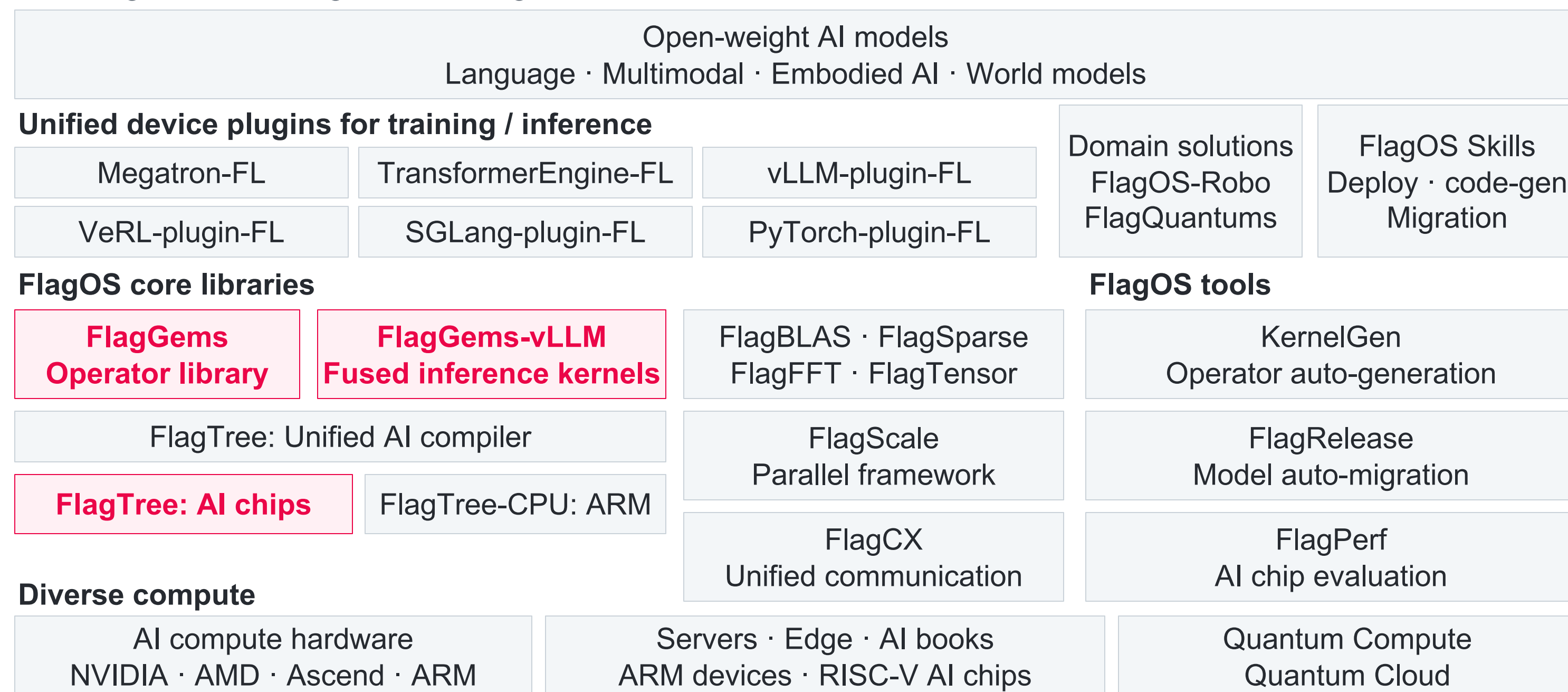


Hang Xiao, Chunlei Men, AI Compiler Researcher

FlagOS: Open-sourced System SW for Multi-Vendor AI chipsets

URL: flagos.io, GitHub: github.com/flagos-ai



2 Layout conversion elimination

Layout conversions can add shared-memory traffic and synchronization.

Four ways to eliminate redundant conversions

- 01 Cost-based layout selection 02 Local backward propagation
03 Small-component solving 04 Write-back rematerialization

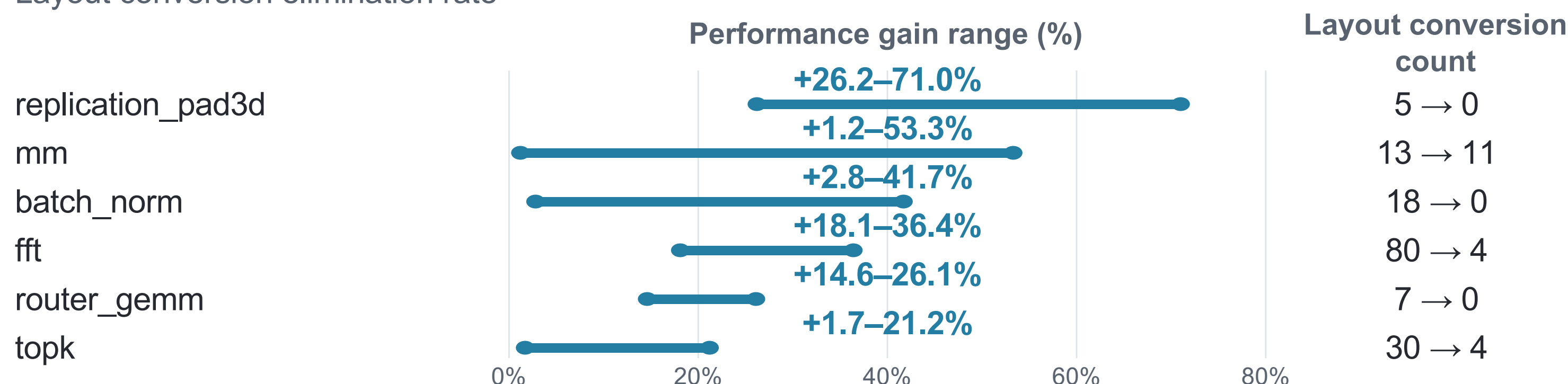


Performance gain (representative operators)

68–79%

Layout conversion elimination rate

100+ operators

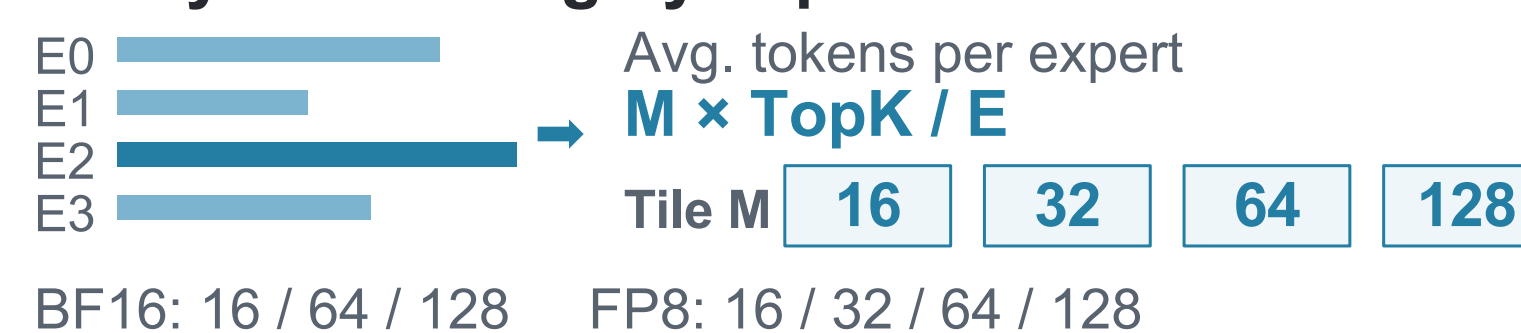


4 End-to-end inference performance

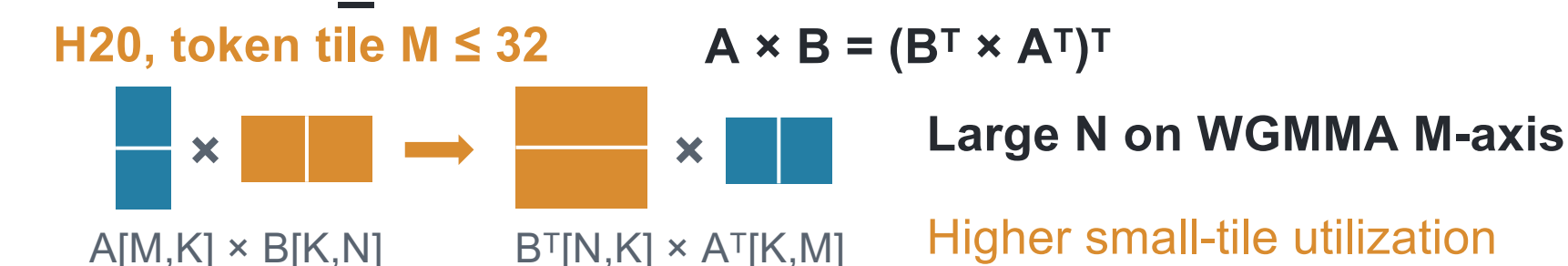
NVIDIA H20		Throughput	TTFT	ITL
DeepSeek V4 Flash	TP=4	+12.63%	-20.22%	-4.19%
fused_marlin_moe	TP=8	+7.65%	-12.87%	-2.99%

1 Fused MoE optimization

01 Dynamic tiling by expert load



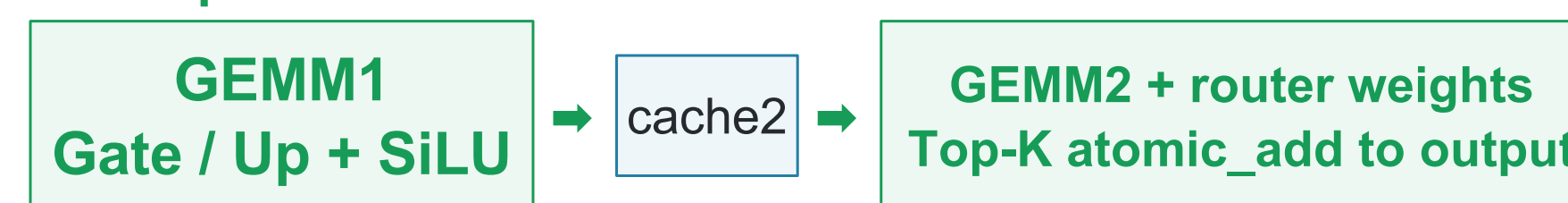
03 SWAP_AB for small FP8 tiles



02 Fused GEMM dataflow



Fused path



Fewer launches and intermediate writes

Fused MoE speedup

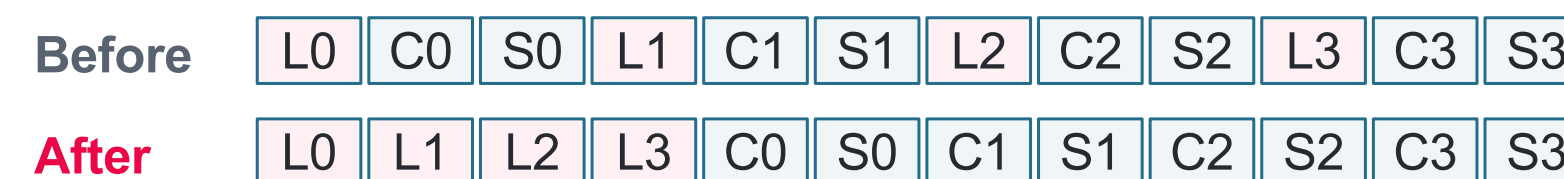
Backend	Mean	Range	Baseline
NVIDIA	1.17×	0.91–1.36×	vLLM
MetaX	1.14×	0.86–1.62×	vLLM
Hygon	1.08×	0.95–1.26×	vLLM
Moore Threads	2.60×	1.05–3.50×	vLLM
T-Head	0.99×	0.75–1.20×	vLLM
Ascend	0.776×	0.277–1.021×	Torch

3 Instruction scheduling

A. Load fronting after loop unrolling

Move independent loads and pointer / mask dependencies earlier. Increase load concurrency to hide memory latency and reduce stalls.

```
tile.range(..., loop_unroll_factor=4, reorder=True)
```



Kernel speedup after unroll + compiler reorder

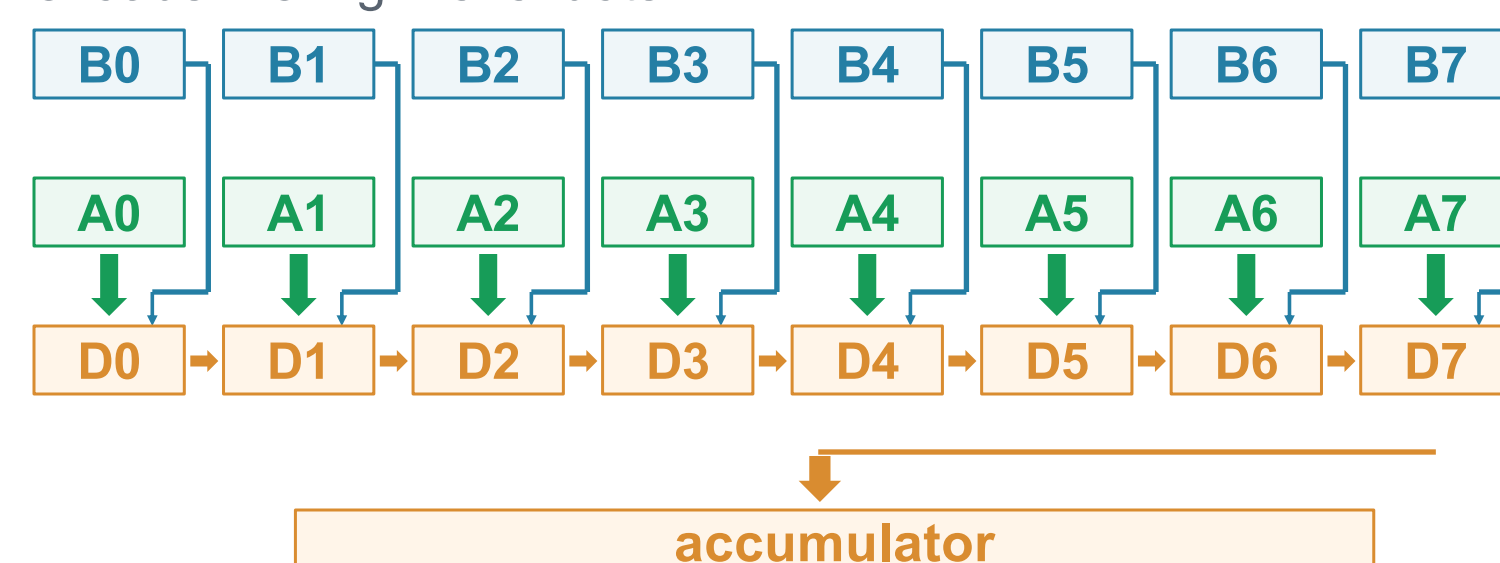
Kernel	Mean	Range
per_token_group_fp8_quant	1.420×	1.040–2.002×
fused_inv_rope_fp8_quant	1.186×	1.142–1.228×
cp_gather_indexer_k_quant_cache	1.609×	1.182–2.270×

B. Dot-fragment merging

Merge dot fragments in registers to reuse activations, reduce address calculations, and avoid shared-memory staging.

Before: 8 × K16

8 loads + 8 high-level dots



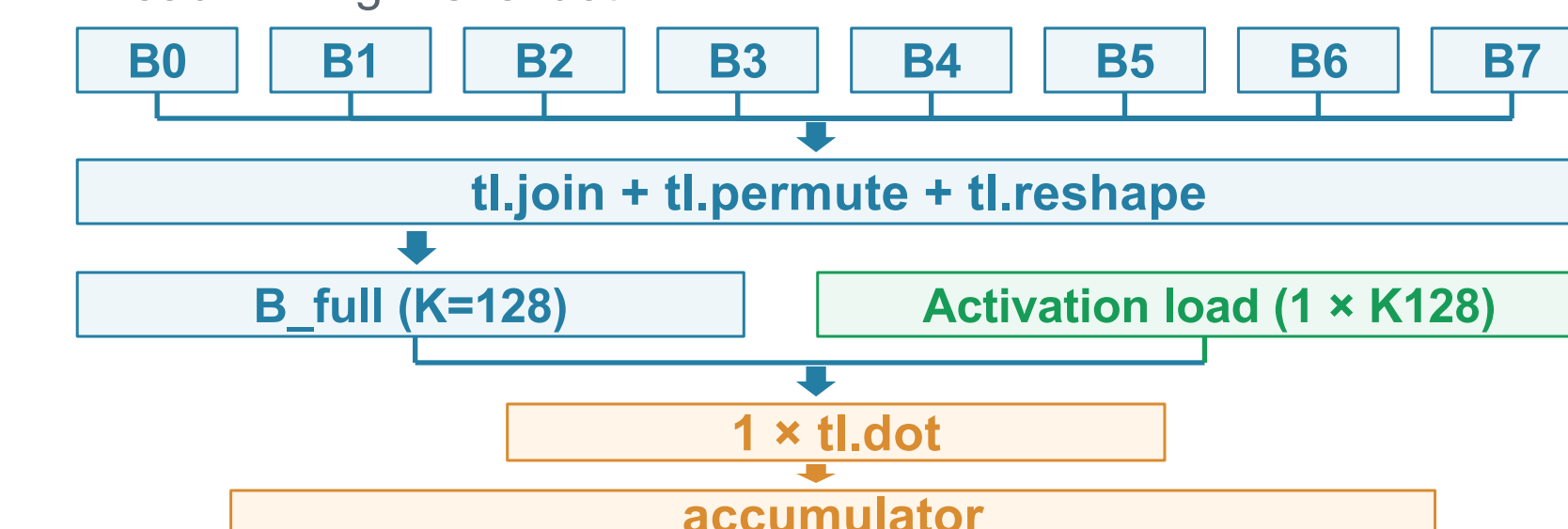
Fused_marlin_moe FP4 performance

DeepSeek-V4-Flash, P32K / D1K. Baseline: vLLM CUDA.

1.168× Mean 1.208× Count-weighted 1.814× Peak

After: full K128

1 load + 1 high-level dot



Higher throughput is better; lower TTFT and ITL are better.

MetaX C550		Throughput	TTFT	ITL
Qwen3.6-35B-A3B	TP=2	+12.85%	-13.08%	-5.55%
fused_moe + GDN + mul				