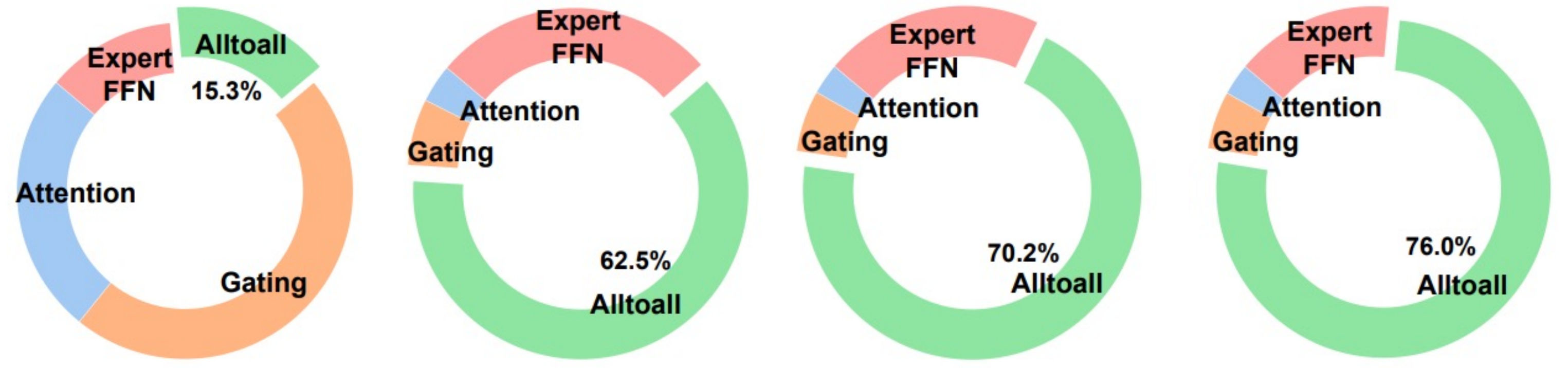
Exploiting Inter-Layer Expert Affinity for Mixture-of-Experts Model Inference

Jinghan Yao, Quentin G. Anthony, Aamir Shafi, Hari Subramoni and Dhabaleswar K. Panda
{yao.877, anthony.301, shafi.16, subramoni.1, panda.2}@osu.edu

MOTIVATION

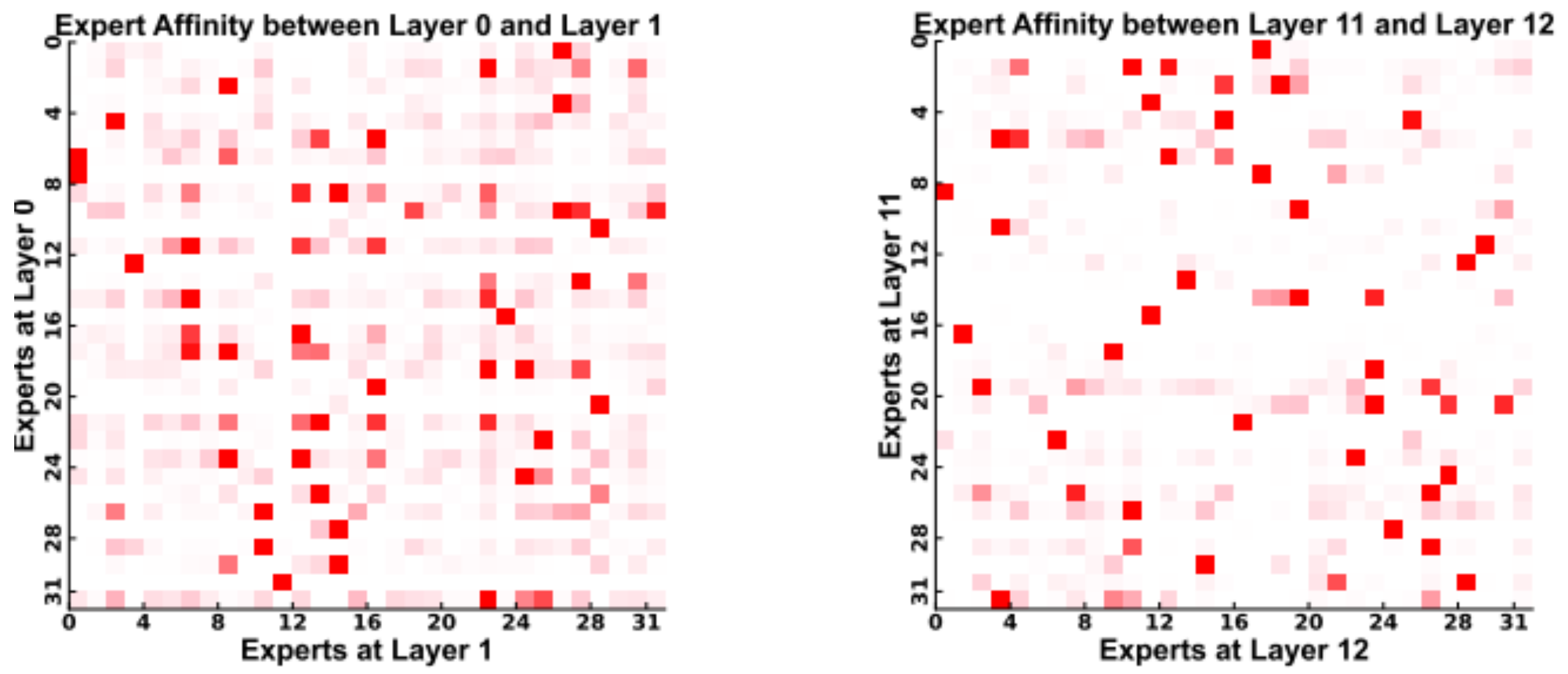
- Mixture-of-Experts (MoE) models are becoming popular in LLM/VLM, expert parallelism is widely used in both training and inference.
- Expert parallelism necessities Alltoall communication to route tokens, which can introduce non-trivial overhead.



(a). 1 node (b). 2 node (c). 4 node (d). 8 node

Proportion of Alltoall overhead to the time spent on computations. We only measure the most significant four operations in the MoE model.

- We found that in pretrained MoE models, experts among different layers exhibit a strong correlation. We name it as Expert Affinity.



Heatmaps illustrating the distribution of the conditional probability of expert routing in different layers of a pretrained GPT model. Color intensity represents the magnitude of the likelihood.

- ***Can we put experts with high affinity together to reduce the inter-layer Alltoall overhead?***

CHALLENGES

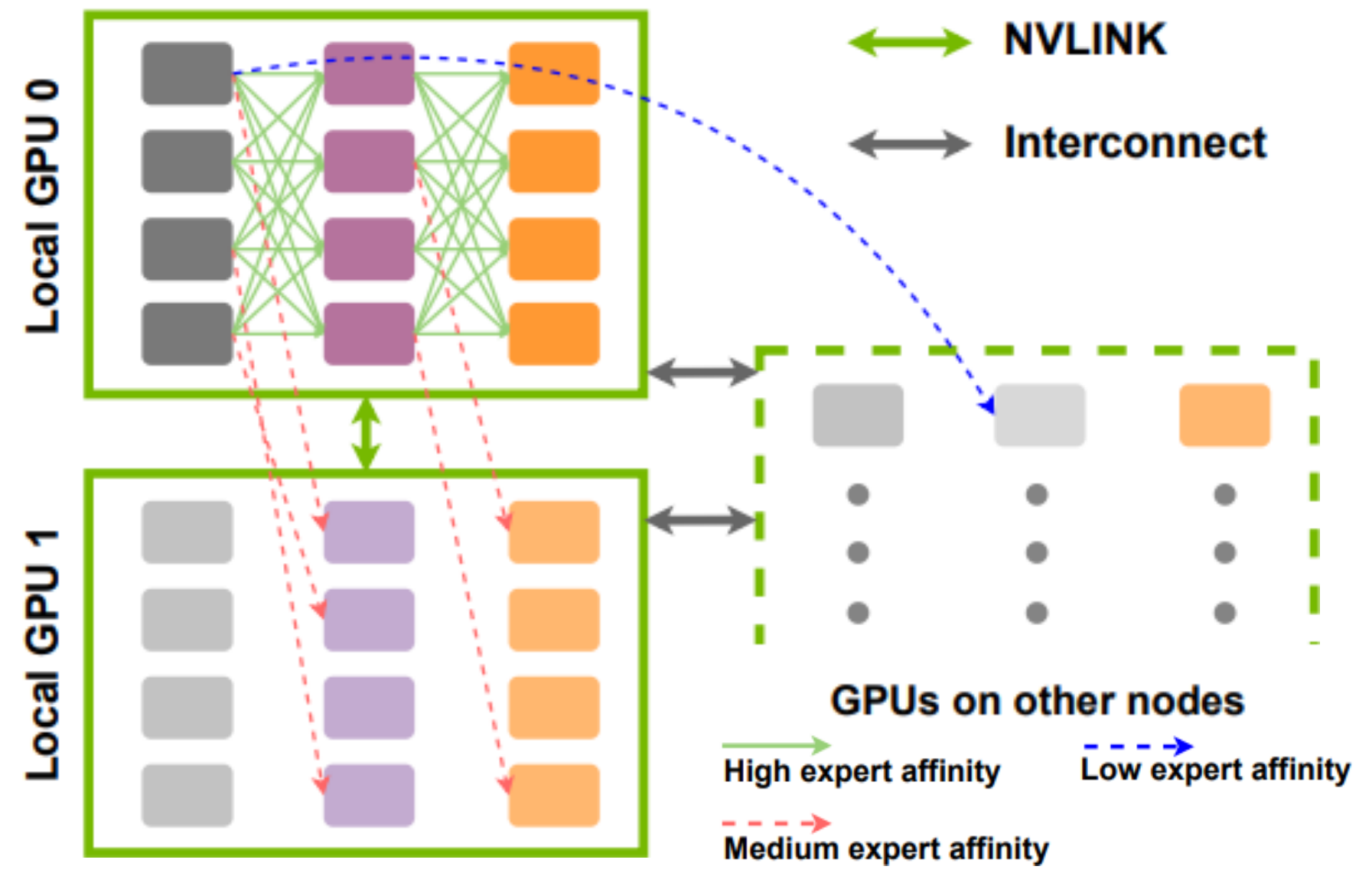
How to leverage expert affinity to accelerate MoE model inference?

- How do we model expert affinity in a deterministic manner?
- How to map expert affinity to a given hardware topology?
- How to eliminate data locality in expert + data parallel schemes?

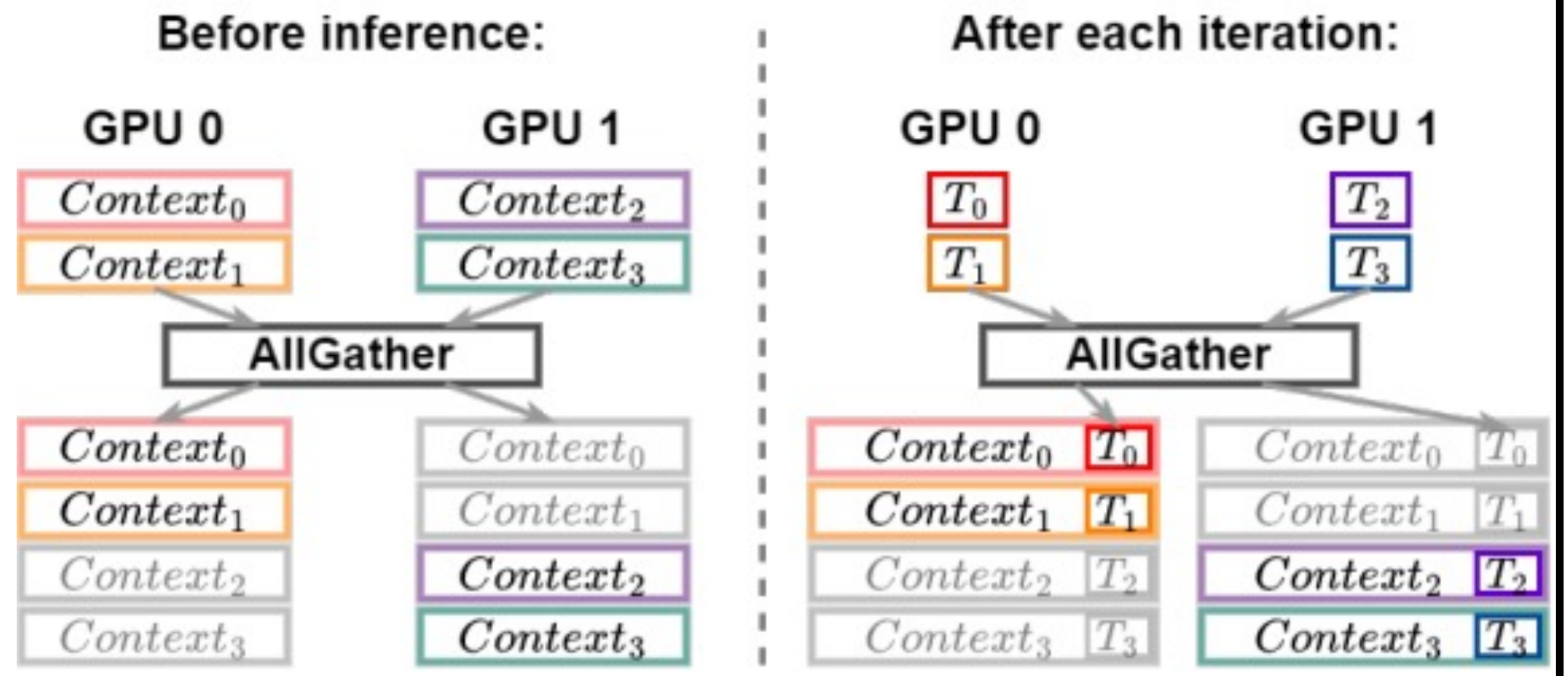
DESIGN – ExFlow

We propose a novel expert parallelism optimization solution based on context coherence and expert affinity, named ExFlow.

- ExFlow defines the expert affinity property that exists implicitly in pretrained MoE models by capturing the combined conditional probability of the expert routing decisions among layers.
- Together with expert affinity, ExFlow introduces a context-coherent data parallelism to largely reduce the Alltoall overhead.



Affinity-aware Expert Parallelism



Context-coherent data parallelism

PERFORMANCE – End-to-end Inference

We test ExFlow based on DeepSpeed MoE implementation. We use Cambridge Wilkes-3 (A100 GPUs) for up to 64 GPUs with IB HDR 200.

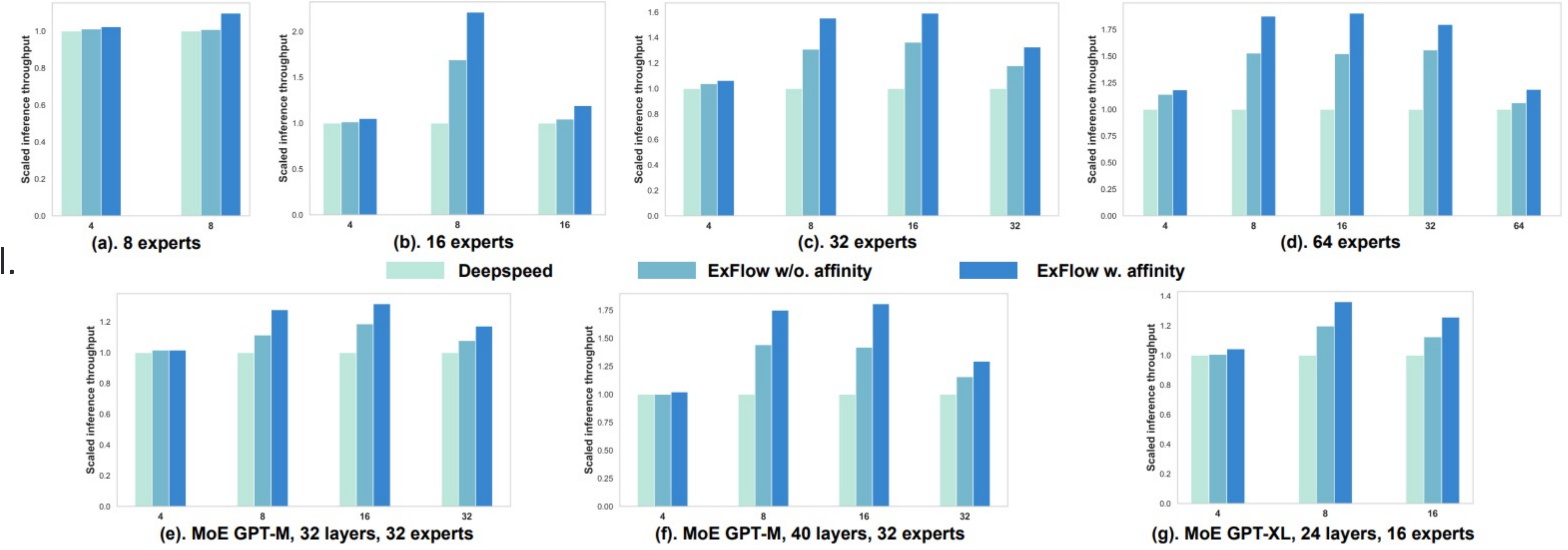
- **Baseline:** DeepSpeed MoE with Tutel optimization.
- **ExFlow w/o. affinity:** Only use context-coherent data parallel.
- **ExFlow w. affinity:** With context-coherent data parallel & expert affinity-aware parallel.

We choose 6 variants of MoE models, as listed:

- **Training:** We use DeepSpeed-Megatron for pretraining with 300B tokens.
- **Inference:** All models are with Top-1 gating and lossless routing.
- **We only use 5000 tokens to profile expert affinity.**

Model	Base	Experts	Layers	D
MoE GPT-M	350M	8 16 32 64	24	1024
MoE GPT-M	470M 590M	32	32 40	1024
MoE GPT-XL	1.3B	16	24	2048

MoE model configurations

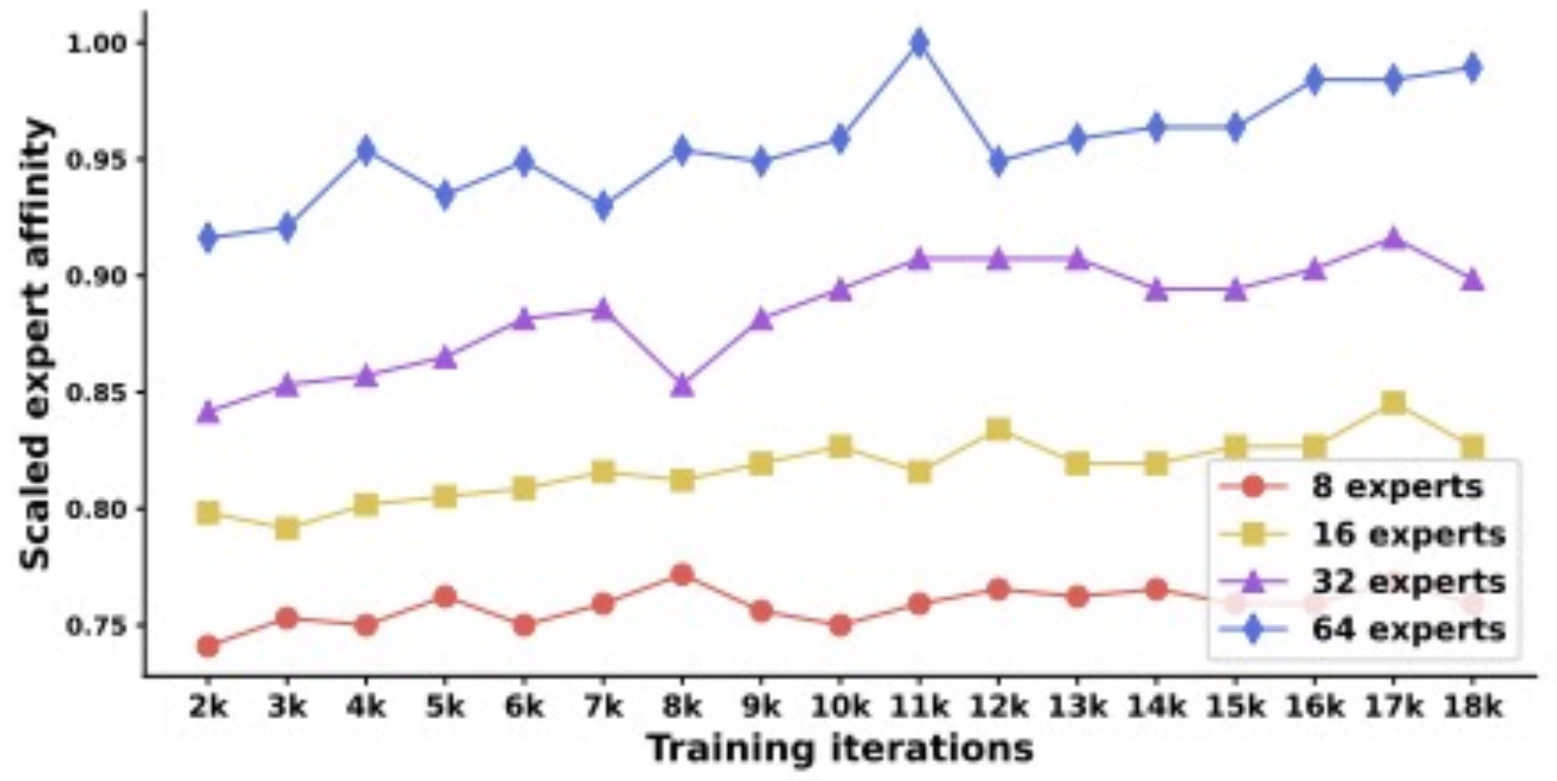


(a). 8 experts (b). 16 experts (c). 32 experts (d). 64 experts
(e). MoE GPT-M, 32 layers, 32 experts (f). MoE GPT-M, 40 layers, 32 experts (g). MoE GPT-XL, 24 layers, 16 experts

End-to-end GPT MoE model inference throughput. Results are normalized for better visualization.

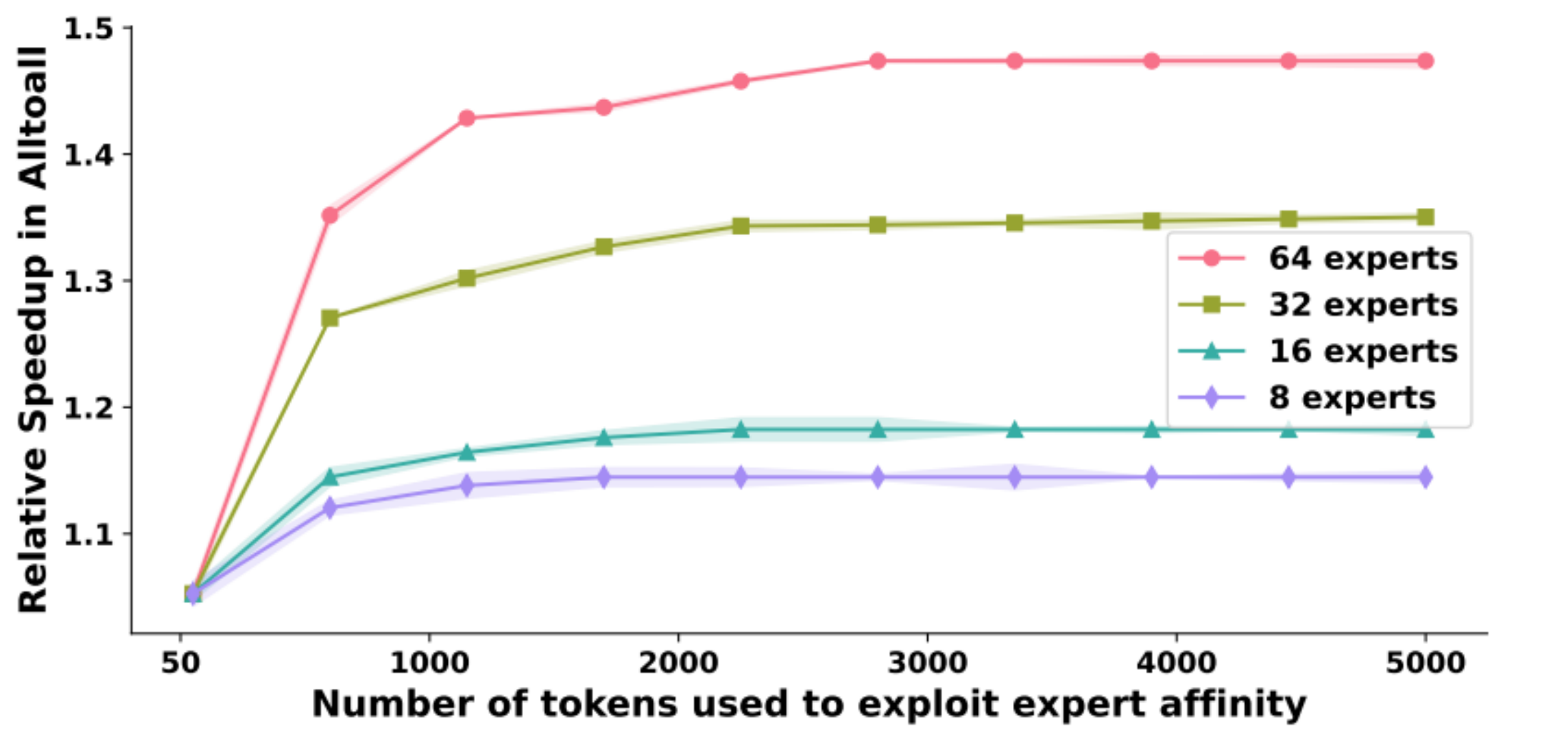
PERFORMANCE – Expert Affinity

- **Expert affinity gets more salient as the training goes on.**



Scaled expert affinity during training iteration 2000 to 18000. Oscillations are observed during the first 2000 iterations. (please refer to our paper for details)

- **We design an efficient model to solve expert affinity. By only examine 5000 tokens, we can get an accurate estimation and gain significant speedup in Alltoall.**



Number of randomly sampled tokens used to estimate the expert affinity and its relative speedup during inference.

CONTRIBUTIONS

- Comprehensive profiling and evaluation of expert affinity in current MoE models.
- Exploit the expert affinity in pretrained MoE models and largely reduce the inference latency in the context of expert parallelism.
- Validate the intrinsic property of expert affinity and its independence and robustness to inference-time data distribution.
- Evaluate proposed designs compared to existing expert parallel libraries using **DeepSpeed Mixture-of-Experts on up to 64 A100 GPUs.**