Power and Energy Considerations for Machine Learning Systems

Jae-Won Chung

April 2nd, 2024

SymbioticLab
ML.ENERGY
UNIVERSITY OF MICHIGAN

About the Speaker

Jae-Won Chung

- Third year PhD student here
- Advised by Professor Mosharaf Chowdhury
- Making energy a first-class systems optimization metric
- But I know a little bit about power as well

Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs

In total, Meta will have the compute power equivalent to 600,000 Nvidia H100 GPUs to help it develop next-generation AI, says CEO Mark Zuckerberg.

By Michael Kan   January 18, 2024

(David Paul Morris/Bloomberg via Getty Images)

Source: https://www.pcmag.com/news/zuckerbergs-meta-is-spending-billions-to-buy-350000-nvidia-h100-gpus

Data Center Planning

A couple considerations

- Land
- Building
- Racks
- Cooling
- Power delivery

Global Data Center Trends 2023

REPORT

New technology is driving record demand but power constraints are inhibiting growth

CBRE RESEARCH
JULY 2023

Source: https://www.cbre.com/insights/reports/global-data-center-trends-2023

Data Center Planning

A couple considerations

- Land
- Building
- Racks
- Cooling
- Power delivery

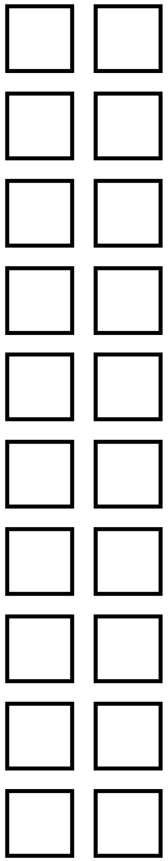
350,000 H100 GPUs?

- One GPU's TDP is 700 W
- 245 MW in total
- 200,000 average households
- Four Ann Arbors

Then, do we allocate 245 MW for GPU power?

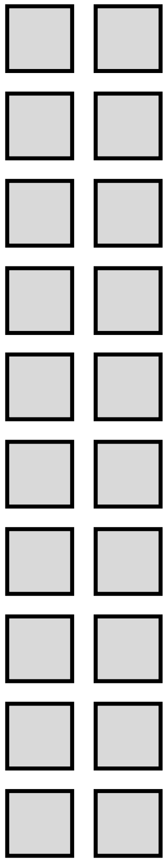
Source: "How much electricity does an American home use?," EIA, https://www.eia.gov/tools/faqs/faq.php?id=97&t=3

Airplane Overbooking



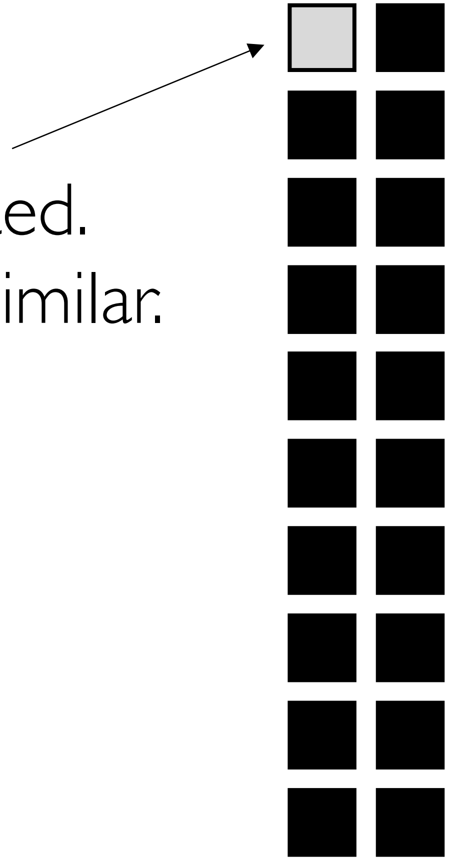
20 seats on an airplane

Airplane Overbooking



Fully booked!

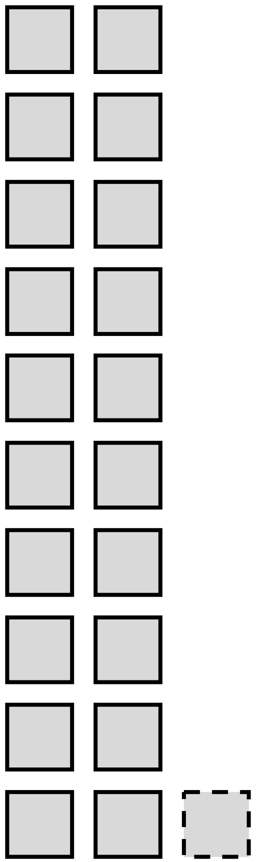
Airplane Overbooking



One empty seat wasted.
Plane operating cost is similar.

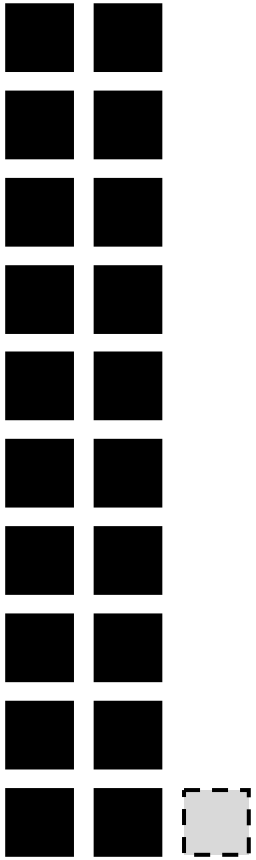
A passenger has on average
a 95% chance of showing up

Airplane Overbooking



105% overbooked!

Airplane Overbooking



Prepare a fallback strategy
just in case everyone shows up

99.75% filled

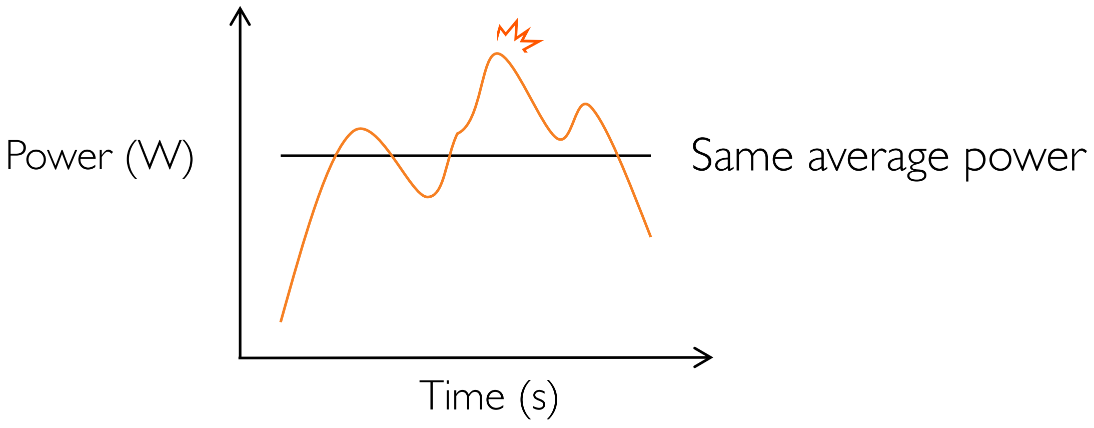
Data Center Power Oversubscription

Will all the 350,000 H100 GPUs consume 700 W all the time?

- Probably not – Average power draw will be lower.

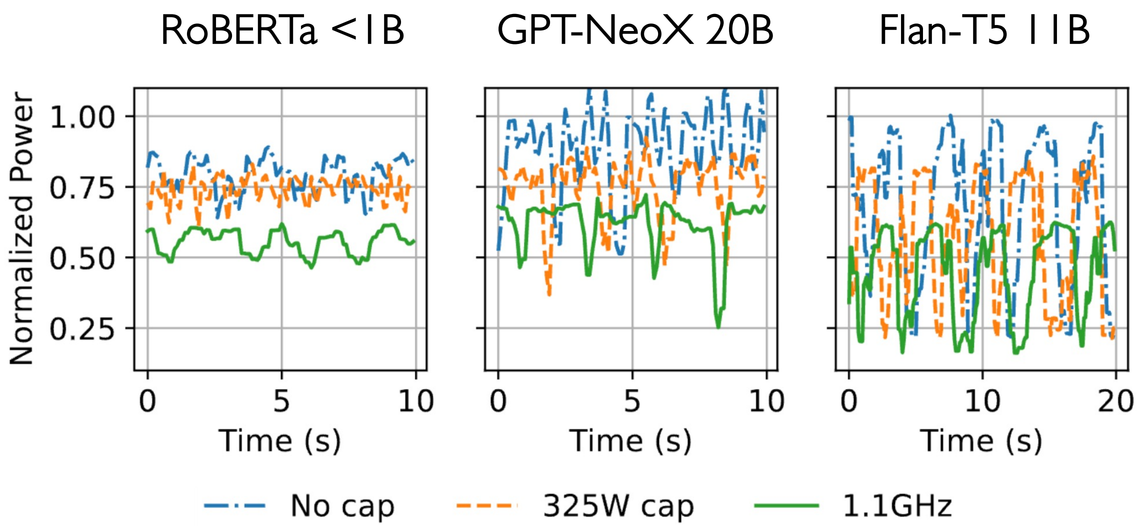
Is it the exact same problem as airplane overbooking?

- The extra *time* axis – It's airplane overbooking *over time*.
- The variability of power draw should be considered.



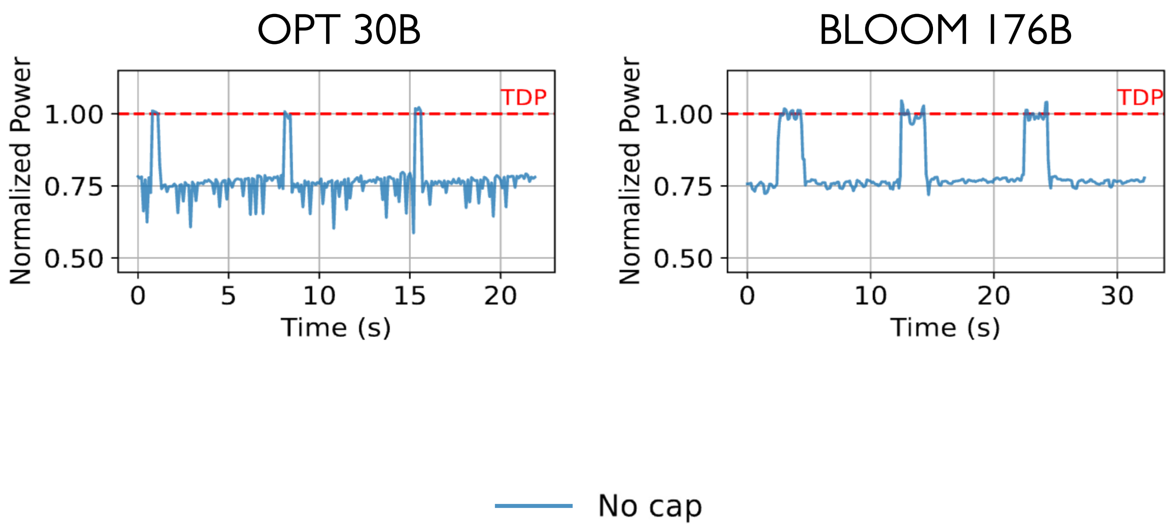
Should We Oversubscribe Power?

LLM training



Average power is close to TDP
High power variability
Hard to run multiple jobs to reduce variability

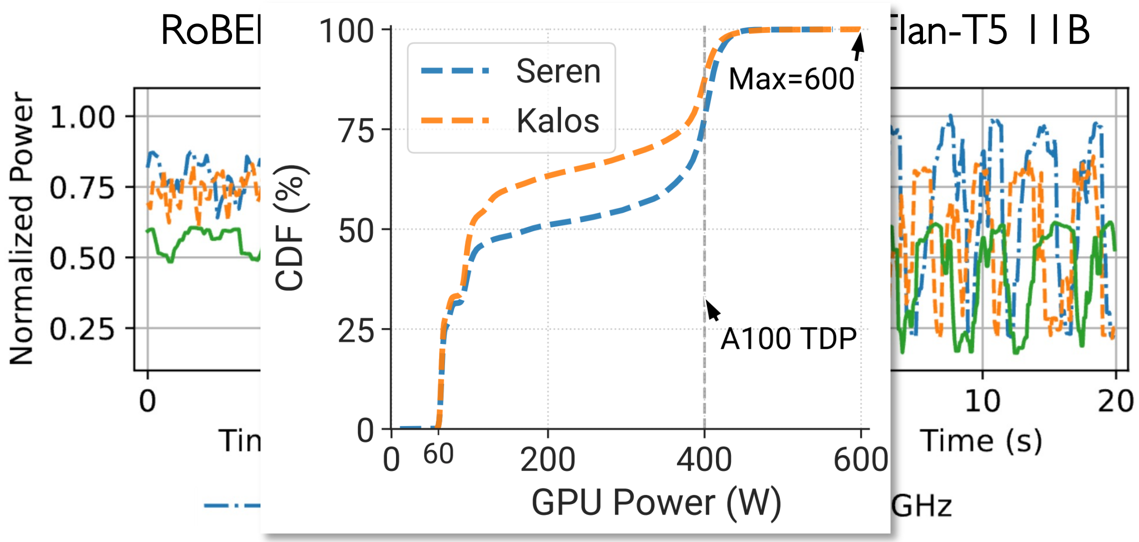
LLM inference



Average power has 20% headroom
High power variability but has clear patterns
Can run multiple servers to reduce variability

Should We Oversubscribe Power?

LLM training

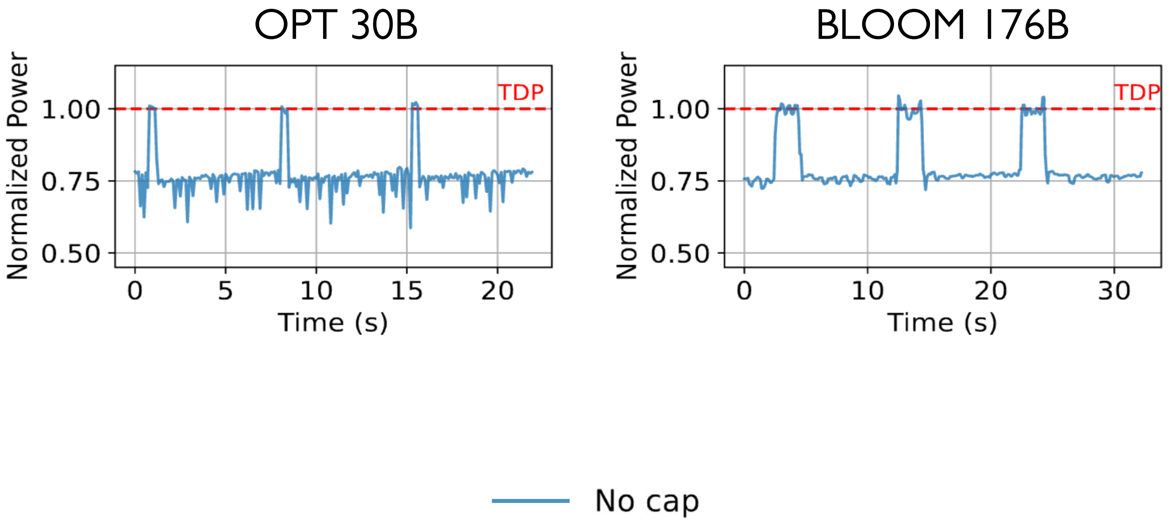


Average power is close to TDP

High power variability

Hard to run multiple jobs to reduce variability

LLM inference



Average power has 20% headroom

High power variability but has clear patterns

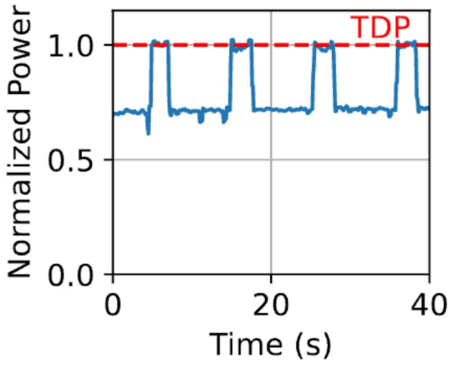
Can run multiple servers to reduce variability

Source: "Characterization of Large Language Model Development in the Datacenter," Hu, Ye, and Wang et al., 2024

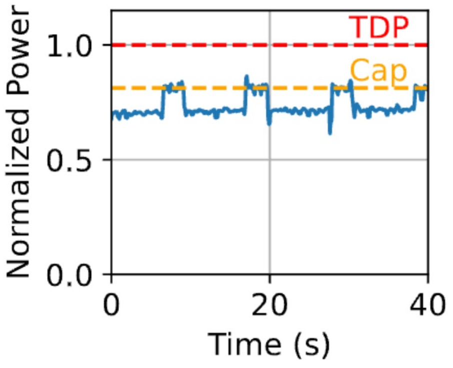
Preventing Power From Exceeding Cap

GPU Power Limiting

BLOOM 176B Inference



No cap

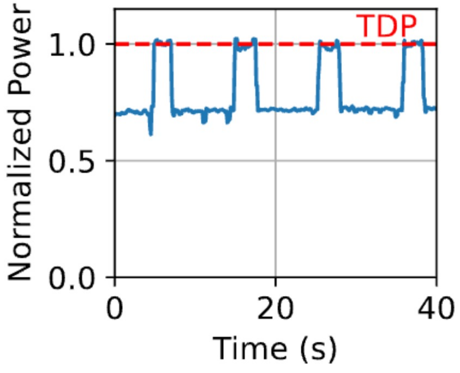


325W/400W cap

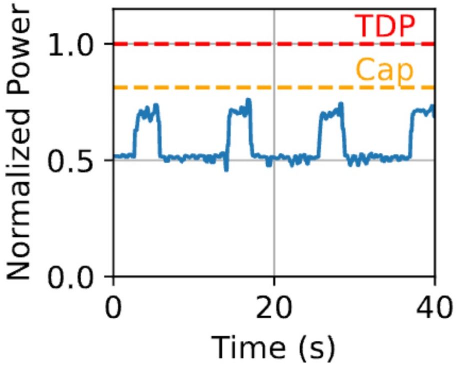
Limited power reduction (only peak)

GPU Frequency Locking

BLOOM 176B Inference



No cap



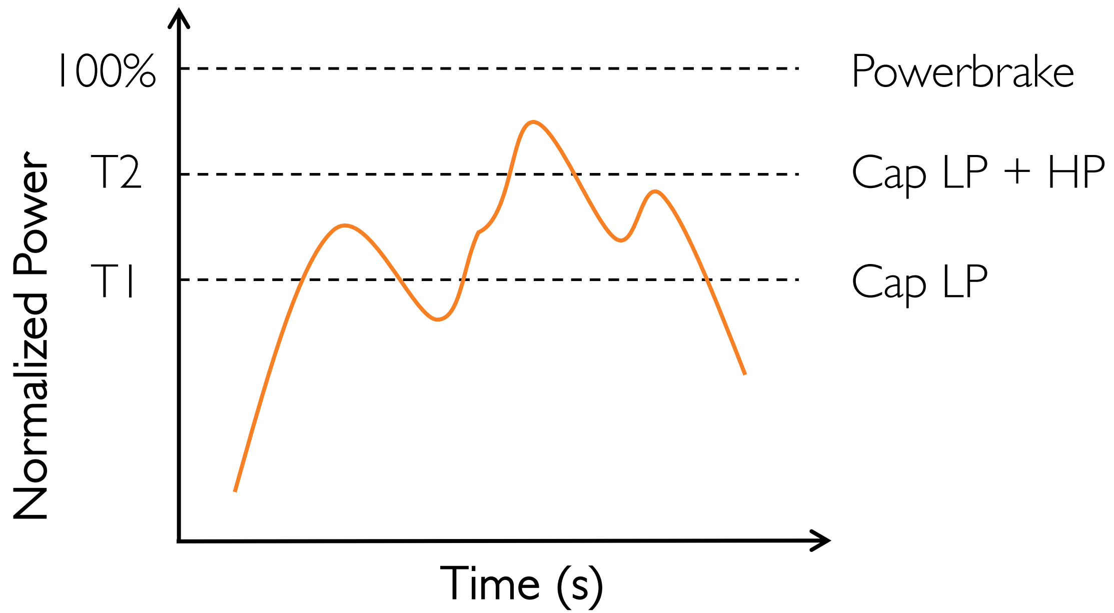
1.1GHz/1.4GHz cap

Reduces power over all phases

Power Oversubscription Policy

Workload	Ratio	Priority
Summarize	25%	Low
Search	25%	High
Chat	50%	50:50

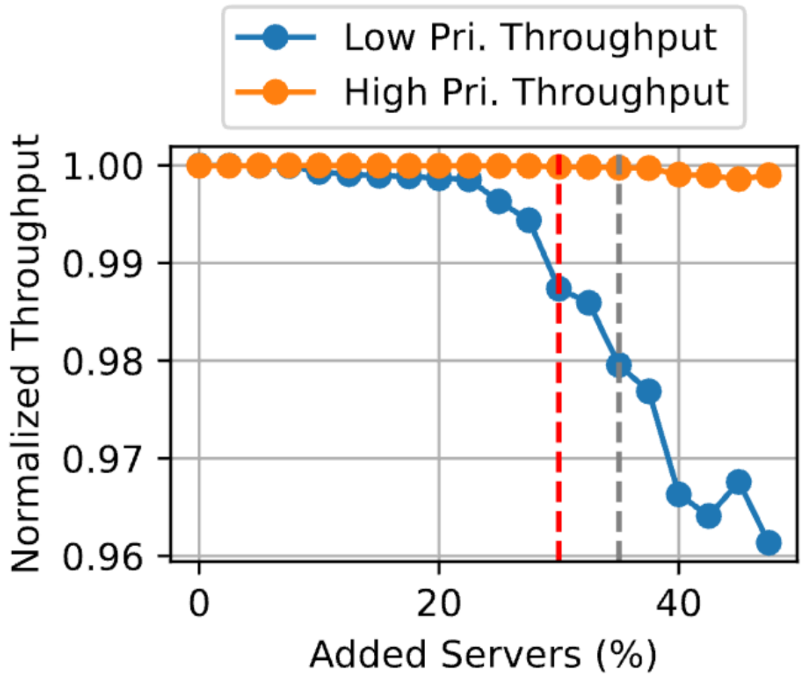
Inference cluster with mixed-priority workloads



Two-threshold policy

Evaluation

What happens as we oversubscribe more and more power?



Can add 30% more servers with very little throughput degradation

Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs

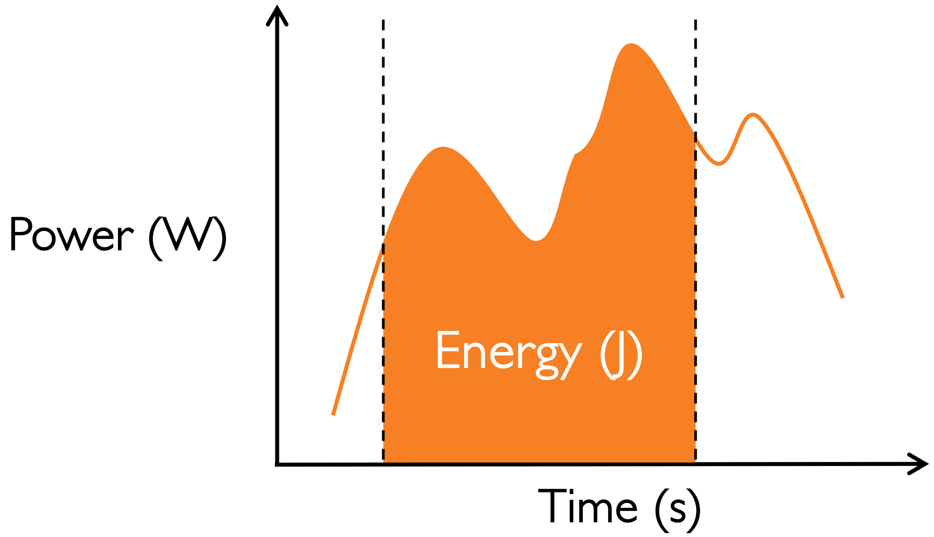
In total, Meta will have the compute power equivalent to 600,000 Nvidia H100 GPUs to help it develop next-generation AI, says CEO Mark Zuckerberg.

By Michael Kan January 18, 2024

(David Paul Morris/Bloomberg via Getty Images)

Source: https://www.pcmag.com/news/zuckerbergs-meta-is-spending-billions-to-buy-350000-nvidia-h100-gpus

Power vs. Energy



We're billed by the amount of energy (electricity) we use.
Power oversubscription doesn't optimize energy.

ML Energy Consumption

Some numbers

- IT consumes 7-8 % of global electricity today[1]
- Amazon consumed ~11.9 GWh to train one 200B LLM[2]
  - Enough to power more than 1000 US households for a year
- Models are periodically re-trained to keep it up to date[3]

[1] *"Digital Economy and Climate Impact – White Paper,"* Schneider Electric, 2021
[2] *"Constraint-driven Innovation (CIDR keynote),"* Hamilton, 2024
[3] *"Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective,"* Hazelwood et al., 2018

Understanding GPU Energy Consumption

Energy to Accuracy **(ETA) for DNN training**

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

Understanding GPU Energy Consumption

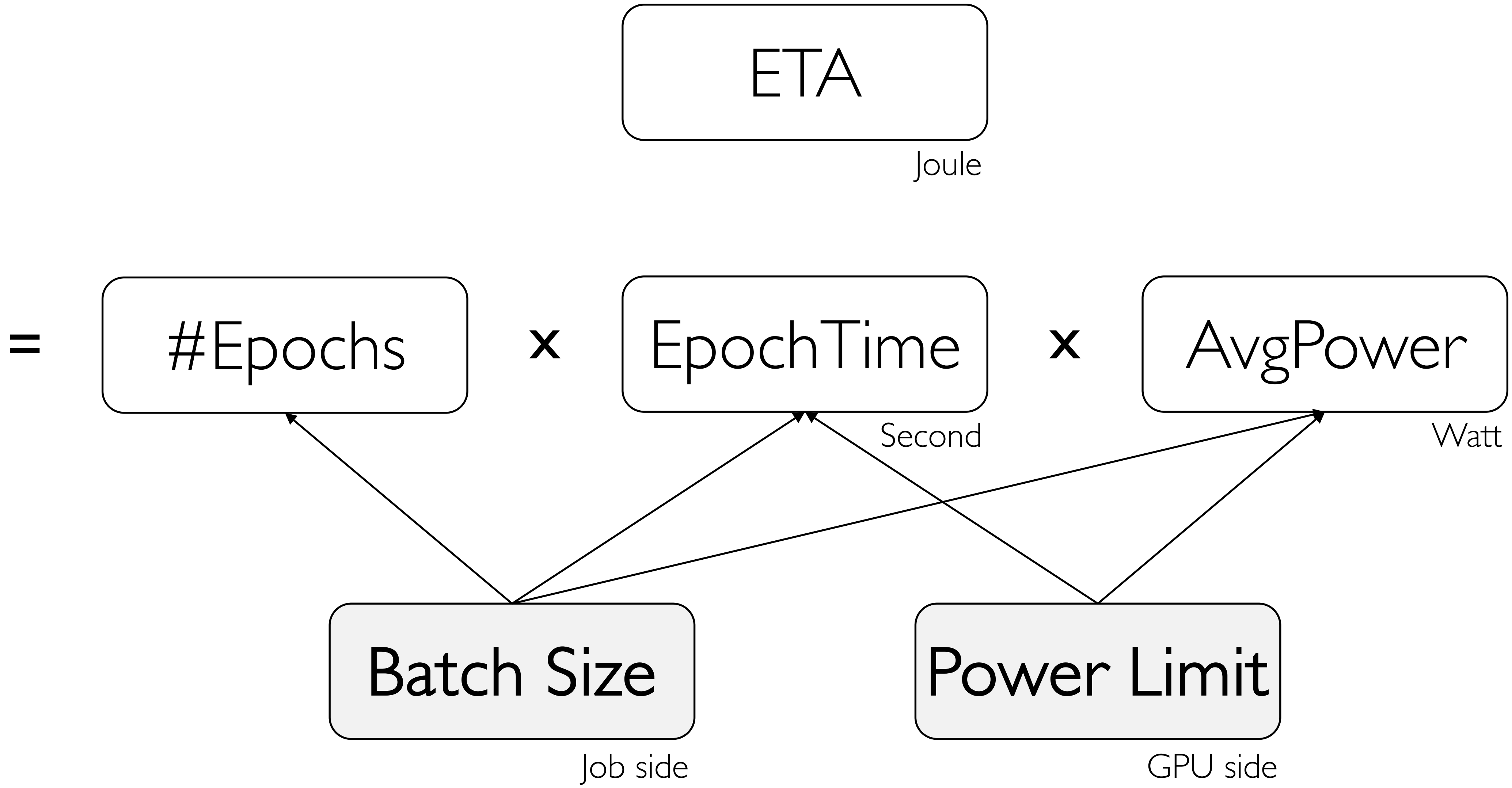
ETA

Joule

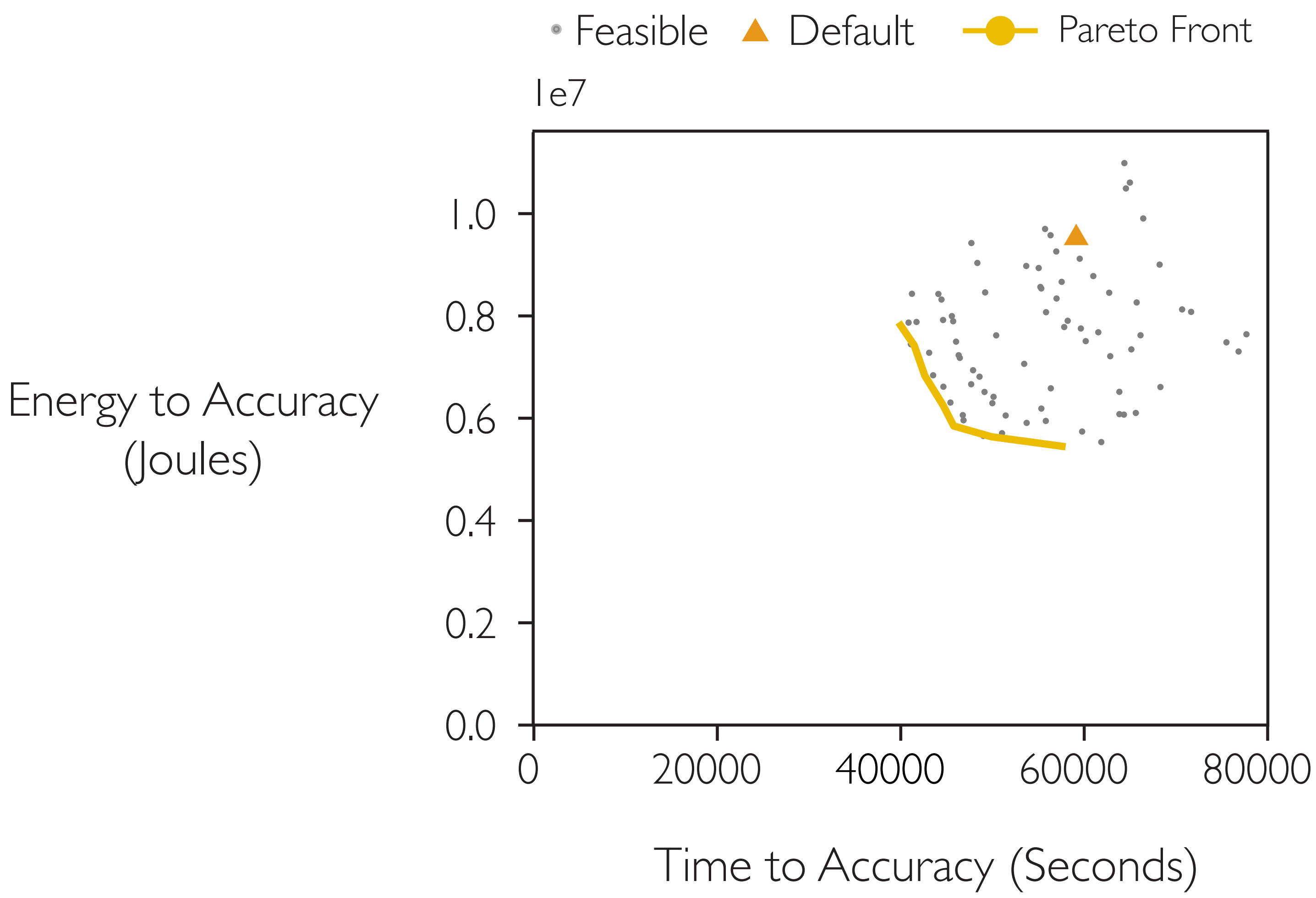
= TTA x AvgPower

Second Watt

Understanding GPU Energy Consumption

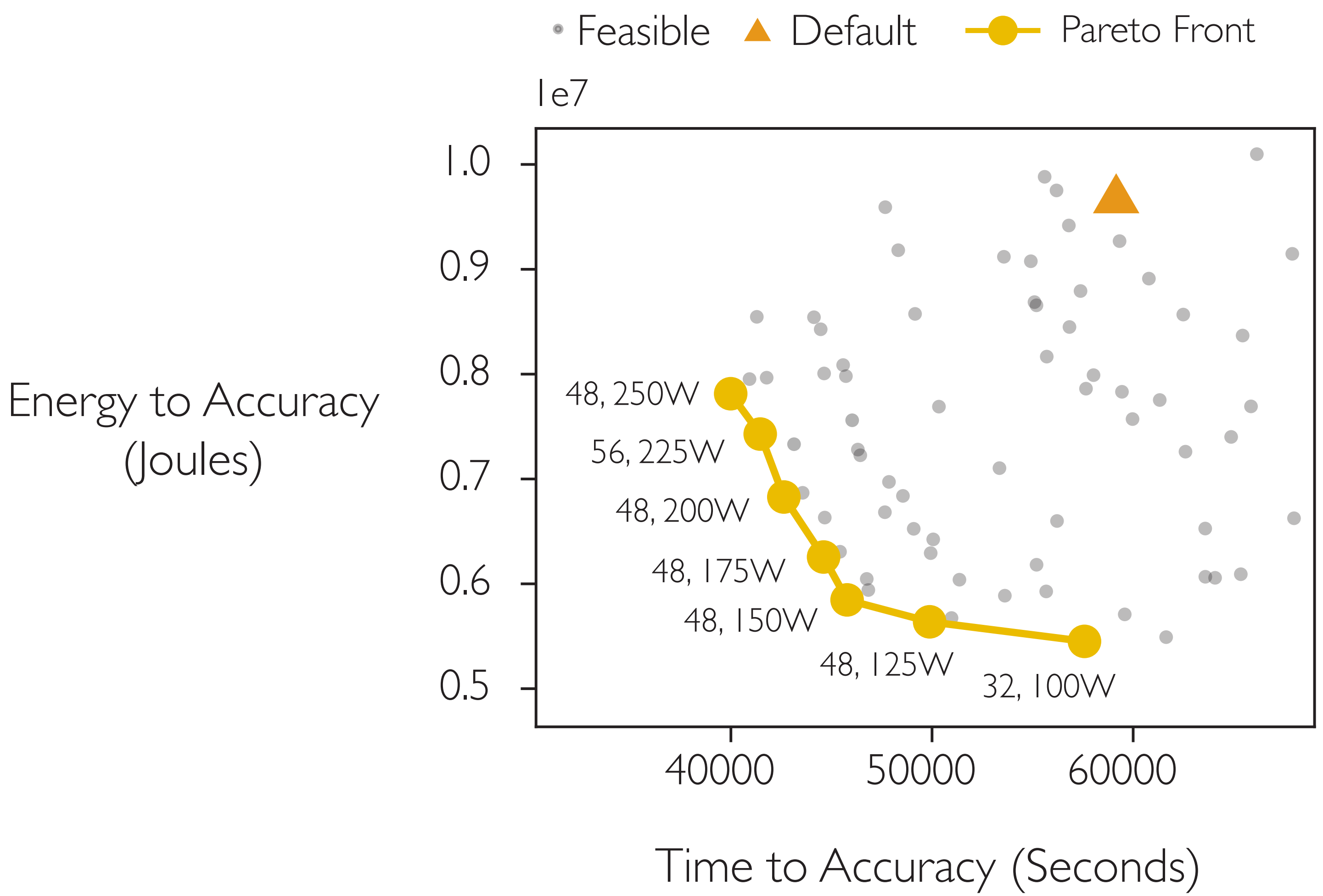


Relationship Between Time and Energy



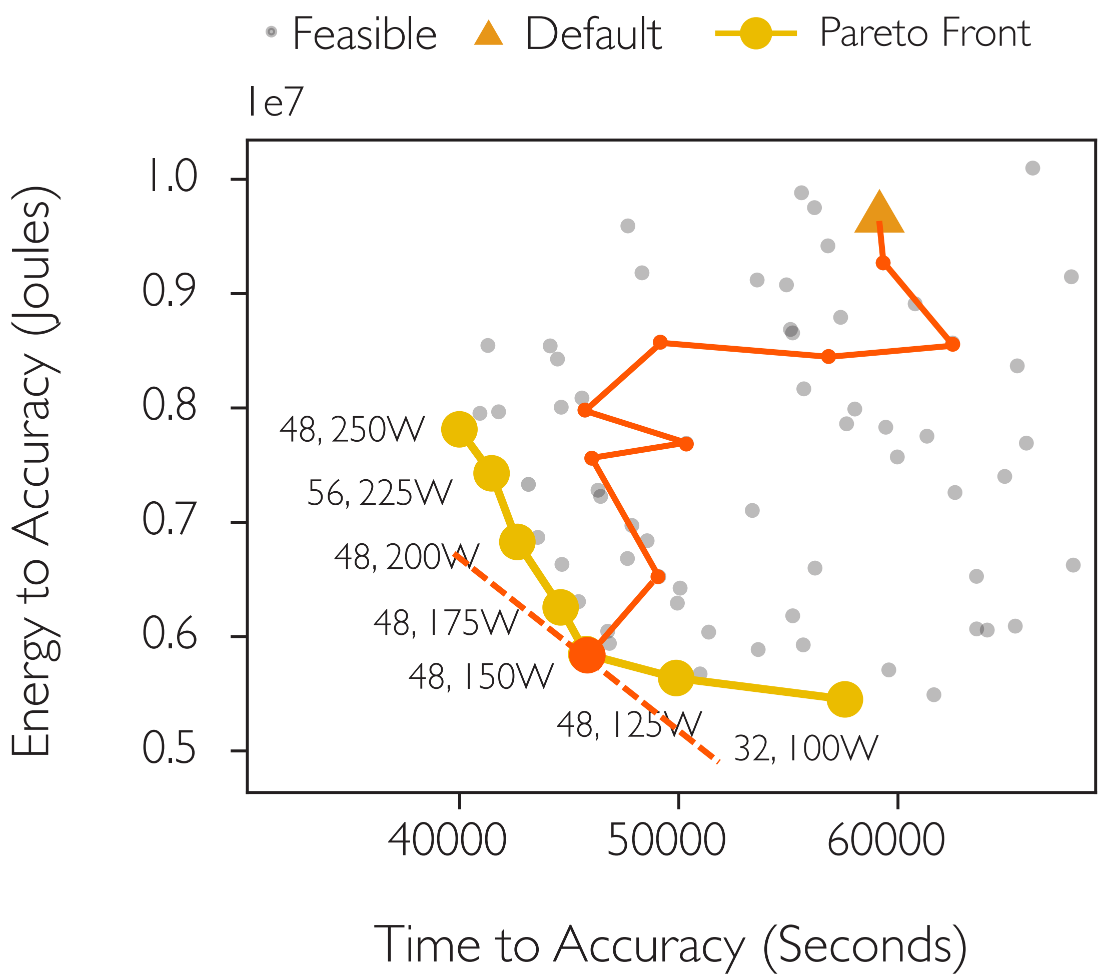
Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100 GPU.
Similar trends found across 6 DL workloads and 4 GPU generations.

Relationship Between Time and Energy



Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100 GPU.
Similar trends found across 6 DL workloads and 4 GPU generations.

Relationship Between Time and Energy



Which yellow point is the best?

$$\text{Cost} = \eta \cdot \text{ETA} + (1 - \eta) \cdot \text{MaxPower} \cdot \text{TTA}$$

Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100 GPU.
Similar trends found across 6 DL workloads and 4 GPU generations.

Finding the Pareto Frontier

Batch size and power limit optimization decoupled

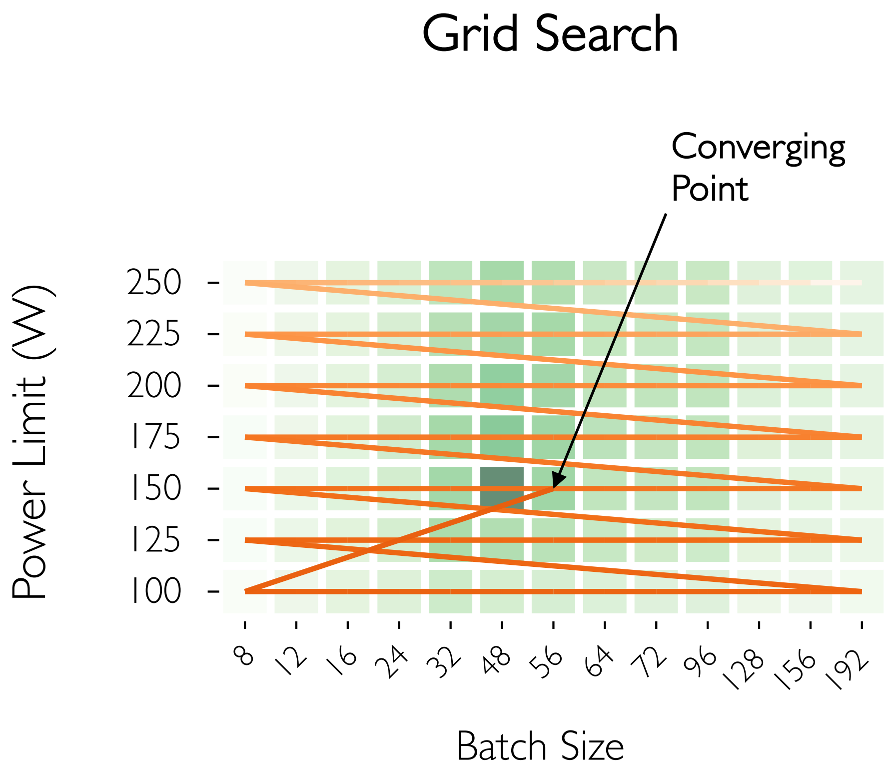
- Find the best batch size across retraining jobs
- Find the best power limit for one batch size during training

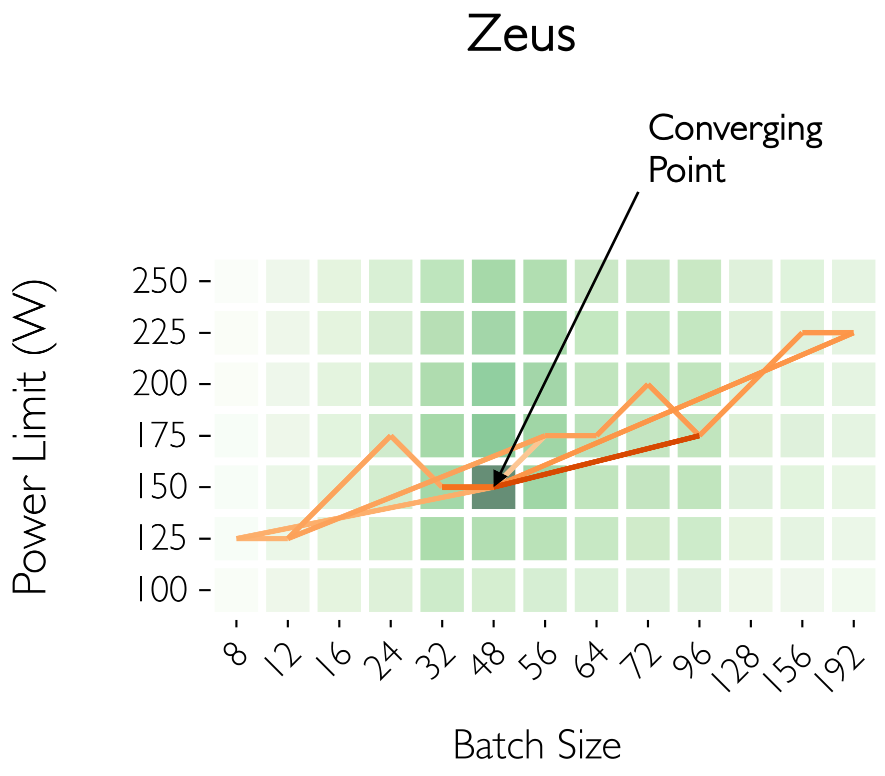
$$\text{Cost} = \eta \cdot \text{ETA} + (1 - \eta) \cdot \text{MaxPower} \cdot \text{TTA}$$

Multi-Armed Bandit formulation

- Learns a stochastic function from batch size to cost
- Automatically trades off exploration and exploitation

Zeus in Action

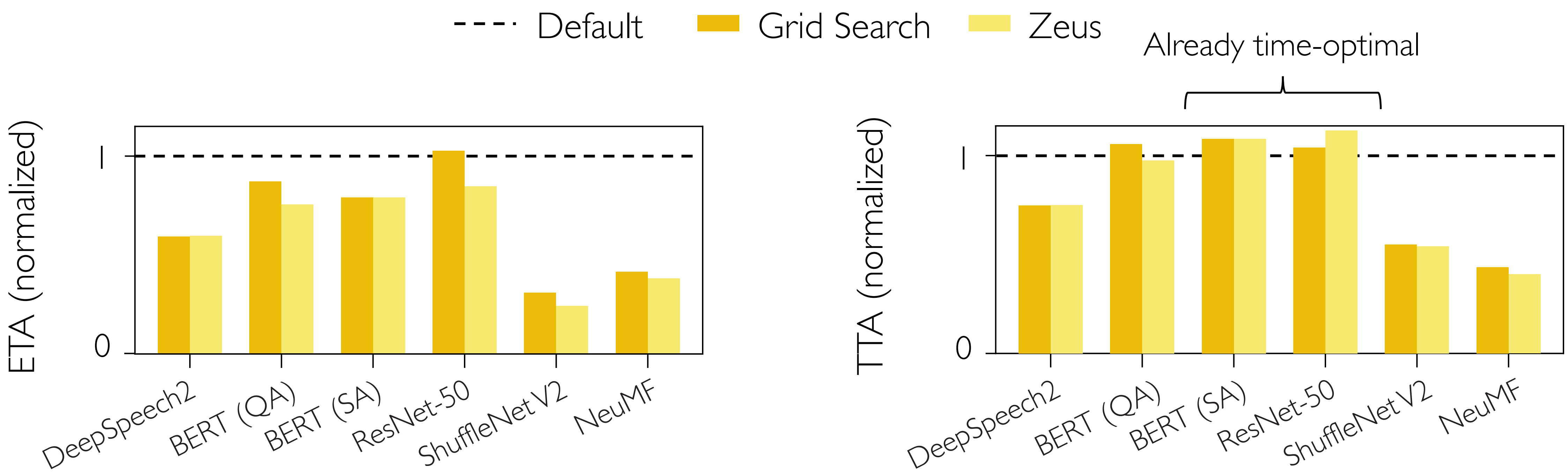




Search Path Training Cost (darker means better)

DeepSpeech2 trained on LibriSpeech on an NVIDIA V100 GPU.

Zeus Leads to Large Benefits

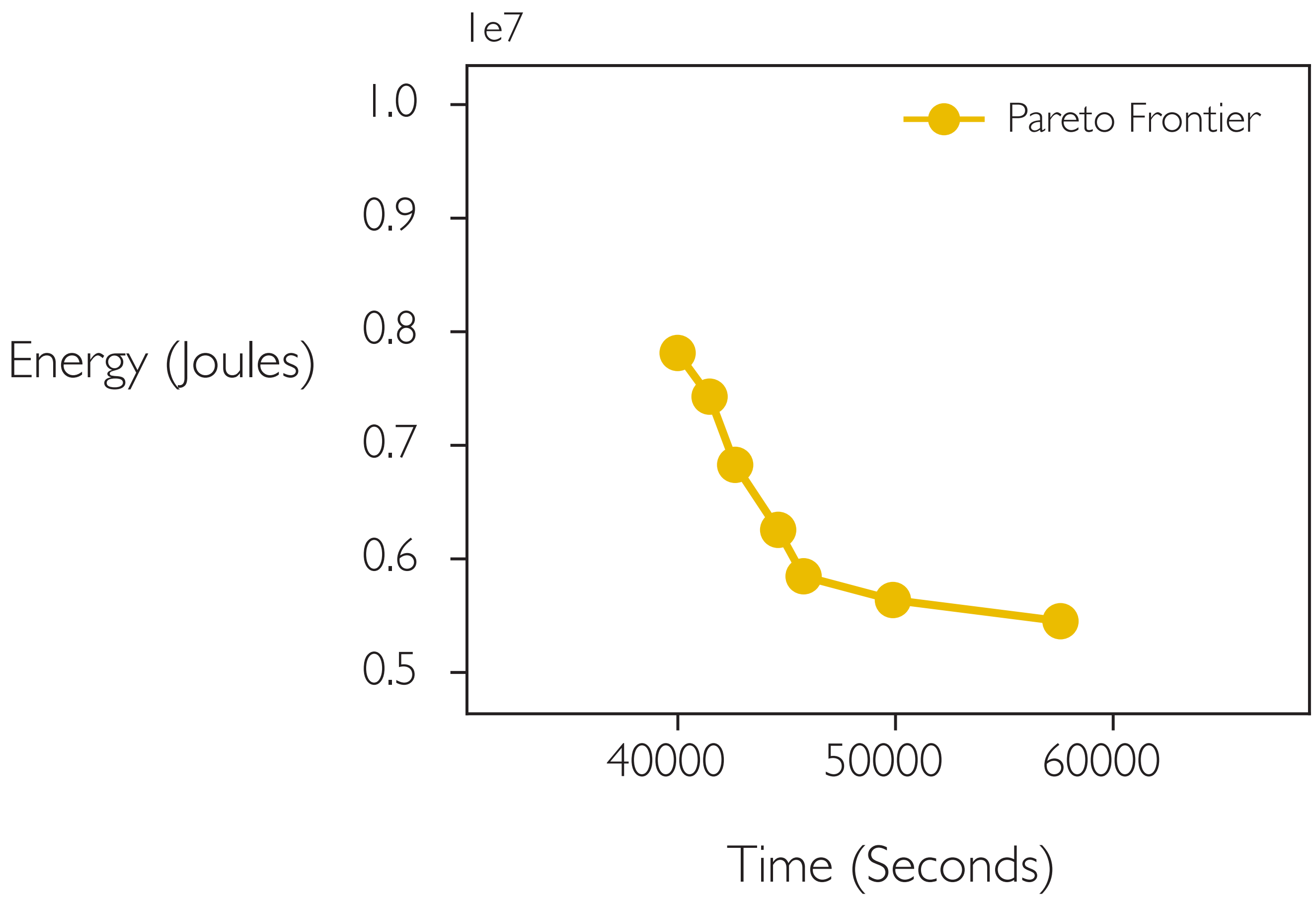


15 ~ 76% energy reduction

Up to 60% time reduction

Results obtained on an NVIDIA V100 GPU

Is Zeus Good Enough for Large Models?



Energy Bloat

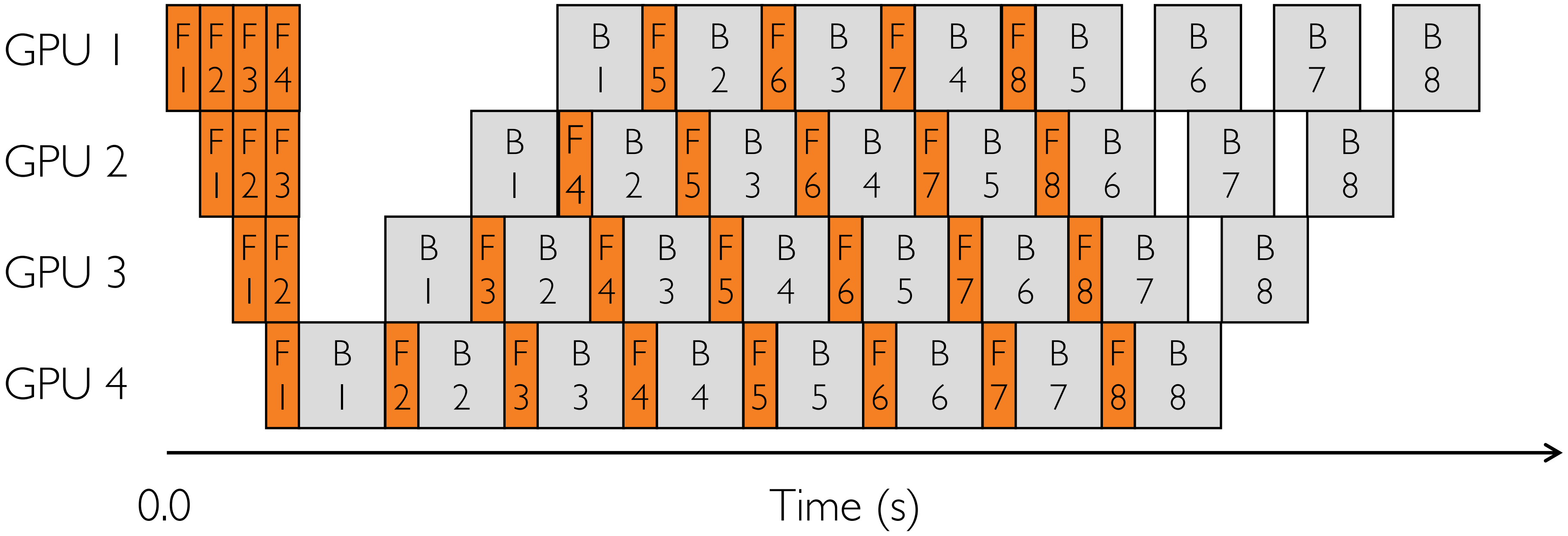
Not all Joules count

- A portion of energy doesn't contribute to throughput
- Removing such energy bloat doesn't affect throughput

Two sources of energy bloat

- Intrinsic to one training pipeline
- Extrinsic to one training pipeline

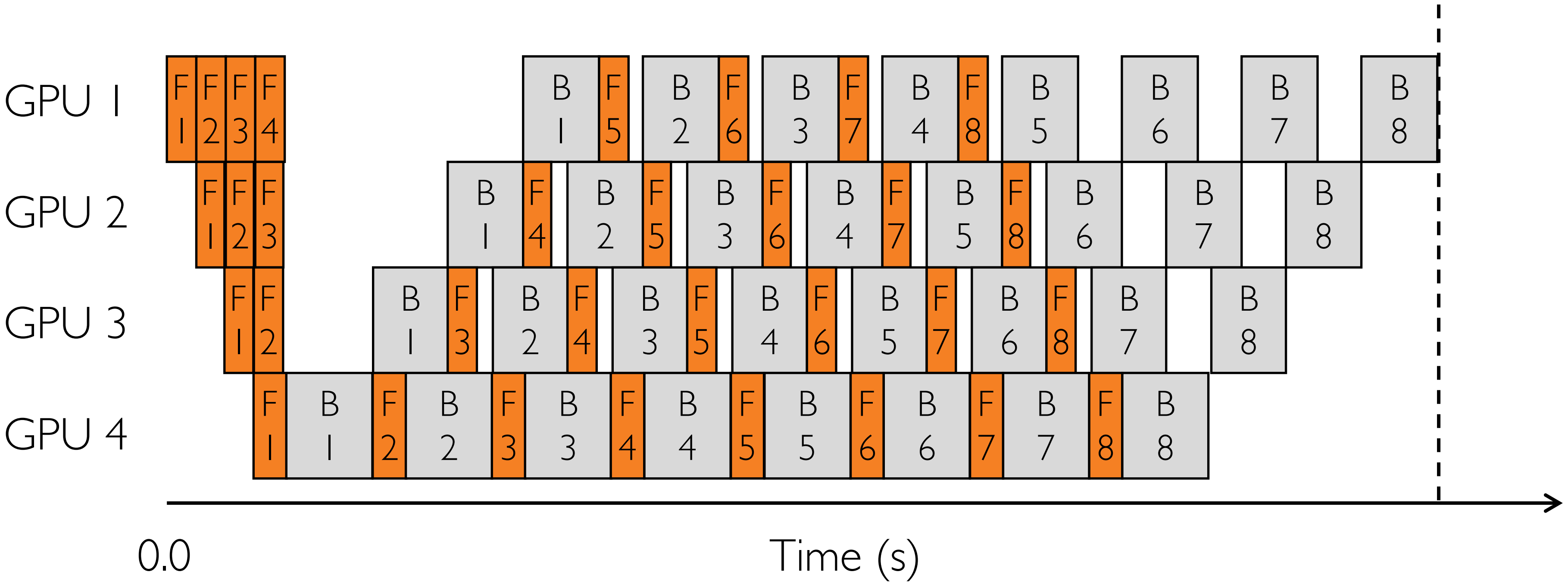
Intrinsic Energy Bloat



F = Forward, B = Backward

Intrinsic Energy Bloat

Some computations run at maximum speed and waste energy

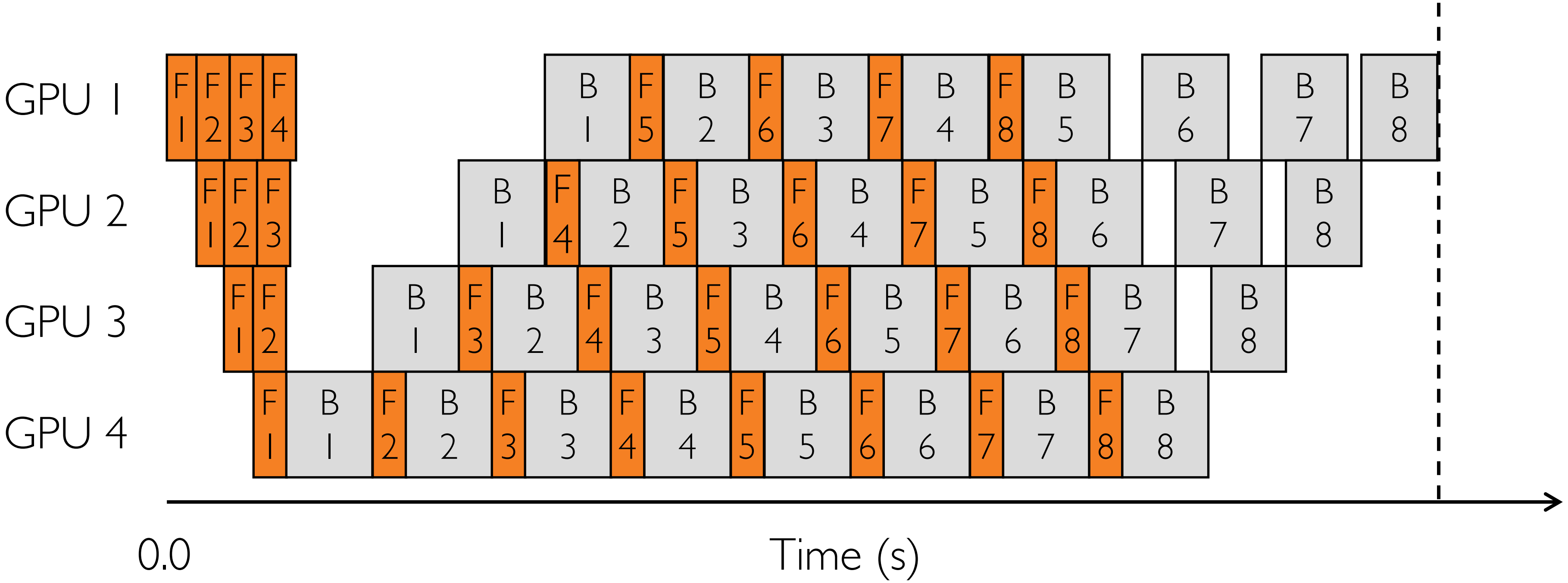


F = Forward, B = Backward

Drawn to scale for GPT-3, measured on NVIDIA A40 GPUs.

Intrinsic Energy Bloat

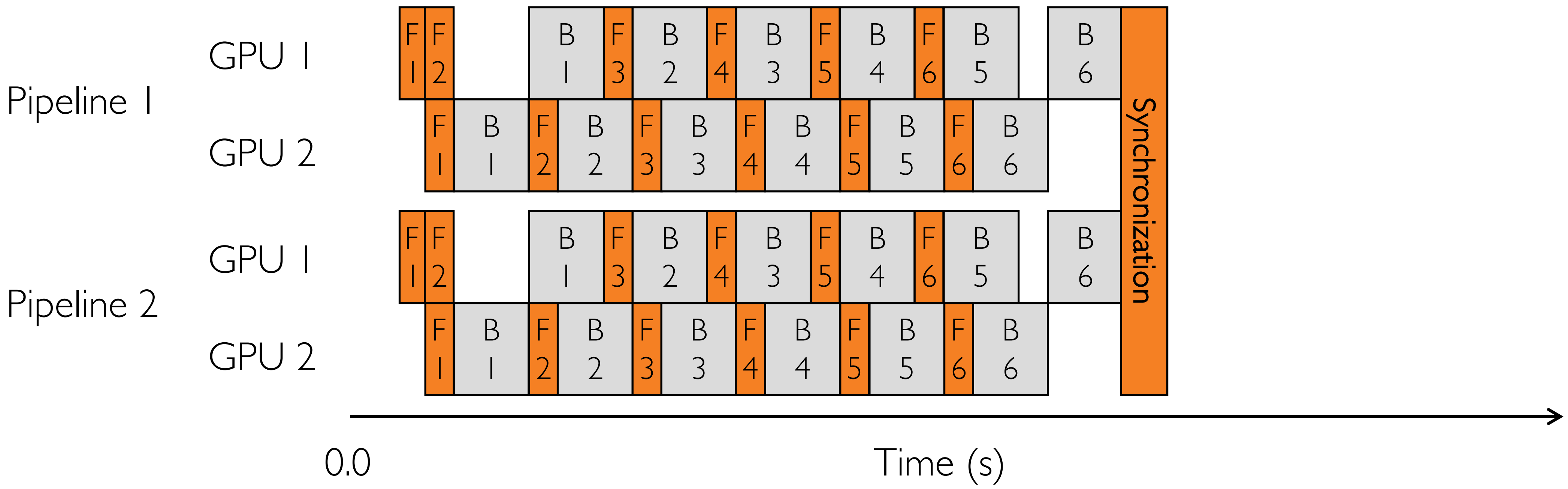
Some computations run at maximum speed and waste energy



F = Forward, B = Backward

Drawn to scale for GPT-3, measured on NVIDIA A40 GPUs.

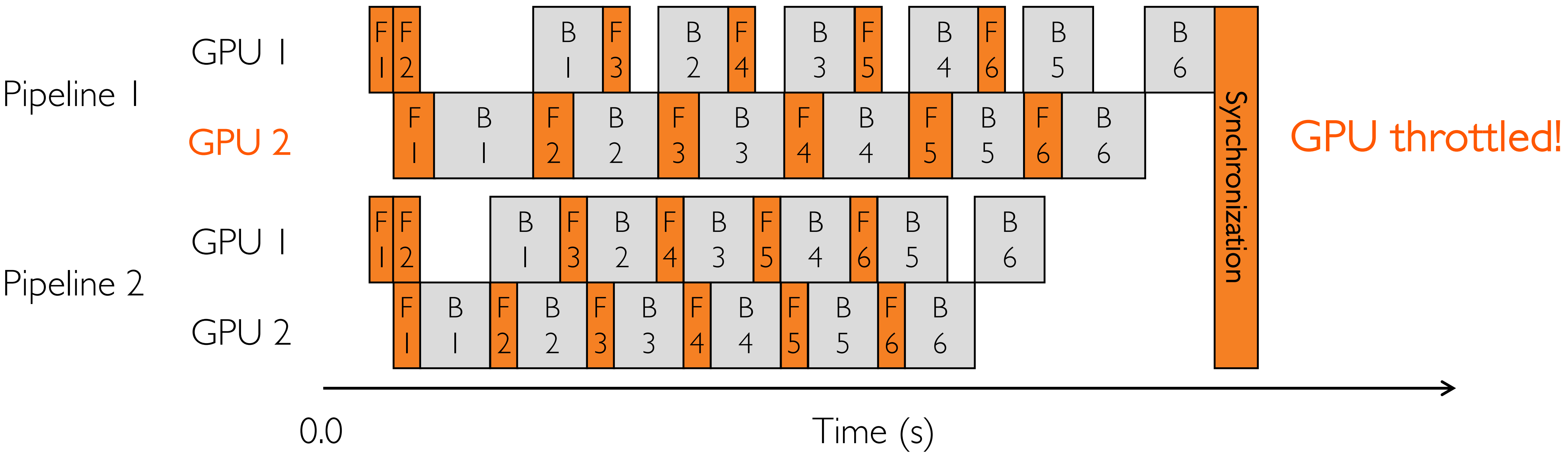
Extrinsic Energy Bloat



F = Forward, B = Backward

Extrinsic Energy Bloat

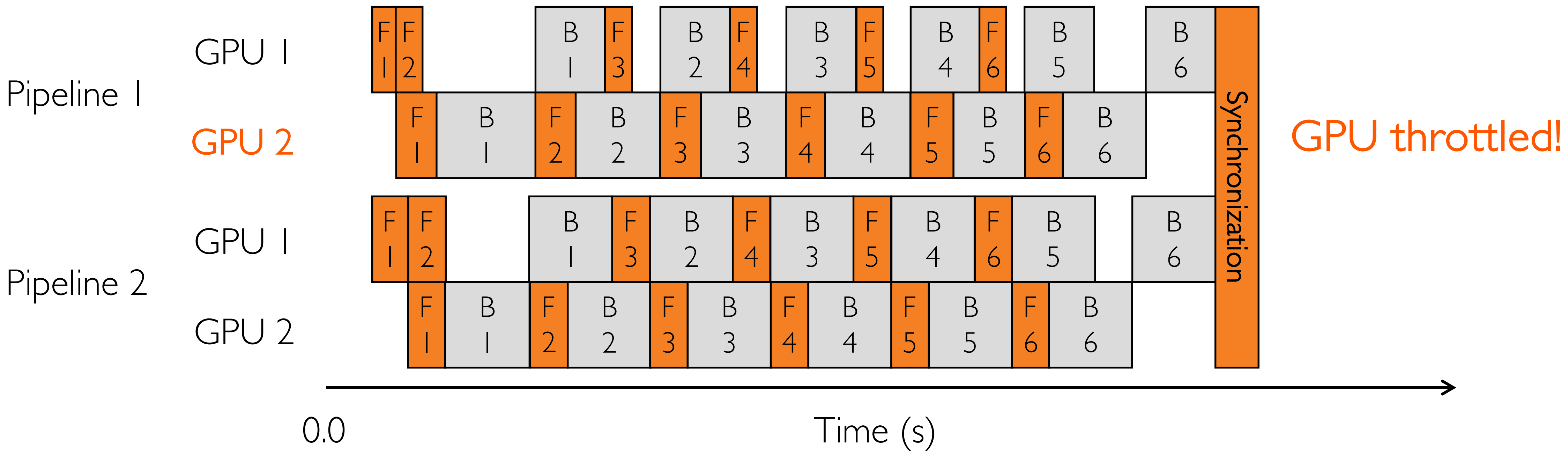
Numerous causes of stragglers in large scale training



F = Forward, B = Backward

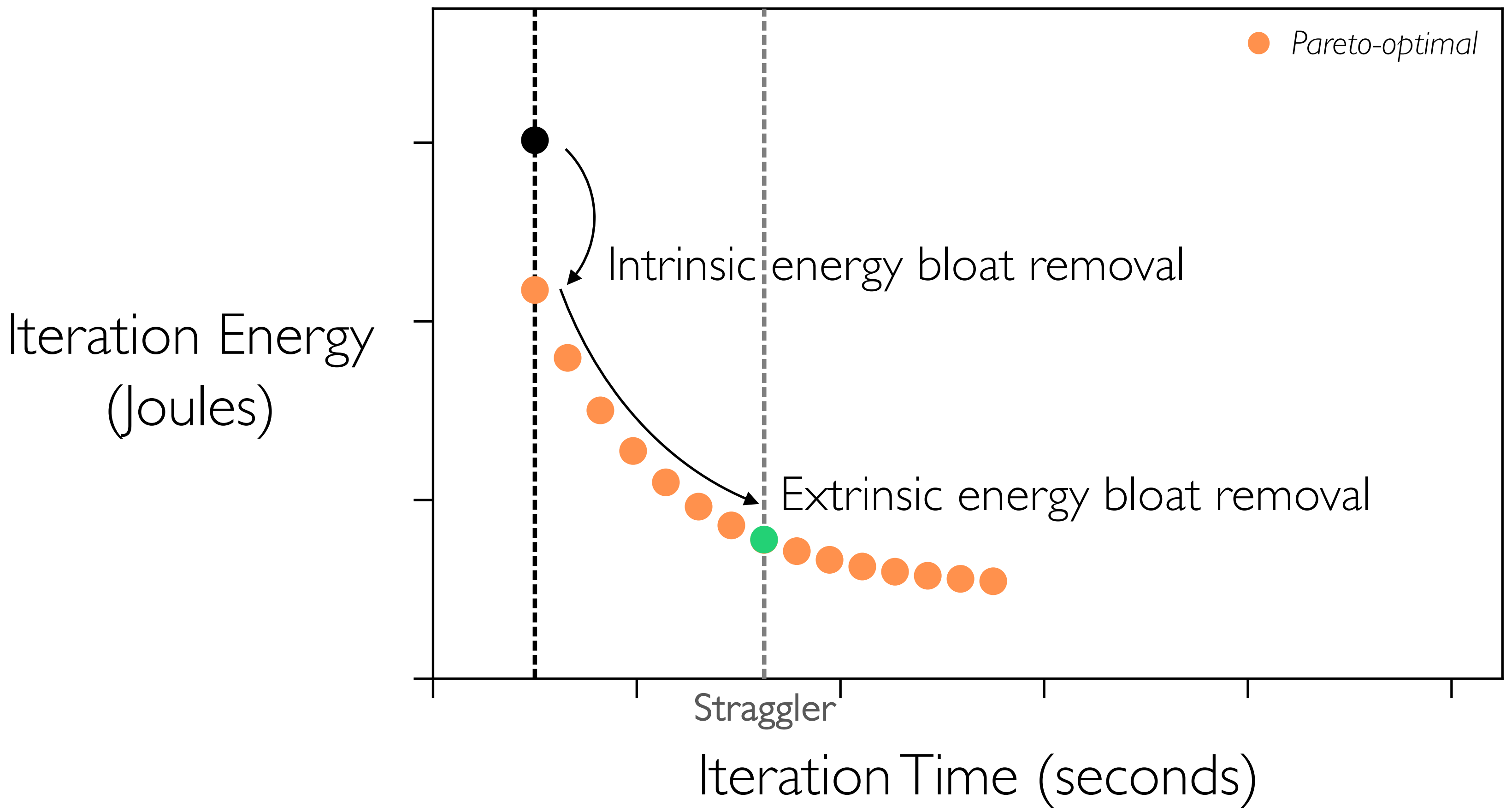
Extrinsic Energy Bloat

Numerous causes of stragglers in large scale training

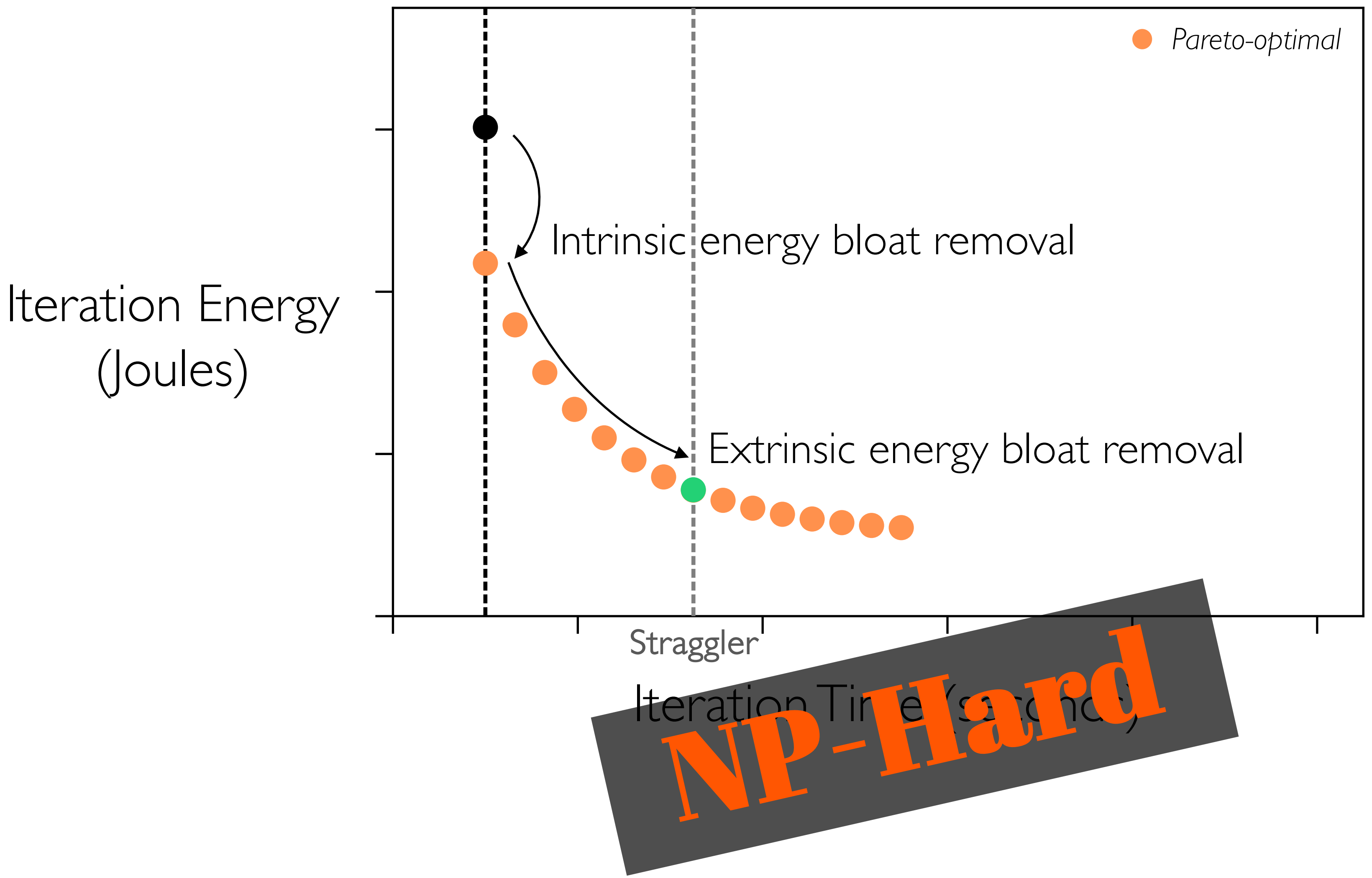


F = Forward, B = Backward

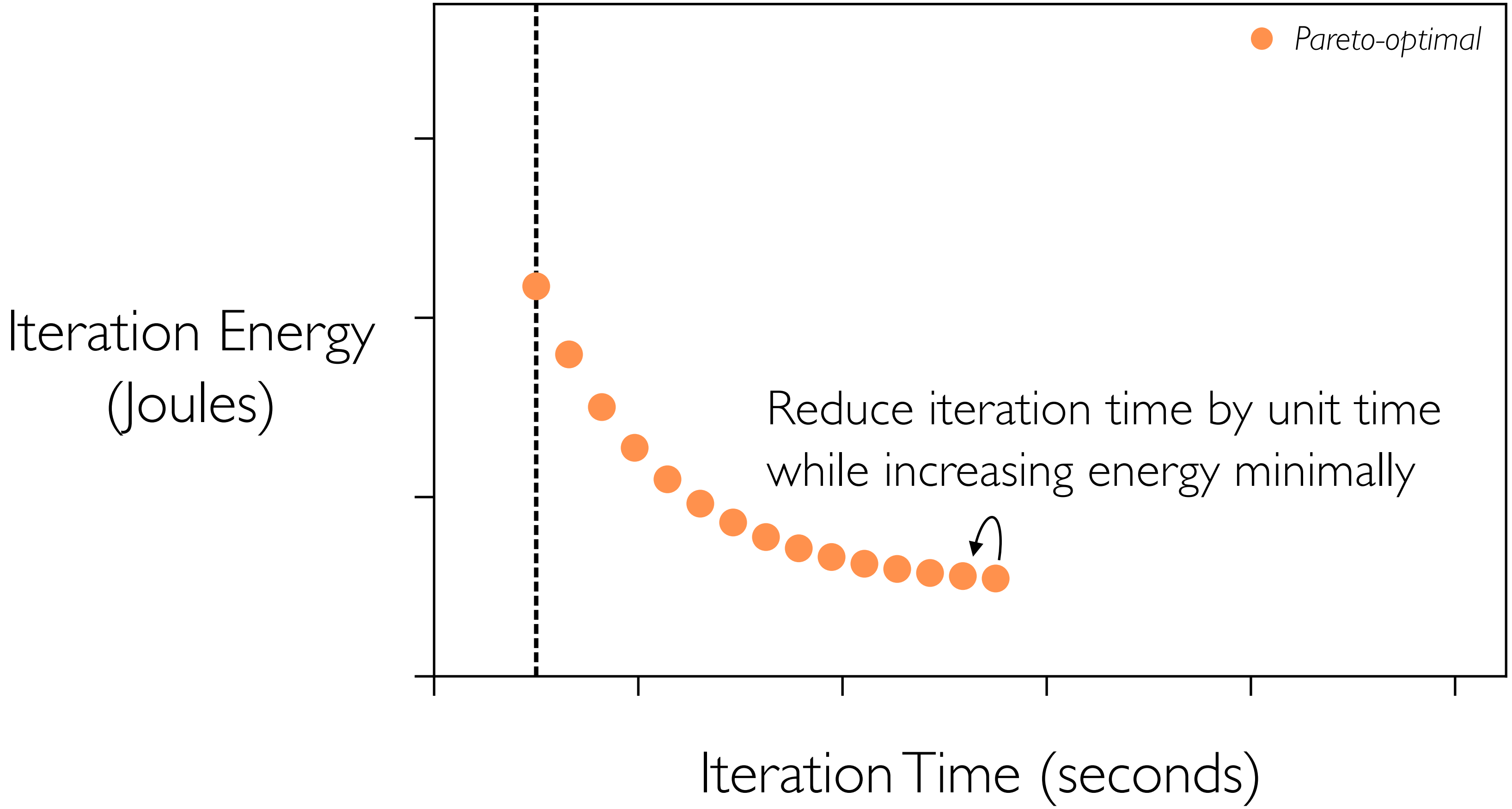
Iteration Time-Energy Pareto Frontier



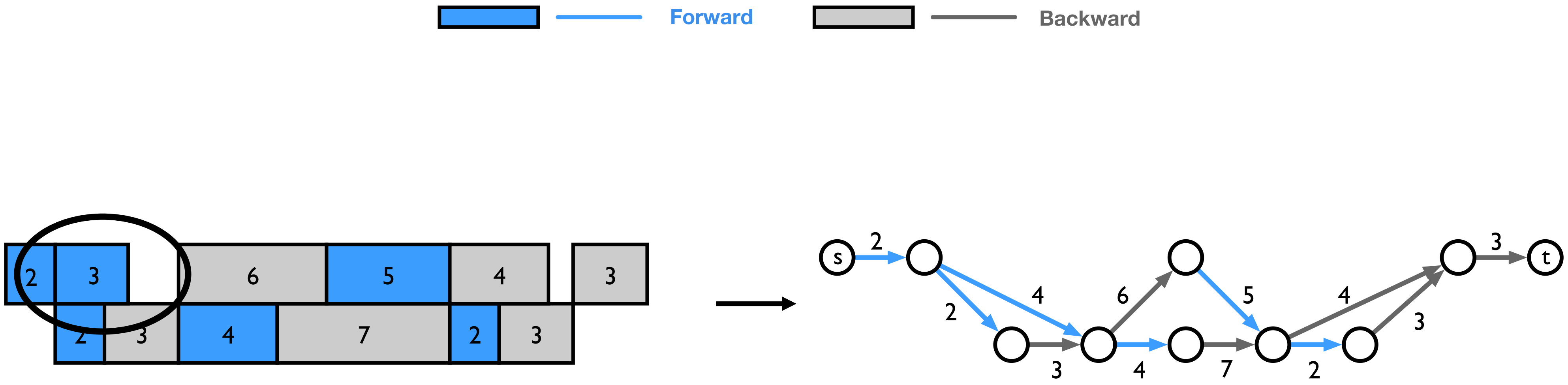
Iteration Time-Energy Pareto Frontier



An Iterative Solution

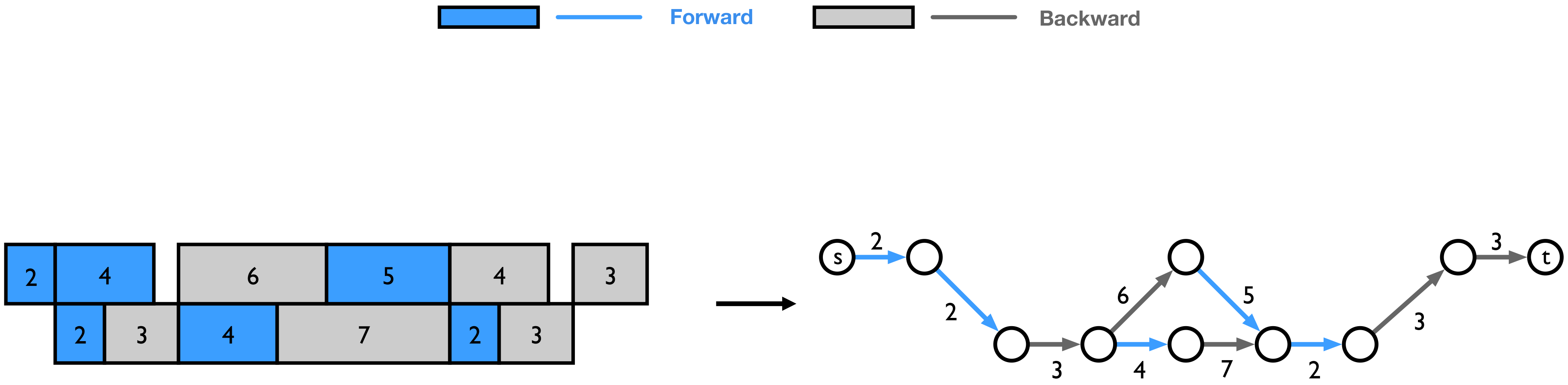


Reducing Time with Minimal Energy Increase



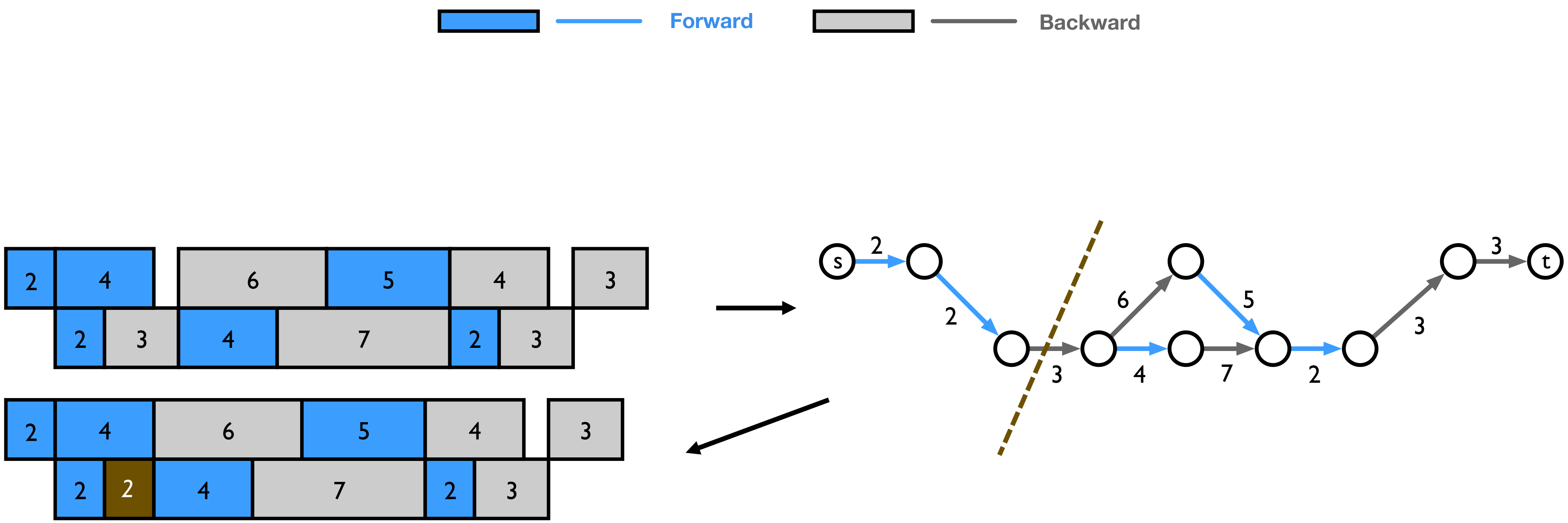
Only leave *critical* edges (computations)

Reducing Time with Minimal Energy Increase



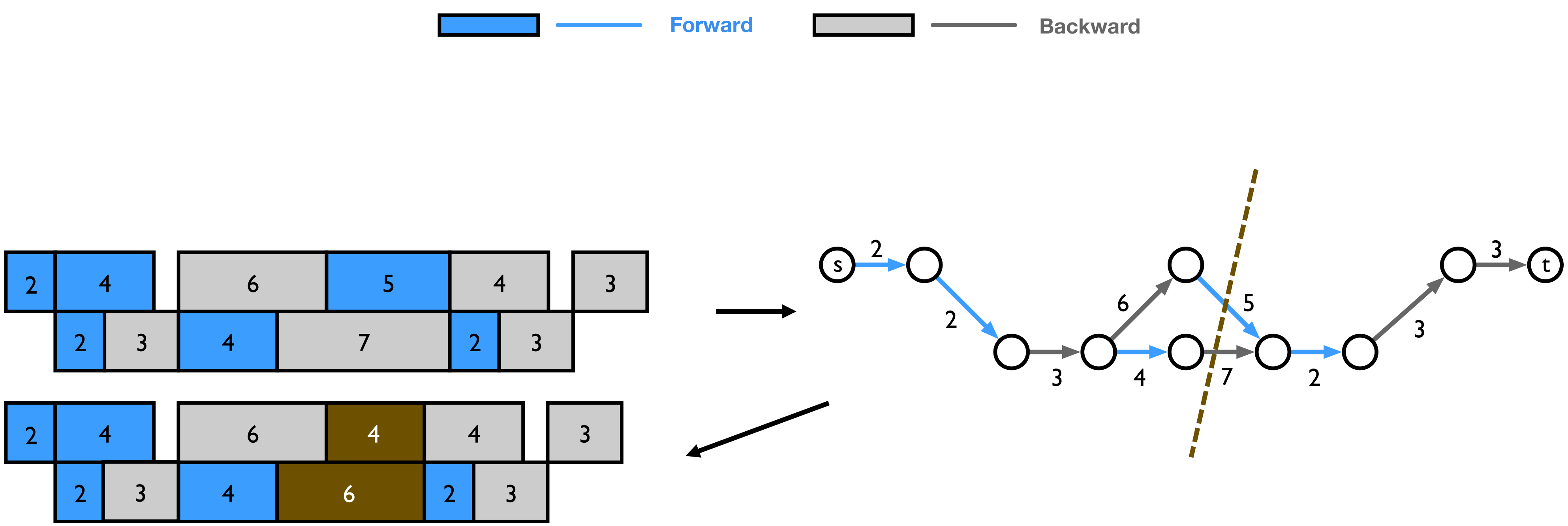
Only leave *critical* edges (computations)

Reducing Time with Minimal Energy Increase



Any *s-t cut* represents a way to
reduce the DAG's end-to-end execution time by 1

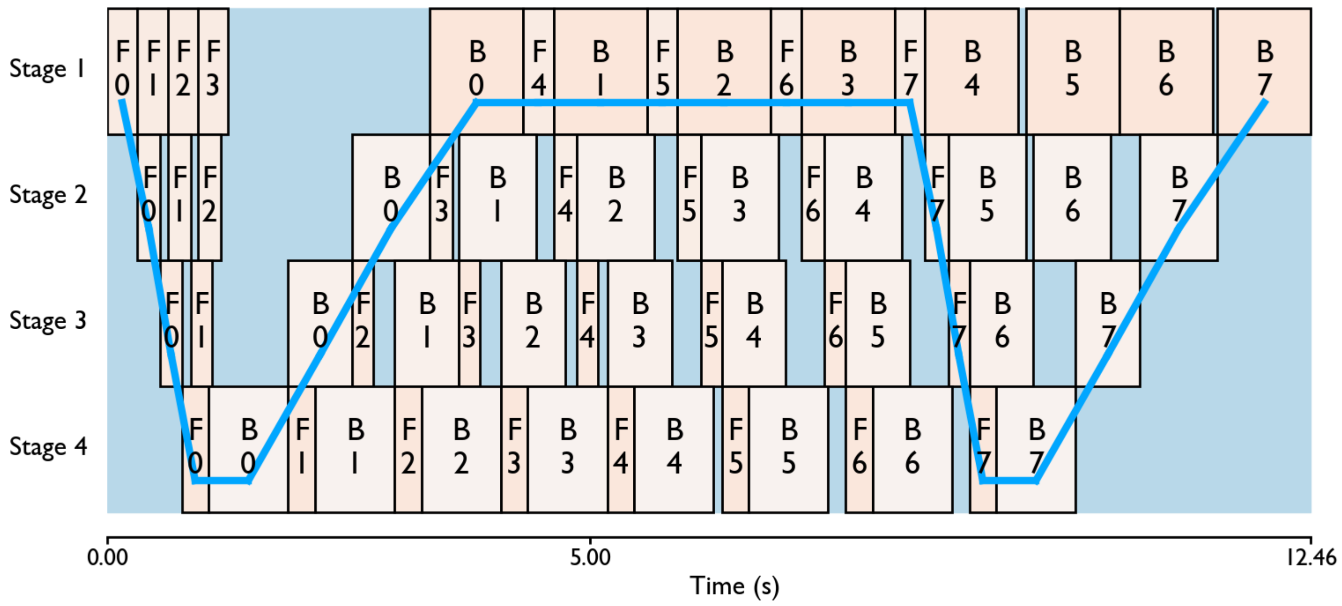
Reducing Time with Minimal Energy Increase

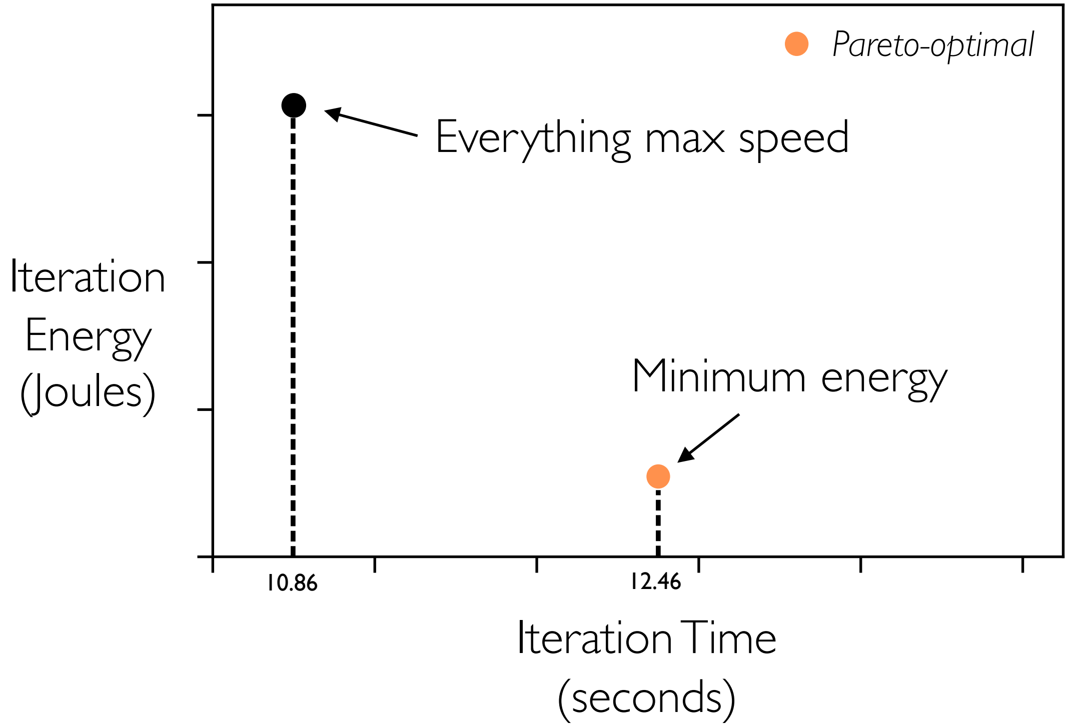


Any *s-t cut* represents a way to reduce the DAG's end-to-end execution time by 1

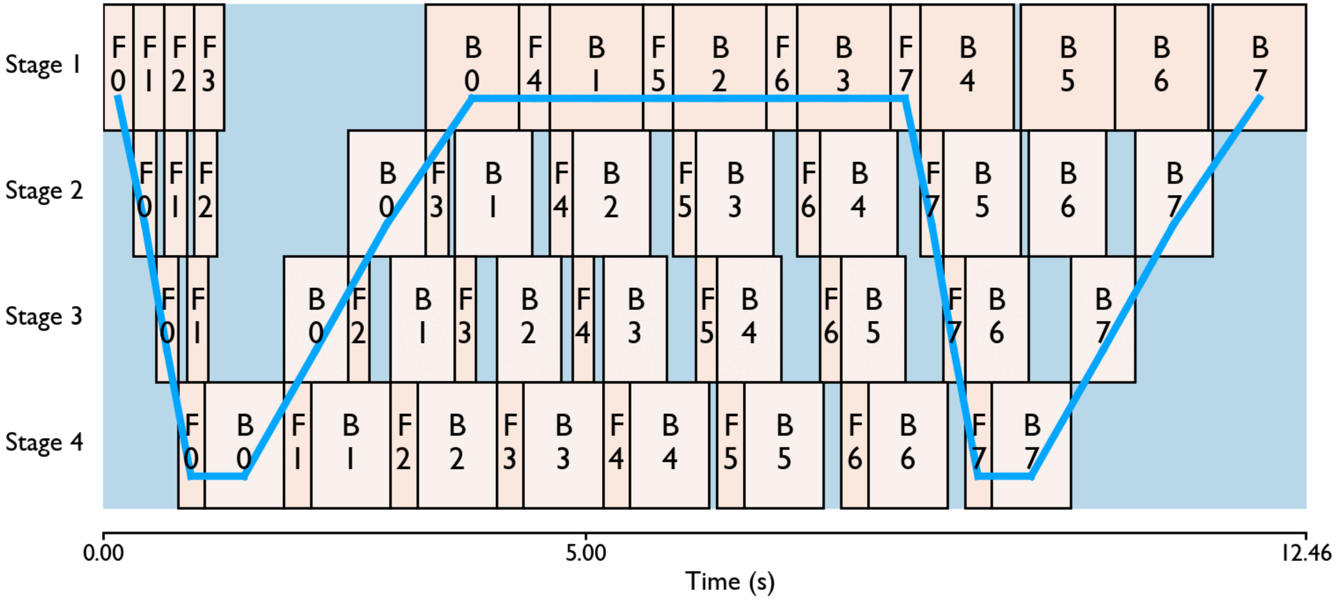
Edge cut capacity ⇔ Energy increase

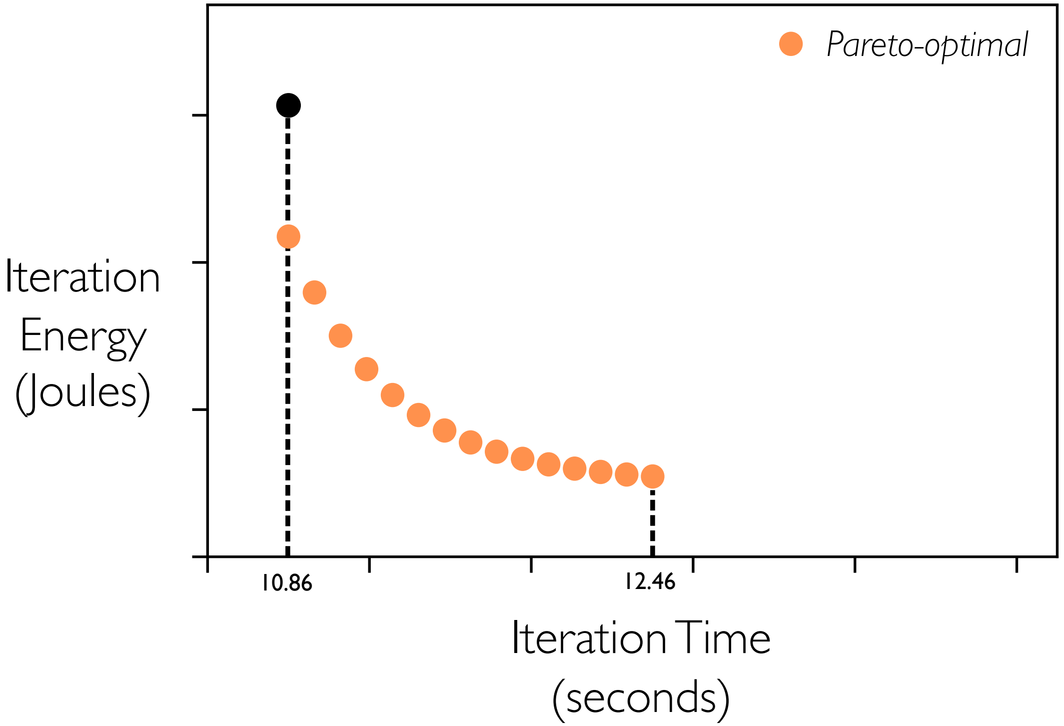
Perseus in Action



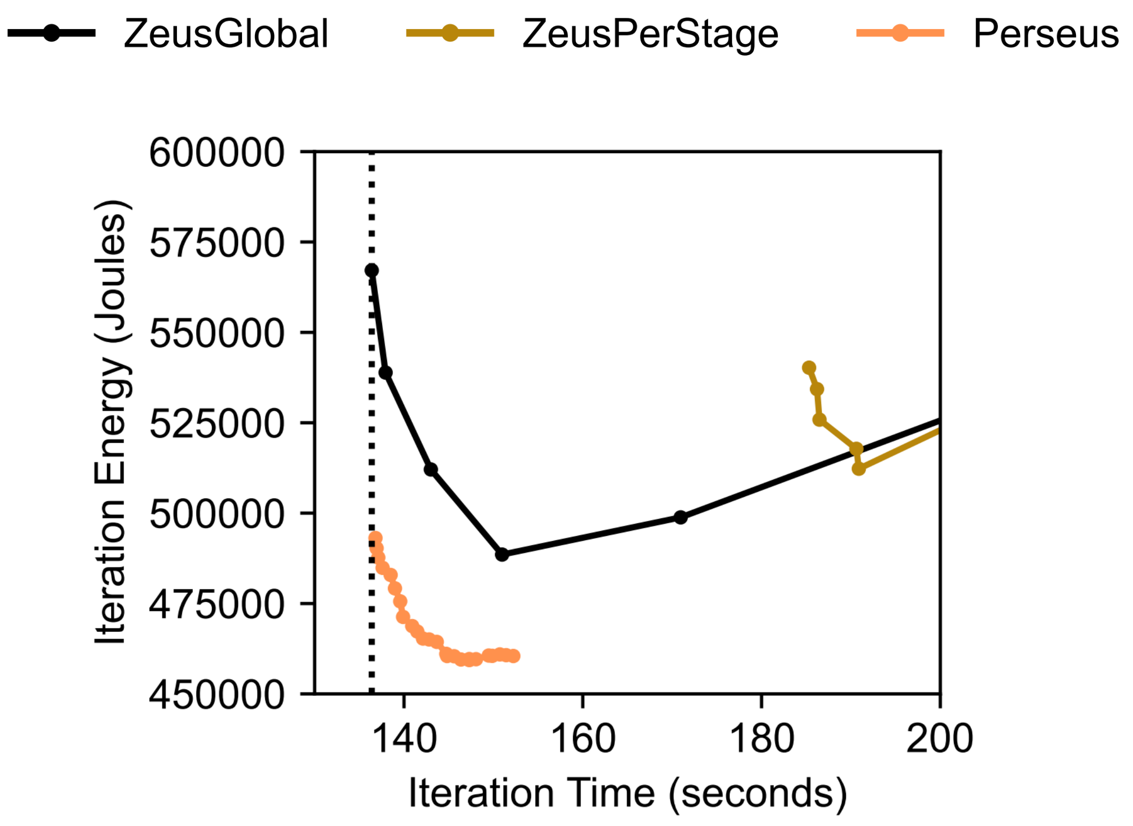


Perseus in Action





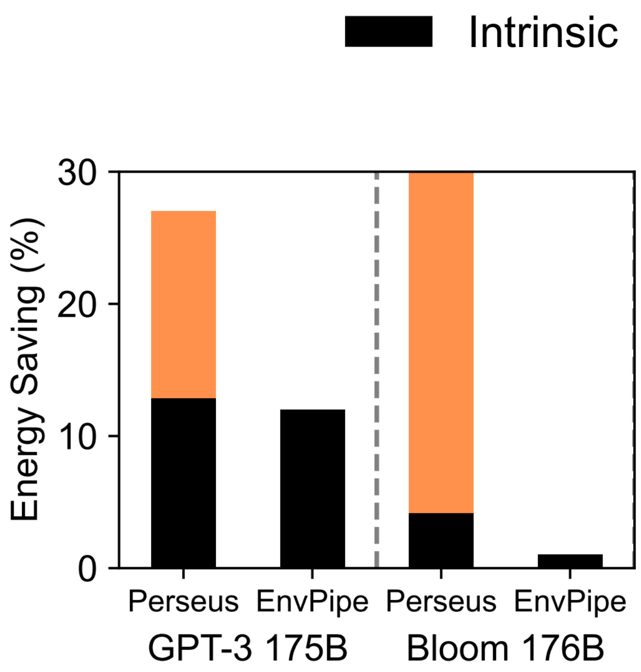
Perseus Pushes the Frontier



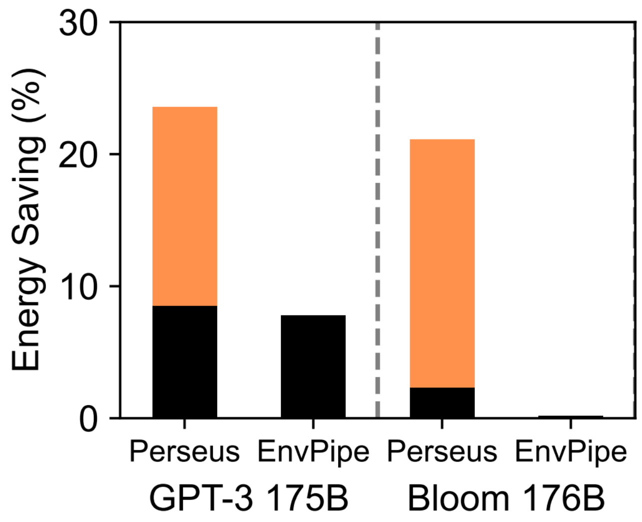
GPT-3 6.7B

NVIDIA A40 GPUs

Perseus Pushes the Frontier



NVIDIA A100 GPUs



NVIDIA A40 GPUs

Conclusion

Power is a growing bottleneck for data centers
that deserves careful management

Energy is a new first-class software systems metric
that is worth optimizing

We're always looking for great collaboration!
https://ml.energy

SymbioticLab ML.ENERGY UNIVERSITY OF MICHIGAN