EuroSys 2026

MegaScale-MoE: Large-Scale Communication-Efficient Training of Mixture-of-Experts Models in Production

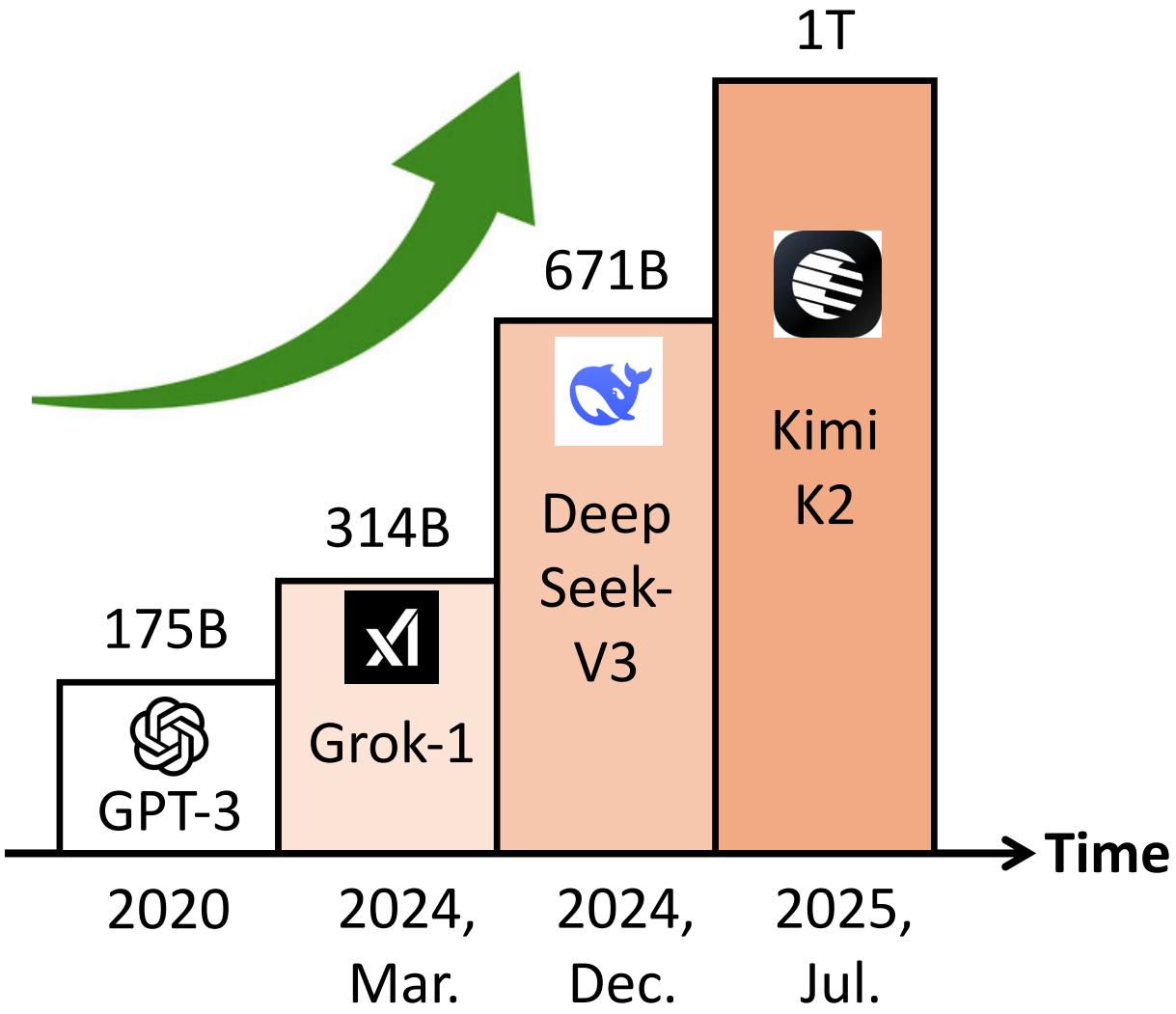
Chao Jin[1*], Ziheng Jiang[2*], Zhihao Bai[2], Zheng Zhong[2], Juncai Liu[2], Xiang Li[2], Ningxin Zheng[2], Xi Wang[2], Cong Xie[2], Qi Huang[2], Wen Heng[2], Yiyuan Ma[2], Wenlei Bao[2], Size Zheng[2], Xuegui Zheng[2], Yanghua Peng[2], Haibin Lin[2], Xuanzhe Liu[1], Xin Jin[1], Xin Liu[2]

[1] 北京大学 PEKING UNIVERSITY

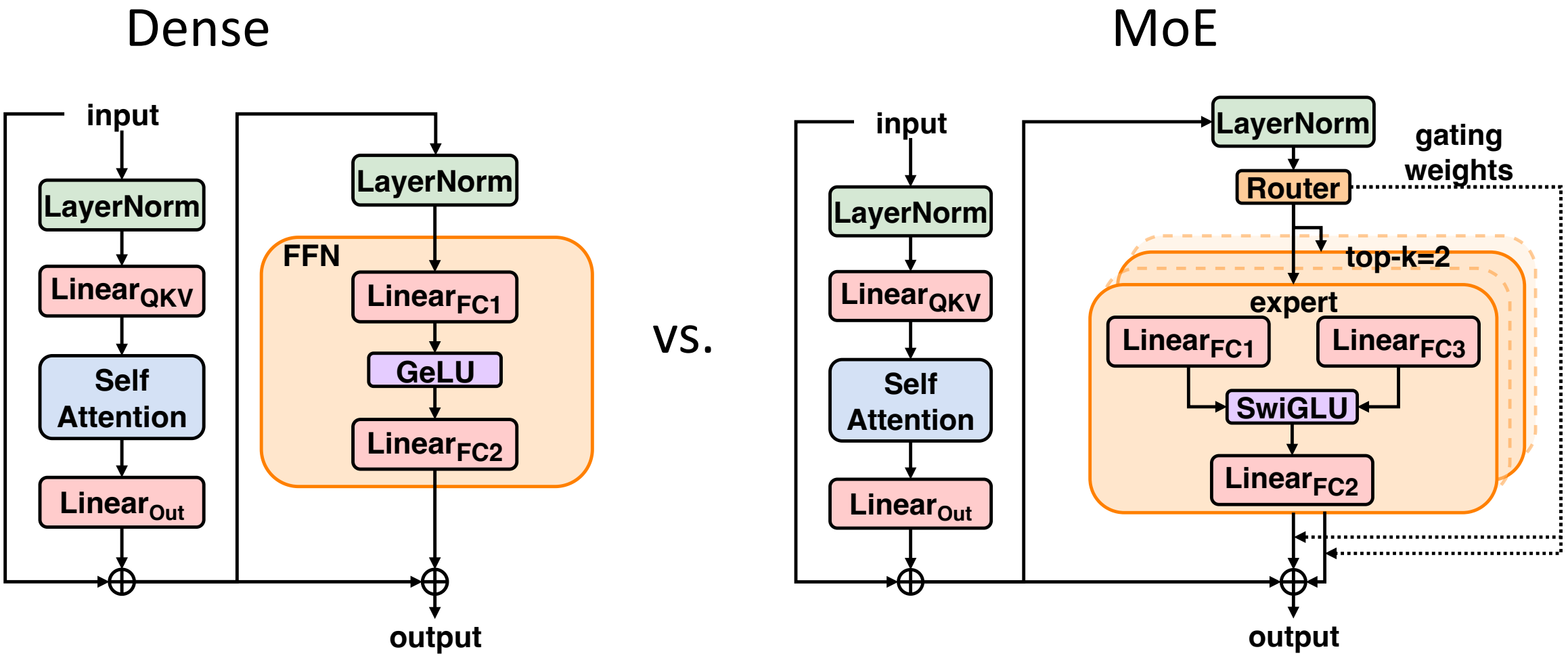
[2] ByteDance | Seed

Scaling for Large Language Models

- LLM Applications
  - conversational assistants
  - code development
  - reasoning
  - tool-augmented agents
  - …

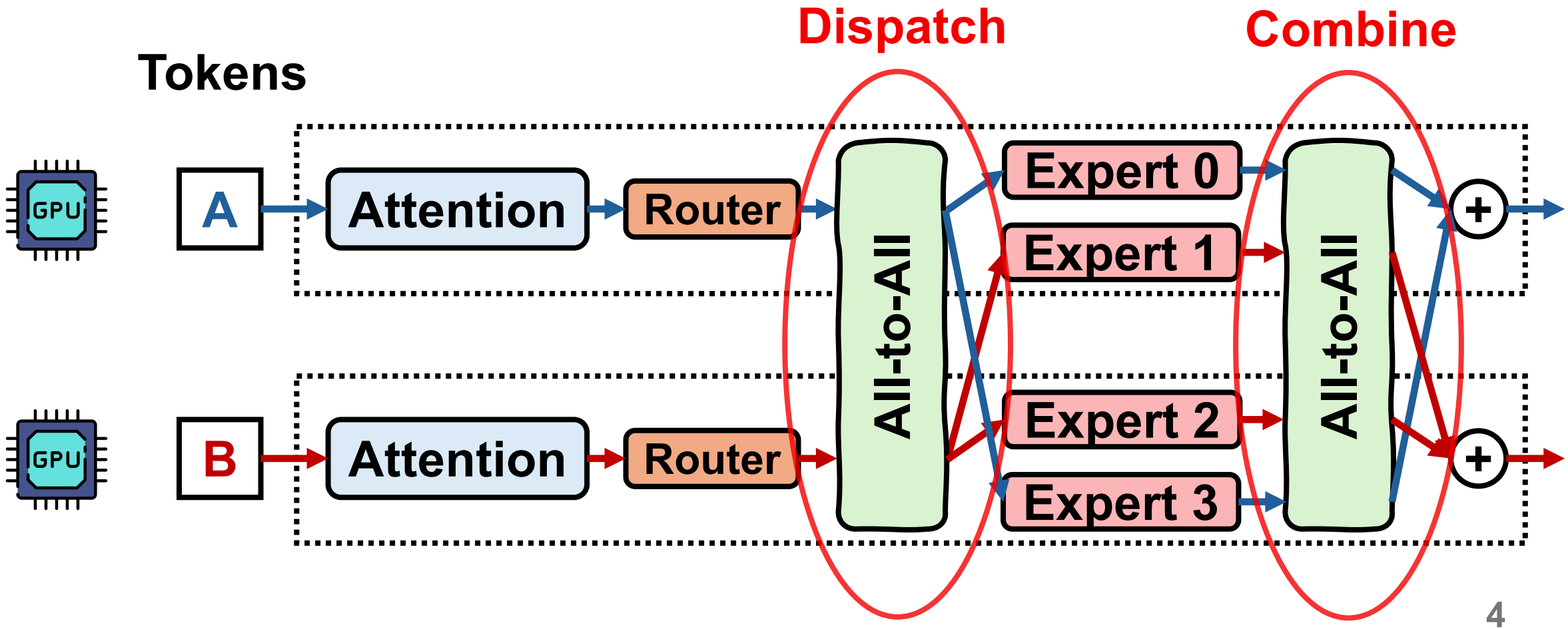


Mixture-of-Experts (MoE)



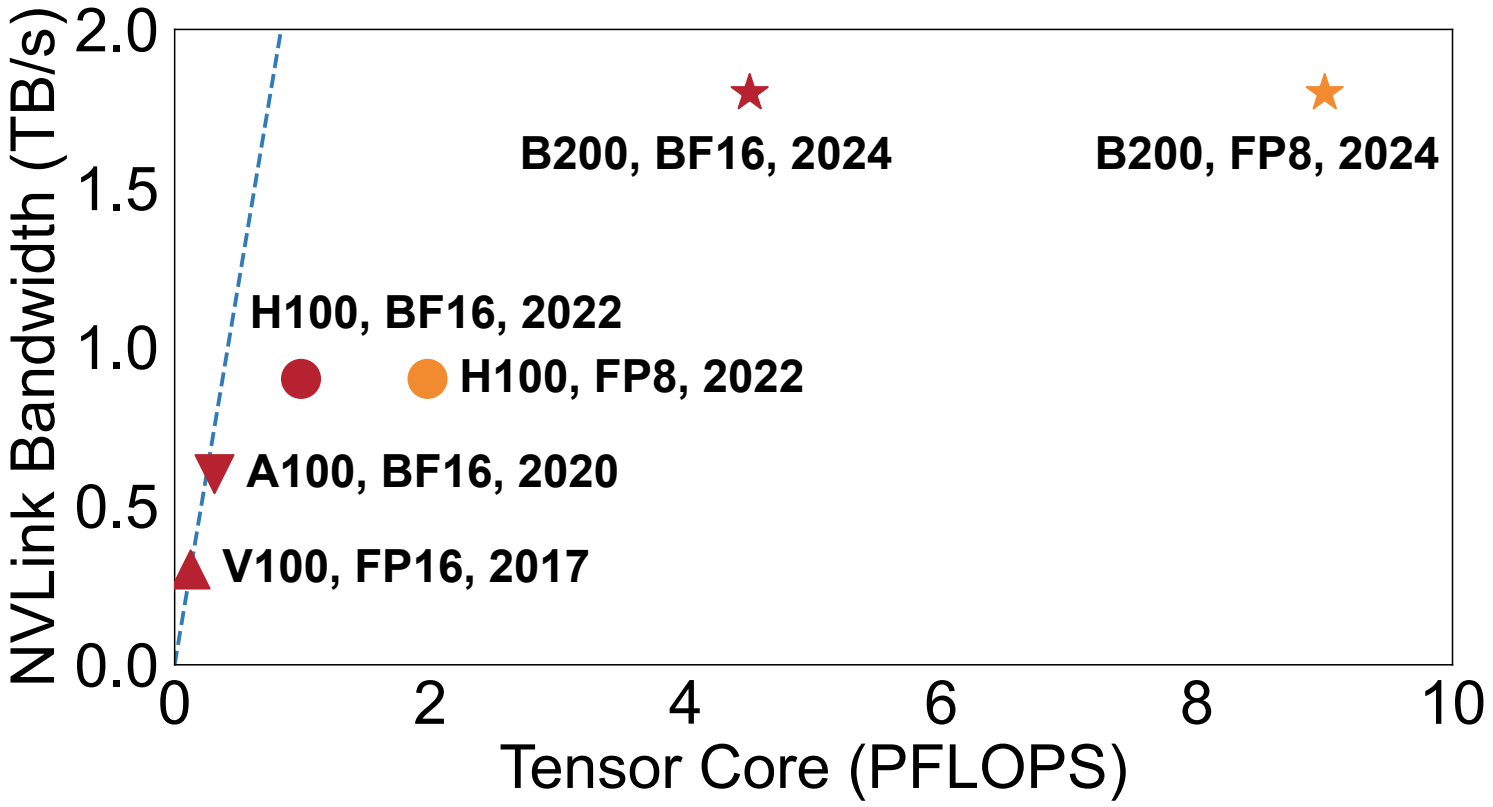
Communication Bottleneck

1. Dispatch & combine communication



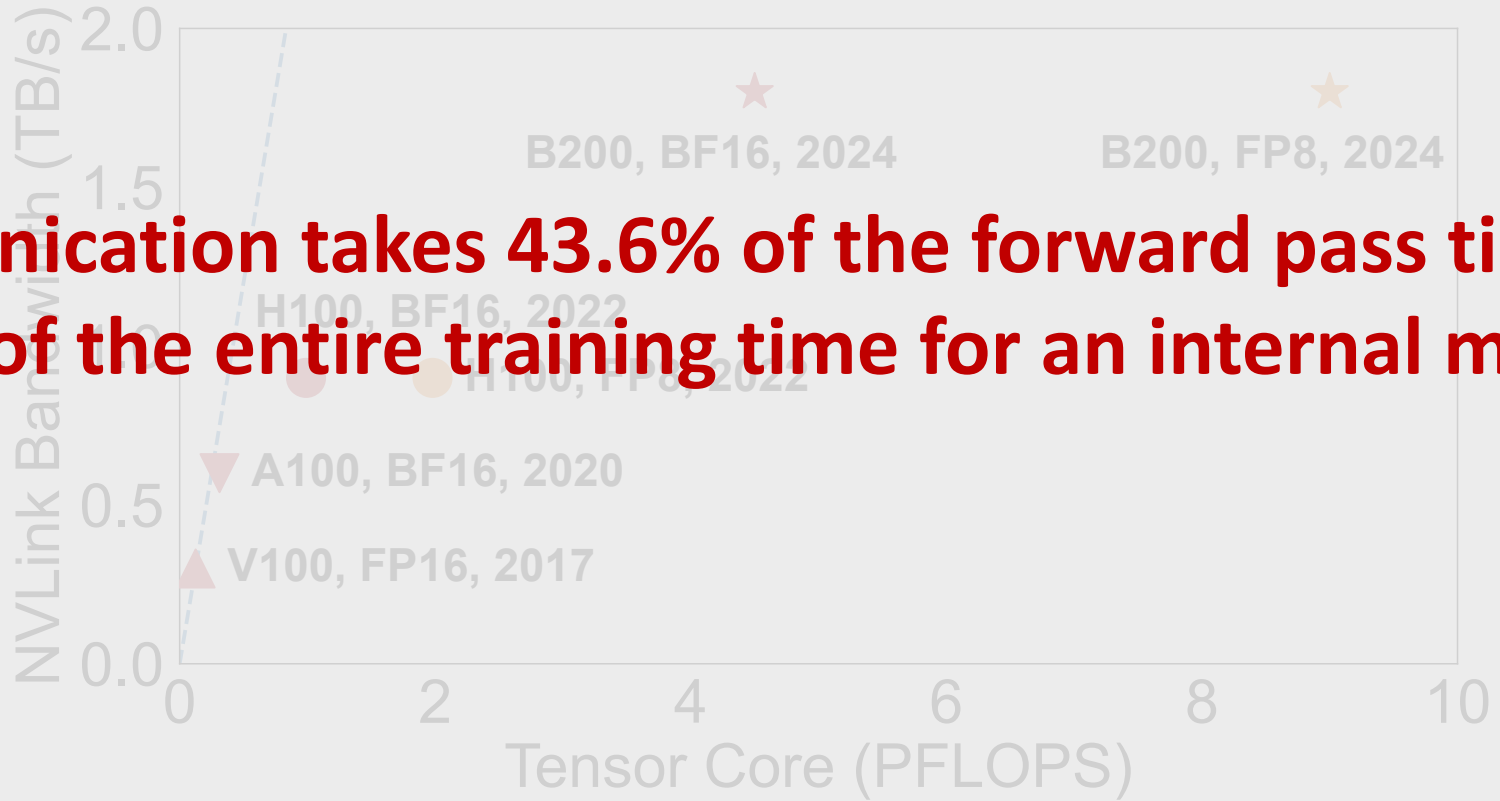
Communication Bottleneck

2. Hardware evolution & low-precision training



Communication Bottleneck

2. Hardware evolution & low-precision training



Communication takes 43.6% of the forward pass time and 32% of the entire training time for an internal model

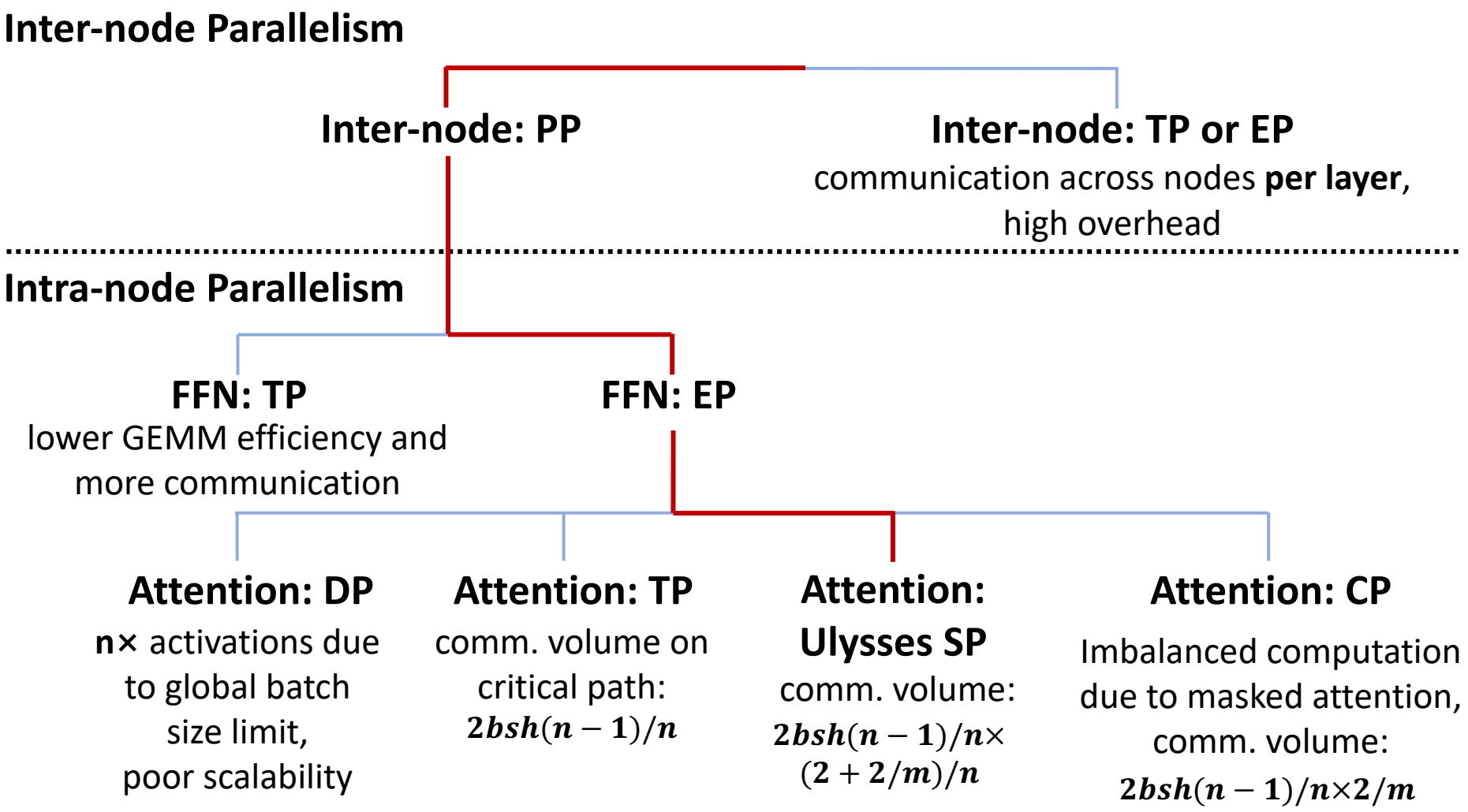
Our solution: MegaScale-MoE

- **Communication-efficient parallelism**
  - **Reduce** communication volume by choosing better parallelism strategies
- **Communication-computation overlap**
  - **Hide** the communication latency through operator scheduling and fusion
- **Communication compression**
  - **Compress** bandwidth cost by lowering communication precision while preserving convergence

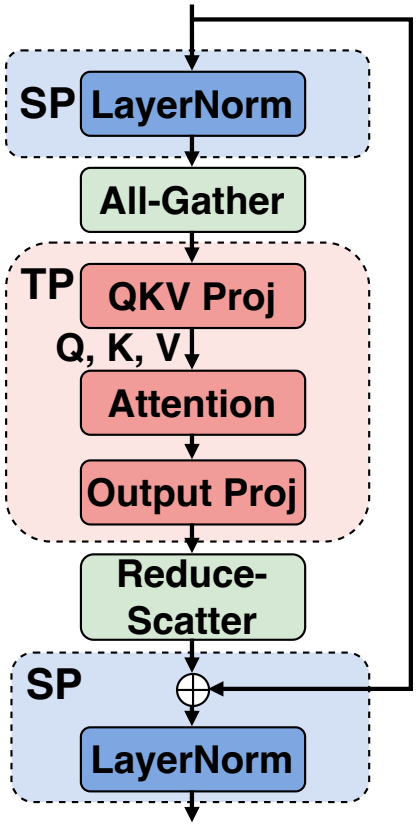
Our solution: MegaScale-MoE

- **Communication-efficient parallelism**
  - **Reduce** communication volume by choosing better parallelism strategies
- **Communication-computation overlap**
  - **Hide** the communication latency through operator scheduling and fusion
- **Communication compression**
  - **Compress** bandwidth cost by lowering communication precision while preserving convergence

Communication-Efficient Parallelism

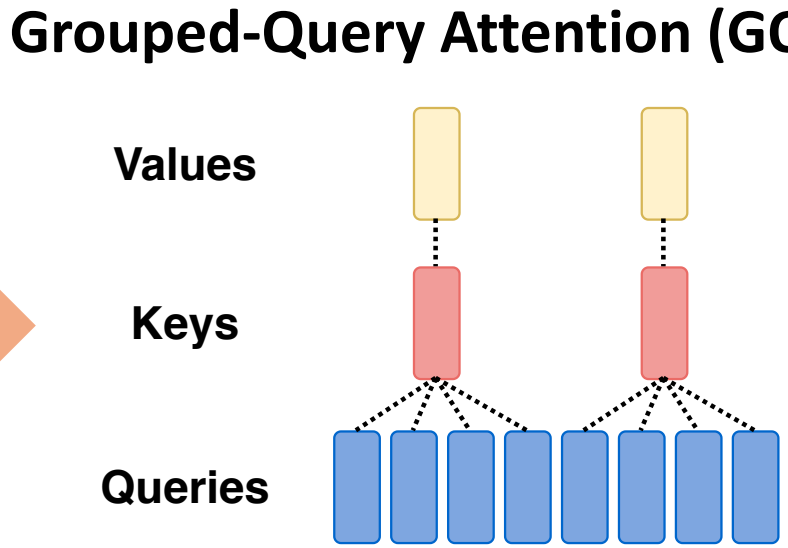


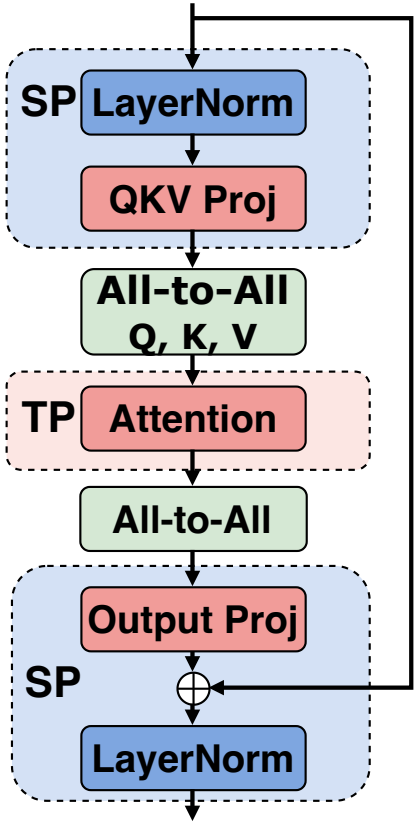
Sequence Parallelism for Attention



(a) Megatron-LM Tensor parallelism.

comm. volume on critical path: $2bsh(n-1)/n$



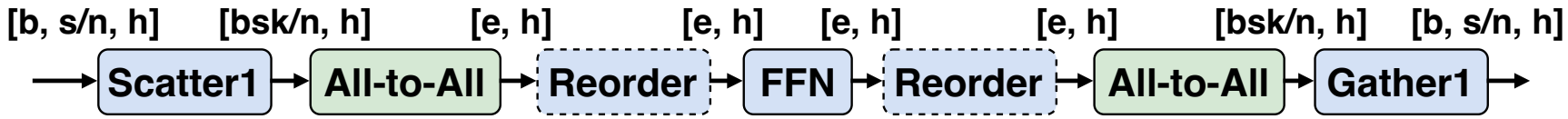


(b) DeepSpeed Ulysses Sequence parallelism.

comm. volume: $2bsh(n-1)/n \times (2+2/m)/n$

Expert Parallelism for FFNs

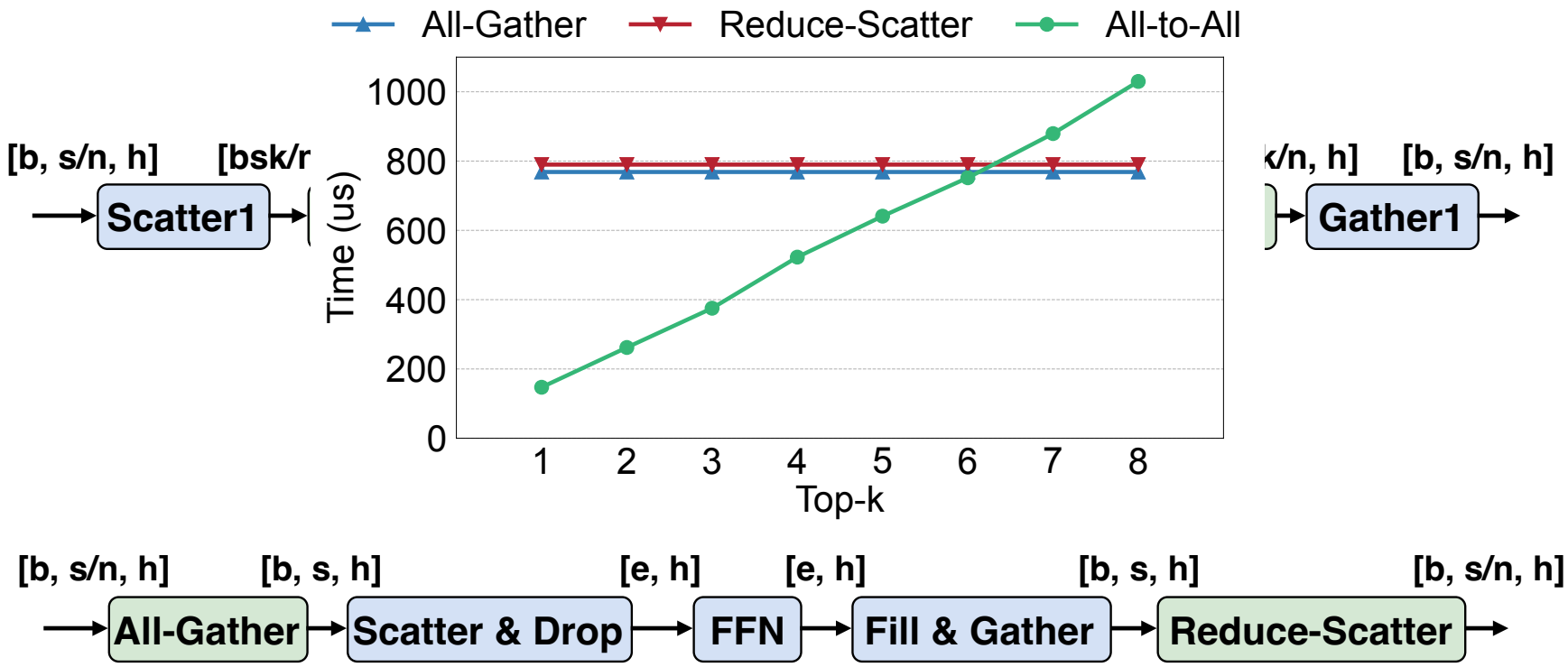
Communication volume: $2k/n \times bsh(n-1)/n$ for EP vs. $2bsh(n-1)/n$ for TP



(a) Typical EP implementation.

Expert Parallelism for FFNs

Communication volume: $2k/n \times bsh(n-1)/n$ for EP vs. $2bsh(n-1)/n$ for TP

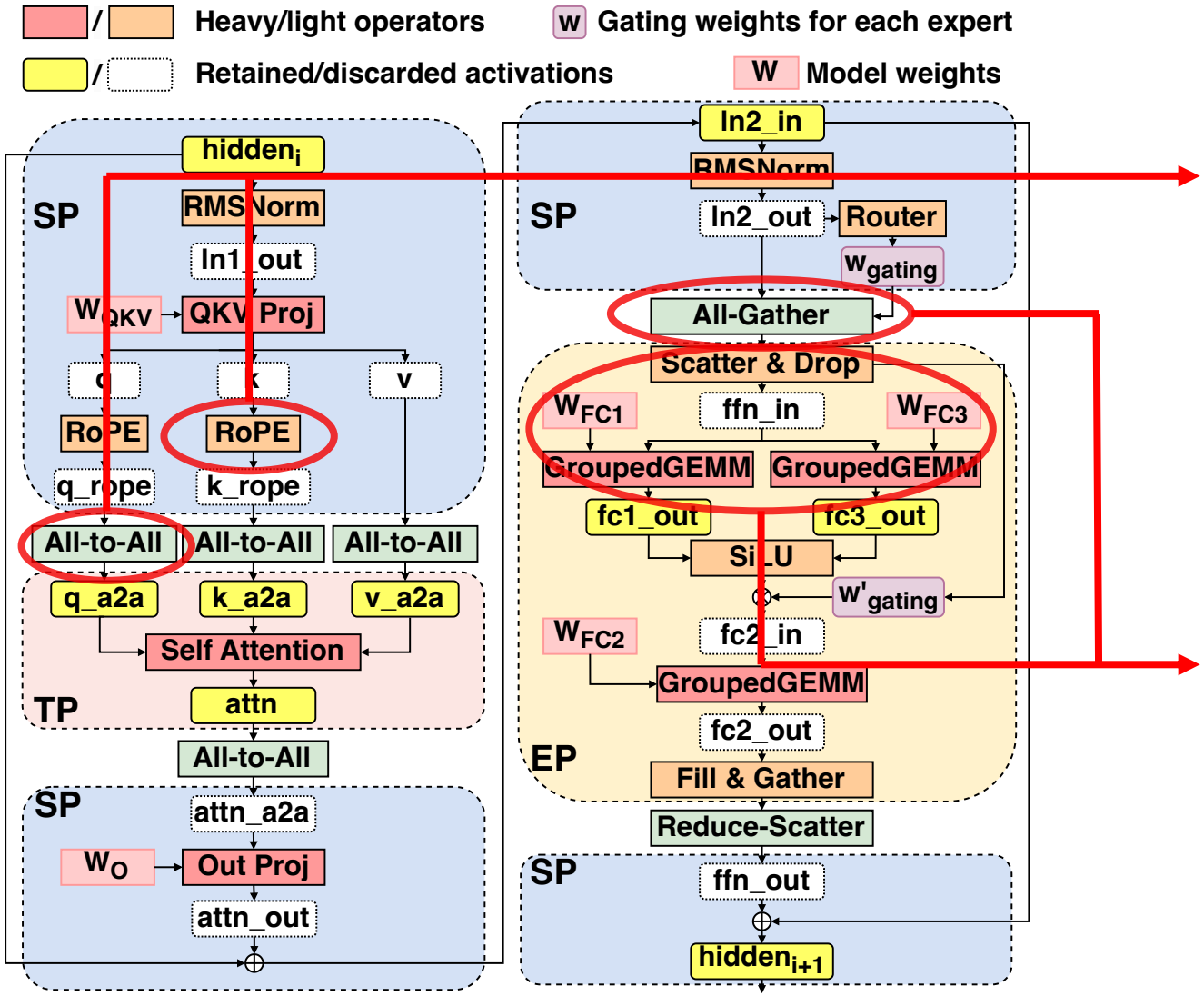


(b) EP Implementation in MegaScale-MoE when top-k > n.

Our solution: MegaScale-MoE

- **Communication-efficient parallelism**
  - **Reduce** communication volume by choosing better parallelism strategies
- **Communication-computation overlap**
  - **Hide** the communication latency through operator scheduling and fusion
- **Communication compression**
  - **Compress** bandwidth cost by lowering communication precision while preserving convergence

Communication-Computation Overlap

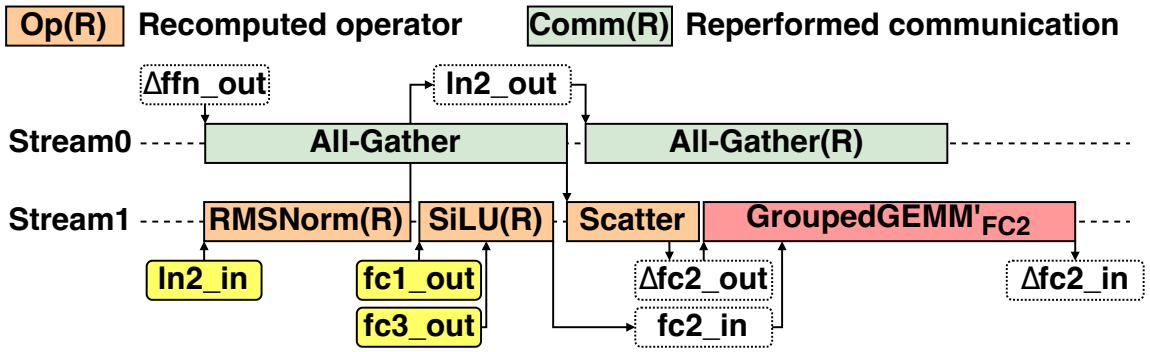


(a) Forward pass of a MoE layer.

- Inter-operator
  - w/o data dependencies
  - reordering/scheduling
- Intra-operator
  - w/ data dependencies
  - fusing

Inter-operator overlap

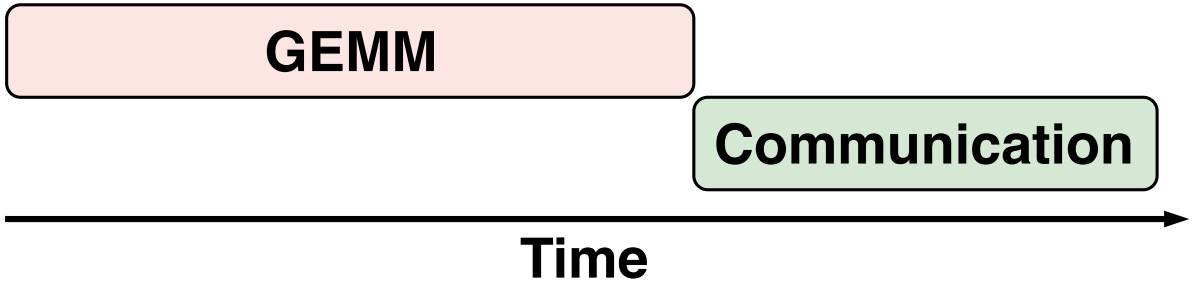
- Holistic Scheduling
  - Fine-grained control over resources to minimize interference
- Selective Activation Rematerialization
  - Minimal activation memory usage with no training performance degradation



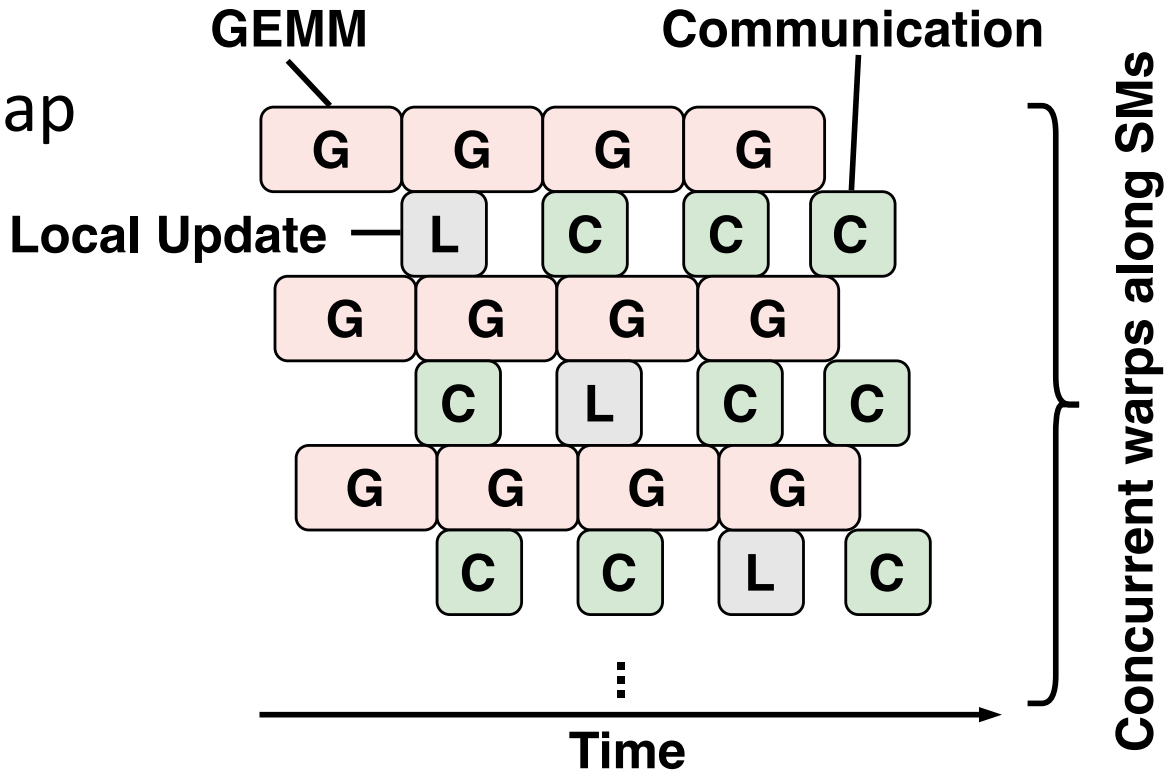
(b) Backward pass snippet with activation rematerialization.

Intra-operator overlap

Without intra-operator overlap



With intra-operator overlap



Our solution: MegaScale-MoE

- **Communication-efficient parallelism**
  - **Reduce** communication volume by choosing better parallelism strategies
- **Communication-computation overlap**
  - **Hide** the communication latency through operator scheduling and fusion
- **Communication compression**
  - **Compress** bandwidth cost by lowering communication precision while preserving convergence

Evaluation

- Training Performance
- Performance Breakdown
- Effectiveness & Additional Overhead of Communication-Efficient Parallelism
- Effectiveness of Inter-Operator Overlap
- Effectiveness of Intra-Operator Overlap
- Effectiveness of Data Parallelism Communication Compression

Training performance

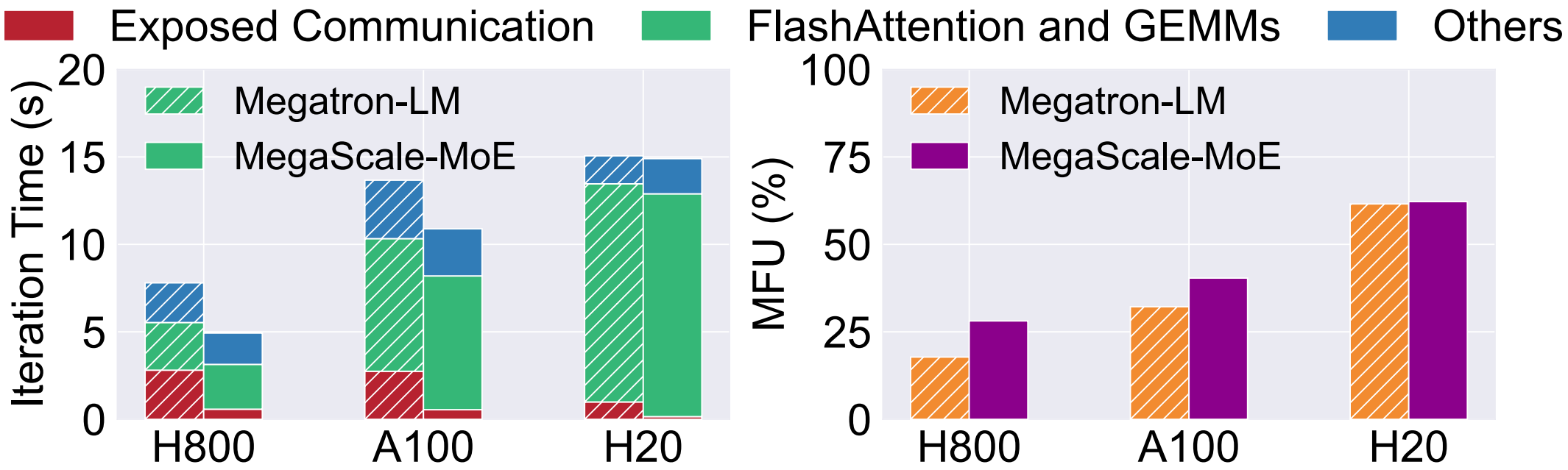
北京大学 PEKING UNIVERSITY | ByteDance | Seed

1.65-1.88x higher training throughput for the 352B MoE model compared to Megatron-LM

System	#GPUs	Iteration Time (s)	Throughput (tokens/s)	Training Time for 1T Tokens (days)
Megatron-LM	240	39.94	151.1k	76.61
	480	19.56	301.1k	38.38
	720	13.70	430.5k	26.88
	960	10.82	550.2k	21.23
	1440	7.90	746.6k	15.50
MegaScale-MoE	240	21.61	272.9k (**1.81**×)	42.41
	480	11.83	498.6k (**1.65**×)	23.21
	720	7.97	740.1k (**1.72**×)	15.64
	960	6.12	963.8k (**1.77**×)	12.01
	1440	4.19	1407.7k (**1.88**×)	8.22

Performance breakdown

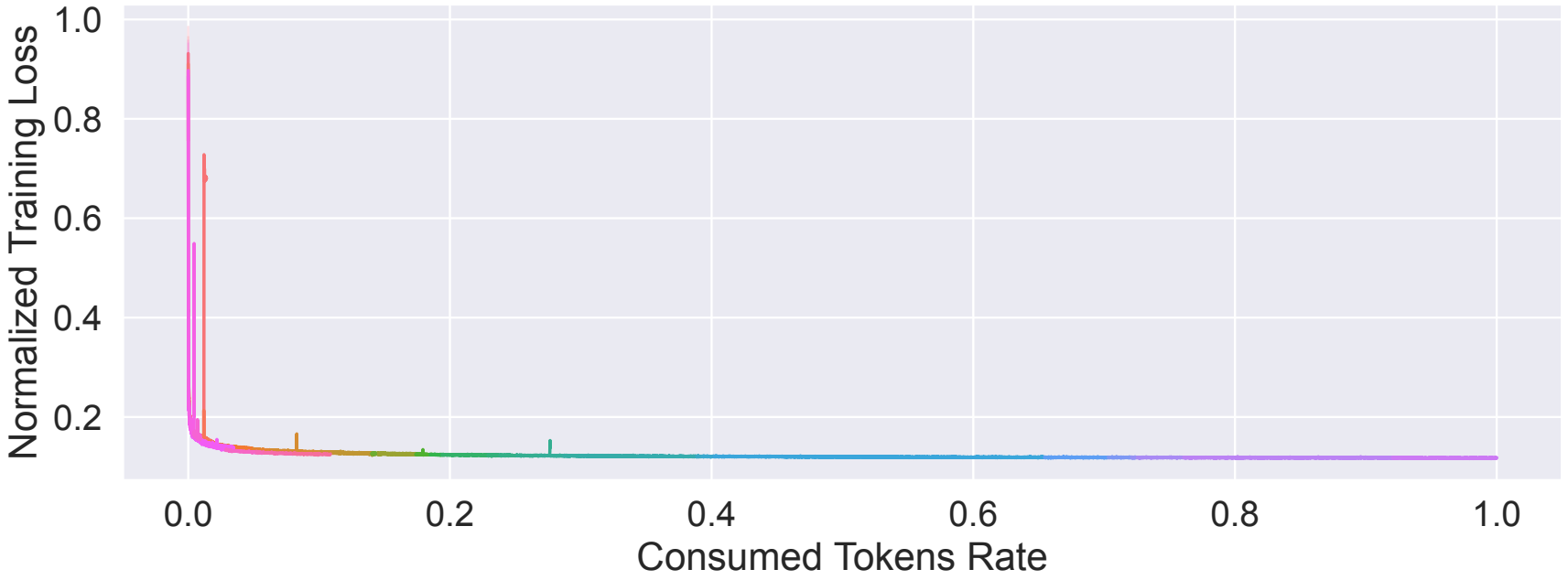
MegaScale-MoE consistently improves MFU by **nearly eliminating exposed communication** across **different GPU platforms**



(a) Iteration time breakdown. **(b) MFU.**

Deployment Experience

- Training loss curve on **>10,000** GPUs for training a 200B-A20B MoE model



- MoE training requires holistic optimization beyond attention and GEMMs, including routing, dispatch/combine, and operators with dynamic shapes.

Conclusion

- Communication is a key bottleneck in large-scale MoE training

- *MegaScale-MoE*: production-grade efficient MoE training system
  - Communication-efficient parallelism
  - Communication-computation overlap
  - Communication compression

- MegaScale-MoE achieves up to 1.88x higher throughput compared to Megatron-LM and is deployed in production

Thank you!


chaojin@pku.edu.cn