FAST '25 Work-in-Progress Reports (WiPs)

Baton: Orchestrating GPU Memory for LLM Training on Heterogeneous Cluster

Yi Zhang Shuibing He **Ping Chen**

Zhejiang University and Zhejiang Lab

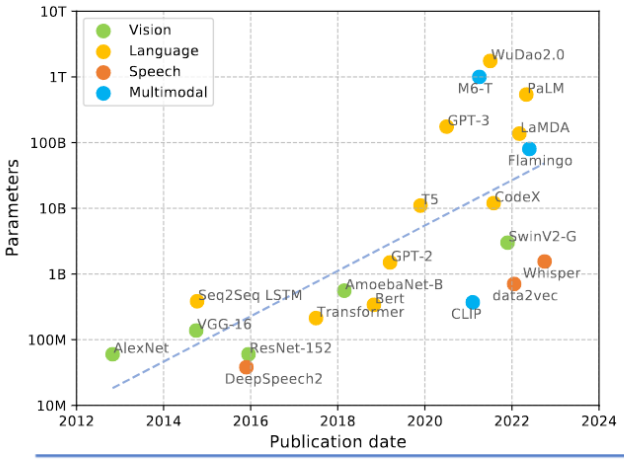
之江实验室 ZHEJIANG LAB

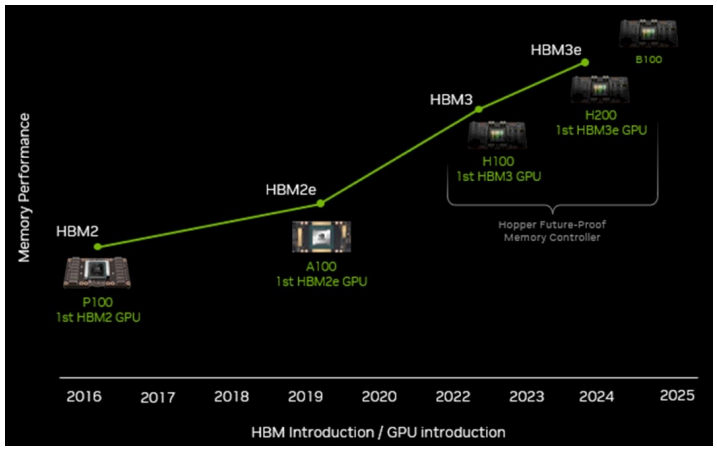
浙江大学 ZHEJIANG UNIVERSITY

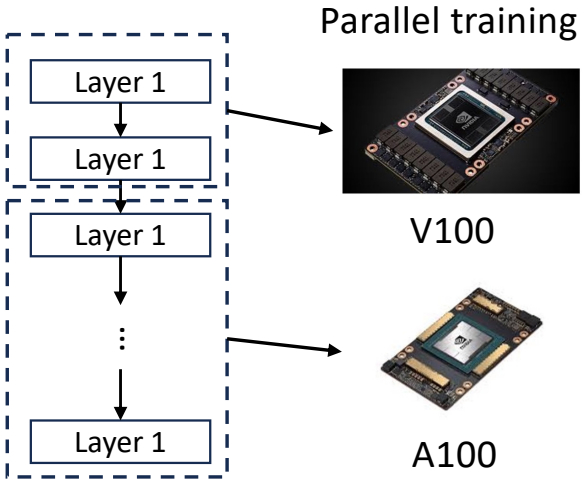
Model Training on Heterogeneous GPU Clusters

New Large Model Training Scenario: Heterogeneous GPU Clusters

- LLMs are growing in size and sequence length
- AI companies buy and deploy new GPUs in training clusters
- Heterogeneous clusters with coexistence of high-end and low-end GPUs



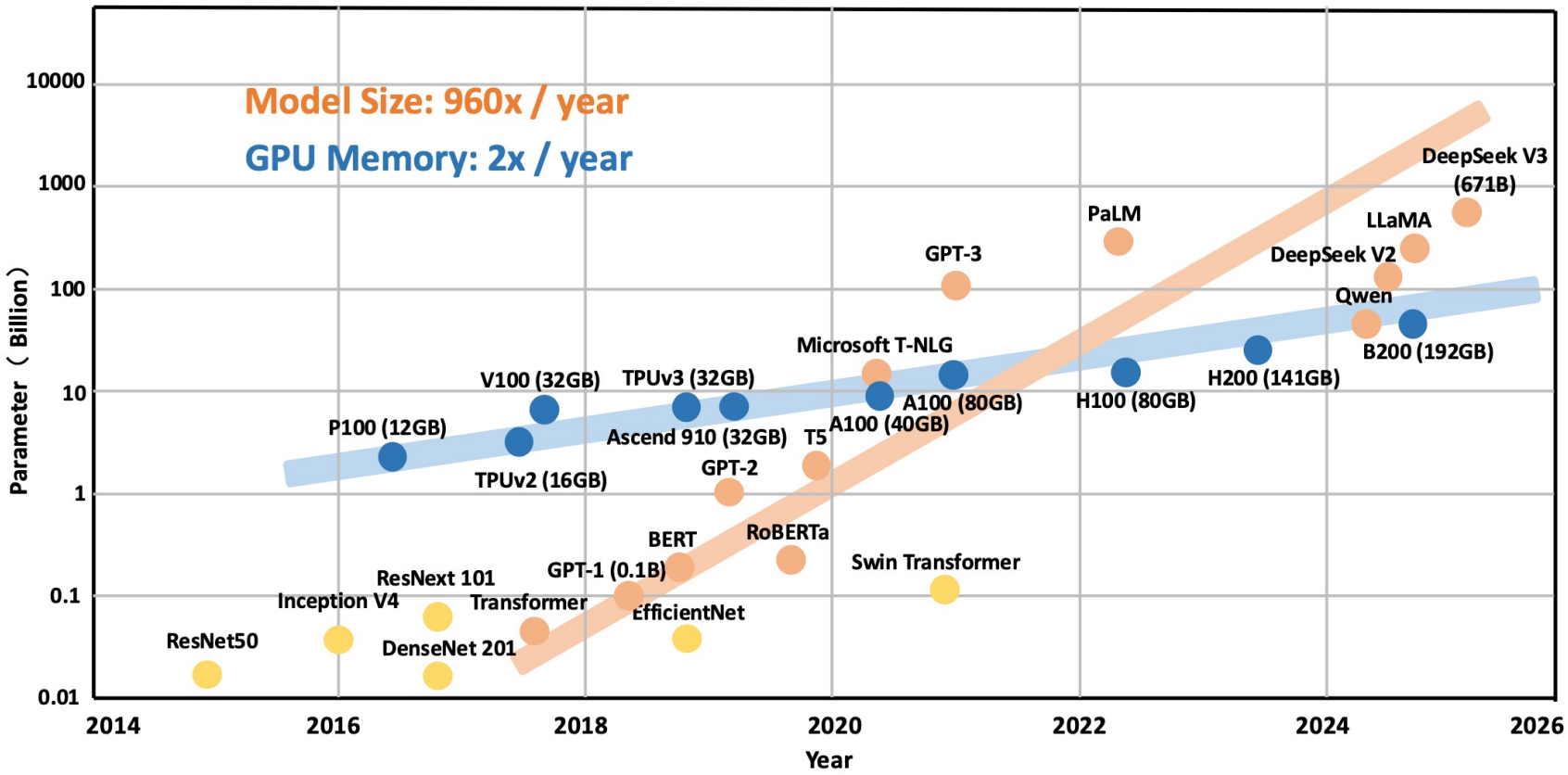




[1] NVIDIA SC23 Special Address

浙江大學 ZHEJIANG UNIVERSITY

Memory Wall in LLM Training

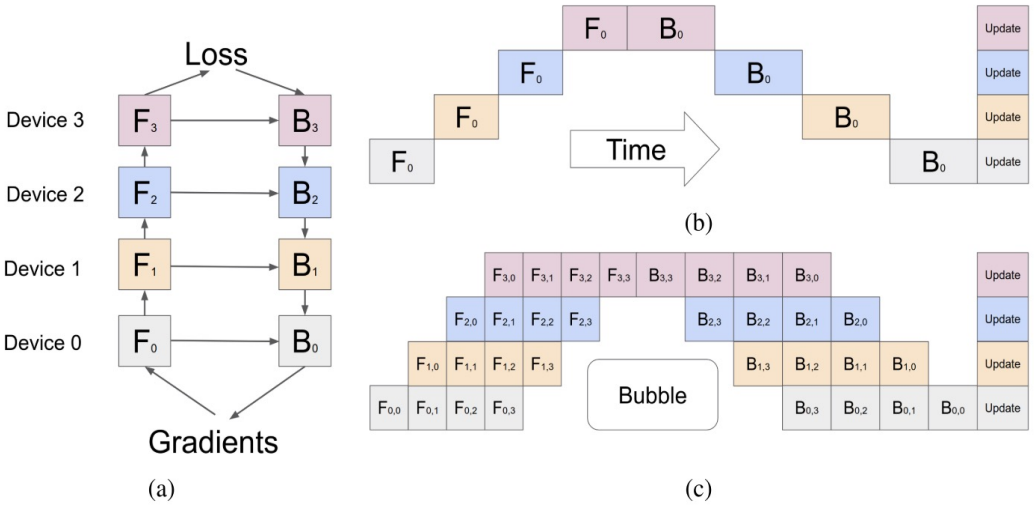


Limited GPU memory restricts the large model training

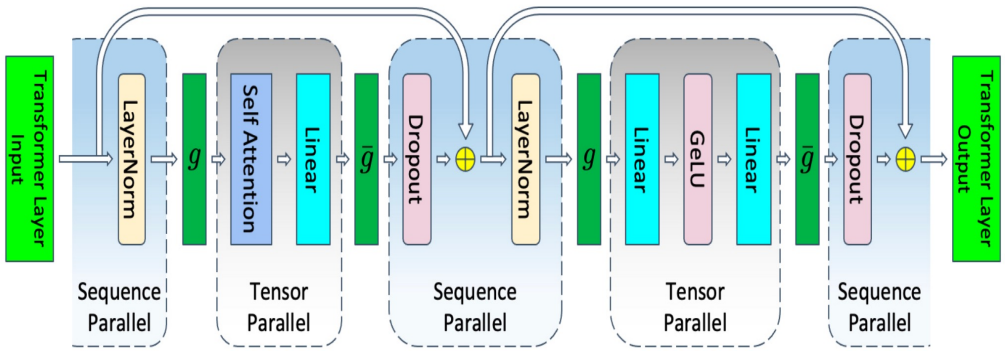
Parallel Training to Reduce Memory Pressure

The foundation distributed LLM training policy

- Pipeline Model Parallelism, vertical partitioning for layers
- Tensor Model Prallelism, horizontal partition for layers



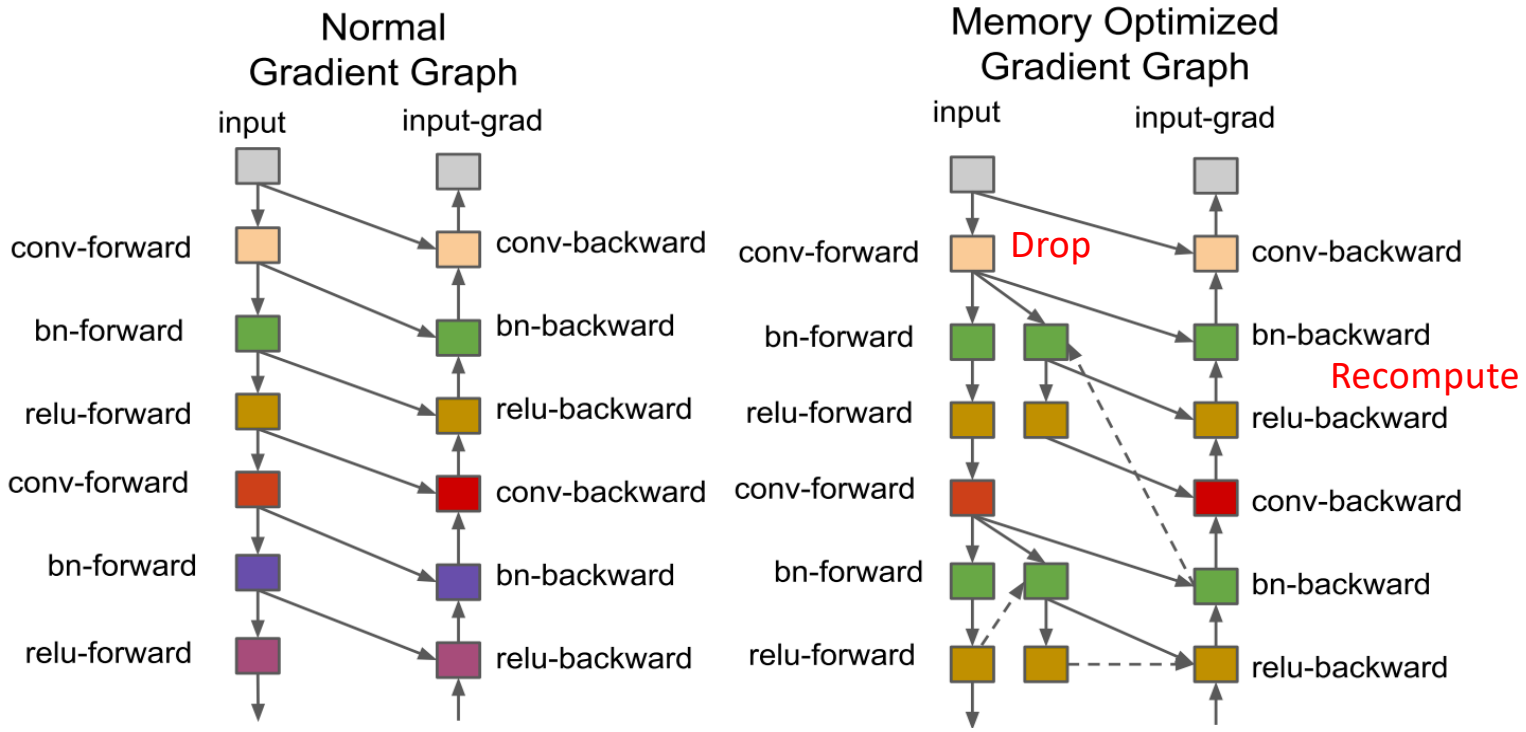
Pipeline Model Parallelism



Tensor Model Prallelism

[1] GPipe: Efficient training of giant neural networks using pipeline parallelism, NIPS19.
[2] REDUCING ACTIVATION RECOMPUTATION IN LARGE TRANSFORMER MODELS, MLsys23.

Activation Recomputation to Reduce Memory Pressure



Activation Recomputation to Reduce Intermediate Data

[1] Training Deep Nets with Sublinear Memory Cost, 2016

浙江大学 ZHEJIANG UNIVERSITY

How to Deploy These Techniques in Heterogeneous Clusters

Pipeline Model Parallelism

Tensor Model Parallelism

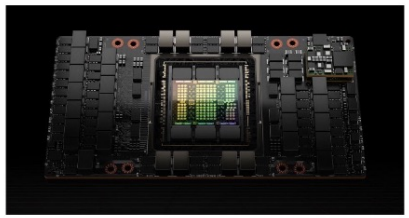
Activation Recomputation

How?

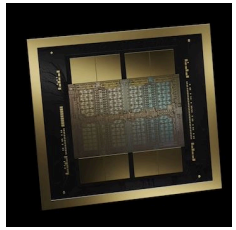
Different memory capacity and computing power

V100

A100



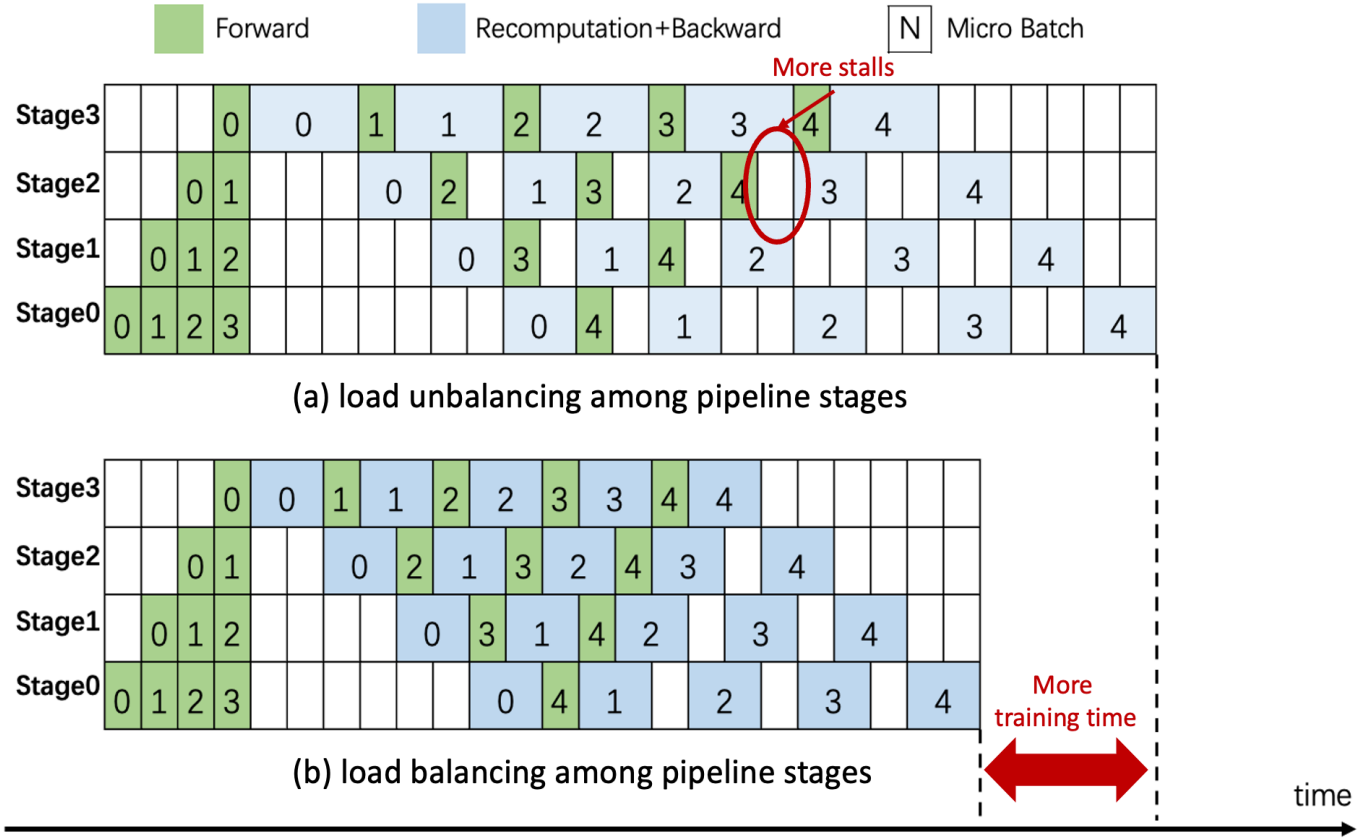
H100



B100

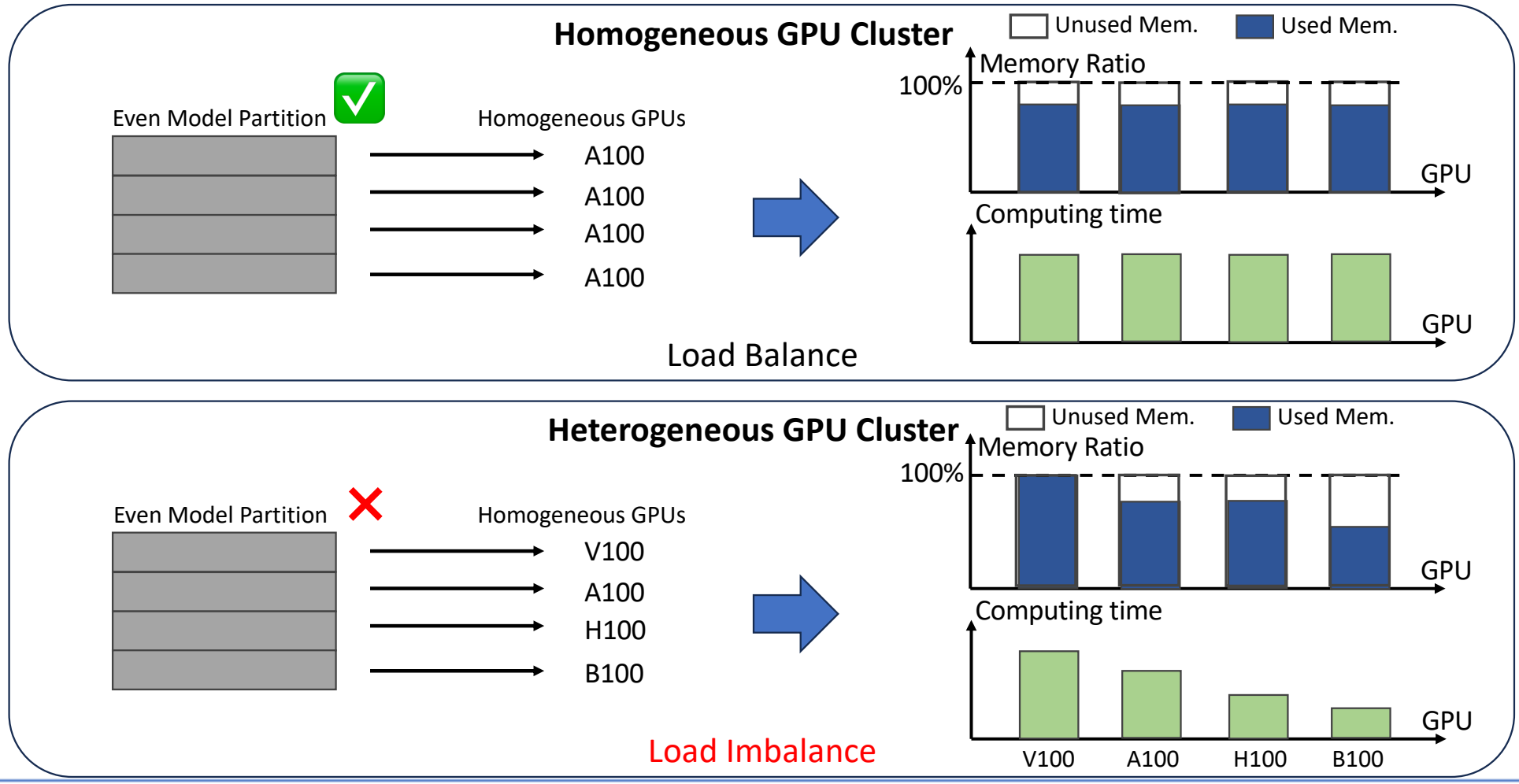
FAST LLM Training

Pipeline Model Parallelism on Heterogeneous Cluster

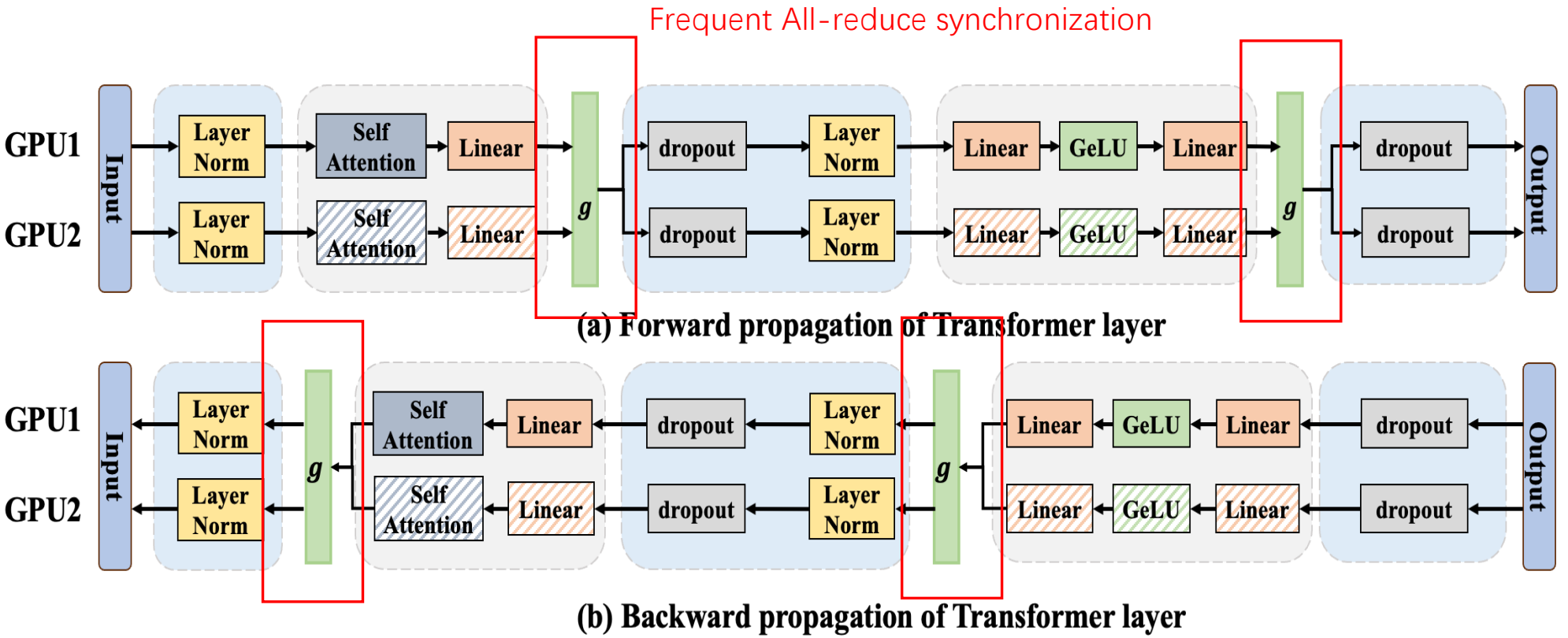


Load Balance for Optimal Pipeline Model Parallelism

Challenge: Pipeline Model Parallelism in Heterogeneous Cluster

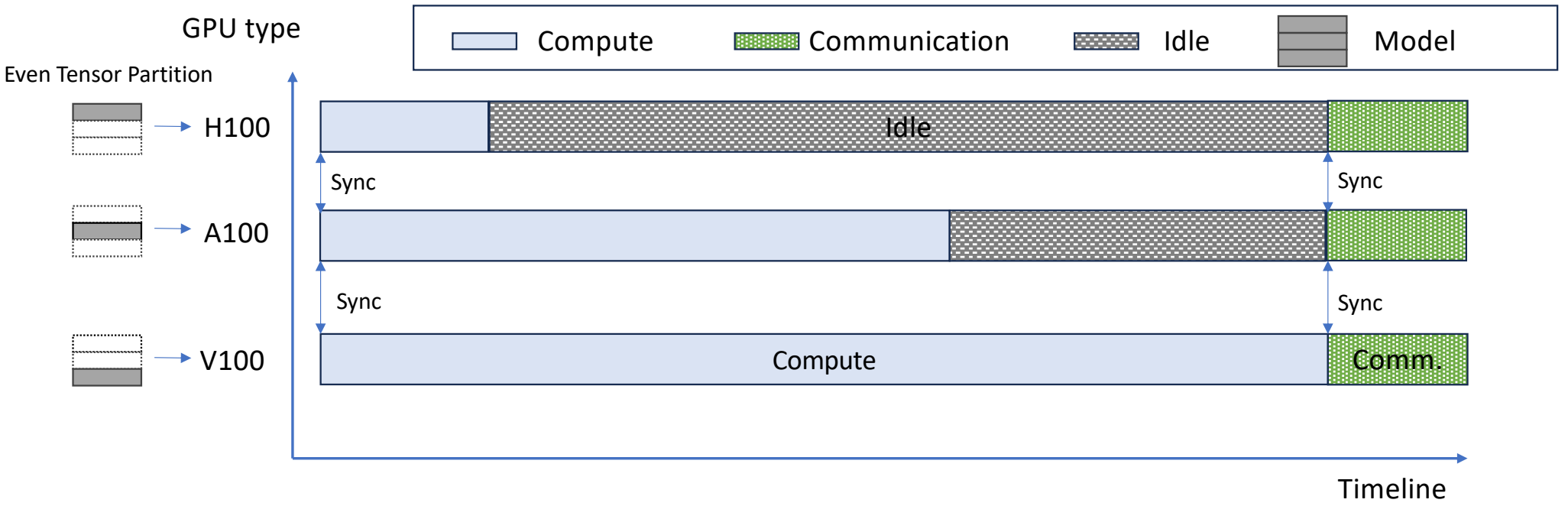


Frequent communications in Tensor Model Parallelism



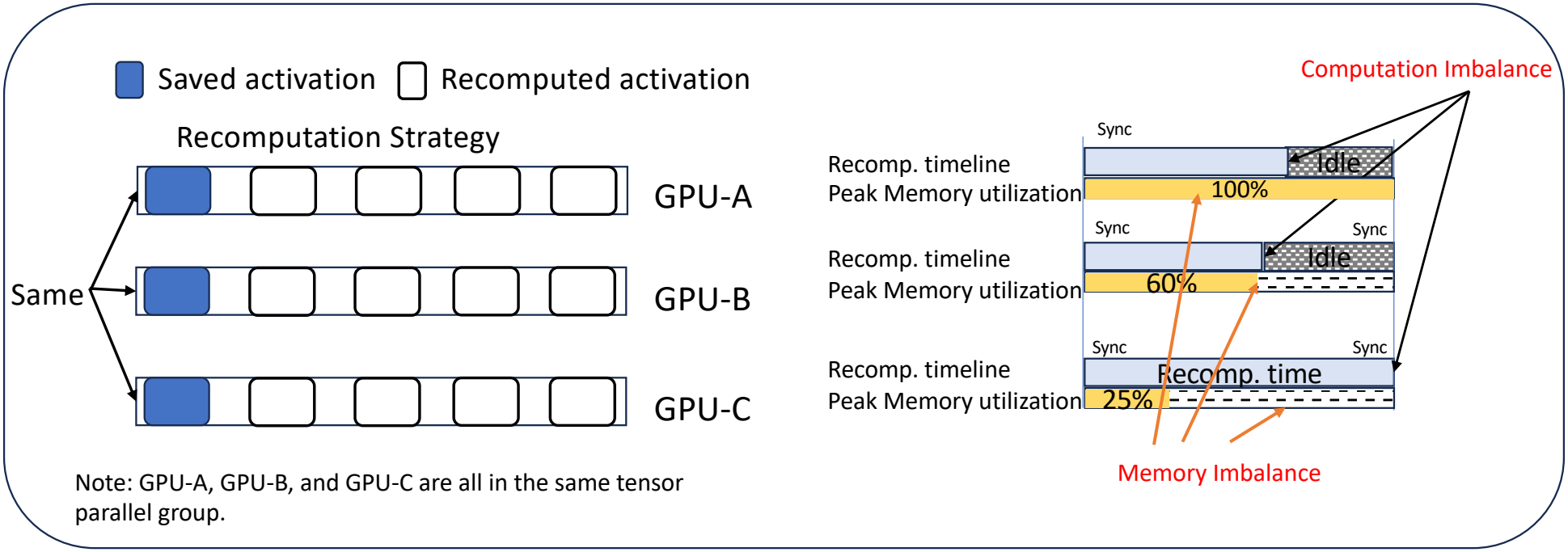
Challenge: Tensor Model Parallelism in Heterogeneous Cluster

Training is constrained by the slowest GPU



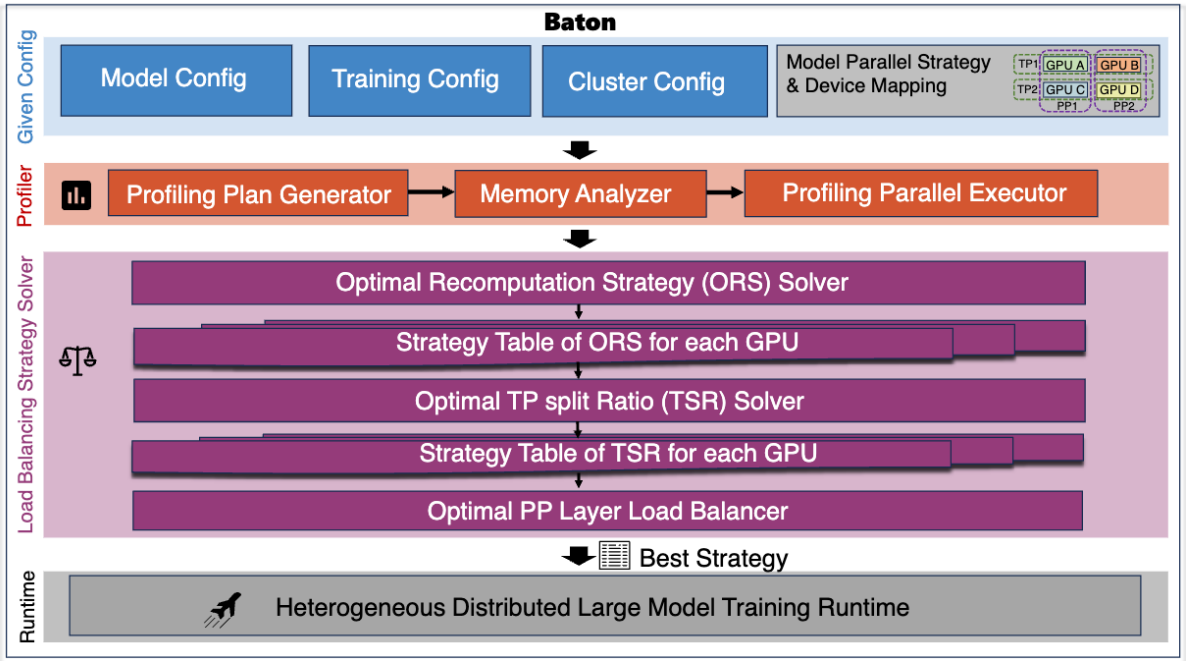
We need a more efficient tensor Partition strategy

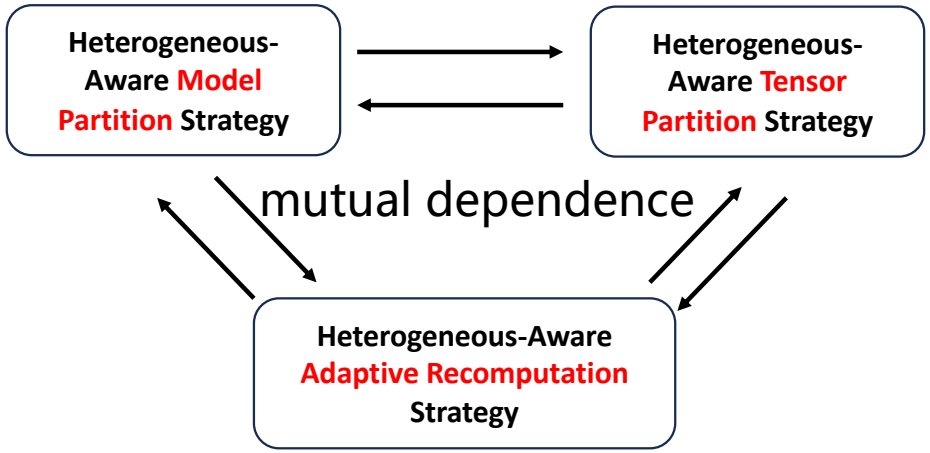
Challenge: Recomputation in Heterogeneous Cluster



Traditional recomputation strategies cause computation and memory imbalance

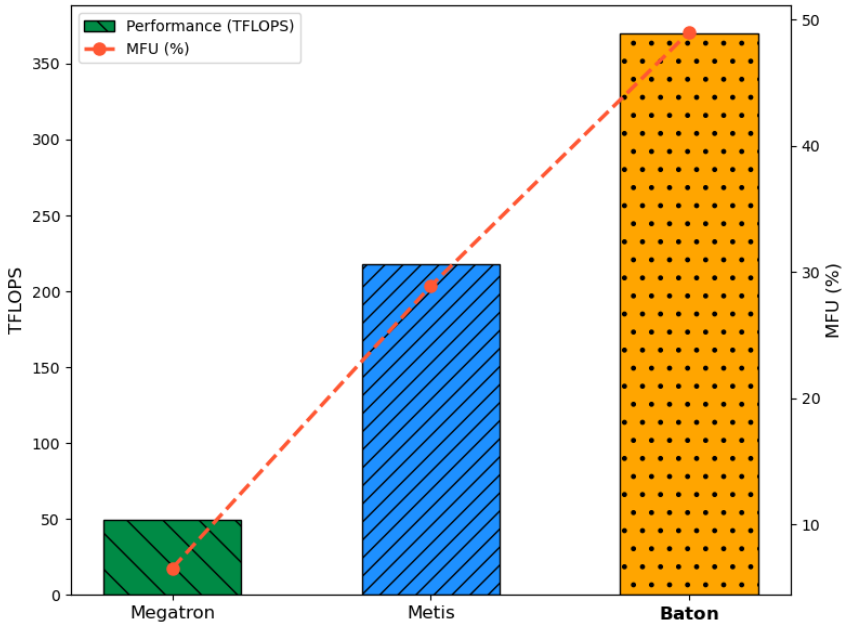
Our Design





Model partitioning + Tensor partitioning + Recomputation

Preliminary Experimental Results



- Setup: 2 A100 and 2 T4 GPUs, training mini GPT-3 1B model
- **Baton** improves the training throughput by **1.69×** compared to SOTA system **Metis[ATC'24]**, and **7.12×** compared to **Megatron-LM**.

Thanks

Please contact zjuchenping@zju.edu.cn for any questions

Yi Zhang Shuibing He **Ping Chen**

Zhejiang University and Zhejiang Lab

之江实验室 ZHEJIANG LAB