Holmes: Localizing Irregularities in LLM Training with Mega-scale GPU Clusters

Zhiyi Yao, Pengbo Hu, Congcong Miao, Xuya Jia, Zuning Liang, Yuedong Xu, Chunzhi He, Hao Lu, Mingzhuo Chen, Xiang Li, Zekun He, Yachen Wang, Xianneng Zou, Juncheng Jiang

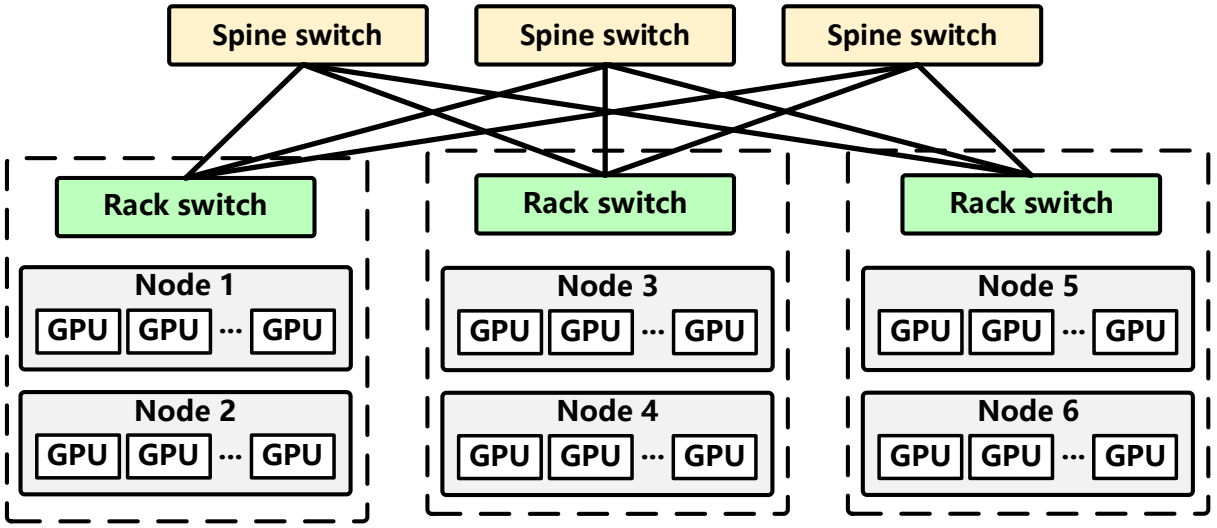
復旦大學
FUDAN UNIVERSITY

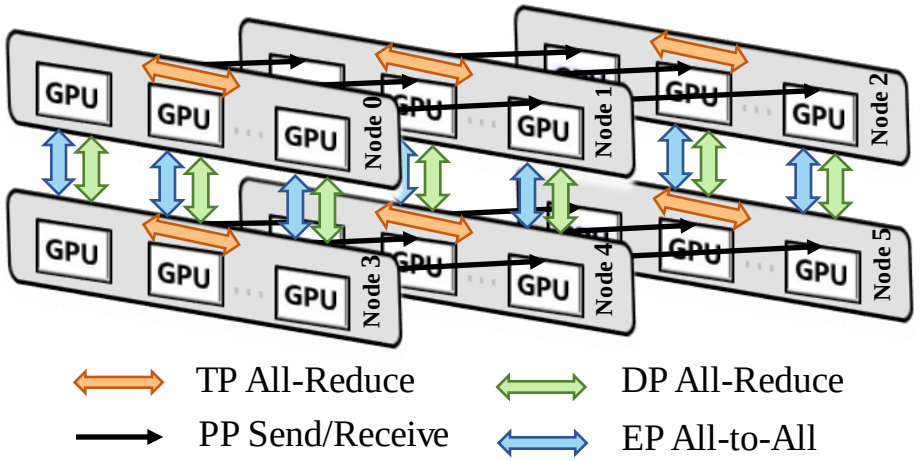
Tencent 腾讯

THE UNIVERSITY OF CHICAGO

Parallel training and communication

- LLM training incorporates multi-dimension parallelisms
  - Data-, Tensor-, Pipeline-, Expert-parallelism, etc.

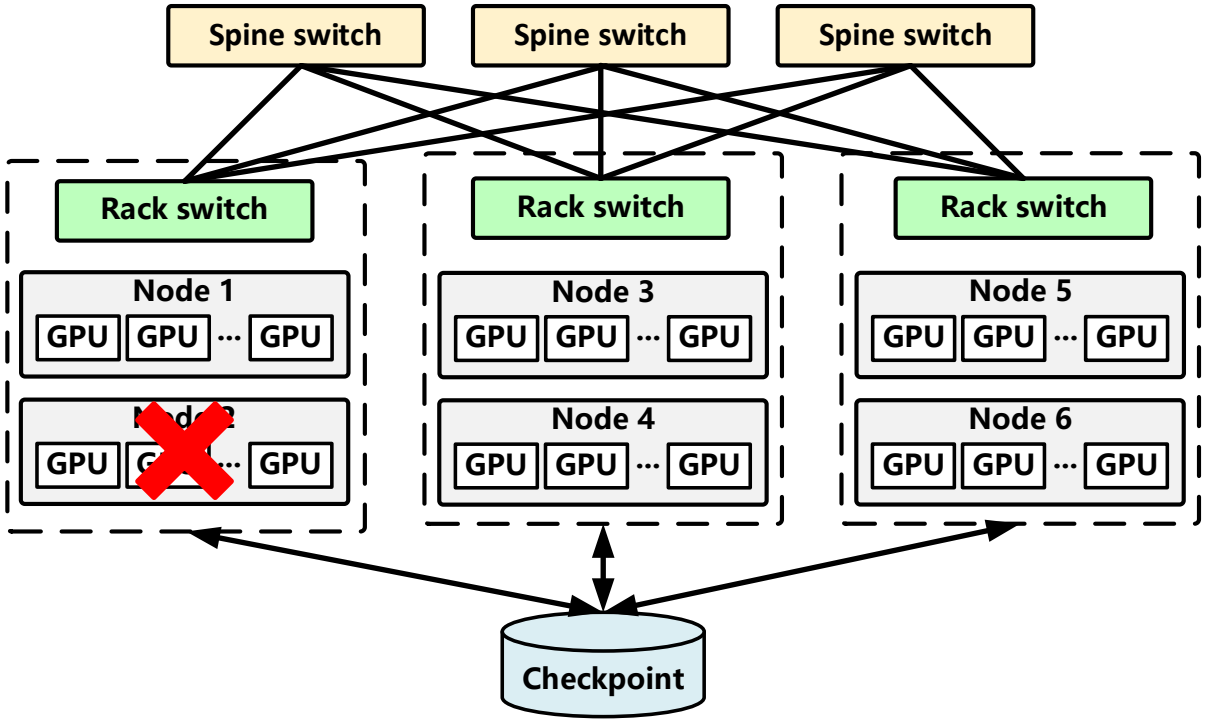




How to ensure the reliability of large-scale training in mega-scale cluster?

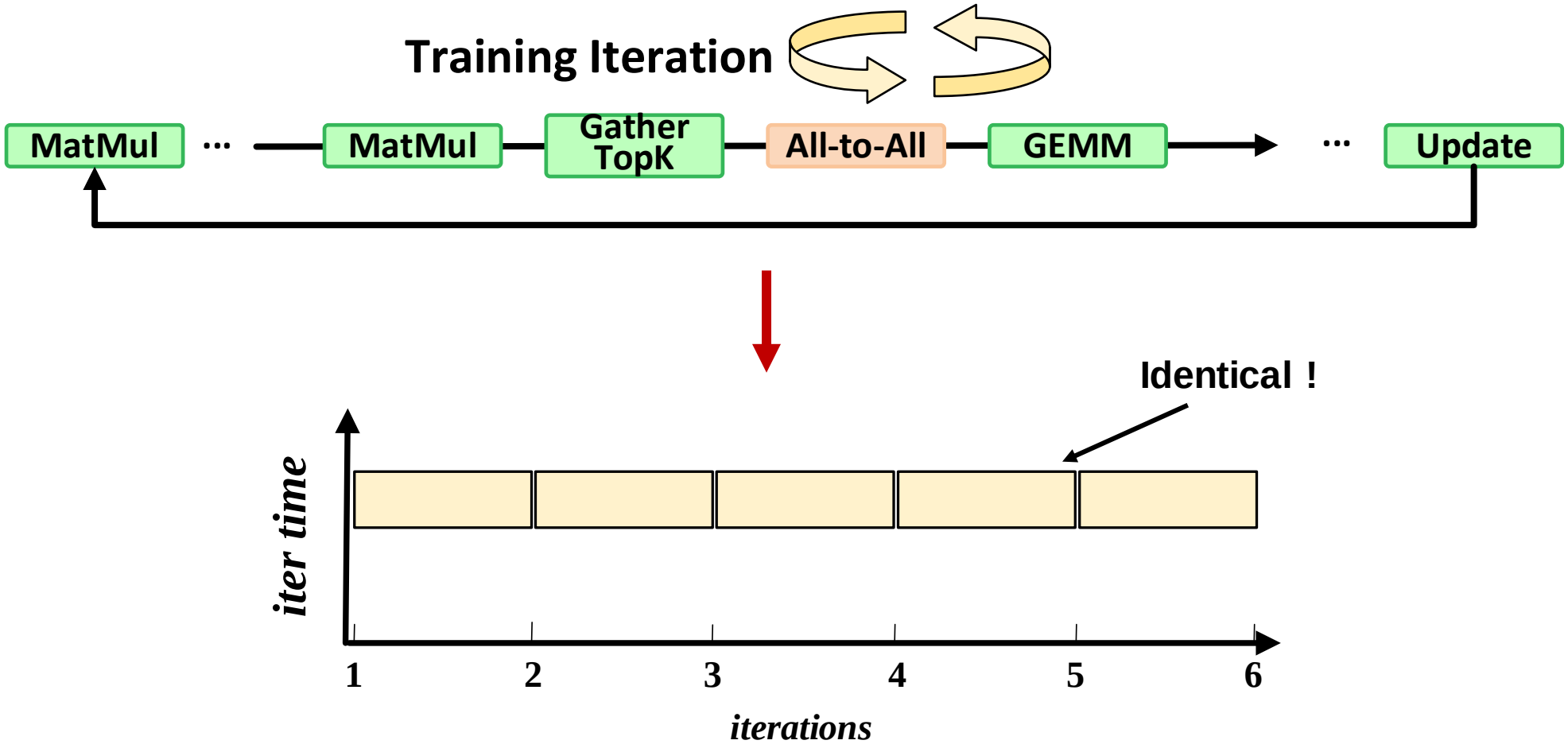
Reliability of LLM training

- Existing approaches focus on addressing failures during training

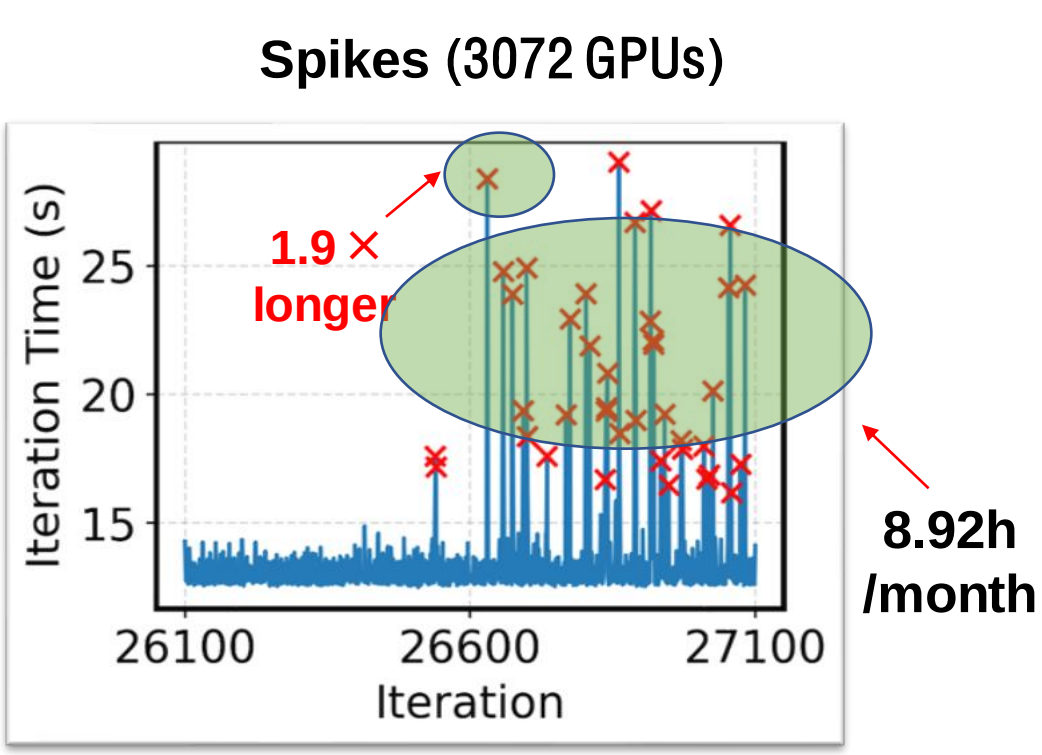


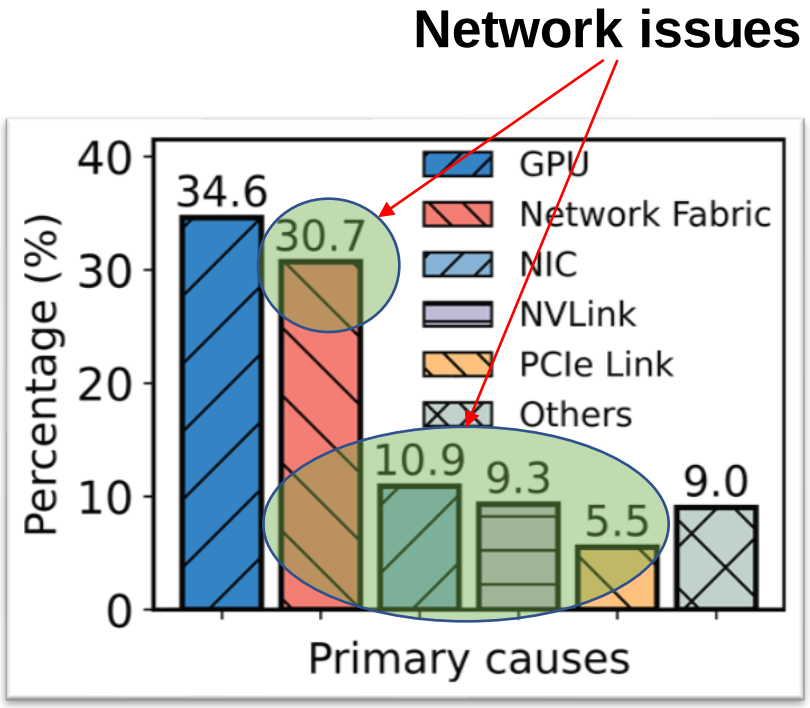
- Checkpointing mechanisms
  - CheckFreq (FAST'21)
  - Gemini (SOSP'23)
  - Check-N-Run (NSDI'22)
  - …
- Reliable/Elastic training mechanisms
  - Oobleck (SOSP'23)
  - Parcae (NSDI'24)
  - Bamboo (NSDI'23)
  - Tenplex (SOSP'24)
  - …

Iteration time is expected to be stable



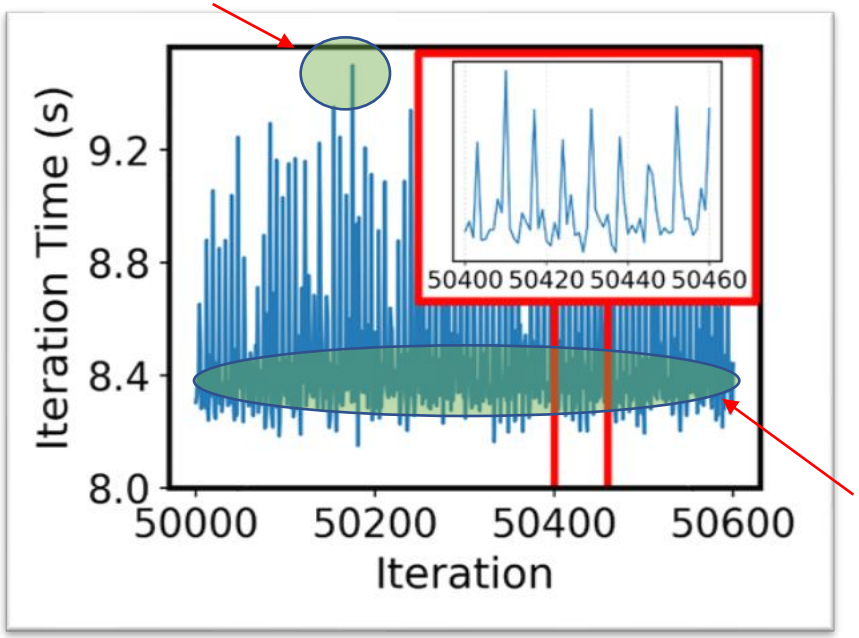
Irregularity: abnormal spikes



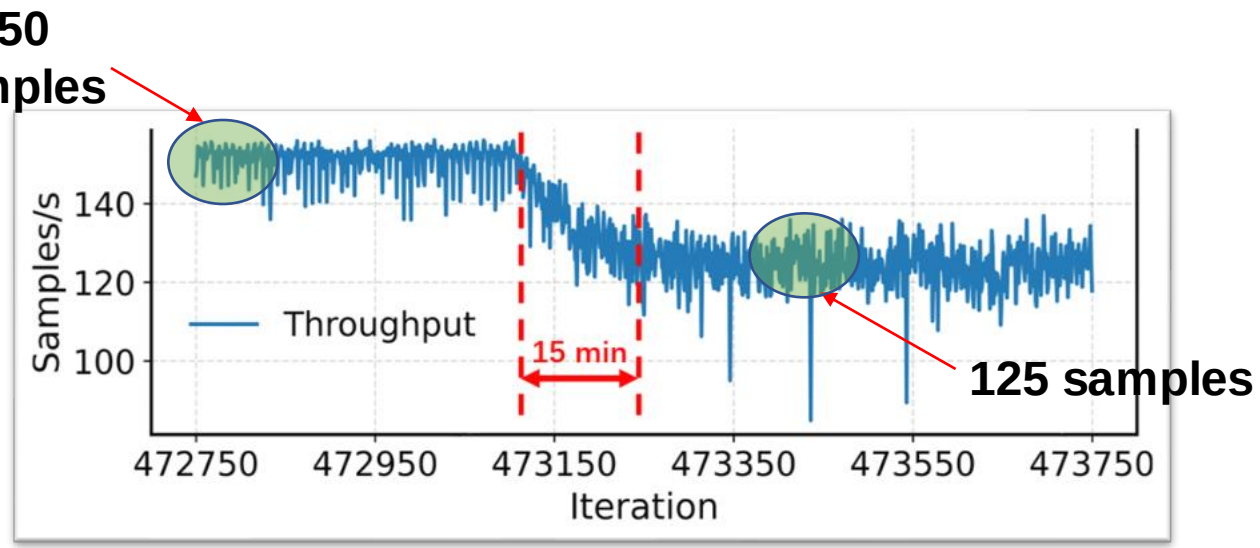


Abnormal spikes: severe impact on individual iteration time

Irregularity: performance fluctuation and persistent degradation



Fluctuation (4096 GPUs)

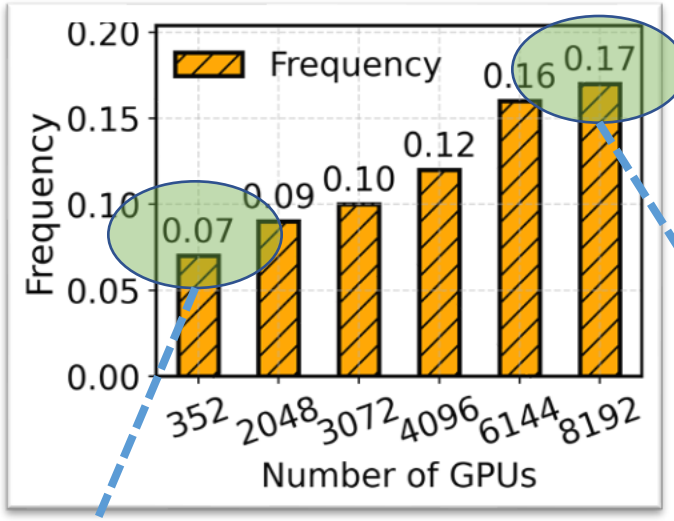


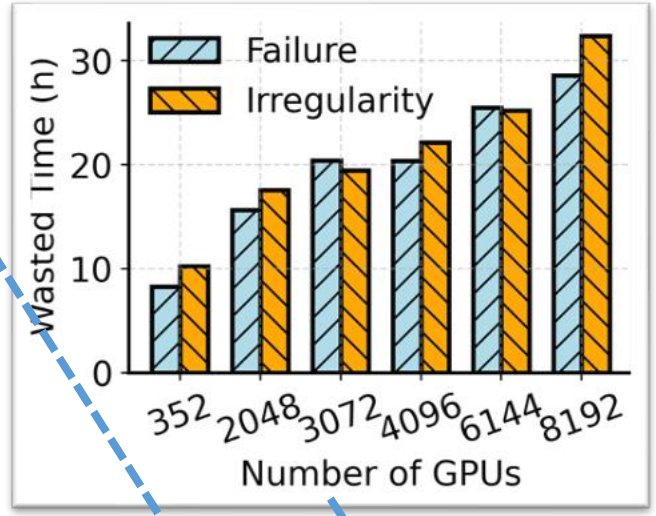
Training throughput degrades in 15 mins

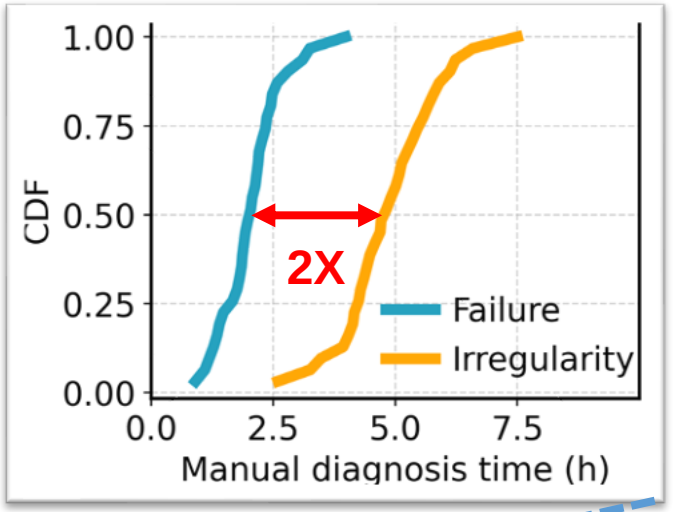
Persistent degradation (4096 GPUs)

Irregularity impacts the training process

Definition: δ-Irregularity refers to the phenomenon where training continues **uninterrupted**, but the training iteration time, is **δ times longer than the average value** within a given time window.





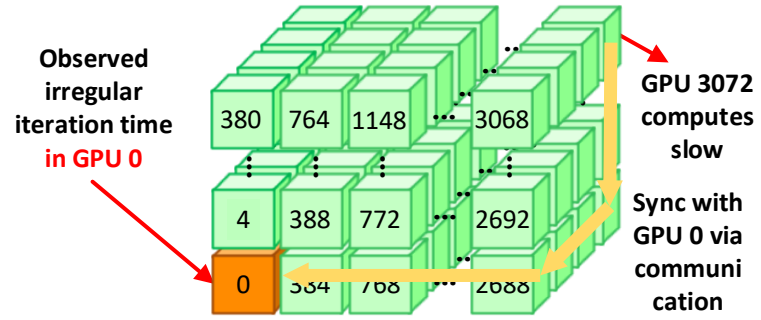


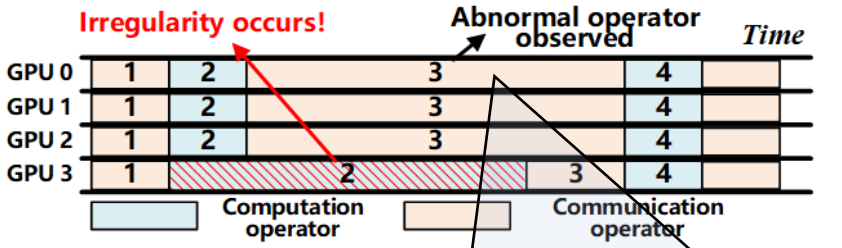
Frequency of Irregularity increase with number of GPUs

Irregularities waste lots of training and operation time

Challenges on pinpointing irregularity

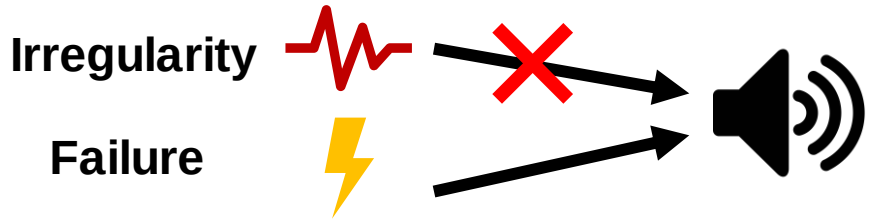
C1: Irregularities behavior can propagate



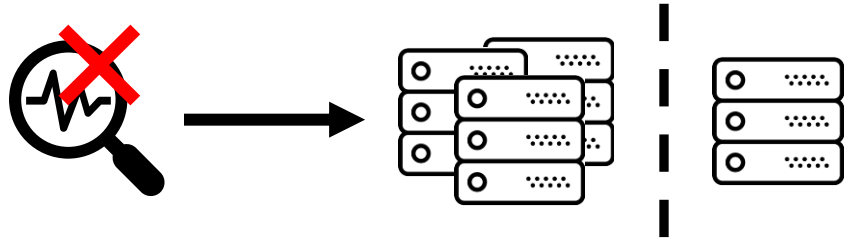


Operator 2 executor abnormally, making Operator 3 of GPU 0 **looks abnormal**

C2: Irregularity occurs silently

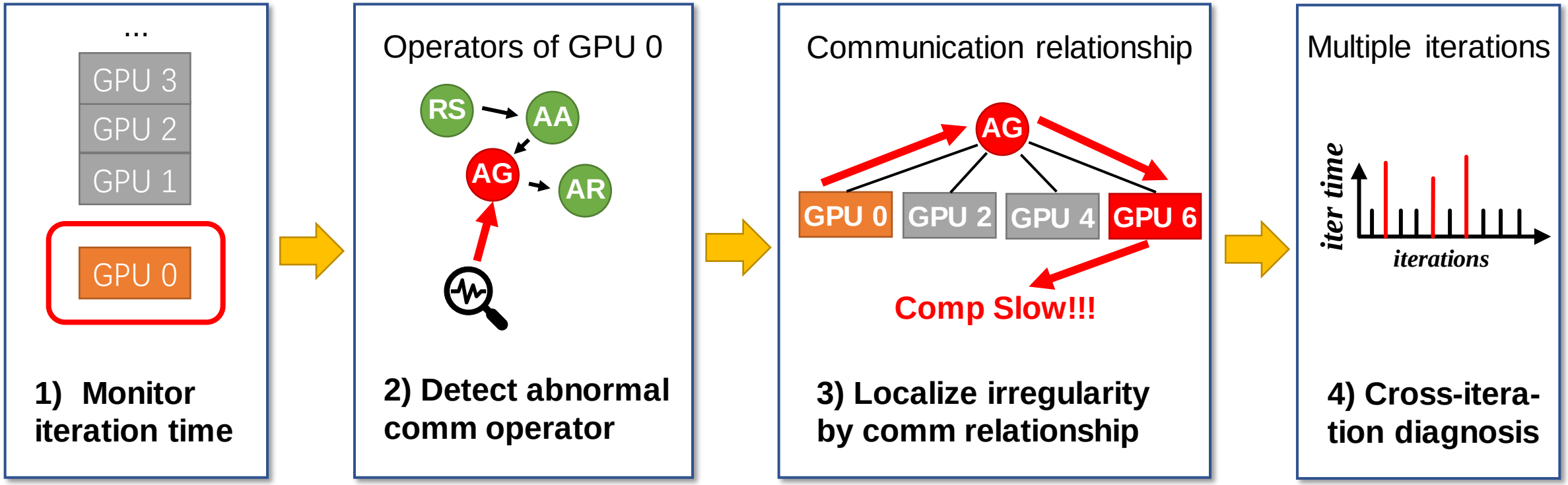


C3: Misidentifying irregularities is costly



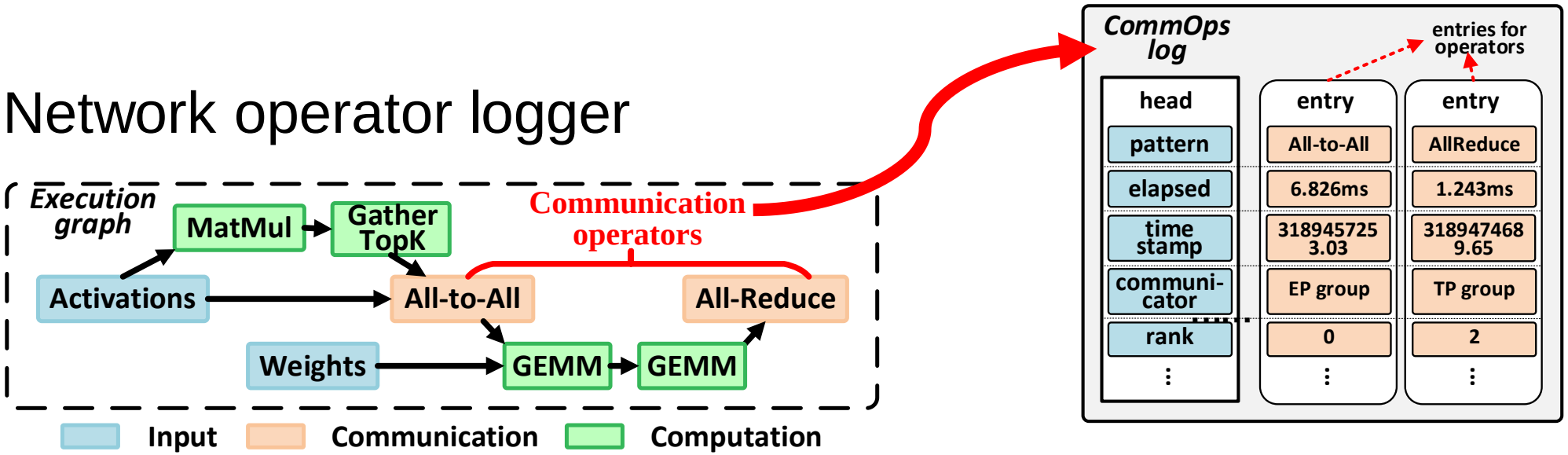
Localizing irregularity with Holmes

- Workflow: localizing the irregularity using communication operators and communications during training.

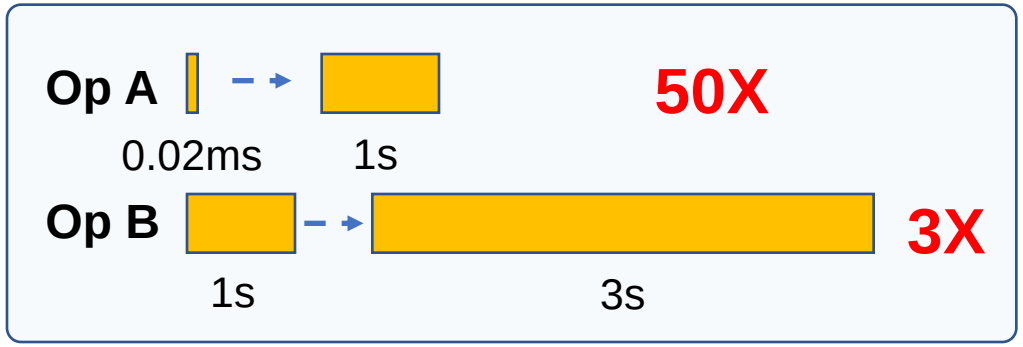


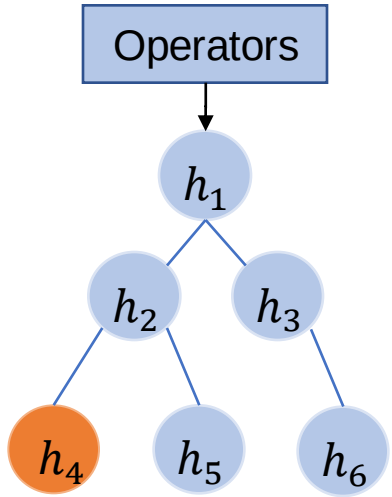
Detect abnormal operators

- Network operator logger



- How to confirm an abnormal operator?



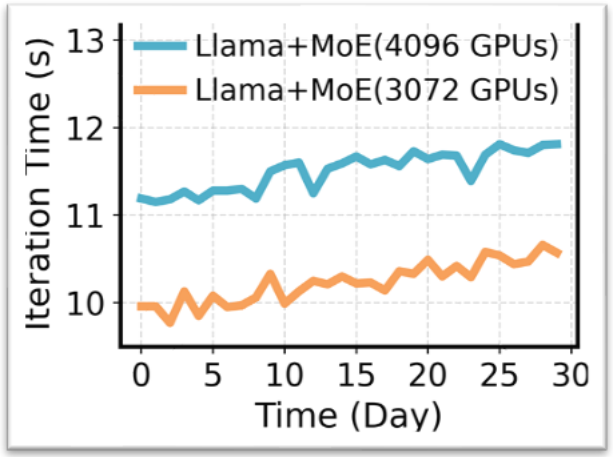


Random Forest

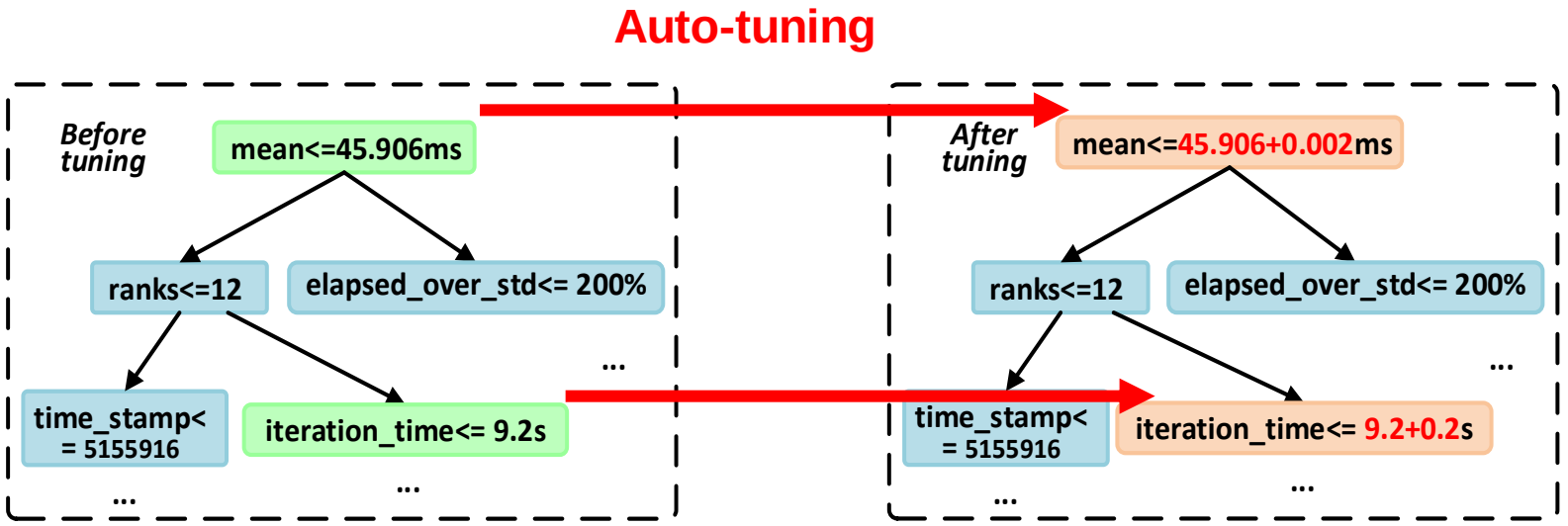
Mean and std.
Z-Score
Quartiles
IQR
Fixed feature
…

Detect abnormal operators

- How to mitigate the precision degradation brought by distribution drift?

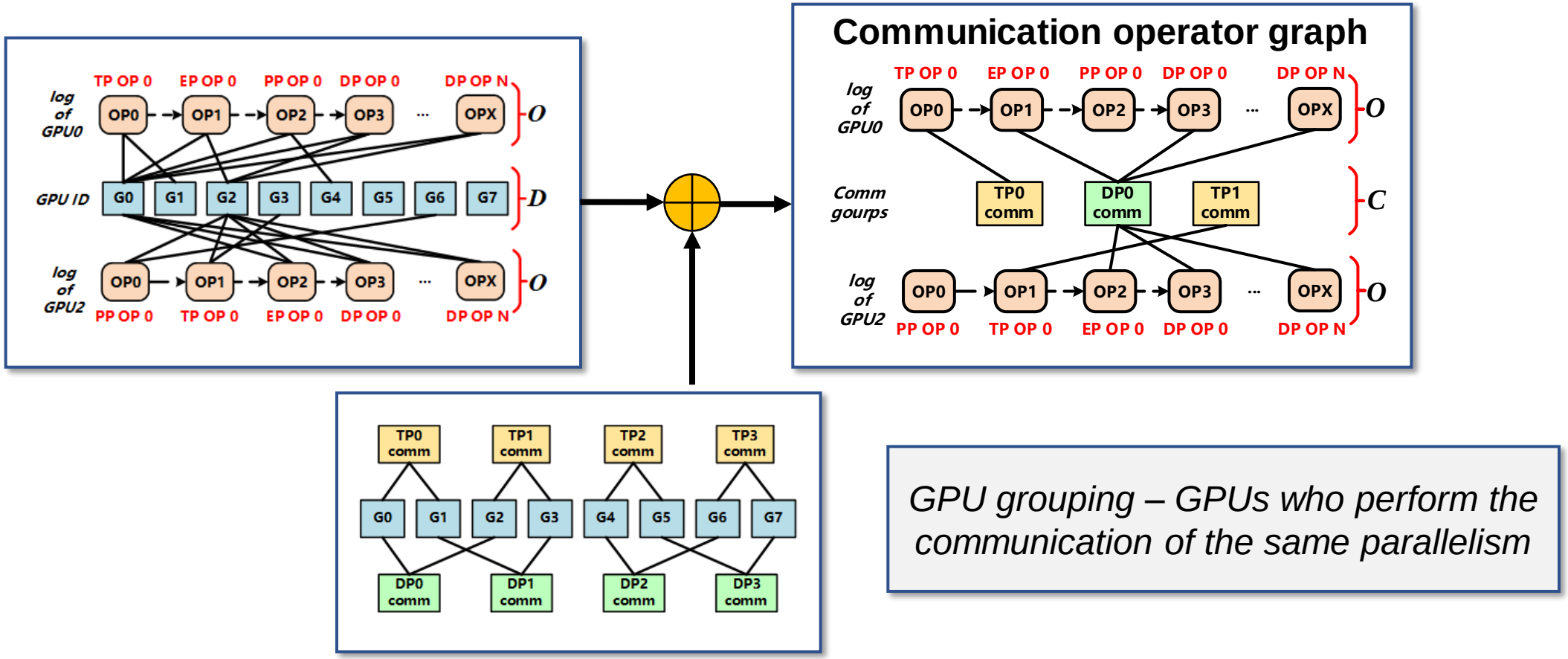


Iteration time varies slightly



Localizing root cause devices

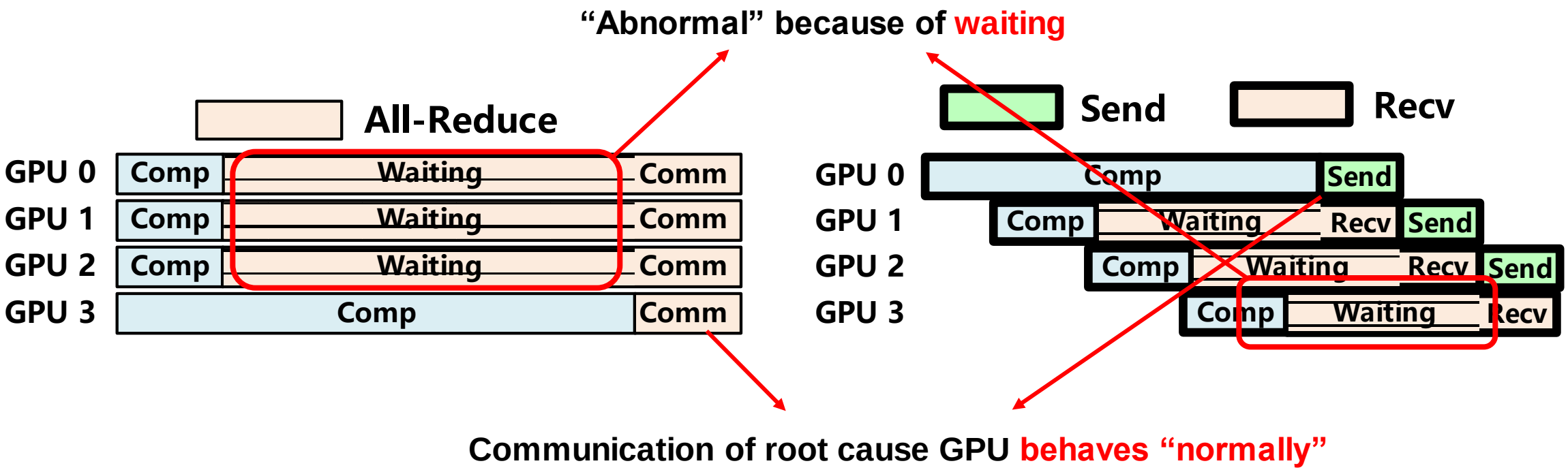
- How to represent the relationship between operators and GPUs?



GPU grouping – GPUs who perform the communication of the same parallelism

Localizing the root cause

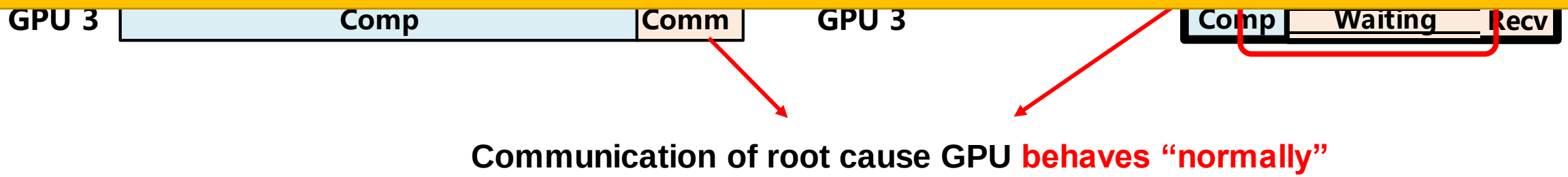
- Observation: “Abnormal” communication operators can be waiting for the straggler.



Localizing the root cause

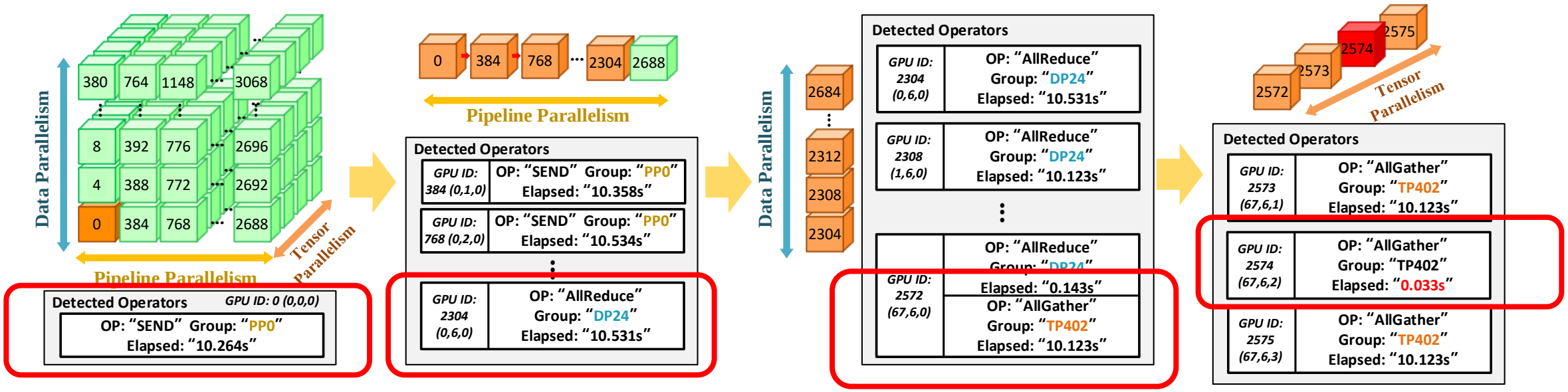
1. For prolonged communication operation, if any participating GPU shows **normal** operator status **while all others exhibit abnormal**, that specific GPU with normal communication operators is identified as the root cause.

2. If **all GPUs** involved in the communication show **abnormal** operator status, we conclude that the communication itself is the source of the problem.

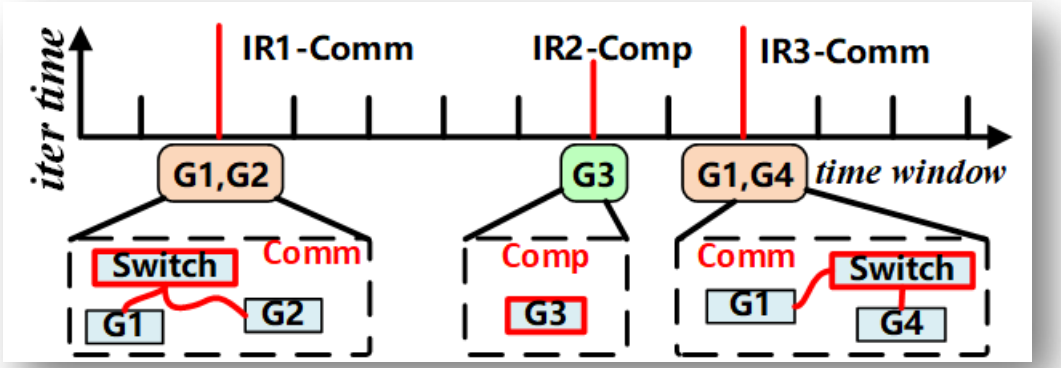


Case Study

- Training with 3072 GPUs
- Degrees: DP(96), PP(8), TP(4)



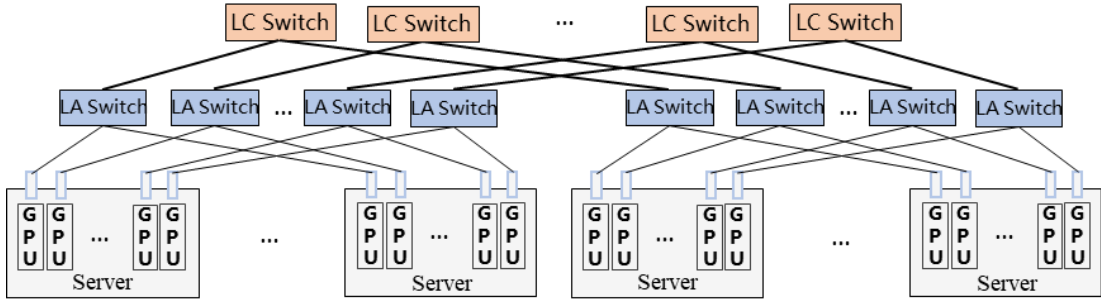
Cross-iteration analysis



- Quantify the influence of abnormal operator (x) on training within time window T
- Accumulation of device irregularity rate cross-iteration
- Identifies the Top-2 anomalous device candidates

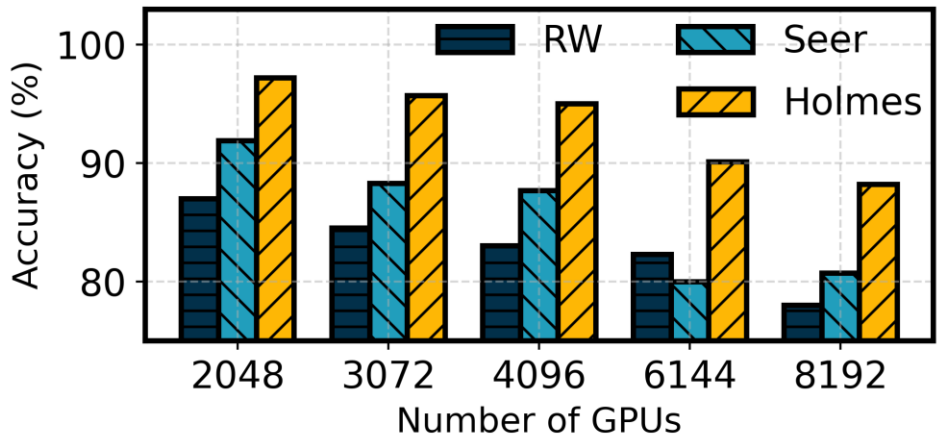
Evaluation setup

- Setup
  - Large-scale trace driven simulation
  - Production-level testbed
- Models
  - LLama2 w and w/o MoE
  - GPT3 w and w/o MoE
- Data
  - Ground Truth: Manual anomaly detection and validation based on expert experience
- Metrics
  - Accuracy and Latency

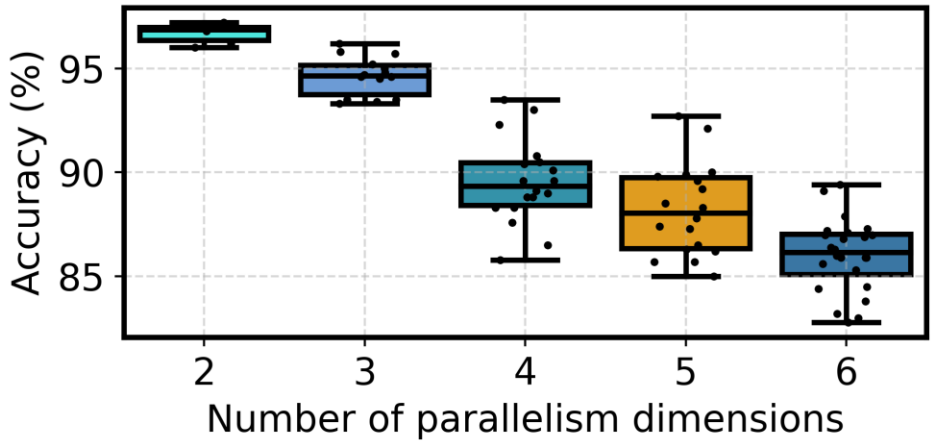


End-to-end localization accuracy

- Overall accuracy up to **97.2%**
- 88.2% of accuracy when training with 8192 GPUs
- Median accuracy of 86.0% when 6-d parallelism enabled



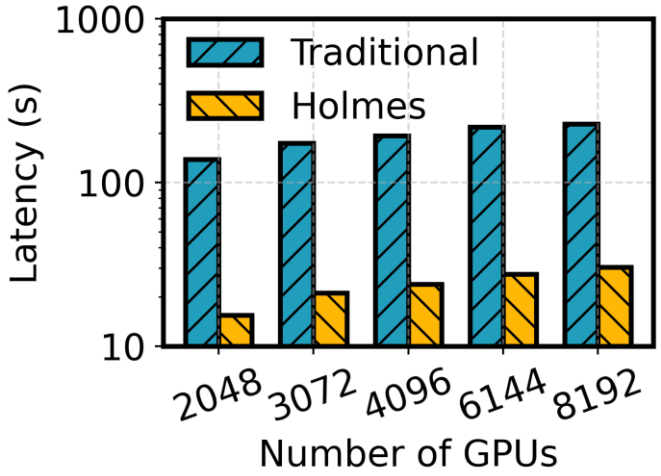
(a) E2E localization accuracy over training with different number of GPUs.



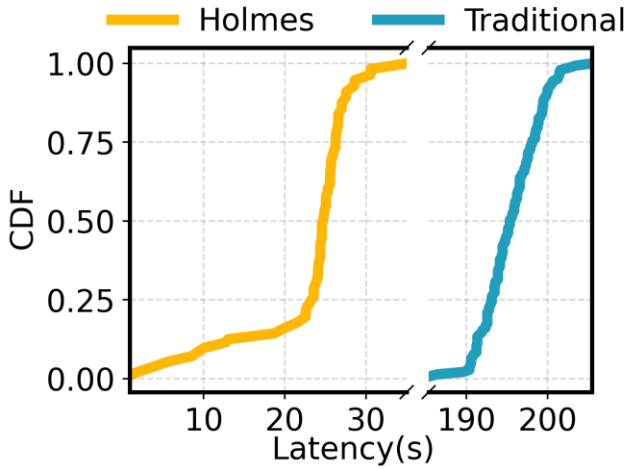
(b) E2E localization accuracy over training with different number of parallelism dimensions.

End-to-end localization latency

- Overall latency within **30.3s**
- An 87.8% reduction compared to full GPU log analysis (baseline)



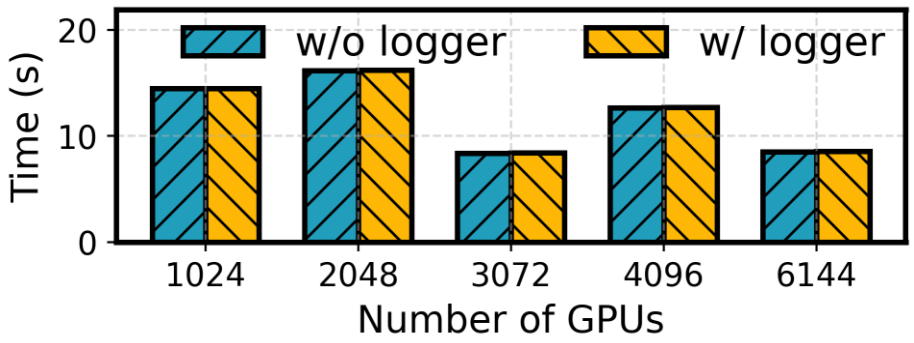
(a) End-to-End analysis latency



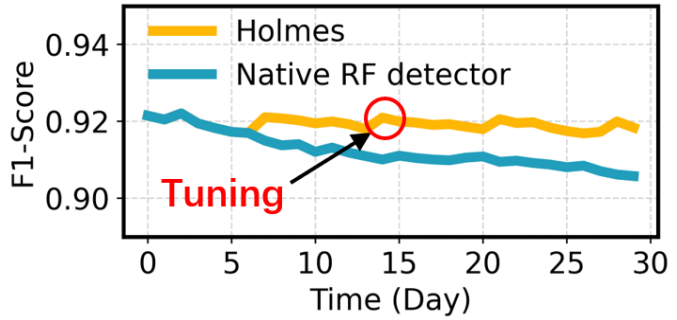
(b) CDF of End-to-End latency

Modular study

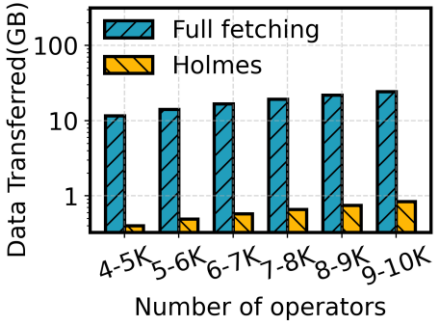
- Per-operator logging time is less than 2μs.
- Better detection accuracy against distribution drift
- 96.6% reduction of the log data transmission



(a) Overhead of CommOps logger



(b) Performance of auto-tuning



(c) Data transferred for localization

More experiments in our paper

- End-to-End evaluation
  - Accuracy over different types irregularity
  - Impact of delta on accuracy
- Study of abnormal operator detection
  - RF compared to other ML models
  - Abnormal operator detection latency
  - Auto tuning features study
  - Parameter search for RF model
- Multi-node monitoring for improving localization accuracy
- Impact of cross-iteration diagnosis

Conclusion

- We systematically measure and analyze the irregularity phenomenon in LLM training

- We design a localization system for irregularity from network perspective and experimentally demonstrate its effectiveness

- We hope our early attempts can inspire more innovations for LLM OPS!

Thank You!

Contact email: ydxu@fudan.edu.cn