Occult: Optimizing Collaborative Communication Across Experts for Accelerated Parallel MoE Training and Inference

Shuqing Luo, Pingzhi Li, Jie Peng, Katie Zhao, Kevin Cao, Yu Cheng, and Tianlong Chen

UNC, UMN, and CUHK

Background: Tendency of Modern MoE-based LLMs

Coarse-Grained MoE
(Small amount of global & activated experts)

Fine-Grained MoE
(Large amount of global & activated experts)

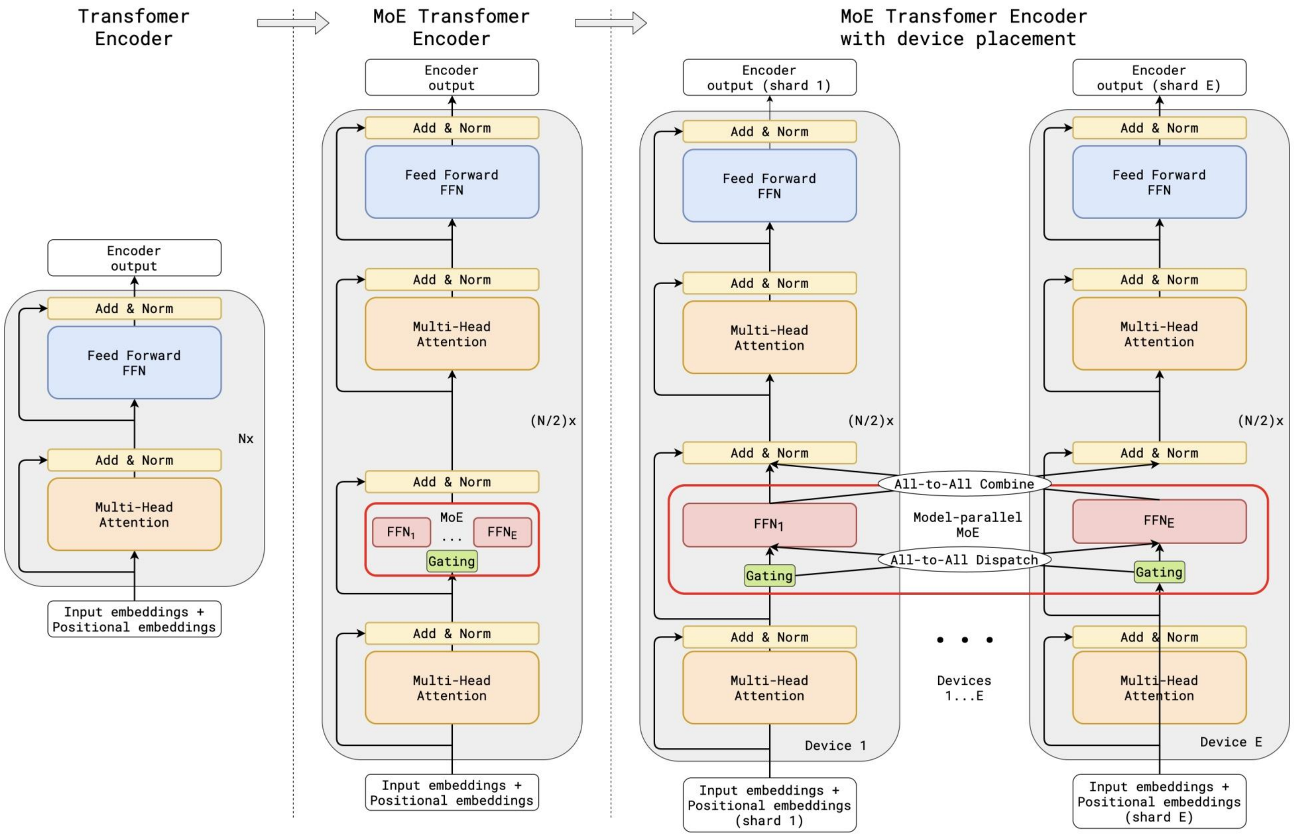
MistralAI

MiniMax-01

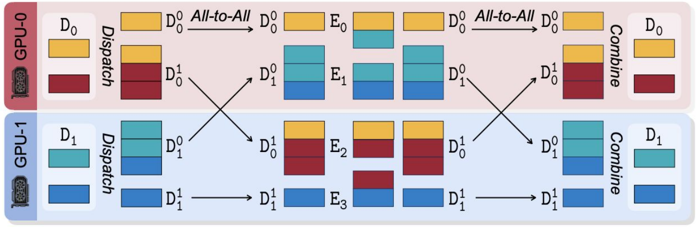
deepseek v3

Qwen 3

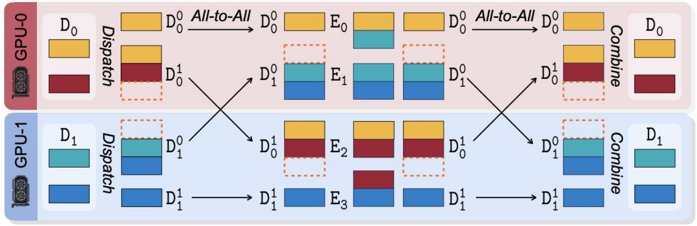
Background: Expert Parallelism for Mixture-of-Experts



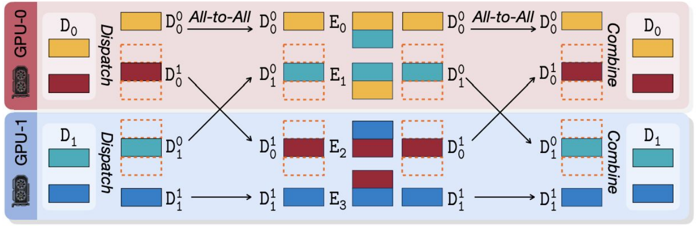
Overview: Optimizing All-to-All Communication Volume



(a) Classical MoE workflow ($C_{\mathcal{T}} = 2$).



(b) `Occult` workflow w/o collaborative pruning ($C_{\mathcal{T}} = 1.5$).



(c) `Occult` workflow w. collaborative pruning ($C_{\mathcal{T}} = 1$).

Core Insights:

- Only send one replica for a token when more than one of its activated experts are kept on a device.
- Optimize all-to-all communication volume with algorithm-system co-design

Methodology: Expert Collaboration for Specialized Layout

Formulate the all-to-all communication as collaborative communication.
For 2 experts co-activated by a token:

- Inter-Collaboration: 2 experts are kept on different devices.
- Intra-Collaboration: 2 experts are kept on the same device.

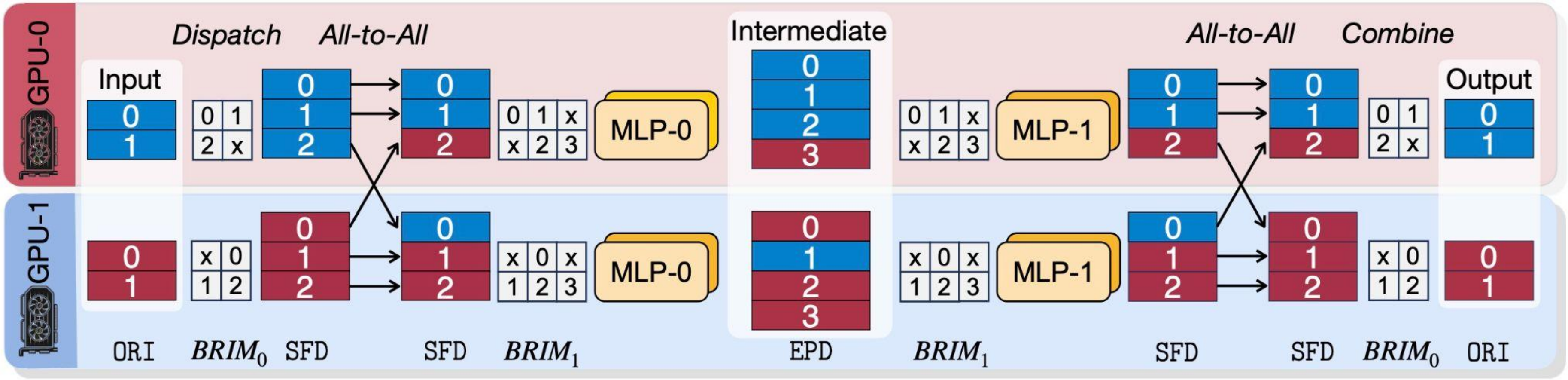
Maximizing intra-collaboration & minimizing inter-collaboration:

- Fully utilize each token replica
- Reduce all-to-all communication volume

Profiling on wikitext to determine the specialized expert layout

- Run the prefilling stage to obtain the routing information
- Construct a collaboration graph for each MoE layer
- Build expert layout through graph partition

Methodology: Sparse MatMul & 2-Stage Top-k Reducing



Methodology: Routing with Collaboration Pruning

Standard routing algorithm cannot achieve ultimate communication efficiency.
Modify the routing choice of each token,
Making it fall into a limited number of devices:

- Keeping the scores of the top-k experts
- Replacing the selected experts with low scores
    - Scheme-1: Replace them using candidates with higher routing score
    - Scheme-2: Replace them using candidates with higher expert similarity

Experiments Setup

Model	Total Params	Activated Params	Top-k	# Routed Experts	# Layers
OLMoE-1B-7B	7B	1B	8	64	16
Qwen1.5-MoE-A2.7B	14B	2.7B	4	60	27
DeepSeek-MoE	16B	2.8B	6	64	24

Datasets:

- Using Alpaca for collaboration pruning

Hardware:

- PCIe-connected NVIDIA A6000 (48 GB) GPUs

Results: Reducing Wall-Clock Latency for Training

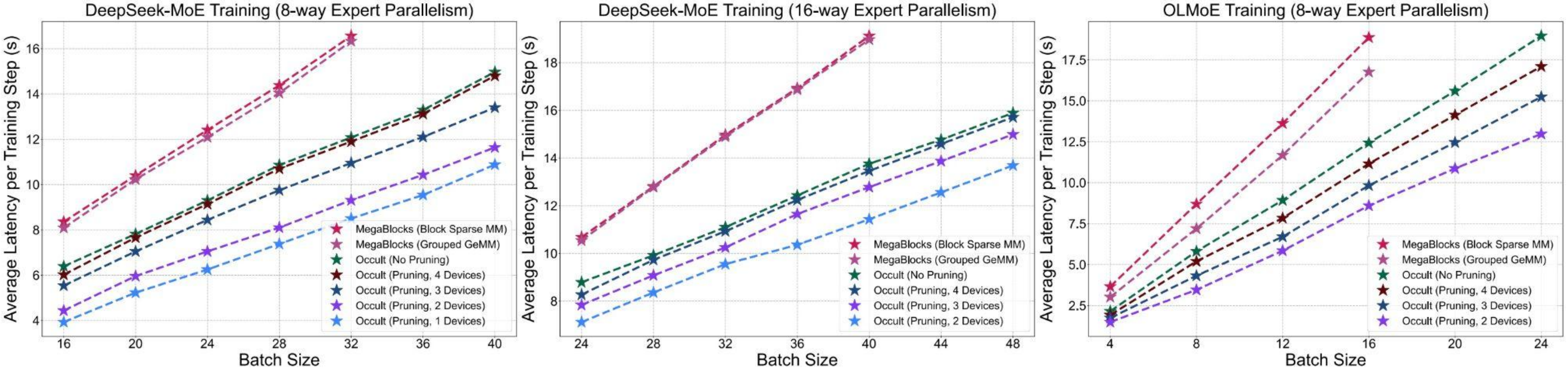


Figure 12. **More training latency comparison for expert parallelism frameworks.** Owning to the communication- and memory-efficient design, `Occult` achieves superior training efficiency under both 8- and 16-way expert parallelism configurations.

Results: Reducing Wall-Clock Latency for Inference

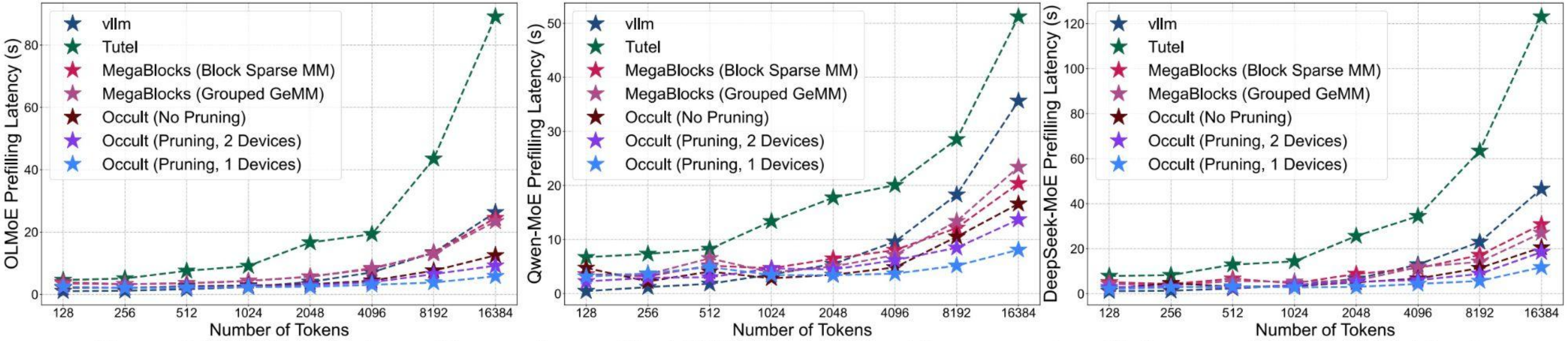


Figure 9. **Prefilling Latency Comparison with** 4 **GPUs.** `Occult` achieves superior efficiency under scaled workload.

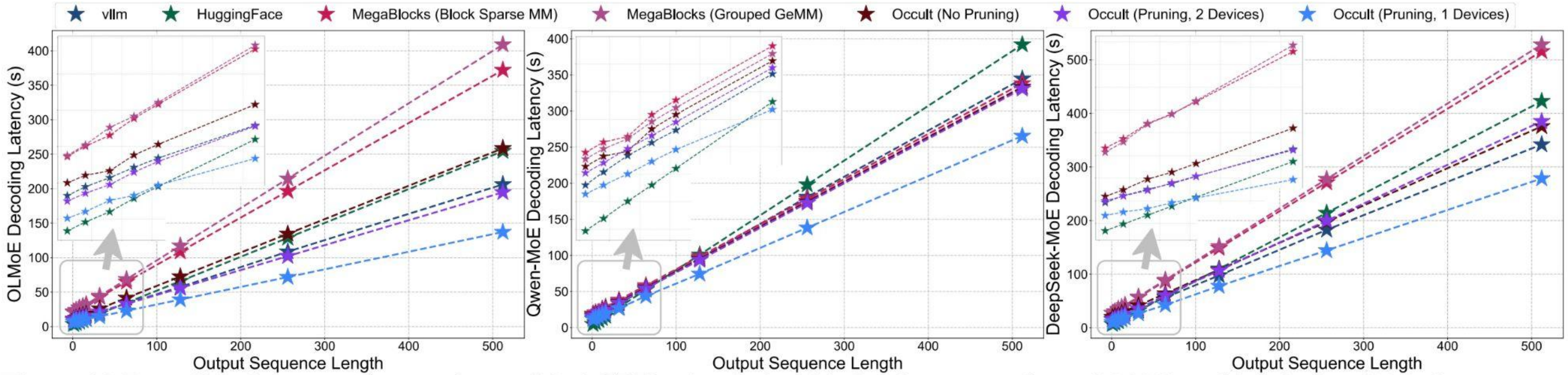


Figure 10. **Decoding Latency Comparison with** 4 **GPUs.** Analysis with fixed prompt tokens (12800) and batch size (512) demonstrates `Occult`'s consistent latency advantages on communication-intensive decoding tasks.

Results: Comparable Performance with Standard Tuning

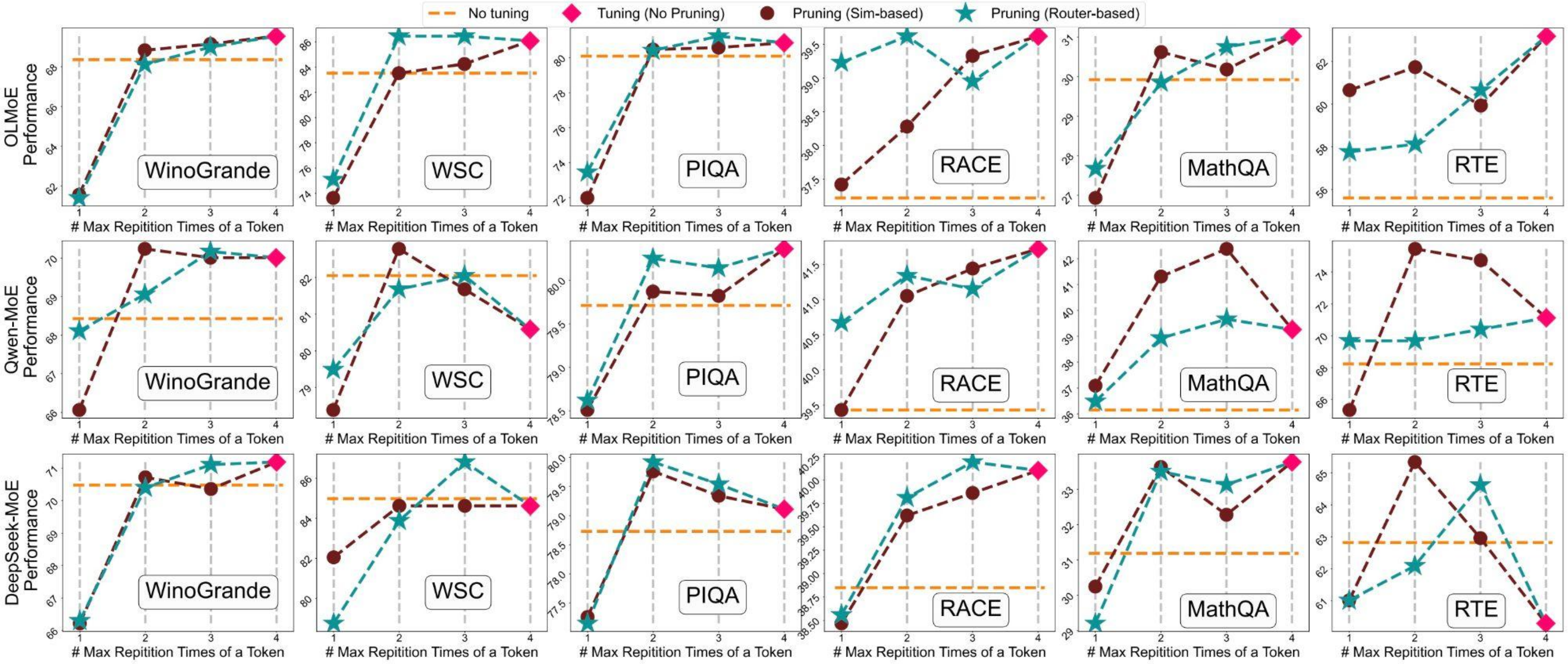


Figure 7. **Performance Comparison for Collaboration Pruning.** Comprehensive evaluation across three MoE architectures shows performance trends under different pruning strategies. Note that 4-device collaboration pruning is equivalent to standard training with original top-k routing.

Thank you!