Accelerating Distributed MoE Training and Inference with Lina

Jiamin Li*, Yimin Jiang‡, Yibo Zhu, Cong Wang*, Hong Xu†

*City University of Hong Kong, ‡ByteDance Inc., †The Chinese University of Hong Kong

香港城市大學 City University of Hong Kong

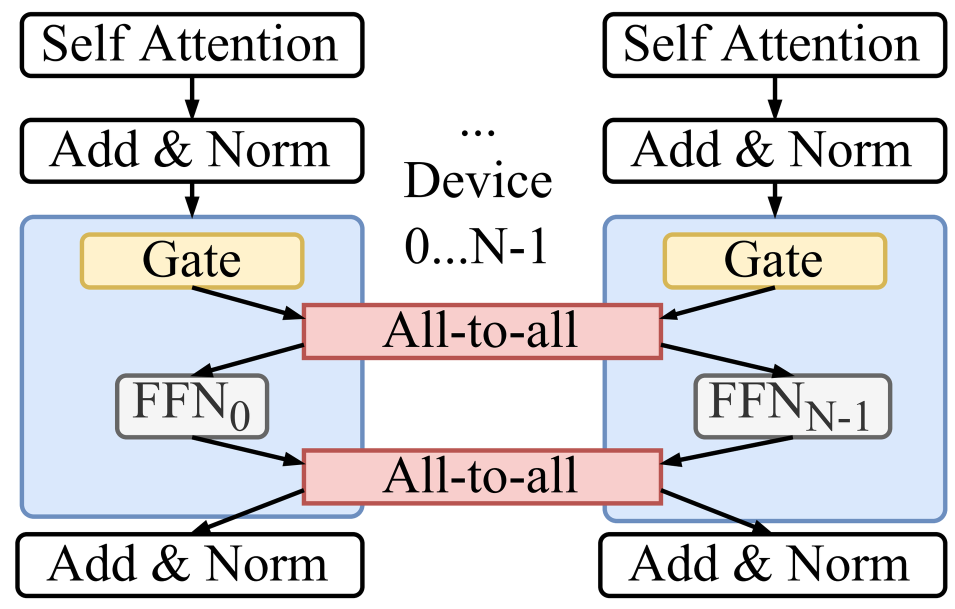
香港中文大學 The Chinese University of Hong Kong

Introduction

Mixture-of-Experts (MoE): a popular way to curb the computation cost of deep learning models.

[MoE in language models] MoE layer replaces the FFN layer in Transformer. It consists of multiple FFNs as experts, and a gating network. The gating network dispatches the token to a small number of experts (top-1, top-2).

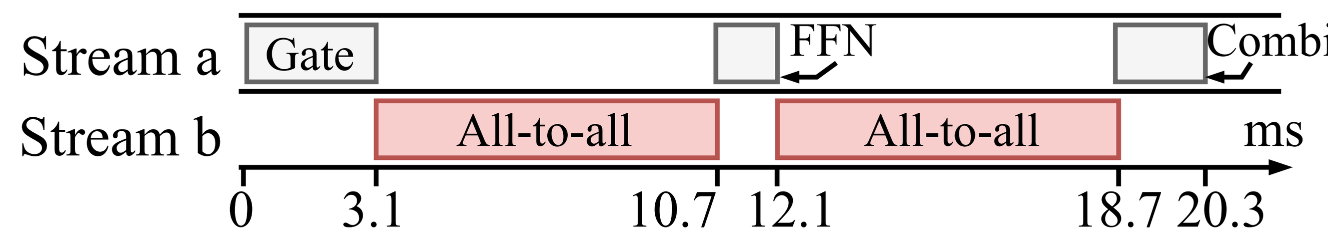
[Distributed MoE] Data parallelism and expert parallelism are applied. It allocates one unique GPU for each expert and use <u>all-to-all</u> to exchange tokens.



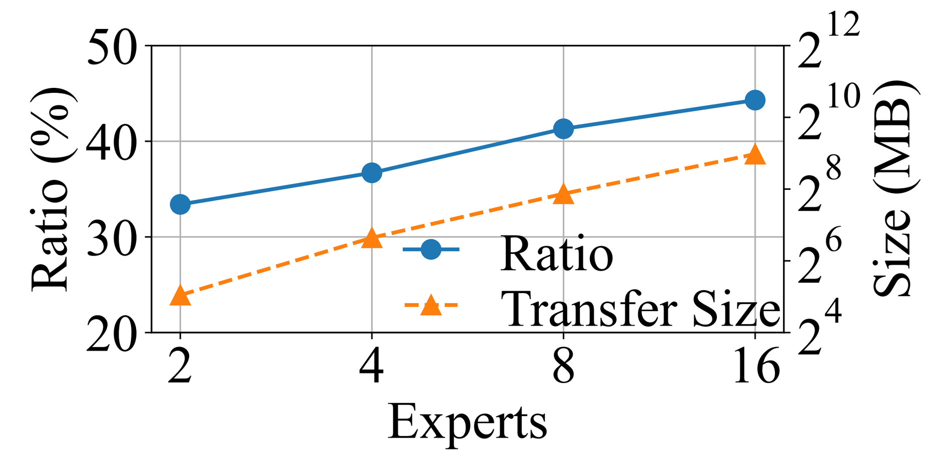
Motivation

Why is all-to-all the bottleneck in distributed MoE?

[Synchronous all-to-all with large data transfer]

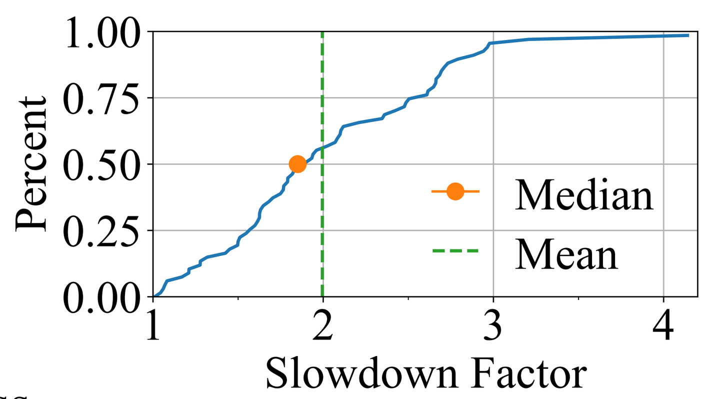


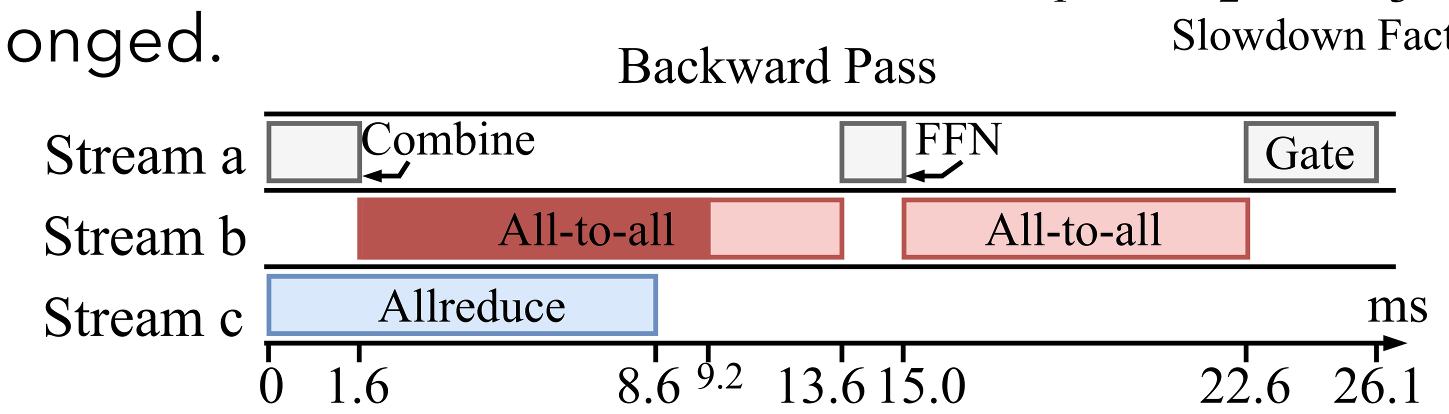
74.9% of the running time of one MoE layer.



Training

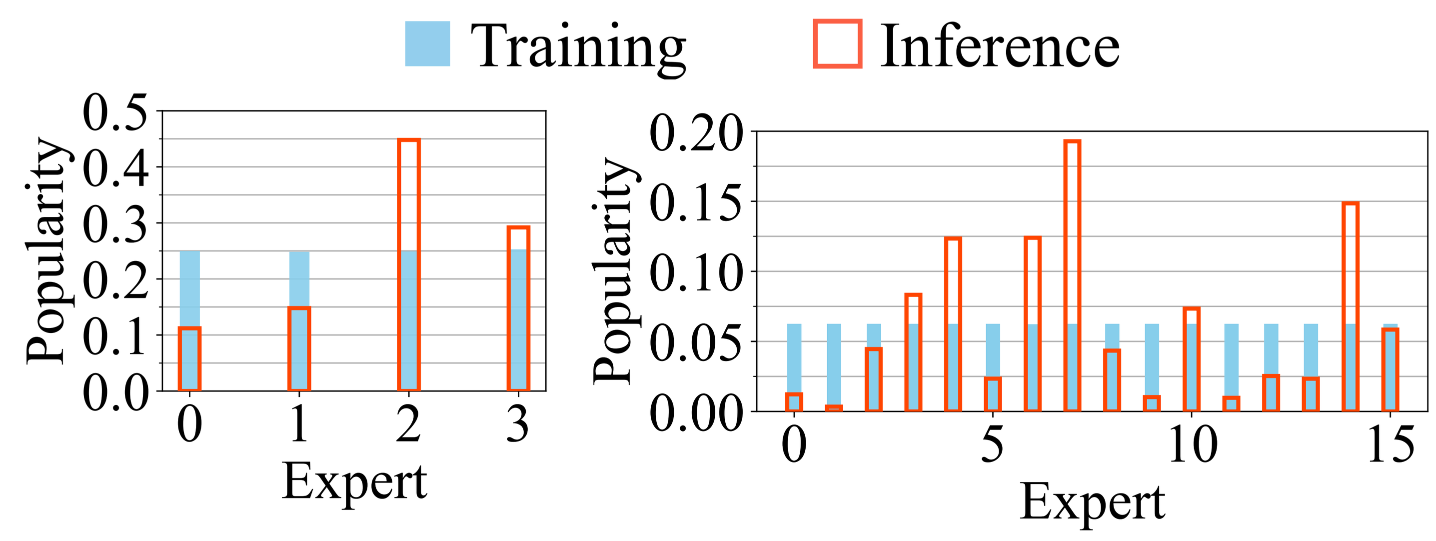
[Prolonged all-to-all with allreduce] In the backward pass, all-to-all and allreduce control their own process group and overlap, they contend for the network bandwidth and their completion times are severely prolonged.



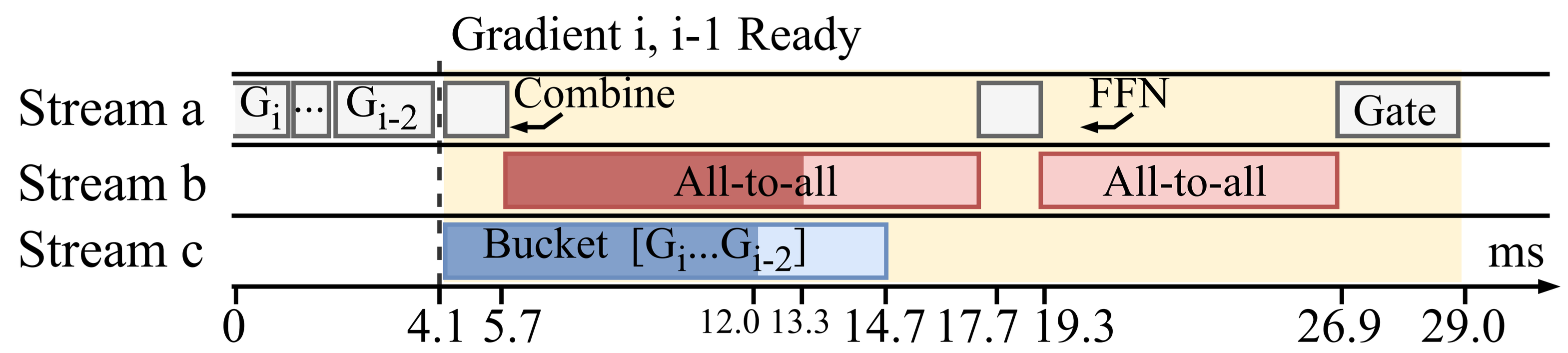


Inference

[Skewed expert popularity] The token-to-expert distribution in inference is purely workload-driven. Expert popularity is <u>highly skewed</u> in sharp contrast to training.

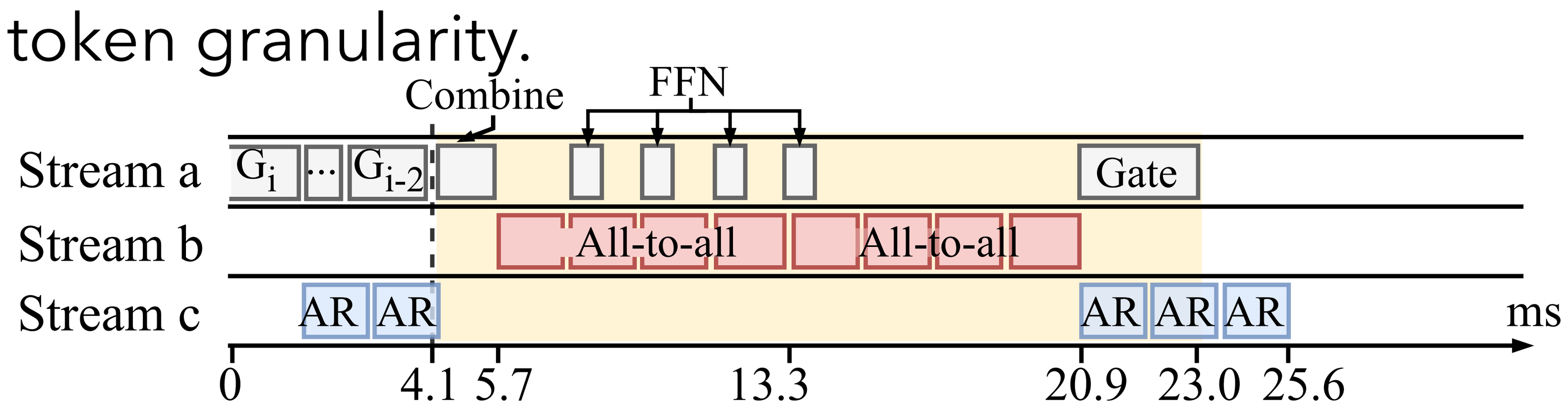


Lina <u>prioritizes all-to-all</u> and avoids concurrent execution with allreduce with priority scheduling.



[Tensor partitioning] Partition each gradient tensor into equal-sized small chunks.

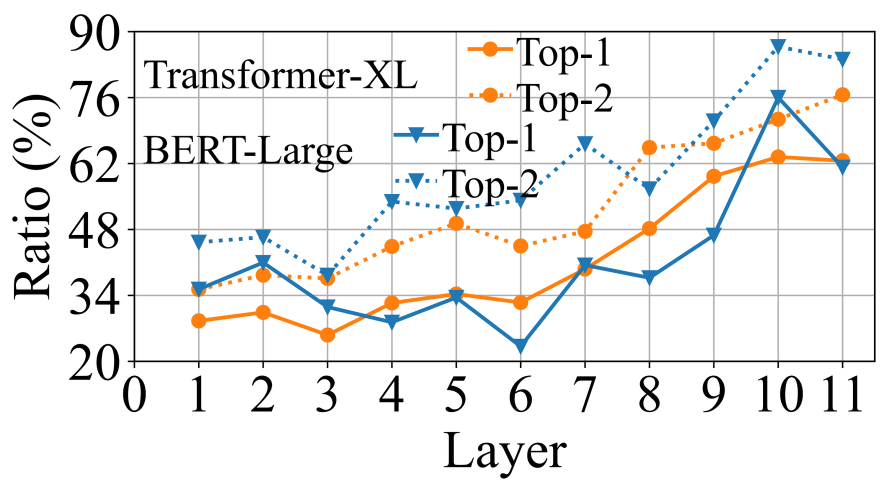
[Pipelining micro-ops] Pipeline the expert computation and all-to-all micro-ops, because the FFN computation is in token granularity.



Design

Lina replicates popular experts on proportionally more devices to <u>balance the workload</u>.

How to know the expert popularity a prior?



[Pattern in expert selection] Tokens that have selected the same expert in layer i tend to select the <u>same</u> expert again in layer i + 1.

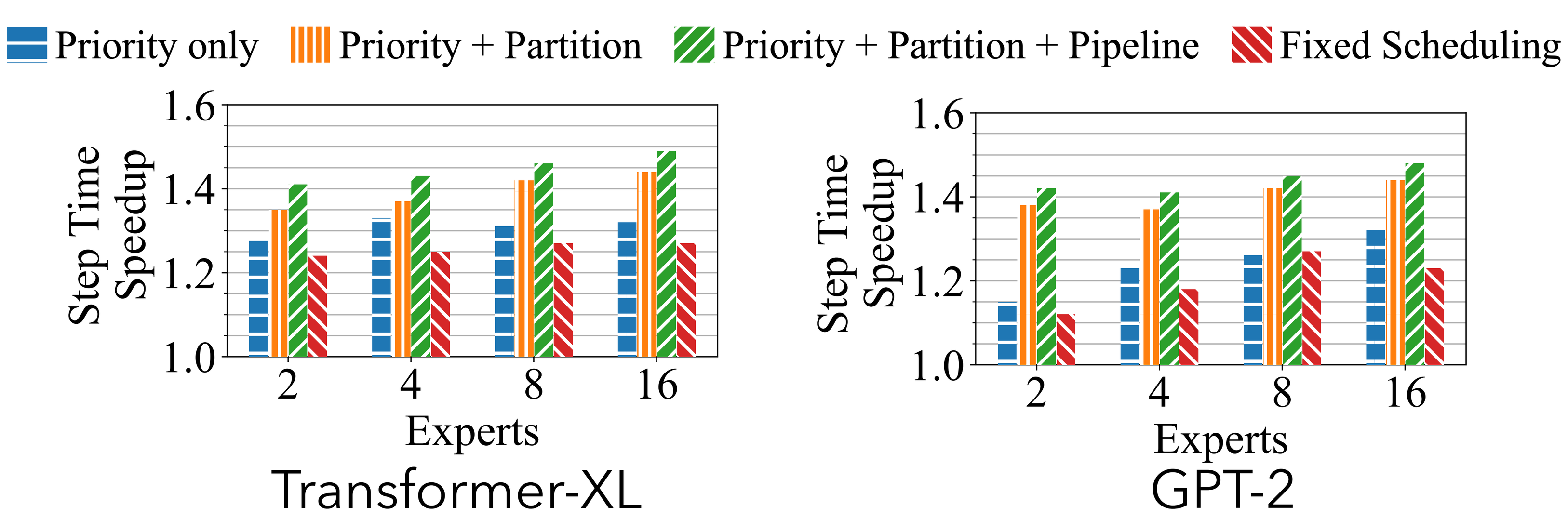
[Two-phase scheduling]

a. Resource scheduling based on estimated popularity
   - Estimate with patterns profiled during training

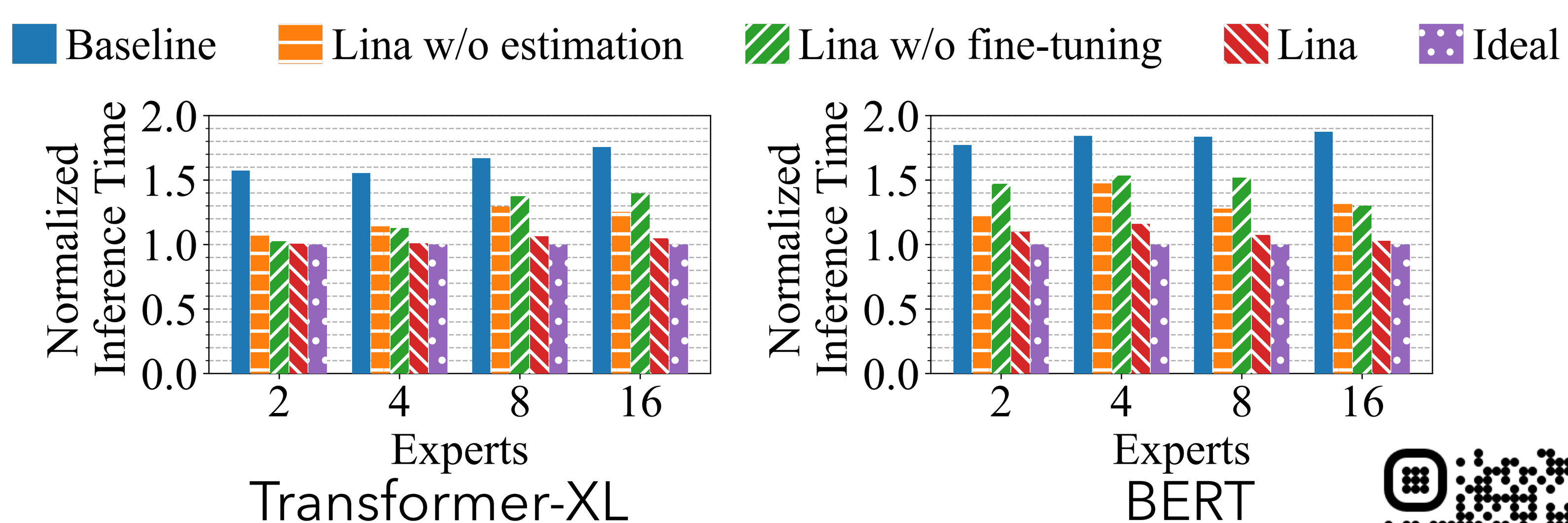
b. Low-overhead fine-tuning on actual routing decision

Evaluation

<u>**Lina reduces the training step time by up to 1.73x.**</u>



<u>**Lina reduces the 95%ile inference time by an average of 1.63x**</u>



[1] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, De- hao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling gi- ant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668, 2020.

[2] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint arXiv:2101.03961, 2021.

[3] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Min- jia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale. arXiv preprint arXiv:2201.05596, 2022.

jiaminli.icy@gmail.com