sigcomm 2025
COIMBRA

# MegaScale-Infer: Efficient Mixture-of-Experts Serving with Disaggregated Expert Parallelism

Ruidong Zhu*, Ziheng Jiang*, Chao Jin*, Peng Wu, Cesar A. Stuardo, Dongyang Wang, Xinlei Zhang, Huaping Zhou, Haoran Wei, Yang Cheng, Jianzhe Xiao, Xinyi Zhang, Lingjun Liu, Haibin Lin, Li-Wen Chang, Jianxi Ye, Xiao Yu, Xuanzhe Liu, Xin Jin, Xin Liu


北京大学
PEKING UNIVERSITY

ByteDance

# Latest Large-Scale LLMs

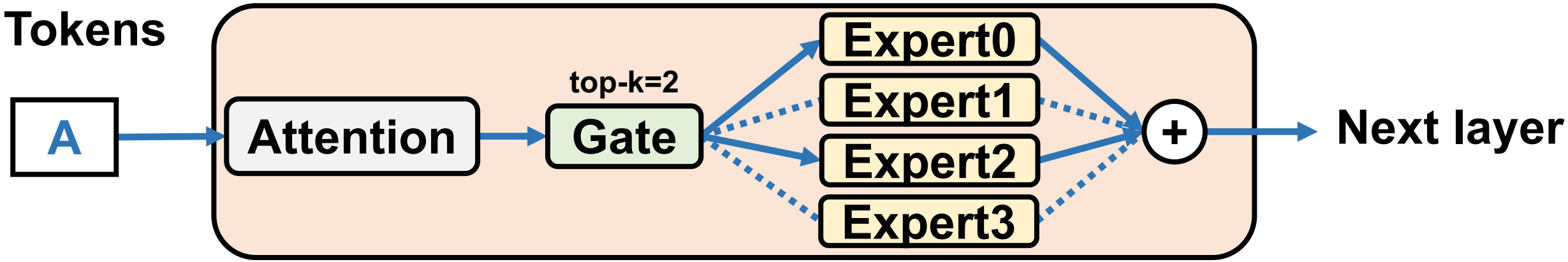
- Doubao-Seed-1.6
- DeepSeek v3.1
- GPT-OSS
- Gemini 2.5 Pro



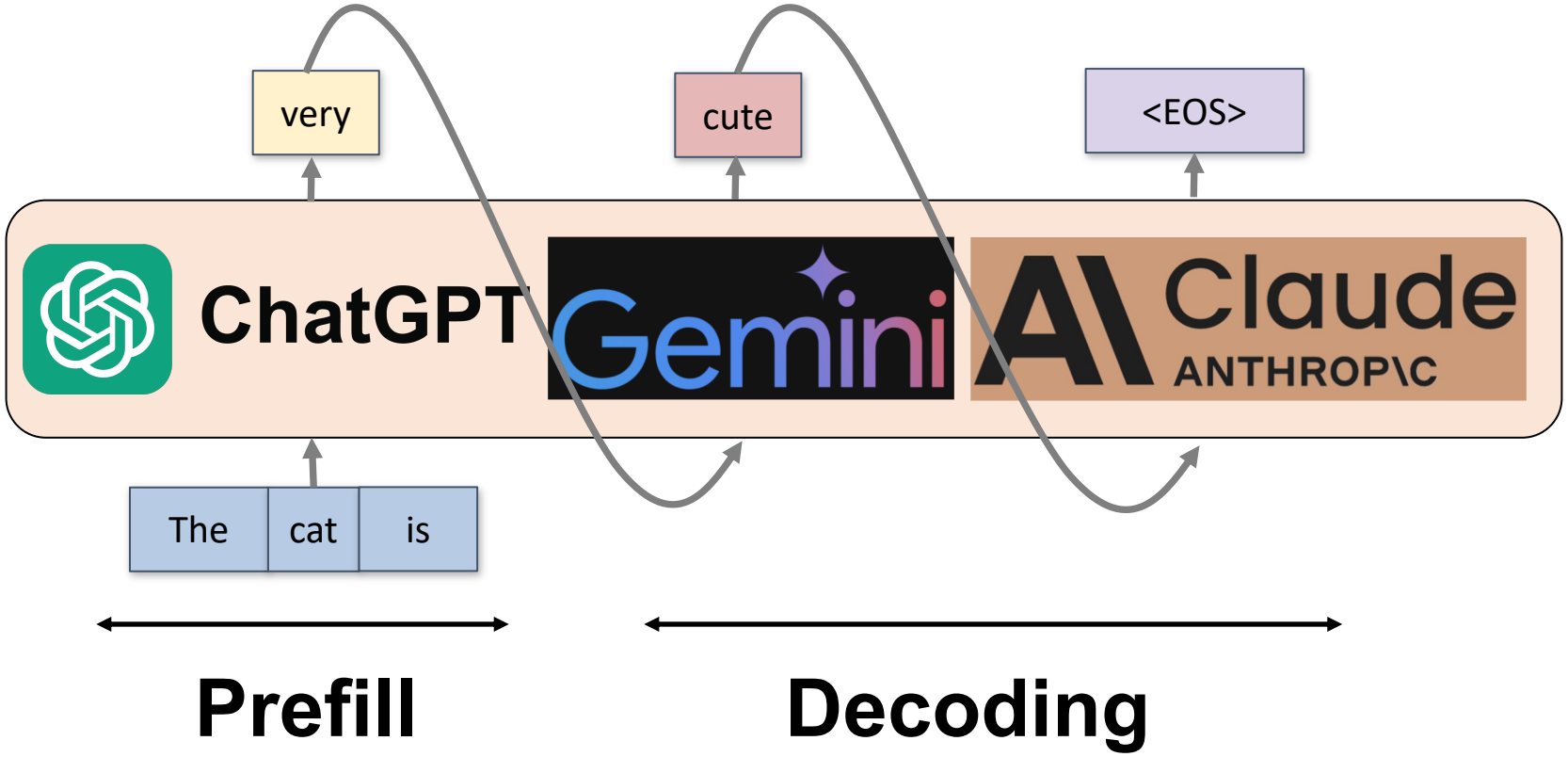
**All adopt the Mixture-of-Experts (MoE) architecture!**

# Mixture-of-Experts (MoE)

- The mainstream architecture of large-scale LLMs
- Each token is only sent to the top-k experts



# Autoregressive Generation of LLM

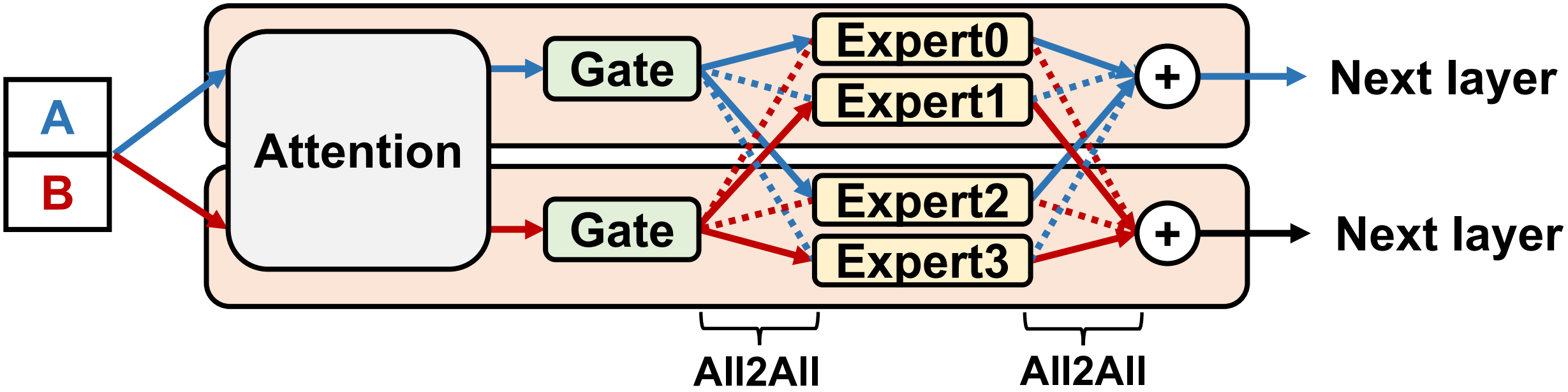


# Characteristics of Two Phases in LLM inference

| | Prefill | Decoding |
|---|---|---|
| **Attention** | Compute-intensive | Memory-intensive |
| **FFN** | Compute-intensive | From memory-intensive to compute-intensive as the batch size increases |

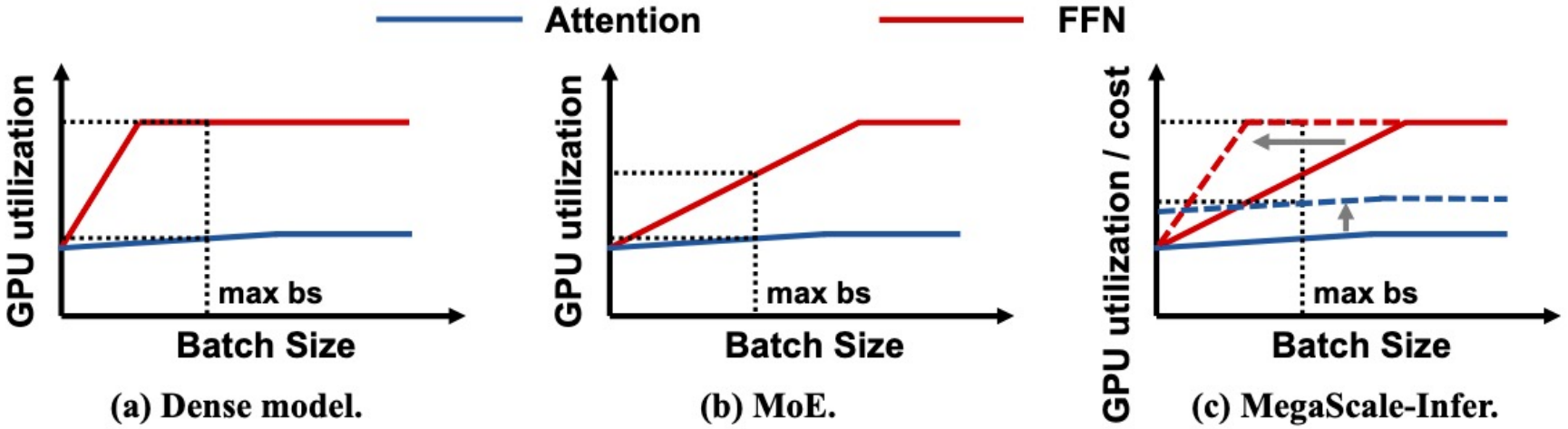
# How to Achieve Efficient MoE Inference

- PD Disaggregation: Execute the prefill and decoding phases on different nodes
- Expert parallelism: Partition the parameters to different nodes at the granularity of experts



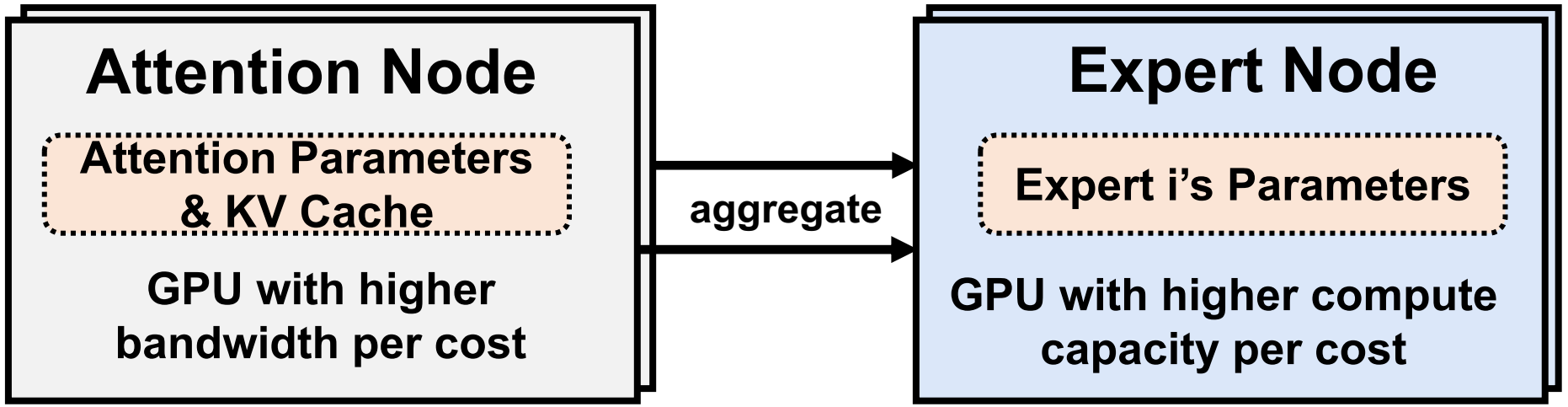
# Issues in the Decoding Phase of MoE Inference

- Low GPU utilization
  - Attention is memory-intensive
  - Sparsity results in a small batch size for each expert and low computational intensity



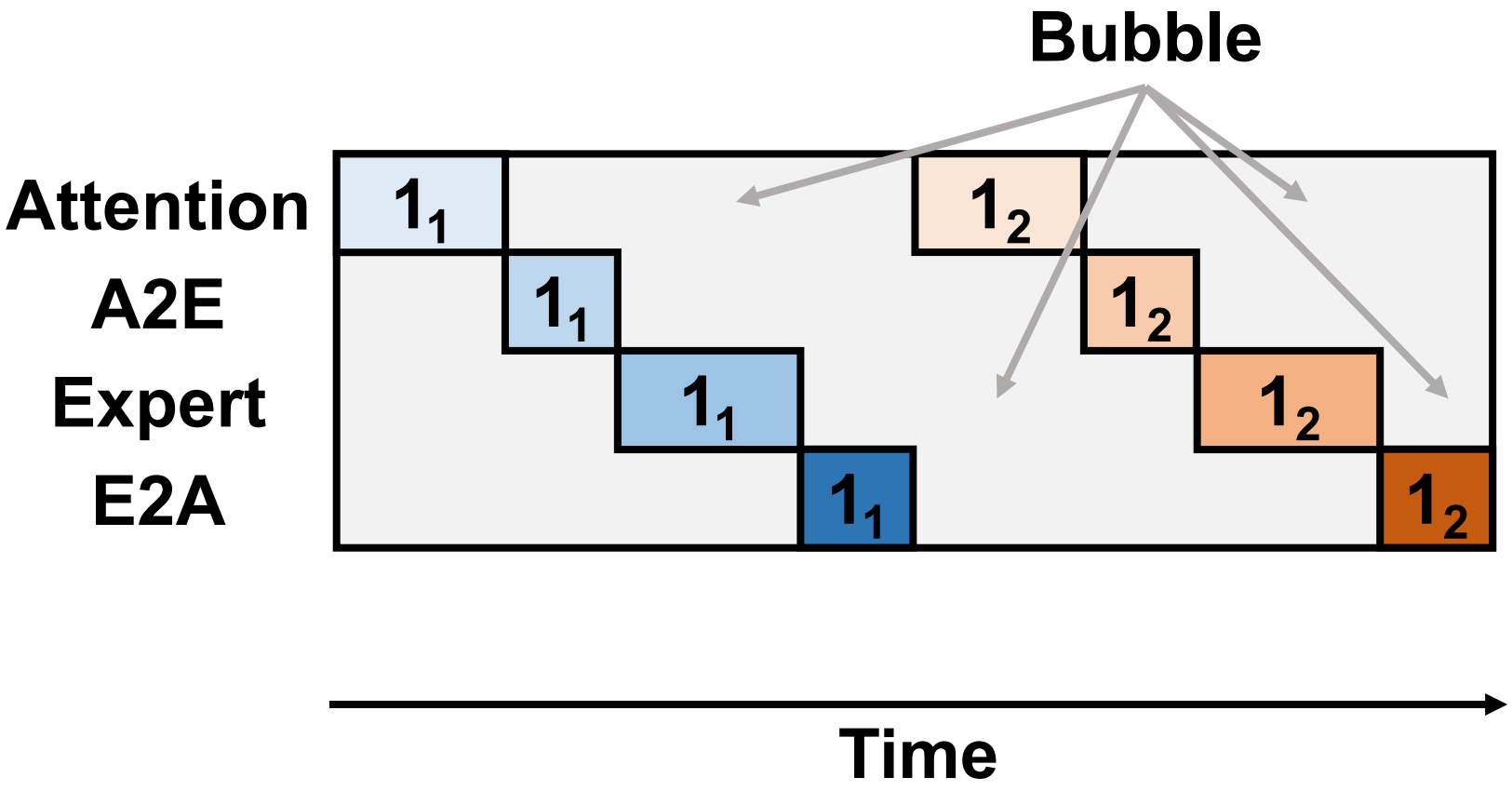
# Solution: Disaggregate Attention and FFN

- Independent scaling: Aggregating multiple attention requests can improve the computational efficiency of FFN
- Heterogeneous deployment: Adopt more cost-effective hardware for each module



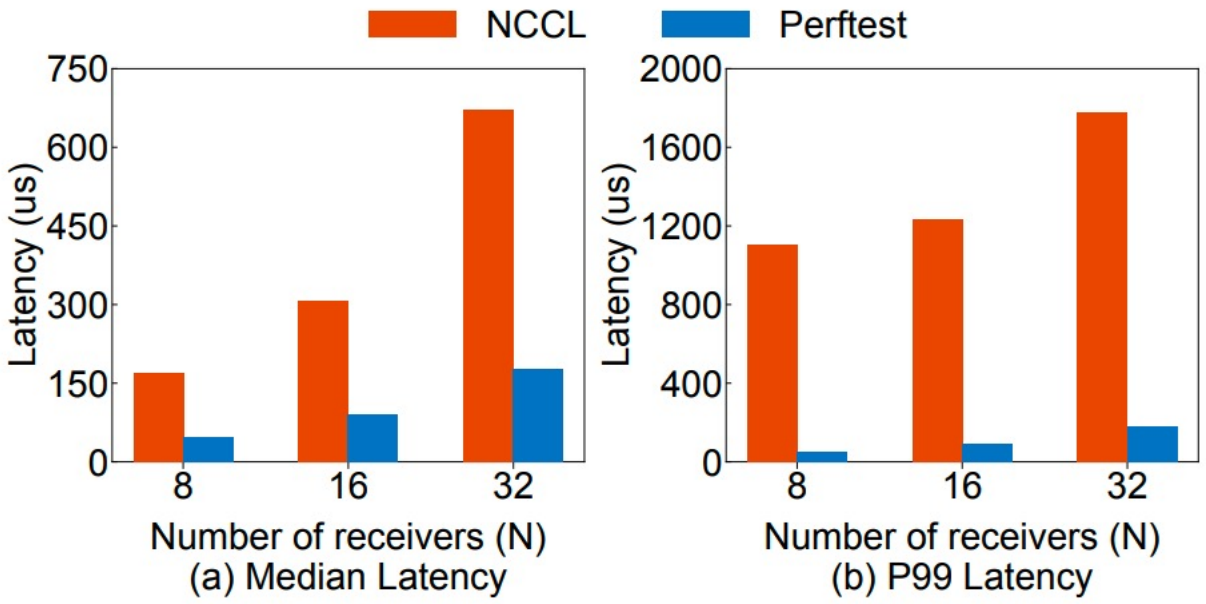
# Challenge 1: Idle Resource Due to Dependencies

- Sequential computation of a batch will result in only a portion of the resources being utilized at the same time



# Challenge 2: Requirement of High-Performance M2N Communication

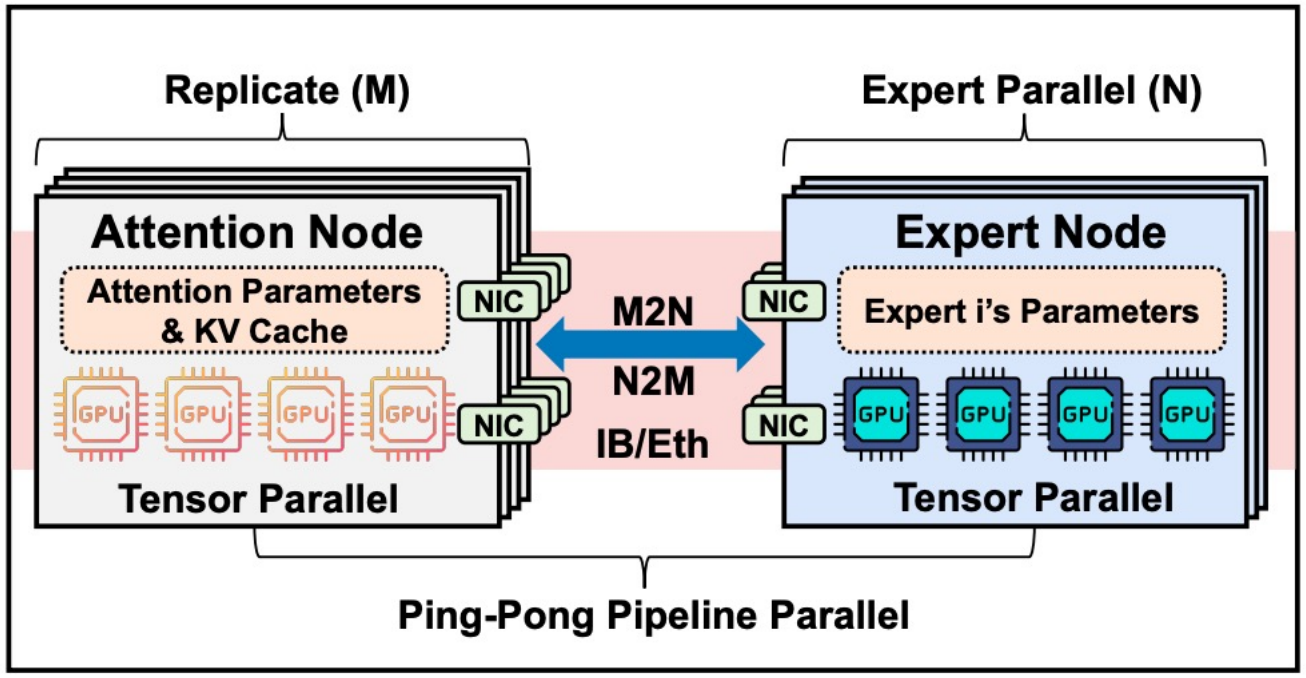
- Problems with existing communication libraries (NCCL):
  - Additional overhead such as data copy and group operation
  - High degree of instability




One sender sends 128K bytes of data to N receivers

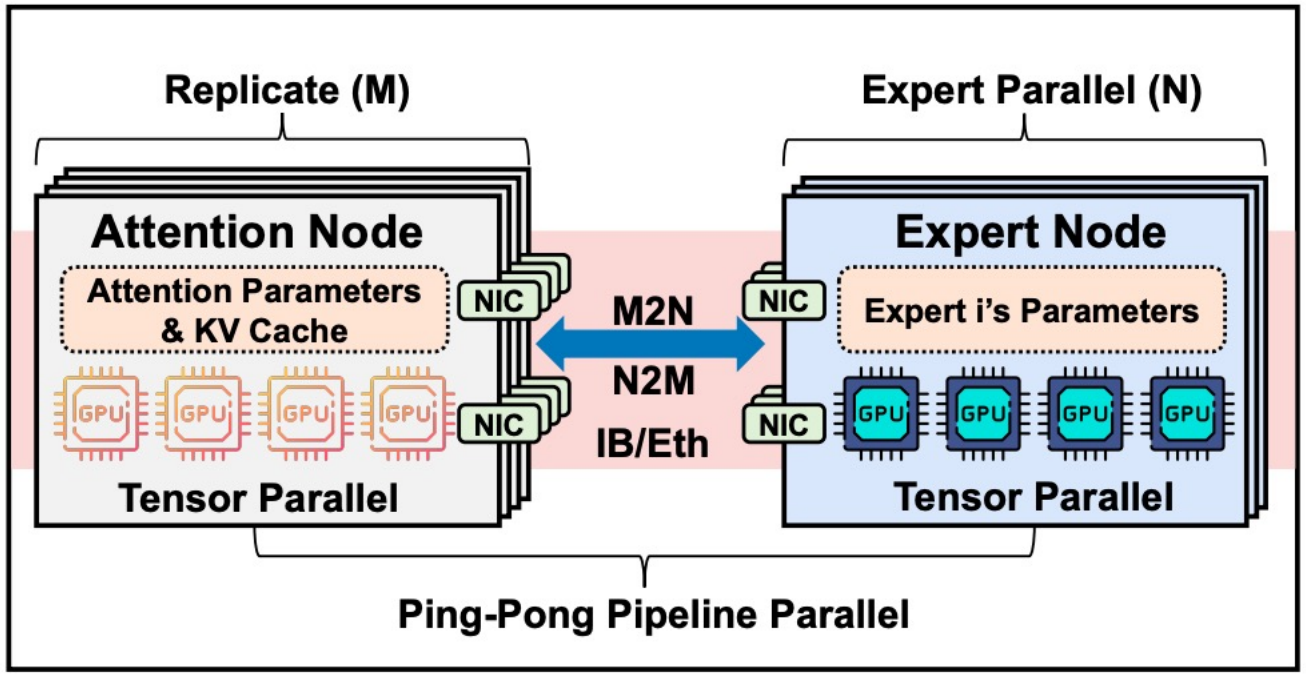
# MegaScale-Infer

- Disaggregated expert parallelism
- Ping-pong pipeline parallel
- High-performance M2N communication library



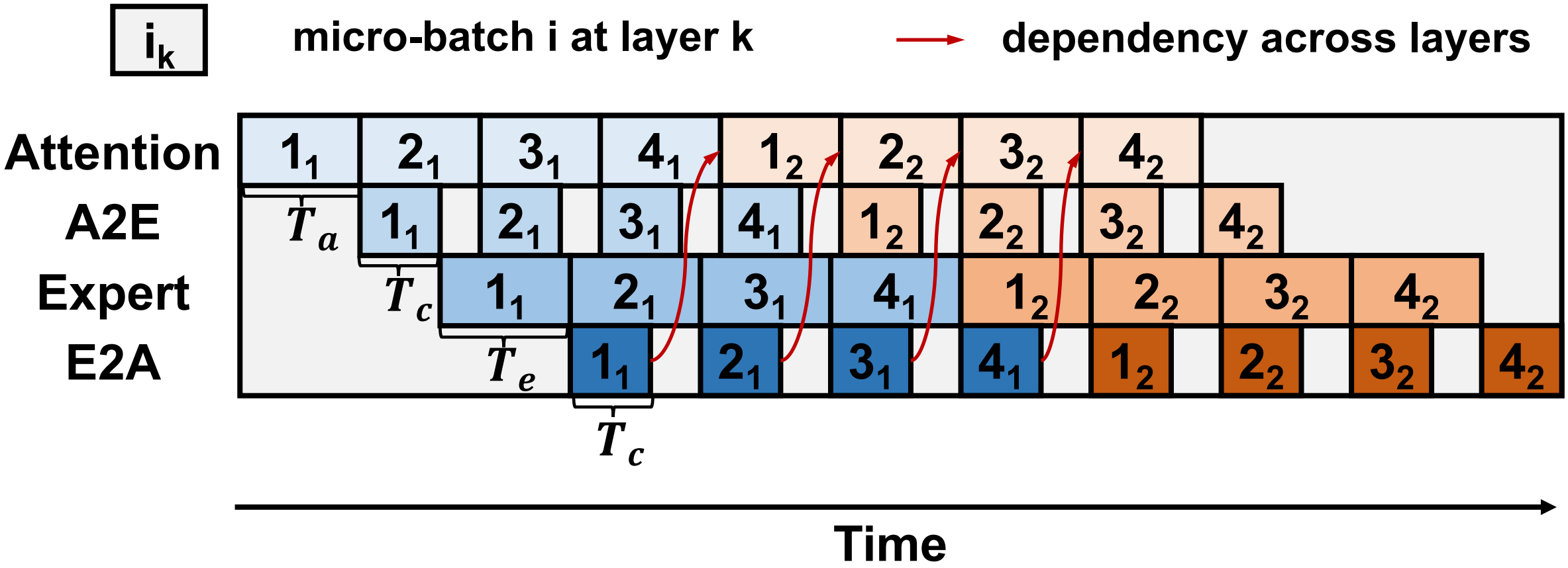
# Disaggregated Expert Parallelism

- Multiple attention nodes process different requests in a replicated manner
- Expert nodes adopt expert parallelism, and each expert receives tokens from multiple attention nodes



# Ping-Pong Pipeline Parallel

- Make full use of resources and overlap communication with computation through multiple micro-batches



# Deployment Plan Search

- Objective: Maximize the throughput per unit cost while meeting the SLOs

- Constraints:
  - $\boldsymbol{T_a \approx T_e}$
  - $\boldsymbol{T_c < T_a}$
  - Sufficient number of micro-batches
  - Enough memory capacity
  - ……

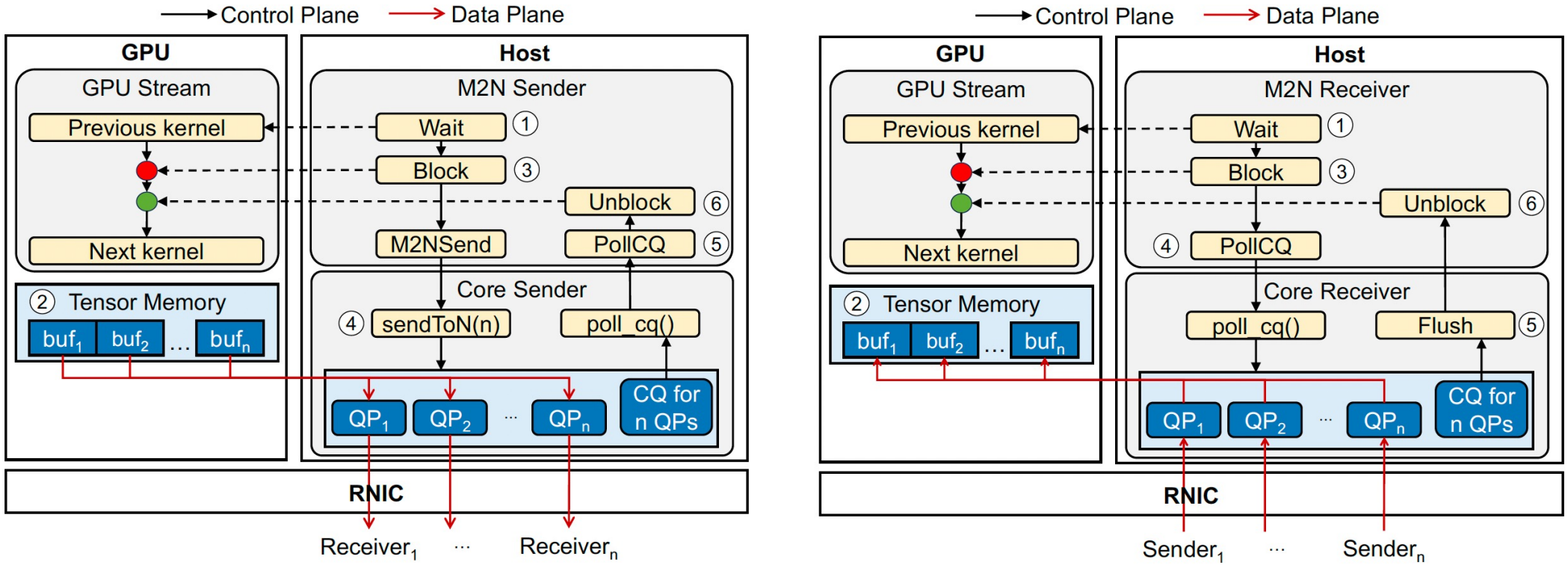
# Heterogeneous Deployment

- Higher bandwidth per unit cost (H20) for attention
- Higher computational capability per unit cost (L40S) for FFN

| Accelerator | Price | Cap. (GB) | Bw. (GB/s) | Comp. (TFLOPS) | Performance per Cost | | |
|---|---|---|---|---|---|---|---|
| | | | | | GB | GB/s | TFLOPS |
| L20 | 1.00 | 48 | 864 | 119.5 | 48 | 864 | 119.5 |
| H800 | 5.28 | 80 | 3430.4 | 989 | 15.2 | 649.7 | 187.3 |
| A800 | 2.26 | 80 | 2039 | 312 | 35.4 | 902.2 | 138.1 |
| H20 | 1.85 | 96 | 4096 | 148 | **51.9** | **2214.1** | 80.0 |
| L40S | 1.08 | 48 | 864 | 362 | 44.4 | 800.0 | **335.2** |

# High-Performance M2N Communication Library

- Minimize overhead such as data copies and group operations
- CPU manages send/receive operations on the control plane
- Data is directly transferred between the GPU and NIC



# Evaluation

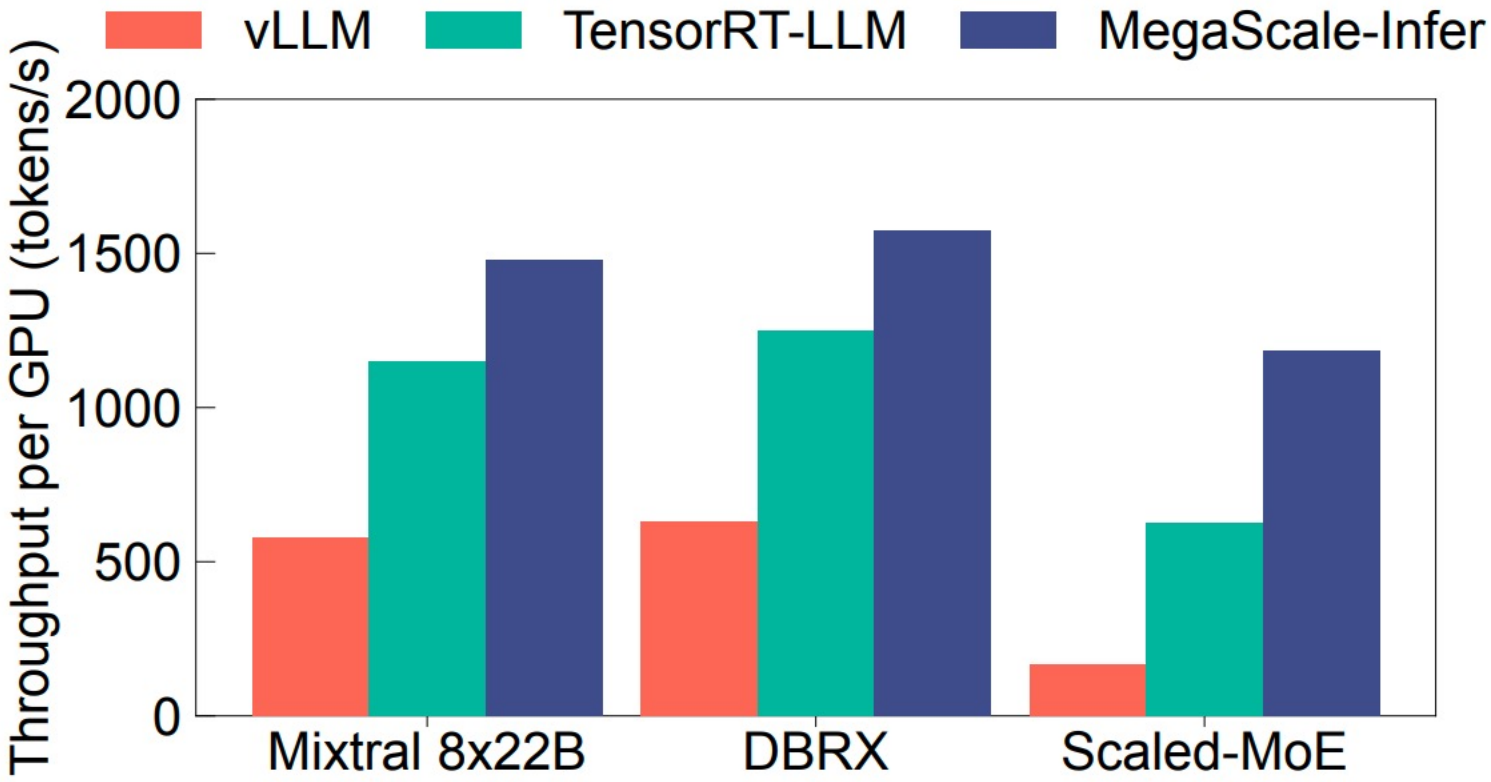
- Testbed
  - Cluster 1: NVIDIA 80GB Ampere GPUs
  - Cluster 2: NVIDIA H20 and L40S GPUs
- Models

| Model | #Layers | Hidden Size | #Experts | top-$k$ | Intermediate Size |
|---|---|---|---|---|---|
| Mixtral-8×22B | 56 | 6144 | 8 | 2 | 16384 |
| DBRX | 40 | 6144 | 16 | 4 | 10752 |
| Scaled-MoE | 48 | 8192 | 32 | 4 | 8192 |

- Baseline
  - vLLM
  - TensorRT-LLM

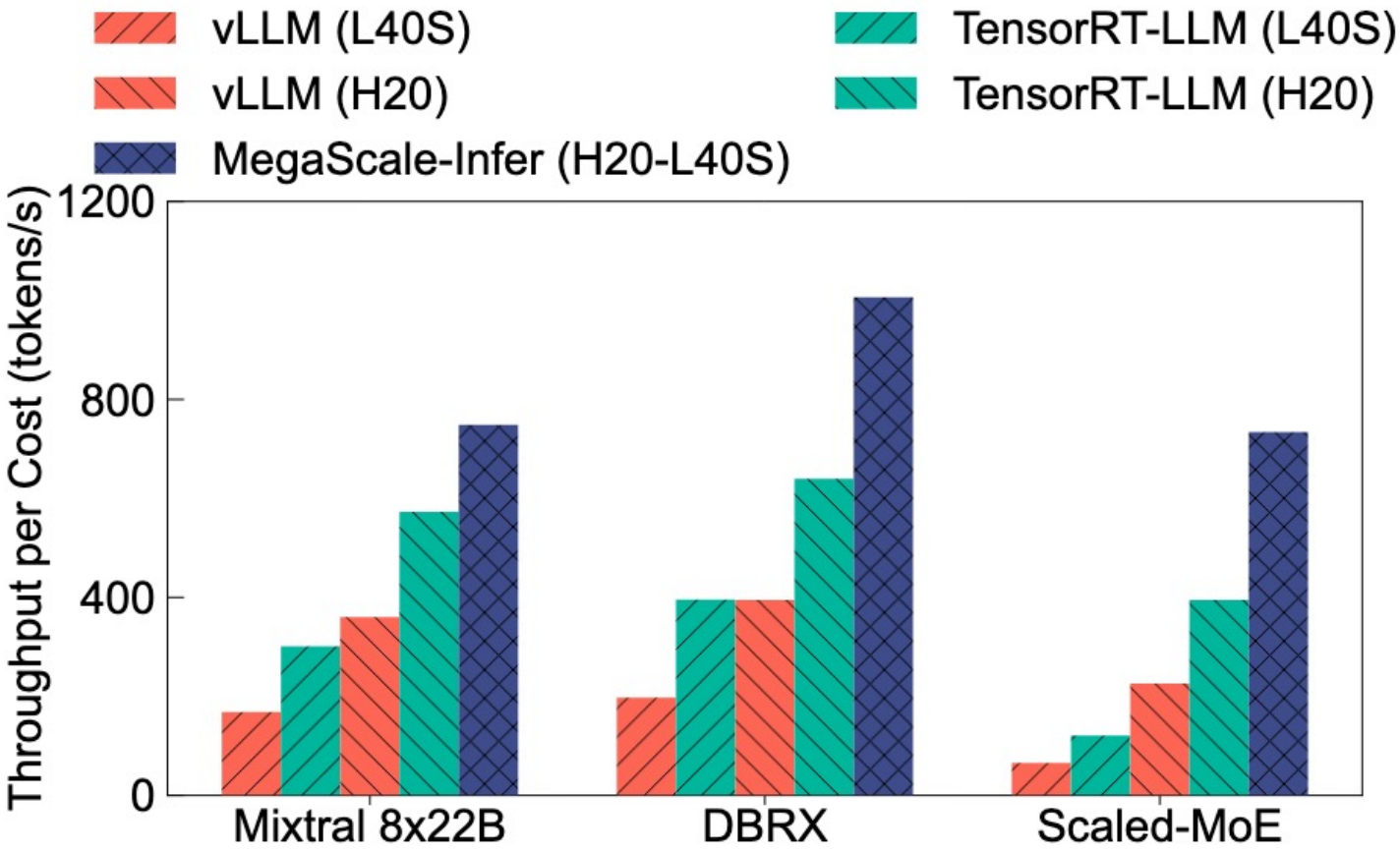
# Homogeneous Cluster

- Improves decoding throughput per GPU by up to **7.11x** and **1.90x** compared to vLLM and TensorRT-LLM



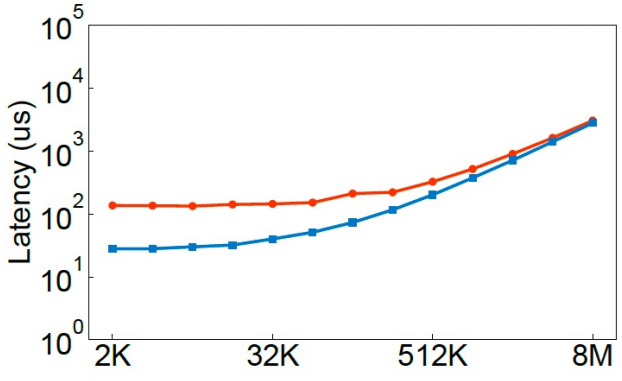
# Heterogeneous Cluster

- Improves decoding throughput per GPU by up to **3.24x** and **1.86x** compared to vLLM and TensorRT-LLM

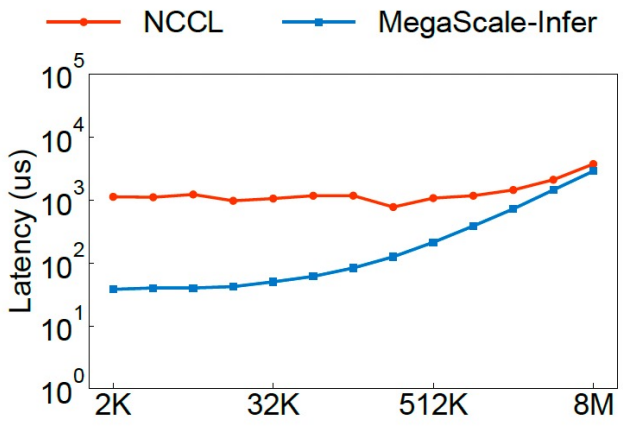


# Performance of M2N Communication

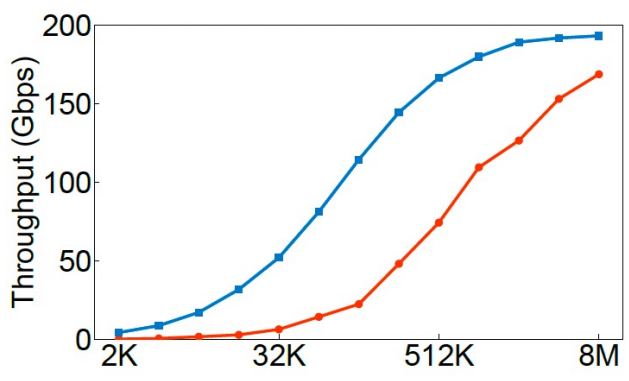
- Both median latency and tail latency are significantly better than NCCL, fully utilizing the network bandwidth



(a) M2N Median Latency



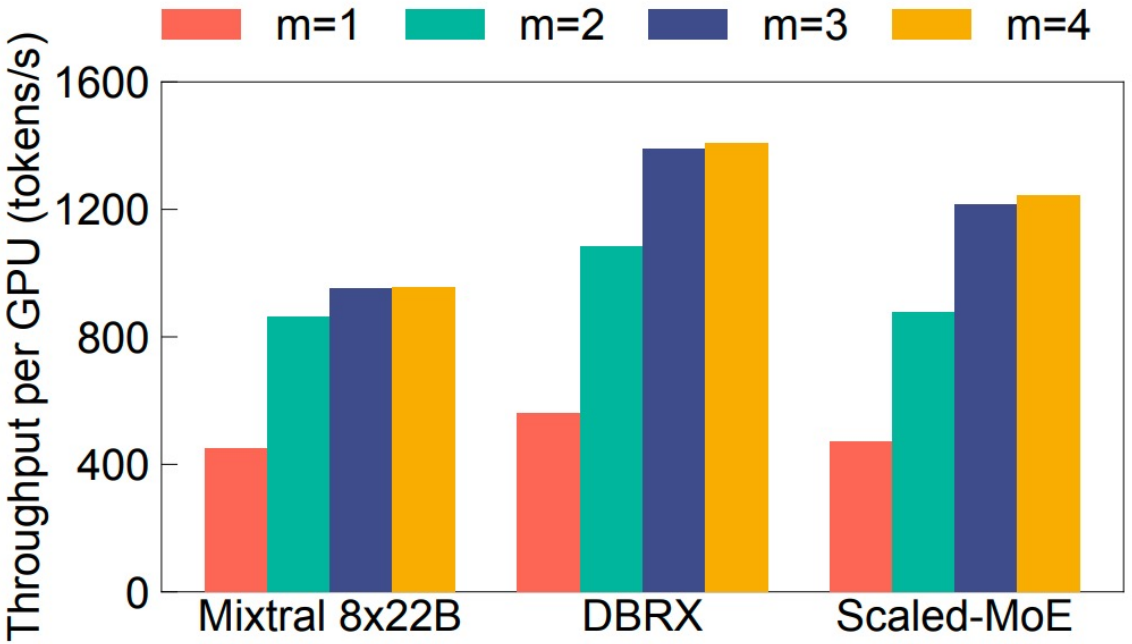
(b) M2N P99 Latency



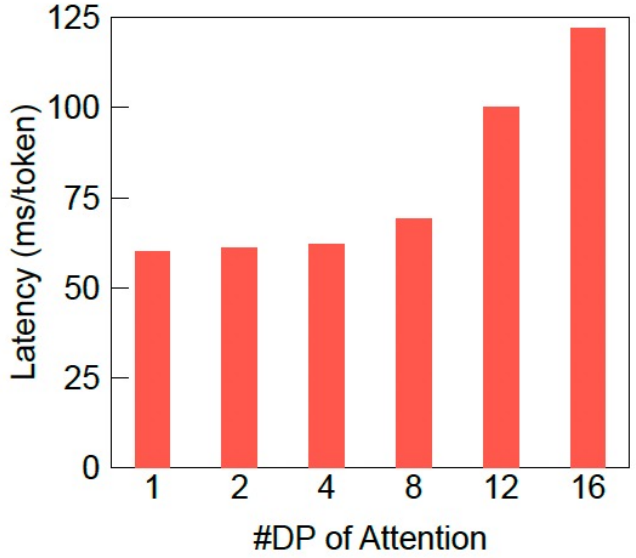
(c) M2N Throughput

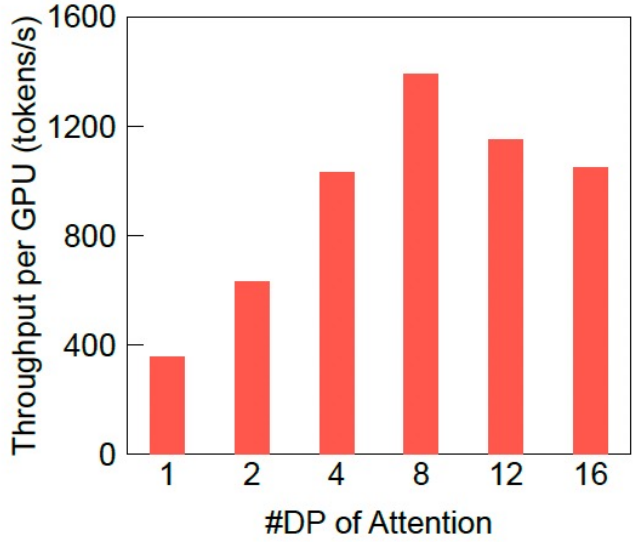
# Influence of Deployment Plan

- Number of micro-batch



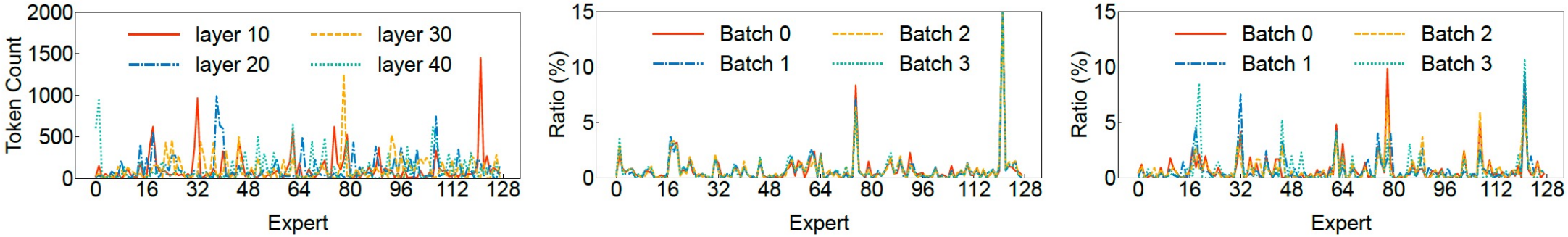

- Number of attention node



# Deployment Experience

- Expert balance
  - Severe load imbalance among experts
  - During the decoding phase, the load on each expert is relatively stable
  - During the prefill phase, the load on each expert varies significantly




(a) Received token count of each expert in a batch. (b) Received token ratio of each expert during decoding. (c) Received token ratio of each expert during prefill.

- Attention balance
  - Differences in sequence length can cause variations in the computation time of different attention nodes
  - Batch composition strategy based on estimated execution time

# Conclusion

- MegaScale-Infer: **Disaggregating attention and FFN** for efficient MoE inference at scale
  - Optimal deployment plan with ping-pong pipeline parallelism
  - High-performance M2N communication

Thanks!


zhurd@pku.edu.cn