree of DP=TP·PP

DP Inference:

$$\textbf{min} \quad \sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(\|c_i\| - \bar{c})^2, \tag{1}$$

$$\textbf{s.t.} \quad N \bmod \lfloor \bar{c} \rceil = 0, \tag{2}$$

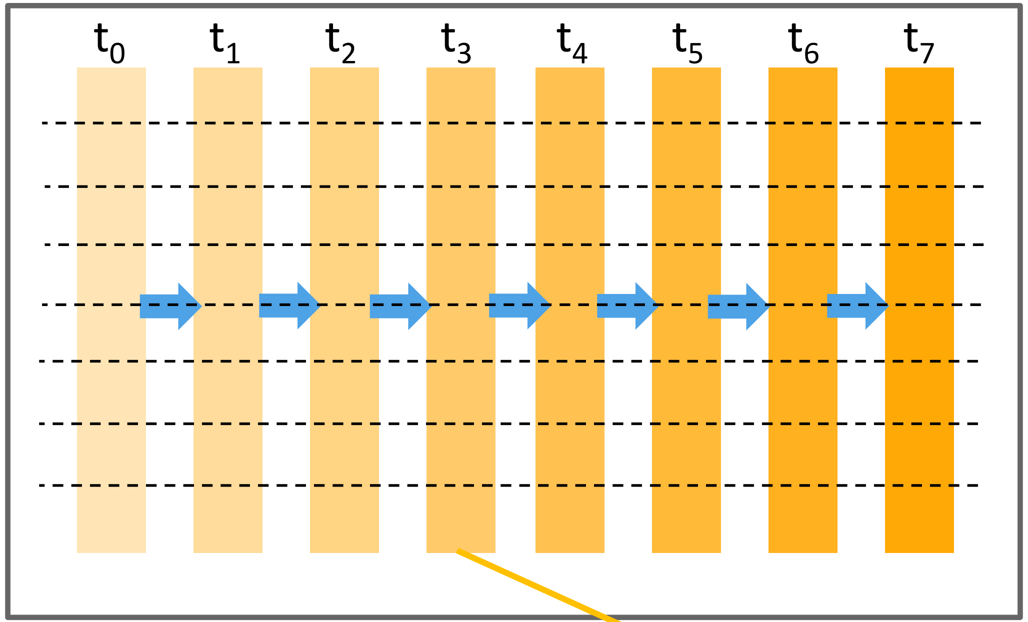
$$r_1, r_2, \cdots r_x \in H_r \Rightarrow \forall c_i, \|c_i \cap H_r\| \leq 1, \tag{3}$$

NVLink/PCIe Communications

3. Design of SkeletonHunter

Traffic skeleton inference

- Runtime: Optimization with inferred **traffic skeletons**



No network activities among the RNICs of the same host

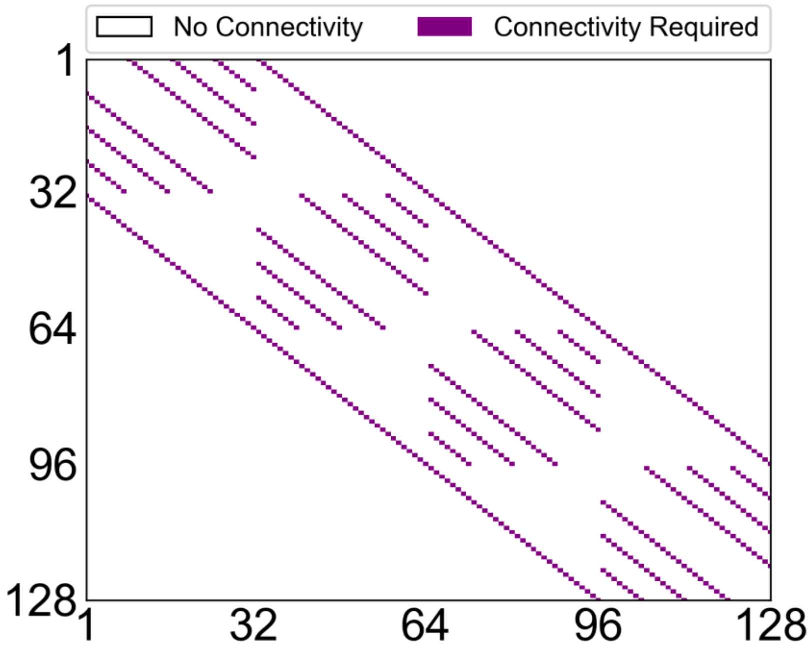
PP Inference:
Time-shifted Send/Recv (➡)

TP Inference:
No network activities for RNICs in the same host

3. Design of SkeletonHunter

Traffic skeleton inference

- Runtime: Optimization with inferred **traffic skeletons**



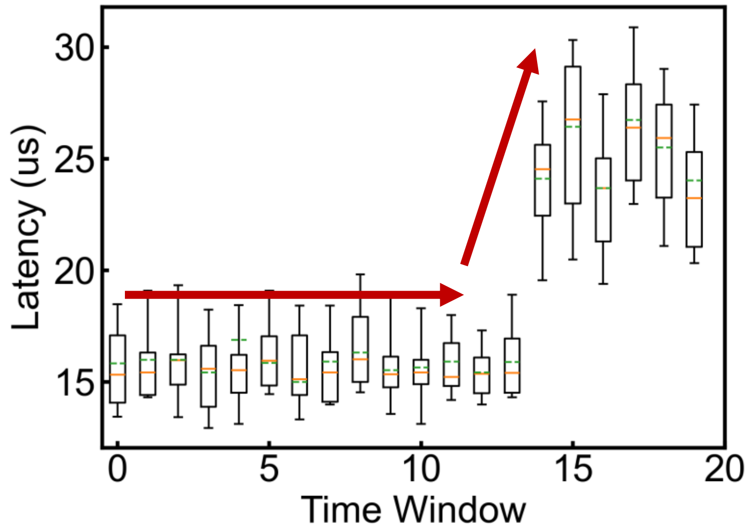
~5% of the all-to-all ping list

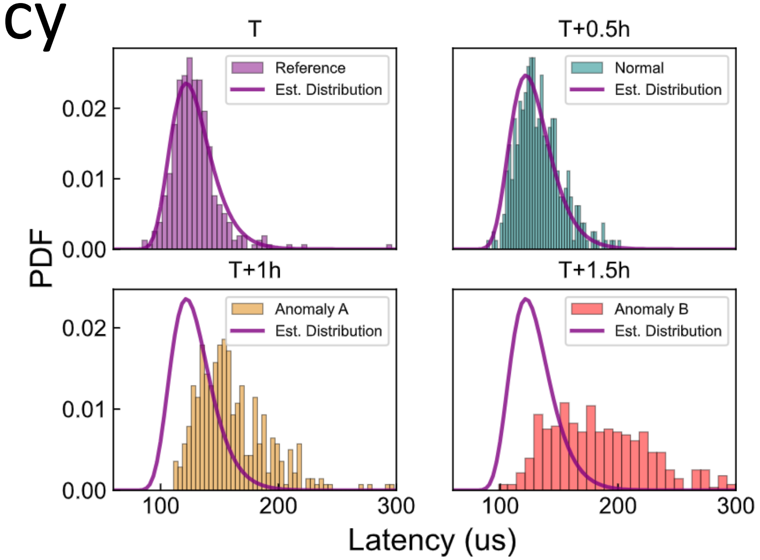
3. Design of SkeletonHunter

Connectivity anomaly detection

- Short-term
  - Percentiles as a feature for time window comparisons
- Long-term
  - Statistical testing to detect latency distribution changes

$$Y = \ln(X) \sim N\left(\mu, \sigma^2\right)$$

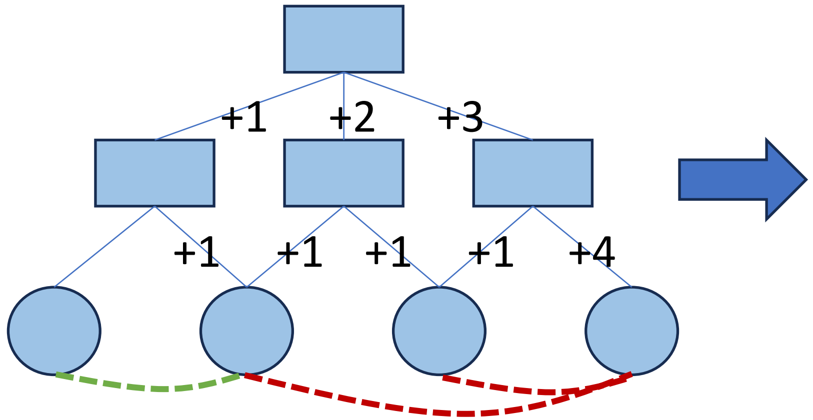




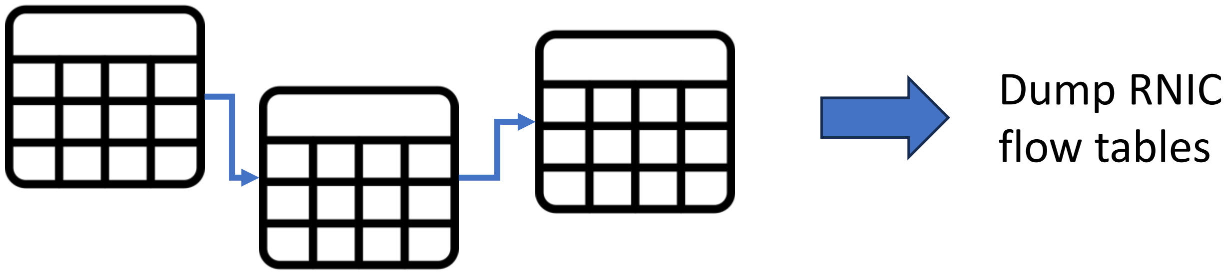
3. Design of SkeletonHunter

Network failure localization

- Optimistic assumption: the root causes of the overlay and the underlay layers are **software-** and **hardware-related** respectively, which **will not propagate to the other layer**
- **Examine the two layers' components separately**



Underlay: Voting-based failure localization

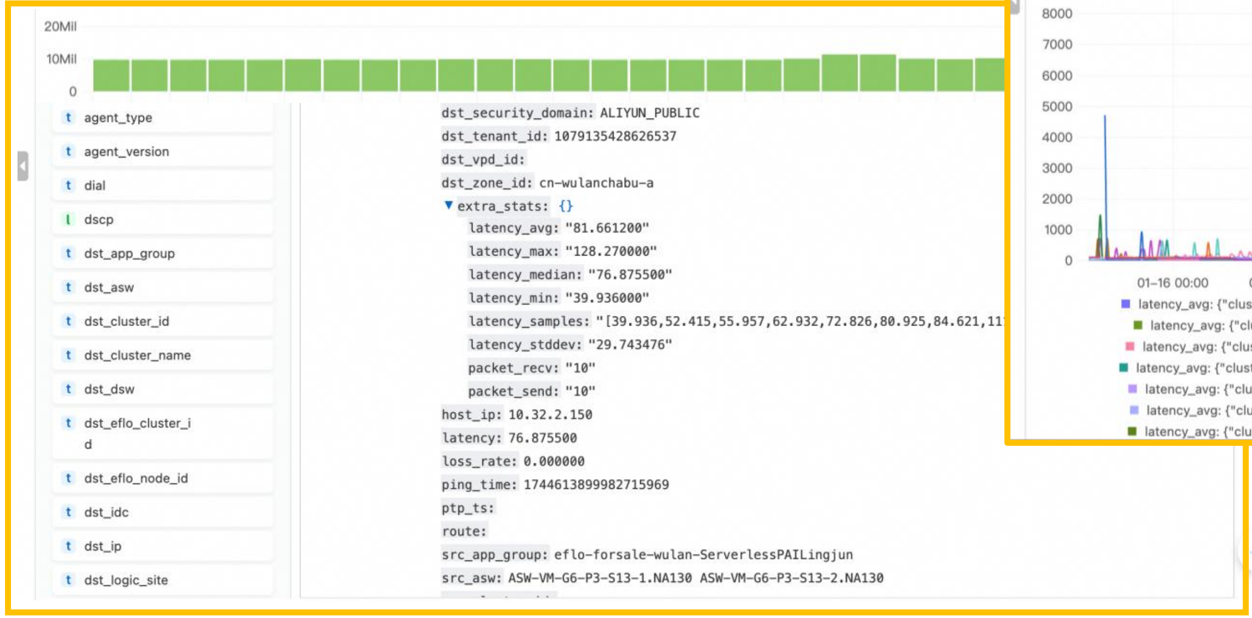


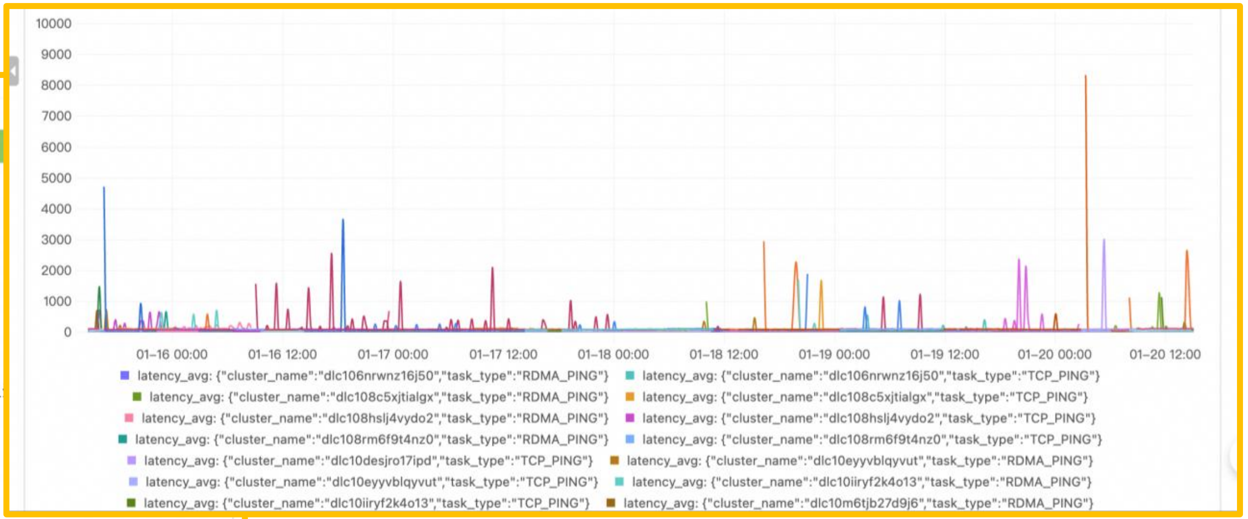
Overlay: Flow table reachability test

4. Evaluation

☐ Real-world deployment

- SkeletonHunter has been deployed in Alibaba Cloud for a year
- **Covering the containers on 5,700+ physical hosts and 40K+ RNICs**
- **1B latency logs among training containers every day**

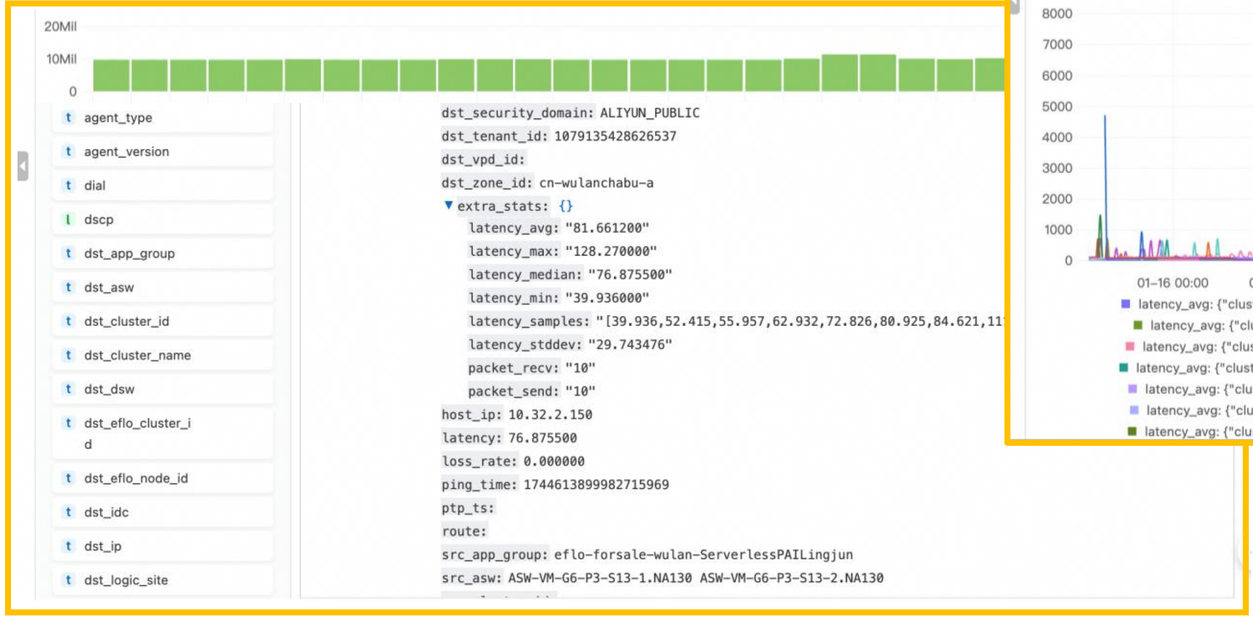


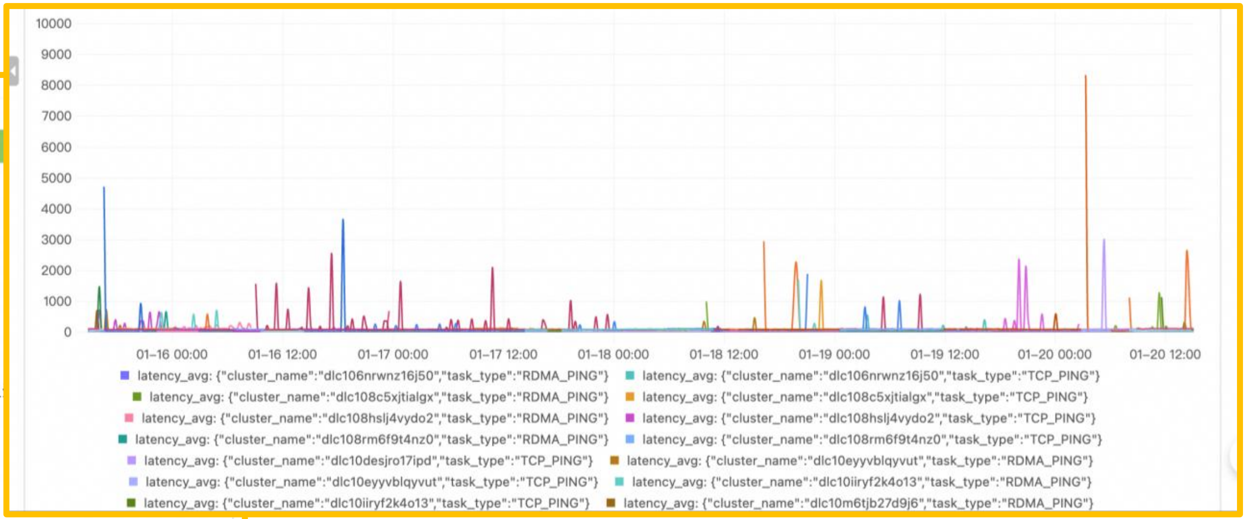


4. Evaluation

Real-world deployment

- Detects and localizes **4,816** network failures with **>98% accuracy**
- Reduces **99%** of failures after fixing corresponding network components

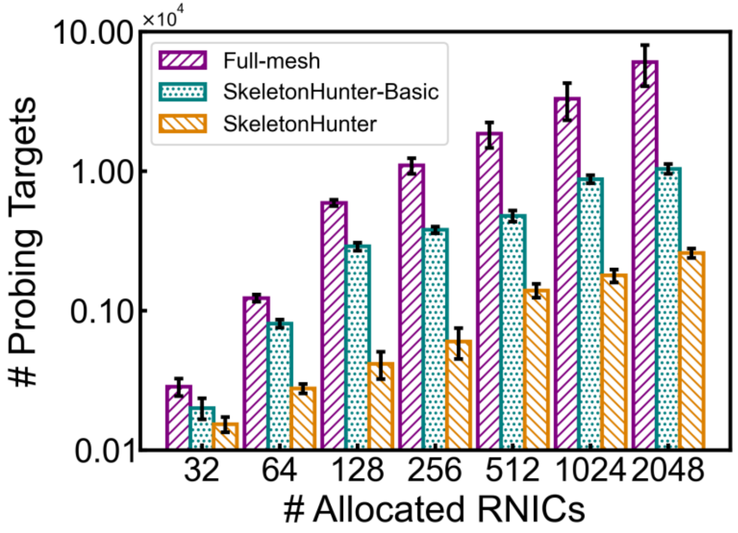


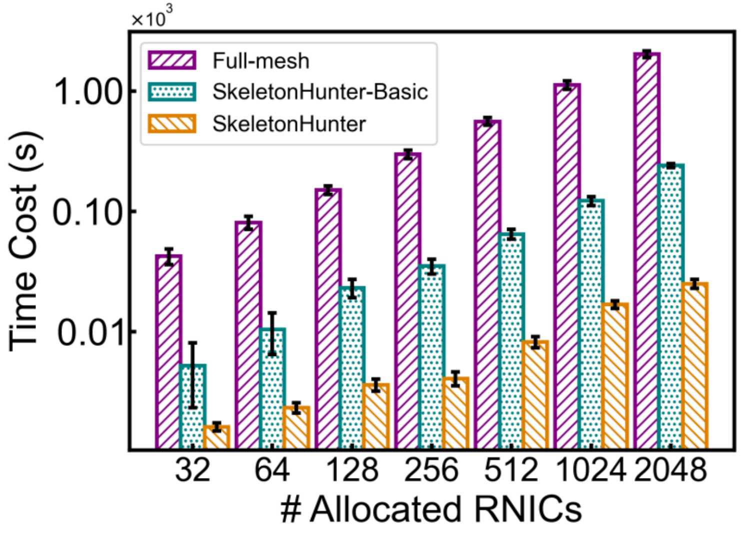


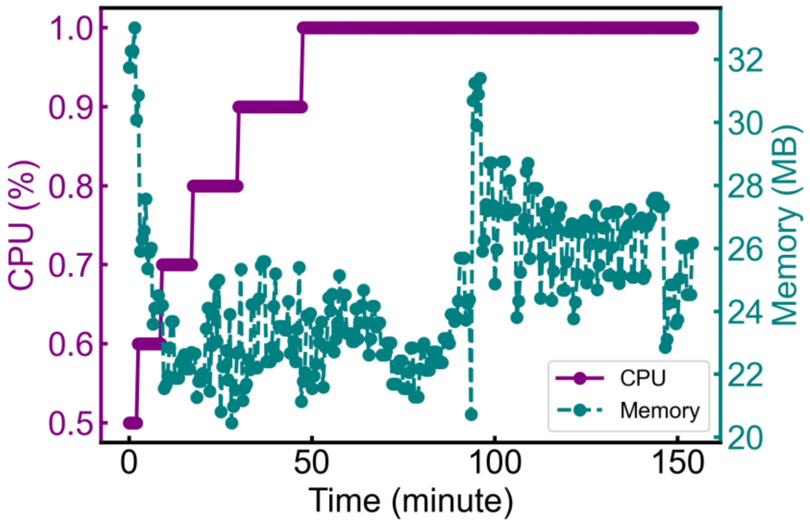
4. Evaluation

Real-world deployment

- Traffic skeletons help reduce detection complexity
- Negligible detection overheads: 8s for a probing round on average







5. Conclusion

- We are the first to point out the **real-world challenges** against **reliable network** support for large-scale **containerized model training**, as well as their **multiplicative effect** on troubleshooting the connectivity issues.
- We propose SkeletonHunter, a container network monitoring and diagnosis system that leverages the unique **traffic patterns of large model training** to accurately and efficiently pinpoint the connectivity issues.
- SkeletonHunter has been **deployed in our production container network** and has helped discover diverse network failures that derive from the problems of different network components. We have fixed most problematic network components and greatly reduced the monthly failure rate.

Thanks!

Q & A