OAK RIDGE National Laboratory

Training Mixture of Experts Models on Frontier

Sajal Dash

Analytics and AI Methods at Scale (AAIMS)

NCCS, ORNL

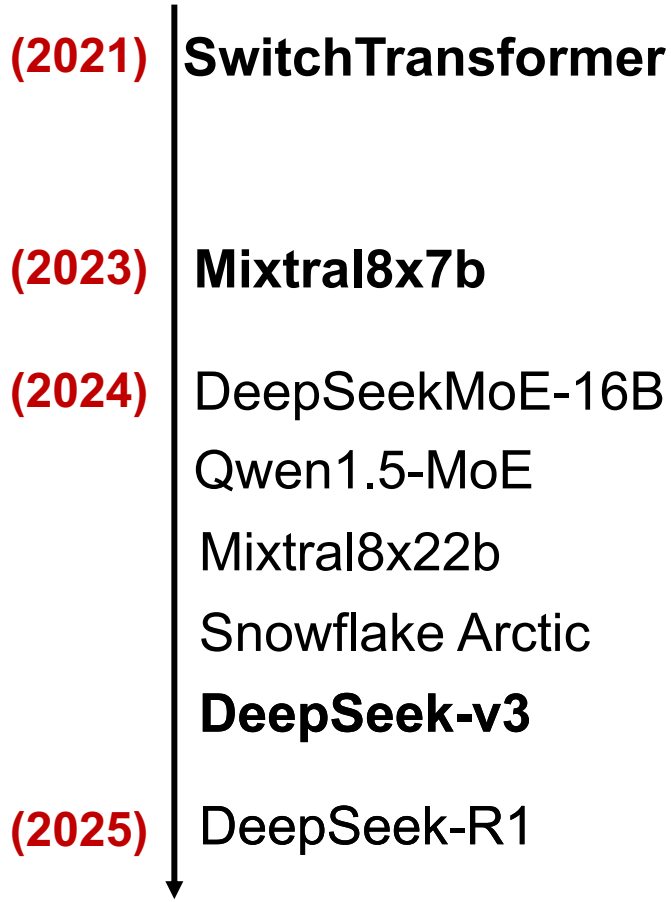
dashs@ornl.gov

ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

A Case for Sparsely Activated Model

- Models with more parameters perform better
- More parameters means more compute time
- "Mixture of Experts" sparsely activates model parameters, thus limiting computational requirement

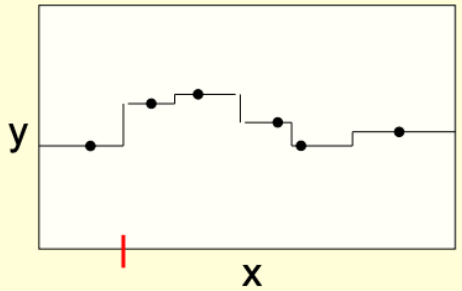


Mixture of Experts Intuition (From Geoffrey Hinton)

A spectrum of models

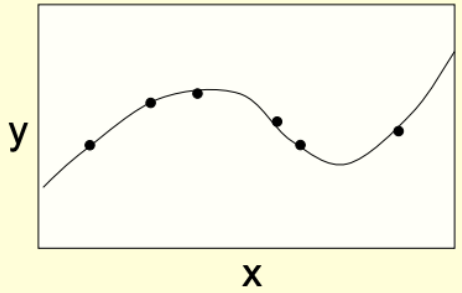
Very local models

- e.g. Nearest neighbors
- Very fast to fit
  - Just store training cases
- Local smoothing obviously improves things



Fully global models

- e. g. Polynomial
- May be slow to fit
  - Each parameter depends on all the data

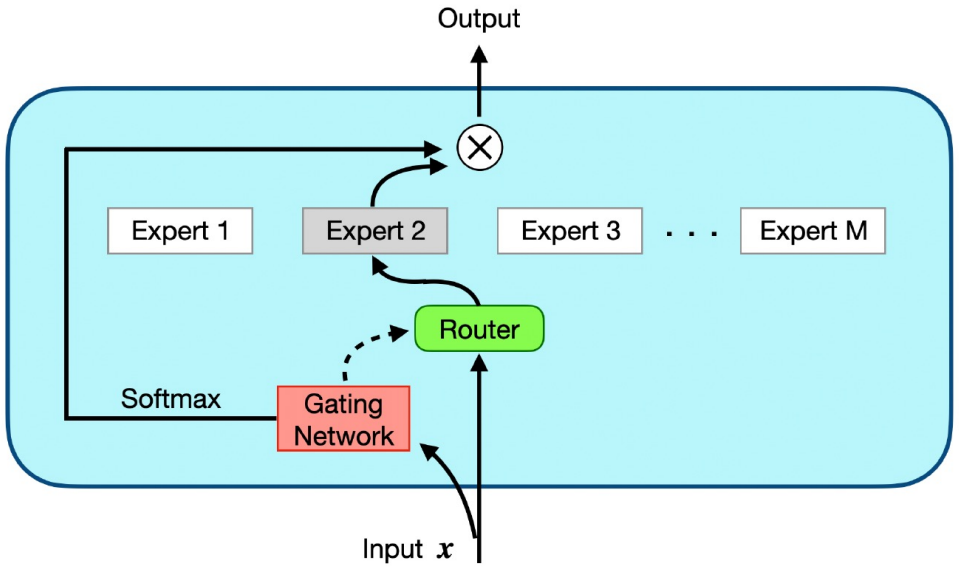


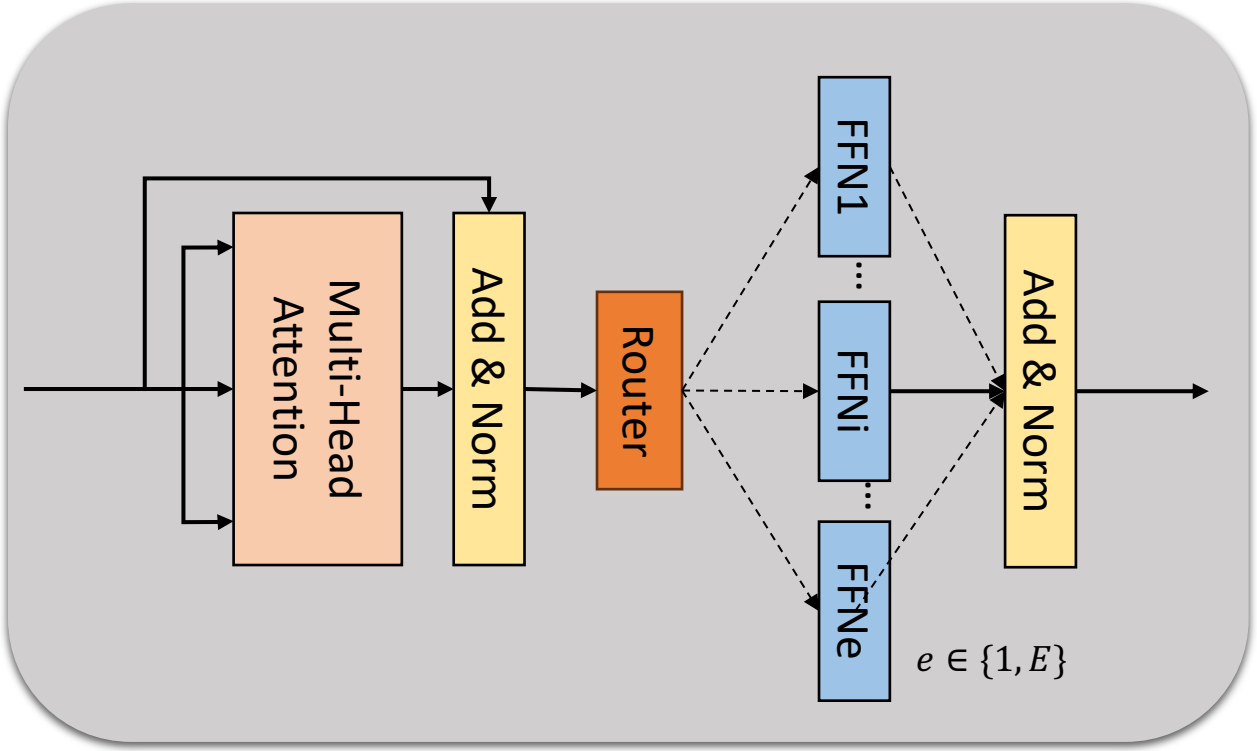
Multiple local models

- Instead of using a single global model or lots of very local models, use several models of intermediate complexity.
  - Good if the dataset contains several different regimes which have different relationships between input and output.
  - But how do we partition the dataset into subsets for each expert?

Transformer to Mixture of Experts

- GPT-style LLMs have Attention blocks and FFN blocks
- Most MoE models implement the FFN layer as Mixture of Experts





Number of Parameters in an MoE Model

- Essential computational budget is "same"+ as the dense model with one expert
-

Design Parameter	Options
Number of Experts (E)	{8, 16, 32, 48, 64, 128}
Experts per token ($topK$)	{1, 2}
Expert frequency (f)	{0.5, 1}
Base Model Size (M)	{1.3B, 7B, 13B}

TABLE I: MoE Design Space

Attention: non-expert

Frequency x num_experts

$$P = 4LD^2 + f \times (E + \frac{1}{f} - 1) \times 8LD^2$$

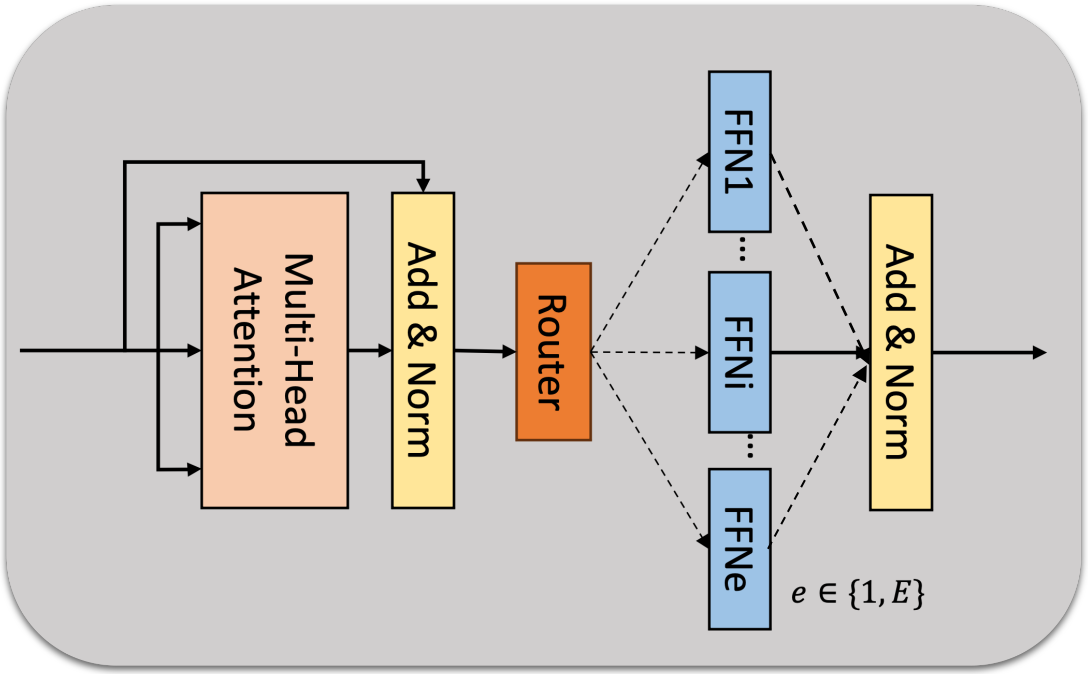
FFN: expert

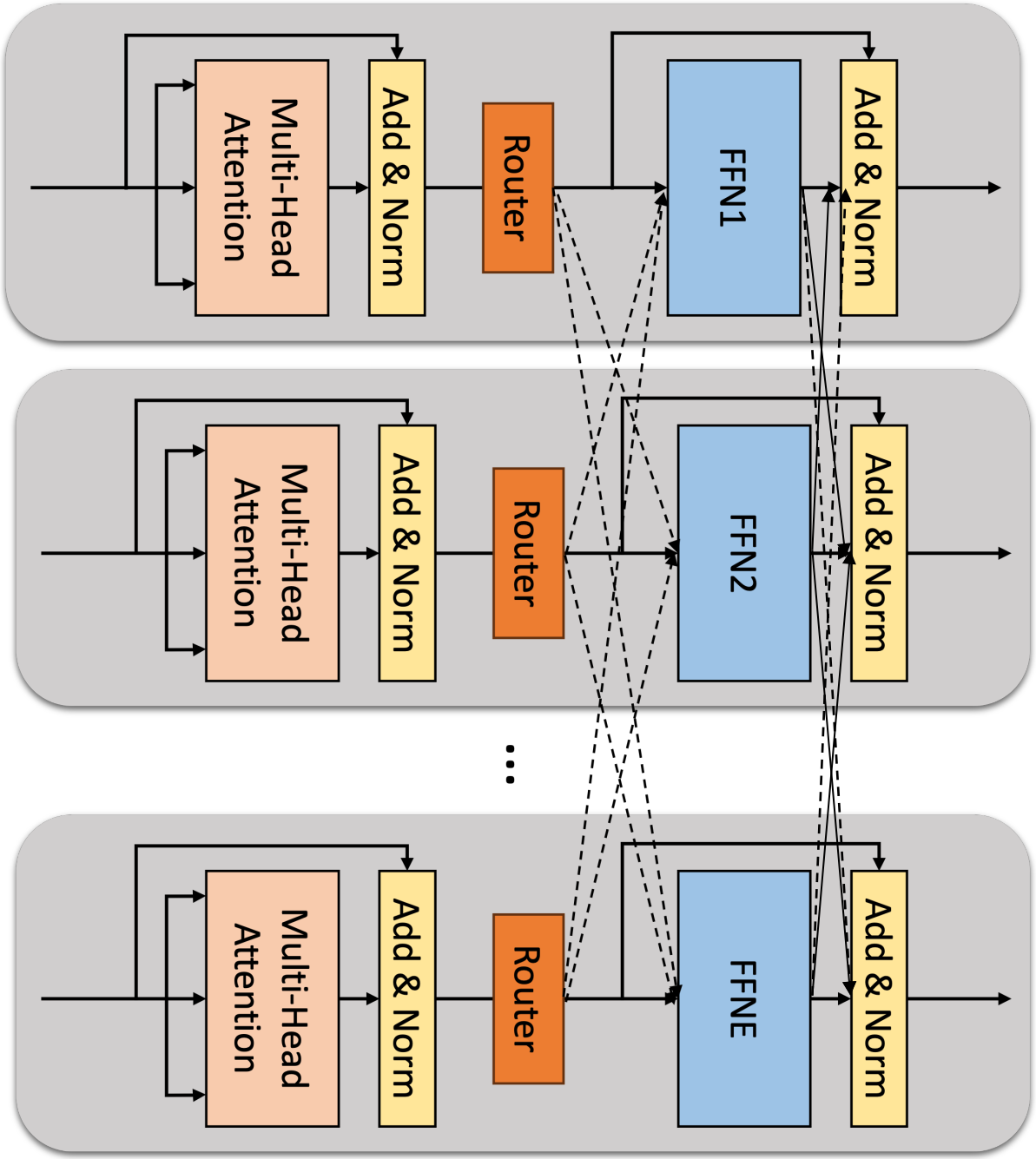
$$= 4LD^2\left(1 + 2f(E + \frac{1}{f} - 1)\right) \quad (1)$$

$$= 4LD^2(3 + 2fE - 2f)$$

Therefore, a target MoE model size is estimated as $P \approx (1 + \frac{2f}{3}E) \times M$.

Expert-Data Parallelism





Comparing parallelization strategies from DeepSpeed-MoE

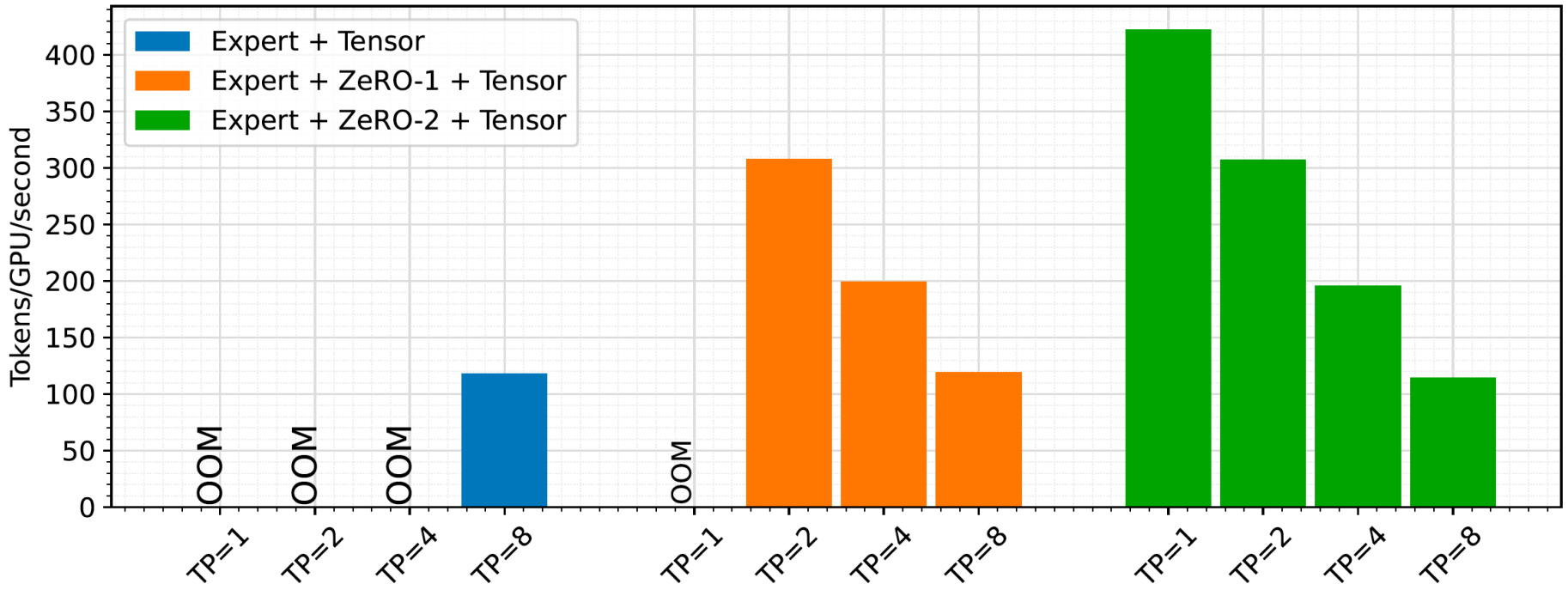


Fig. 6: Comparing combinations of parallelization strategies for training SMORE-6.7B-8E. Expert-Data Parallelism + ZeRO-2 Sharding without any model/tensor parallelism, performs the best.

Tutel (An adaptive parallelization tool) with DeepSpeed

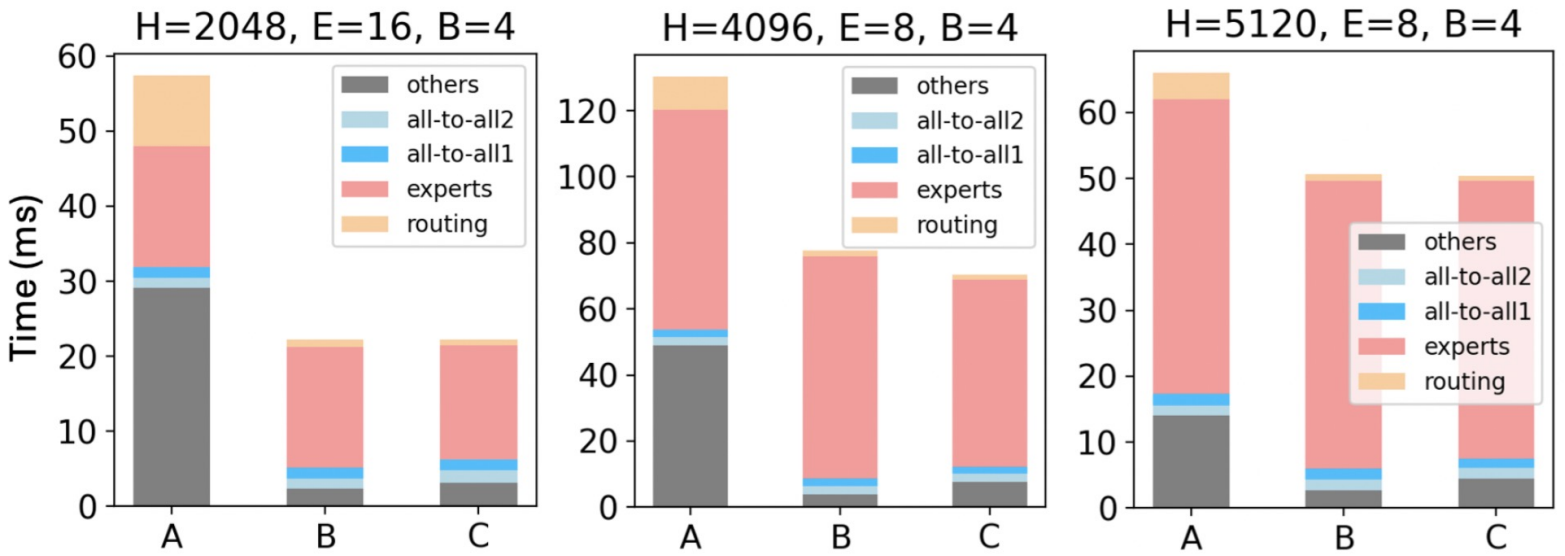


Figure 5: MoE forward layer time breakdown using DeepSpeed-MoE and Tutel. A: DeepSpeed-MoE layer. B: DeepSpeed-MoE layer + Tutel's Fast Encode & Decode. C: Tutel MoE layer.

Training Large Models with Tutel

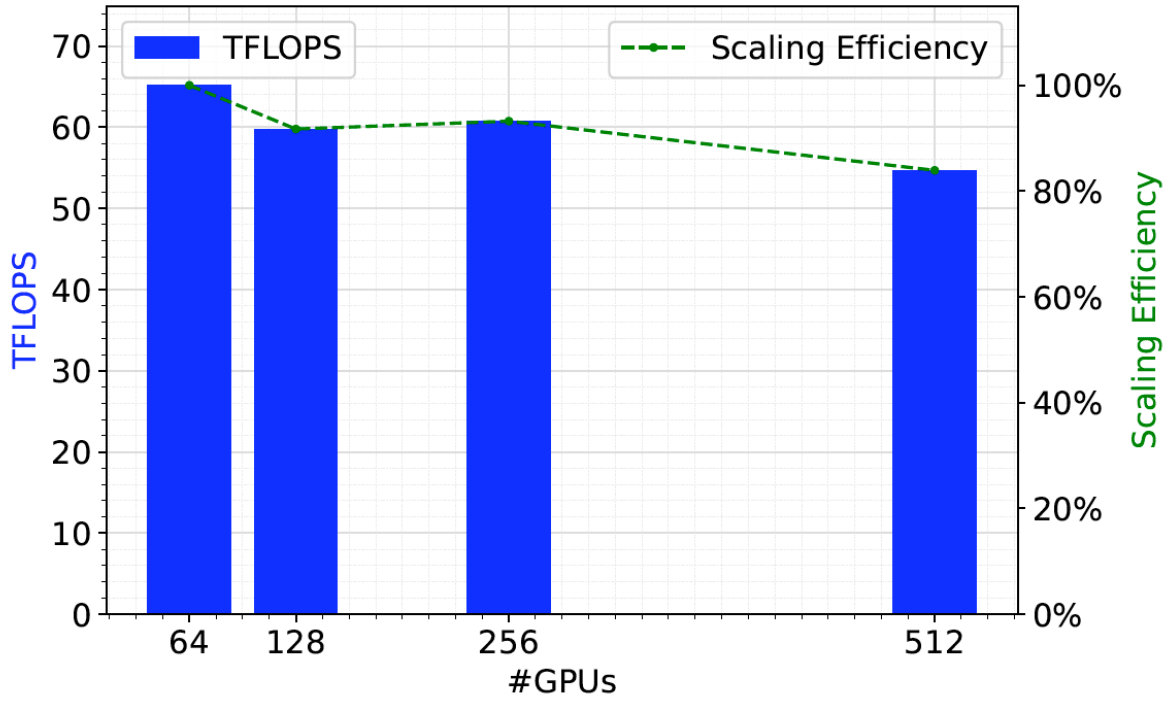


Fig. 8: Scaling the 13B model from 8 experts (on 64 GPUs) to 64 experts (on 512 GPUs). The total parameters vary from 82 Billion parameters to 562 Billion parameters.

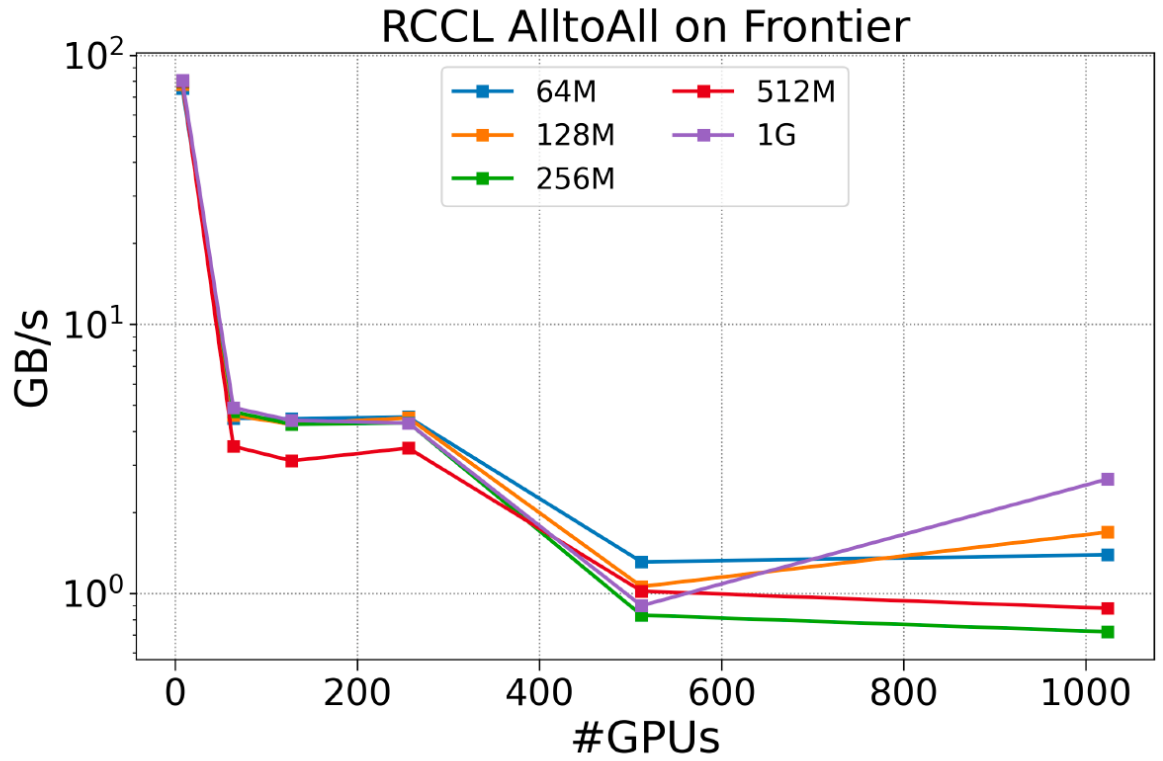
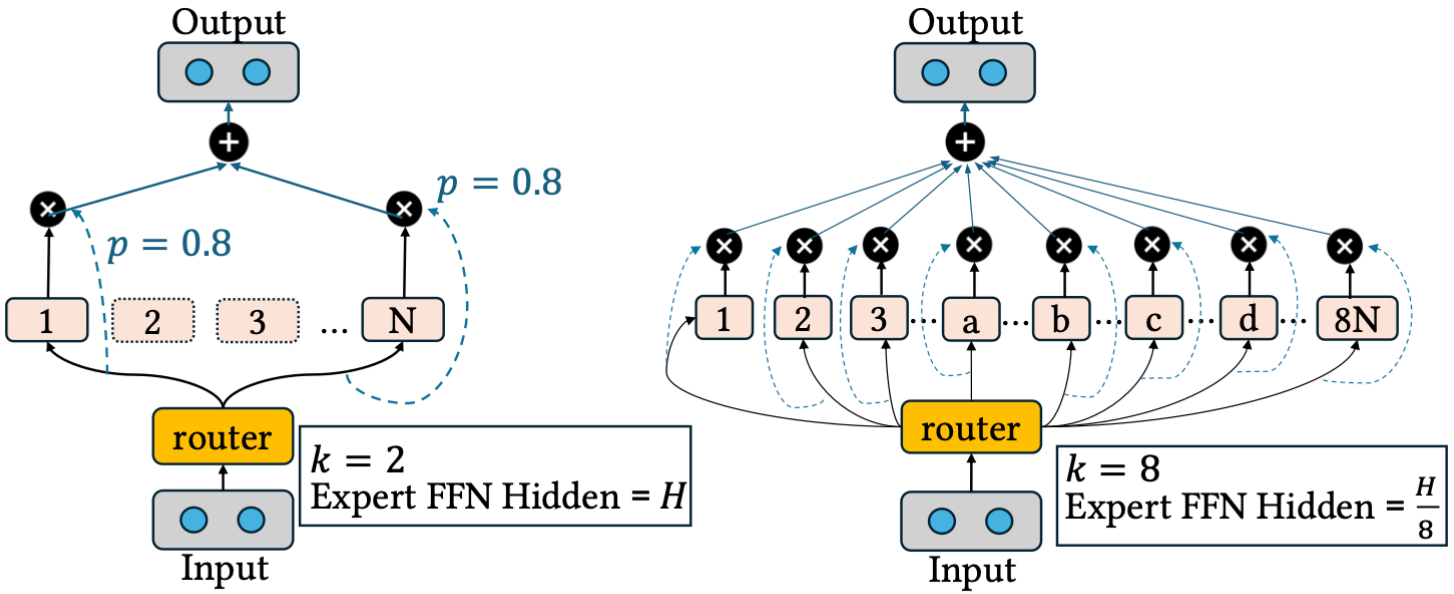


Fig. 10: Scaling performance of RCCL all-to-all communication across GPUs. Achieved bandwidth reduces with increasing number of GPUs participating in all-to-all collective communication.

New model trend: Expert-specialized MoE



(a) Few experts + small top-k routing

(b) Fine-grained experts + large top-k routing

Conventional MoE vs. Expert-specialized MoE

More experts, smaller size per expert

More activated experts per token

Popular expert-specialized MoE models

DeepSeek-MoE

DeepSeek-v3

Qwen3-MoE

Challenges: Training on HPC platforms

Many popular HPC platforms are equipped with
- non-NVIDIA GPUs
- hierarchical network topology

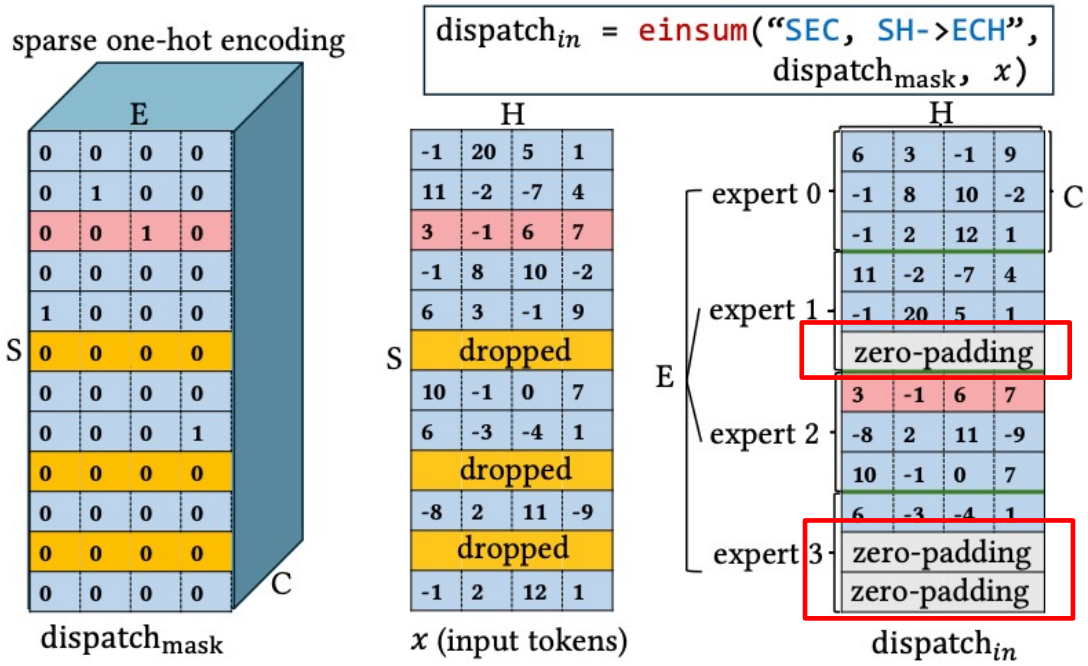
- When training **expert-specialized MoE** using DeepSpeed-MoE…

  <10 TFLOPs when training with 64 GPUs on *Frontier* (AMD MI250X GPUs)

- Other existing frameworks (e.g., Tutel, Megablocks): rely on NVIDIA ecosystem

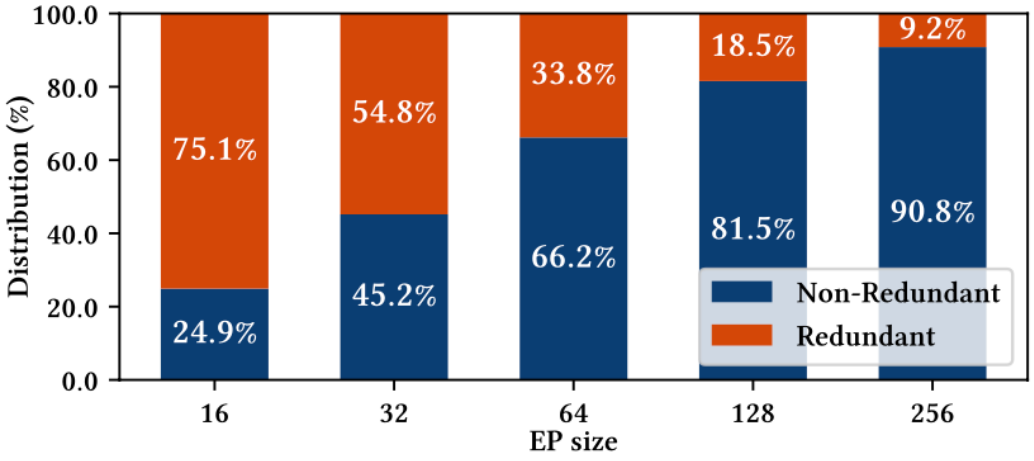
  Cannot run on non-NVIDIA platform or suffer from inefficient port

Observation: All-to-all inefficiency



Zero-paddings in dense representations: large all-to-all overhead for expert-specialized MoE

Duplicated tokens may be sent to the same destination node when k is large:

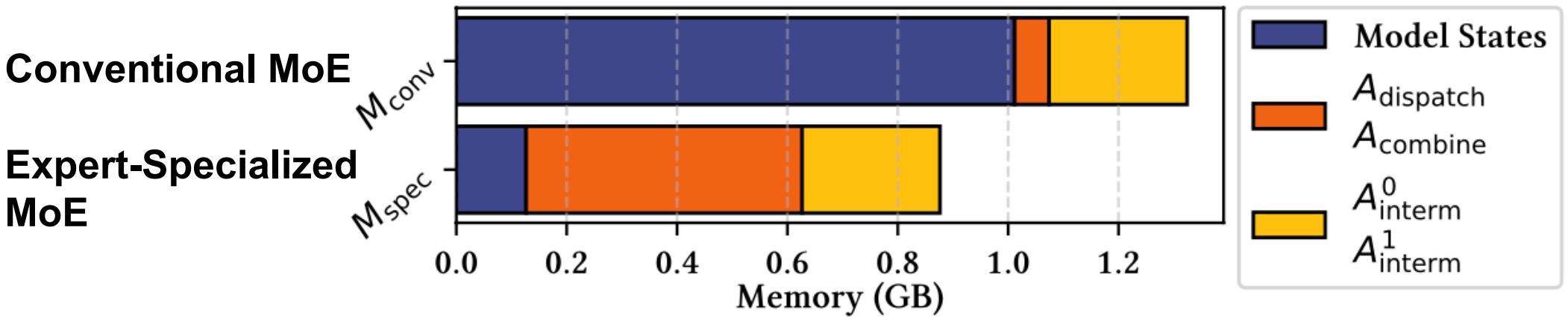


Redundancy on inter-node network caused by large k

Observation: Memory bottleneck shift

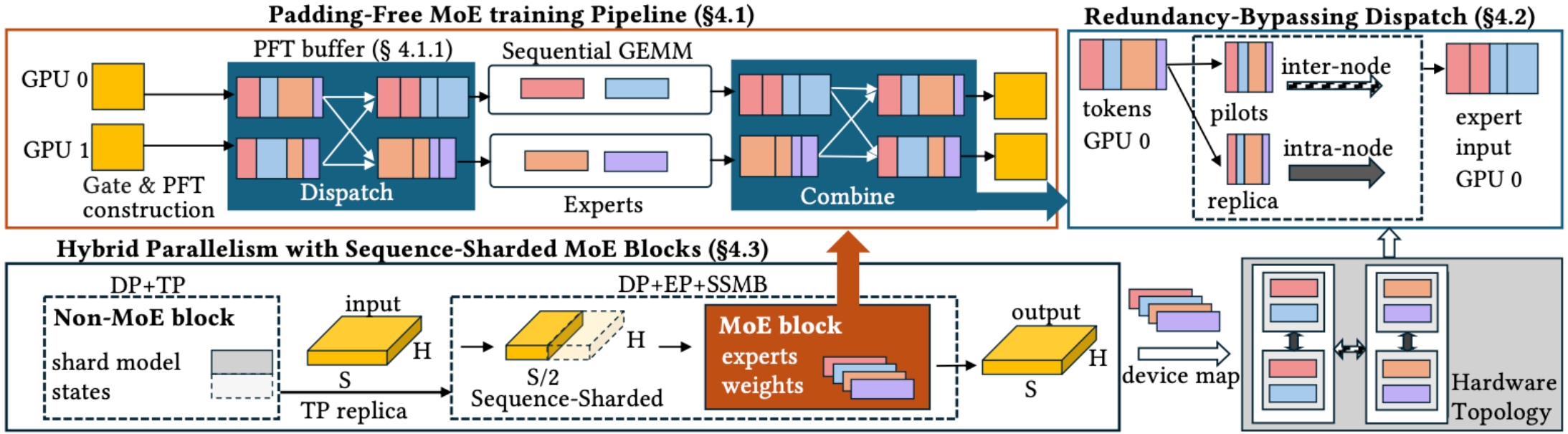
- **Lower model states usage**
- **Higher activation memory usage (**  **)**

Memory consumption with EP (GB per GPU)



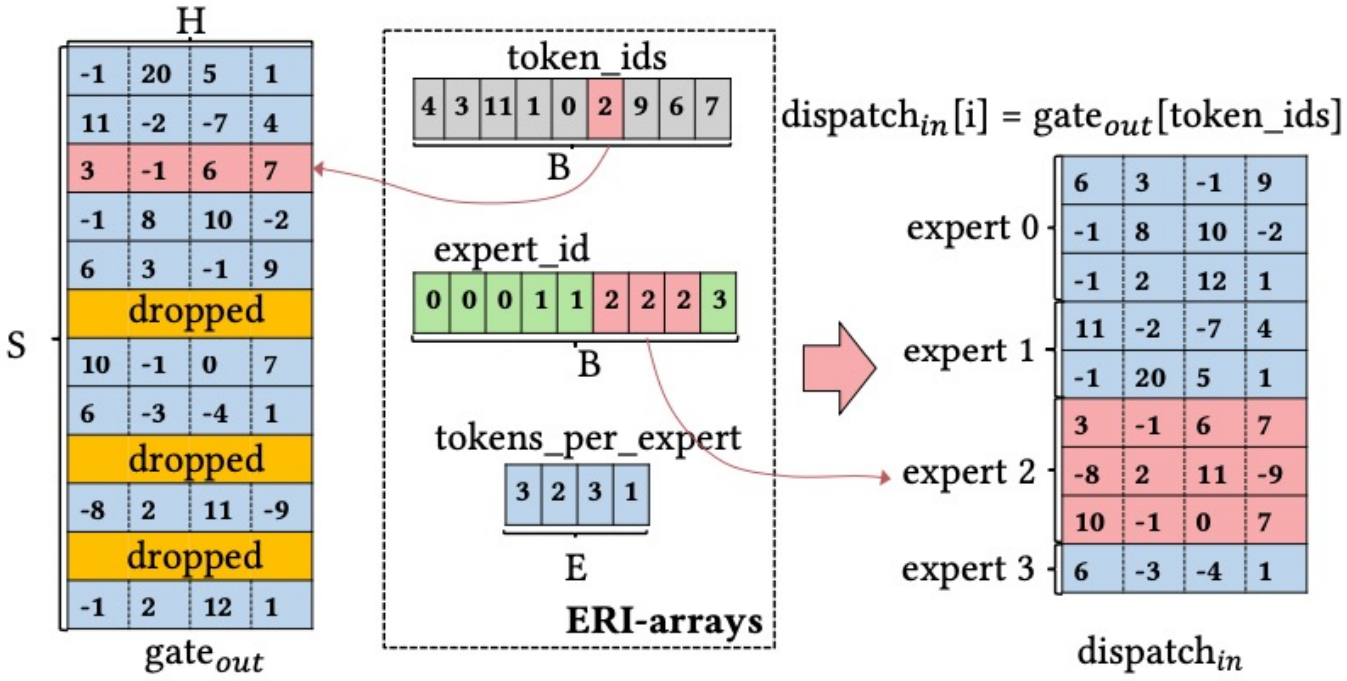
X-MoE

- **Cross-platform**: optimized with Triton, friendly to non-NVIDIA GPUs
- Designed for **expert-specialized MoE** features



Padding-Free Token Buffers (PFT)

- Fully **padding-free**
- Optimized with hardware-agnostic **Triton kernels**

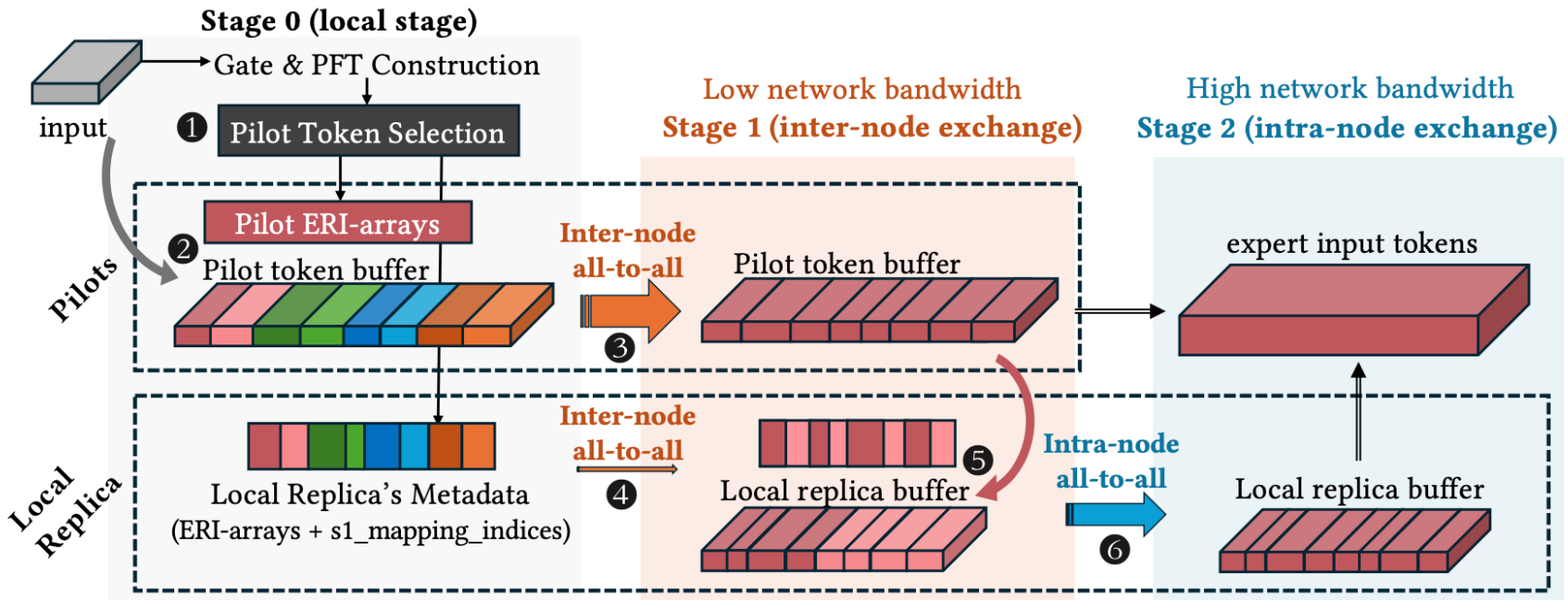


Advantages:

- Eliminate all-to-all overhead on paddings
- Memory-efficient

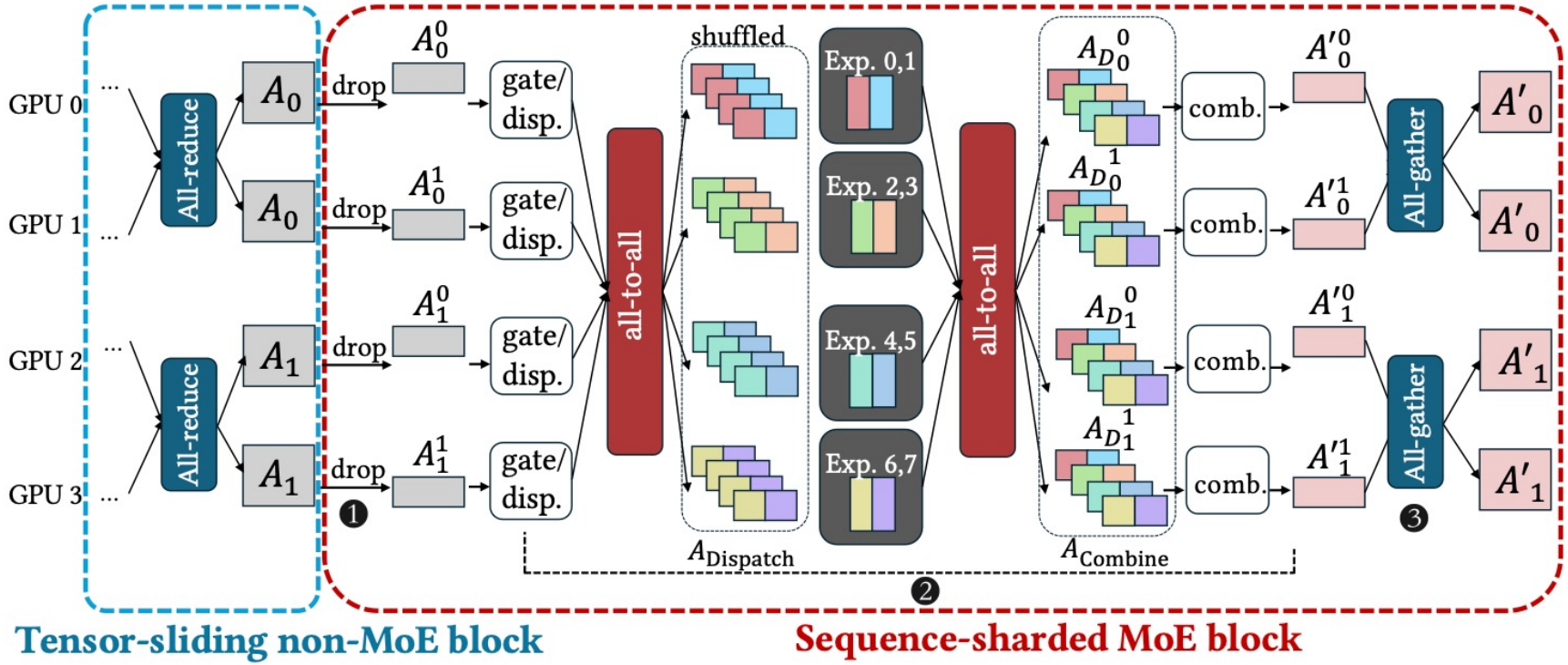
Redundancy-Bypassing Dispatch (RBD)

- Eliminate **redundant inter-node communication**
- Use two-staged, inter-node then intra-node all-to-alls



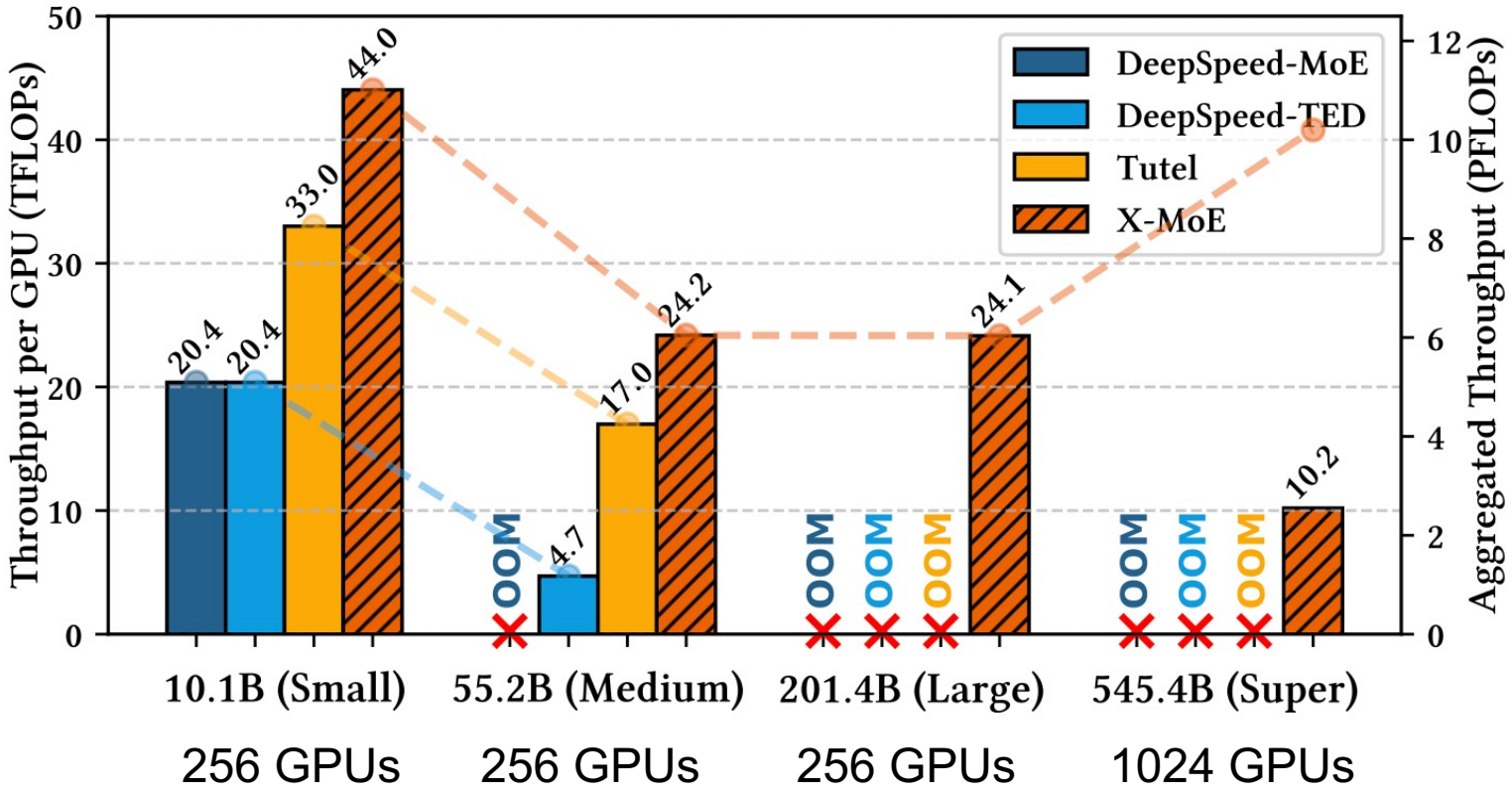
Sequence-Sharded MoE Blocks (SSMB)

- Hybrid parallelism combined with tensor-sharding (TP)
- **Shard sequences** in expert-specialized MoE block



Evaluation

- Scales DeepSeek-style MoEs up to **545** billion parameters
- Up to **1.42x** higher training throughput than baselines



Conducted on Frontier (AMD MI250X GPUs, Slingshot interconnect)