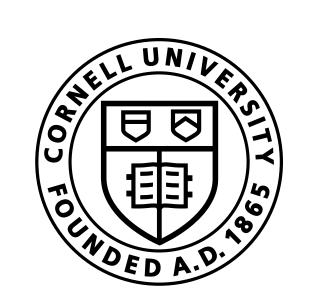
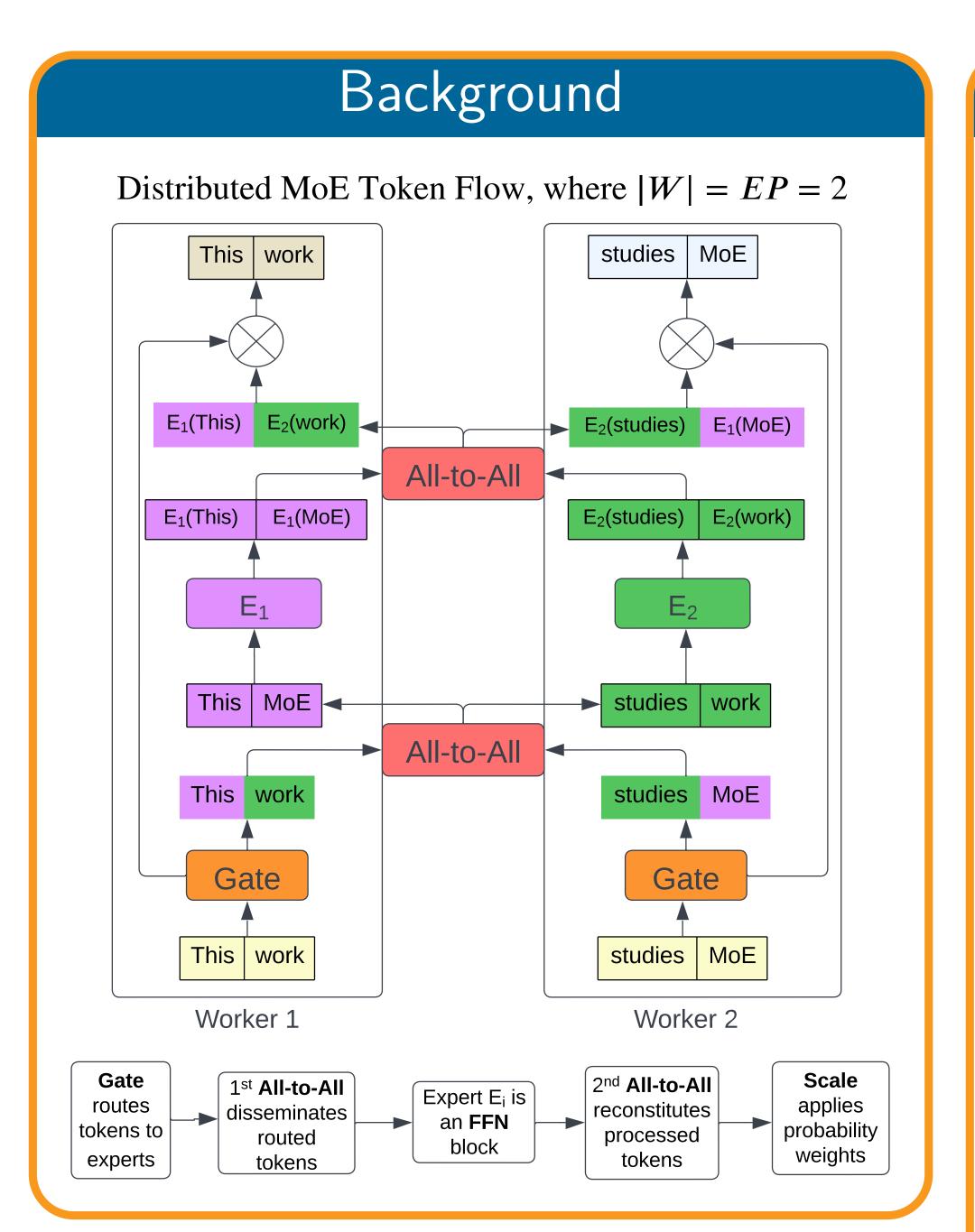
Aristos: Pipelining One-sided Communication in Distributed Mixture of Experts (MoE)



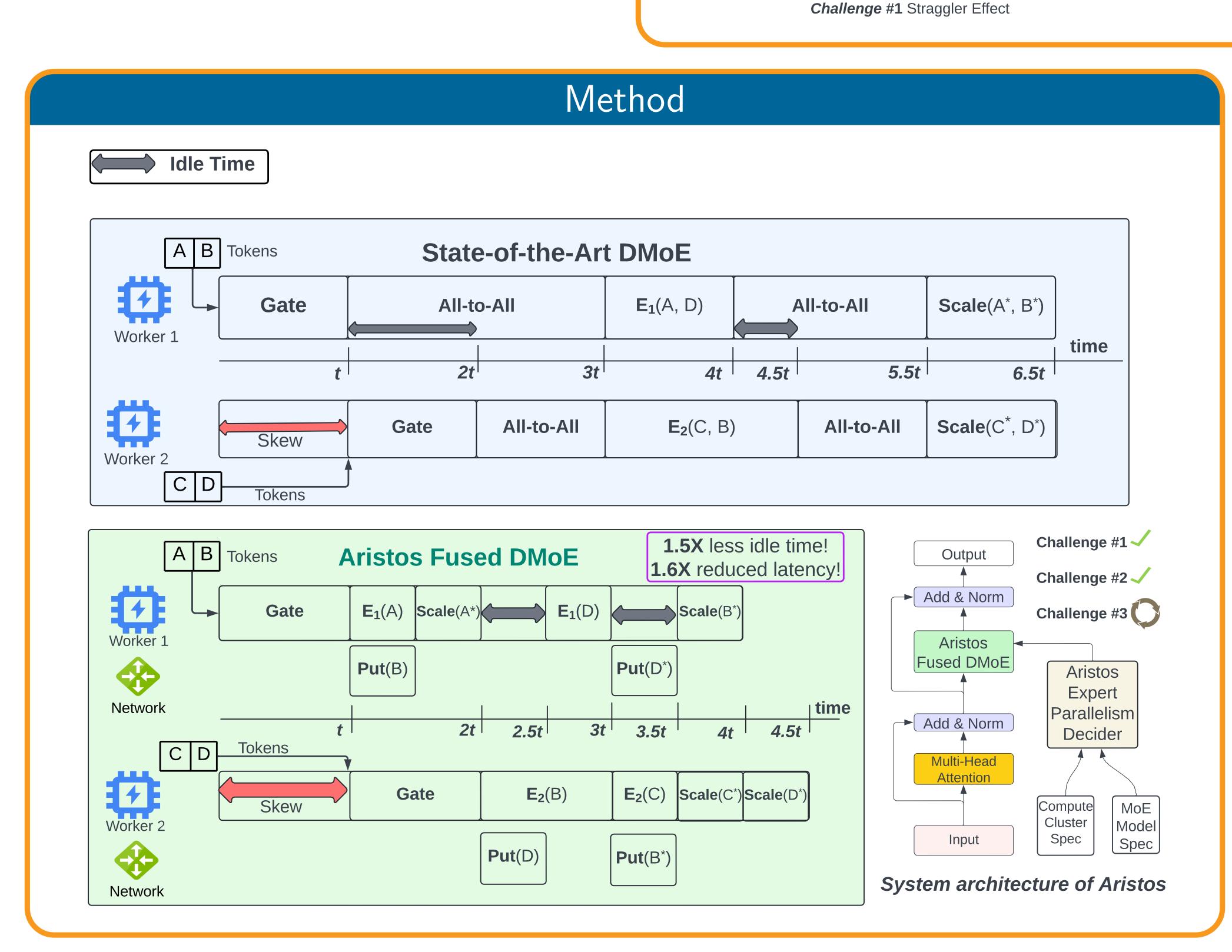
Cornell Bowers C·IS Computer Science

Osayamen Jonathan Aimuyo[†]

oja7@cornell.edu † Cornell Ann S. Bowers College of Computing and Information Science, Cornell University



Challenges The widely-adopted [2, 3, 4] MoE architecture, promising 5x faster training and 9x reduced inference costs [6], is currently plagued by three open challenges [5, 1] in the distributed setting. nx4 A100: n Perlmutter GPU nodes, each with 4 A100 80GB **1x8 V100:** Azure NDv2 VM **GPT-3 8x2.7B Forward Pass Kernel Summary: 2x4 A100** Challenge #2 GPT-3 MoE 32x350M All-to-All Straggler Effect: 8x4 A100 GPT-3 MoE 8x350M All-to-All Straggler Effect: 1x8 V100 FFN: 21.5% No overlap of communication Multi-head Attn: 35.8% and computation. max delay = 8.79 msavg actual = 0.98ms**NB:** All-to-All is the *second* mos Gate: 20.4% time-consumind kernel, only Scale: 10.1% behind the All-to-All: 11.9% Attn/FFN GEMM. Expert Parallel (EP) vs Iteration Latency (↓ is better) 750 1000 1250 1500 1750 800 1000 1200 All-to-All Steps All-to-All Steps GPT-3 350M MoE on 4x4 A100 GPUs GPT-3 MoE 32x350M All-to-All Straggler Effect: 8x4 A100 GPT-3 MoE 8x350M All-to-All Straggler Effect: 1x8 V100 4.00 +Challenge #3 EP = 4**ECDF** Deciding an optimal expert \rightarrow EP = 16 parallel strategy. ළ 3.40 - $\mu = 0.23$ ms $\mu = 1.17$ ms ;= 3.20 -**NB**: *EP* = 4 P95 = 0.95 msP95 = 0.95 msyields an OOM 00.8 at error at 256 ē 2.89 experts 0.2 0.2 2.67 2.50 0.0 0.0 2.38 Delay (ms) Delay (ms) 16 32 256 Num of Experts



Microbenchmarks Multi-Node P2P BW (↑ is better) Multi-Node P2P Latency (↓ is better) Single-Node P2P BW (↑ is better) 22.5 21.0 → NCCL — NVSHMEM put **NVSHMEM** get \widehat{s} 15.0 <u>5</u> 10.0 Single Node P2P Latency (↓ is better) 55 42 **NVSHMEM** put $\overset{\mathsf{ wo}}{\hspace{0.1cm}\mathsf{ wo}}\hspace{0.1cm}\mathsf{NCCL}$ **NVSHMEM** get —— NVSHMEM get NVSHMEM put (§ 2500 16B 64B 256B 1KB 4KB 16KB 64KB 56KB 1MB ু 2000 型 1500 Bytes Transmitted 1000 Multi-Node P2P Latency (↓ is better) Single Node P2P Latency (↓ is better) Non-blocking Put Latency (↓ is better) **NVSHMEM Multi-node** 3550 → NCCL → NCCL NVSHMEM put → NVSHMEM get NVSHMEM get **NVSHMEM** put Bytes Transmitted <u>></u> 2000 Laten .02 필 1500 1000 16B 64B 256B 1KB 4KB 16KB 64KB 56KB 1MB

Ongoing Work

Define G = (V, E) as the cluster topology, where V denotes devices and E communication links. Equation 1 captures the objective of challenge #3: finding $G^* = \{g = (V_g, E_g) \mid g \subseteq G\}$ the set of optimal expert parallel groups or the topology-aware sharding specification.

$$\min \max_{g \in G^*} \kappa \left(\pi(g) + \max_{i \in V_g} C_i \right) + T_{\rho}(|G^*|) \quad (1)$$

subject to,

$$\sum_{j \in V_q} m_j \ge |\mathcal{X}| \qquad \forall g \in G^* \tag{2}$$

where, \mathcal{X} is the set of all experts and m_j is expert memory capacity for device j. Also, κ identifies the frequency of MoE computation, $\pi(g)$ is the compute cost of g, C_i denotes the communication cost of device i and $T_{\rho}(|G^*|)$ is the cost of inter-group all-reduce on MoE parameters due to data parallelism. We also note that CUDA development for Aristos Fused is underway.

Acknowledgements

We thank Dr. Rachee Singh for her guidance; Dr. Guila Guidi for Perlmutter access under award DDR-ERCAP0027296 of the National Energy Research Scientific Computing Center (NERSC); and Julian Bellavita for invigorating discussions.

References

- [1] Fedus et al. "Switch Transformers". In: JMLR 23 (2022).
- 2] Gale et al. "MegaBlocks". In: *MLSys.* 2023.
- [3] GeminiTeam et al. Gemini 1.5. 2024.
- [4] MosaicResearch. Introducing DBRX. 2024.
- [6] Liu et al. "Janus". In: SIGCOMM '23.
- [6] Rajbhandari et al. "DeepSpeed-MoE". In: ICML '22.