Kareus: Joint Reduction of Dynamic and Static Energy in Large Model Training

Ruofan Wu

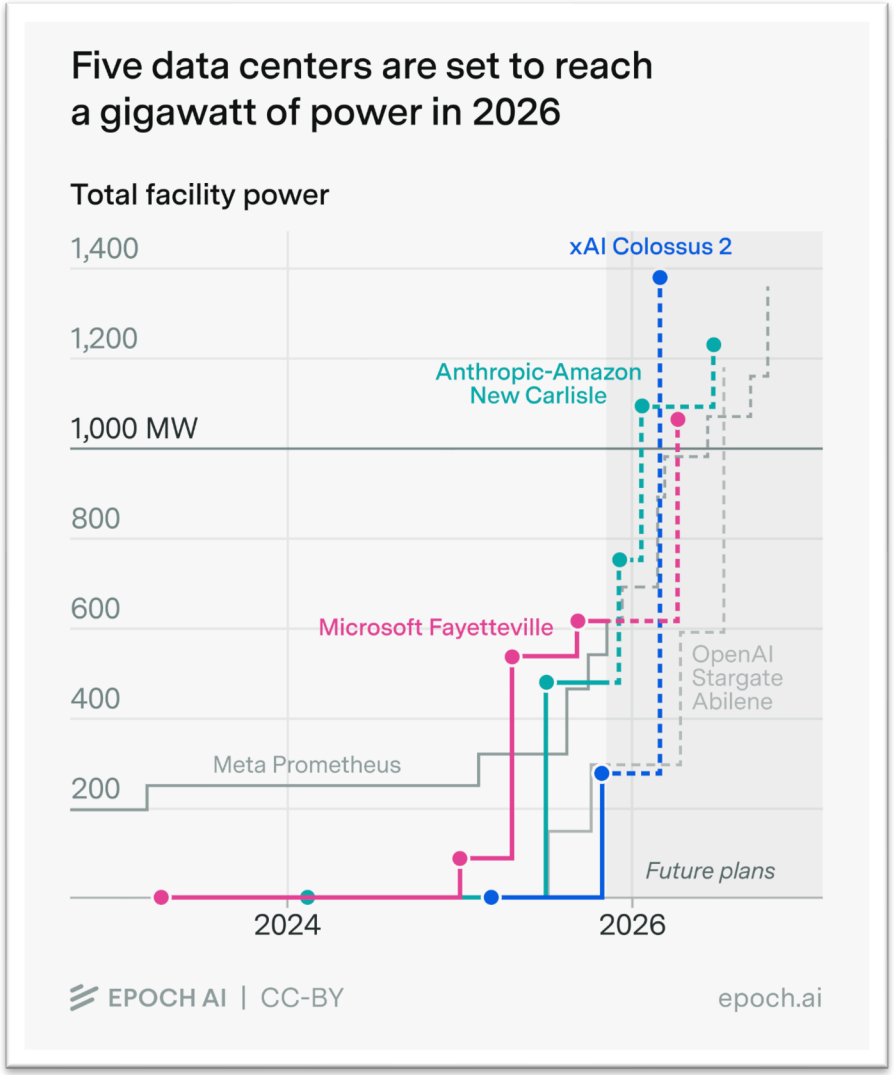
With Jae-Won Chung and Mosharaf Chowdhury

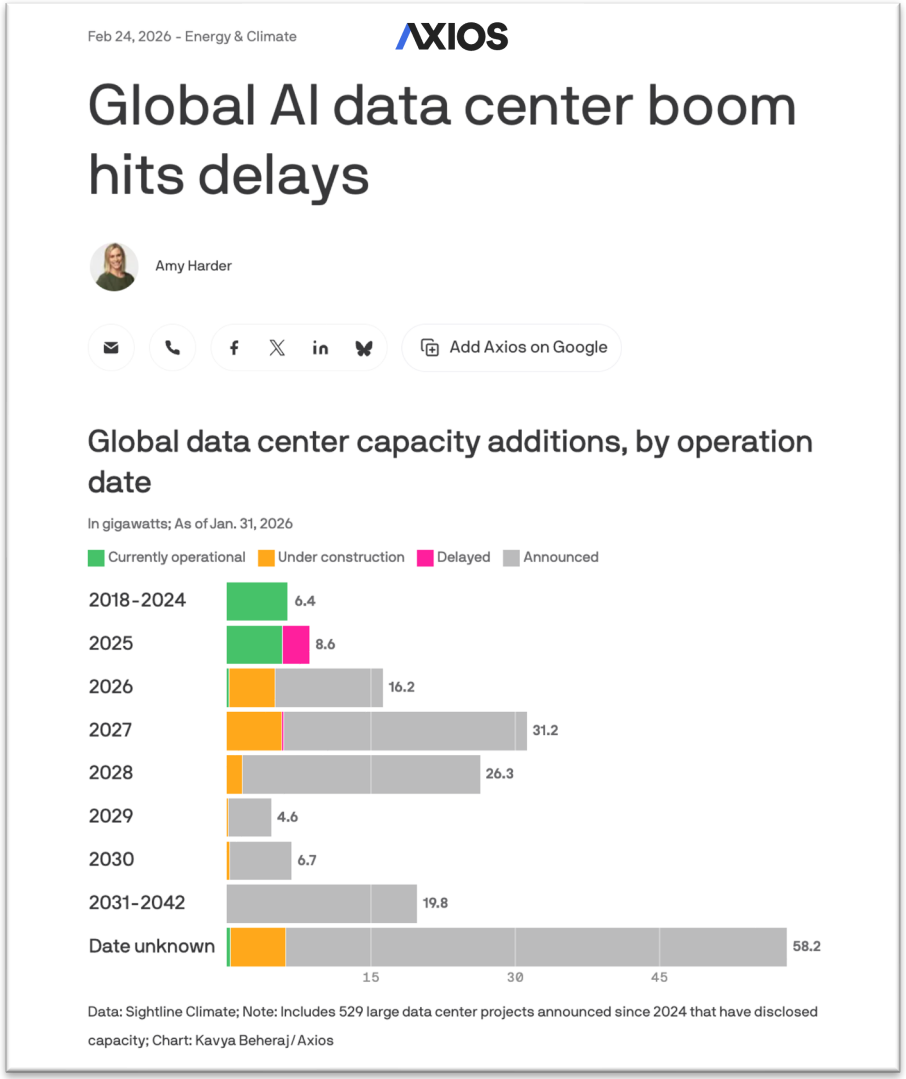






Energy is the Bottleneck for AI Scaling





Energy Breakdown

GPU energy

= Dynamic energy + Static energy

↑

Total work

@

GPU frequency

↑

Static power

x

Time

Energy Breakdown

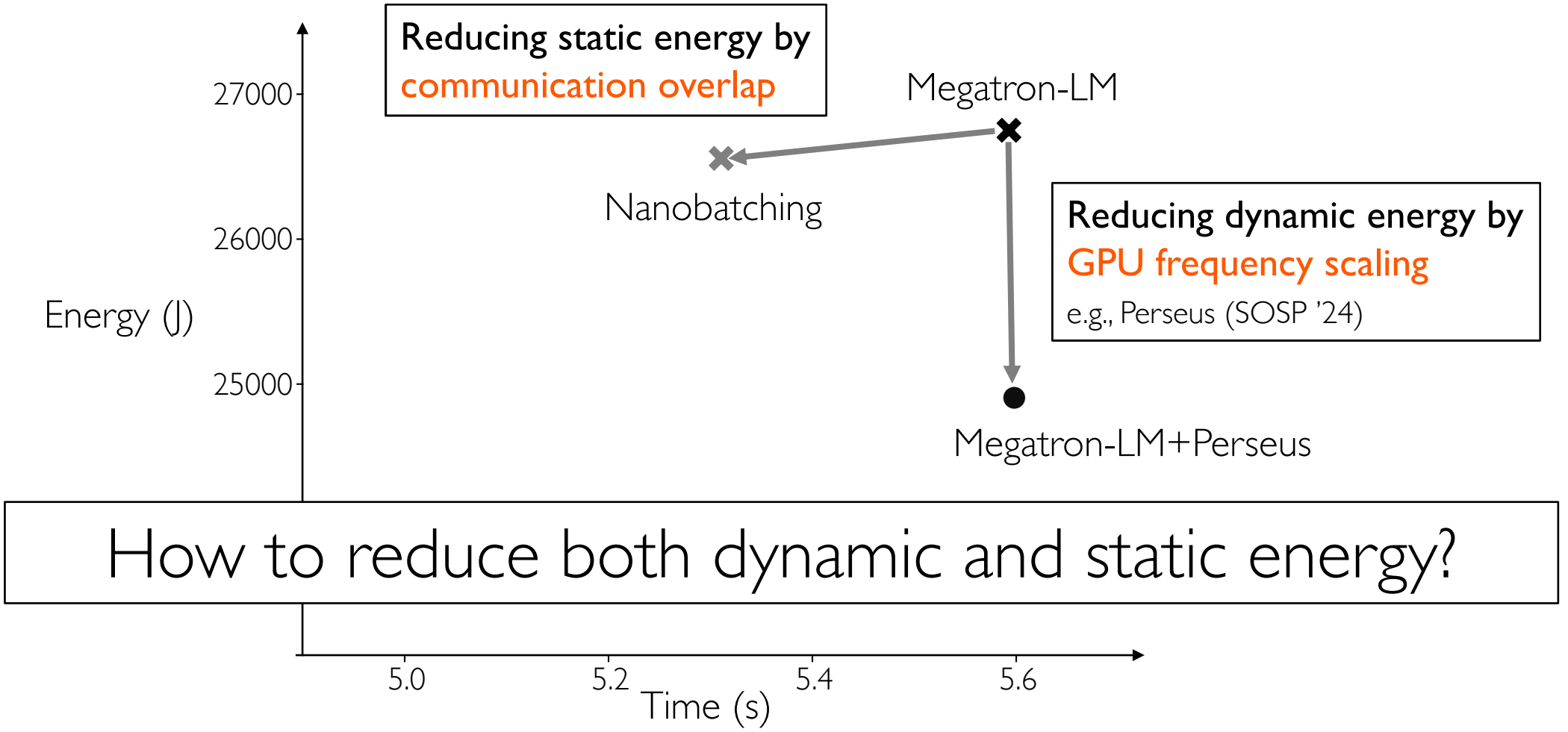
GPU energy

= Dynamic energy + Static energy

Total work @ GPU frequency

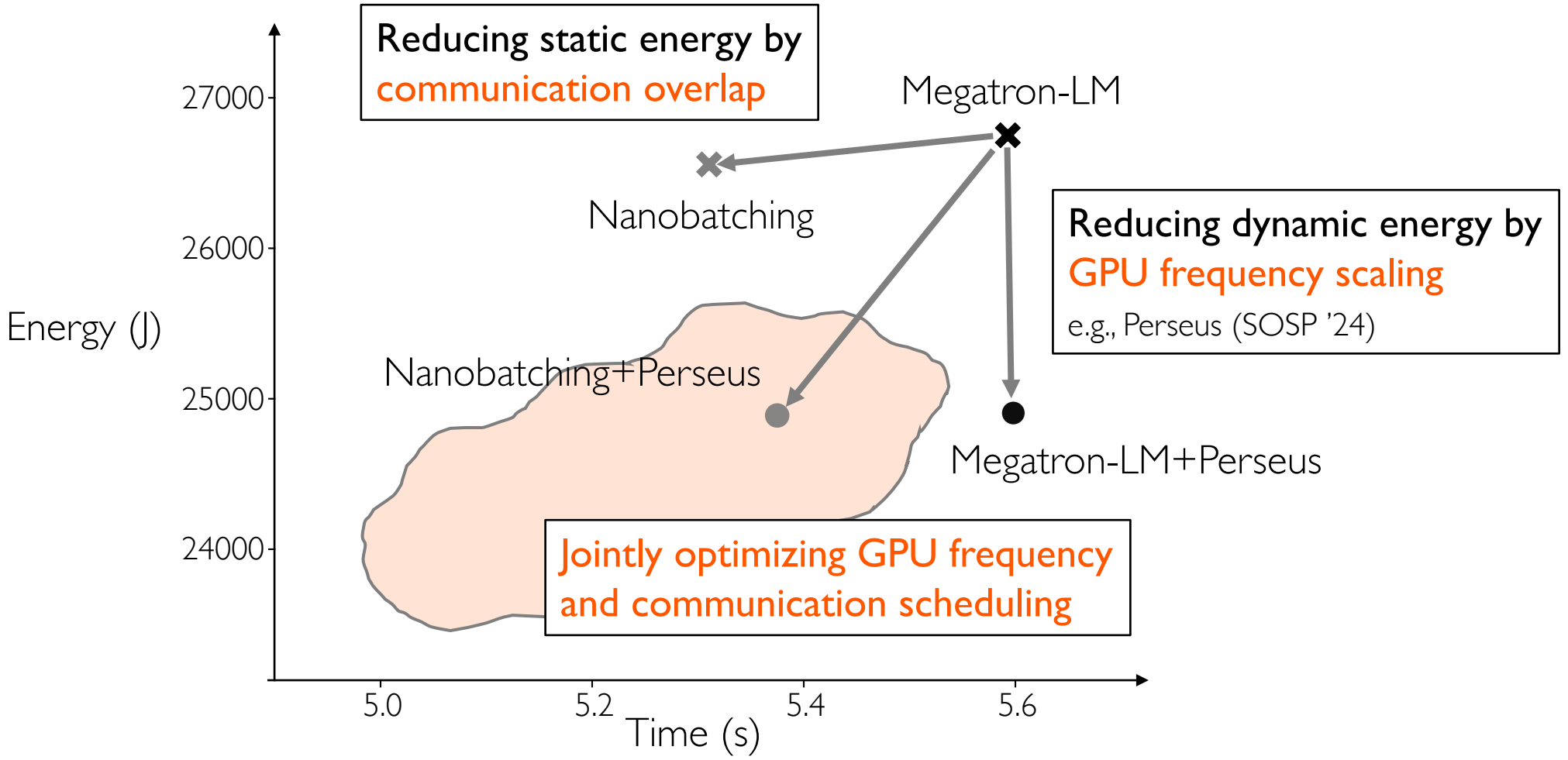
Static power × Time

Existing Work Targets Either Time or Energy



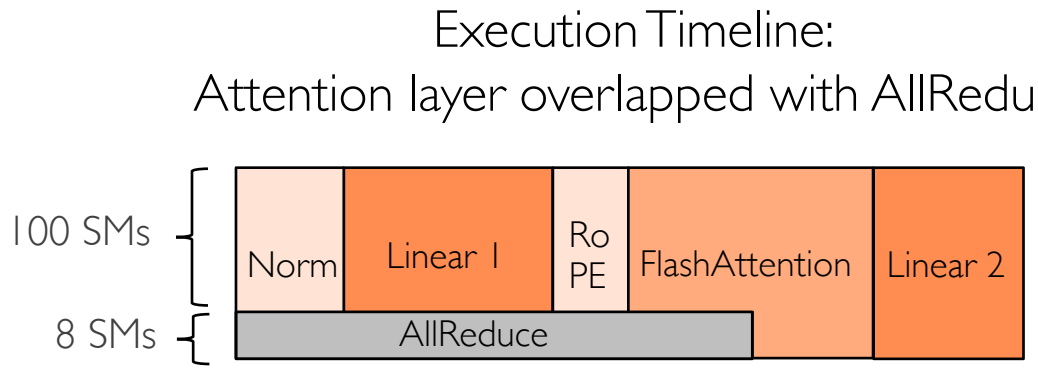
Iteration time and energy for Qwen 3 1.7B

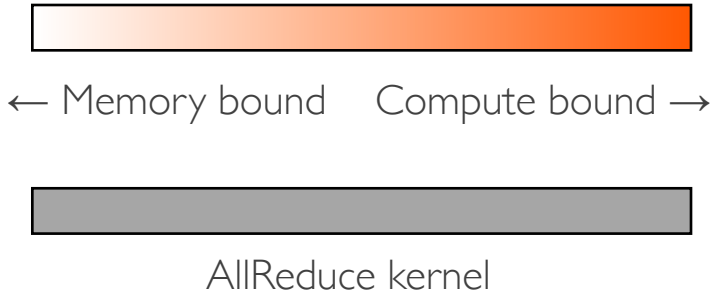
Existing Work Targets Either Time or Energy



Iteration time and energy for Qwen 3 1.7B

Zooming into Execution Schedule





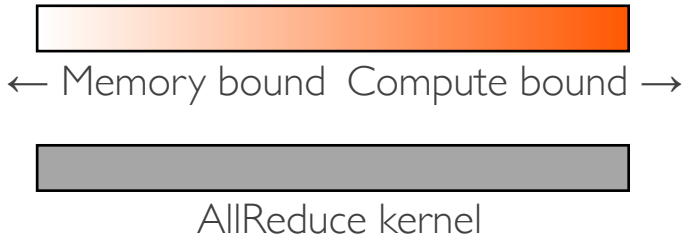
Kernel scheduling

- Number of SMs allocated to AllReduce
- When to launch AllReduce?

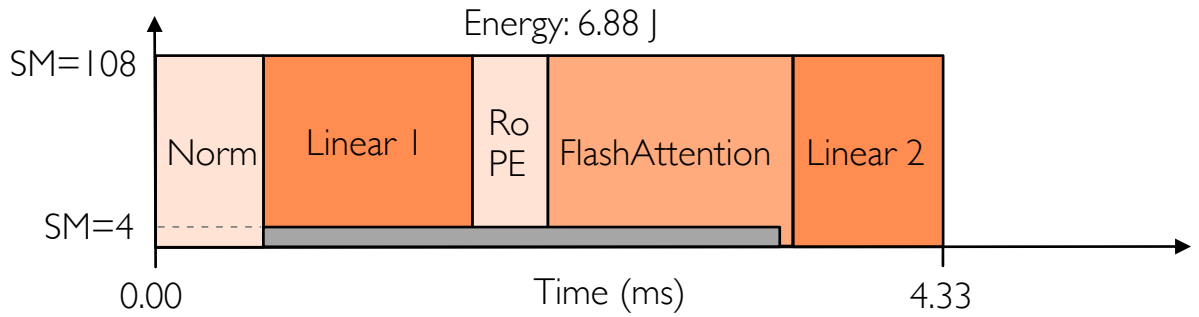
GPU frequency

- SM/core clock

SM Allocation Matters

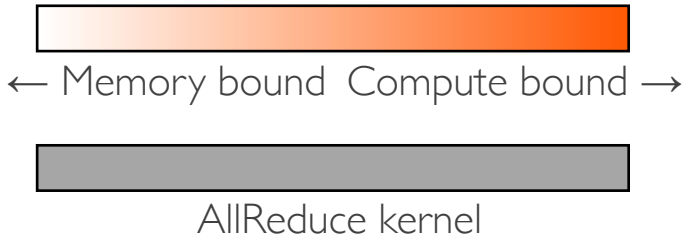


SM allocation

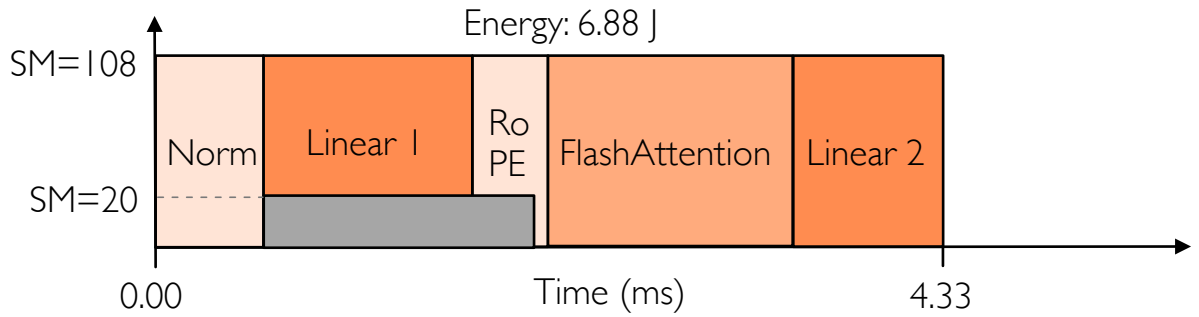


Llama 3.2 3B Attention Layer on 4 NVIDIA A100 GPUs

SM Allocation Matters

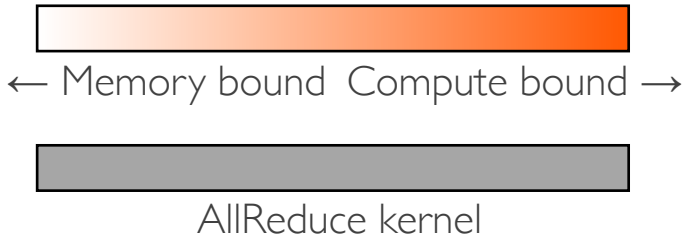


SM allocation

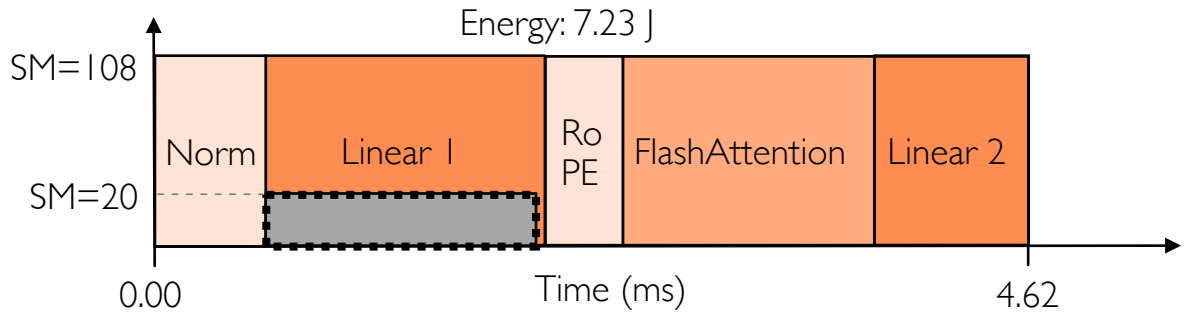


Llama 3.2 3B Attention Layer on 4 NVIDIA A100 GPUs

SM Allocation Matters



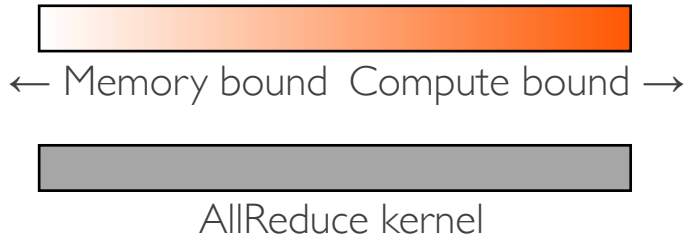
SM allocation



Llama 3.2 3B Attention Layer on 4 NVIDIA A100 GPUs

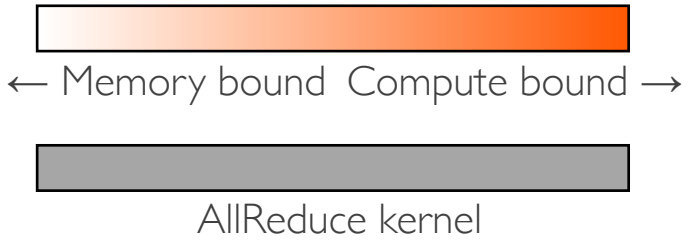
SM underutilization → Static power wasted

SM Allocation Matters



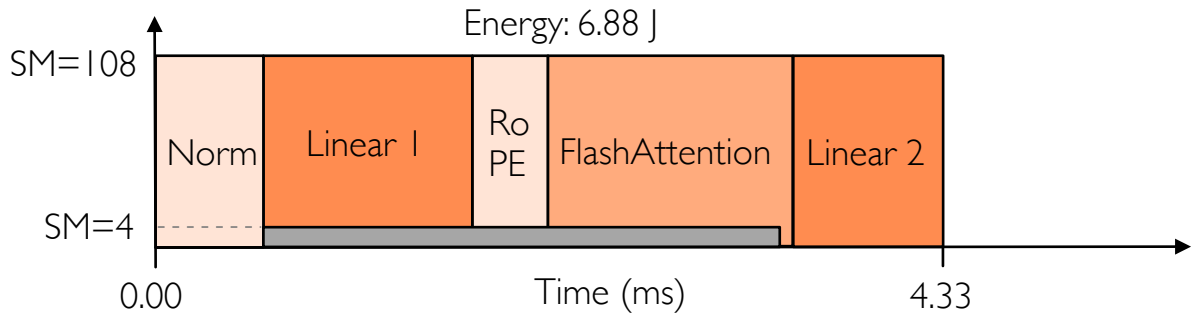
SM allocation SM underutilization → Static power wasted

Launch Timing Matters

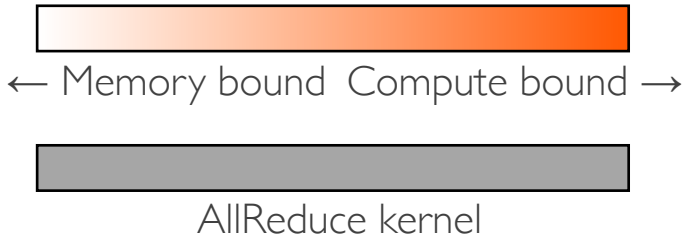


SM allocation SM underutilization → Static power wasted

Launch timing

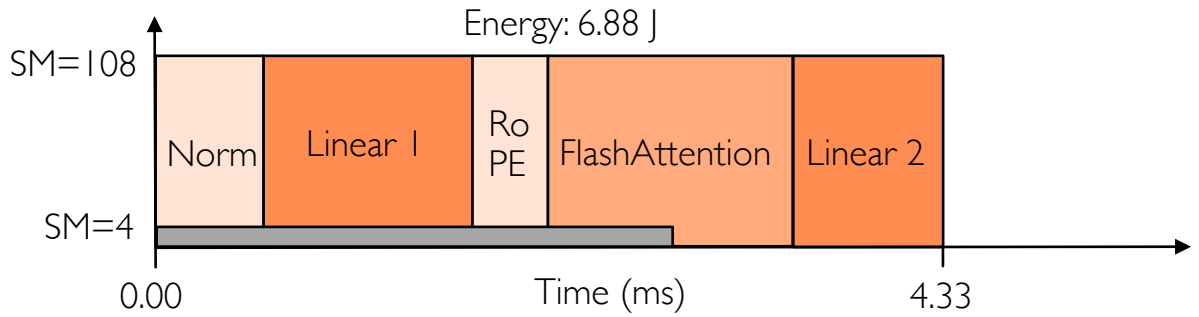


Launch Timing Matters

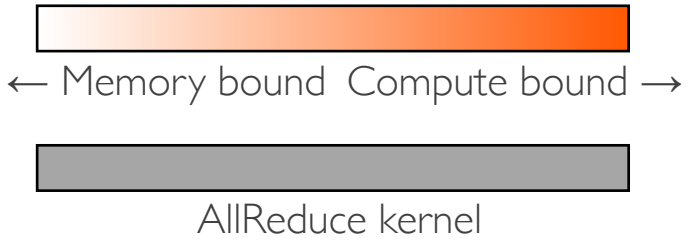


SM allocation SM underutilization → Static power wasted

Launch timing

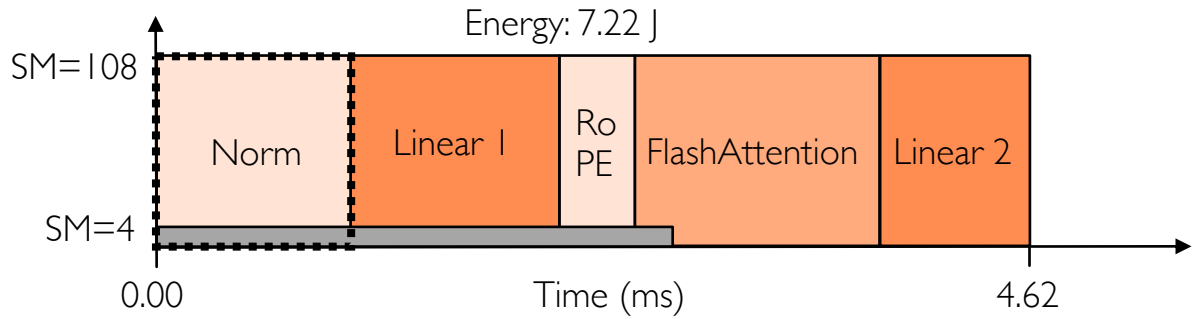


Launch Timing Matters



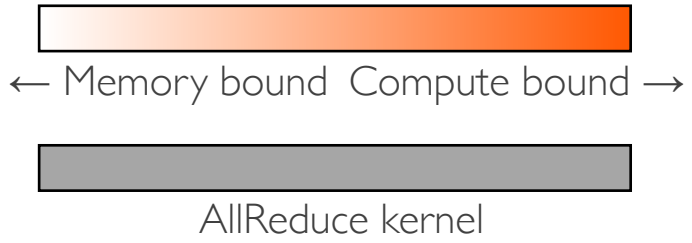
SM allocation SM underutilization → Static power wasted

Launch timing



Memory bandwidth interference → Static power wasted

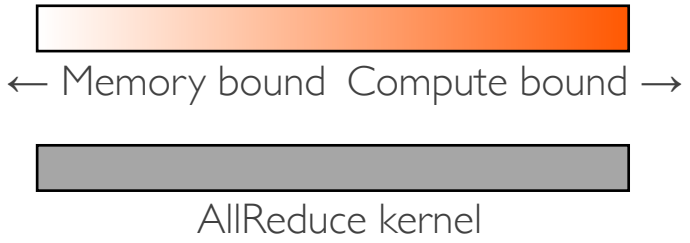
Launch Timing Matters



SM allocation SM underutilization → Static power wasted

Launch timing Memory bandwidth interference → Static power wasted

Frequency Reshapes Scheduling

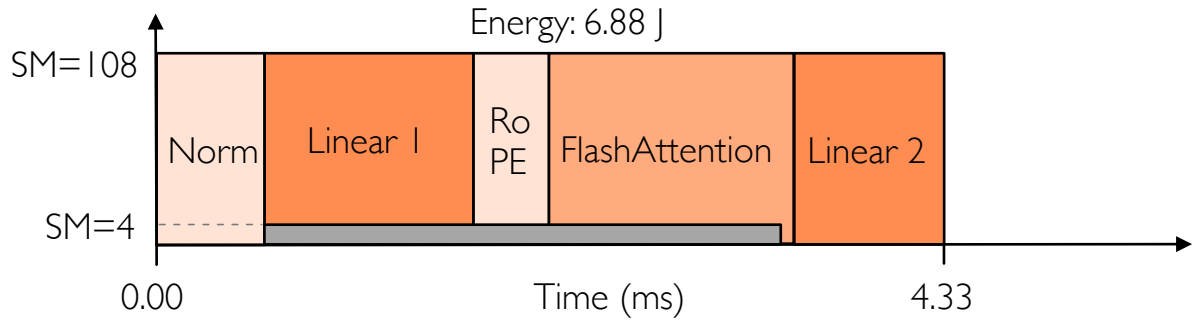


SM allocation SM underutilization → Static power wasted

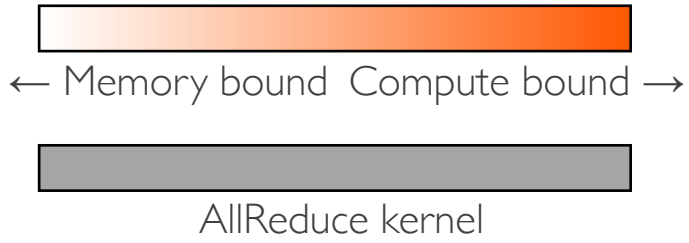
Launch timing Memory bandwidth interference → Static power wasted

SM frequency

@ 1410 MHz



Frequency Reshapes Scheduling

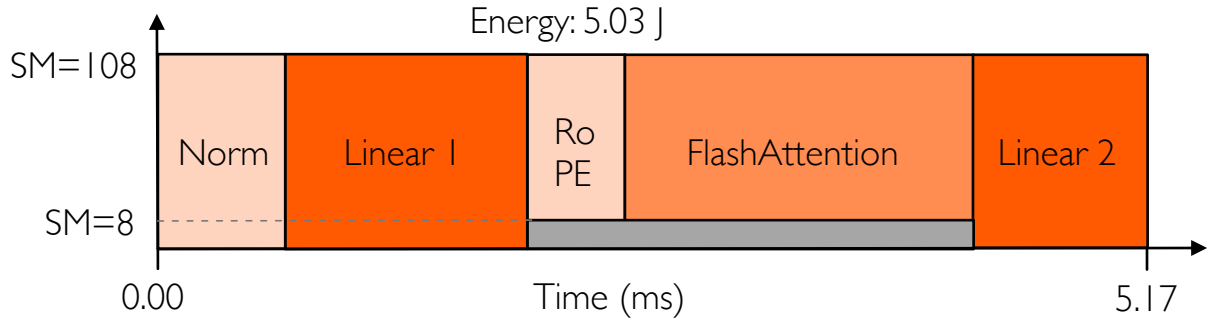


SM allocation SM underutilization → Static power wasted

Launch timing Memory bandwidth interference → Static power wasted

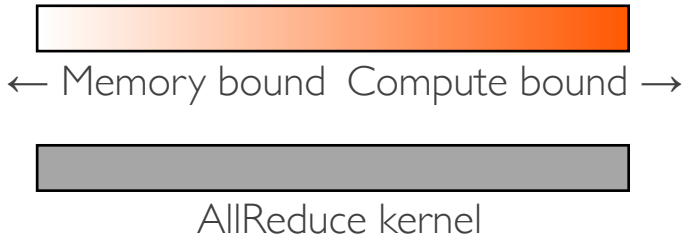
SM frequency

@ 1100 MHz



Compute speed reduced → Avoid overlapping with Linear

Frequency Reshapes Scheduling



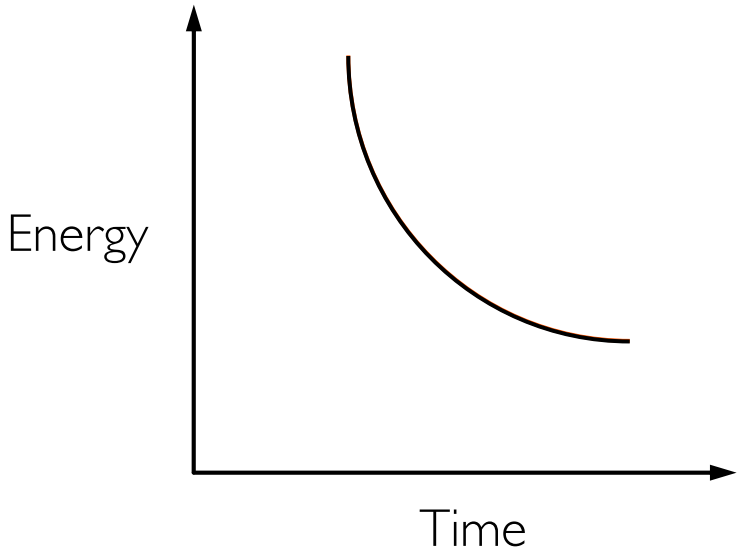
SM allocation SM underutilization → Static power wasted

Launch timing Memory bandwidth interference → Static power wasted

SM frequency Compute speed reduced → Avoid overlapping with Linear

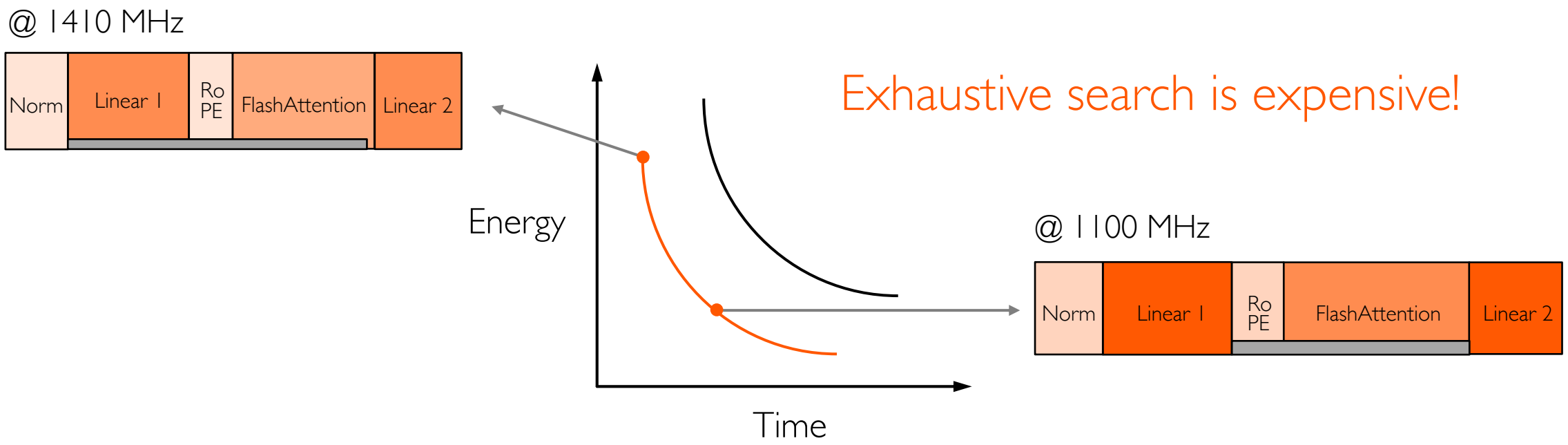
Kareus

GPU frequency and kernel scheduling interact.
Joint optimization reduces both dynamic and static energy.



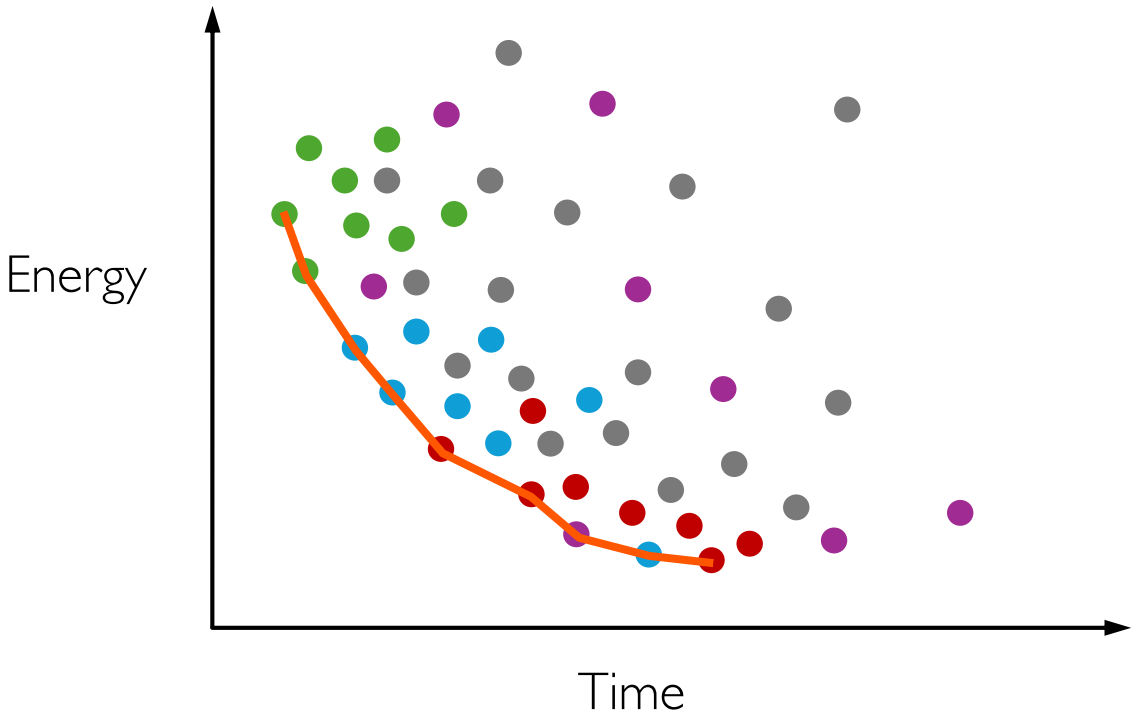
Kareus

GPU frequency and kernel scheduling interact.
Joint optimization reduces both dynamic and static energy.



Multi-Objective Bayesian Optimization

- Surrogate models (T, E) guide the search toward promising regions iteratively

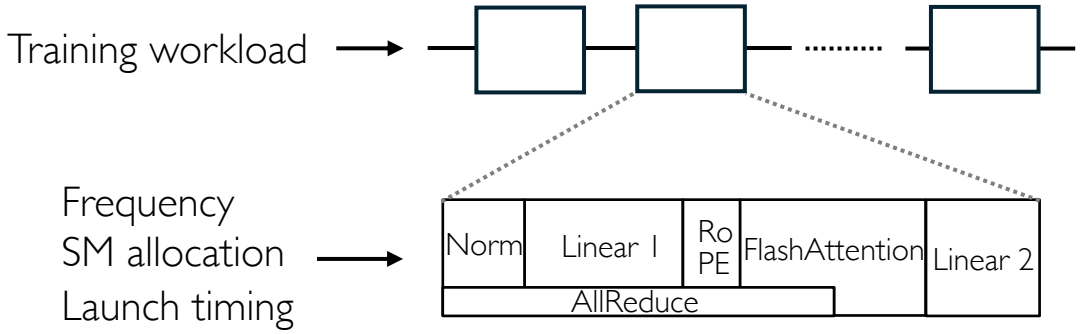


- Random initialization
- Multi-pass selection
  - Total energy pass
  - Dynamic energy pass
  - Static energy pass
  - Uncertainty pass

System Workflow

1. Optimize

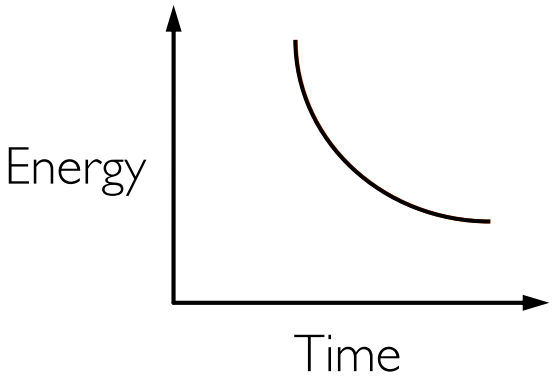
Multi-objective
Bayesian Optimization

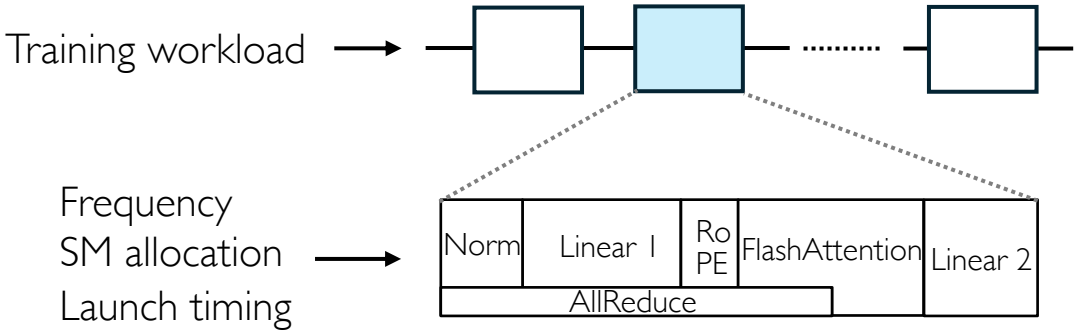


System Workflow

1. Optimize

Multi-objective
Bayesian Optimization

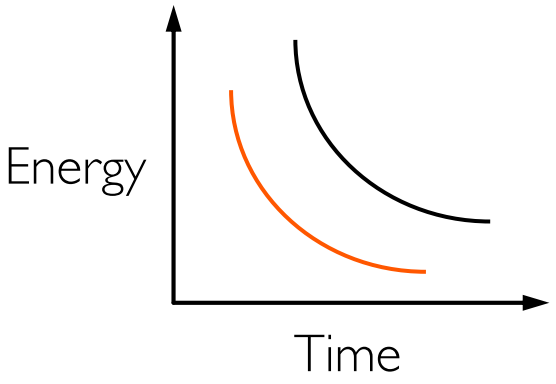


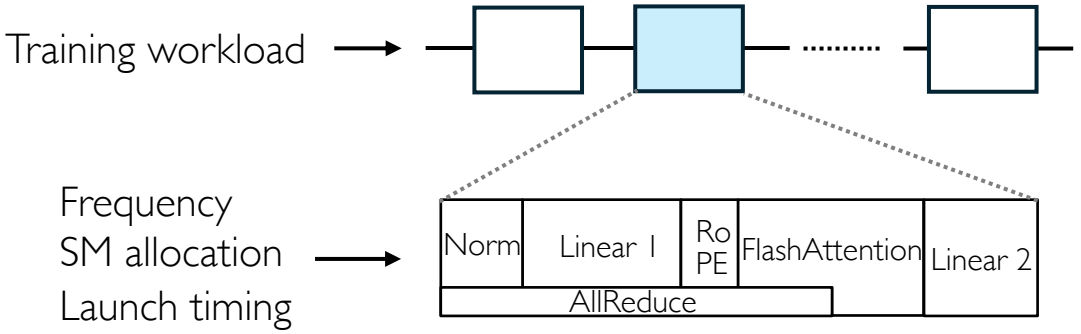


System Workflow

1. Optimize

Multi-objective
Bayesian Optimization

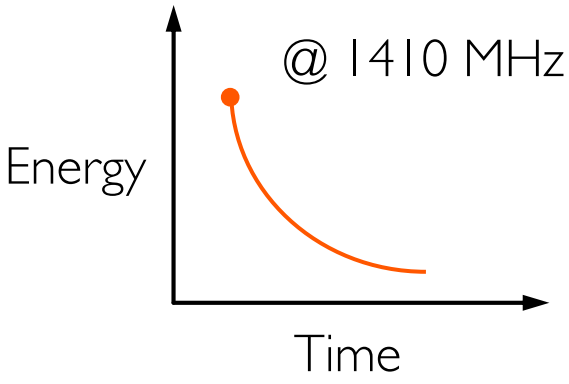




System Workflow

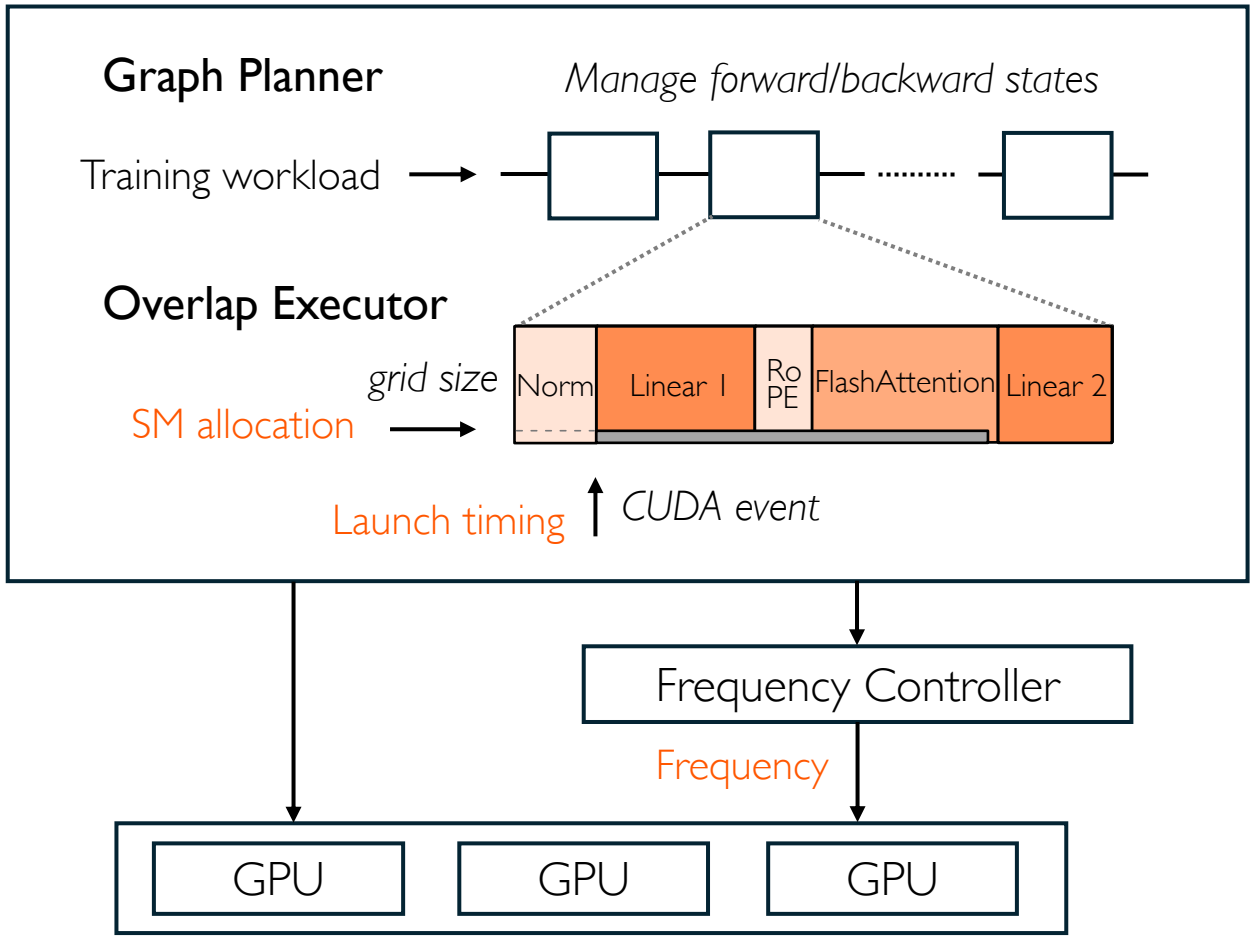
1. Optimize

Multi-objective Bayesian Optimization



Job requirement

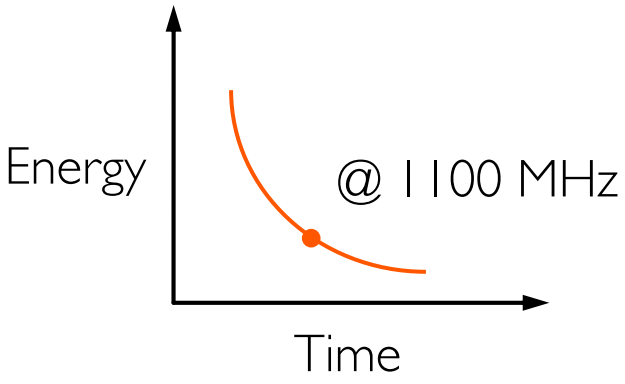
2. Training Runtime



System Workflow

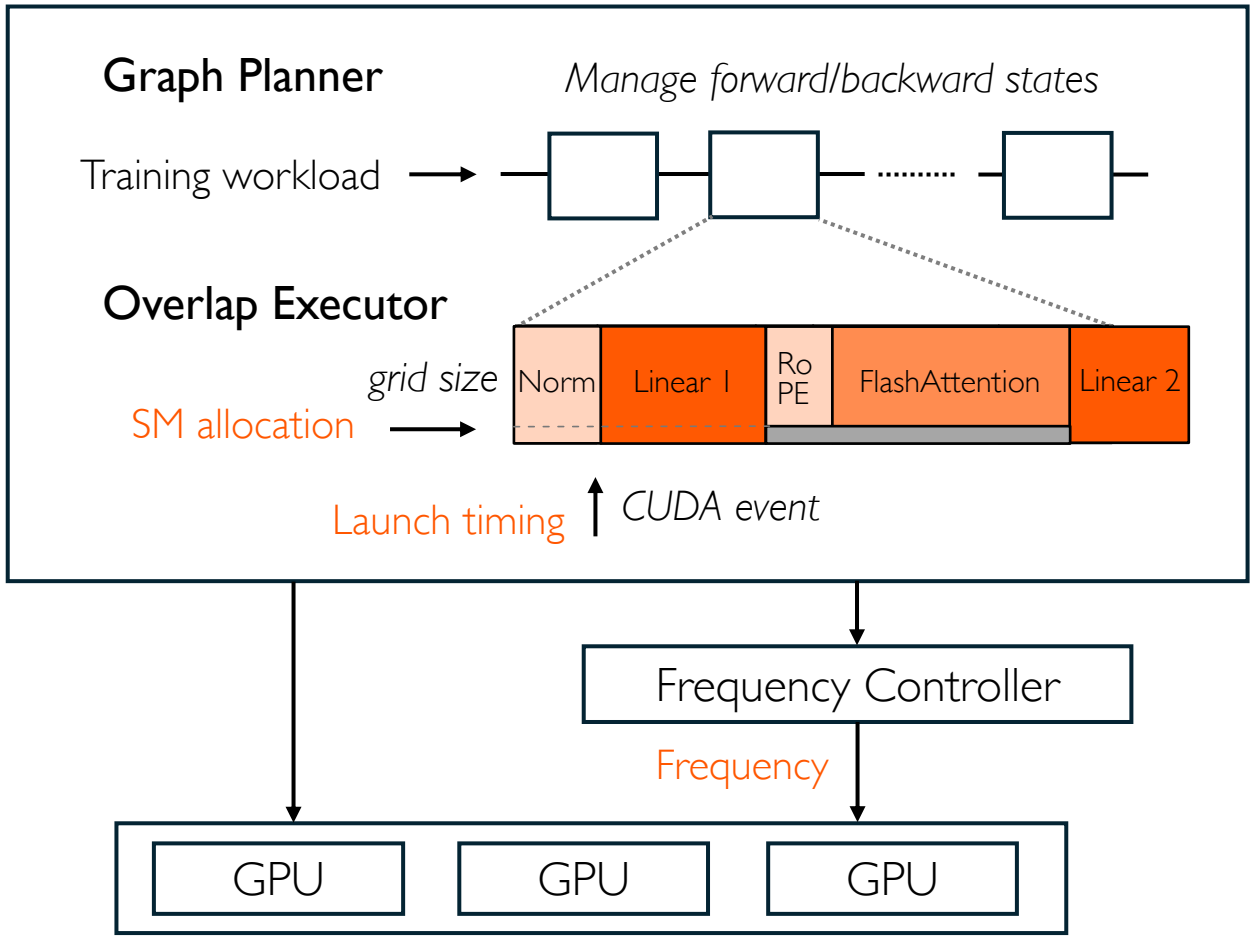
1. Optimize

Multi-objective Bayesian Optimization



Job requirement

2. Training Runtime



Evaluation Setup

Setup and workloads

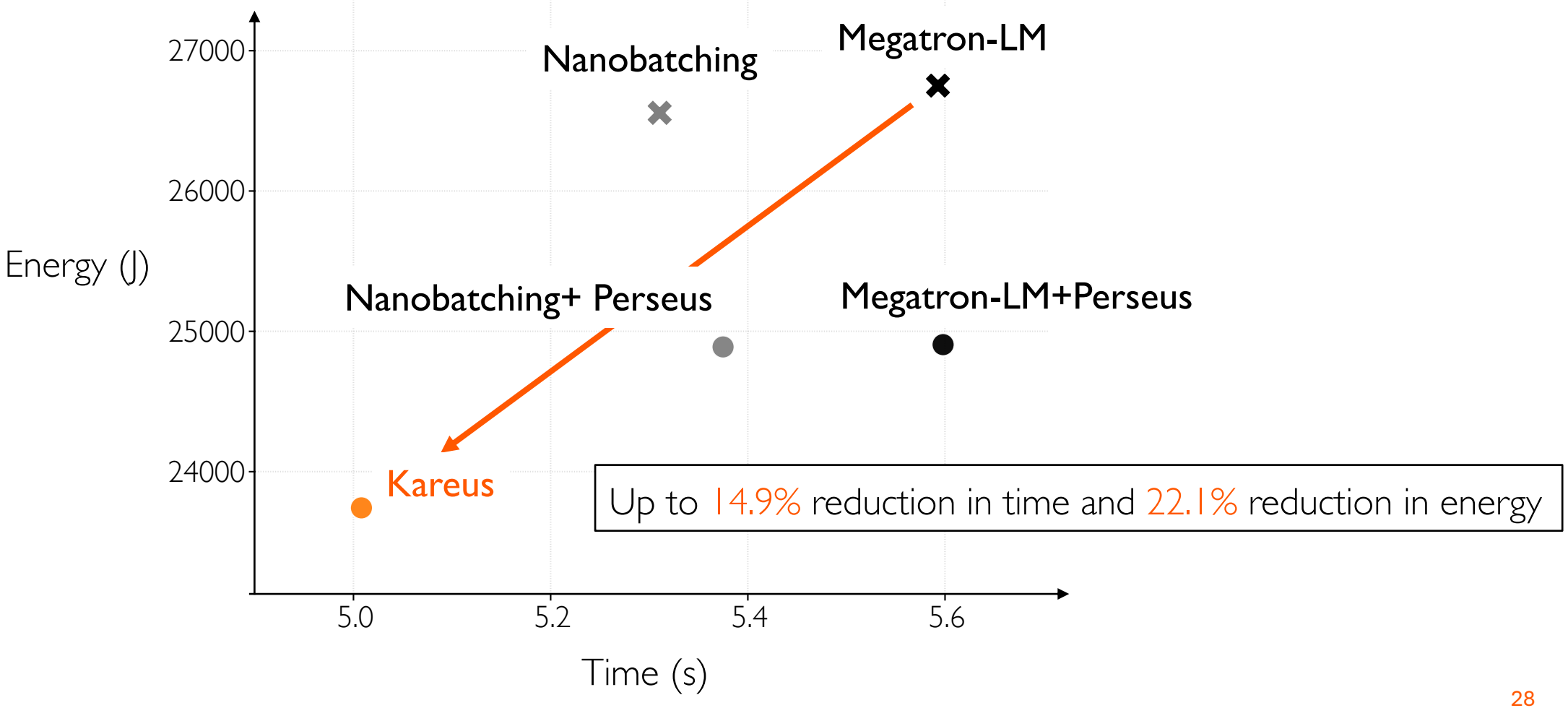
- Testbed: Two collocated 8×A100 nodes with NVSwitch
- 14 workloads on Llama 3.2 3B and Qwen 3 1.7B

Questions

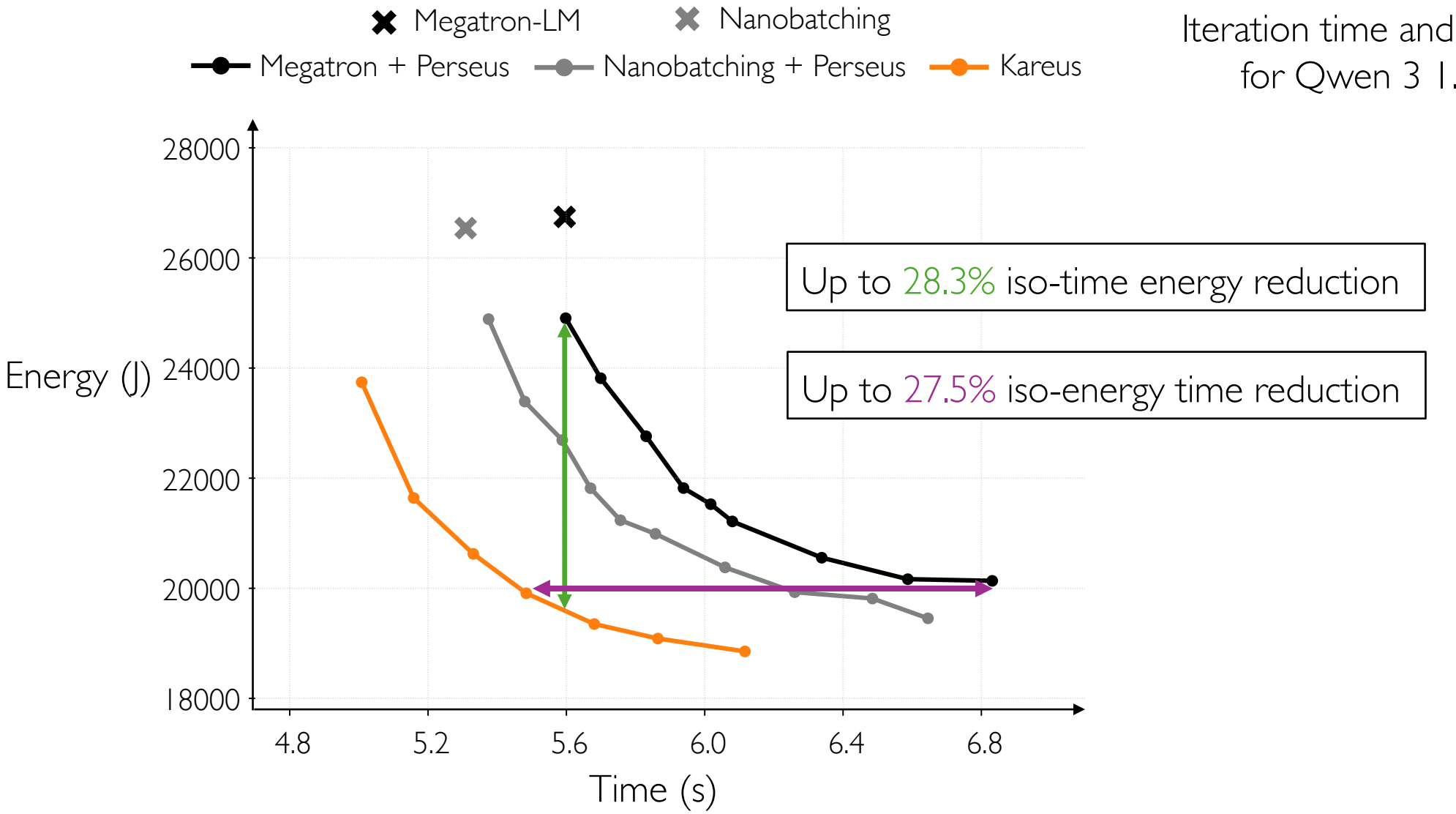
- Does Kareus reduce both time and energy in end-to-end experiments?
- Is joint optimization of dynamic and static energy necessary?
- Does this benefit generalize to large-scale training?
- Is Kareus effective across different workloads?
- Is the optimization overhead reasonable?

Kareus Reduces Both Time and Energy

Iteration time and energy
for Qwen 3 1.7B



Kareus Pushes the Time–Energy Frontier



Iteration time and energy for Qwen 3 1.7B

Looking Forward

- We are collaborating with NVIDIA to optimize energy efficiency in production training stack.
- *Energy Calculus* provides an analytical framework for reasoning about time and energy as composable system resources.

NVIDIA blog post



Energy Calculus

ml.energy/calculus

ML.ENERGY