Optimizing Energy with Performance in Mind

Ruofan Wu

December 2nd, 2025

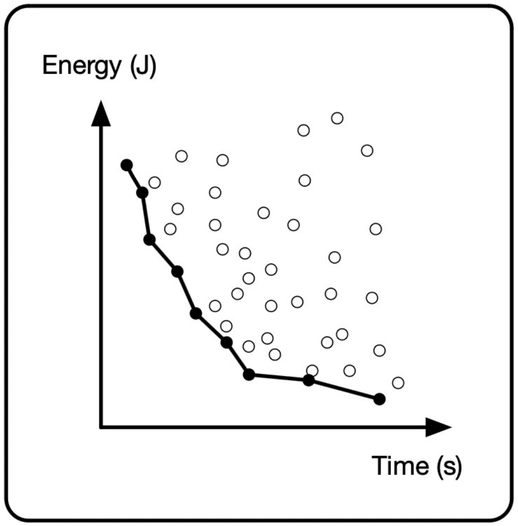
SymbioticLab
ML.ENERGY
UNIVERSITY OF MICHIGAN

Energy Optimization for AI

Principles

- The time–energy trade-off frontier is a key object for reasoning.

    - Same computation
    - Different ways



Overview of Existing work

Serving

DynamoLLM (HPCA '25)

The ML.ENERGY Benchmark (NeurIPS '25 D&B)

Training

Zeus (NSDI '23)

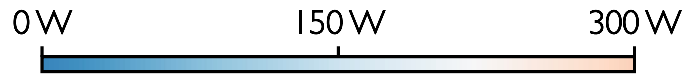
Perseus (SOSP '24)

DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency

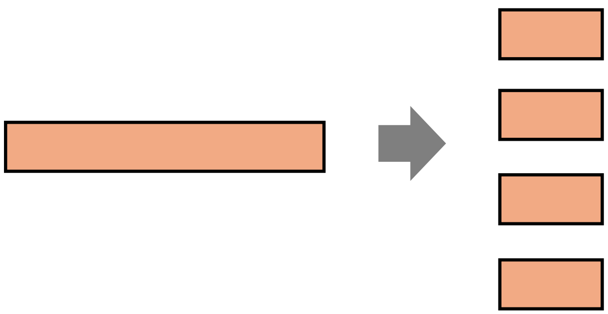
Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Josep Torrellas, Esha Choukse

How to minimize energy consumption given latency deadlines?

Time vs. Energy Trade-off



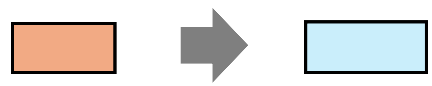
Model parallelism



Time: 0.8s ⟶ 0.25s

Energy: 200J ⟶ 250J

GPU frequency



Time: 0.25s ⟶ 0.27s

Energy: 75J ⟶ 60J

LLM Inference

Prefill

What is DynamoLLM ?

Compute-bound

Decode

It's a dynamic energy-management system

Memory-bound

Model parallelism

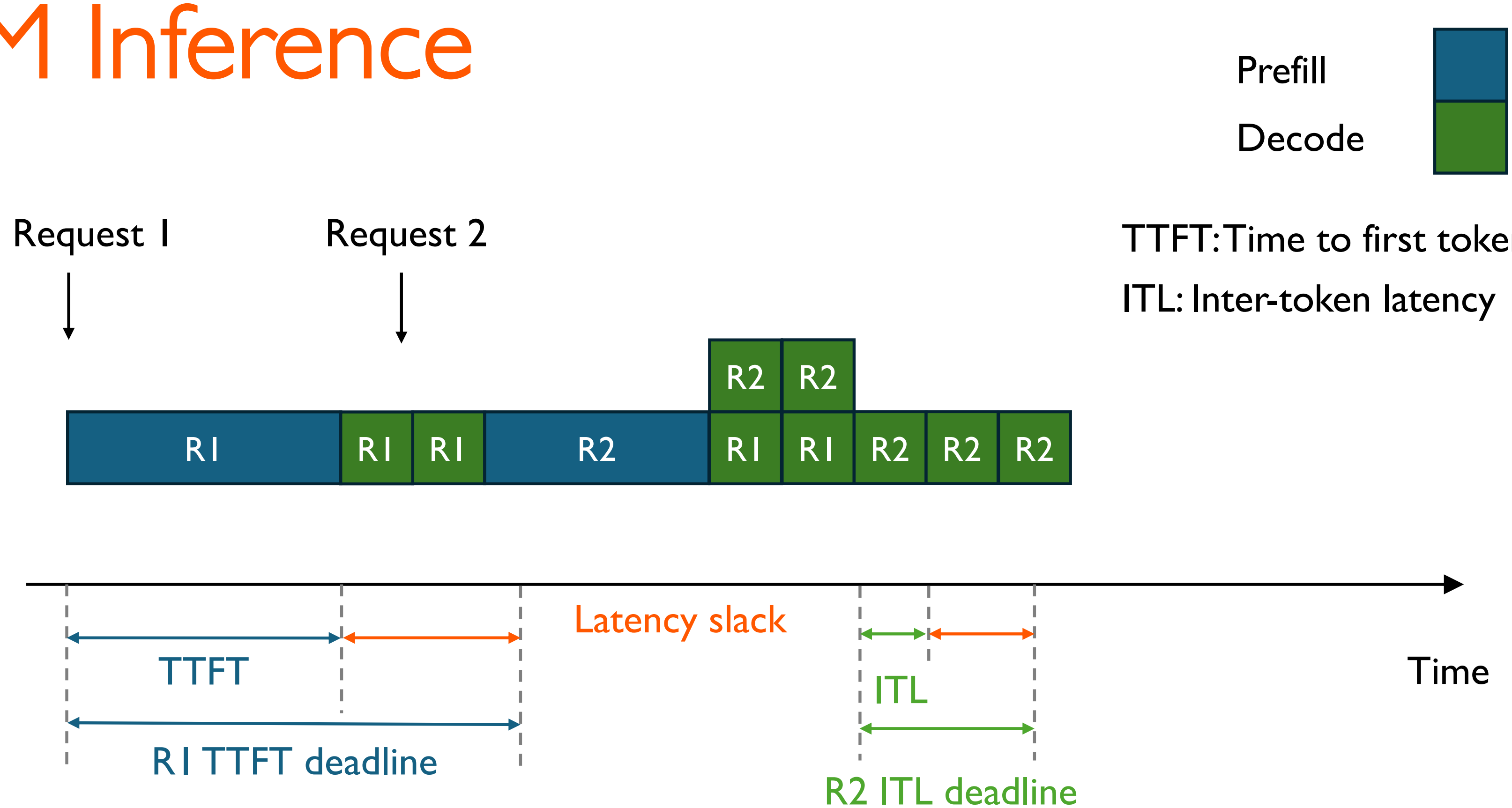
GPU frequency

Time sensitive

Time insensitive

DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency, HPCA, 2025

LLM Inference



TTFT: Time to first token

ITL: Inter-token latency

Heterogeneous Request Behavior

What is DynamoLLM?

It's a dynamic energy-management system designed for large-scale LLM inference clusters. It observes that different inference requests have vastly different compute and energy characteristics.

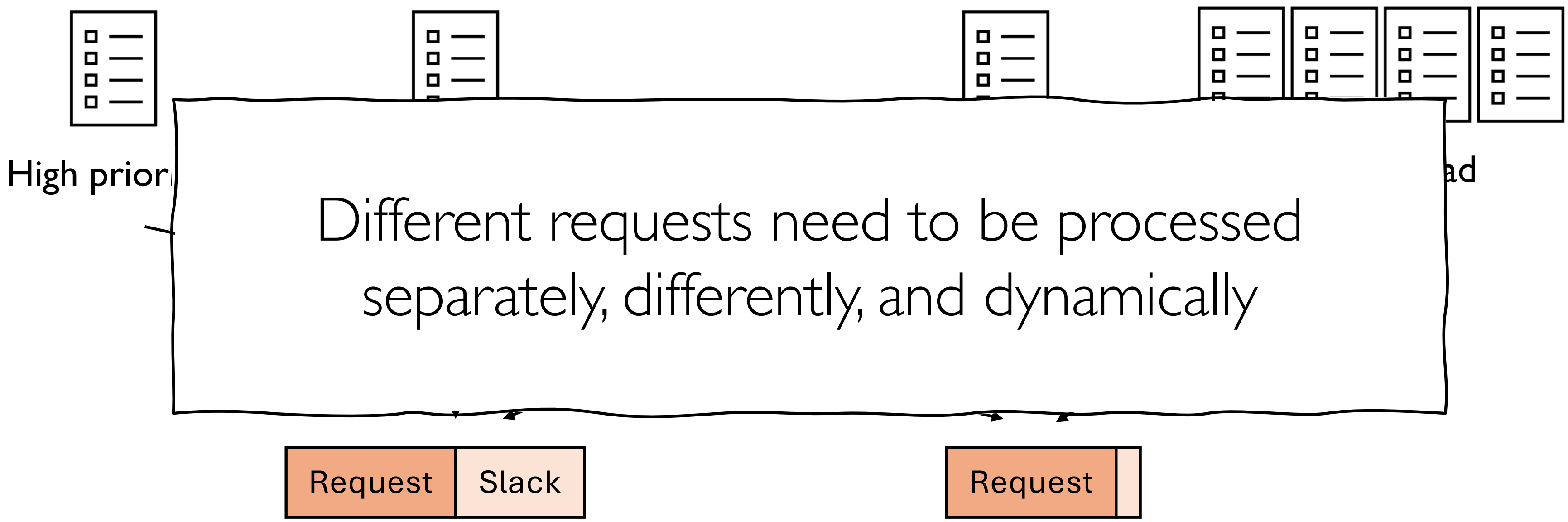
Memory-bound

I have 5 apples. I gave 2 to my friend and then bought 4 more. After that, I ate 3. How many apples do I have now?

4

Compute-bound

Heterogeneous Request Behavior



Different requests need to be processed separately, differently, and dynamically

Large slack

Small slack

Hierarchical Control

Request length
Request load
Time–energy frontier

Level	Decision	Time scale
Cluster	Number of instances	Minutes
Pool	Model parallelism	Minute
Instance	GPU frequency	Seconds

Up to 53% energy reduction
while meeting latency deadlines

DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency, HPCA, 2025

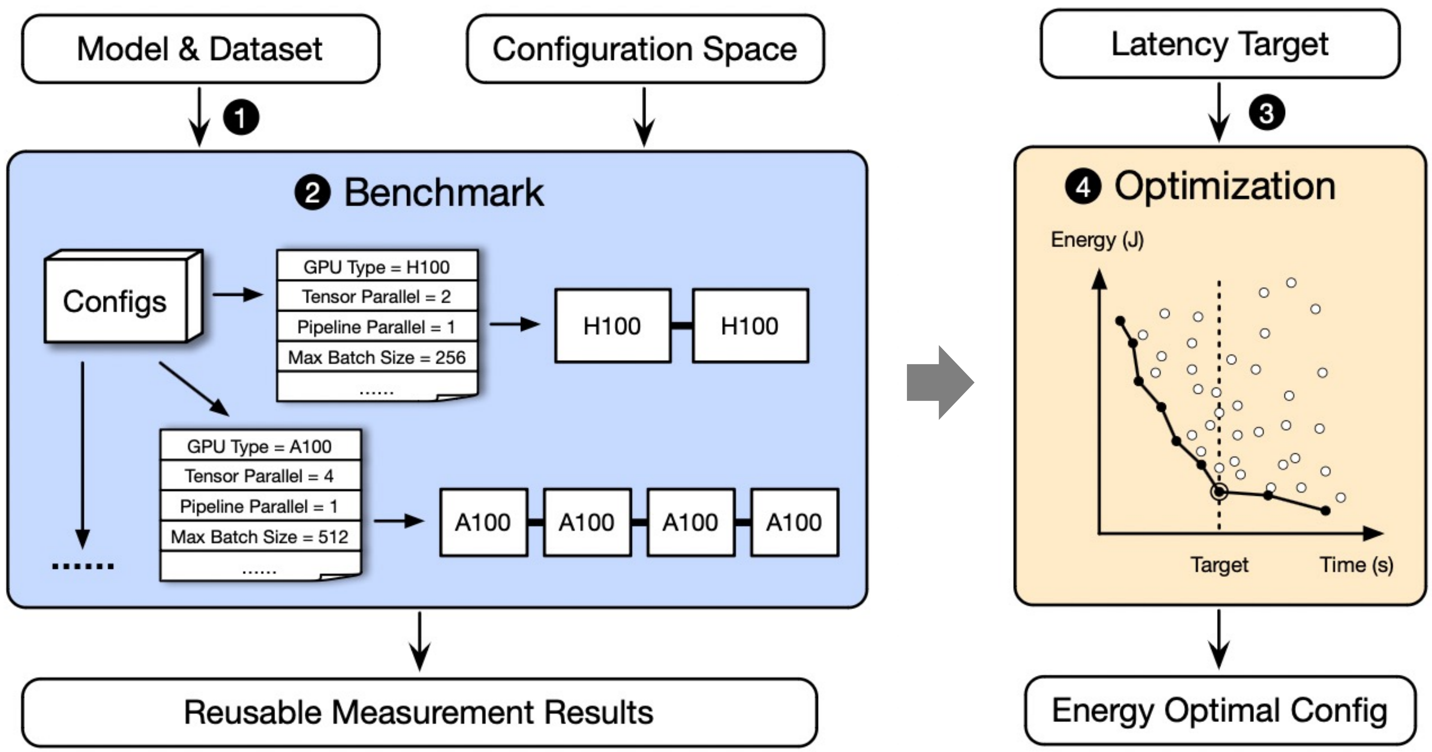
The ML.ENERGY Benchmark: Toward Automated Inference Energy Measurement and Optimization

https://ml.energy/leaderboard/

Jae-Won Chung, Jeff J. Ma, Ruofan Wu, Jiachen Liu,
Oh Jun Kweon, Yuxuan Xia, Zhiyu Wu, Mosharaf Chowdhury

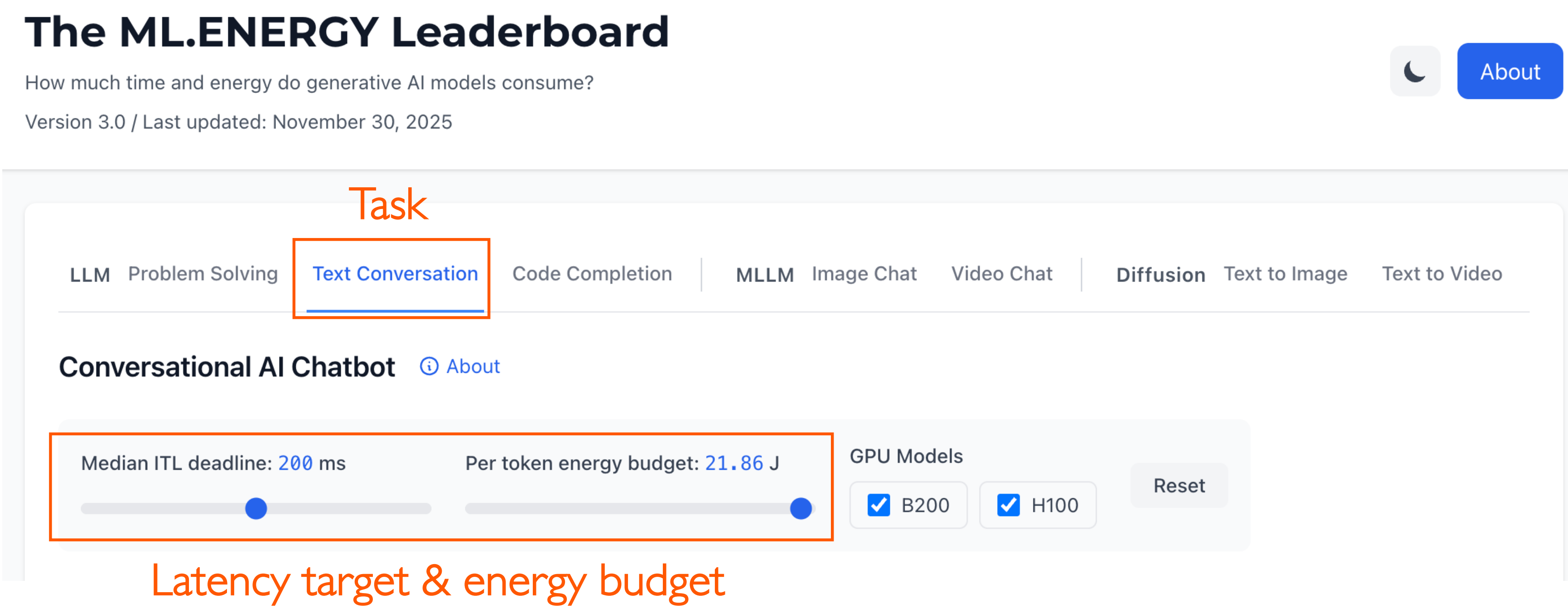
"What are the energy implications of the choices we make?"

Automated Optimization Recommendation



The ML.ENERGY Benchmark: Toward Automated Inference Energy Measurement and Optimization, NeurIPS D&B Spotlight, 2025

The ML.ENERGY Leaderboard



Time vs. Energy Trade-off

https://ml.energy/leaderboard/

Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training

Jie You, Jae-Won Chung*, and Mosharaf Chowdhury*

"How does energy interact with time?"

Understanding GPU Energy Consumption

***Energy to Accuracy* (ETA)**

- Energy needed to reach the user-specified target accuracy
- Energy-counterpart of *Time to Accuracy* (TTA)

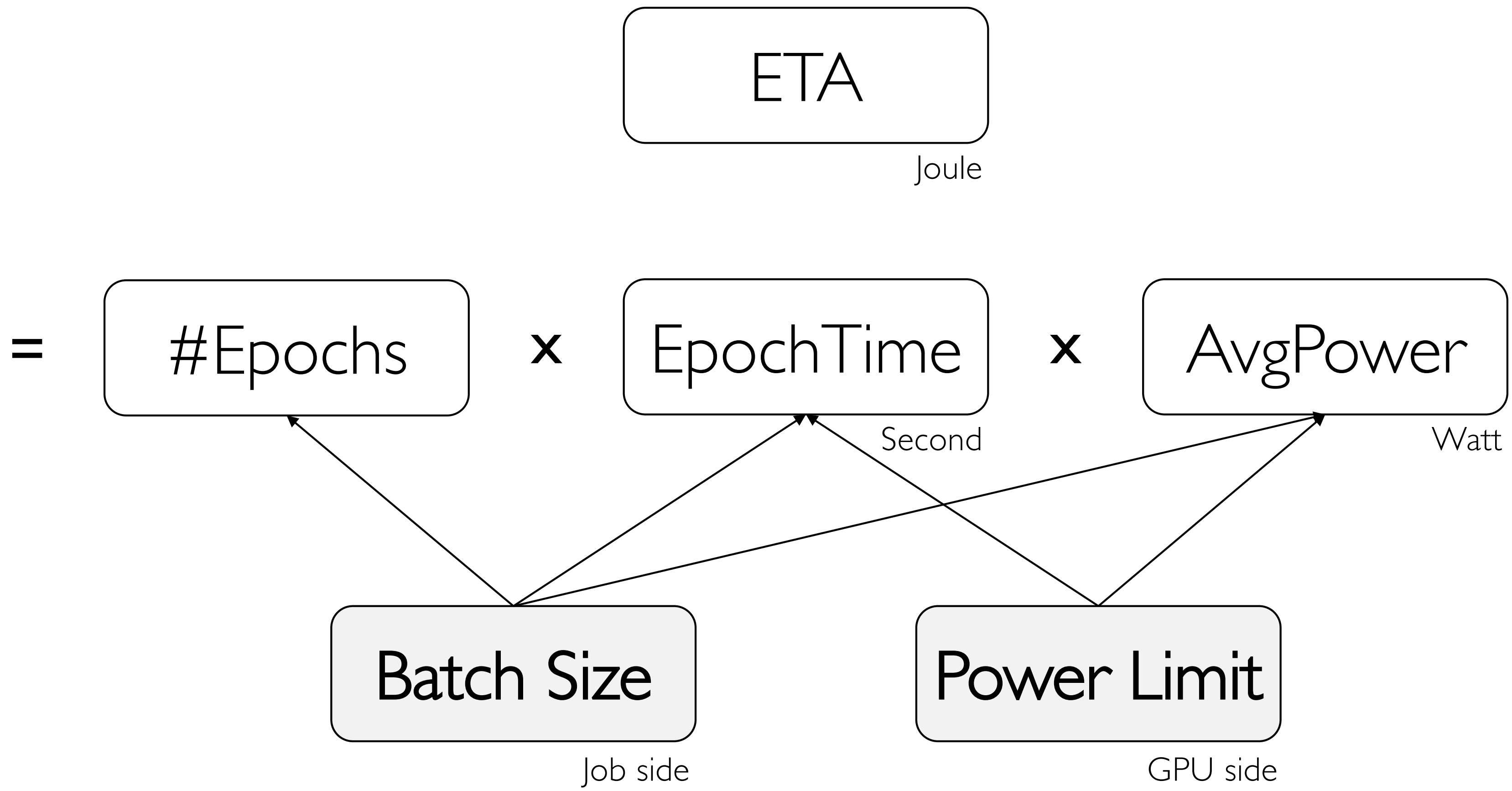
Zeus: Understanding and Optimizing Energy GPU Energy Consumption of DNN Training, NSDI, 2023

Understanding GPU Energy Consumption

$$\underbrace{\boxed{\text{ETA}}}_{\text{Joule}} = \underbrace{\boxed{\text{TTA}}}_{\text{Second}} \times \underbrace{\boxed{\text{AvgPower}}}_{\text{Watt}}$$

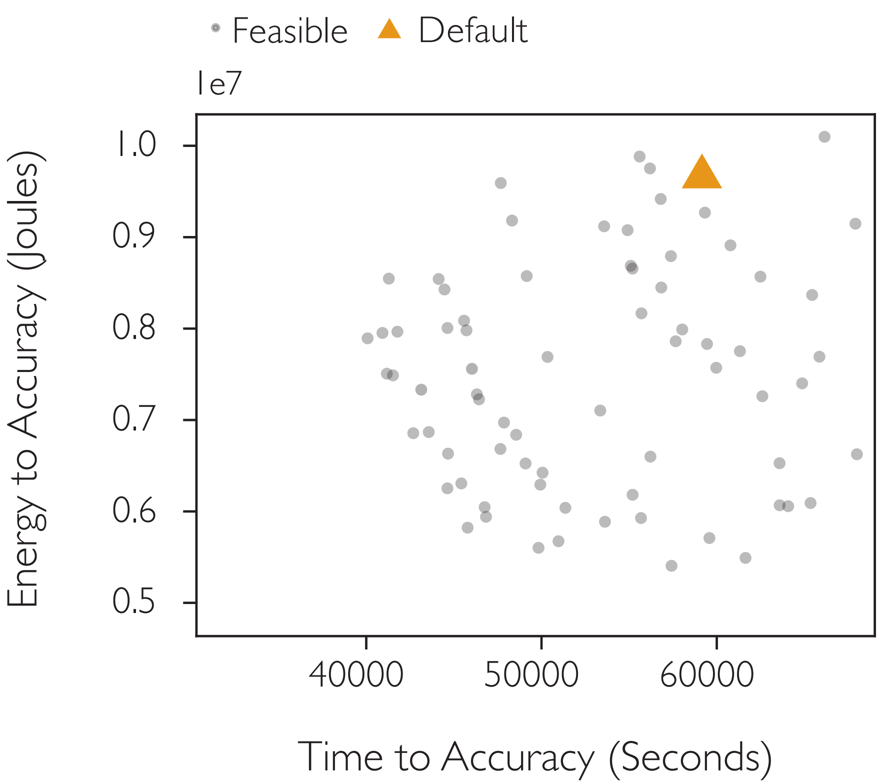
Zeus: Understanding and Optimizing Energy GPU Energy Consumption of DNN Training, NSDI, 2023

Understanding GPU Energy Consumption



Zeus: Understanding and Optimizing Energy GPU Energy Consumption of DNN Training, NSDI, 2023

Opportunity for Energy Savings

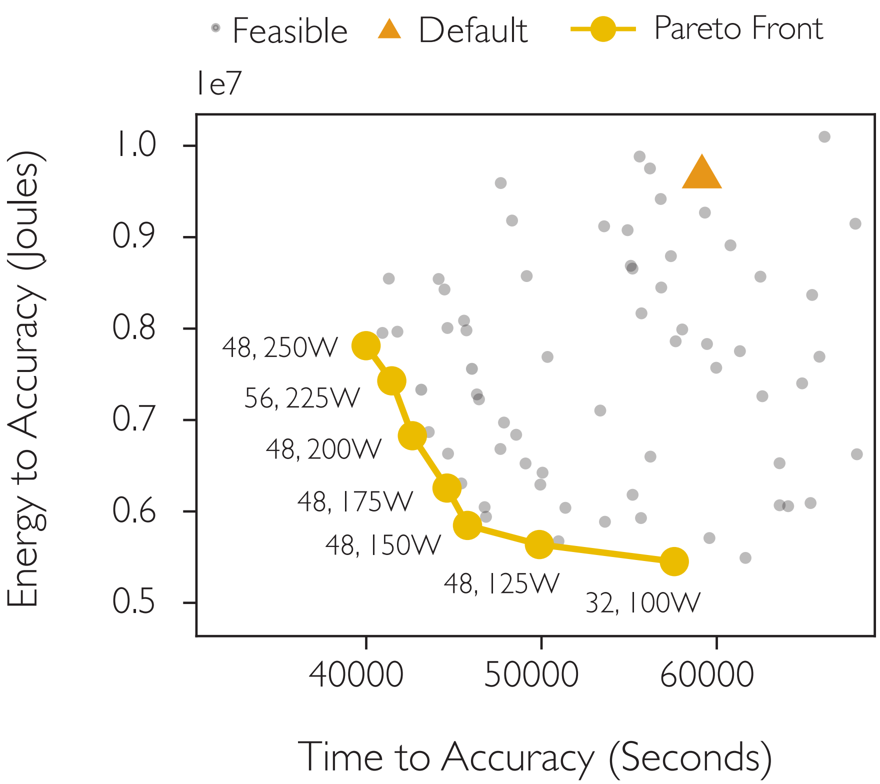


Training time and total energy affected by batch size and GPU power limit

Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100.
Similar trends found across four GPU generations.

Zeus: Understanding and Optimizing Energy GPU Energy Consumption of DNN Training, NSDI, 2023

Time vs. Energy Trade-off

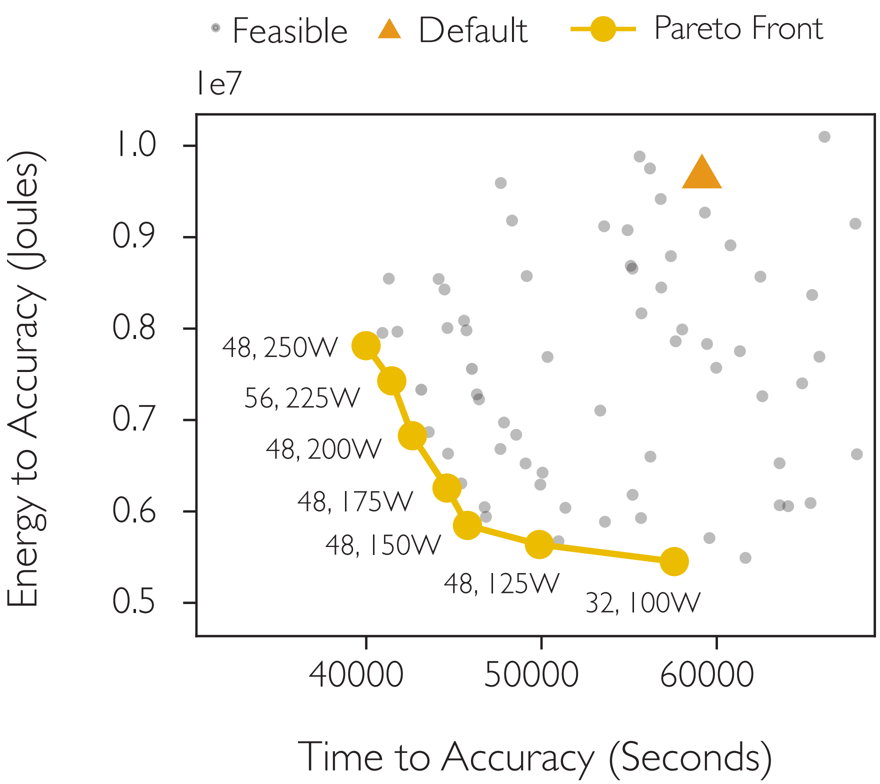


Training time and total energy affected by batch size and GPU power limit

Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100.
Similar trends found across four GPU generations.

Zeus: Understanding and Optimizing Energy GPU Energy Consumption of DNN Training, NSDI, 2023

Time vs. Energy Trade-off

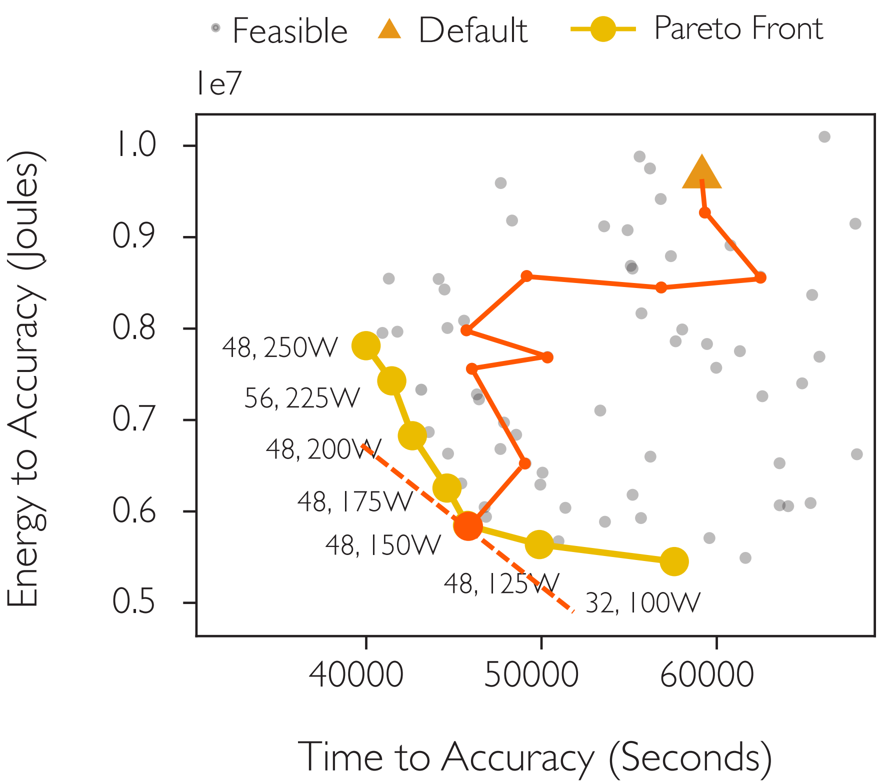


Which yellow point is the best?

Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100.
Similar trends found across four GPU generations.

Zeus: Understanding and Optimizing Energy GPU Energy Consumption of DNN Training, NSDI, 2023

Multi-Armed Bandit Algorithm

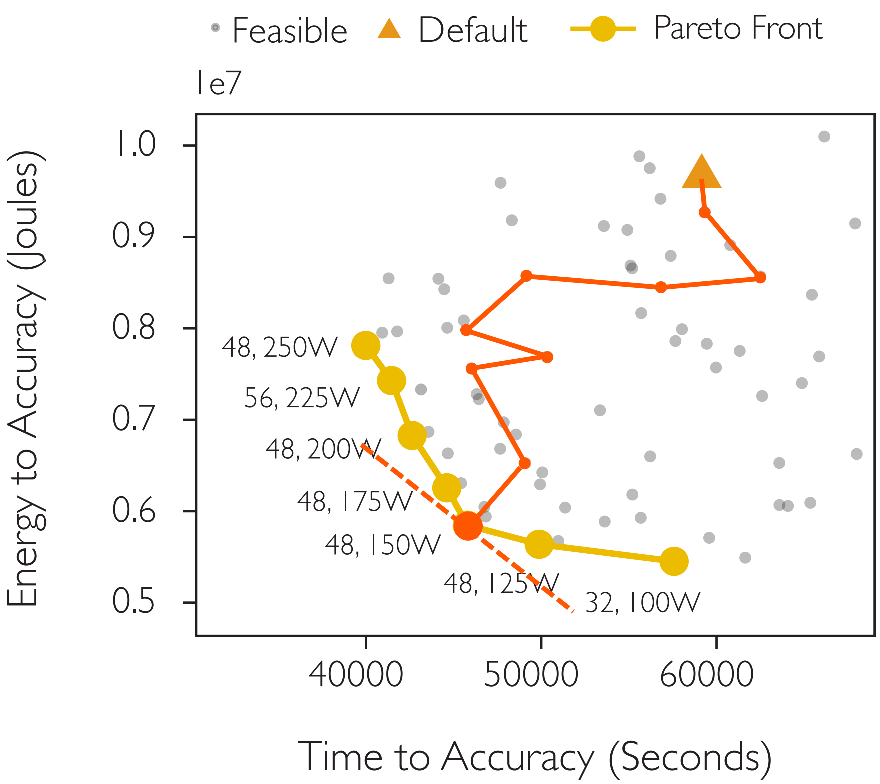


- Objective
  = Linear combination of time & energy

- Arm = Batch size
- Horizon = Recurring training
- Thompson sampling

Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100.
Similar trends found across four GPU generations.

Multi-Armed Bandit Algorithm



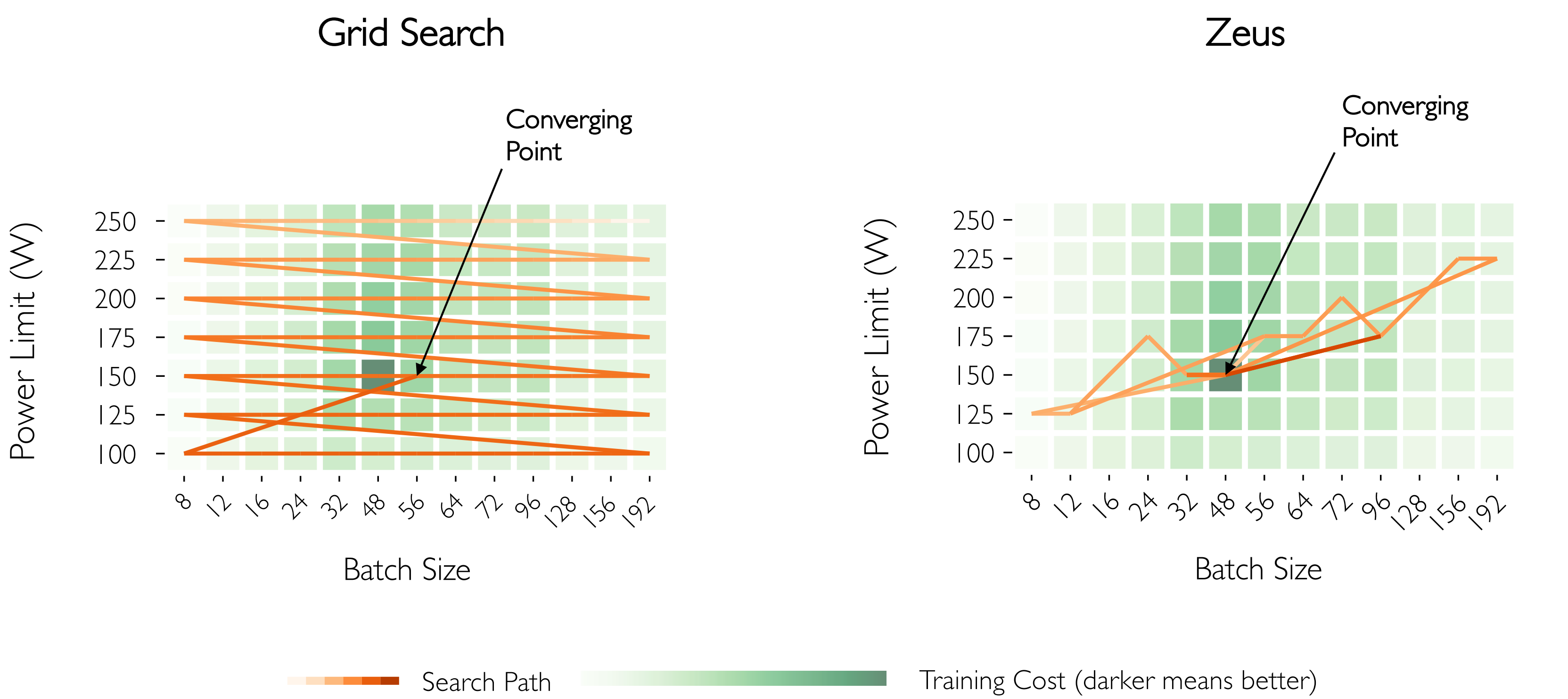
15% to 76% energy reduction

across diverse models
and multiple GPU generations

Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100.
Similar trends found across four GPU generations.

Zeus: Understanding and Optimizing Energy GPU Energy Consumption of DNN Training, NSDI, 2023

Zeus in Action



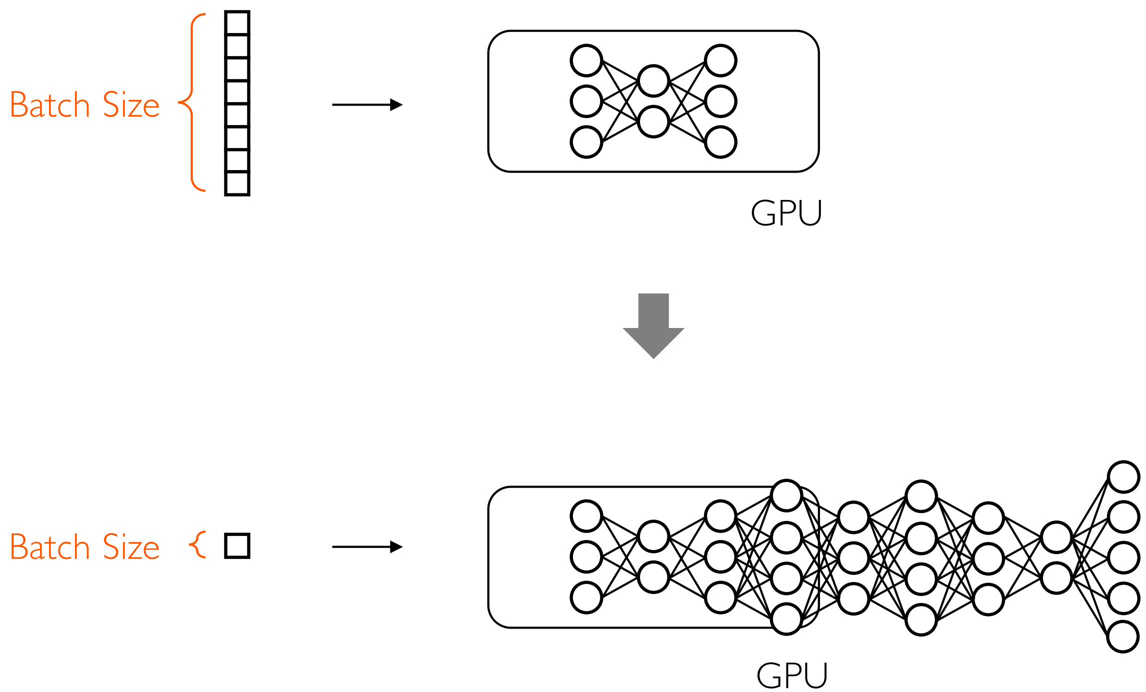
DeepSpeech2 trained on LibriSpeech on an NVIDIA V100 GPU.

Reducing Energy Bloat in Large Model Training

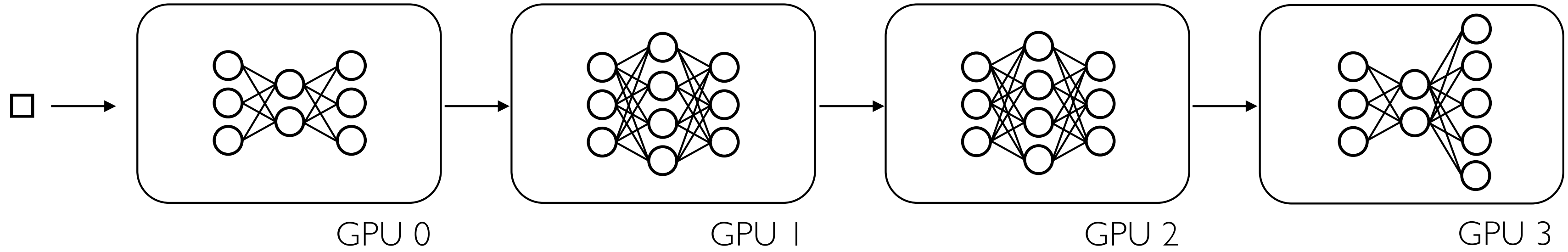
Jae-Won Chung, Yile Gu, Insu Jang,
Luoxi Meng, Nikhil Bansal, Mosharaf Chowdhury

"Any way to reduce energy consumption without slowdown?"

Explosive Growth in Model Size

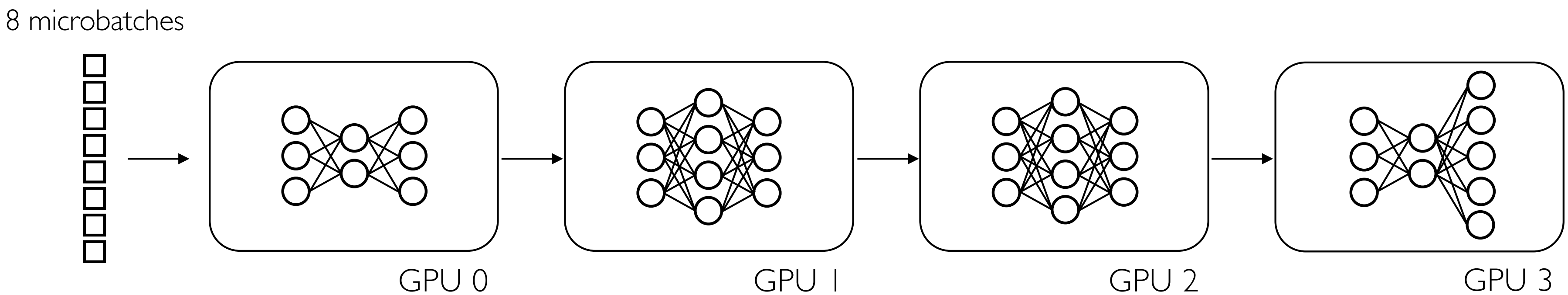


Model Parallelism

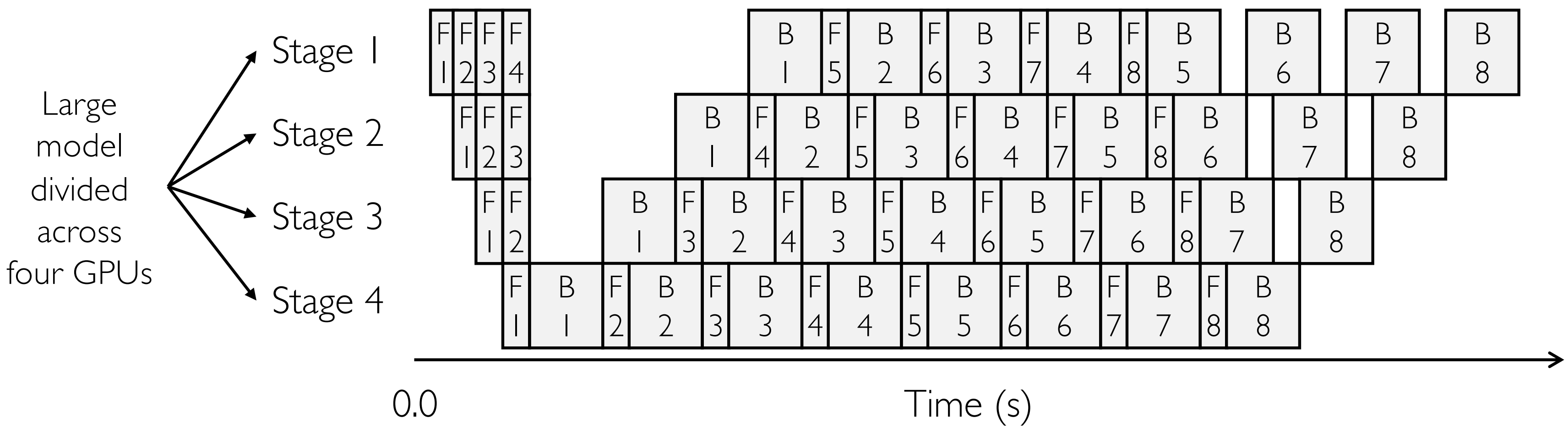


Model Parallelism

Pipeline Parallel training

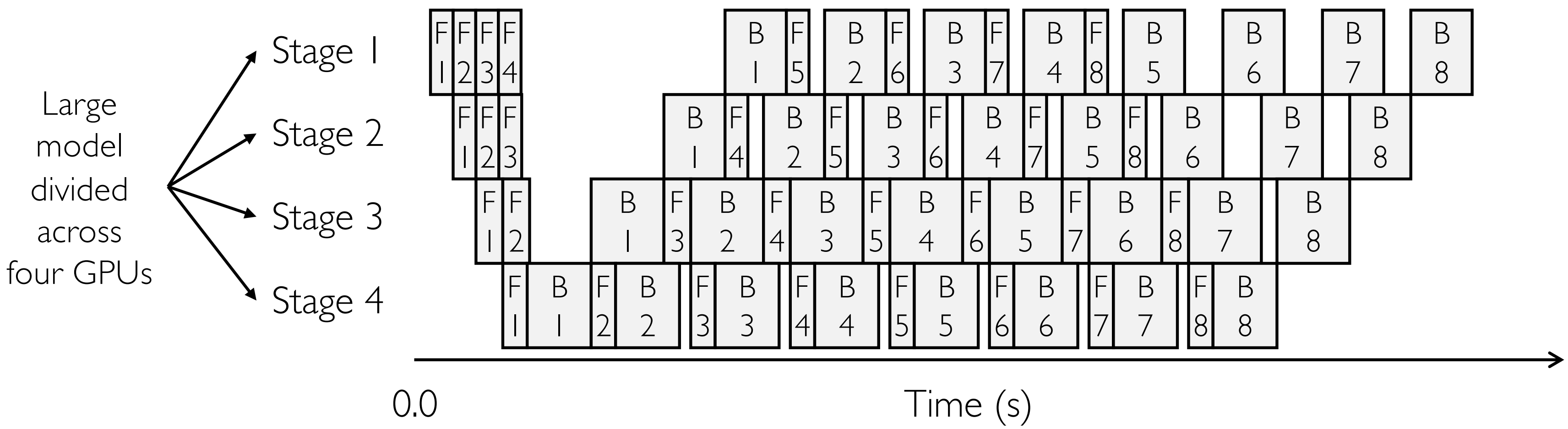


Pipeline Parallel Training



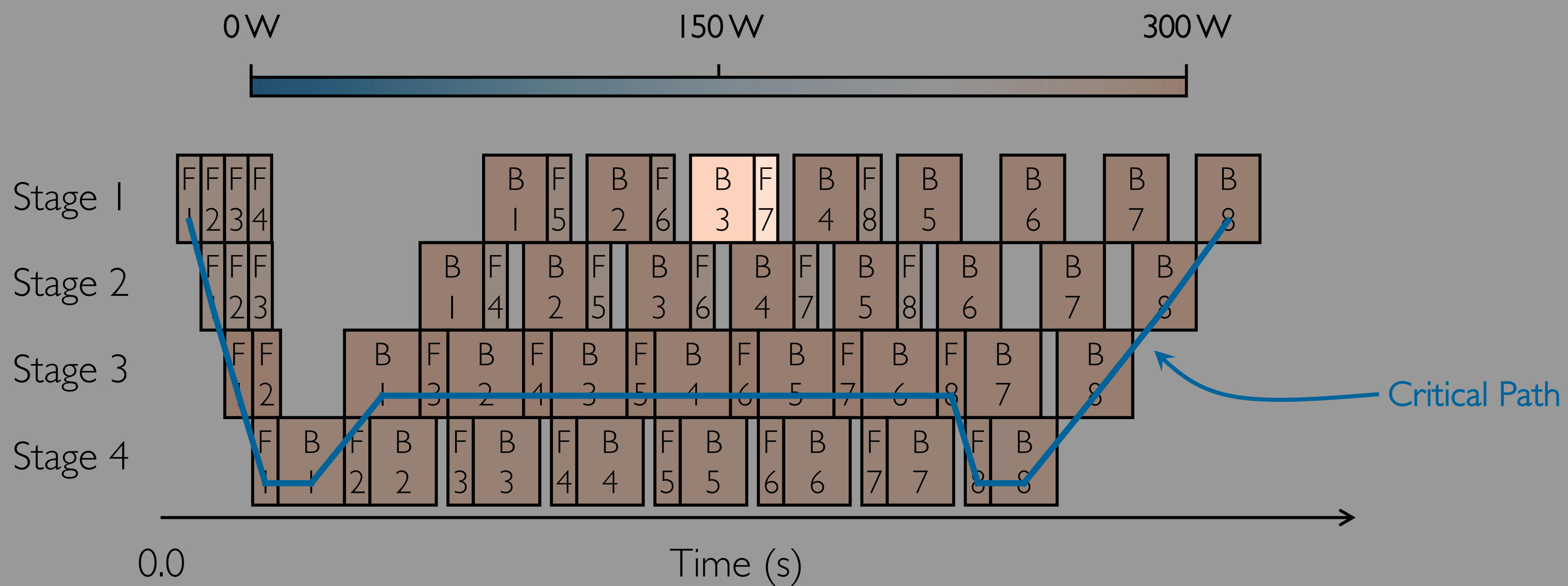
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).

Fundamental Computation Imbalance



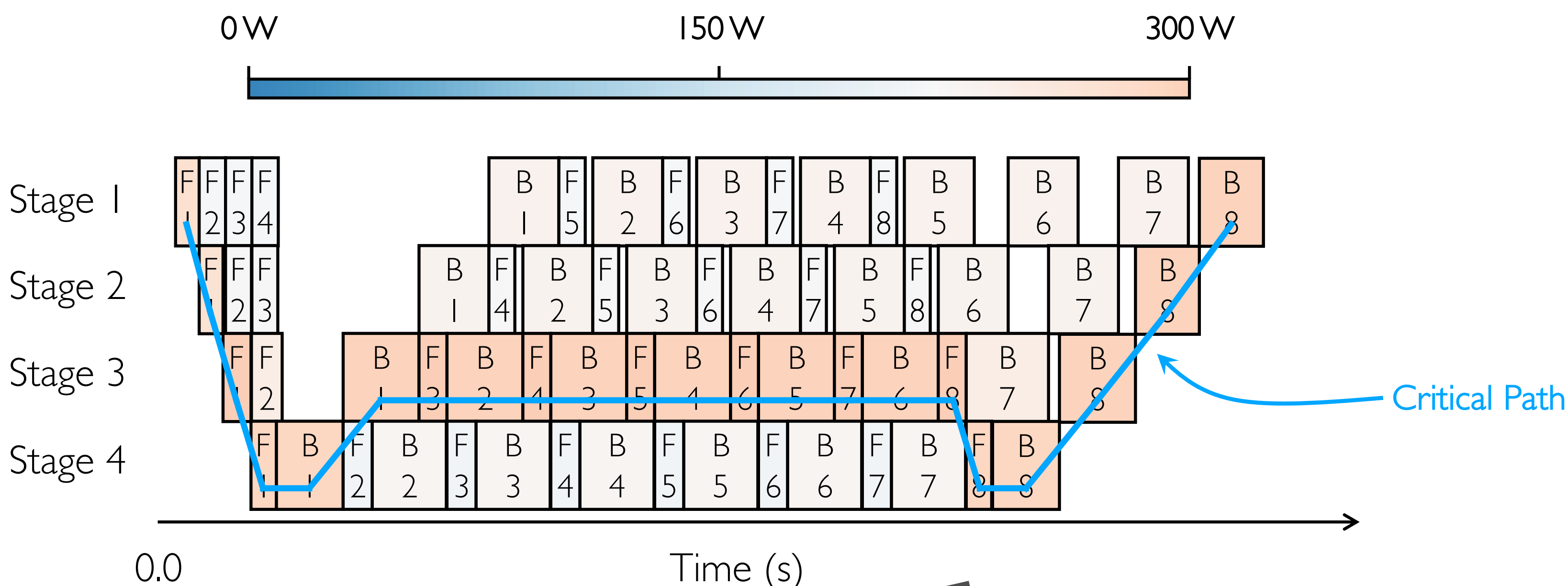
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).
Drawn to scale for GPT-3 1.3B on NVIDIA A100 GPUs.

Where Do the Joules Go?



One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).
Drawn to scale for GPT-3 1.3B on NVIDIA A100 GPUs.

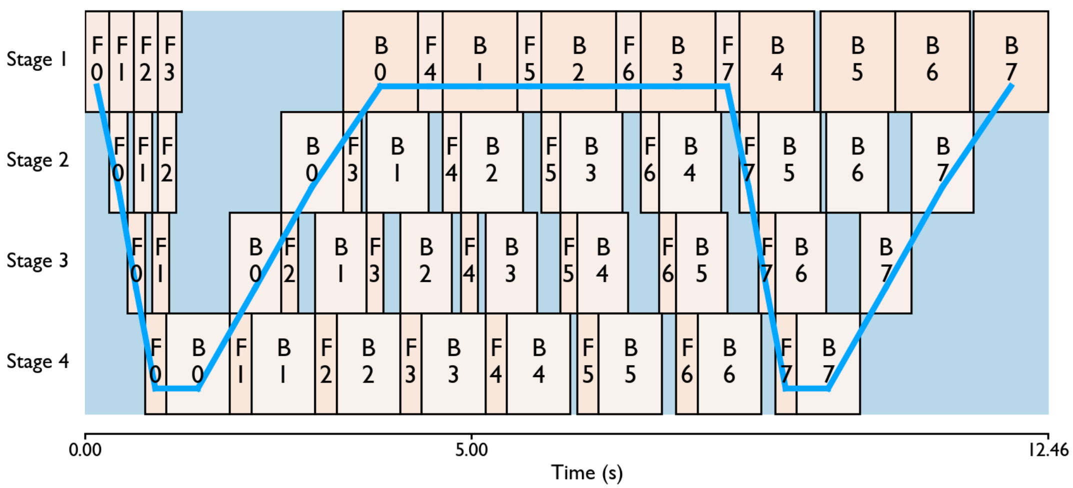
Cutting 30% Energy Bloat

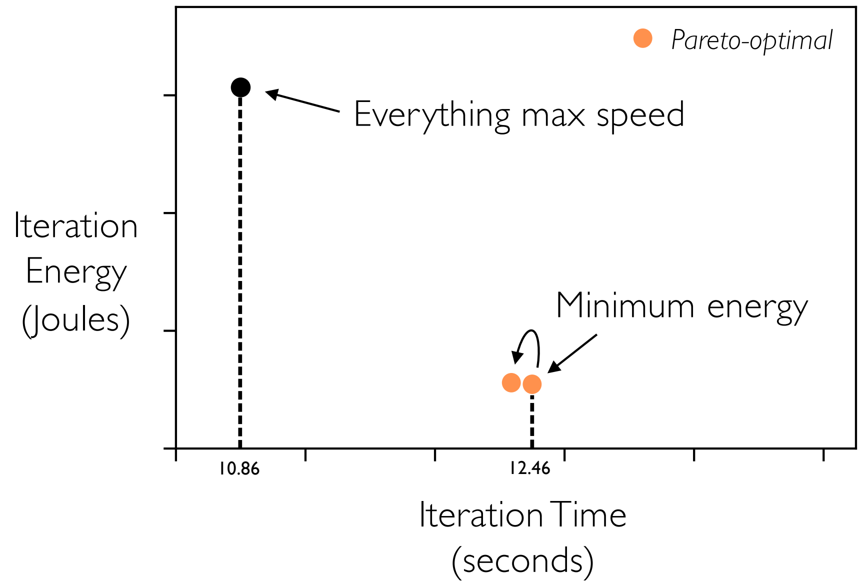


One training iteration of GPT-[illegible] B with four pipeline stages
and eight microbatches on NVIDIA A100 GPUs, drawn to scale.

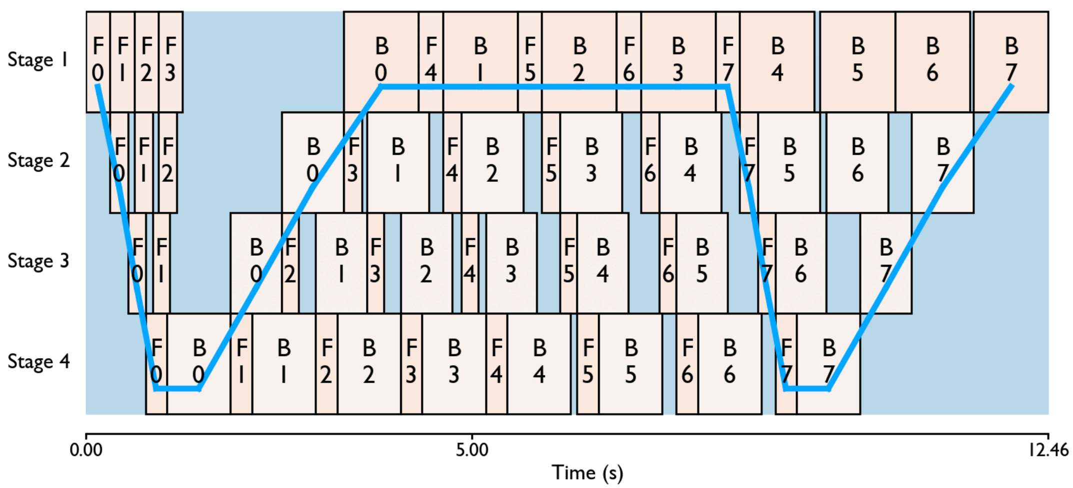
NP-Hard

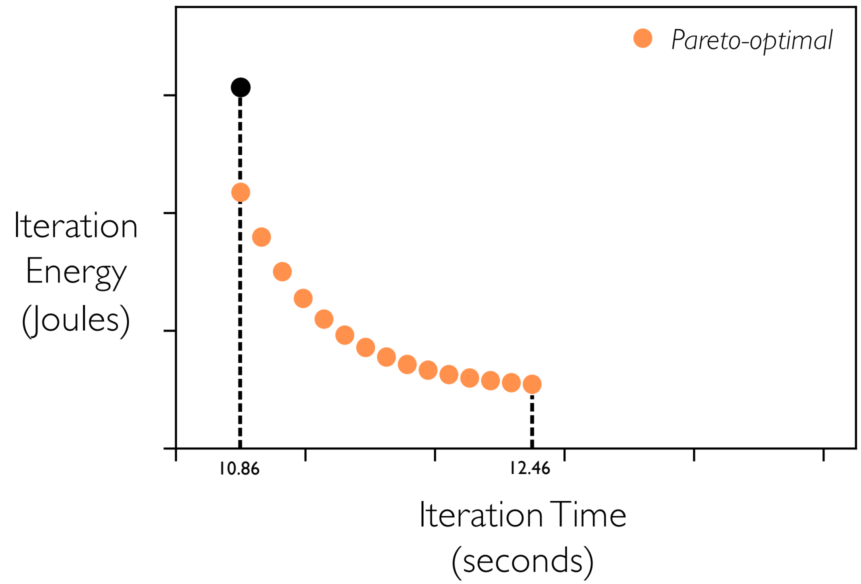
Time vs. Energy Trade-off





Time vs. Energy Trade-off





Summary: Decisions Across the AI Stack

ML Algorithm

MHA vs. GQA
Sparse vs. Dense VideoGen

ML Job

Batch size for training
Input/Output length
Image resolution

Framework

Model parallelism
Batch size for serving
PD disaggregation etc.
Execution scheduling

Cluster

Number of instances
GPU type

Hardware (GPU)

Frequency scaling
Power capping

→ Jointly influence energy and time

Summary

- Optimize energy along the time–energy trade-off frontier.
- Leverage available latency slack for energy savings.