Large-scale Deployment of Emerging LLMs

Cheng Wan
cwan@x.ai

August 22, 2025

New Trends in LLM Design

Model	num_kv_heads	num_experts	Release Date
Gemma-1 7B	16	1	June 2024
Llama-3.1 405B	16	1	July 2024
Gemma-2 27B	16	1	August 2024
Qwen-2.5 72B	8	1	September 2024
Phi-4 15B	10	1	December 2024
DeepSeek-3 685B	1 (MLA)	256	December 2024
Llama-4 402B	8	128	April 2025
Qwen-3 235B	4	128	May 2025
GLM-4.5 355B	8	160	July 2025
Kimi-K2 1T	1 (MLA)	384	July 2025
gpt-oss 120B	8	128	August 2025

New Trends in LLM Design

Model	num_kv_heads	num_experts	Release Date
Gemma-1 7B	16	1	June 2024
Llama-3.1 405B	16	1	July 2024
Gemma-2 27B	16	1	August 2024
Qwen-2.5 72B	8	1	September 2024
Phi-4 15B	10	1	December 2024
DeepSeek-3 685B	1 (MLA)	256	December 2024
Llama-4 402B	8	128	April 2025
Qwen-3 235B	4	128	May 2025
GLM-4.5 355B	8	160	July 2025
Kimi-K2 1T	1 (MLA)	384	July 2025
gpt-oss 120B	8	128	August 2025

- **Trends in Model Architecture**
  - Less heads for KV cache
  - Larger expert size

New Trends in LLM Design

Model	num_kv_heads	num_experts	Release Date
Gemma-1 7B	16	1	June 2024
Llama-3.1 405B	16	1	July 2024
Gemma-2 27B	16	1	August 2024
Qwen-2.5 72B	8	1	September 2024
Phi-4 15B	10	1	December 2024
DeepSeek-3 685B	1 (MLA)	256	December 2024
Llama-4 402B	8	128	April 2025
Qwen-3 235B	4	128	May 2025
GLM-4.5 355B	8	160	July 2025
Kimi-K2 1T	1 (MLA)	384	July 2025
gpt-oss 120B	8	128	August 2025

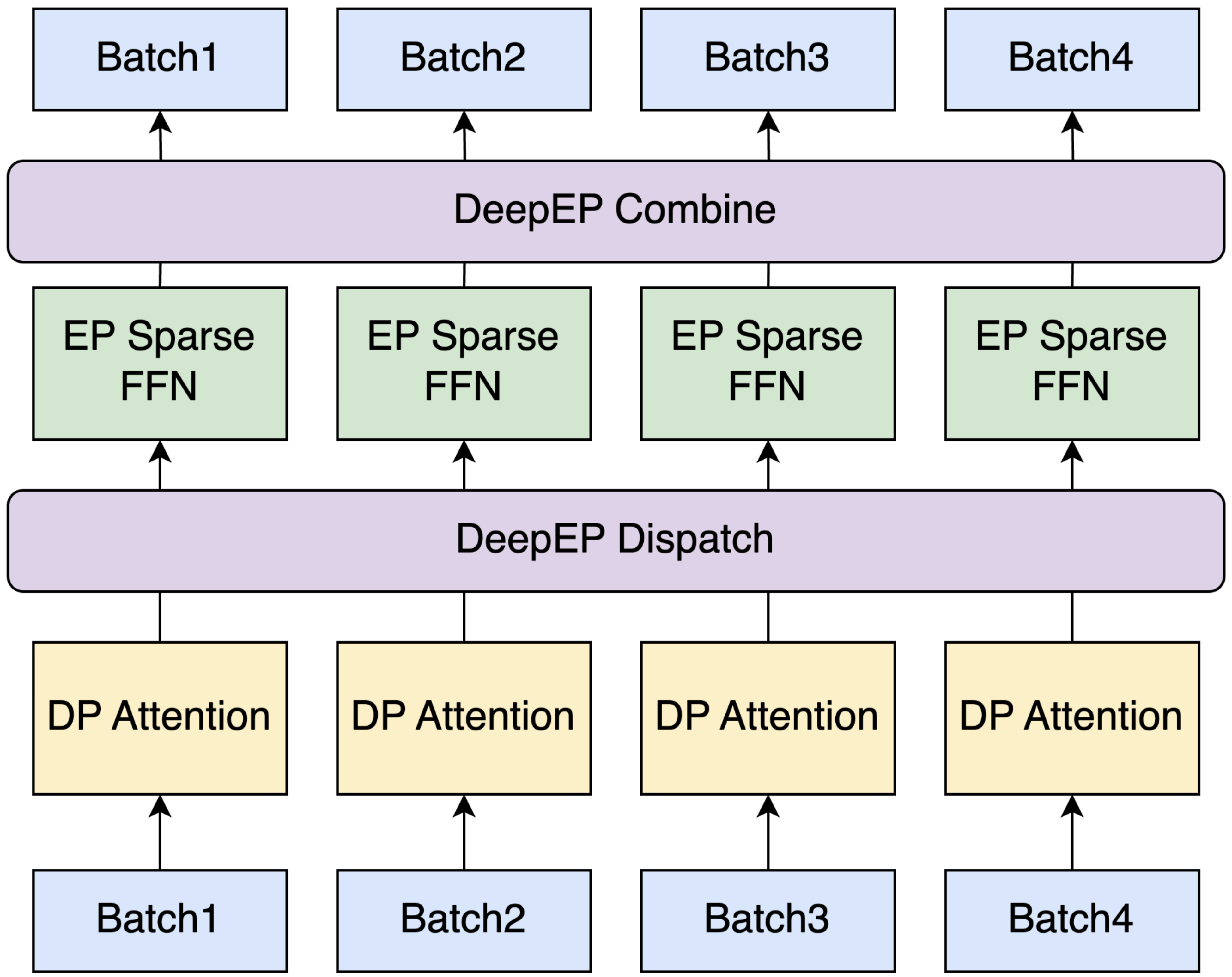
- **Trends in Model Architecture**
  - Less heads for KV cache
  - Larger expert size
- **Challenges in Large-scale Deployment**
  - Large TP for attention is not scalable: KV cache duplication
  - Large TP for FFN is expensive: scaled-out communication volume for all-reduce

New Trends in LLM Design

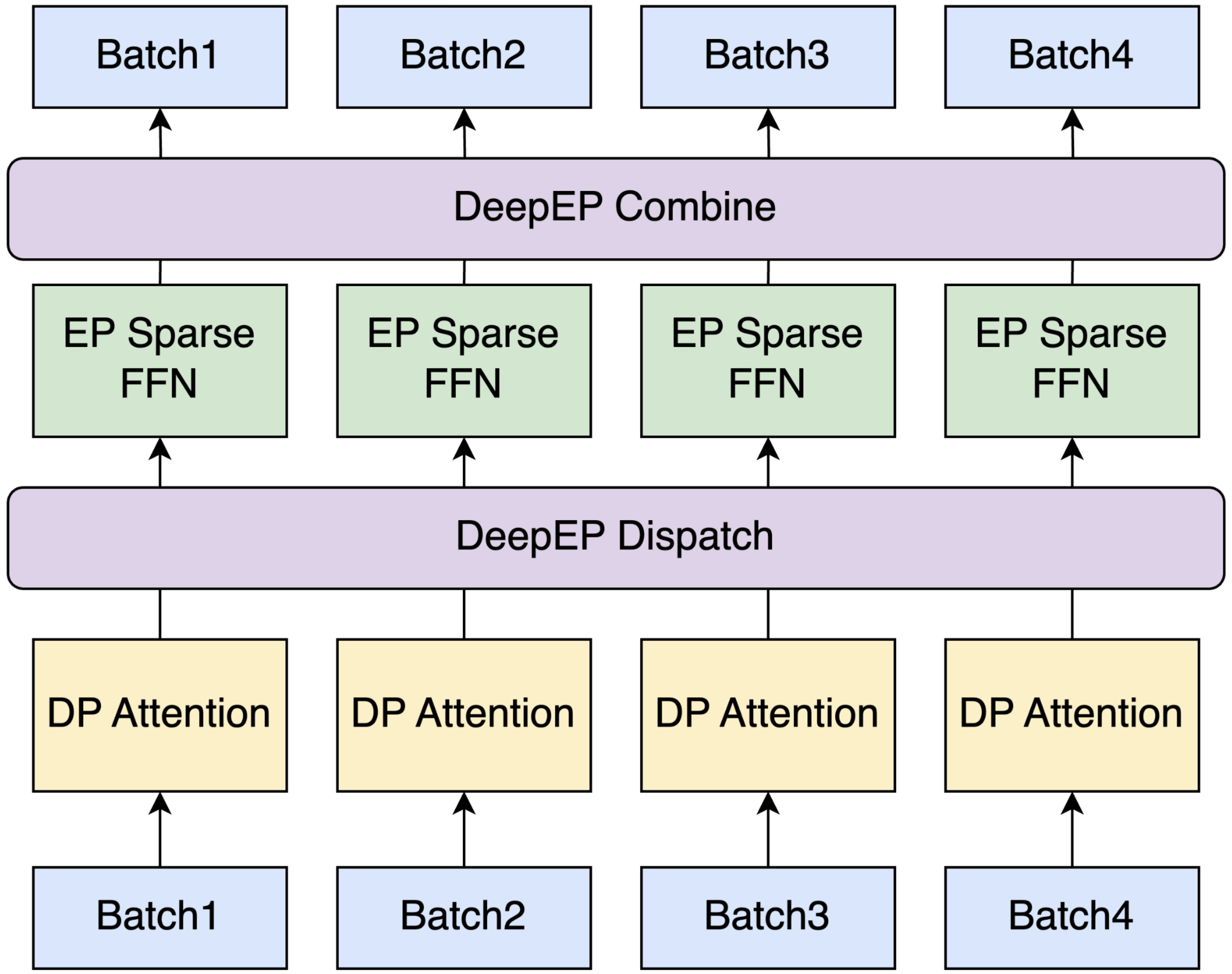
Model	num_kv_heads	num_experts	Release Date
Gemma-1 7B	16	1	June 2024
Llama-3.1 405B	16	1	July 2024
Gemma-2 27B	16	1	August 2024
Qwen-2.5 72B	8	1	September 2024
Phi-4 15B	10	1	December 2024
DeepSeek-3 685B	1 (MLA)	256	December 2024
Llama-4 402B	8	128	April 2025
Qwen-3 235B	4	128	May 2025
GLM-4.5 355B	8	160	July 2025
Kimi-K2 1T	1 (MLA)	384	July 2025
gpt-oss 120B	8	128	August 2025

- **Trends in Model Architecture**
  - Less heads for KV cache
  - Larger expert size
- **Challenges in Large-scale Deployment**
  - Large TP for attention is not scalable: KV cache duplication
  - Large TP for FFN is expensive: scaled-out communication volume for all-reduce
- **New System Design**
  - DP for attention layers
  - EP for FFN layers

DP Attention + EP FFN

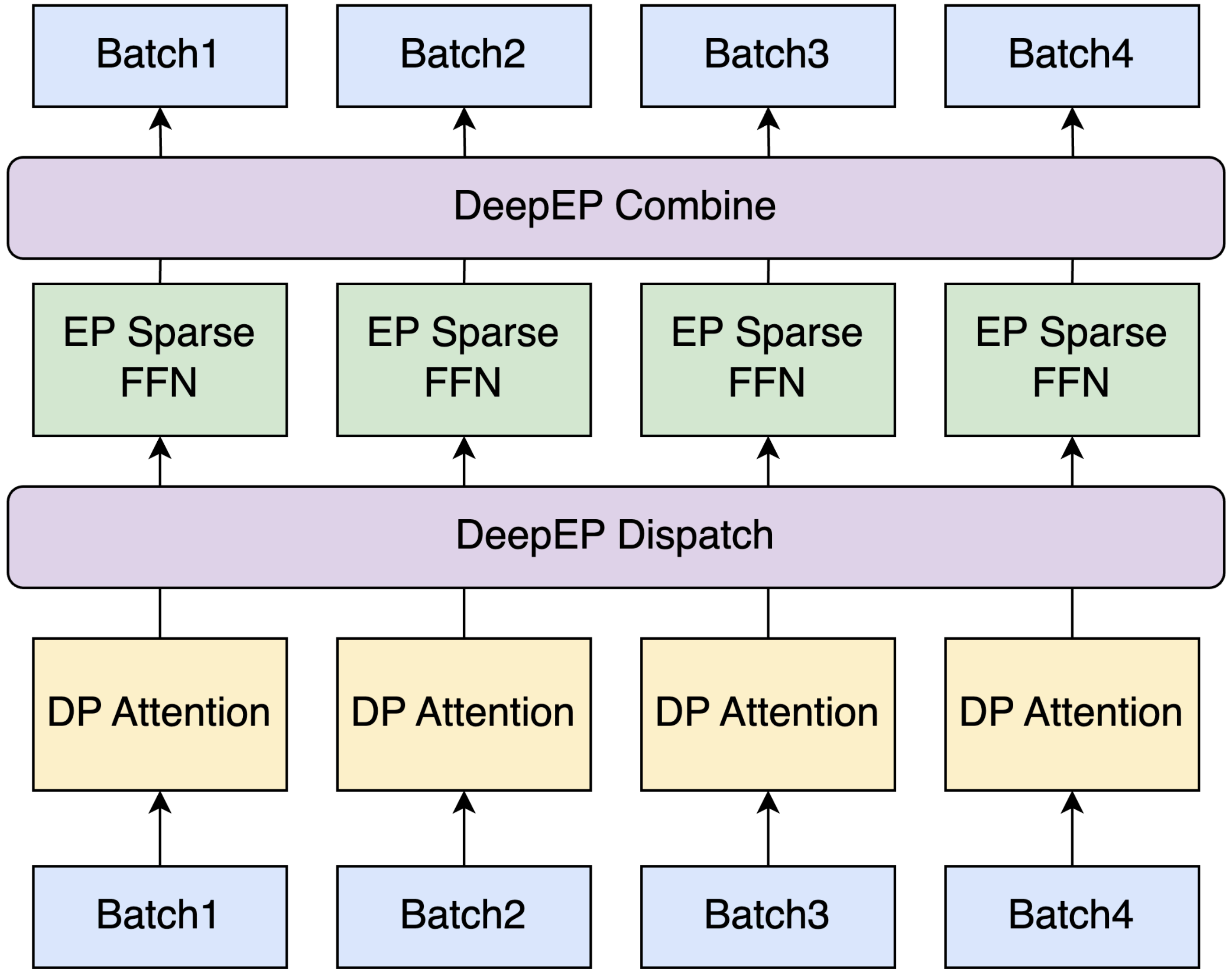


DP Attention + EP FFN



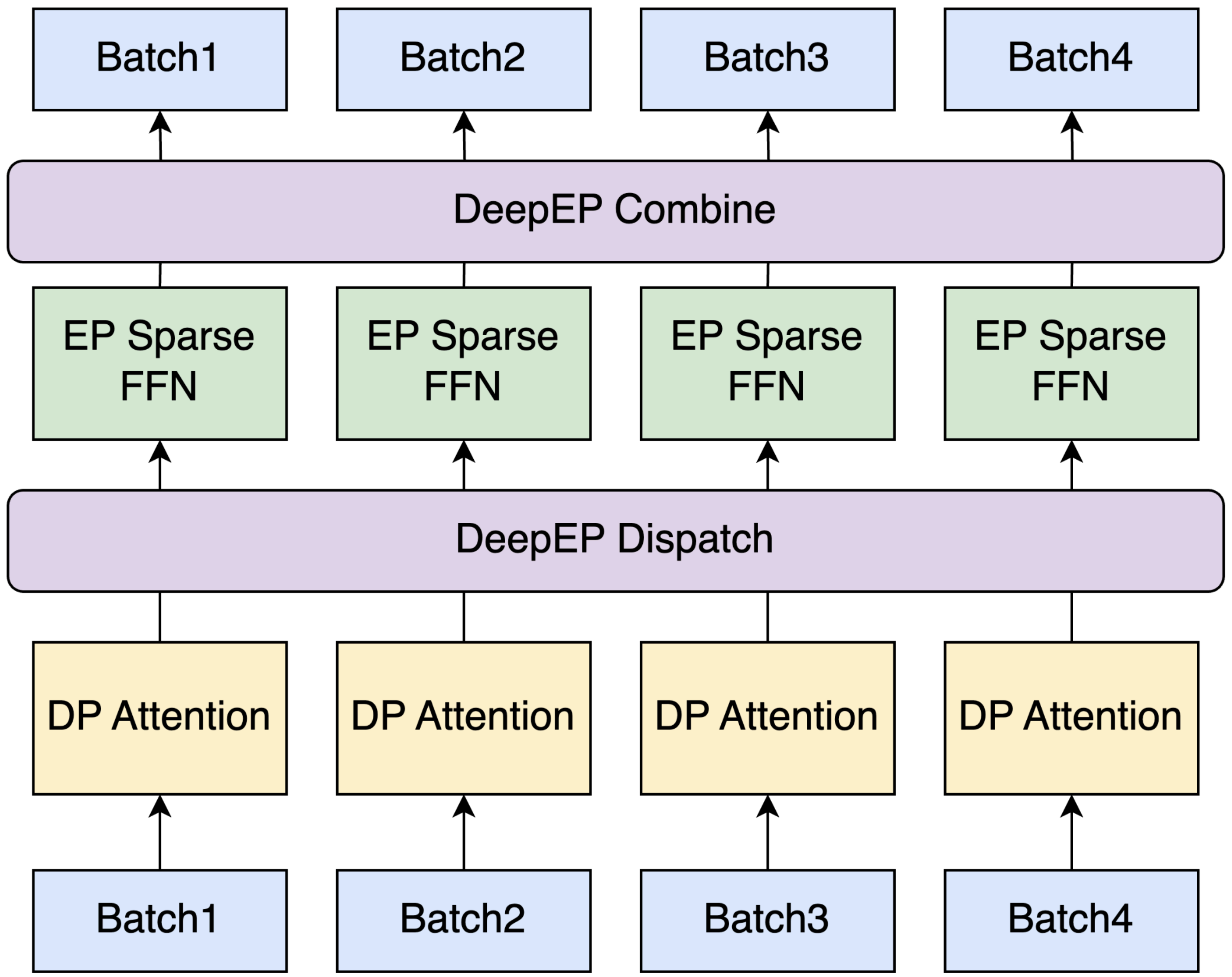
- **Scalable KV Cache:** DP attention avoids duplication for KV cache.

DP Attention + EP FFN



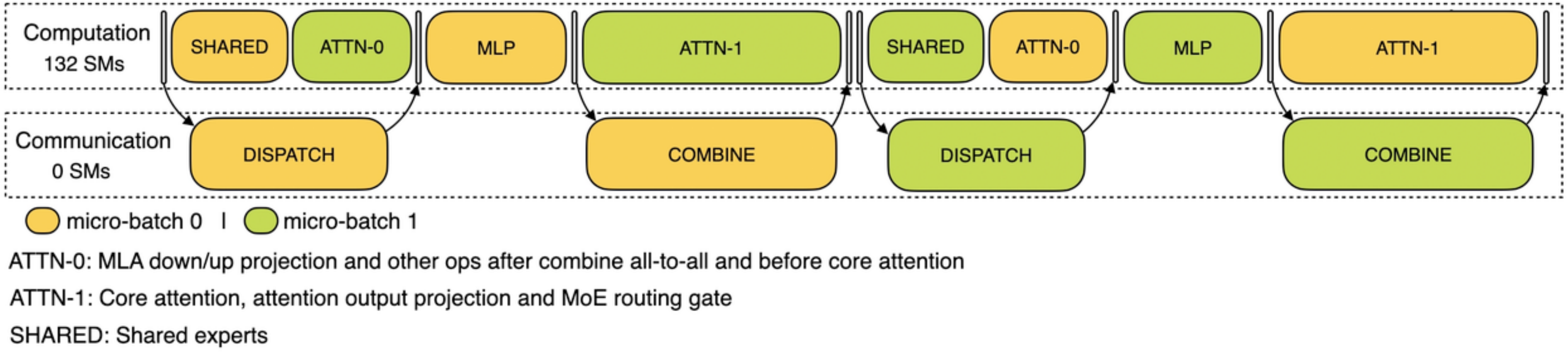
- **Scalable KV Cache:** DP attention avoids duplication for KV cache.
- **Scalable Model Capacity:** Expert weights are partitioned across devices using Expert Parallelism, removing memory bottlenecks.

DP Attention + EP FFN



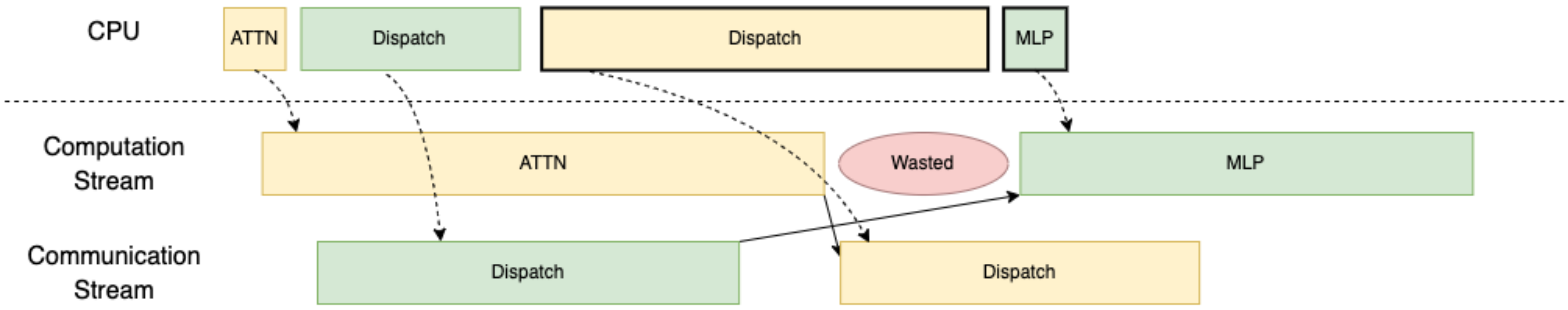
- **Scalable KV Cache:** DP attention avoids duplication for KV cache.
- **Scalable Model Capacity:** Expert weights are partitioned across devices using Expert Parallelism, removing memory bottlenecks.
- **Optimized Communication:** Follows a Dispatch → Expert → Combine pattern; powered by DeepEP and **Two-Batch-Overlap** to minimize latency and overhead.

Two-batch Overlap (TBO)



- **Two Batch Overlap (TBO)**: Executing communication and computation simultaneously.

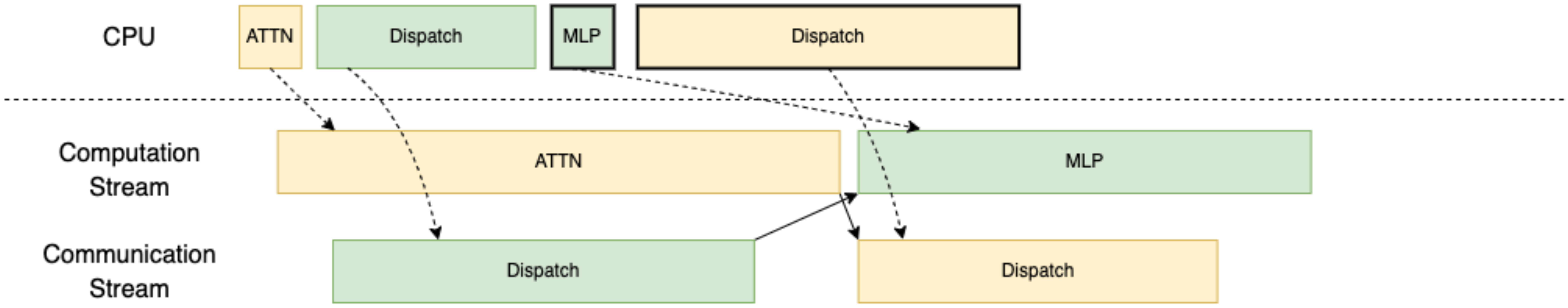
Improper Launch Order of TBO



(a) Two-batch overlap with an improper launch order

- **Two Batch Overlap (TBO)**: Executing communication and computation simultaneously.
- Dispatch brings **synchronization**, which blocks the CPU until the GPU receives metadata (required for allocating correctly sized tensors).
- **Improper launch order**, e.g. dispatch before MLP, will block the launching and leave the computation stream idle.

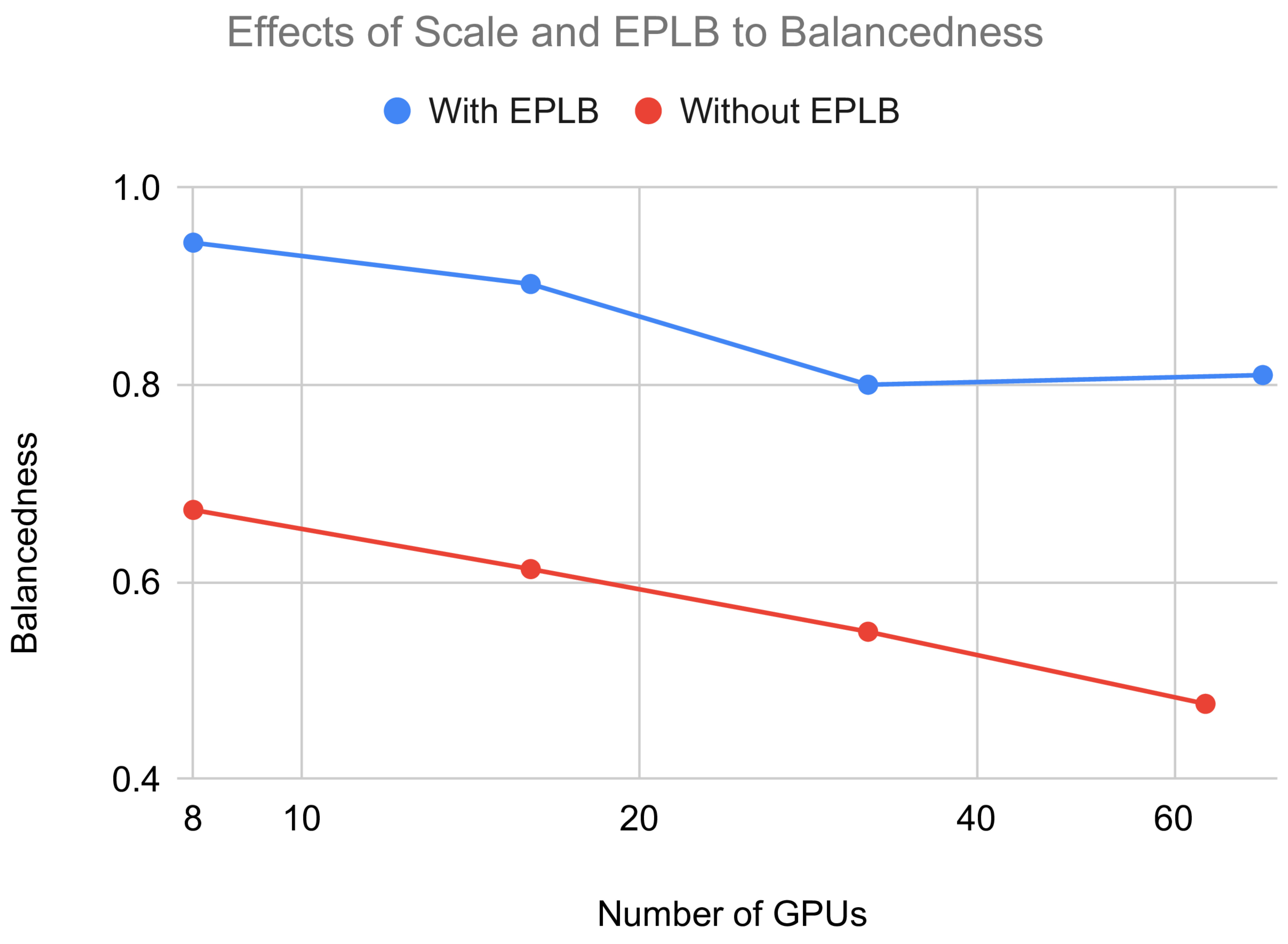
Proper Launch Order of TBO



(b) Two-batch overlap with a proper launch order

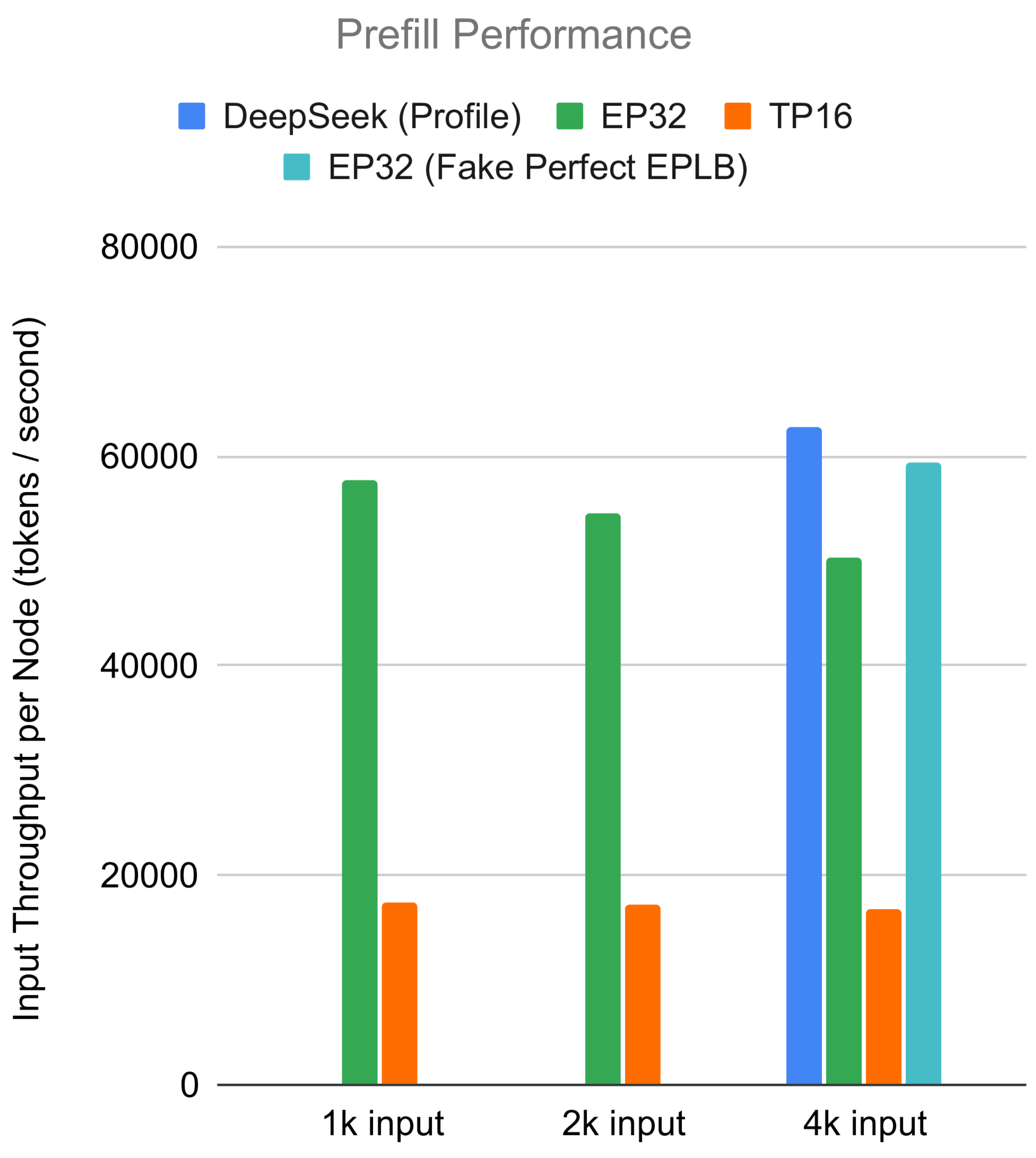
- Proper launch order: submitting computation tasks to the GPU **before launching CPU-blocking communication.**
- Computation → Communication: enabling GPU to remain active during communication.

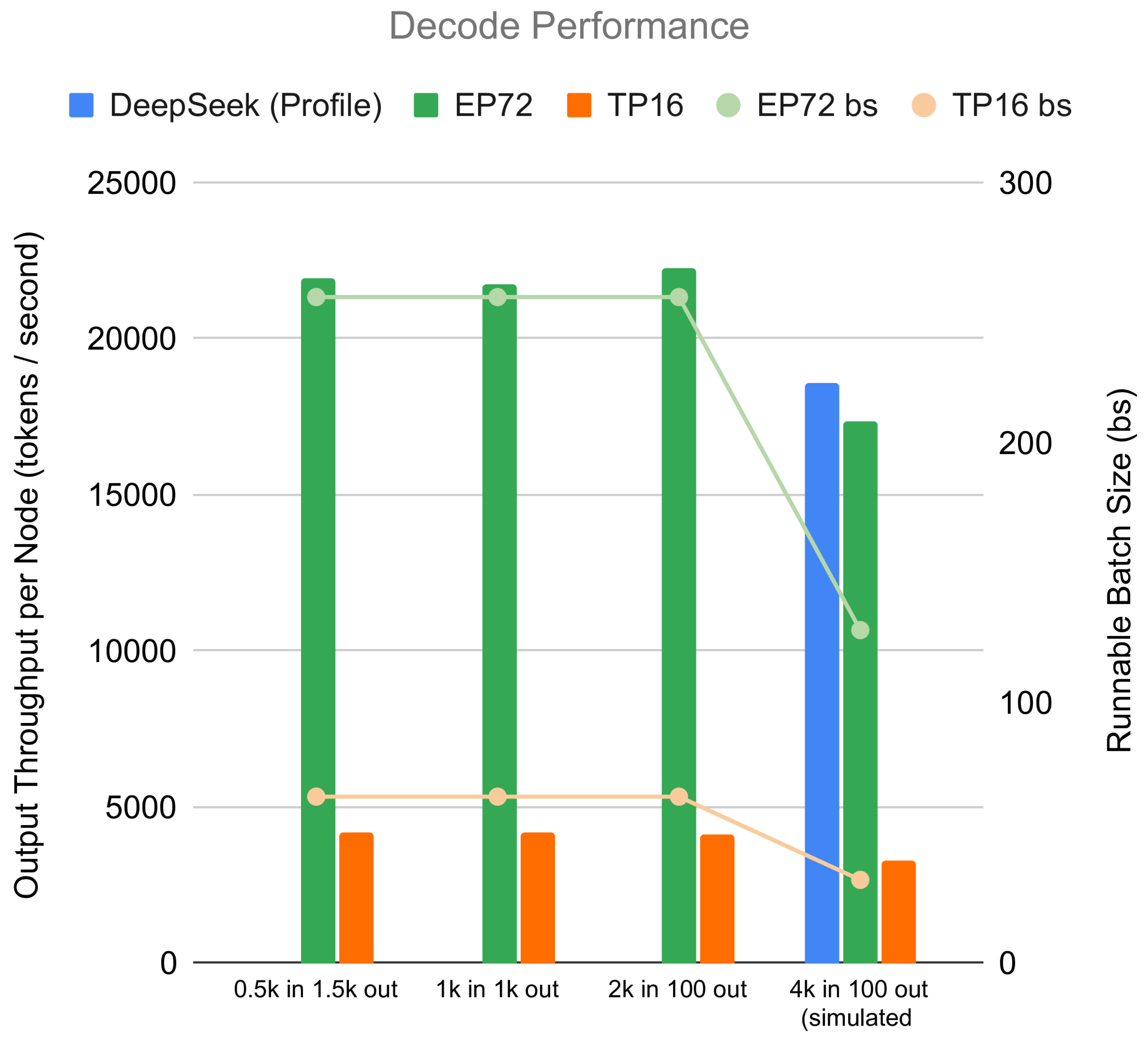
EPLB for Balancing Workload



- **Balancedness:** the ratio between mean computation time and maximum computation time for a MoE layer among GPUs.
- Balancedness decreases when the system scales with the number of nodes.
- Enabling EPLB significantly improves the balancedness.

Throughput Performance





Throughputs of prefill (P) and decode (D) phases are evaluated independently, assuming unlimited resources for the non-tested phase to isolate and maximize the load on the tested nodes—mirroring the setup used by DeepSeek.

Deployment of DeepSeek R1 and Kimi K2

Model	Experts	GPUs	Prefill Throughput (tokens/sec)	Decode Throughput (tokens/sec)	Cost per 1M Output Tokens
DeepSeek R1	256	96 × H100	52.3k / node	22.3k / node	$0.20
Kimi K2	384	128 × H200	56k / node	24.0k / node	$0.21

- With large-scale DP attention and EP FFN, trillion-scale models achieve 20+k output throughput per node.
- Cost: ~$0.20 per 1M tokens

Q & A