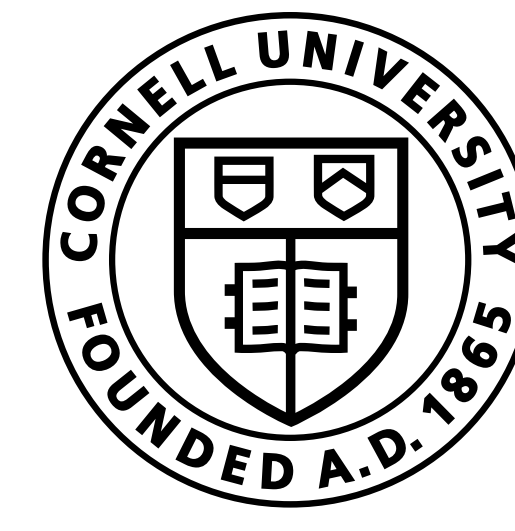


# Aristos: Pipelining One-sided Communication in Distributed Mixture of Experts (MoE)

Osayamen Jonathan Aimuyo<sup>†</sup>

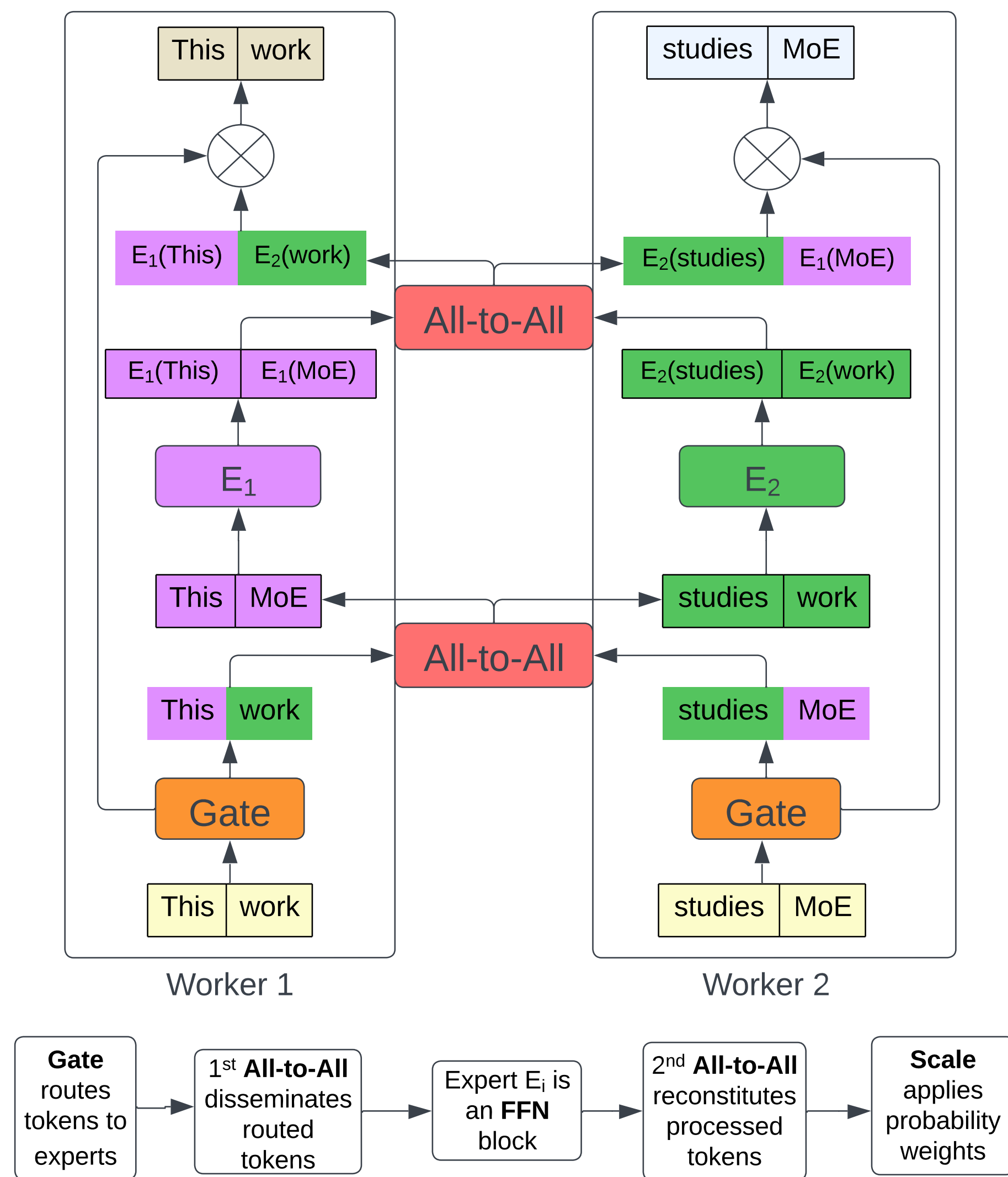
oja7@cornell.edu <sup>†</sup> Cornell Ann S. Bowers College of Computing and Information Science, Cornell University



Cornell Bowers CIS  
Computer Science

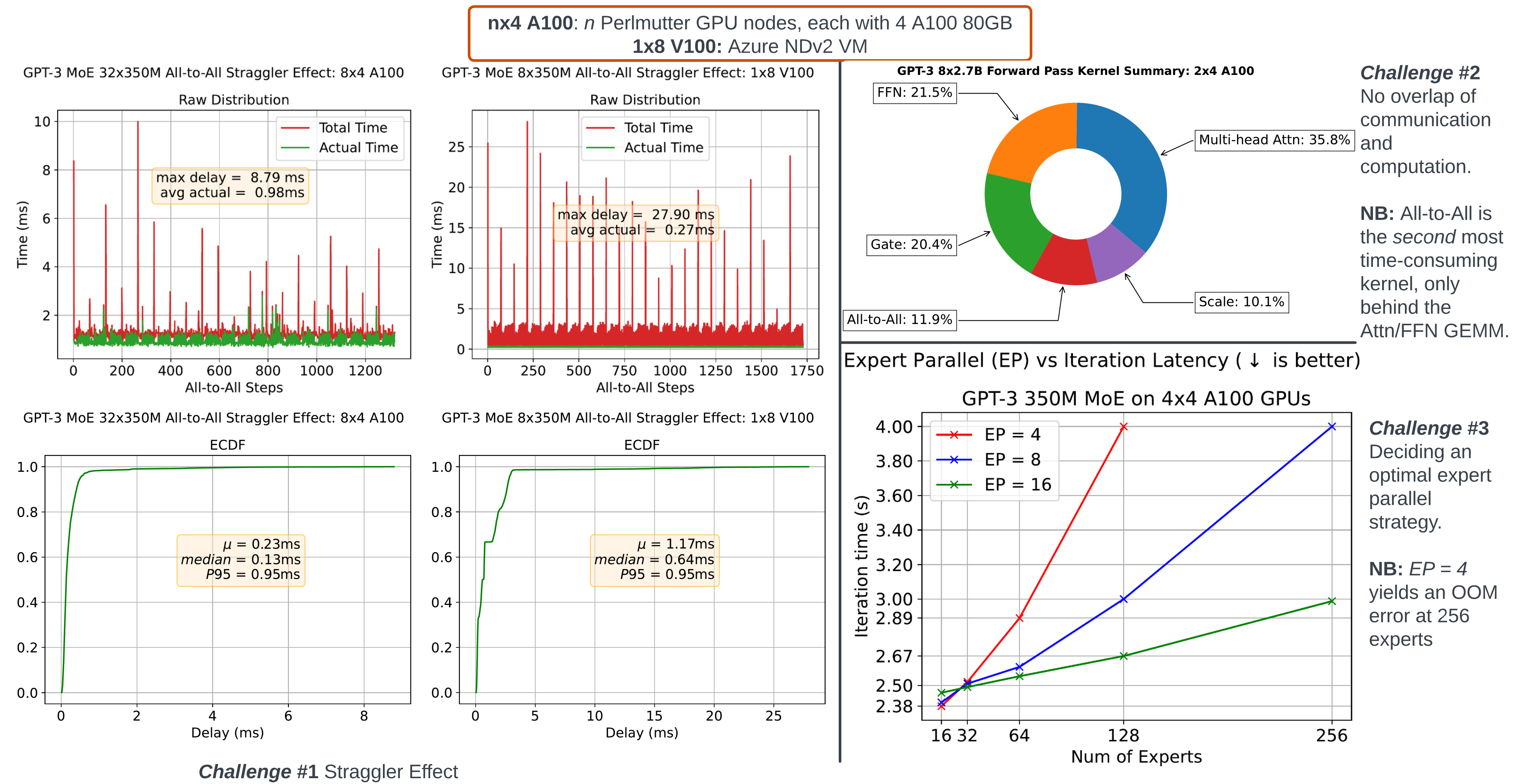
## Background

Distributed MoE Token Flow, where  $|W| = EP = 2$



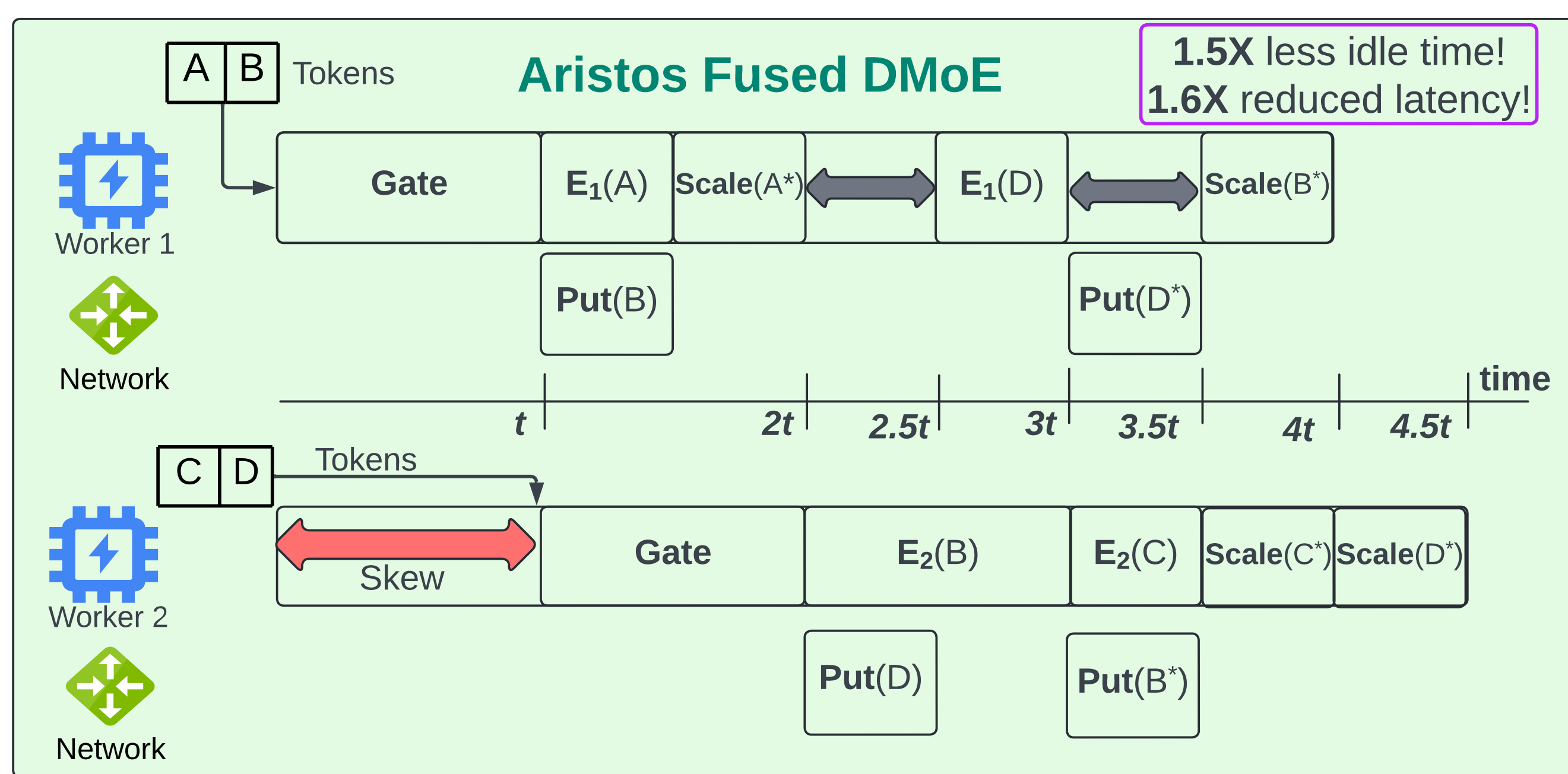
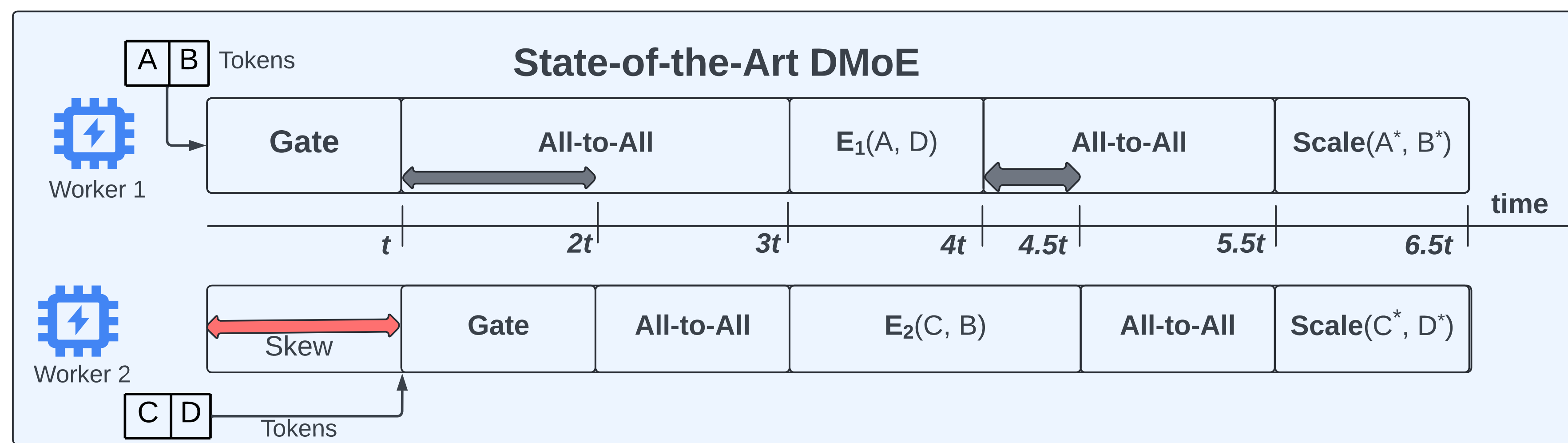
## Challenges

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.



## Method

Idle Time



System architecture of Aristos

## 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*.

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

subject to,

$$\sum_{j \in V_g} m_j \geq |\mathcal{X}| \quad \forall g \in G^* \quad (2)$$

where,  $\mathcal{X}$  is the set of all experts and  $m_j$  is expert memory capacity for device  $j$ . Also,  $\kappa$  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 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.
- [5] Liu et al. "Janus". In: *SIGCOMM '23*.
- [6] Rajbhandari et al. "DeepSpeed-MoE". In: *ICML '22*.

## Microbenchmarks

