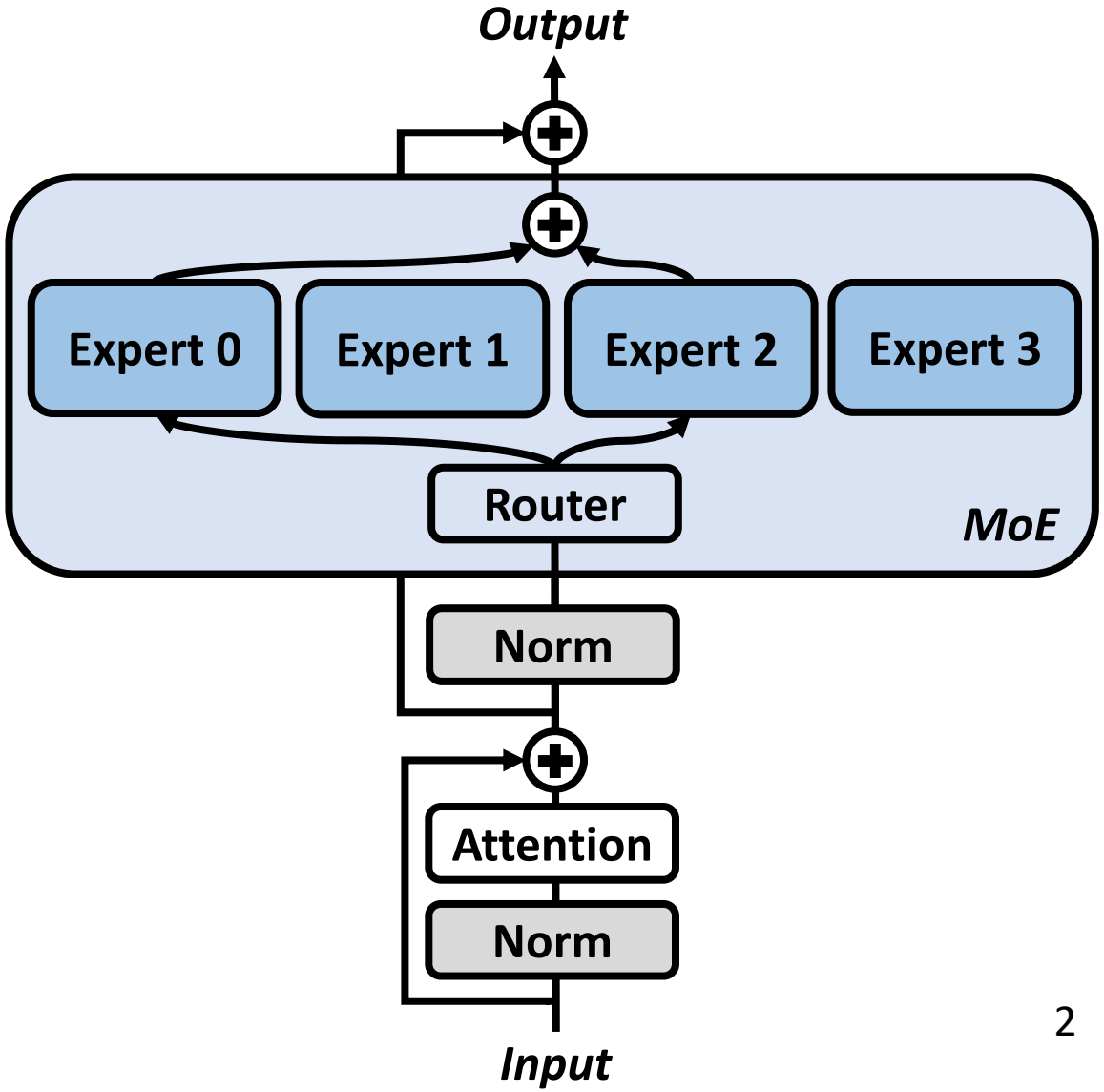
CRAFT: Fine-Grained Cost-Aware Expert Replication For Efficient Mixture-of-Experts Serving

Adrian Zhao, *Zhenkun Cai, Zhenyu Song, Lingfan Yu,*
Haozheng Fan, Jun Wu, Yida Wang, Nandita Vijaykumar

UNIVERSITY OF TORONTO

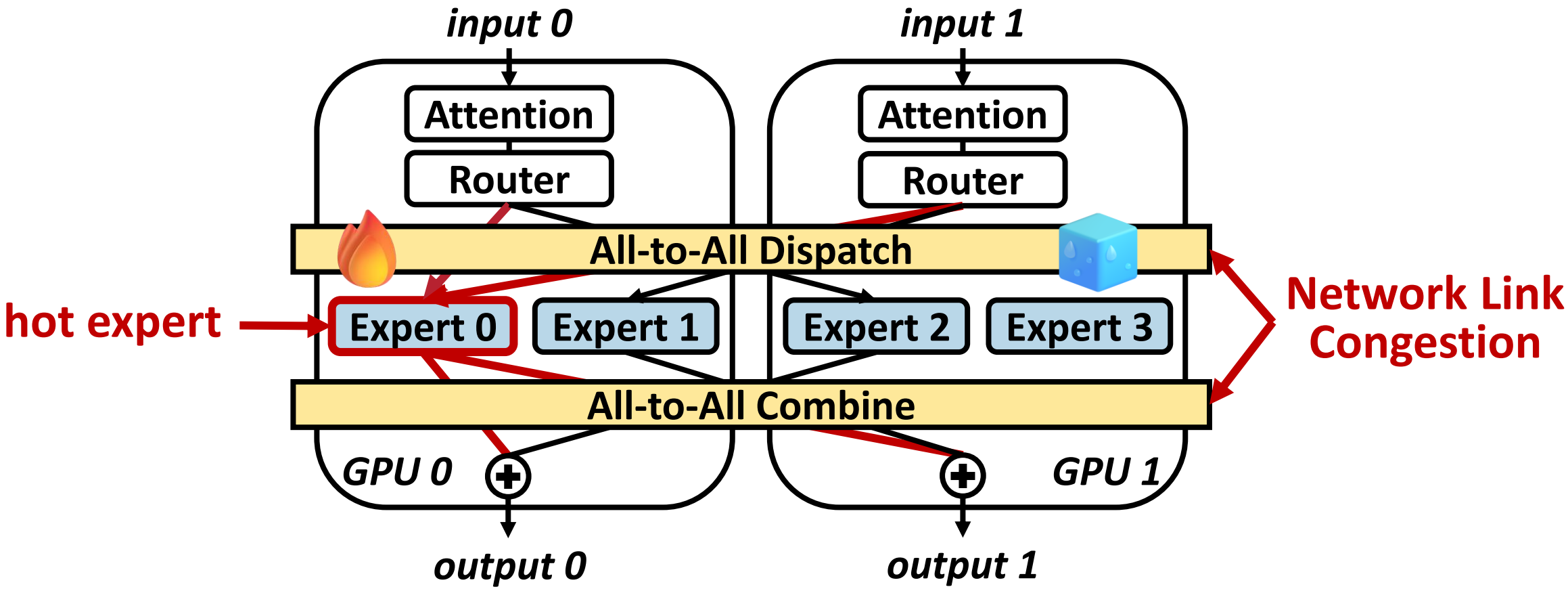
MoE: Sparse Activation for Efficient Scaling

- Activating all parameters for each token is expensive
- MoE's *sparse activation* scales parameters at constant compute
- Industrial standard powering the SOTA models:



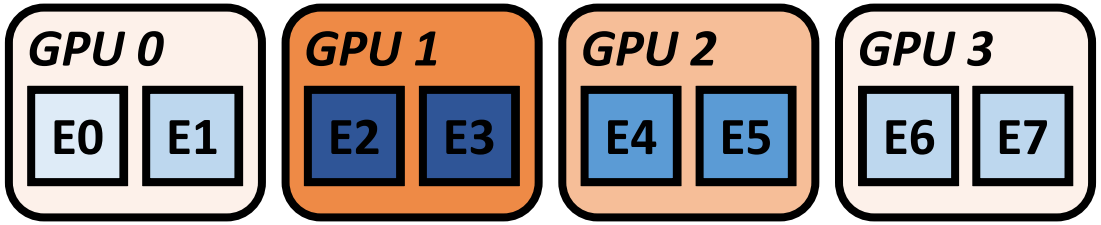
Expert parallelism introduces load imbalance

- Real-world tokens are skewed, leading to **expert load imbalance**



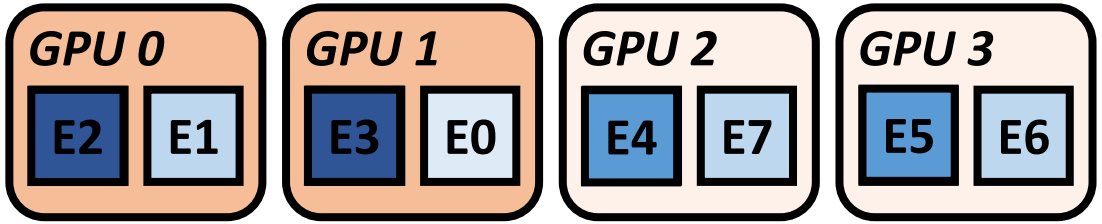
Existing fix: Expert placement & replication

Trivial EP baseline (by ID)
No load balancing



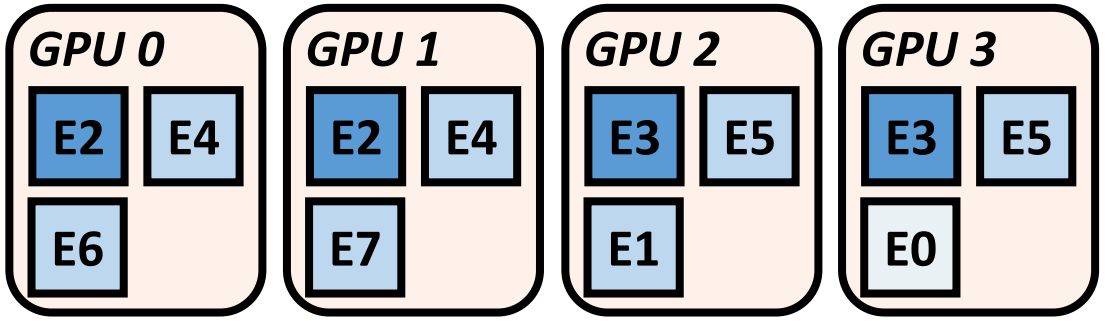
Placement

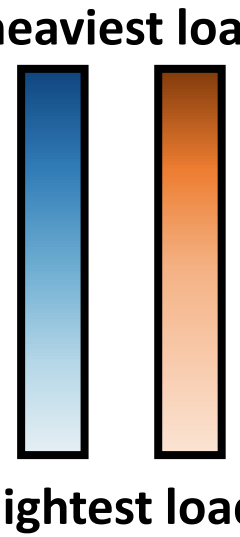
× imbalance remains



Placement & Replication

× high memory cost

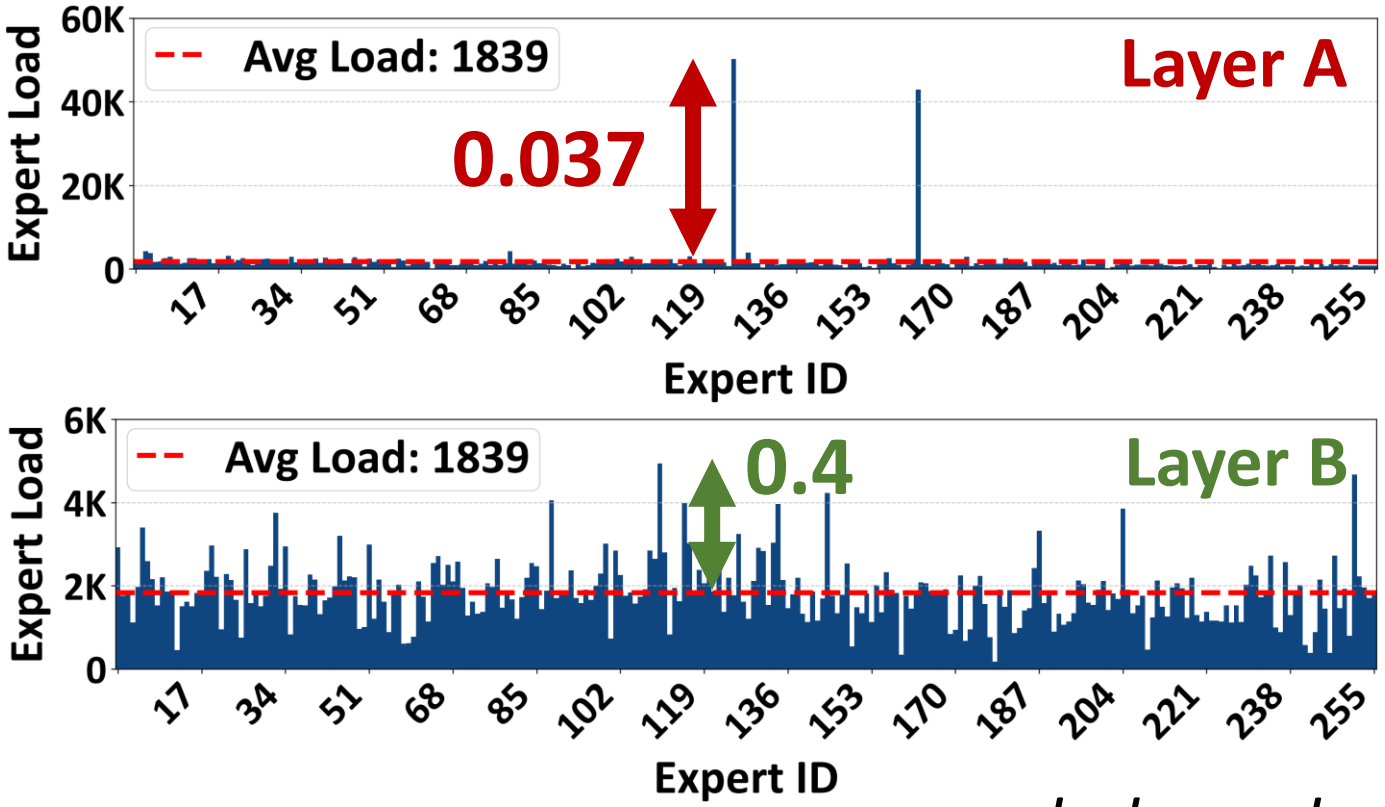




Our Goal:
✓ load balance ↑
✓ memory cost ↓

OB 1: Replication benefit varies across layers

- Different layers exhibit different level of expert load imbalance



Replication

+ 0.579 = 0.616
high benefit

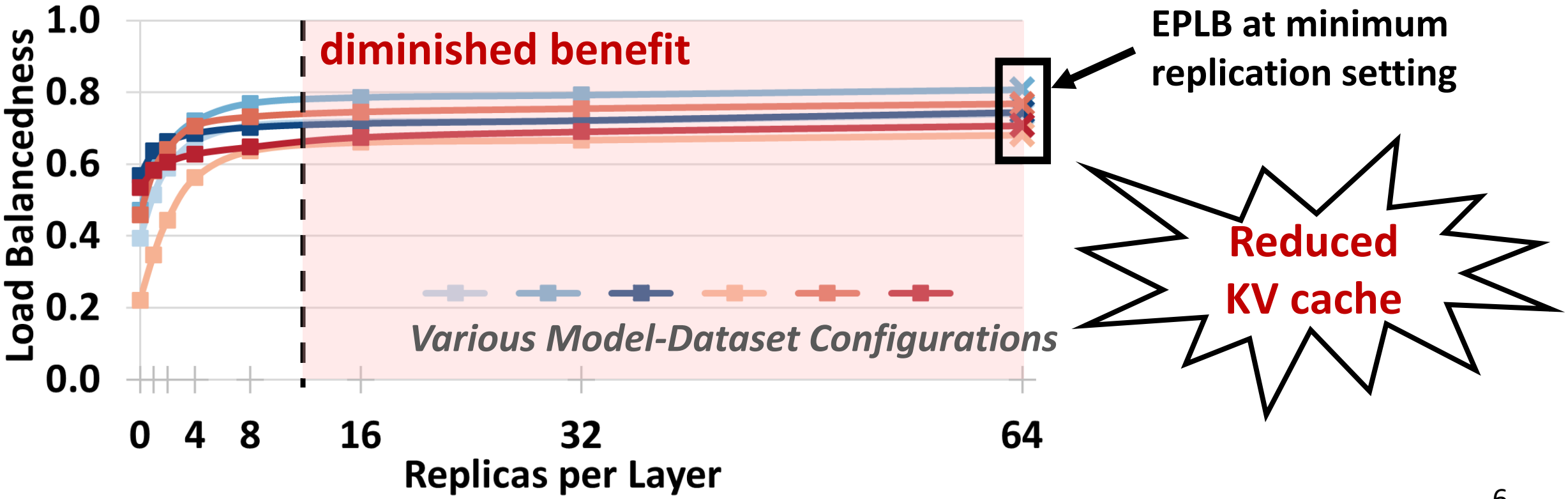
+ 0.031 = 0.431
low benefit

balancedness = mean load / max load

same batch, same token load

OB 2: Replication benefit diminishes rapidly

- Load balancedness gain does not scale linearly with # of replicas
- Existing replication scheme replicates **excessively**

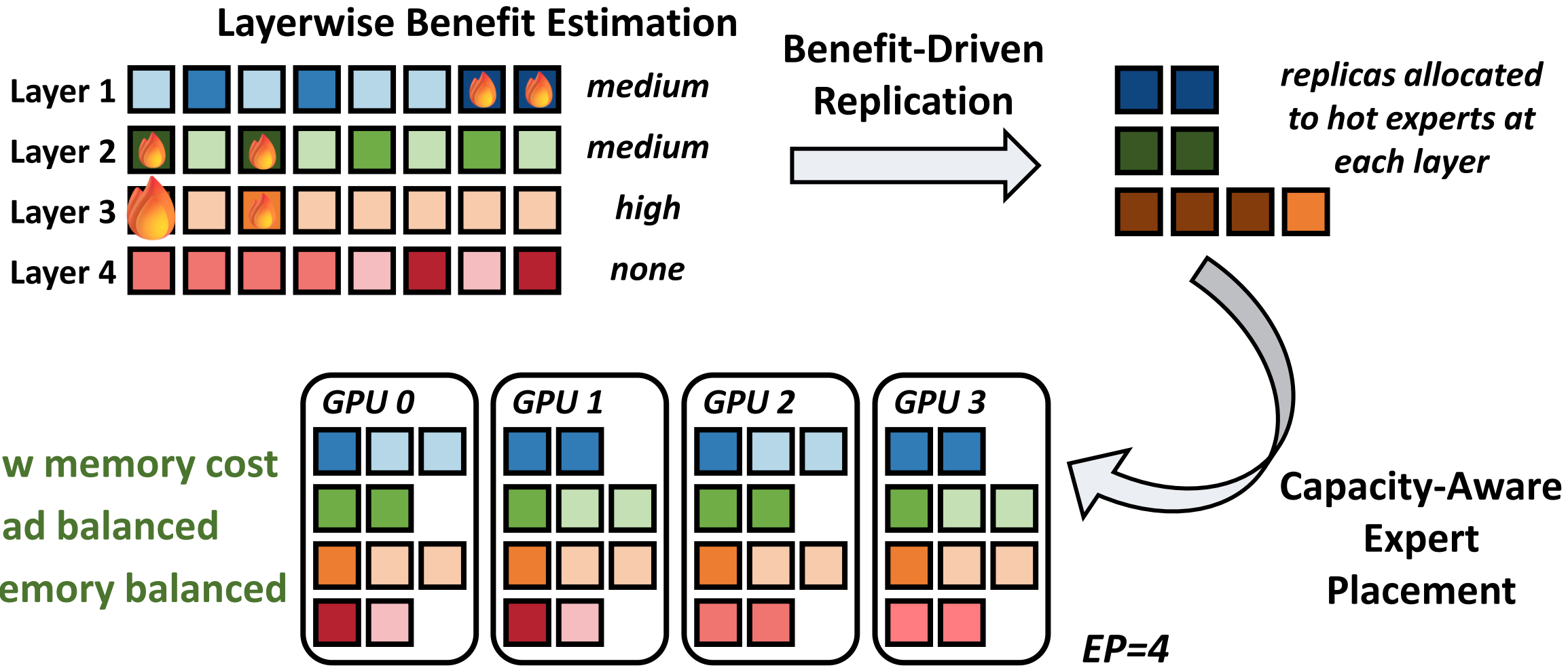


Liu, Aixin, et al. "Deepseek-v3 technical report." *arXiv preprint arXiv:2412.19437* (2024).

Challenges: Fine-grained replication strategy

- **Goal**: maximize serving goodput
- **Challenge 1**: Determining optimal per-layer replica count
  - More replicas: load balance (speed) ↑, KV cache (concurrency) ↓
- **Challenge 2**: Placing replicas onto GPUs while balancing both token load and total expert count (memory usage) across GPUs
  - Different number of experts per layer
  - Maintain symmetric GPU memory layout while balancing load

CRAFT: Benefit-driven fine-grained replication

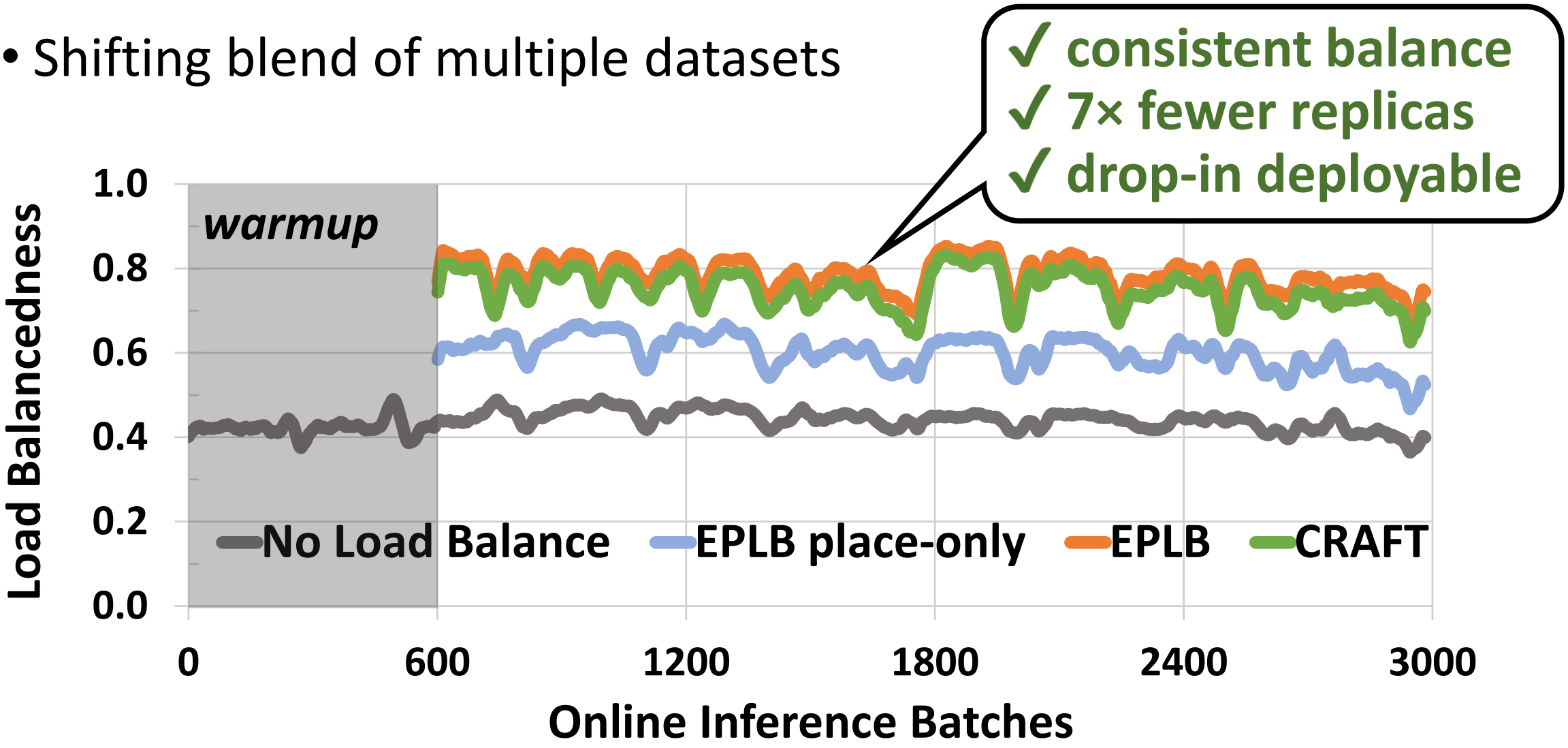


Evaluation Methodology

- **Hardware:** 8-node, 8 NVIDIA A100 80G per node
- **Model:** DeepSeek-R1-671B
- **Parallelism:** DP=8, TP=8, EP=64
- **Network:** Inter-node EFA, intra-node NVLink
- **Serving Framework:** SGLang
- **Datasets:** FinePDF, Lambada, RedPajama-1T

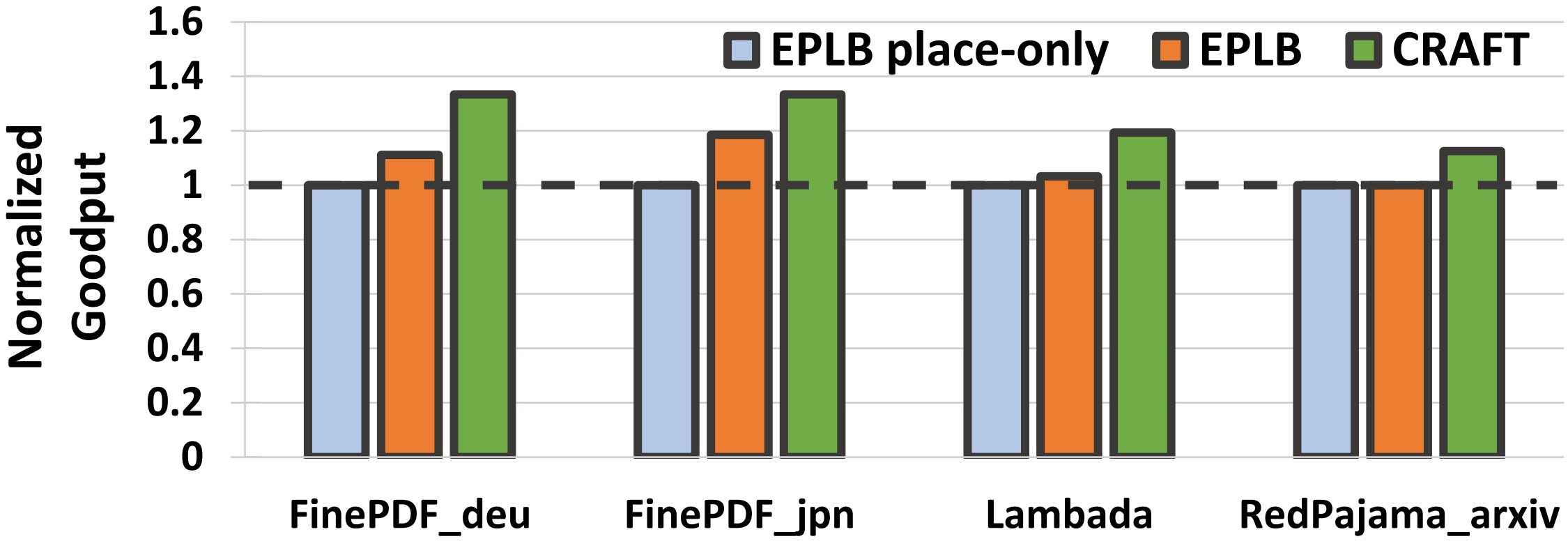
Evaluation: Balanced expert load at low cost

- Shifting blend of multiple datasets



Evaluation: CRAFT improves serving goodput

- Up to **1.2× (1.15× on average)** higher goodput than EPLB



Summary

- **Problem:** Existing expert replication schemes trade off memory for expert load balancedness
- **Observations:**
  - Layers benefit from replication **differently**
  - Expert load balance scales **sub-linearly** with number of replicas
- **Our Solution: CRAFT**: A flexible, memory-efficient replication method that:
  - Approximates per-layer replication benefit headroom
  - Formulates layerwise fine-grained replica allocation as a knapsack problem
  - Drop-in replace EPLB and supports flexible memory budget / online rebalancing
- **Key Results:** CRAFT achieves consistently high expert load balance with **efficient replica allocation**, improving **serving goodput by up to 1.2×**