HexiScale: Facilitating Large Language Model Training over Heterogeneous Hardware

Ran Yan[1*], Youhe Jiang[1*], Xiaonan Nie[2], Fangcheng Fu[3], Bin Cui[2], Binhang Yuan[1]

1. HKUST, 2. Peking University, 3. Shanghai Jiao Tong University

Introduction & Motivation

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

The homogeneous GPUs are expensive:

- Training LLMs (e.g., Qwen, Llama) demands thousands of GPUs, while homogeneous high-end GPUs are scarce and expensive.

The existing resources are underutilized:

- GPU generations update fast: Turing (2018) → Ampere (2020) → Hopper (2022) → Blackwell (2024).
- Older GPUs remain in service for years (e.g., K80 still on AWS p2 instances).

Key question:

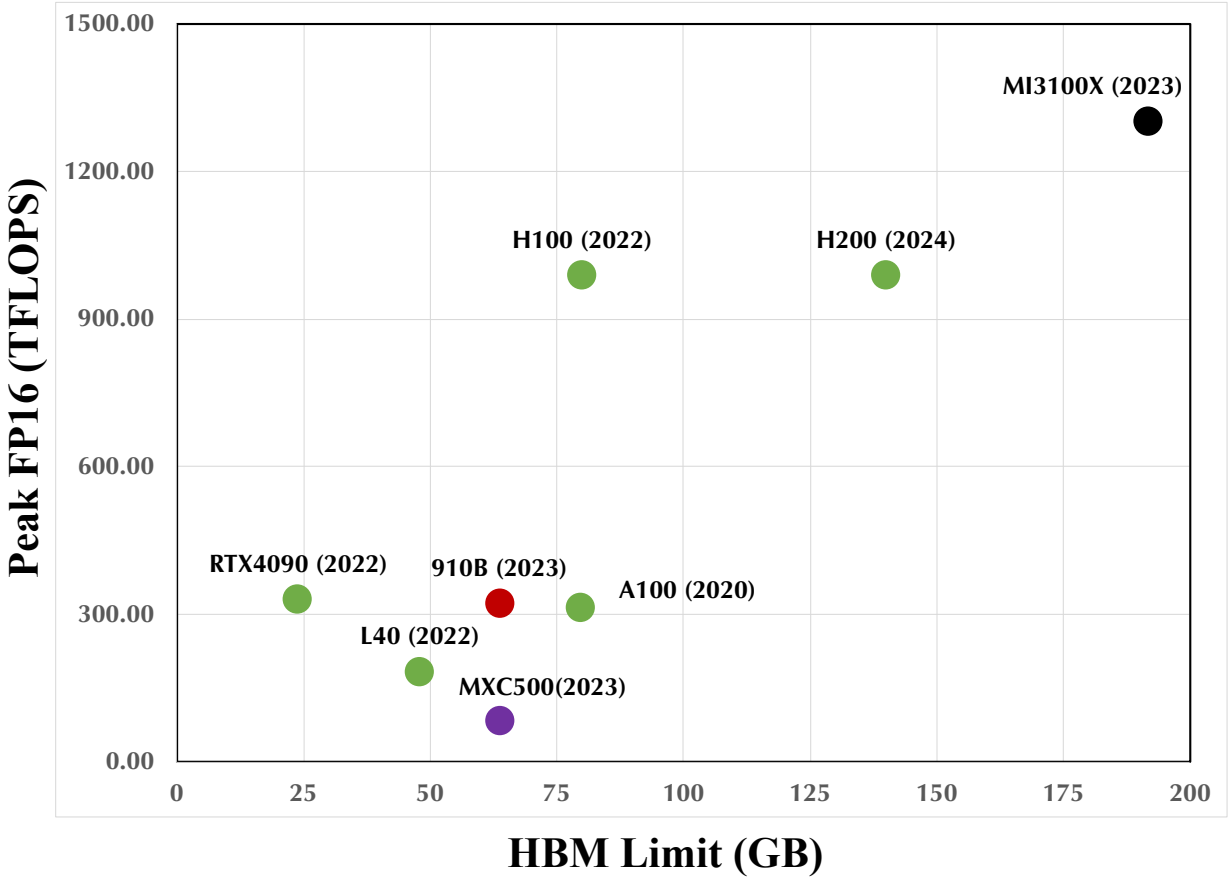
How can we achieve high performance for LLM traning using heterogeneous GPUs?

Challenges

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Challenge 1: Heterogeneous GPUs have varied hardware specs:

- Compute and memory resources demonstrate high level of heterogeneity.
- More capable GPUs may be underutilized without proper management.

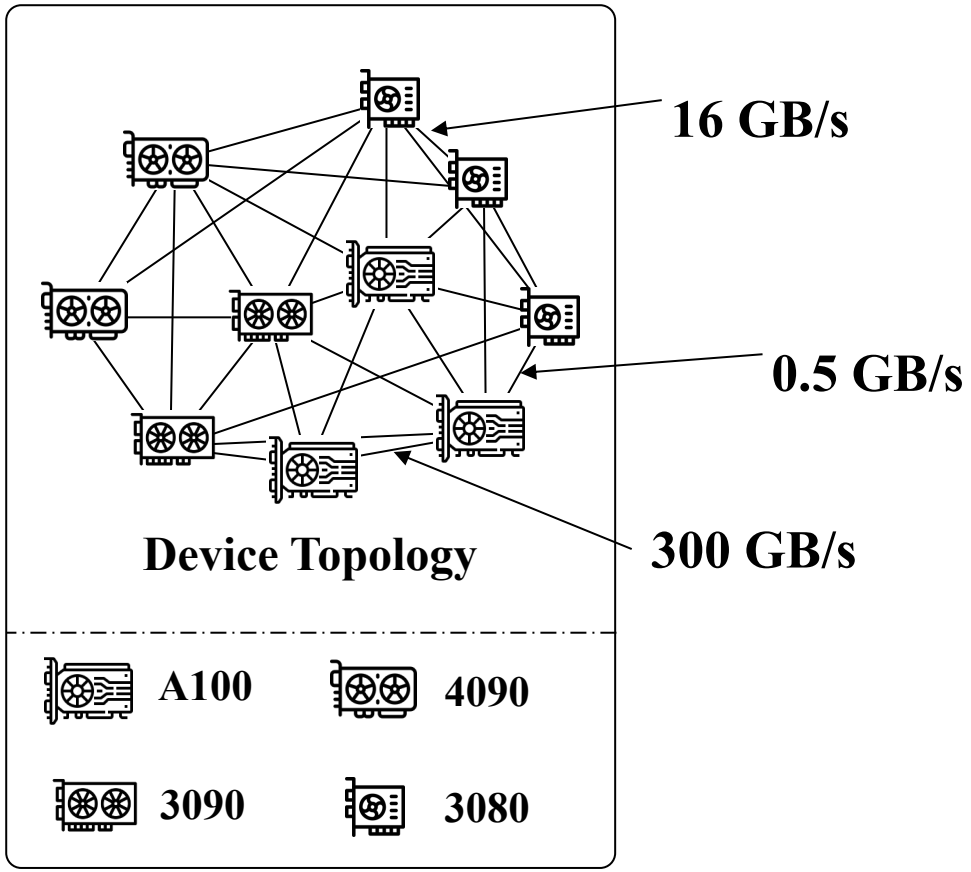


Challenges

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Challenge 2: Heterogeneous GPUs have varied network bandwidth:

- GPU-GPU network bandwidth can vary significantly.
- Effective management of communication overhead is essential to prevent faster connections from stalling.

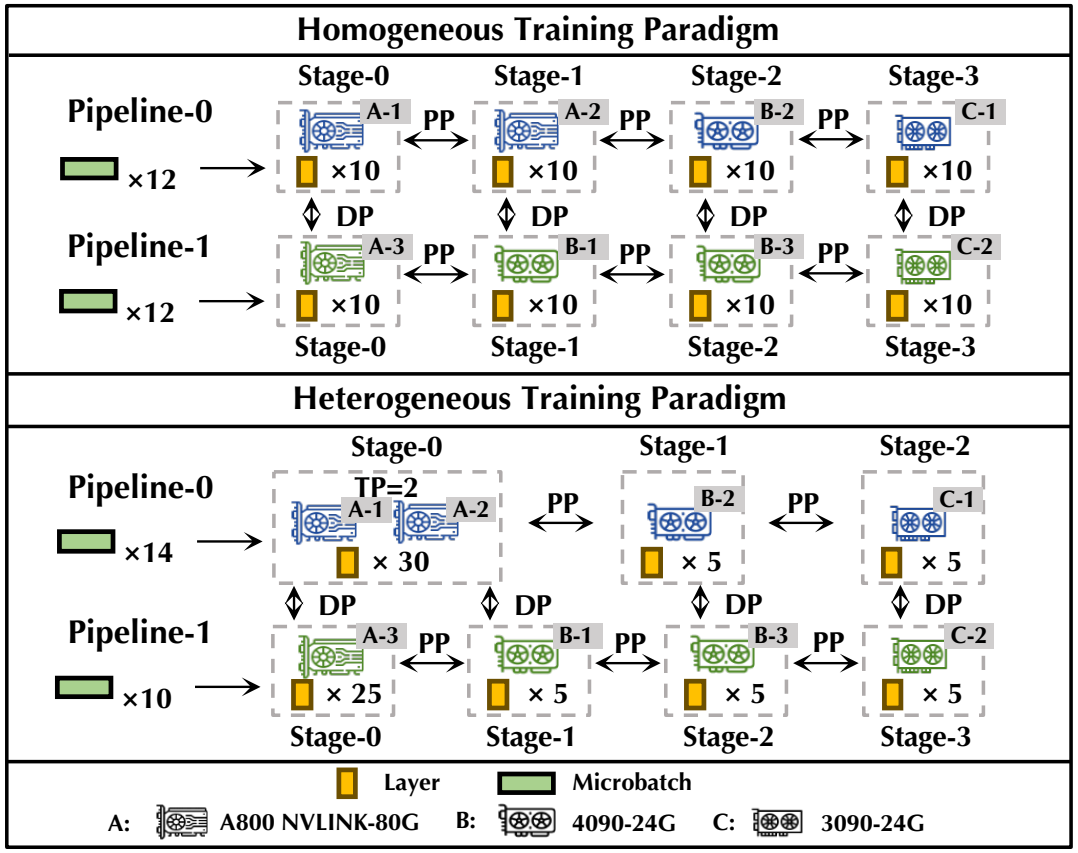


System Design

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Asymmetric parallel support of HexiScale:

- Within one pipeline, PP stages can have varied TP degrees.
- PP stages from different pipelines can have distinct TP degrees.
- Each PP stage can handle arbitrary number of layers.
- Each pipeline can handle uneven batch-size.



Scheduling Algorithm

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Problem Formulation:

- Objective: Given a heterogeneous cluster, we find the optimal placement that have the highest training throughput.
- Formally:

$$\sigma^* = \min_{\sigma} \text{Comm}-\text{Cost}(\sigma) + \text{Comp}-\text{Cost}(\sigma)$$
$$s.t.\ \ \text{Mem}-\text{Cumsum}(d) \leq m_d \ \forall d \in \boldsymbol{D}$$

- d is the GPU device, m_d is the memory limit, $\boldsymbol{D}$ is the heterogeneous devices set.
- σ is the parallel configuration, σ^* is the optimal parallel configuration.
- $\text{Comm}-\text{Cost}(\sigma)$ is the communication cost, $\text{Comp}-\text{Cost}(\sigma)$ is the computation cost, $\text{Mem}-\text{Cumsum}(d)$ is the memory cost of device d.

Scheduling Algorithm

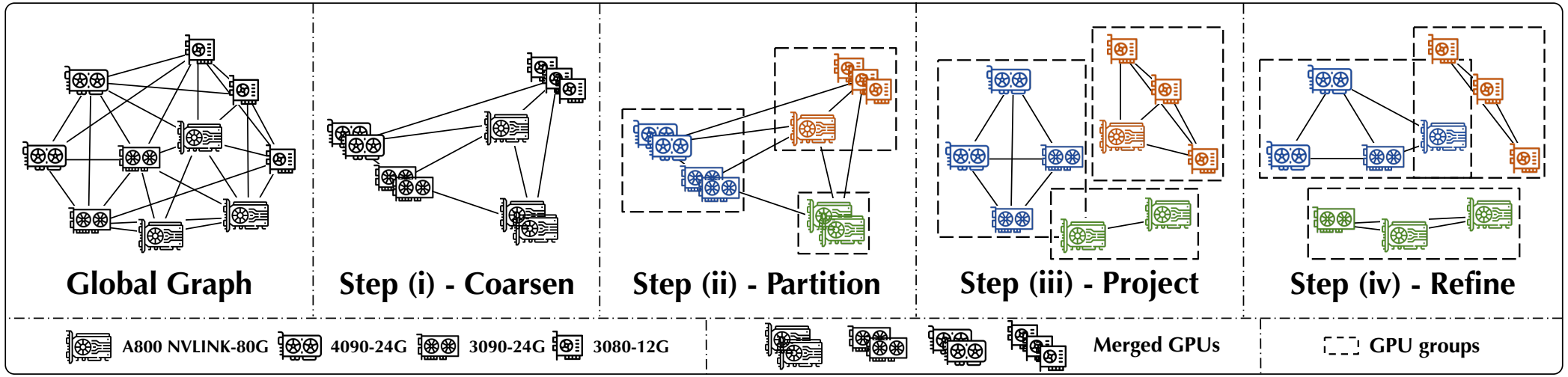
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Phase-1, global graph partitioning:

- Model GPUs as graph: vertices weighted by FLOPS, edges by pairwise bandwidth.
- Partition into D_{dp} groups (each for one pipeline) via 4-step graph partition algorithm.

Iterative optimization from phase-1:

- Enumerate D_{dp} to optimize the number of pipelines.
- Adaptively allocate bandwidth for DP and PP.



Scheduling Algorithm

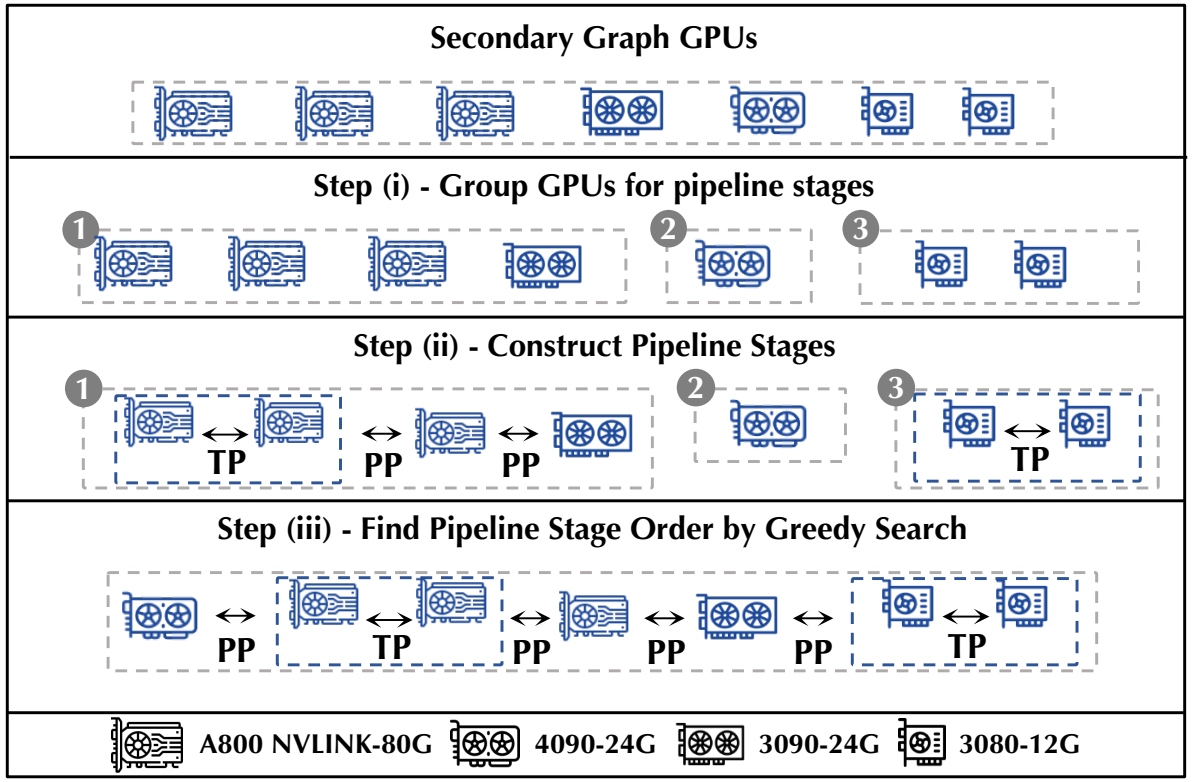
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Phase-2, pipeline construction:

- (i) Group high-bandwidth GPUs via secondary graph partition.
- (ii) Search intra-group TP/PP/layer strategy per machine.
- (iii) Top-τ greedy search for optimal stage ordering.

Iterative optimization from phase-2:

- Construct different GPU groups through different secondary graph partition.



Evaluations

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Compare with baselines over homogeneous clusters:

- Baseline frameworks: Megatron, Galvatron and FSDP.
- Comparison scheme: Under the same total FLOPS, we compare the PFLOPS of HexiScale over *heterogeneous clusters* and baselines over *homogeneous clusters*.

Experimental settings:

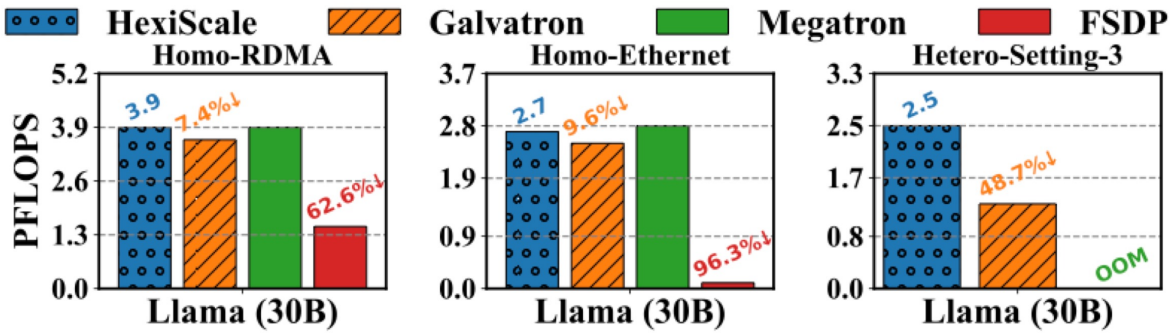
Setting	GPU Configuration	FLOPS vs Baseline	Target Homogeneous Baseline Cluster	Target Model(s)
Hetero-Setting 1	1×(8×3080Ti) + 1×(8×3090) + 3×(8×4090)	**+1.36%** higher total FLOPS	2×(8×A800 PCIe-80G)	Llama-2 7B / 13B
Hetero-Setting 2	1×(8×3080Ti) + 1×(8×3090) + 1×(8×4090) + 1×(8×A800 NVLINK-80G)	**−1.59%** lower total FLOPS	2×(8×A800 PCIe-80G)	Llama-2 7B / 13B
Hetero-Setting 3	1×(8×3090) + 2×(4×3090) + 4×(8×4090) + 1×(8×A800 NVLINK-80G)	**−4.67%** lower total FLOPS	4×(8×A800 PCIe-80G)	Llama 30B

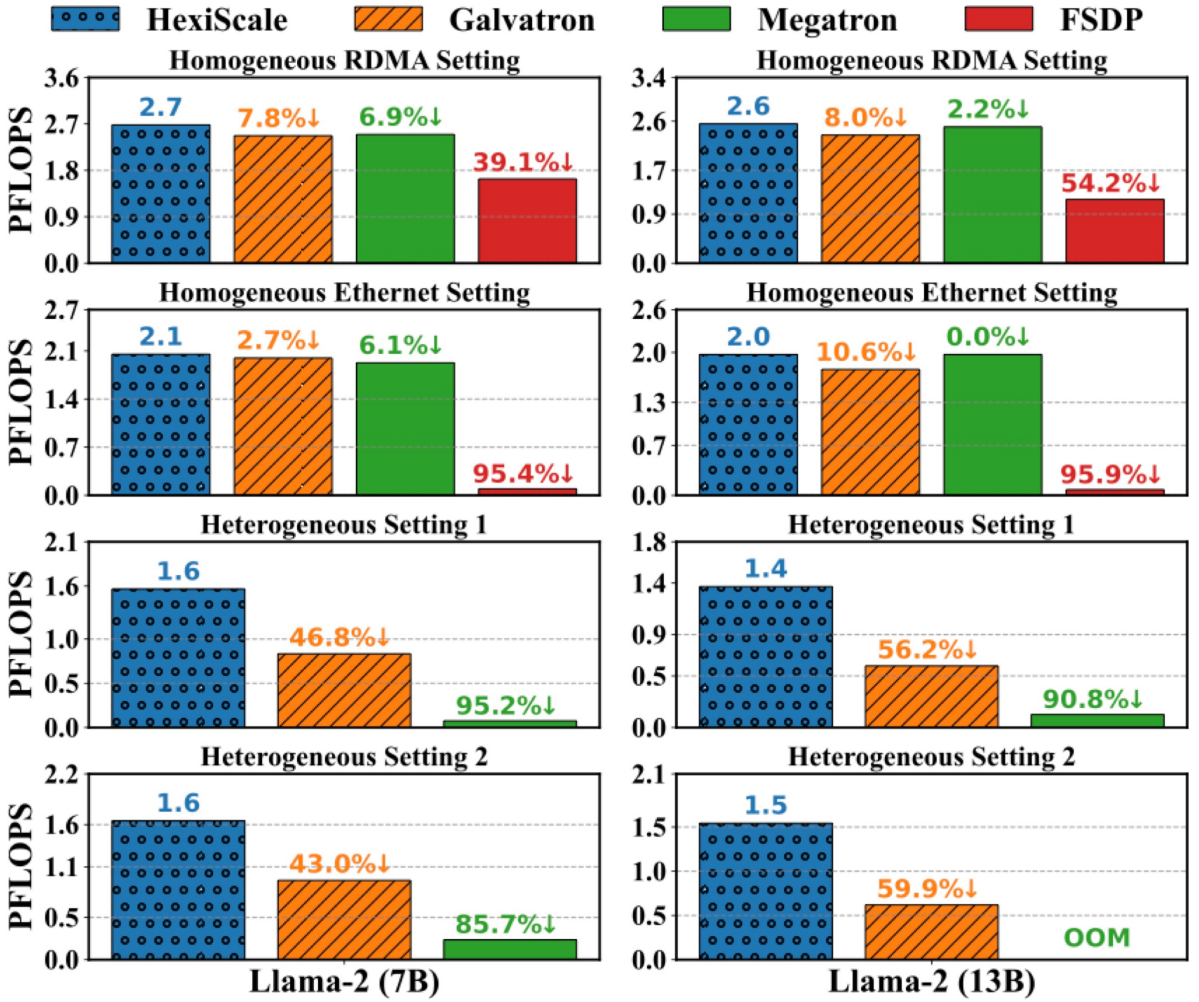
Evaluations

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Evaluation results against homogeneous baselines:

- HexiScale exhibits up to 1.01× throughput and an average throughput gap of 0.83× compared to homogeneous scenarios.





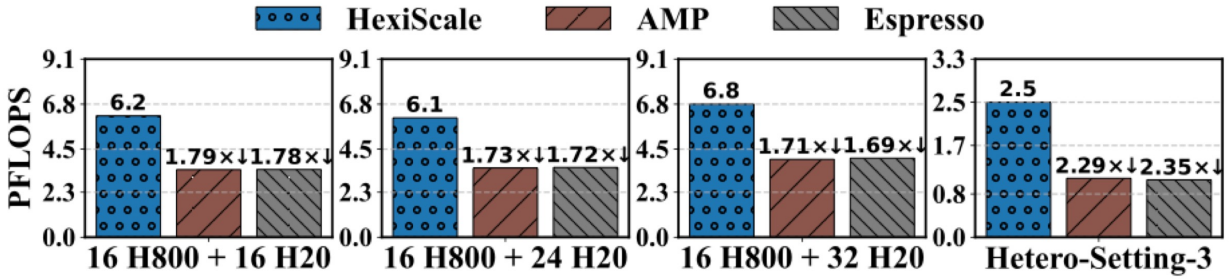
Evaluations

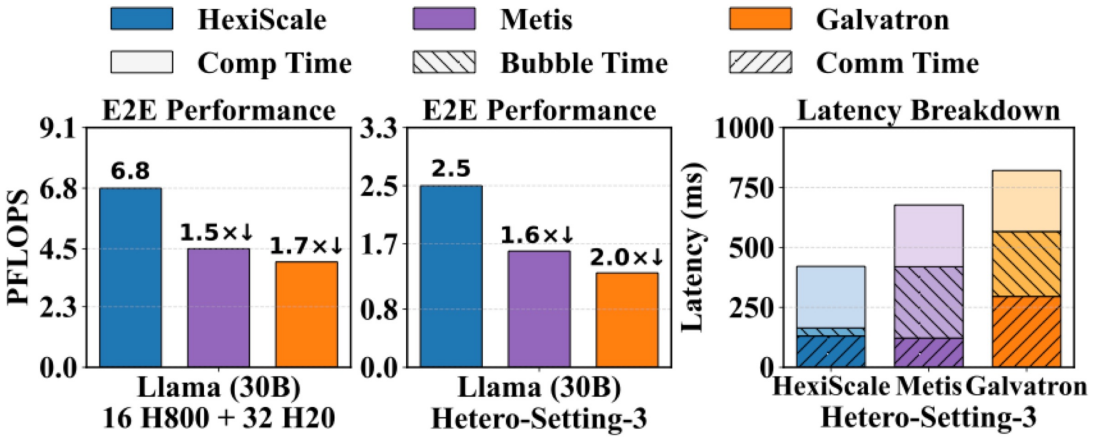
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Compare with heterogeneity-aware baselines over heterogeneous clusters:

- Baseline frameworks: Metis, AMP, and Espresso.
- Comparison scheme: Under the same heterogeneous clusters, we compare the PFLOPS of HexiScale against the heterogeneity-aware baselines.

Experimental results:



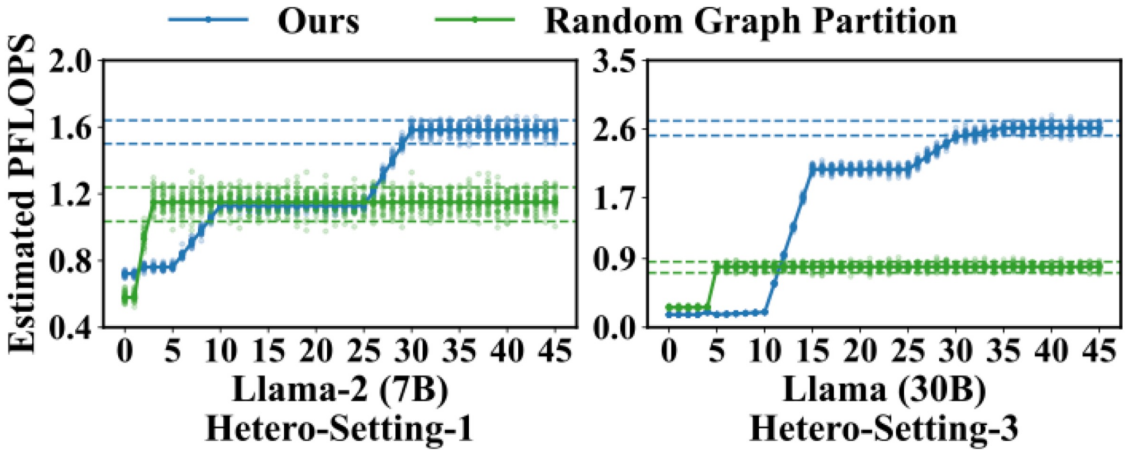


Evaluations

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

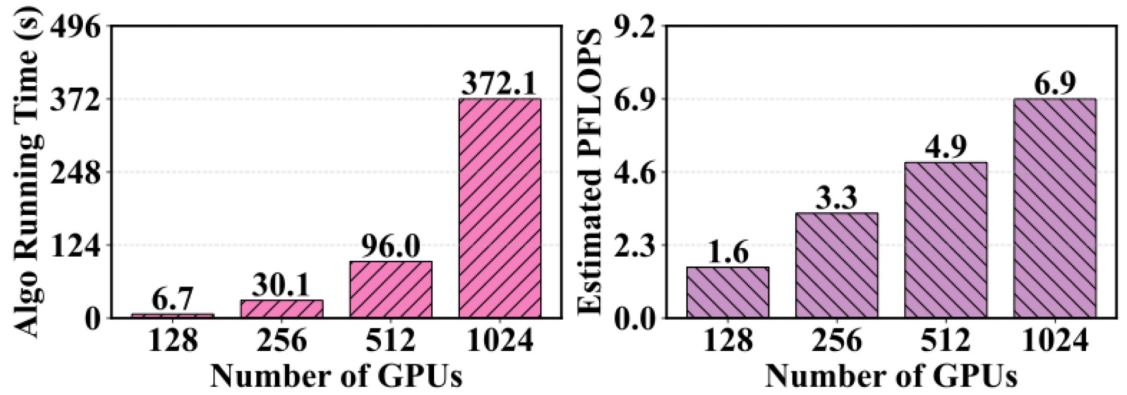
Algorithm effectiveness evaluation:

- Compared to Random Graph Partition algorithm, HexiScale's algorithm demonstrates earlier convergence and 1.3x – 3.3x higher estimated throughput.



Algorithm efficiency evaluation:

- HexiScale's scheduling time for large clusters remain manageable.
- HexiScale demonstrates strong scalability across large GPU clusters.



Summary

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Benefit of using heterogeneous GPUs for LLM training:

- Integrate computational resources from different GPU generations.

Key challenges:

- Heterogeneity in hardware specs and network connections may lead to underutilization of resources.

HexiScale solutions:

- Design and implement flexible asymmetric parallelism for heterogeneous GPUs.
- Design a novel graph partition algorithm to identify the efficient parallel strategies.

Thank You

Questions & Discussion

Contact: ryanaf@connect.ust.hk