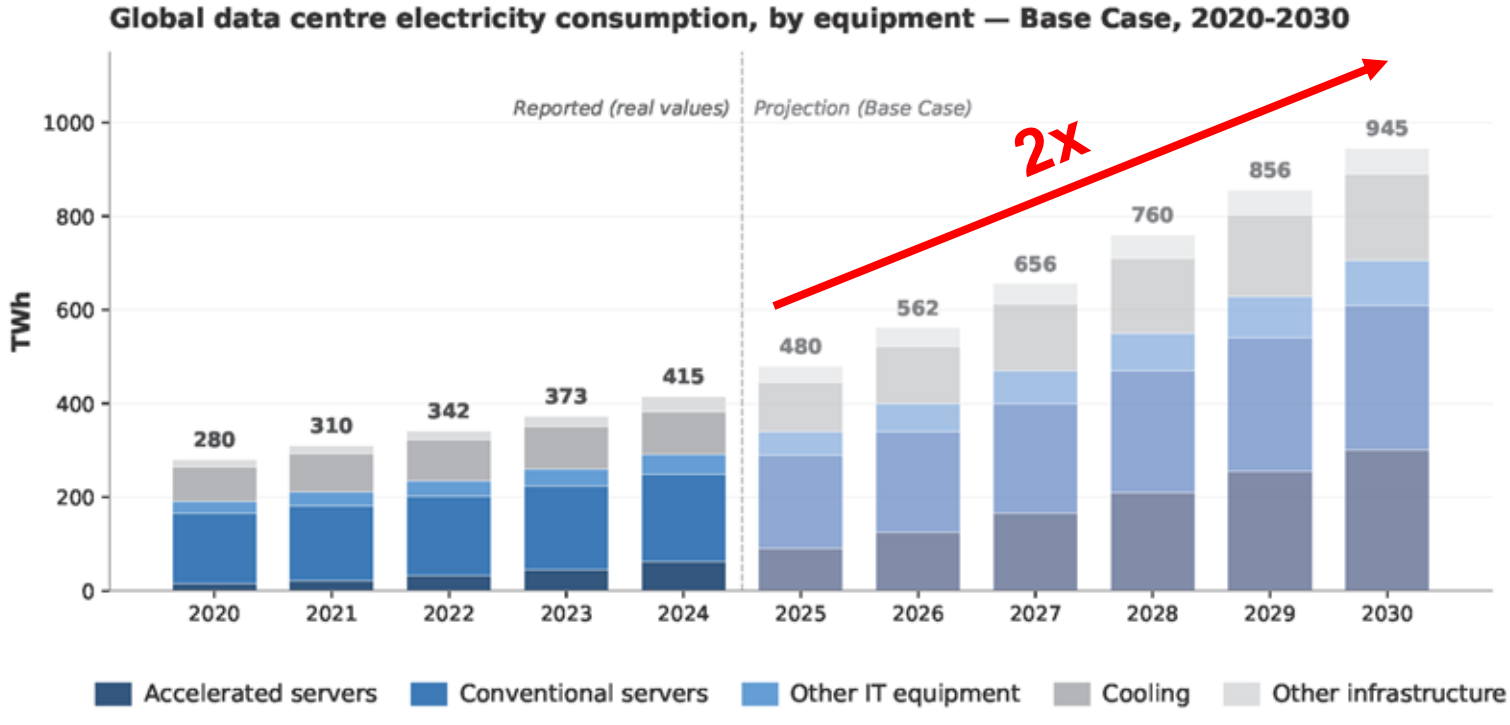
BEAM: Joint Resource–Power Optimization for Energy-Efficient LLM Inference under SLO constraints

Hyunjae Lee, Sangjin Choi, Seungjae Lim, Youngjin Kwon

KAIST

Motivation

Energy Crisis



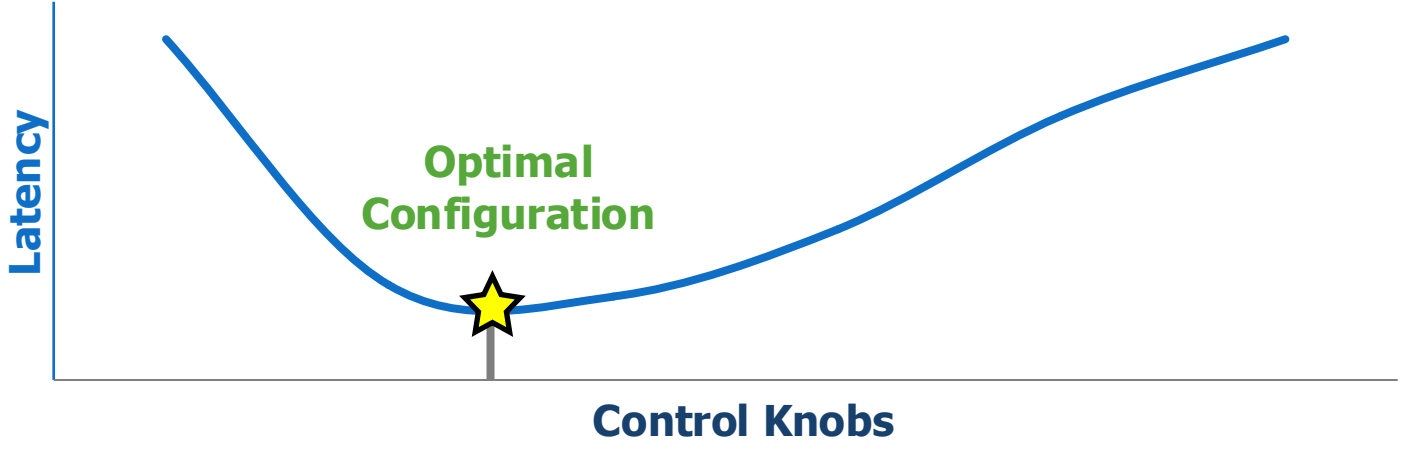
Source: IEA (2025), Energy and AI special report (Figure 2.11). Licence: CC BY 4.0.

Motivation

Performance-Oriented Objective

Performance-Oriented Goal

Minimize TTFT / TBT
Maximize Throughput

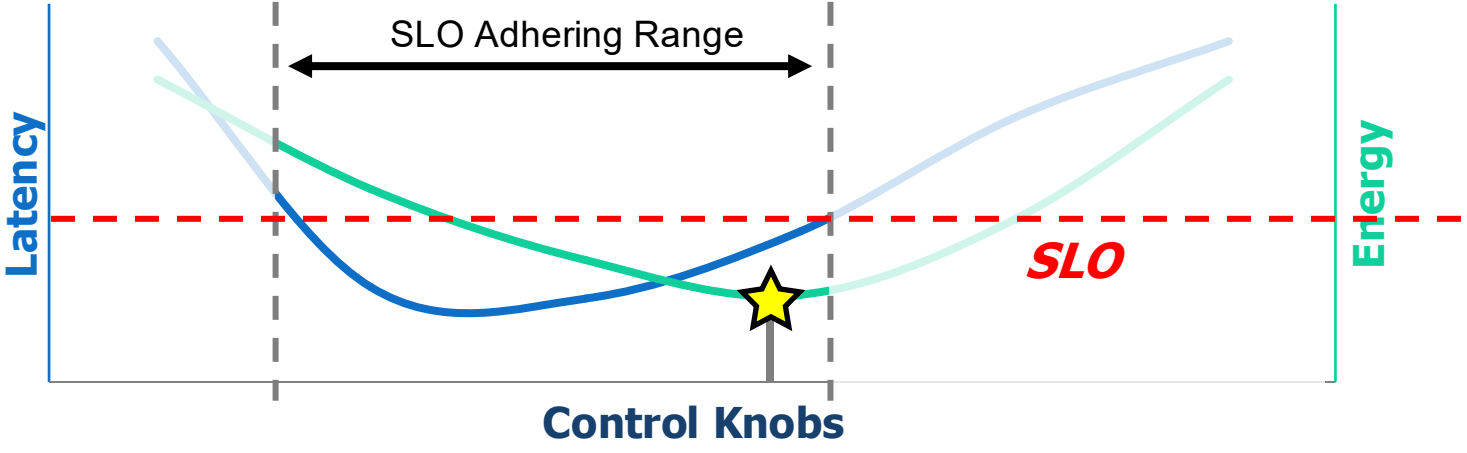


Motivation

Energy-Oriented Objective

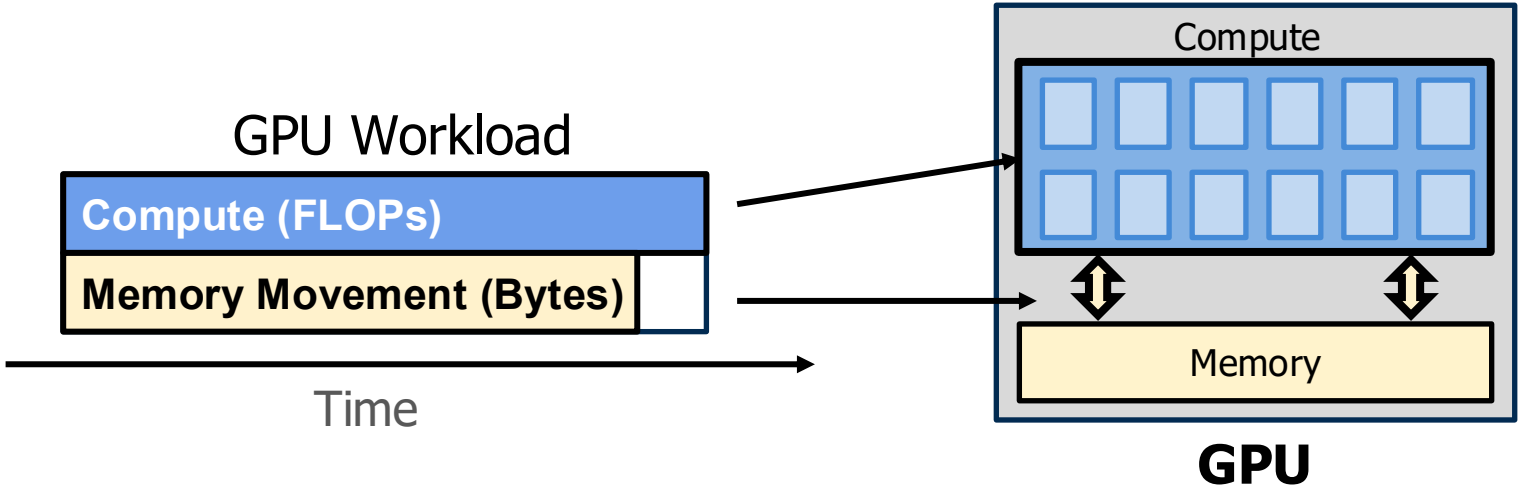
Energy-Oriented Goal

Minimize Energy Usage
Adhere to TTFT/TBT SLO



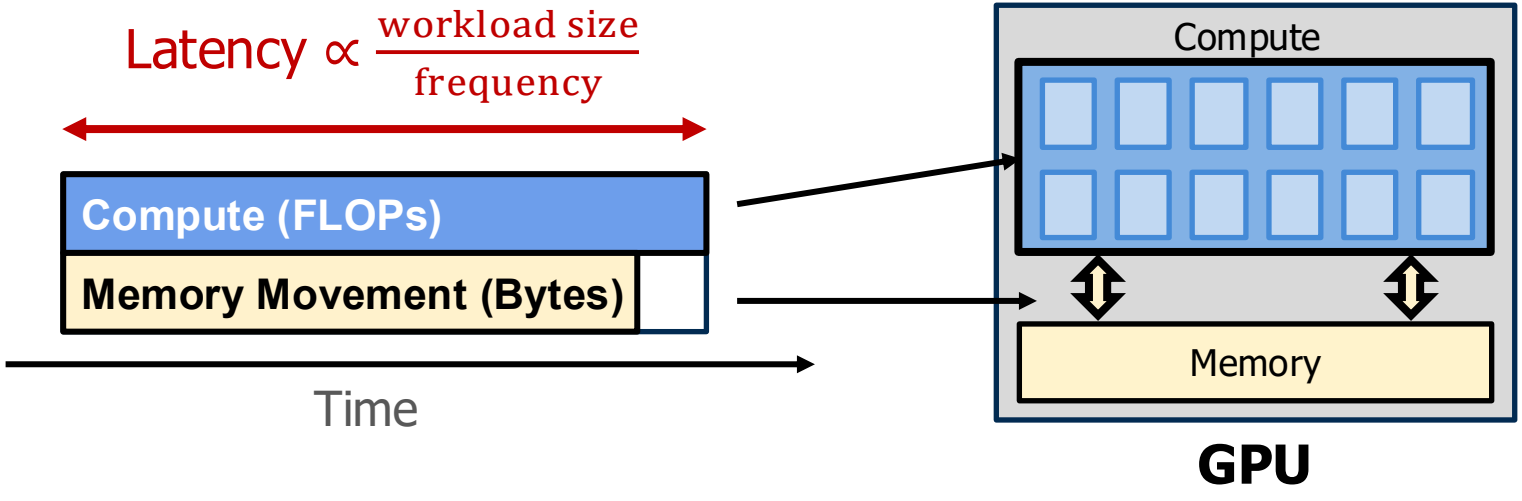
Motivation

Energy usage of GPU workloads



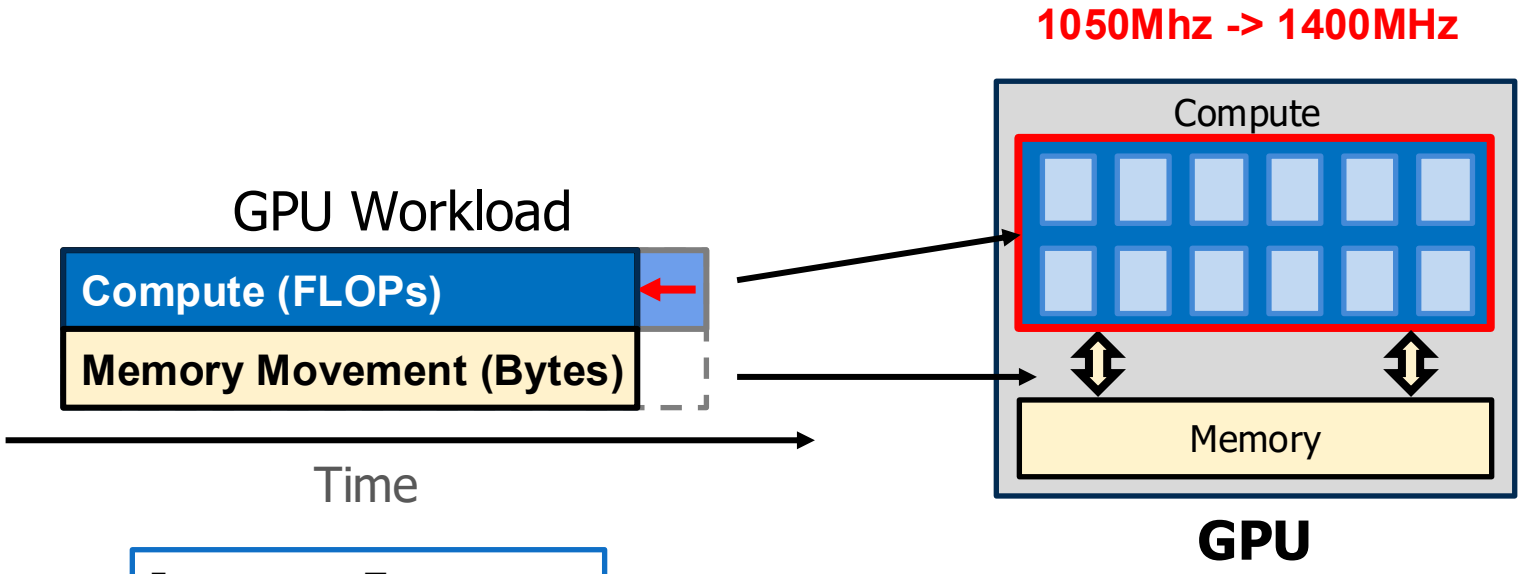
Motivation

Energy usage of GPU workloads



Motivation

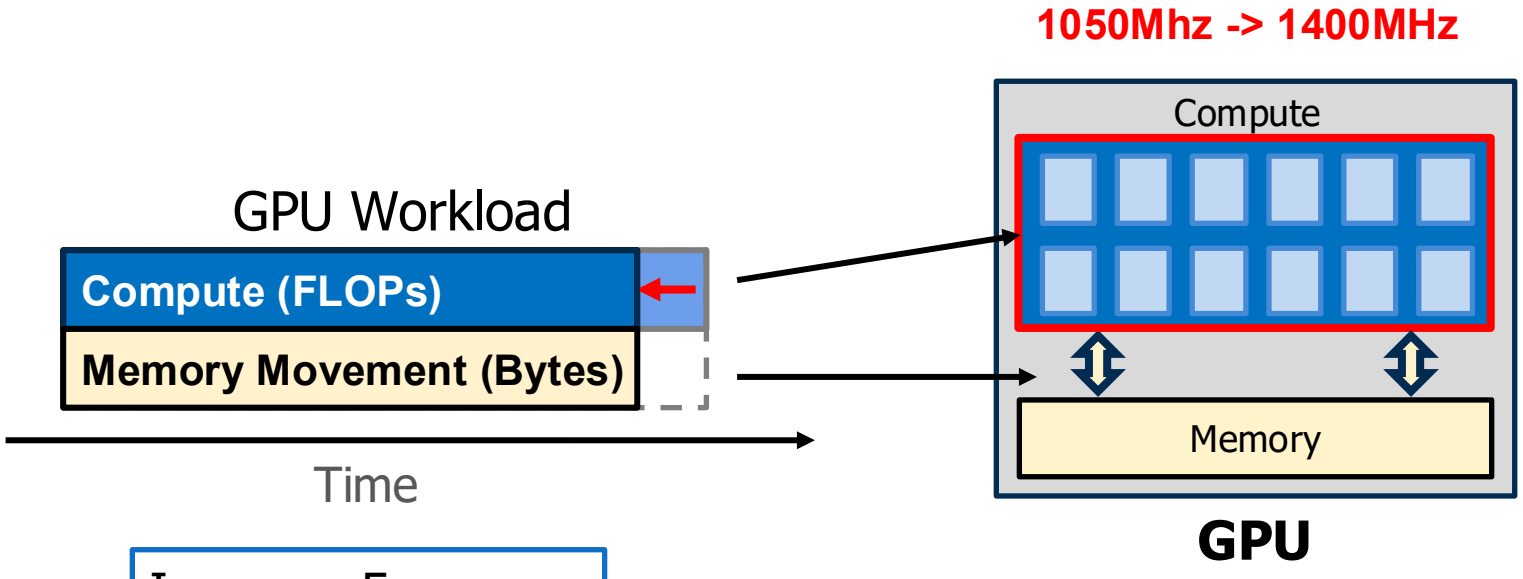
Energy usage of GPU workloads



Increase Frequency
Increase Power
Decrease Time

Motivation

Energy usage of GPU workloads

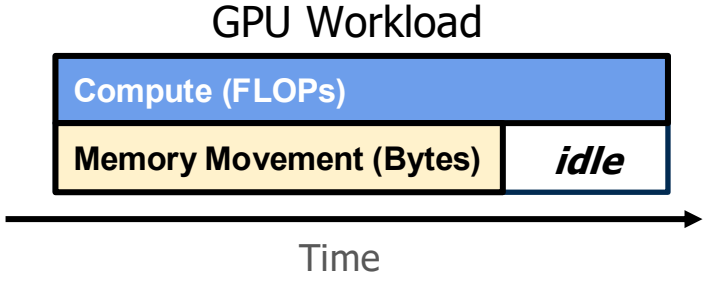


Increase Frequency
Increase Power
Decrease Time

Energy = Power x Time

Motivation

Energy usage of GPU workloads



Latency

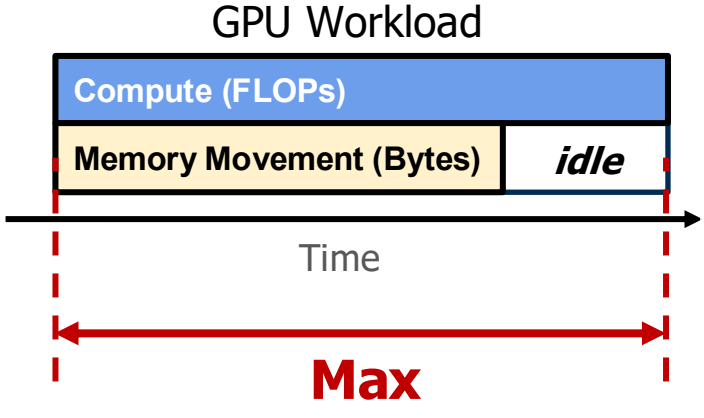
$$L_{\text{total}} = \max\left(L_{compute}, L_{memory}\right) + C$$

Energy

$$E_{total} = E_{\text{compute}} + E_{\text{memory}} + C$$

Motivation

Energy usage of GPU workloads



Latency

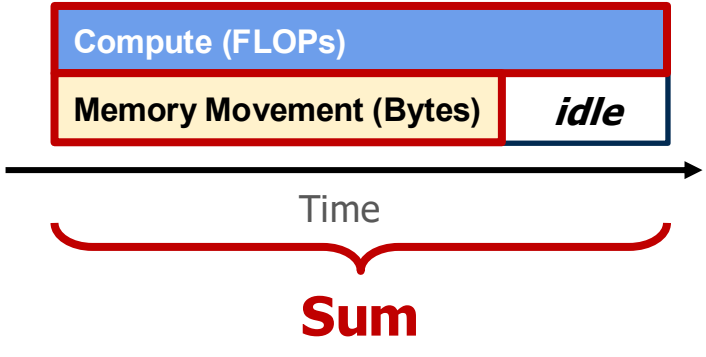
$$L_{\text{total}} = \max(L_{compute}, L_{memory}) + C$$

Energy

$$E_{total} = E_{\text{compute}} + E_{\text{memory}} + C$$

Motivation

Energy usage of GPU workloads



Latency

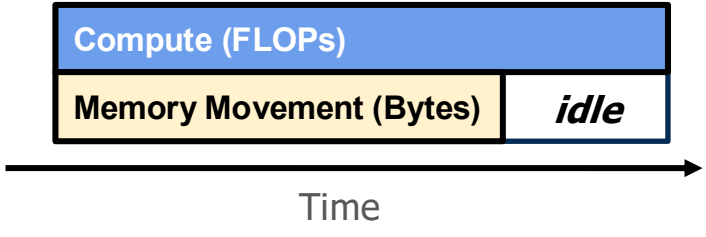
$$L_{\text{total}} = \max(L_{compute}, L_{memory}) + C$$

Energy

$$E_{total} = E_{\text{compute}} + E_{\text{memory}} + C$$

Motivation

Energy usage of GPU workloads



Latency

$$L_{\text{total}} = \max\left(L_{compute}, L_{memory}\right) + C$$

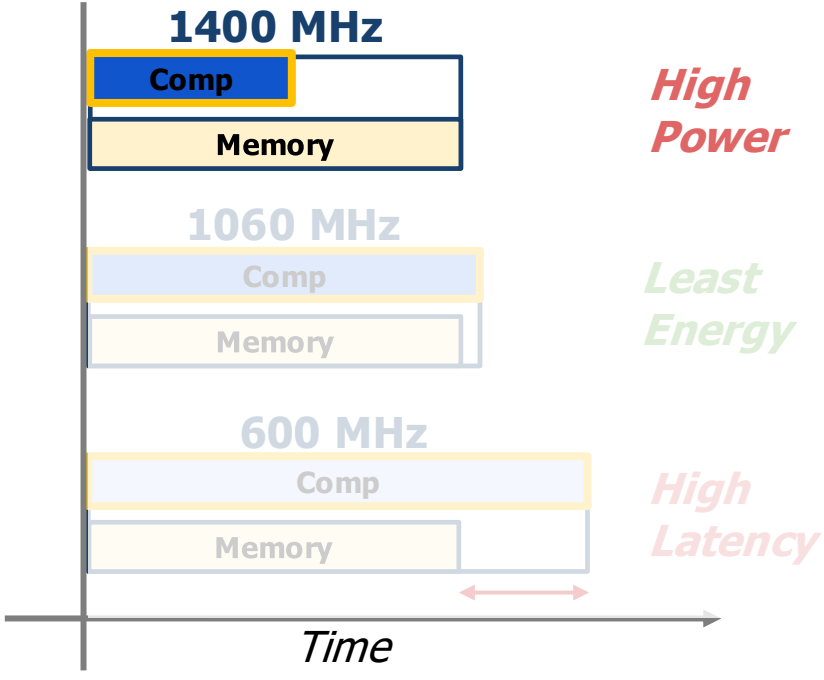
Energy

$$E_{total} = E_{\text{compute}} + E_{\text{memory}} + C$$

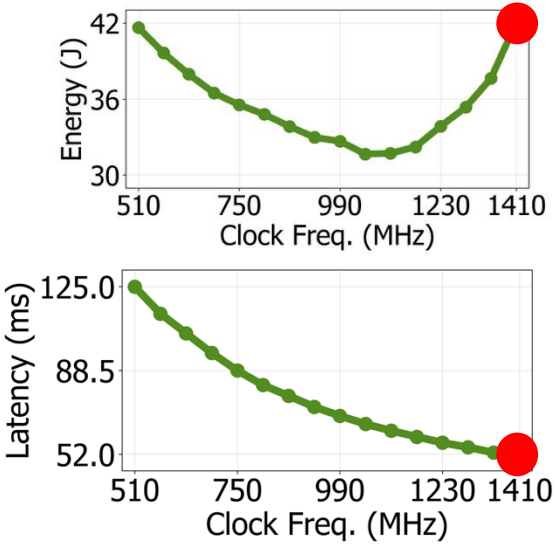
Workload composition determines performance-energy landscape

Motivation

DVFS : Dynamic Voltage Frequency Scaling

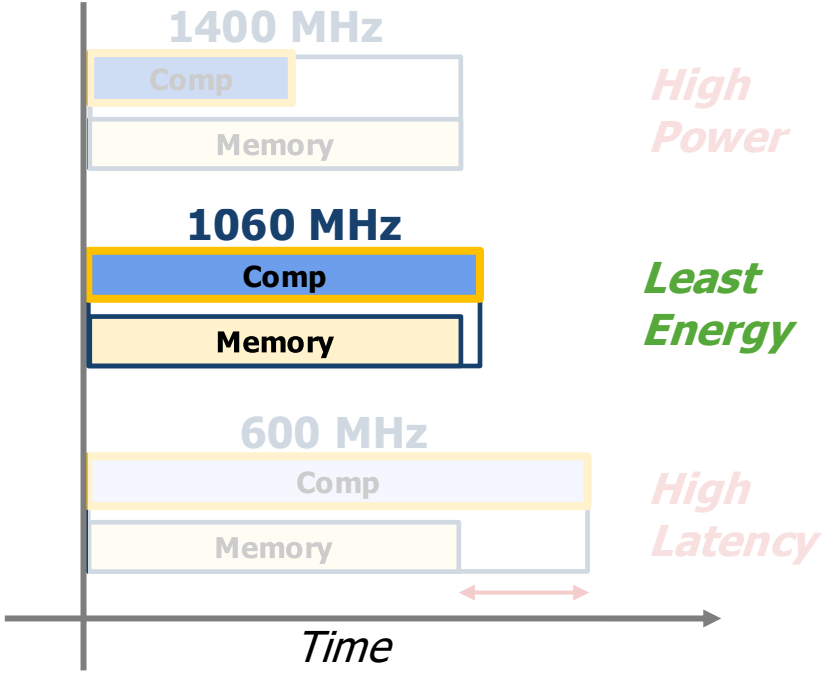


Energy = Power x Time

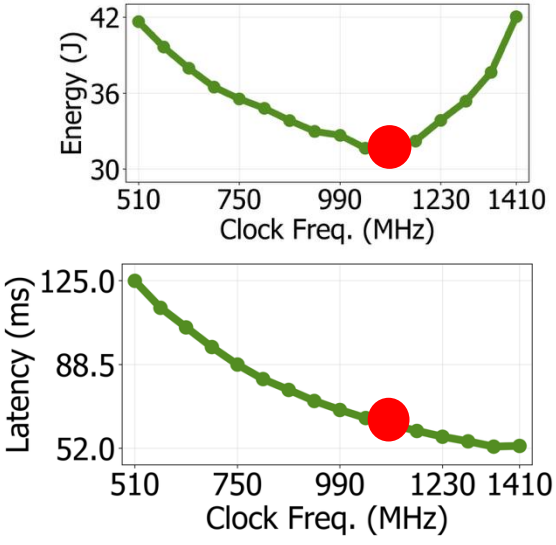


Motivation

DVFS : Dynamic Voltage Frequency Scaling

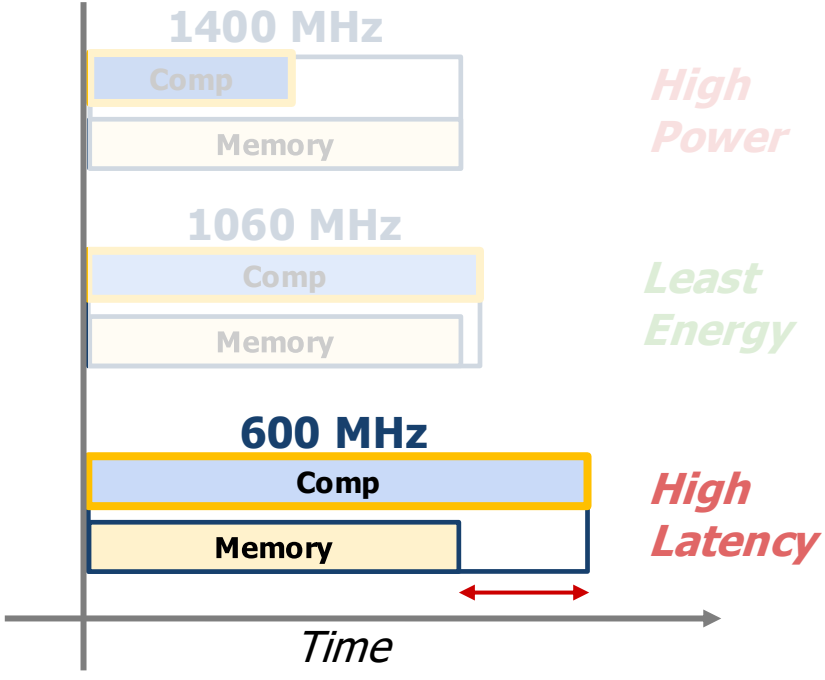


Energy = Power x Time

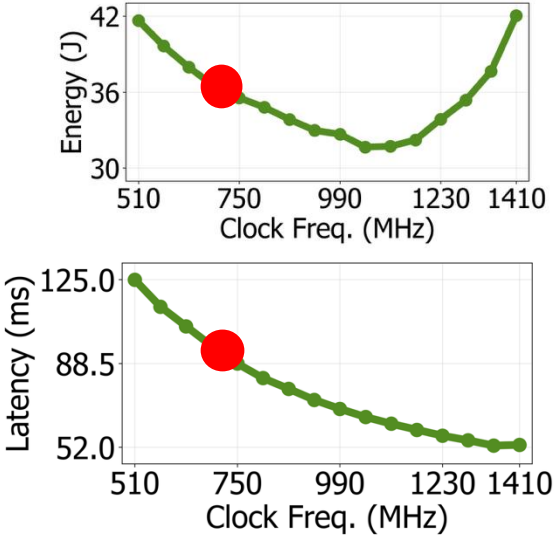


Motivation

DVFS : Dynamic Voltage Frequency Scaling

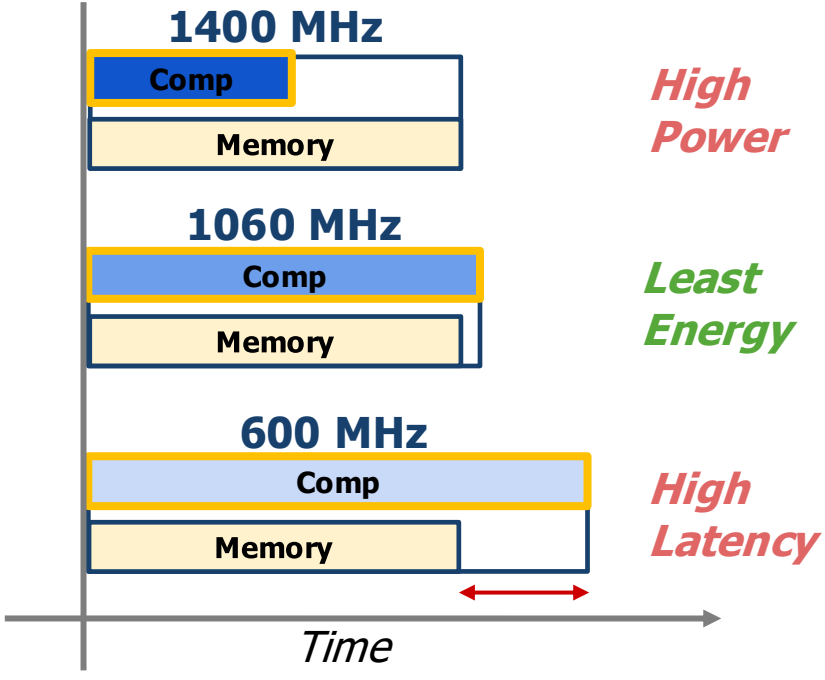


Energy = Power x Time



Motivation

DVFS : Dynamic Voltage Frequency Scaling

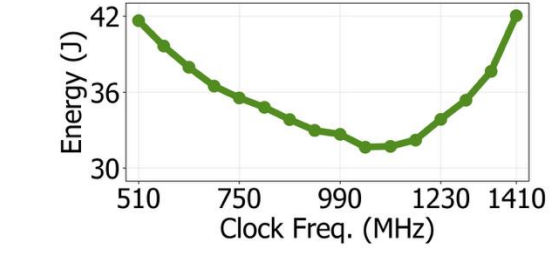


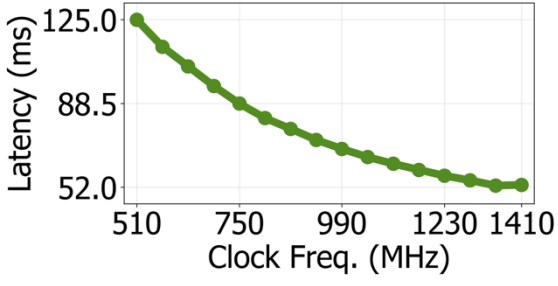
High Power

Least Energy

High Latency

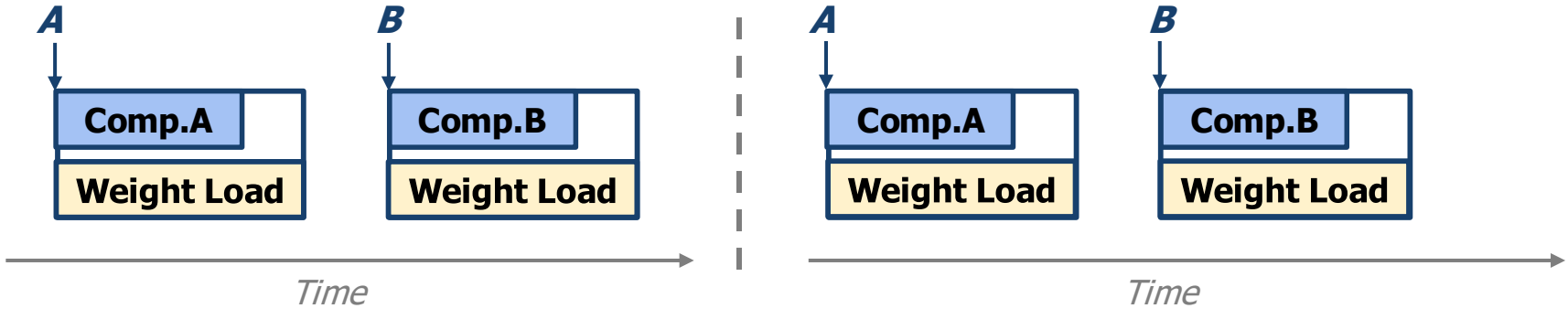
Energy = Power x Time





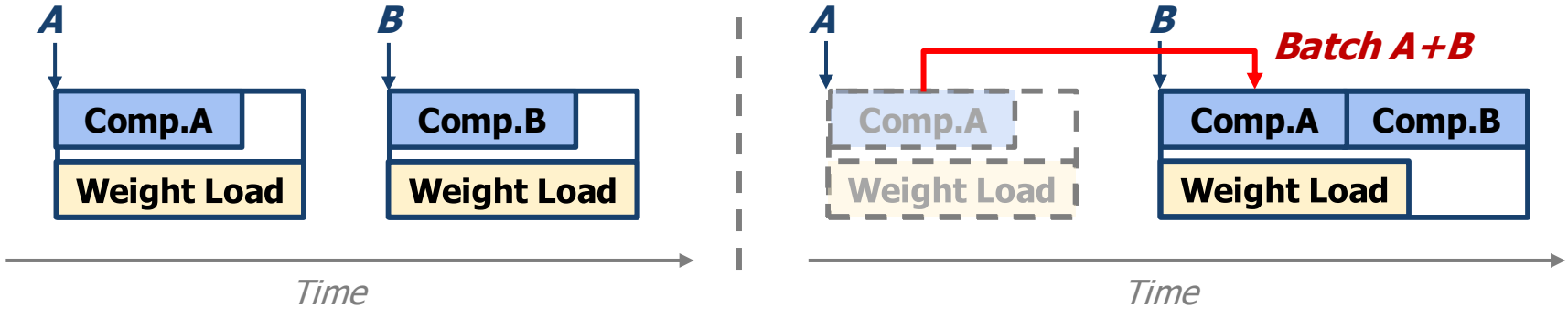
Workload balance x <u>DVFS</u> defines Perf-Energy landscape

Opportunity
Batching



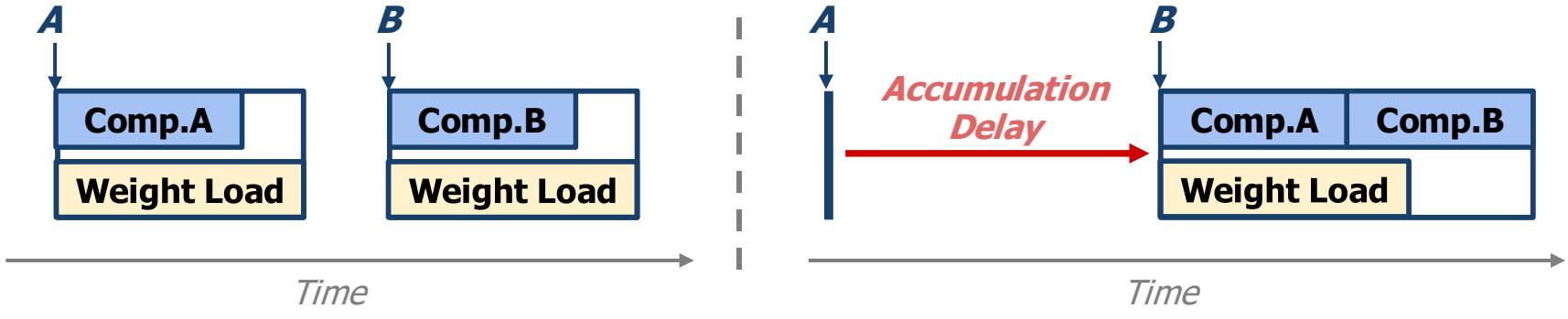
Workload balance x DVFS defines Perf-Energy landscape

Opportunity
Batching



Workload balance x DVFS defines Perf-Energy landscape

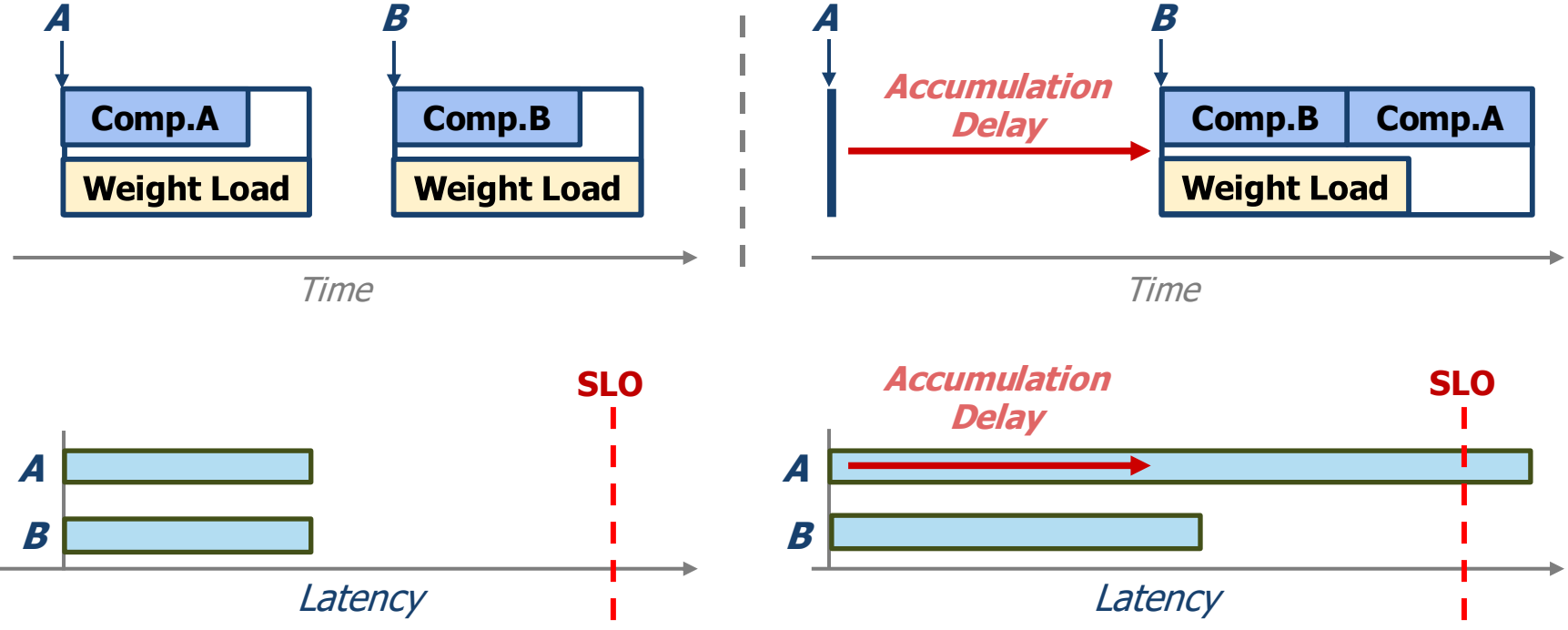
Opportunity
Batching



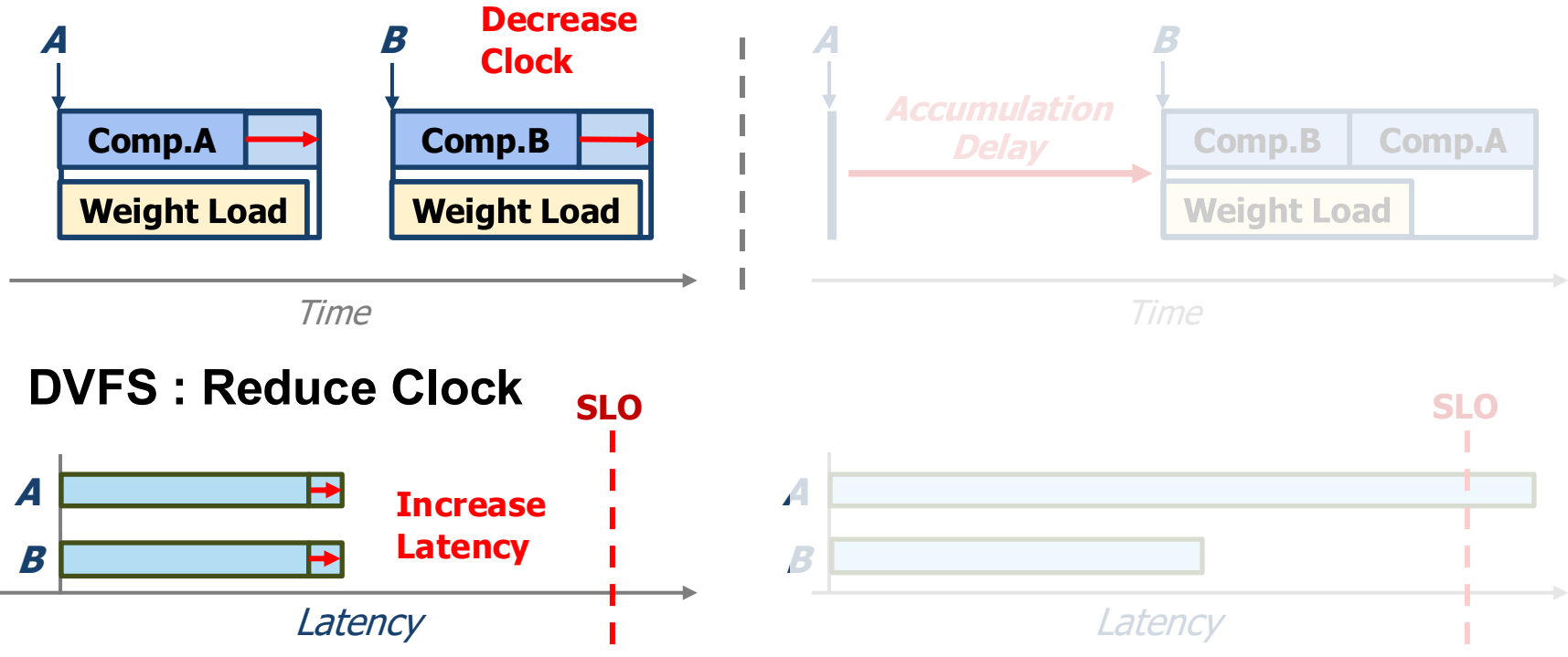
Energy Comp.A + Comp.B **+ 2 x Weight**

Energy Comp.A + Comp.B **+ Weight**

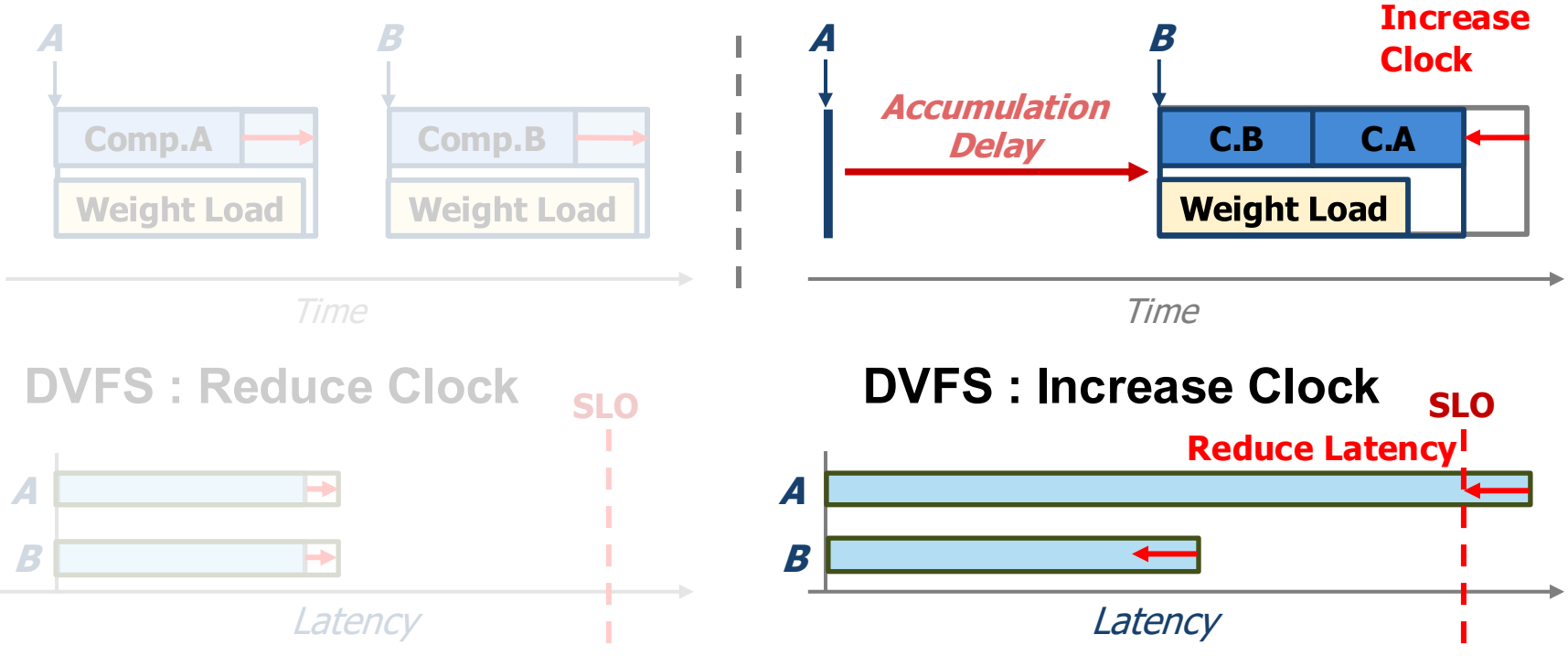
Opportunity
Batching



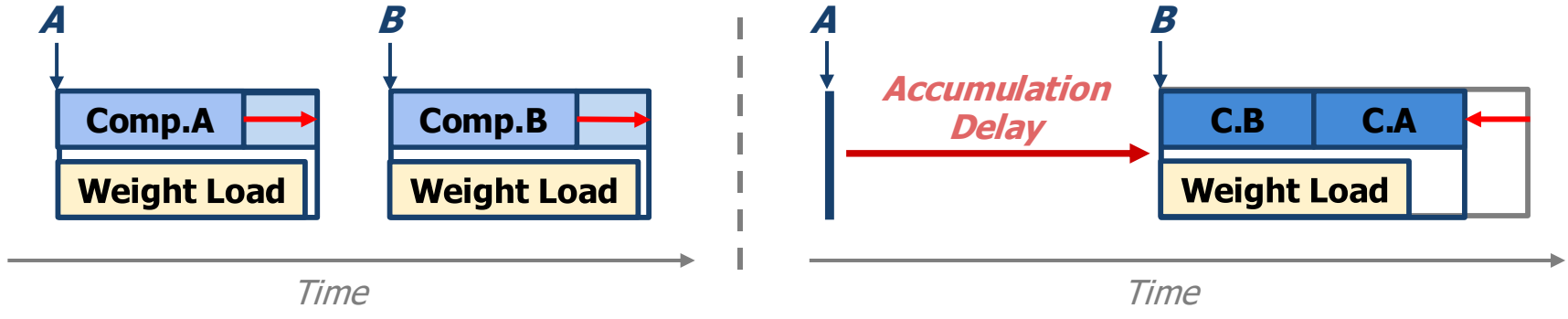
Opportunity
Batching x DVFS

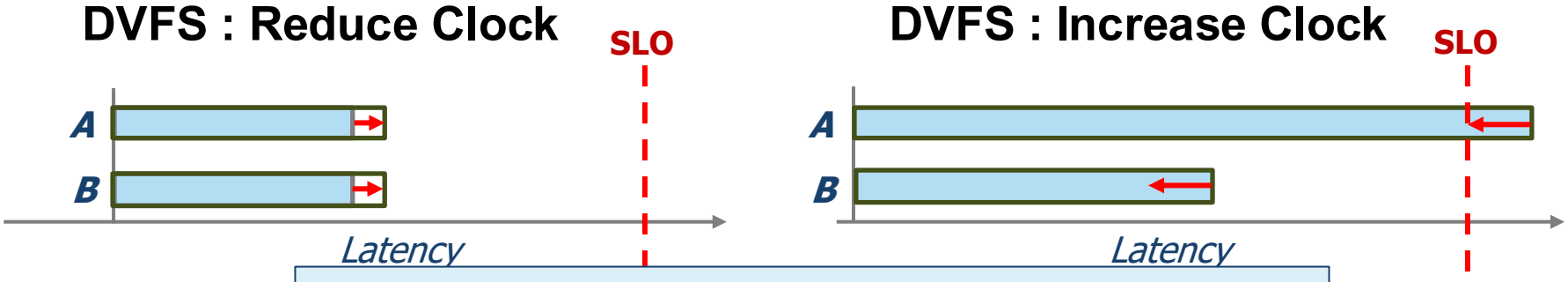


Opportunity
Batching x DVFS



Opportunity
Batching x DVFS



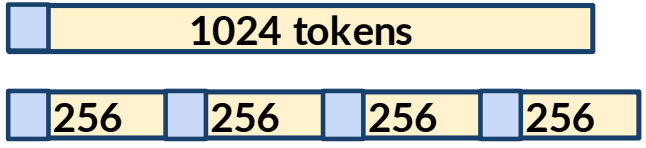


Batching and DVFS must be co-optimized

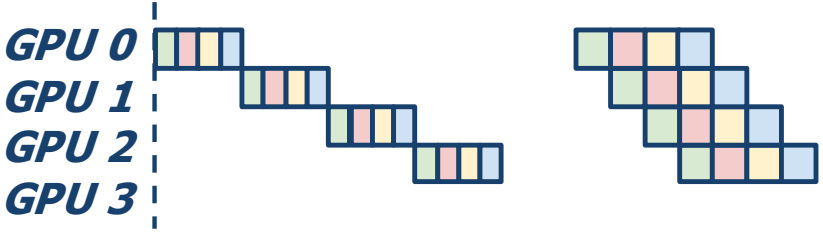
Opportunity

Batch-related control knobs

Chunked Prefill



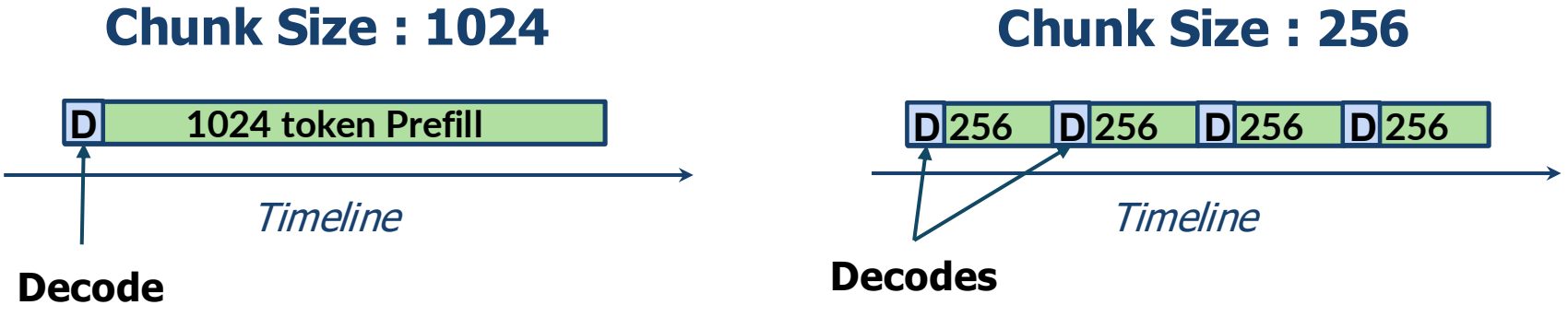
Micro Batching



Opportunity

Control Knob 1 : Chunked Prefill

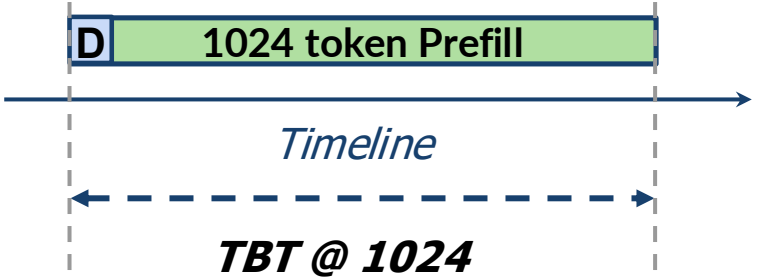
Processing Prefill of Size 1024



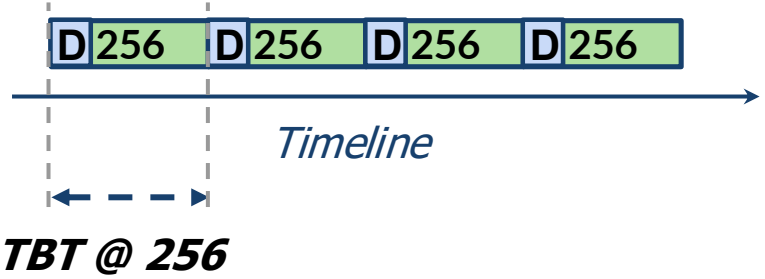
Opportunity

Control Knob 1 : Chunked Prefill

Chunk Size : 1024



Chunk Size : 256

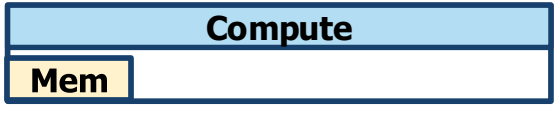


- **Lower TBT**

Opportunity

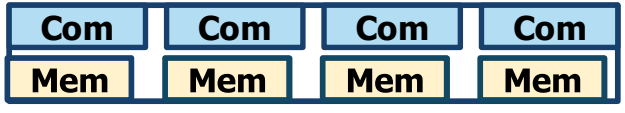
Control Knob 1 : Chunked Prefill

Chunk Size : 1024



Workload Visualization

Chunk Size : 256



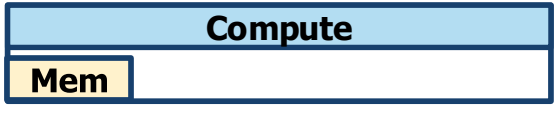
Workload Visualization

- **Lower TBT**

Opportunity

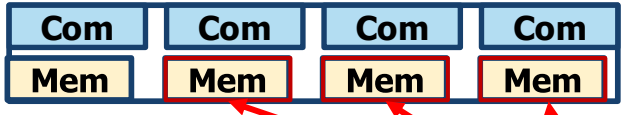
Control Knob 1 : Chunked Prefill

Chunk Size : 1024



Workload Visualization

Chunk Size : 256



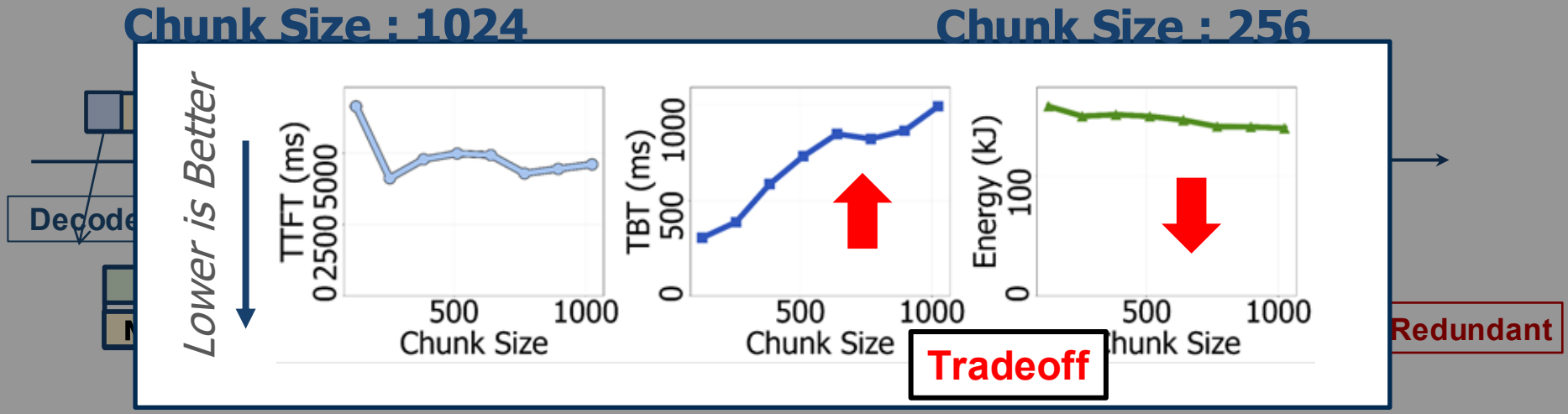
Workload Visualization

Redundant

- **Lower TBT**
- **More Energy Usage**

Opportunity

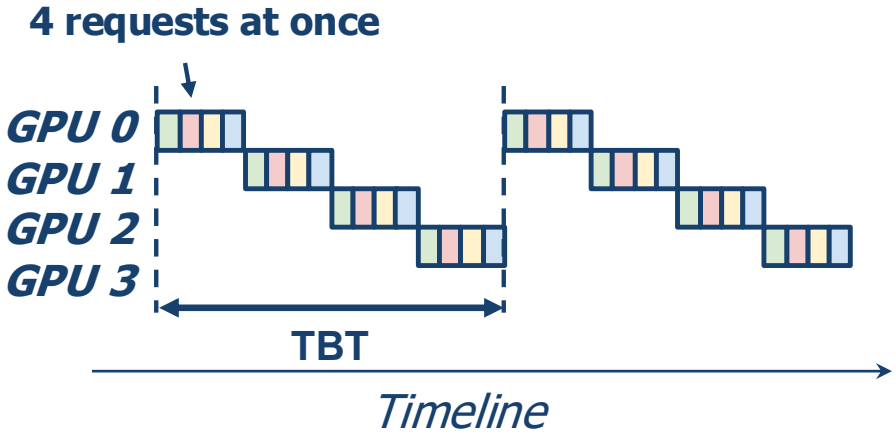
Control Knob 1 : Chunked Prefill



Chunked Prefill impacts performance-energy tradeoff

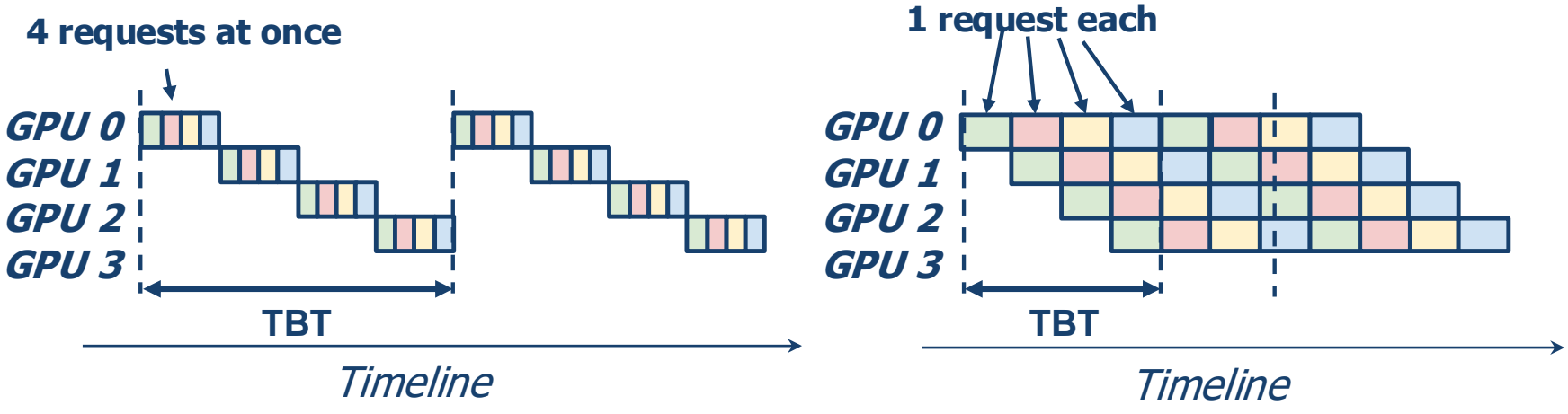
Opportunity

Control Knob 2 : Pipeline Parallel Microbatch



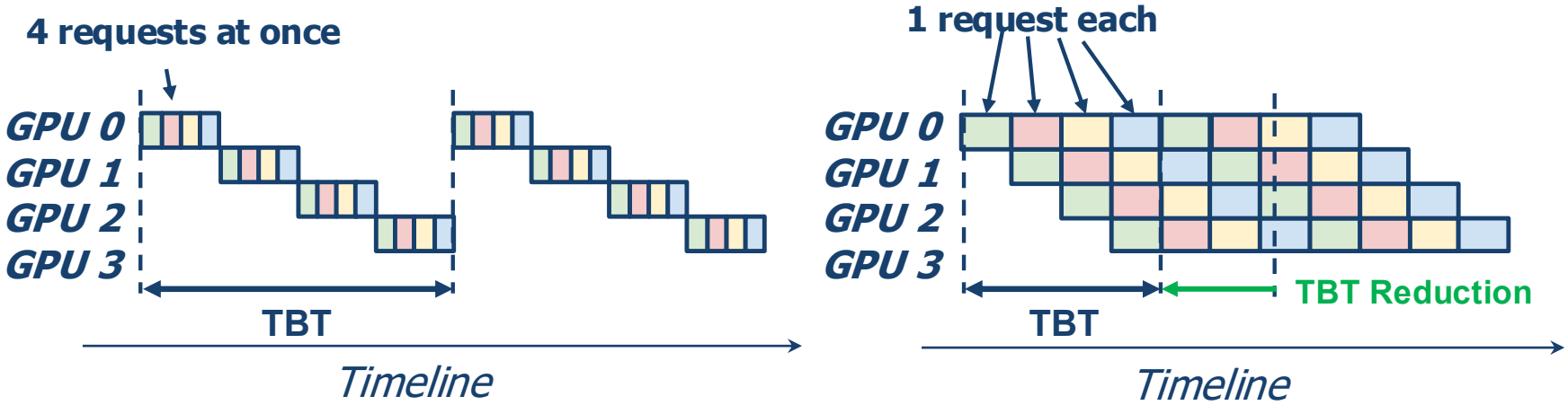
Opportunity

Control Knob 2 : Pipeline Parallel Microbatch



Opportunity

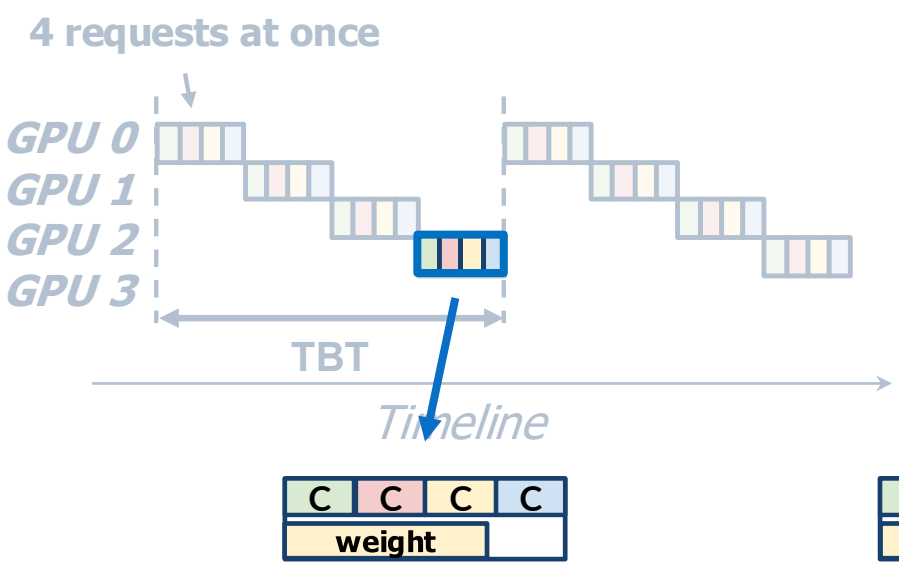
Control Knob 2 : Pipeline Parallel Microbatch



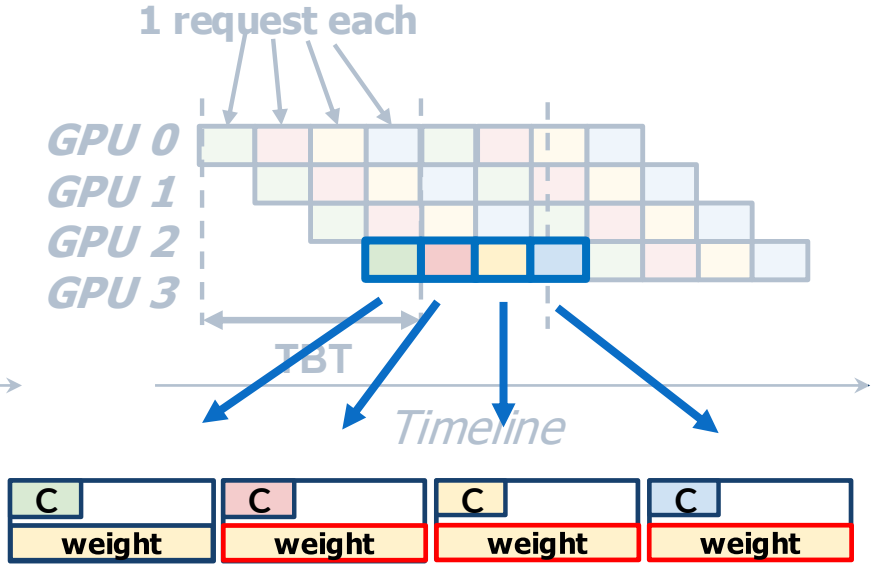
- **Higher Utilization**
- **Better Performance**

Opportunity

Control Knob 2 : Pipeline Parallel Microbatch



Workload Visualization

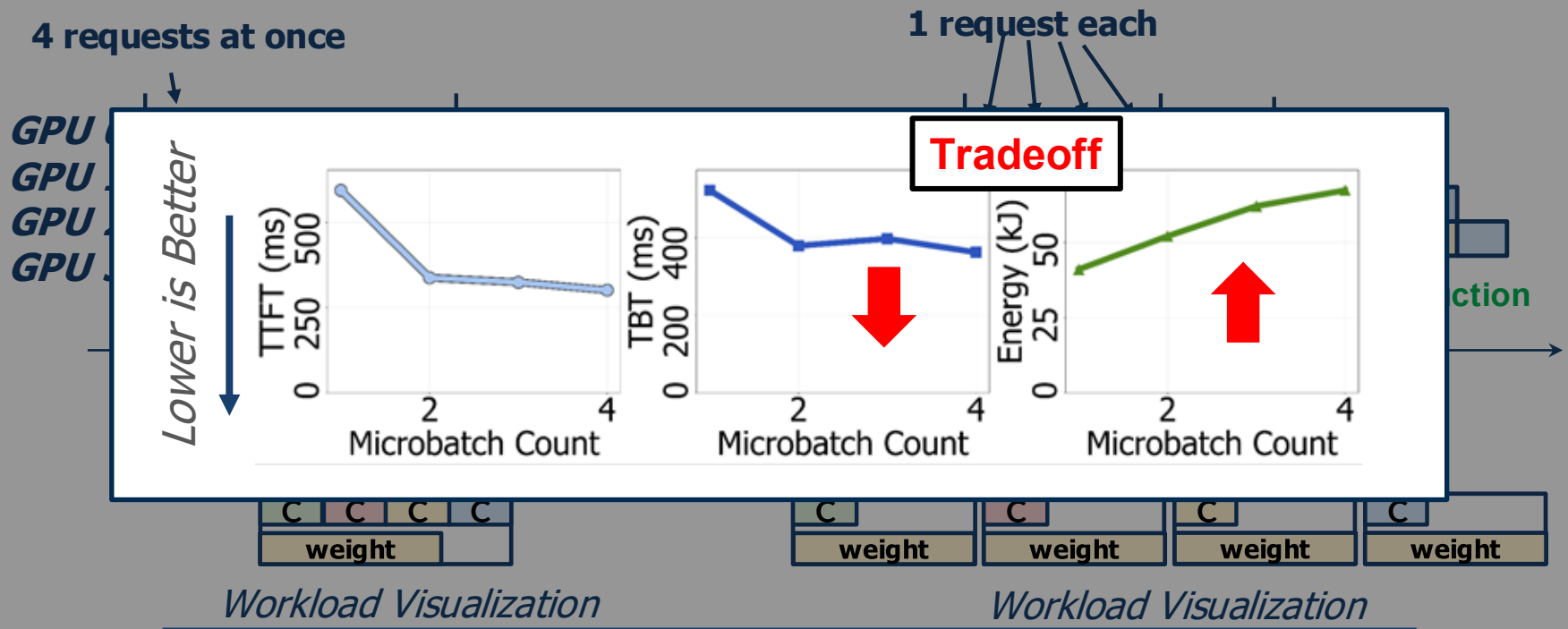


Workload Visualization

- **Higher Utilization**
- **Better Performance**
- **More Energy Usage**

Opportunity

Control Knob 2 : Pipeline Parallel Microbatch



Micro batching impacts performance-energy landscape

BEAM : Batch Energy Aware Manager

Overview

Goal : Minimize energy usage under SLO constraints

Control Knobs

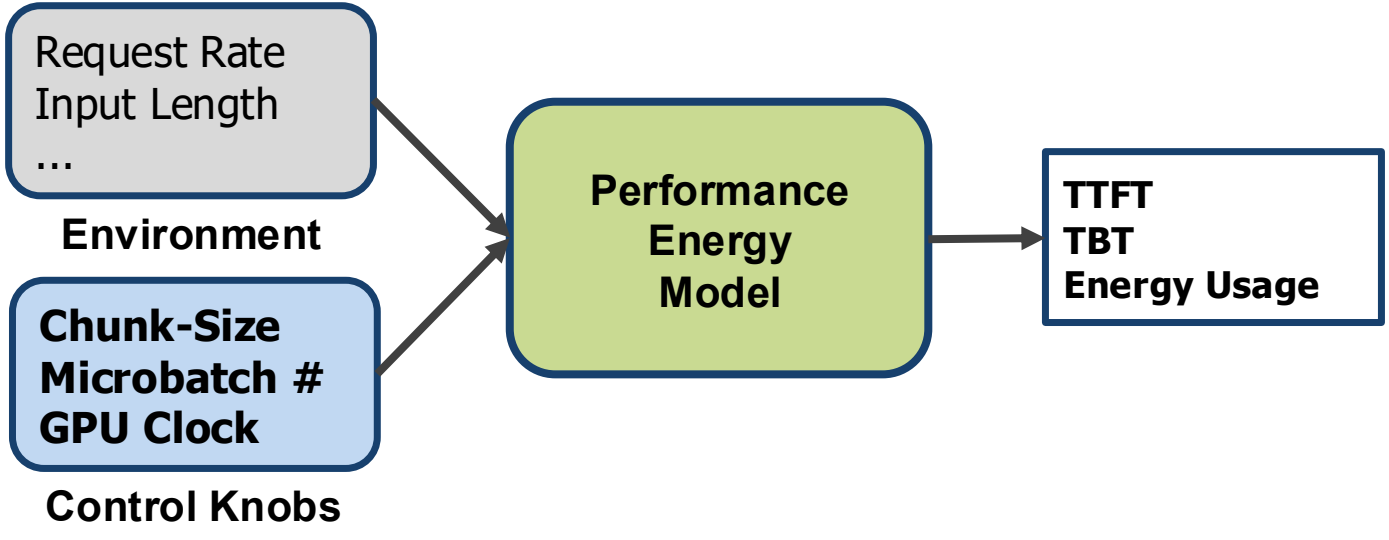
Chunk-Size	Microbatch #	GPU Clock	*Co-optimize*

BEAM : Batch Energy Aware Manager

Overview

Goal : Minimize energy usage under SLO constraints

Challenge : Navigate Performance-Energy Landscape

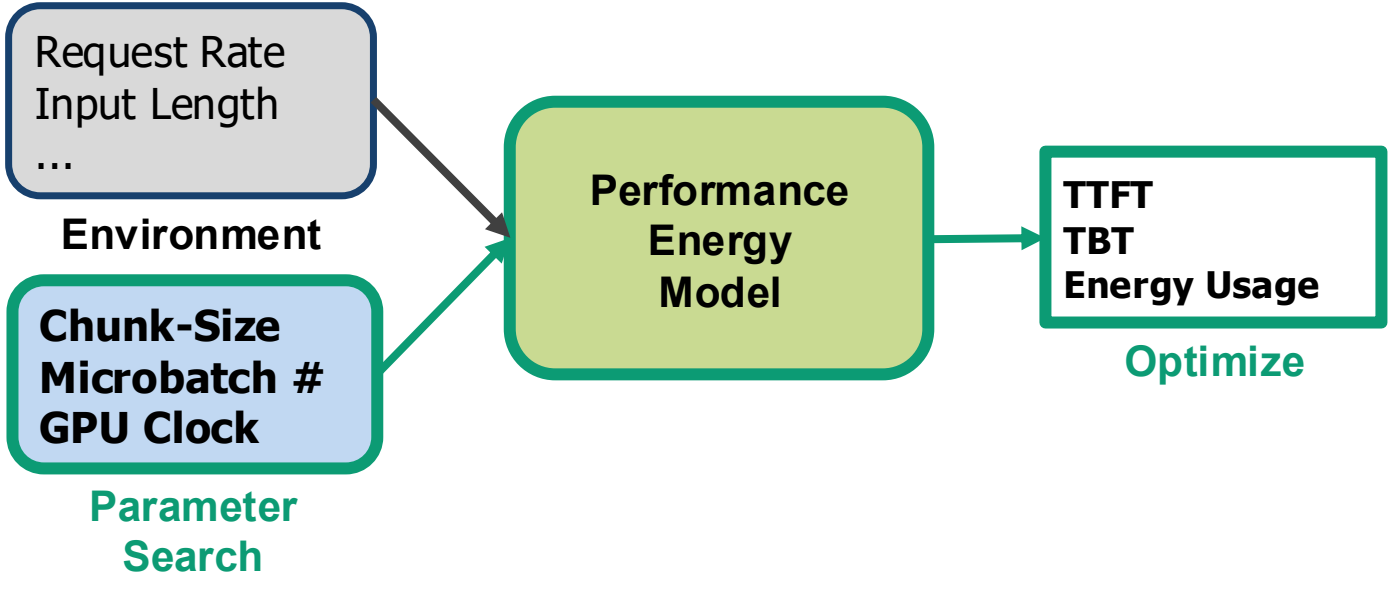


BEAM : Batch Energy Aware Manager

Overview

Goal : Minimize energy usage under SLO constraints

Challenge : Navigate Performance-Energy Landscape



BEAM : Batch Energy Aware Manager

Overview

Goal : Minimize energy usage under SLO constraints

Control Knobs

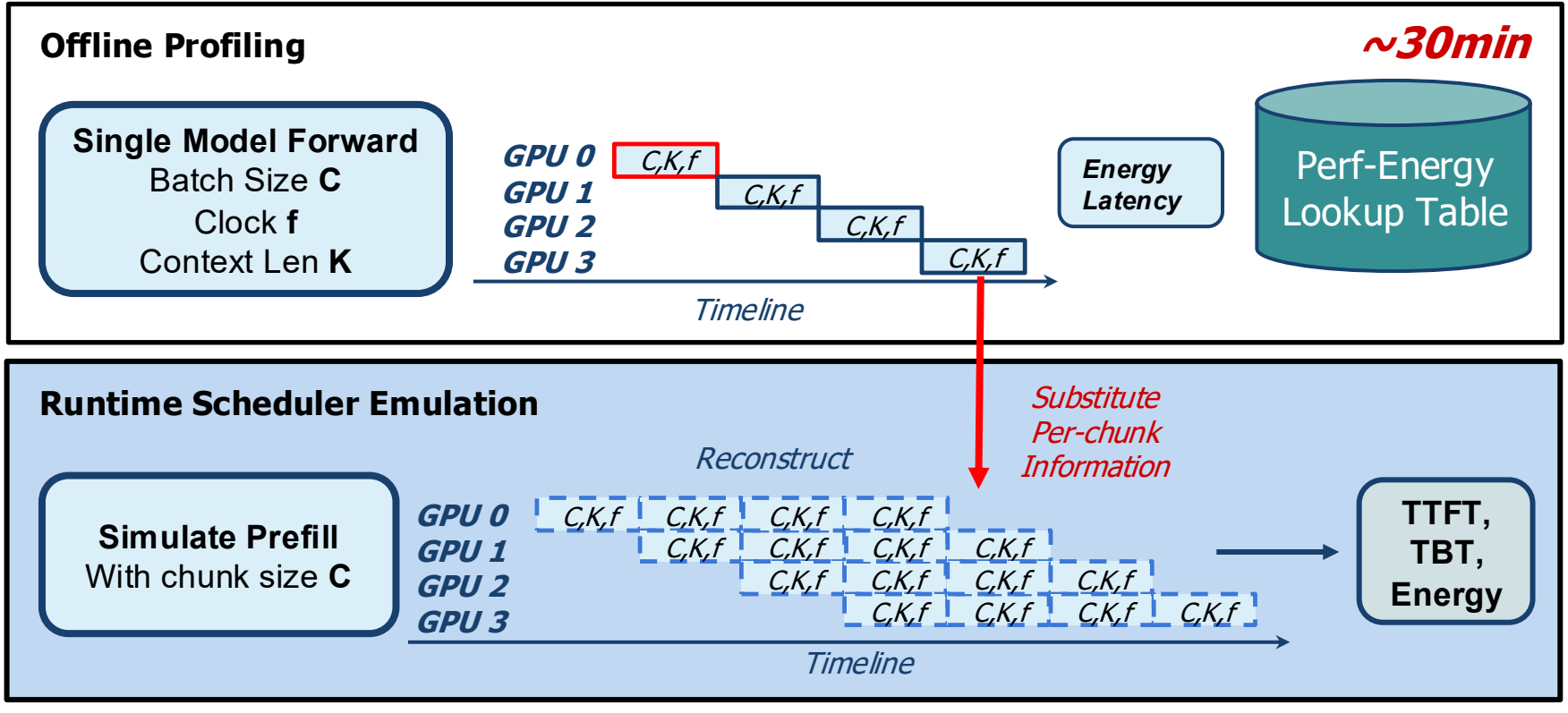
Chunk-Size | **Microbatch #** | **GPU Clock** *Co-optimize*

Design Principles

P1 : Component Based Modelling with Emulation

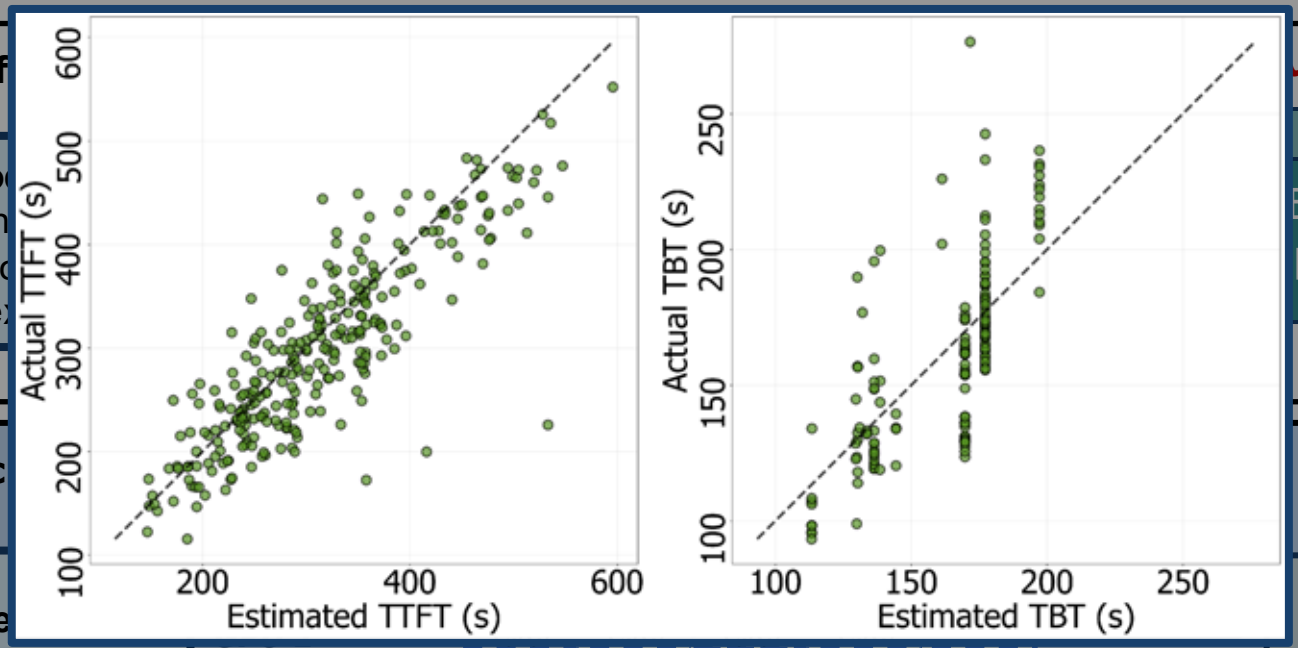
Design

P1 : Component-Based Modelling



Design

P1 : Component-Based Modelling



Offline Prof

~30min

Single Mo
Batch
Clo
Contex

Energy
p Table

Runtime Sc

Simulate
With chunk size **C**

C,K,f C,K,f C,K,f C,K,f

TTFT,
TBT,
Energy

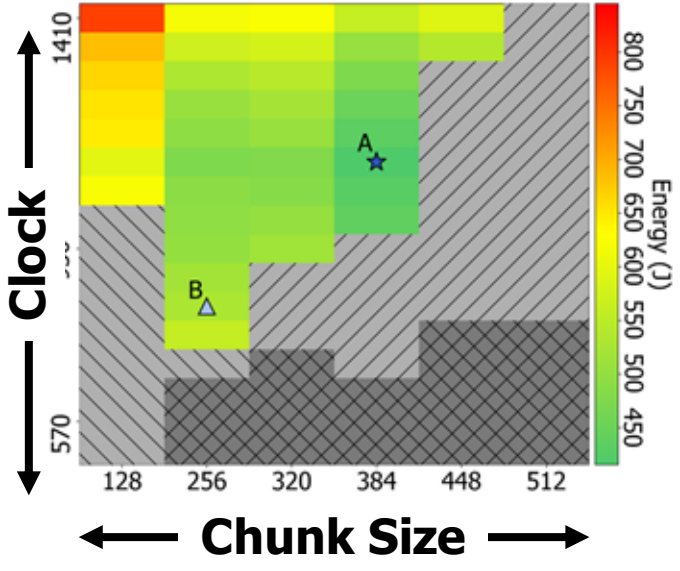
TTFT MAPE ~26%, TBT MAPE ~15%.

KAIST

CASYS | KAIST Computer Architecture & System Lab

Design
Co-optimization Method

Search Space

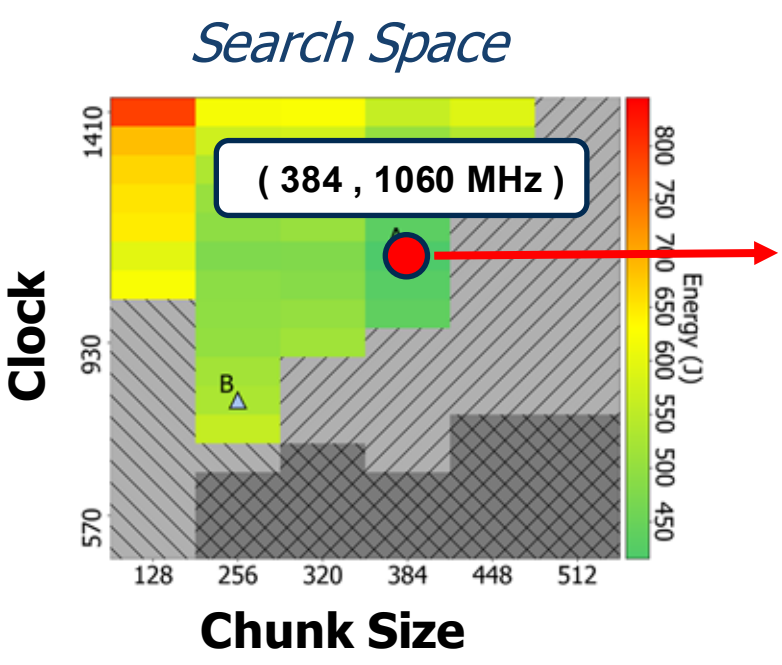


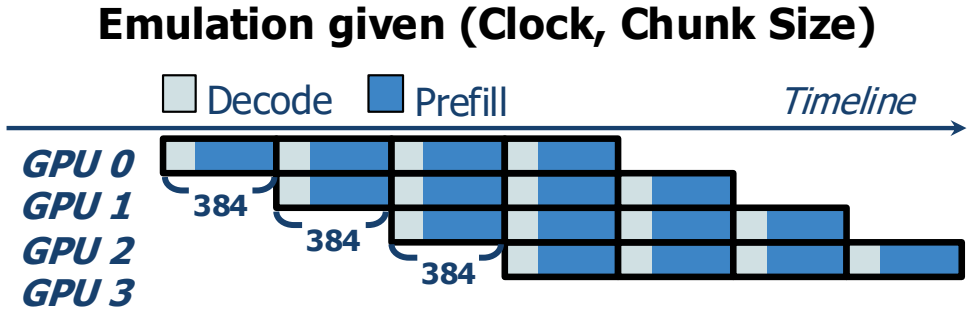
Event
- New prefill arrives (1500 tokens)

Objective
- Find **Clock x Chunk Size**
- Adhering to **SLO**
- Yielding **Minimum Energy**

Design

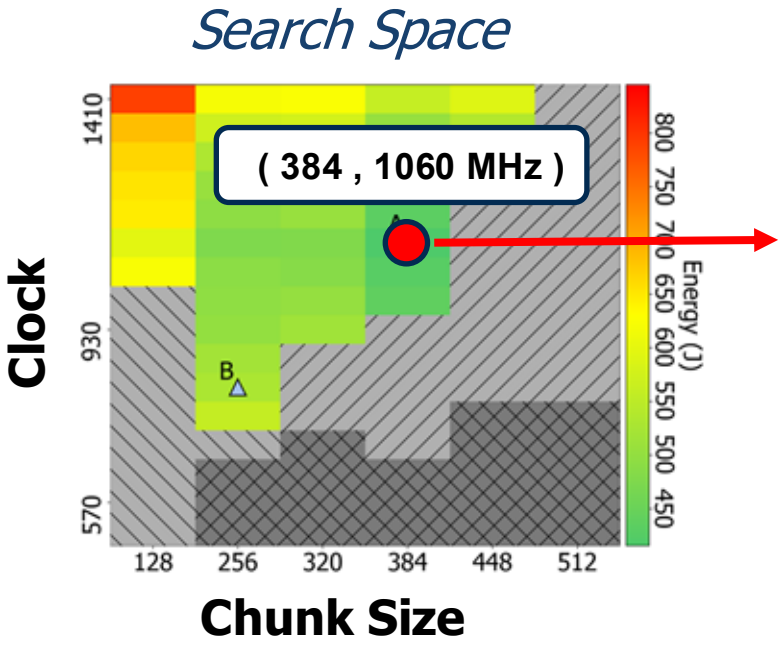
Co-optimization Method

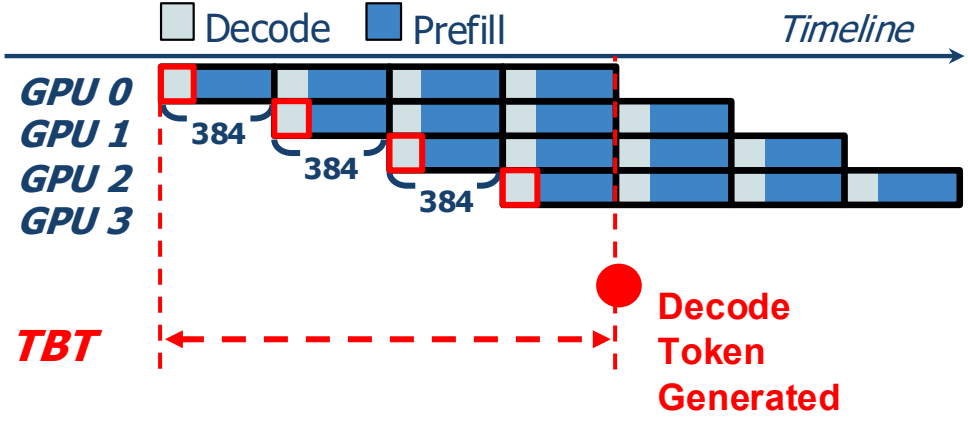




Design

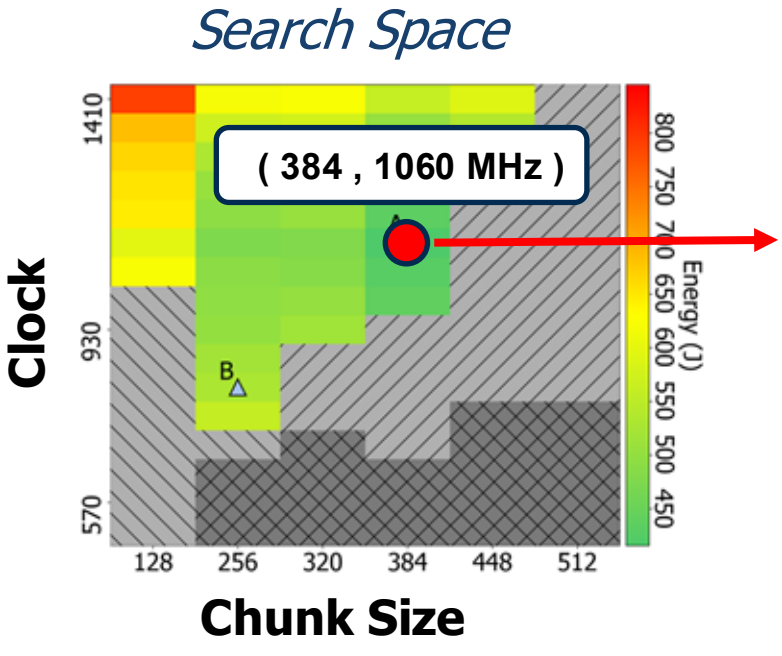
Co-optimization Method

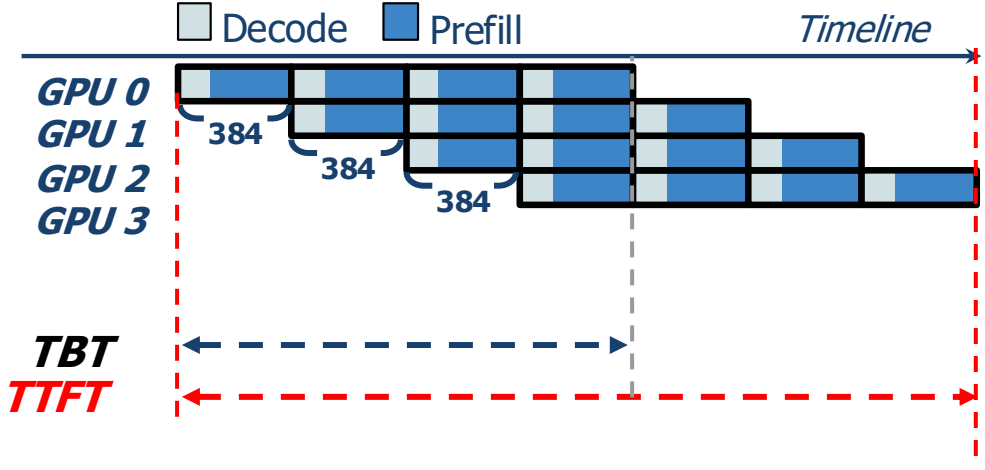




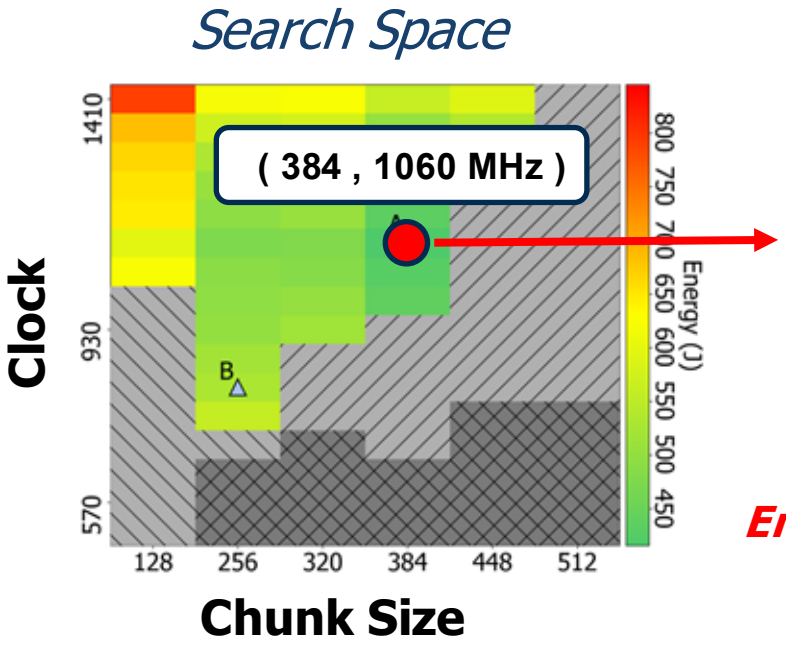
Design

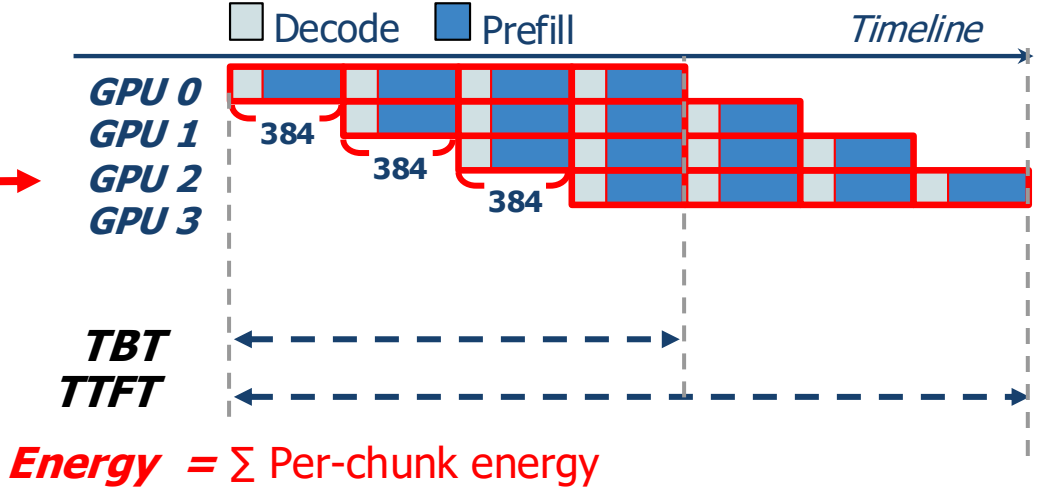
Co-optimization Method





Design
Co-optimization Method

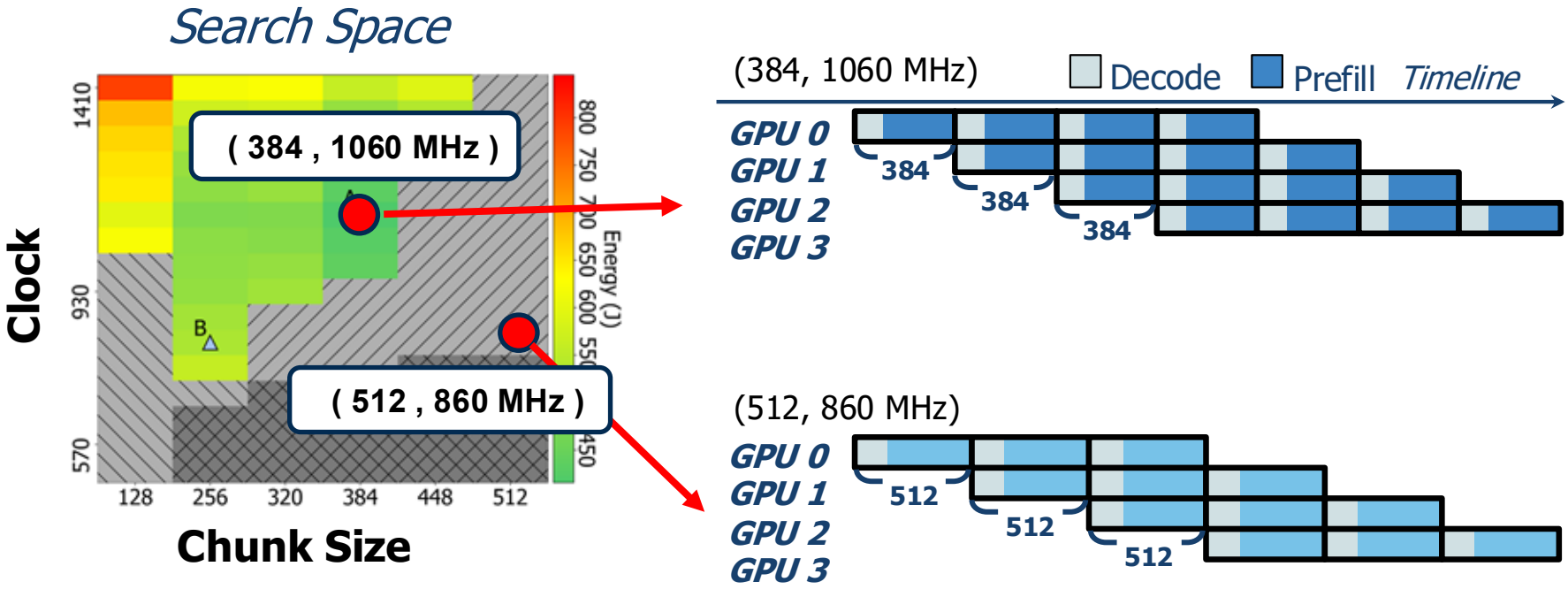




Energy = ∑ Per-chunk energy

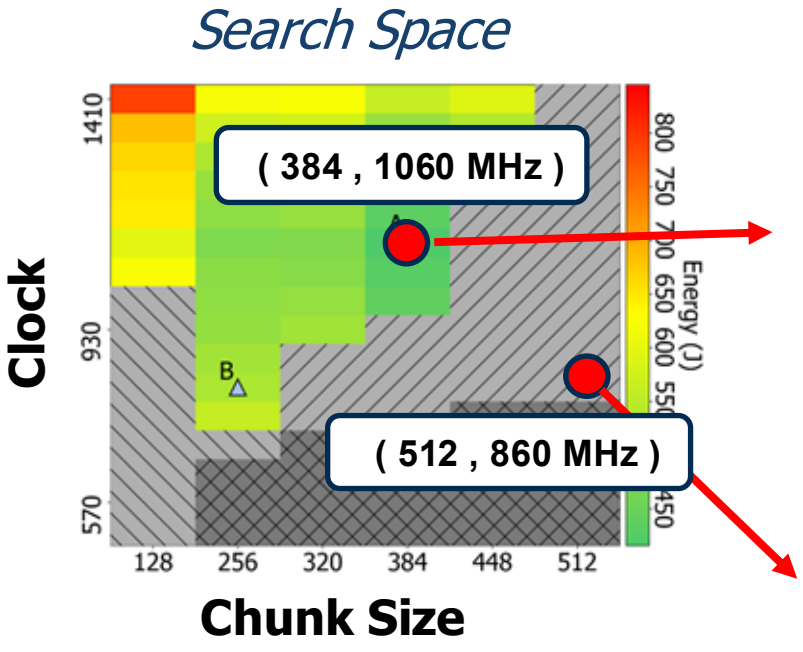
Design

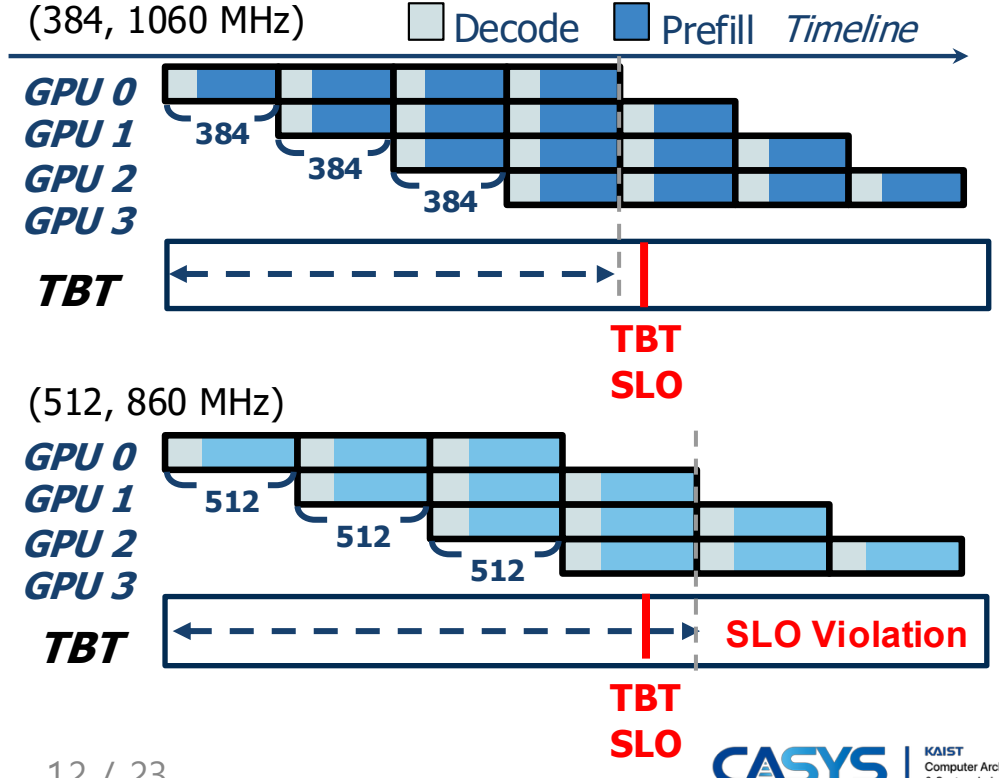
Co-optimization Method



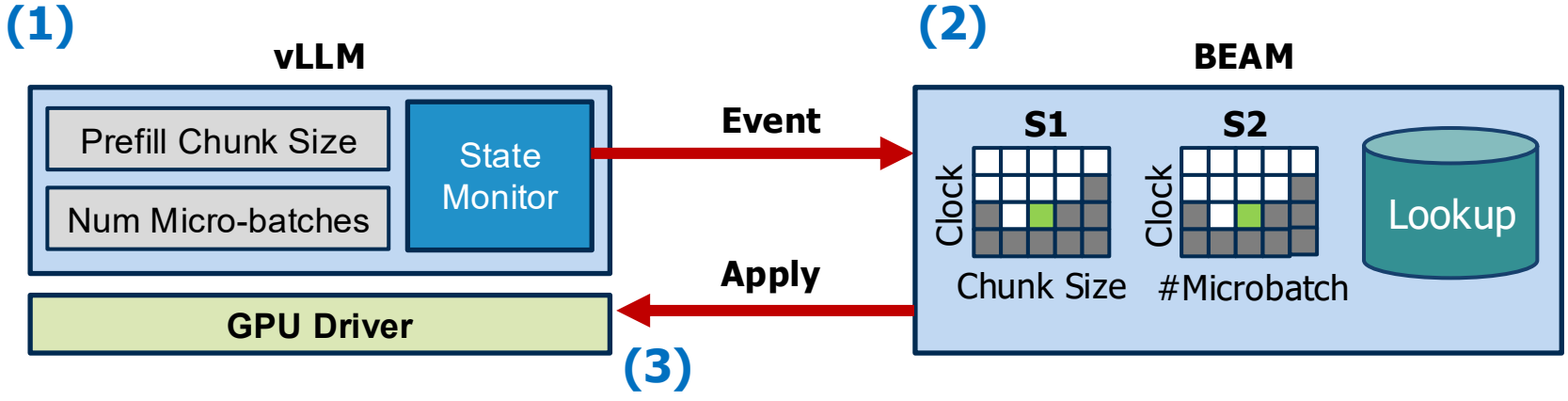
Design

Co-optimization Method





Design
System Overview



(1)State Monitor Hook
(2)Search and Optimization
(3)Knob application

Evaluation
Evaluation Overview

Evaluation Settings

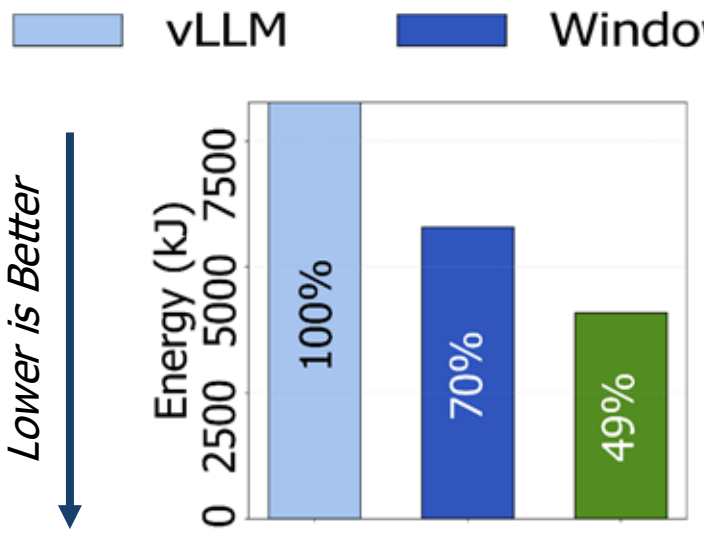
- **Hardware** :
  - DGX A100x8 Server
- **Model** :
  - LLama-3 70B Instruct PP4xTP2
- **Serving Trace**
  - Alibaba ServeGen Trace (E2E)
  - Synthetic (Others)
- **SLO** :
  - TTFT 1s, TBT 0.2s

Baselines

- Vanilla vLLM
- Window-DVFS
- BEAM

Evaluation

End-to-end Analysis – Energy Usage



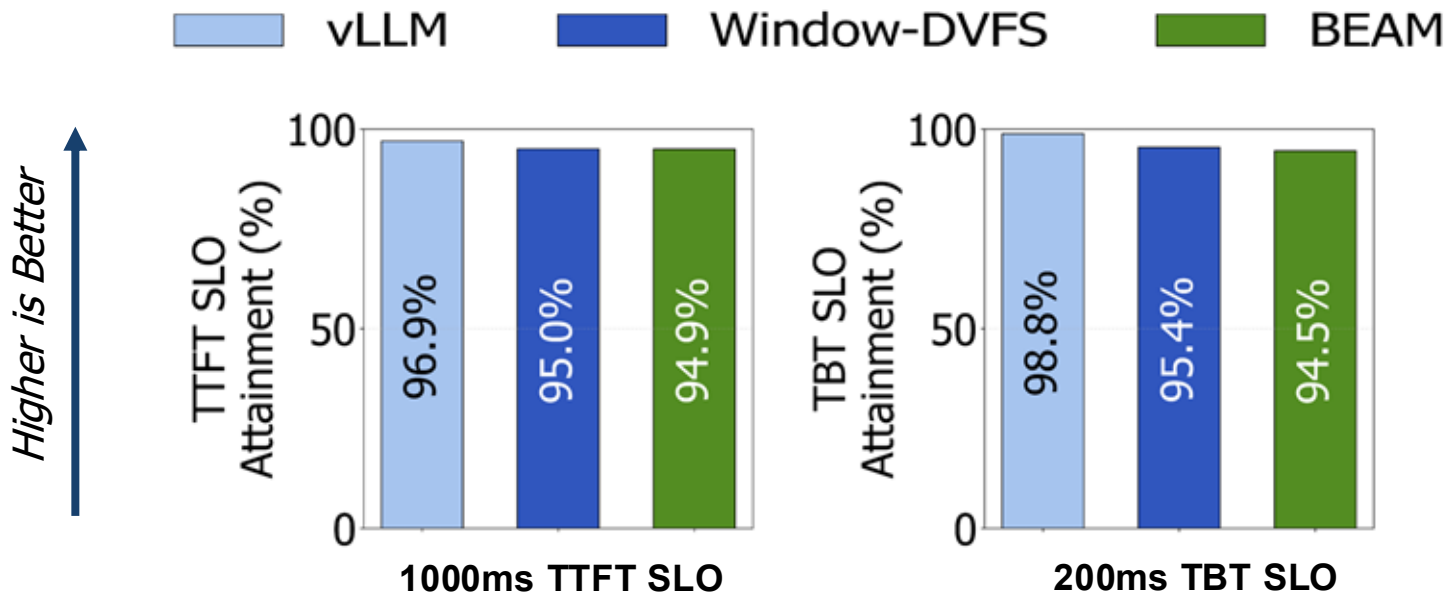
Experiment Settings

- 1h on ServeGen (19:00 ~ 20:00)
- Mean 1800 TPS
- P90 2800 TPS

- **Beam uses 70% energy compared to SOTA**

Evaluation

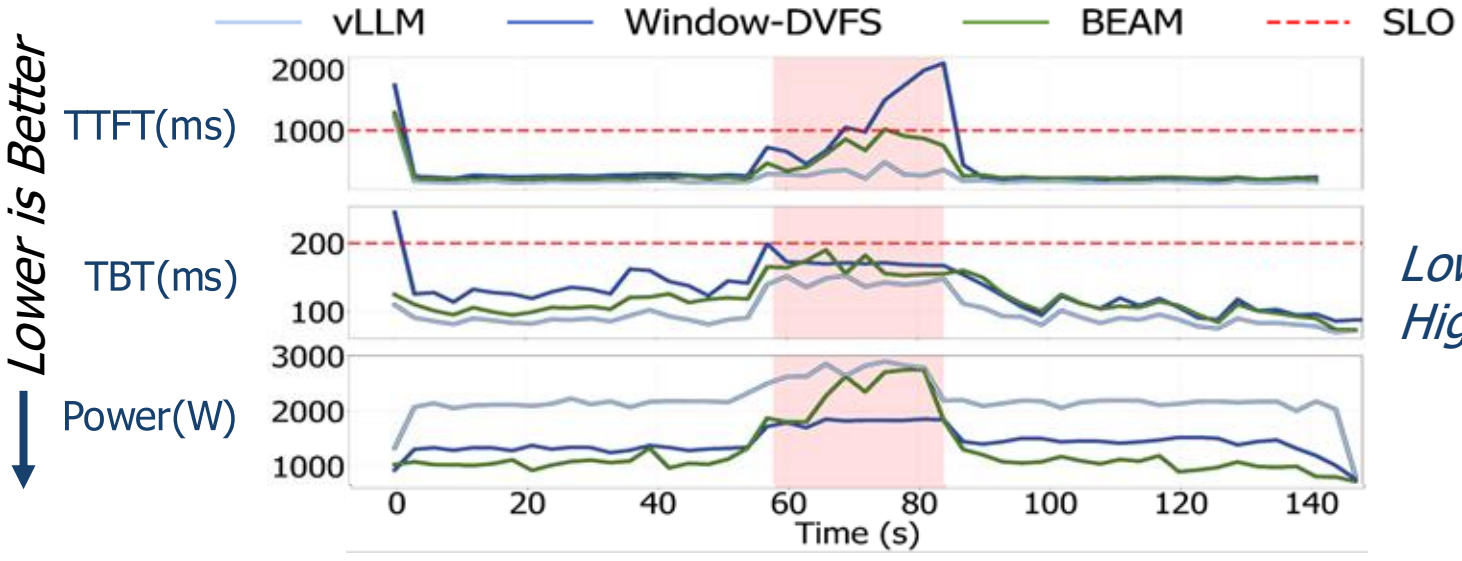
End-to-end Analysis – SLO Adherence



- **BEAM Adheres to TTFT / TBT SLO**

Evaluation

Sensitivity to burst



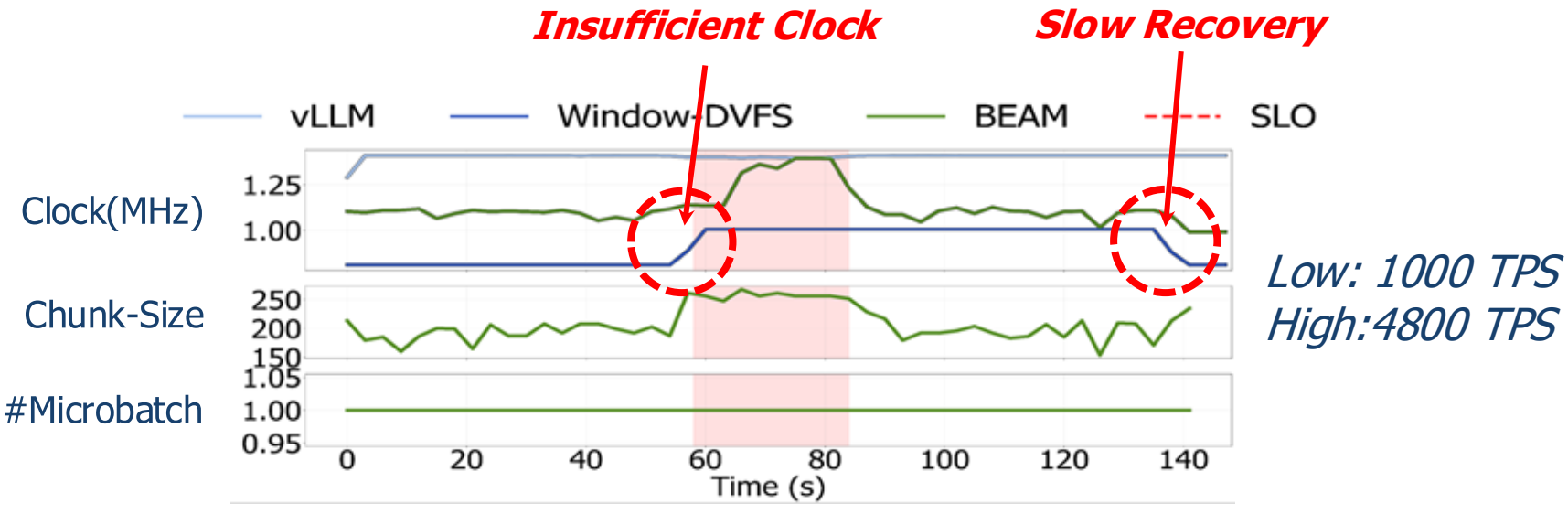
Low: 1000 TPS
High:4800 TPS

Low Energy

- **BEAM rapidly reacts to burst scenarios**

Evaluation

Sensitivity to burst : Control knob view



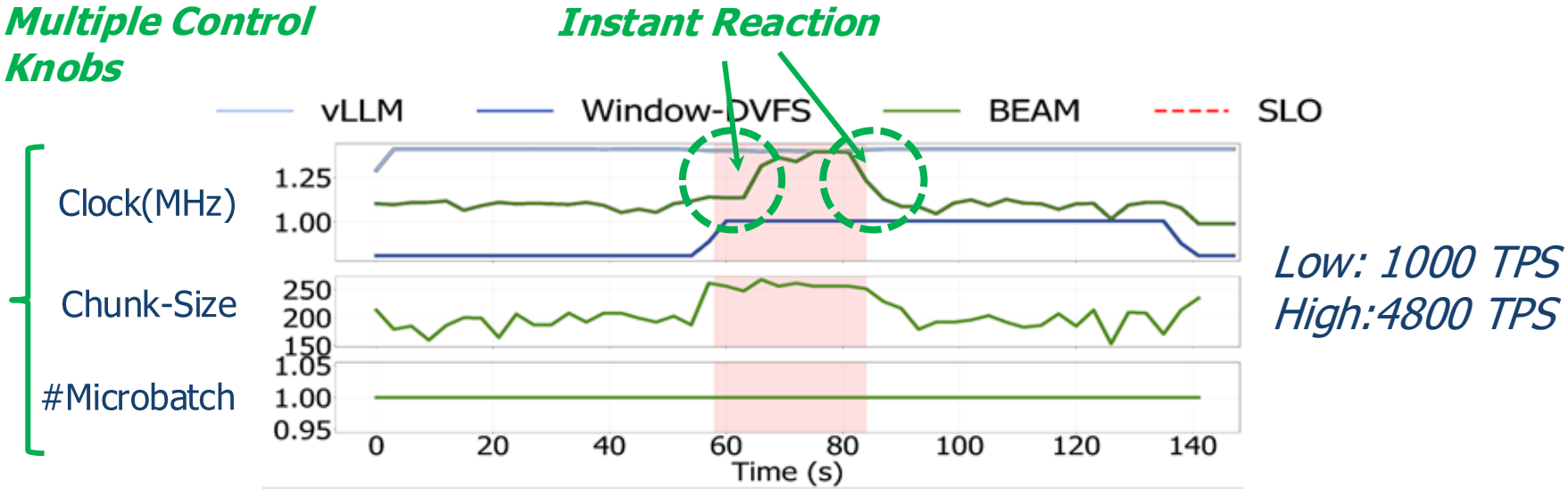
- **BEAM rapidly reacts to burst scenarios**

Evaluation

Sensitivity to burst : Control knob view

Multiple Control Knobs

Instant Reaction



Low: 1000 TPS
High:4800 TPS

- **BEAM rapidly reacts to burst scenarios**

Conclusion

Goal

- Minimize energy under SLO constraints

Method

- Co-optimizes **DVFS** with **Batching**
- By scheduling **Chunk-Size**, **#Microbatch, Clock**
- **Scheduler emulation** to navigate complex landscape

Results

- Energy usage **50% of vanilla** vLLM, **70% of SOTA**

Thank You

Paper PDF

Artifact Github

Contact : hjlee@casys.kaist.ac.kr