Reducing Energy Bloat in Large Model Training

Jae-Won Chung

With Yile Gu, Insu Jang, Luoxi Meng, Nikhil Bansal, and Mosharaf Chowdhury

SymbioticLab

ML.ENERGY

UNIVERSITY OF MICHIGAN

Why AI Energy?

Energy demand of AI

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)

Why AI Energy?

Energy demand of AI

Datacenter power delivery


Global Data Center Trends 2023

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

CBRE RESEARCH
JULY 2023

Why AI Energy?

Energy demand of AI

Datacenter power delivery


Global Data Center Trends

2024

Limited power availability drives rental rate growth worldwide

CBRE RESEARCH
JUNE 2024

Why AI Energy?

Energy demand of AI

Datacenter power delivery

SEPTEMBER 12, 2024

Readout of White House Roundtable on U.S. Leadership in AI Infrastructure

BRIEFING ROOM ▸ STATEMENTS AND RELEASES

Today, as part of the Biden-Harris Administration's comprehensive strategy for responsible innovation, the White House convened leaders from hyperscalers, artificial intelligence (AI) companies, datacenter operators, and utility companies to discuss steps to ensure the United States continues to lead the world in AI. Participants considered strategies to meet clean energy, permitting, and workforce requirements for developing large-scale AI datacenters and power infrastructure needed for advanced AI operations in the United States.

Our Goal

Let's optimize the energy consumption of large model training

- without changing what is being computed
- on the same GPU hardware
- without slowdown

Energy Bloat

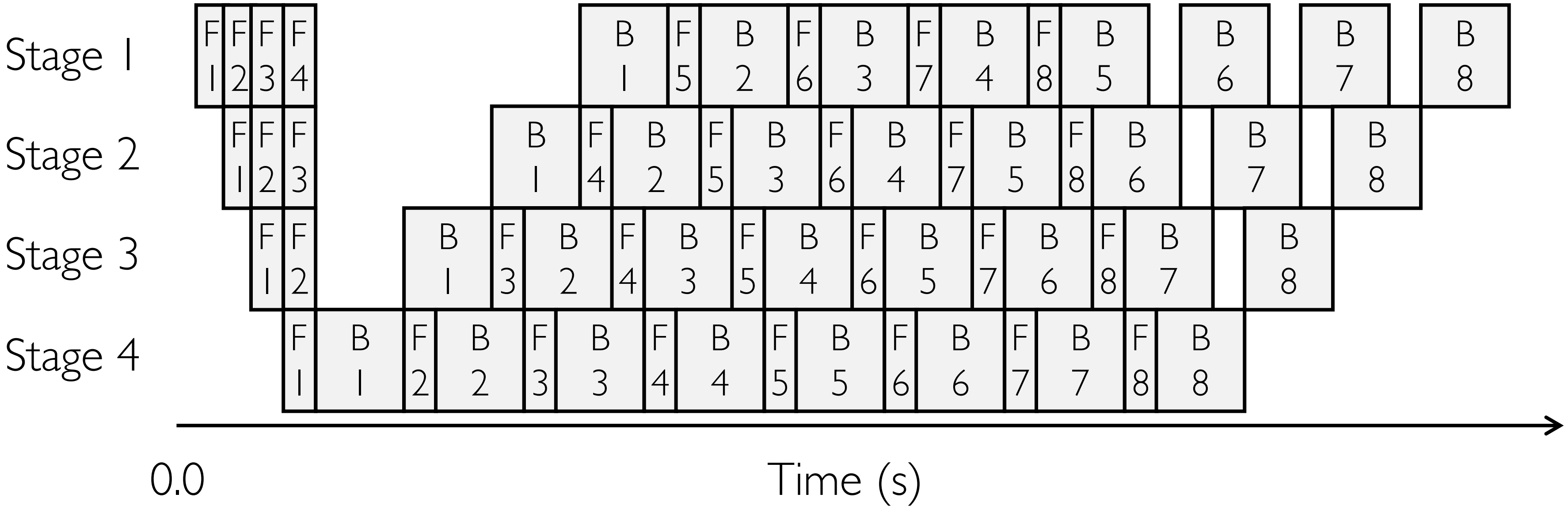
Not all Joules are equal

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

Two sources of energy bloat in large model training

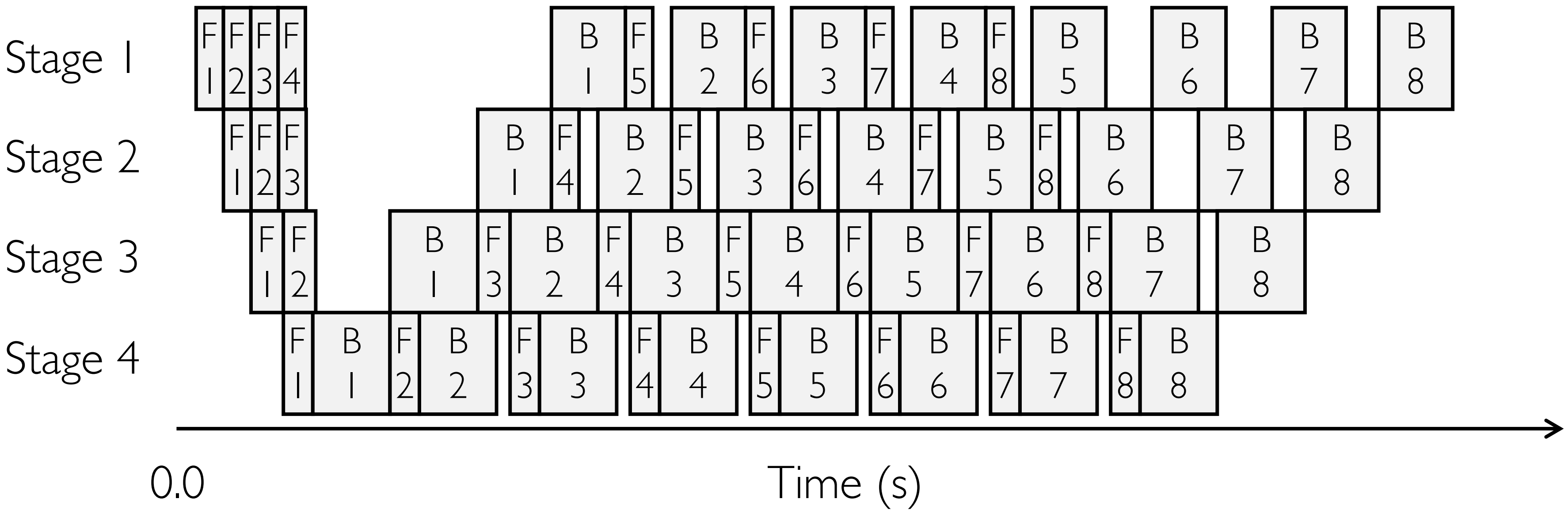
- Intrinsic to one training pipeline
- Extrinsic to one training pipeline

Intrinsic Energy Bloat



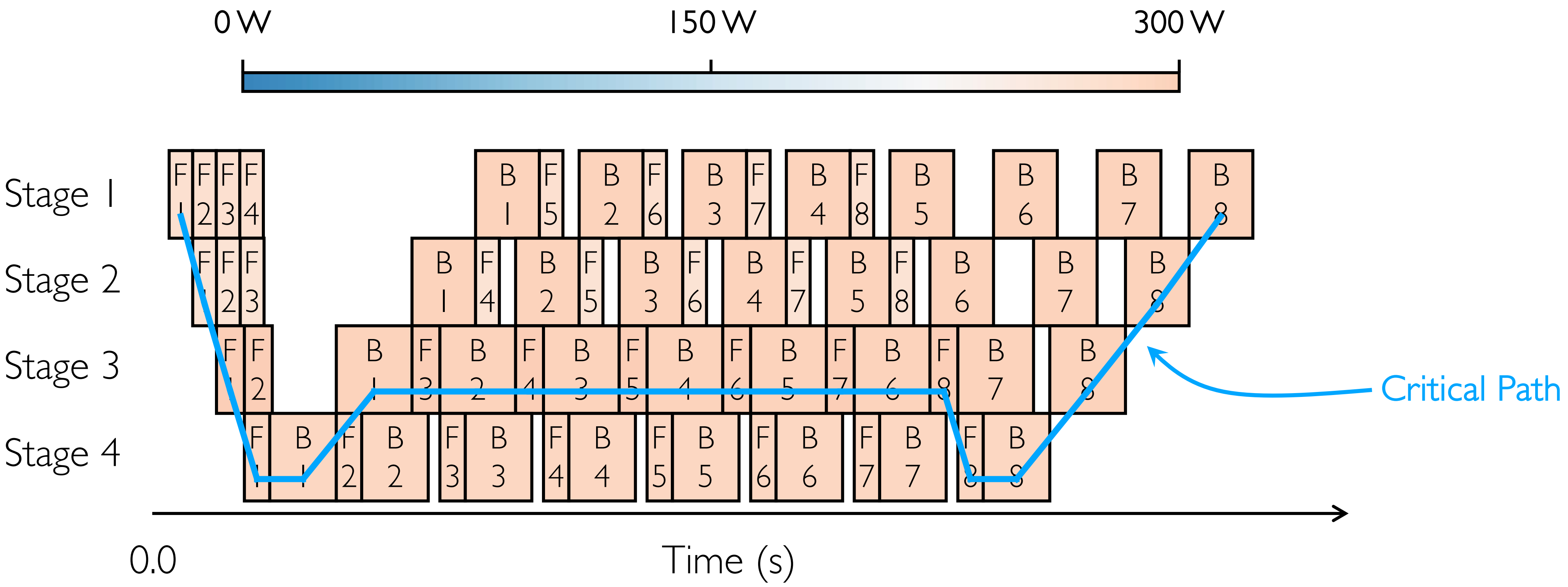
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).

Intrinsic Energy Bloat



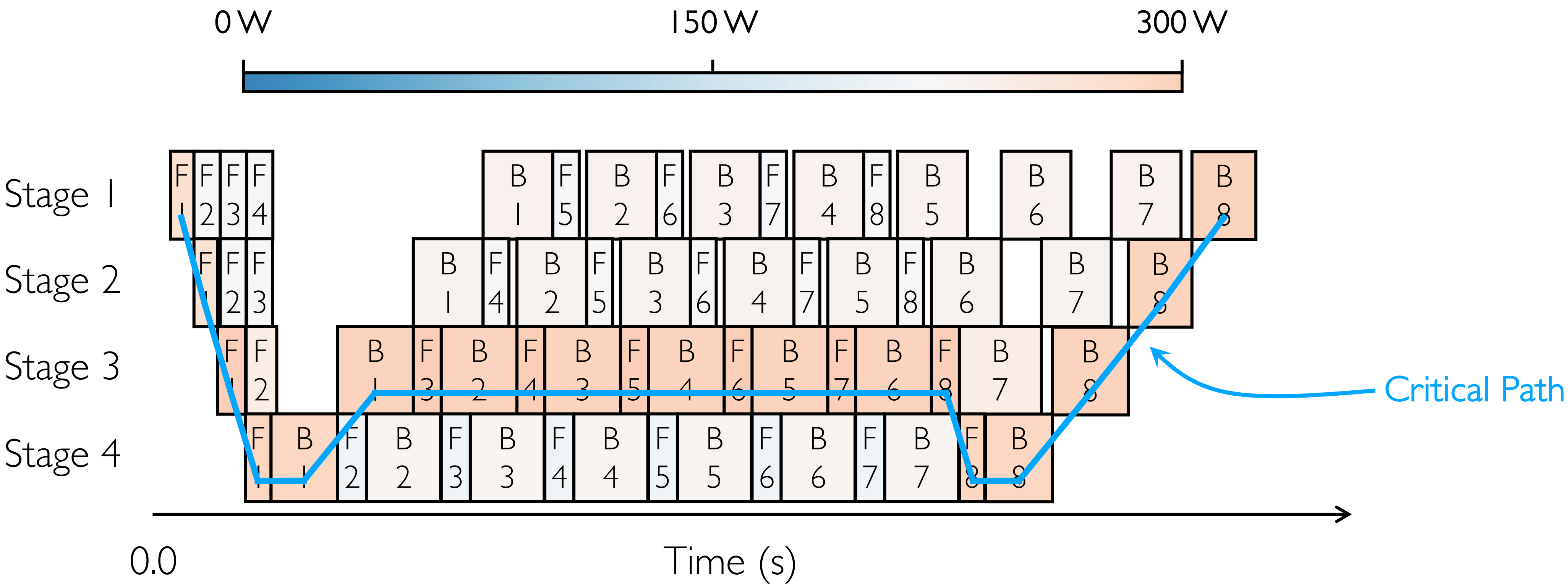
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).
Drawn to scale for GPT-3 1.3B on NVIDIA A100 GPUs.

Intrinsic Energy Bloat



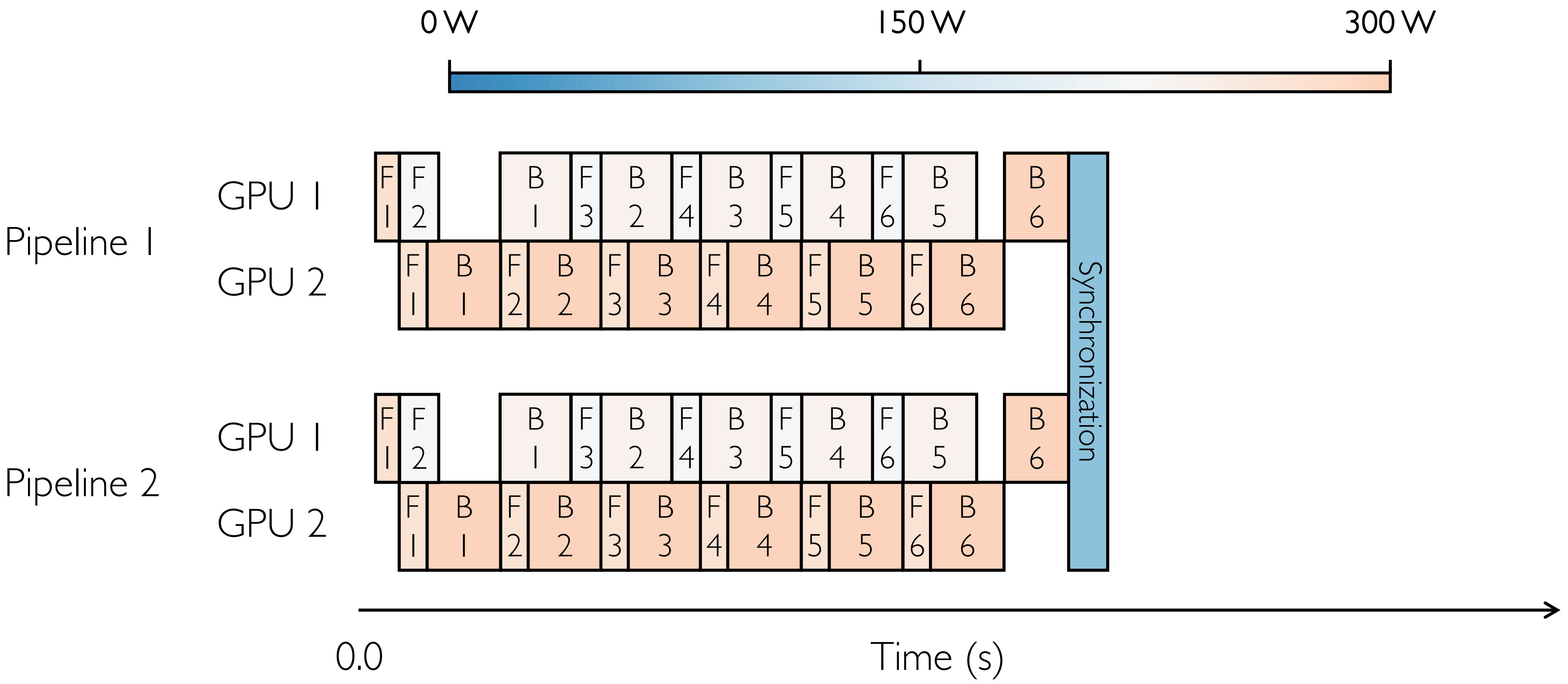
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).
Drawn to scale for GPT-3 1.3B on NVIDIA A100 GPUs.

Intrinsic Energy Bloat

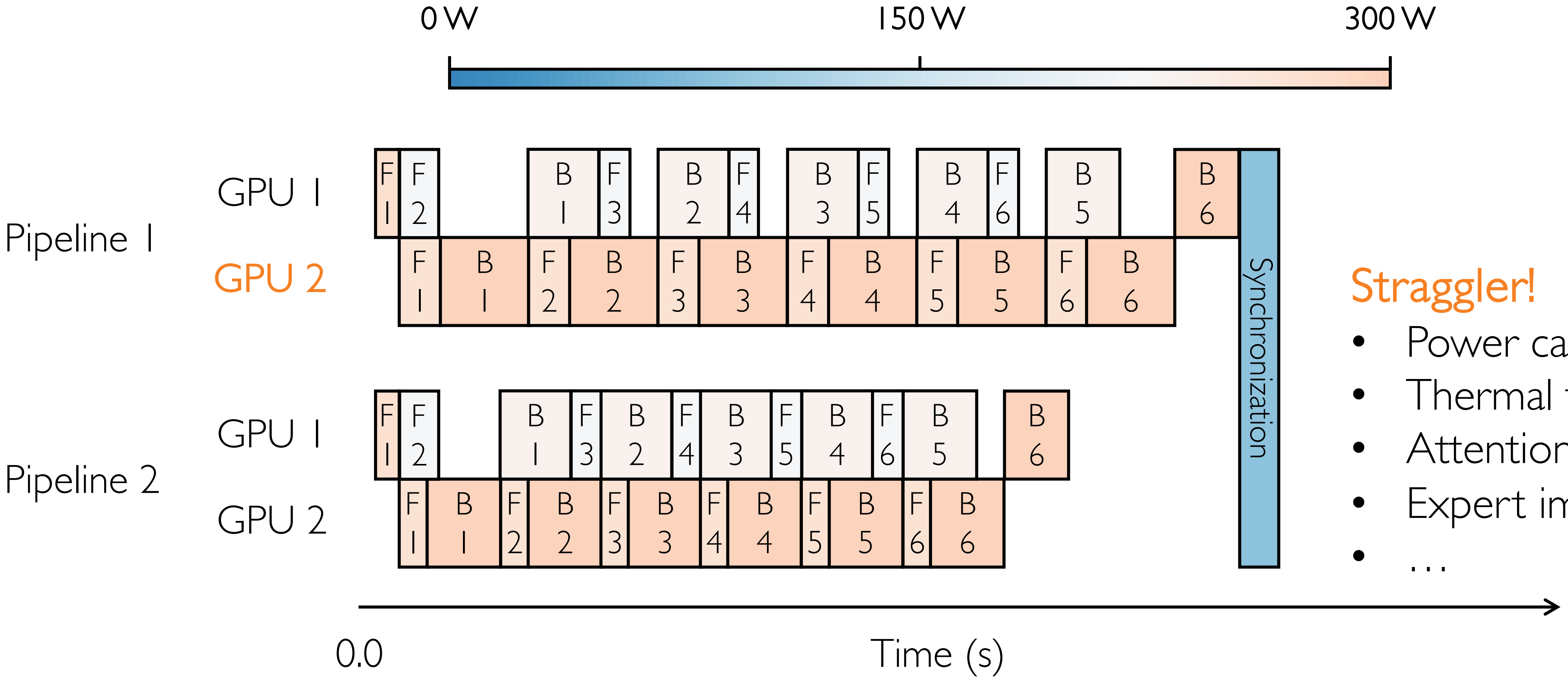


One training iteration of GPT-3 1.3B with four pipeline stages
and eight microbatches on NVIDIA A100 GPUs, drawn to scale.

Extrinsic Energy Bloat



Extrinsic Energy Bloat

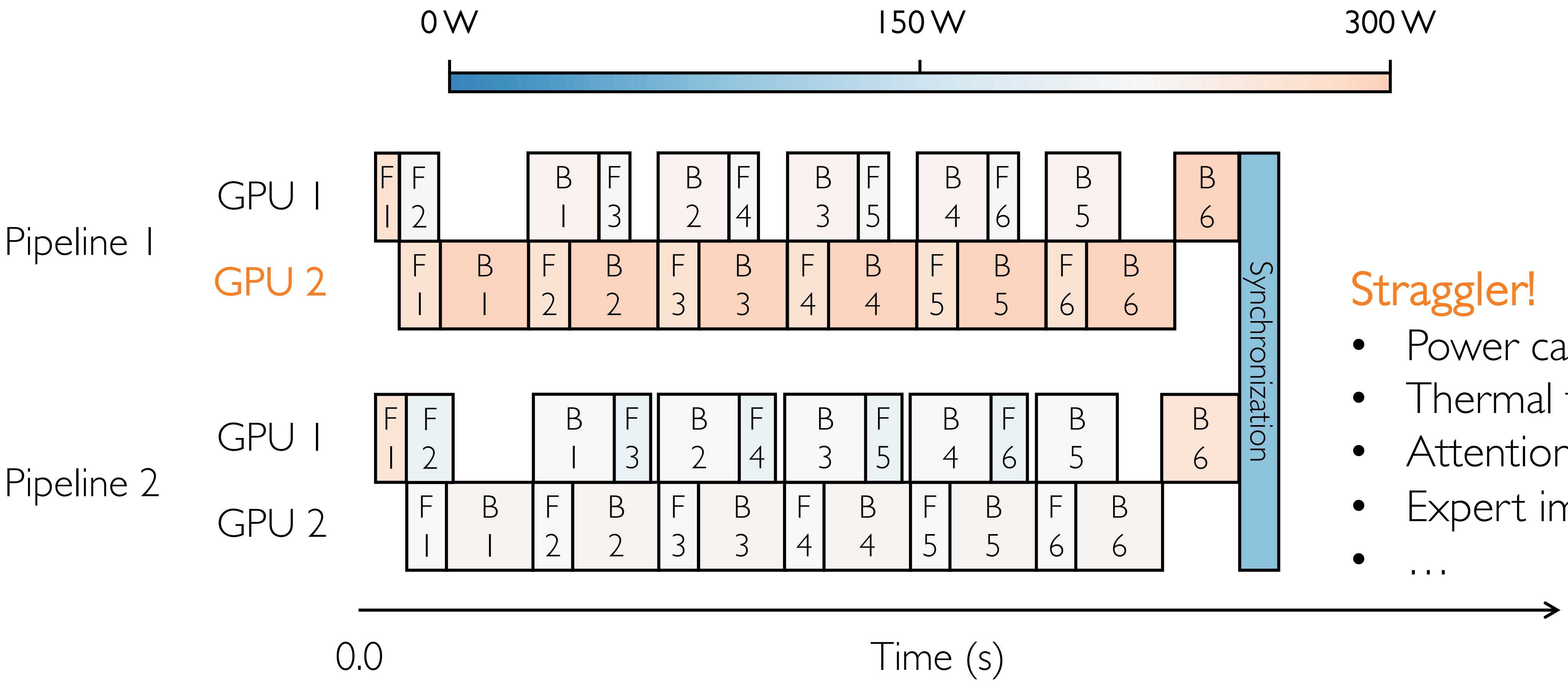


Straggler!

- Power capping
- Thermal throttling
- Attention masks
- Expert imbalance
- …

Stragglers in literature: MegaScale (NSDI '24), SuperBench (ATC '24), Llama 3 (Meta), Falcon (Alibaba)

Extrinsic Energy Bloat

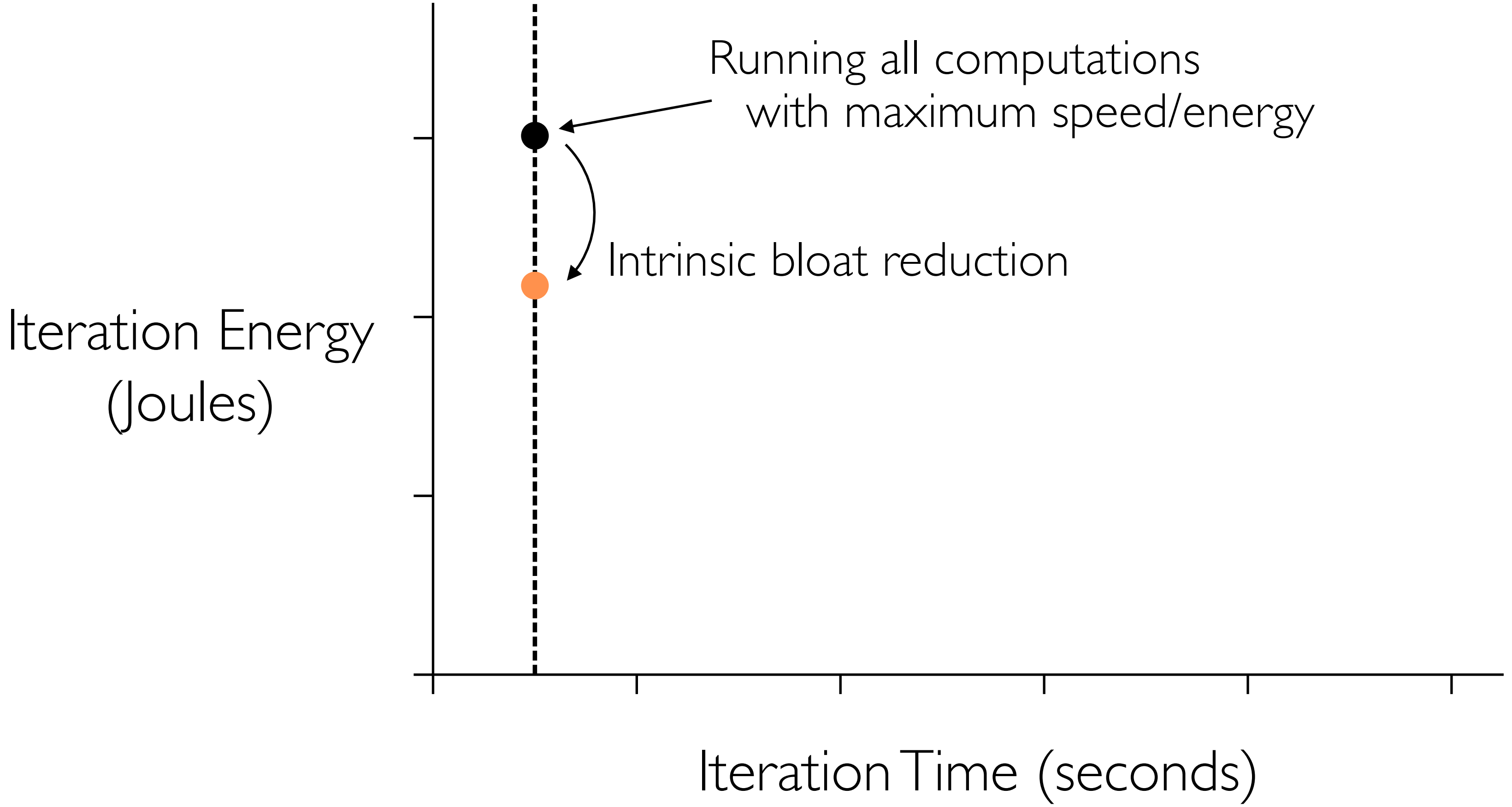


Straggler!

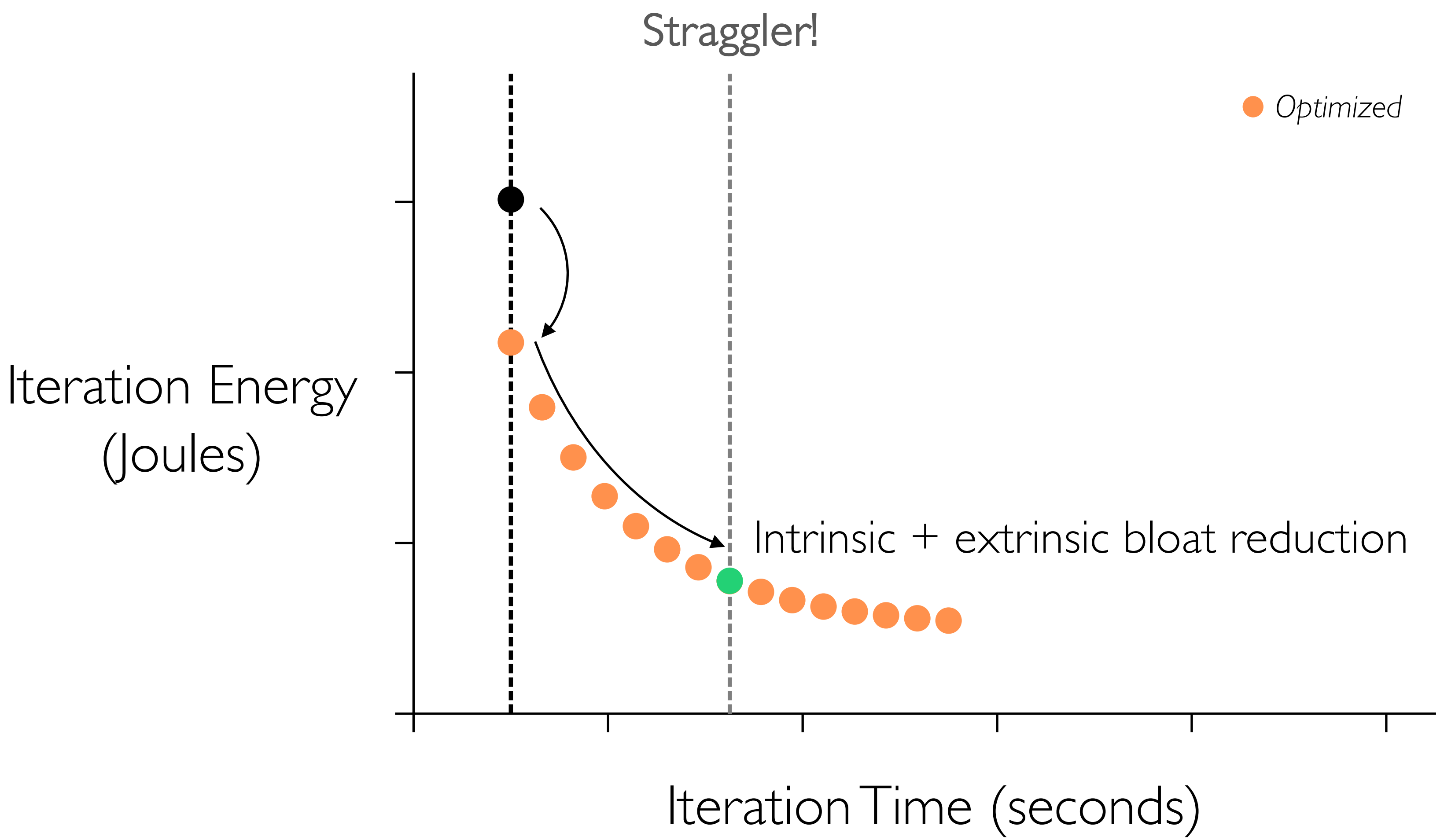
- Power capping
- Thermal throttling
- Attention masks
- Expert imbalance
- …

Stragglers in literature: MegaScale (NSDI '24), SuperBench (ATC '24), Llama 3 (Meta), Falcon (Alibaba)

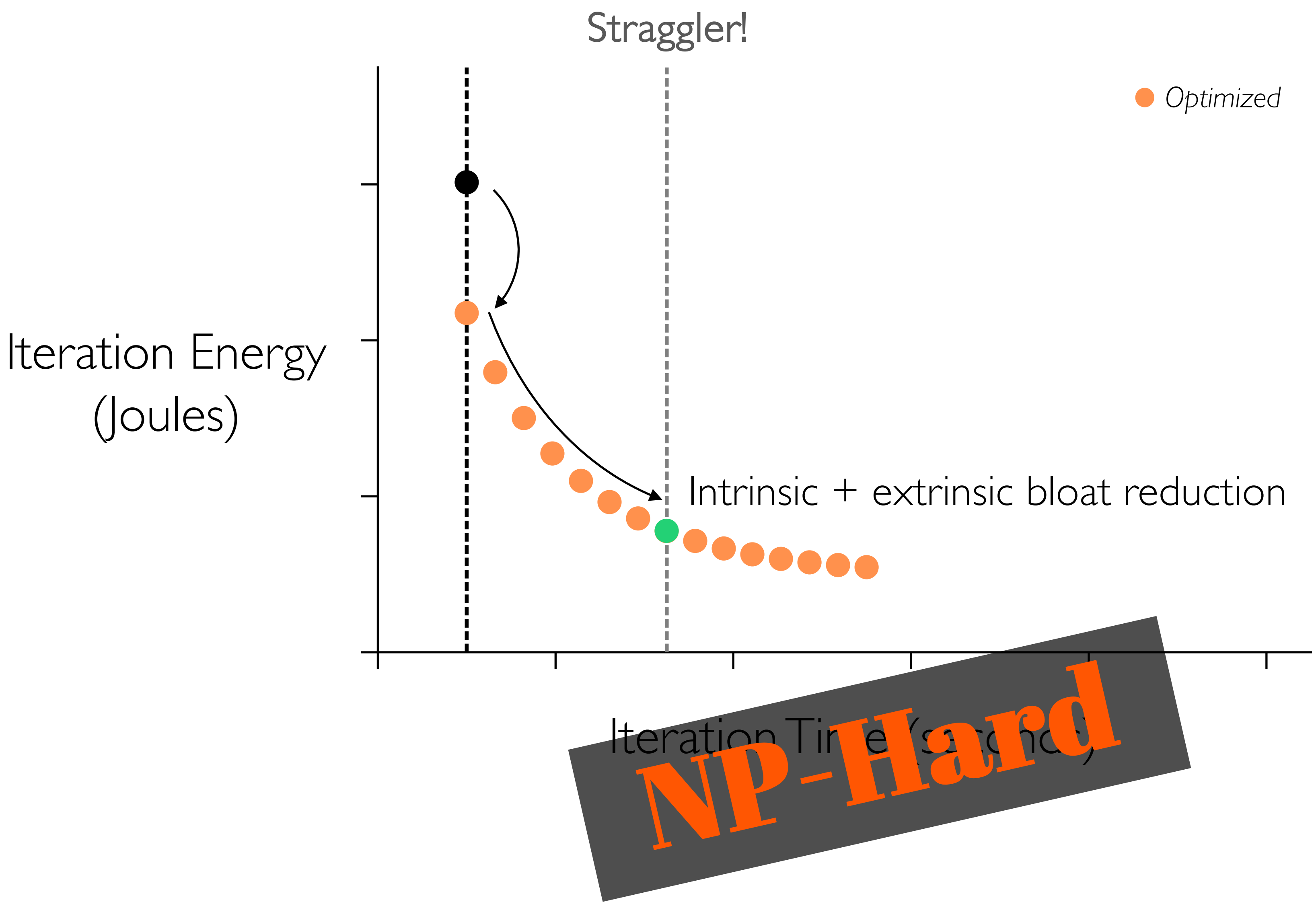
Iteration Time-Energy Frontier



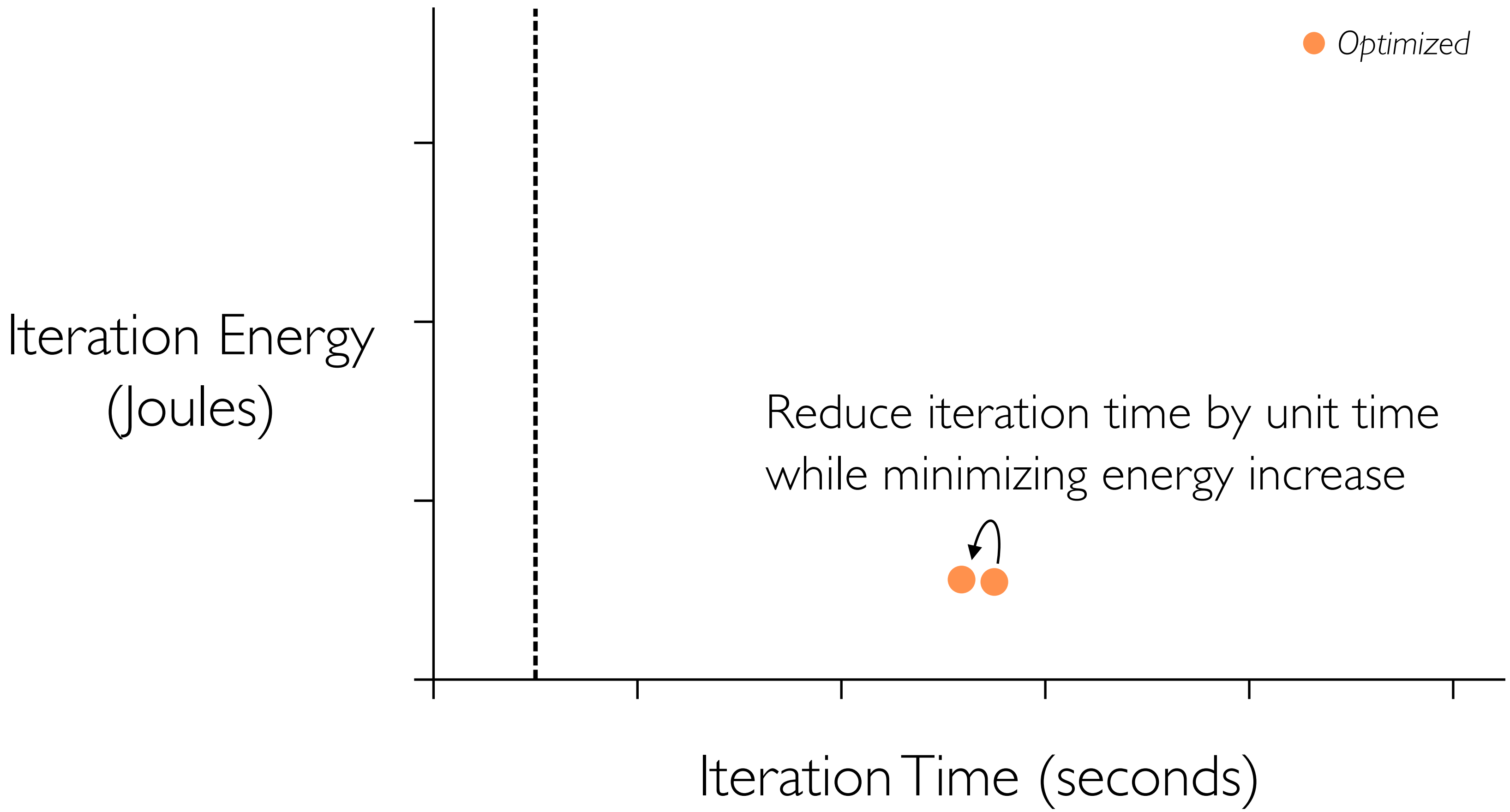
Iteration Time-Energy Frontier



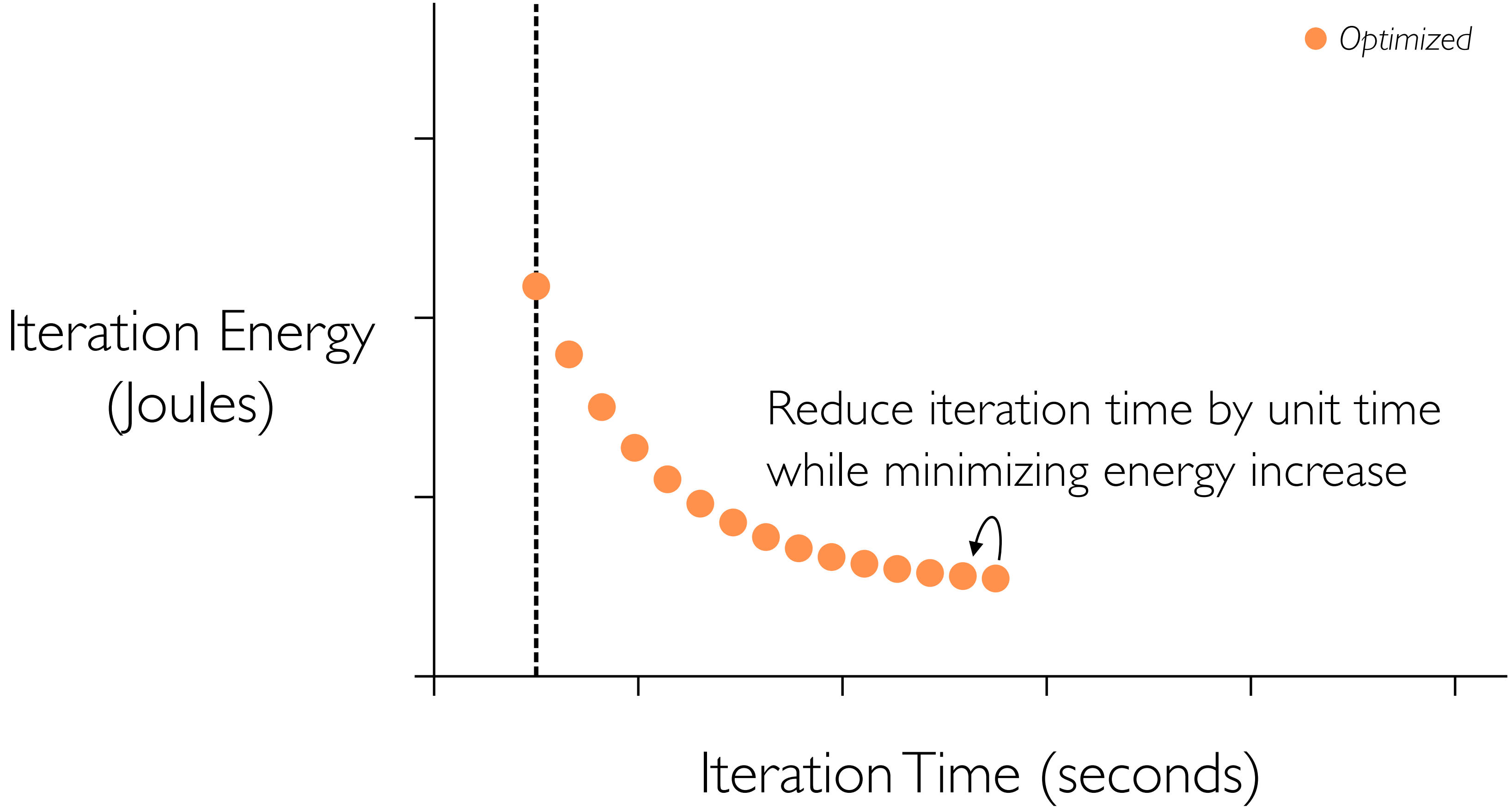
Iteration Time-Energy Frontier



An Iterative Solution

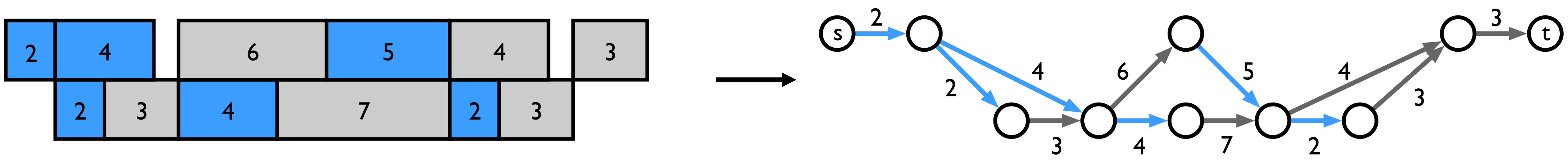


An Iterative Solution



Allocating Energy with Graph Cut

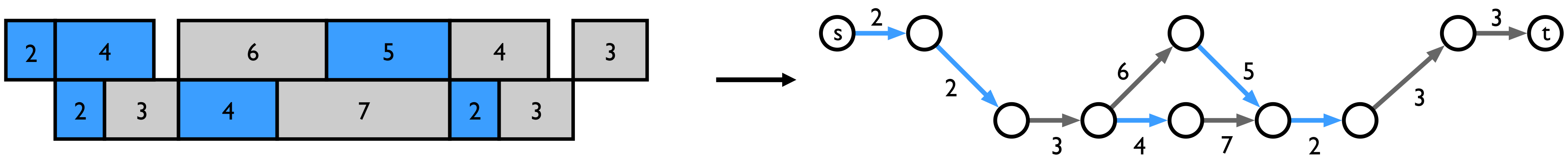
* Numbers are execution latencies



Only leave *critical* edges (computations)

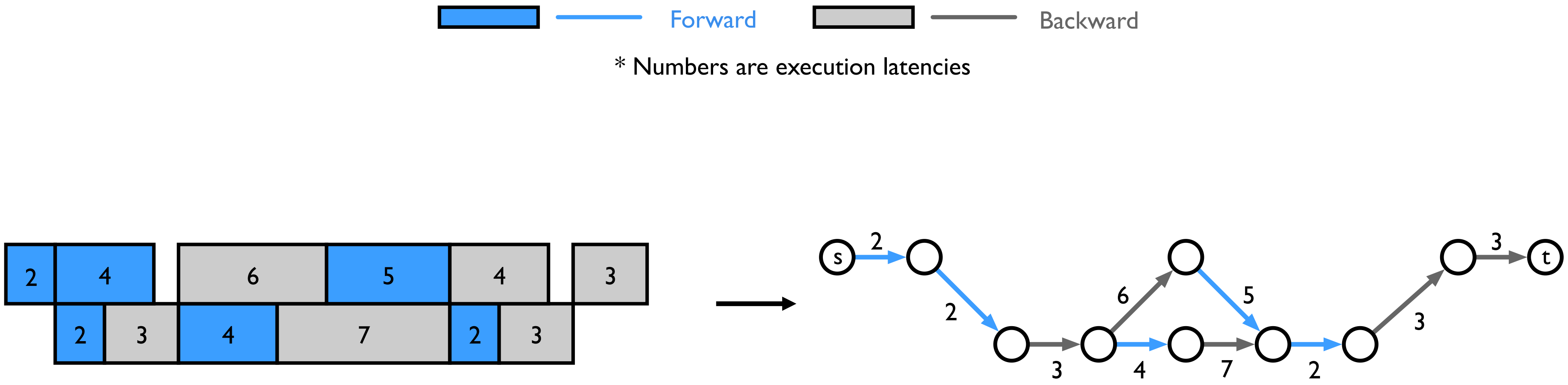
Allocating Energy with Graph Cut

* Numbers are execution latencies



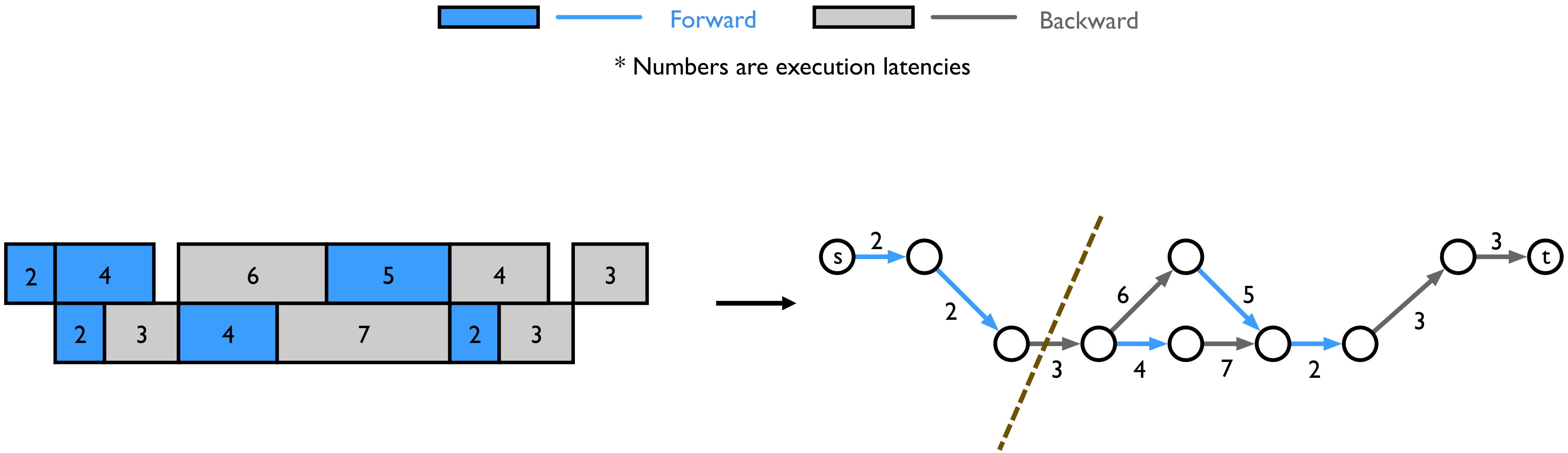
Only leave *critical* edges (computations)

Allocating Energy with Graph Cut



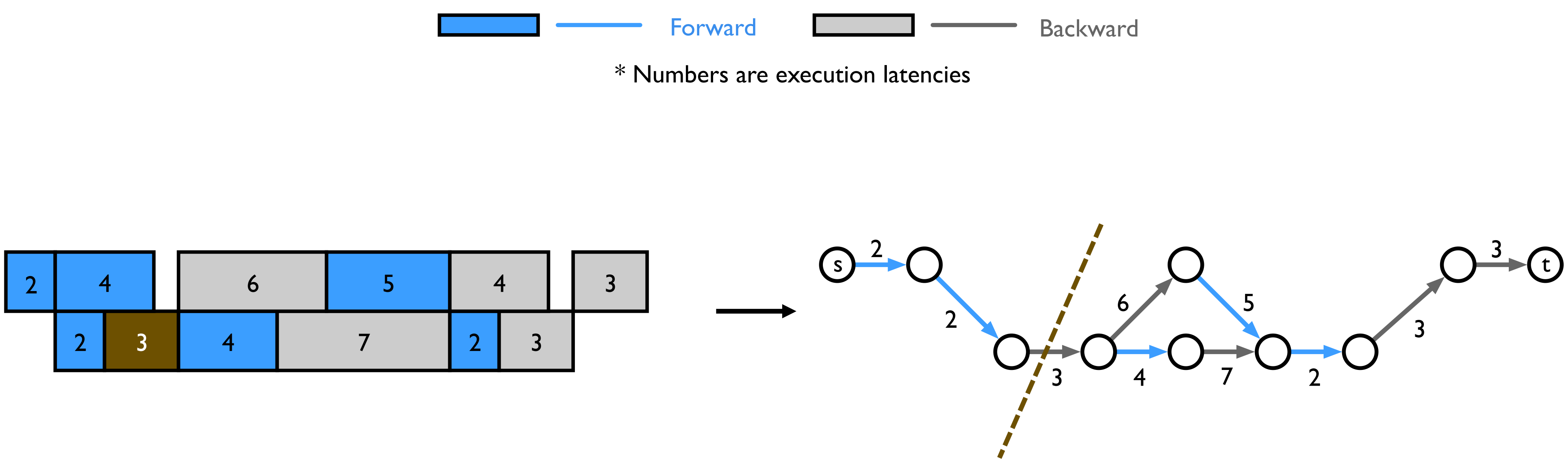
Any *s-t cut* represents a way to
reduce the DAG's end-to-end execution time by 1

Allocating Energy with Graph Cut



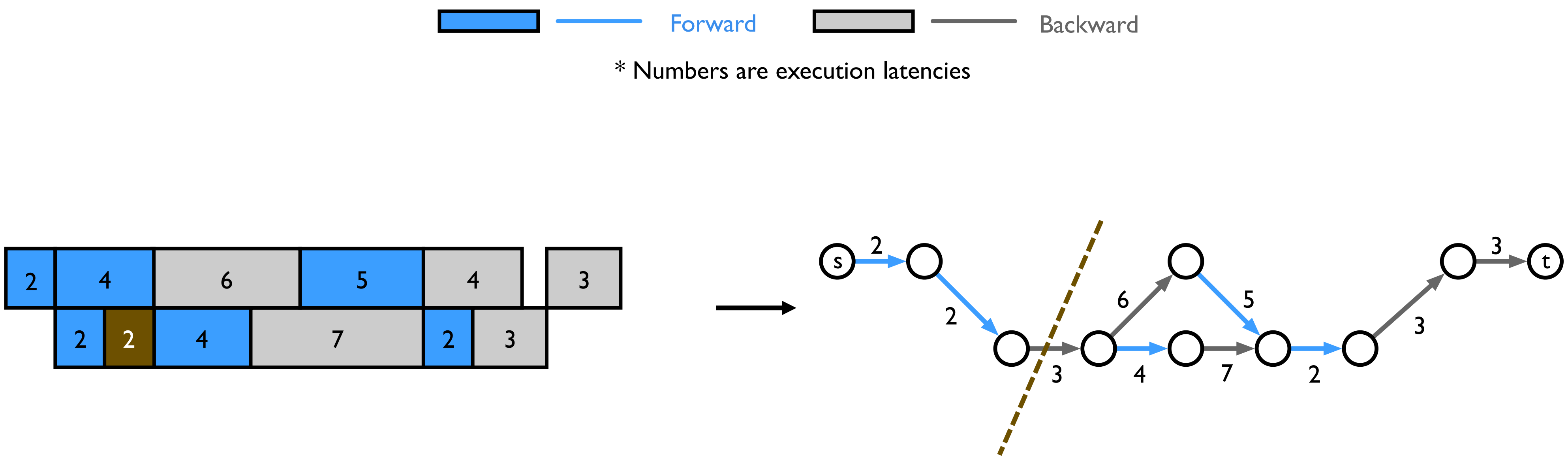
Any *s-t cut* represents a way to
reduce the DAG's end-to-end execution time by 1

Allocating Energy with Graph Cut



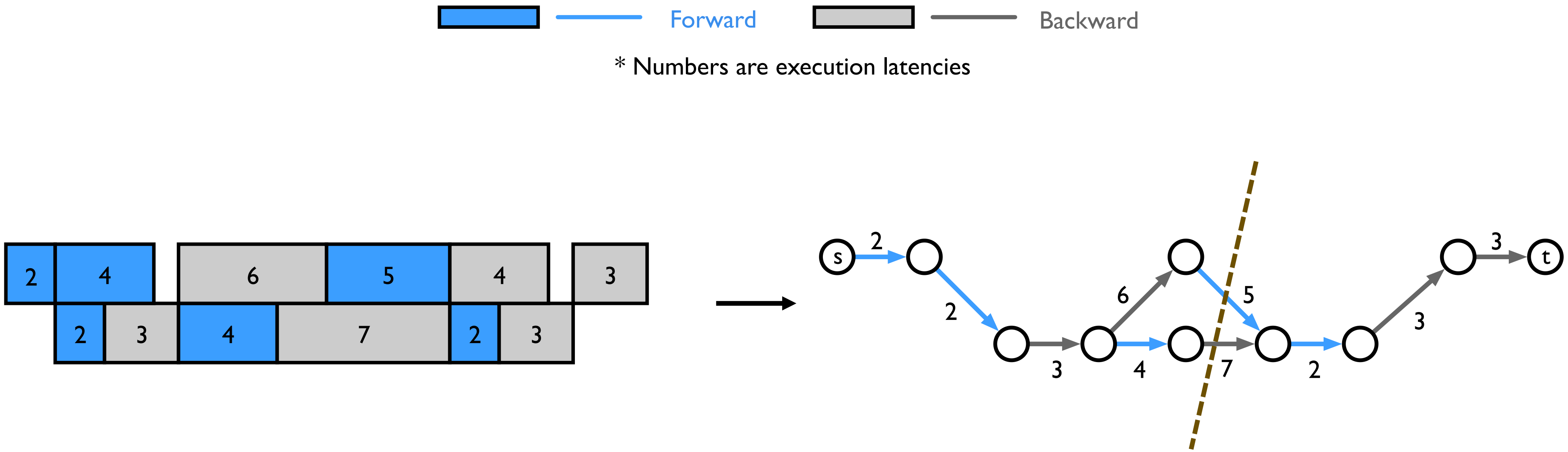
Any *s-t cut* represents a way to
reduce the DAG's end-to-end execution time by 1

Allocating Energy with Graph Cut



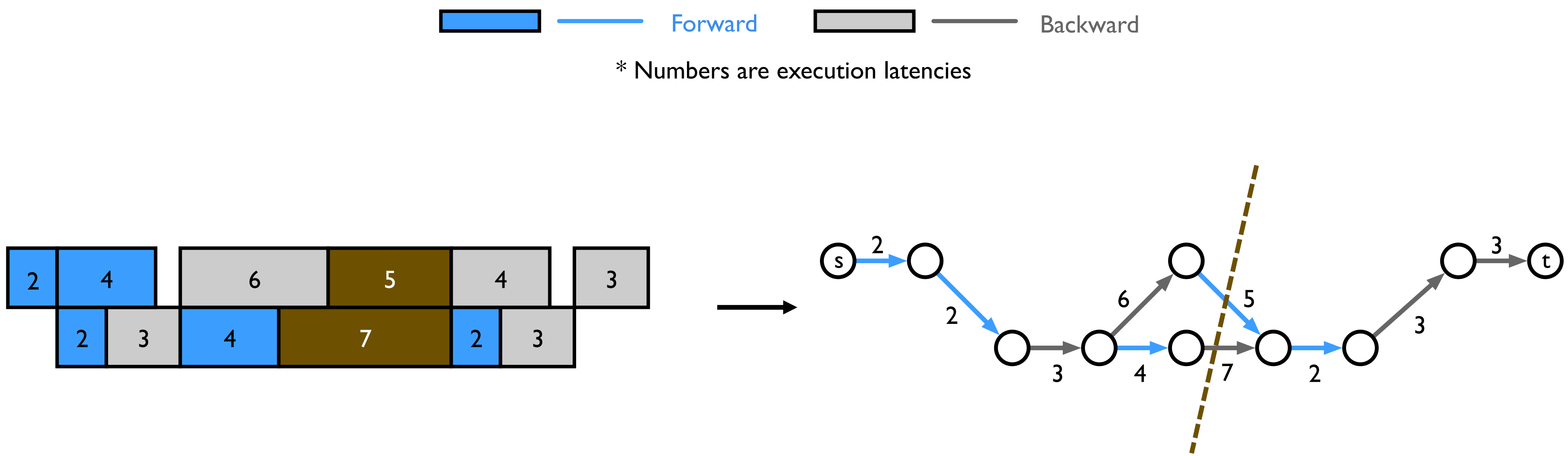
Any *s-t cut* represents a way to reduce the DAG's end-to-end execution time by 1

Allocating Energy with Graph Cut



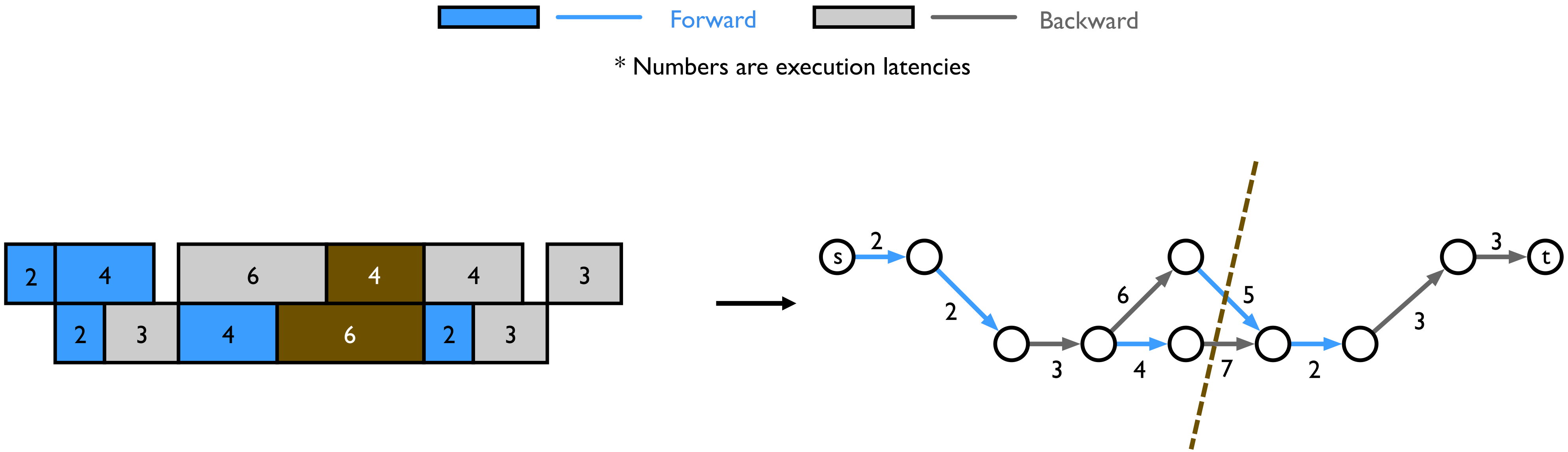
Any *s-t cut* represents a way to
reduce the DAG's end-to-end execution time by 1

Allocating Energy with Graph Cut



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

Allocating Energy with Graph Cut

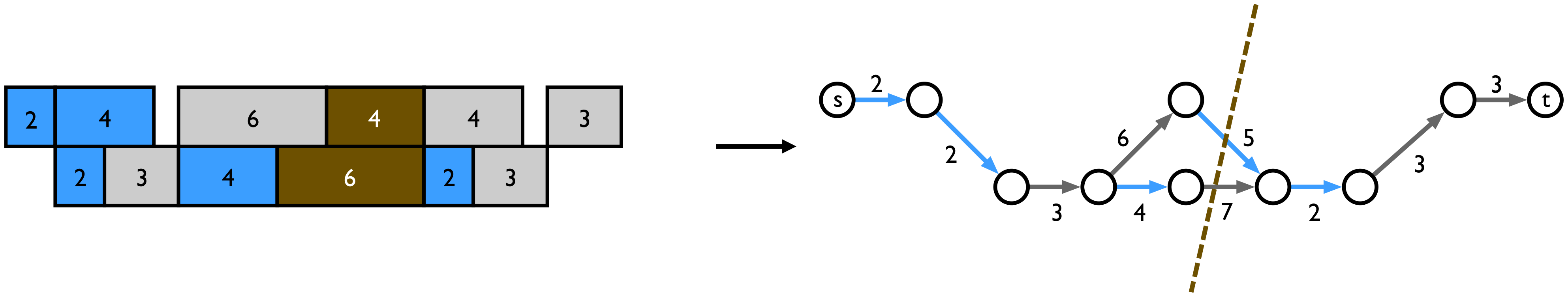


Any *s-t cut* represents a way to
reduce the DAG's end-to-end execution time by 1

Allocating Energy with Graph Cut

Forward Backward

* Numbers are execution latencies



Edge flow capacity = Extra energy needed to speed up by 1

Finding the minimum cut ⇔ Minimizing energy increase

Evaluation

Questions

- How much energy bloat reduction is possible?
- What does the time-energy frontier look like?

Setup and workloads

- Measurement on A100 and A40 GPUs and large-scale emulation
- GPT-3, BLOOM, BERT, T5, Wide-ResNet

Baselines

- Zeus (NSDI ’23)
- EnvPipe (ATC ’23)

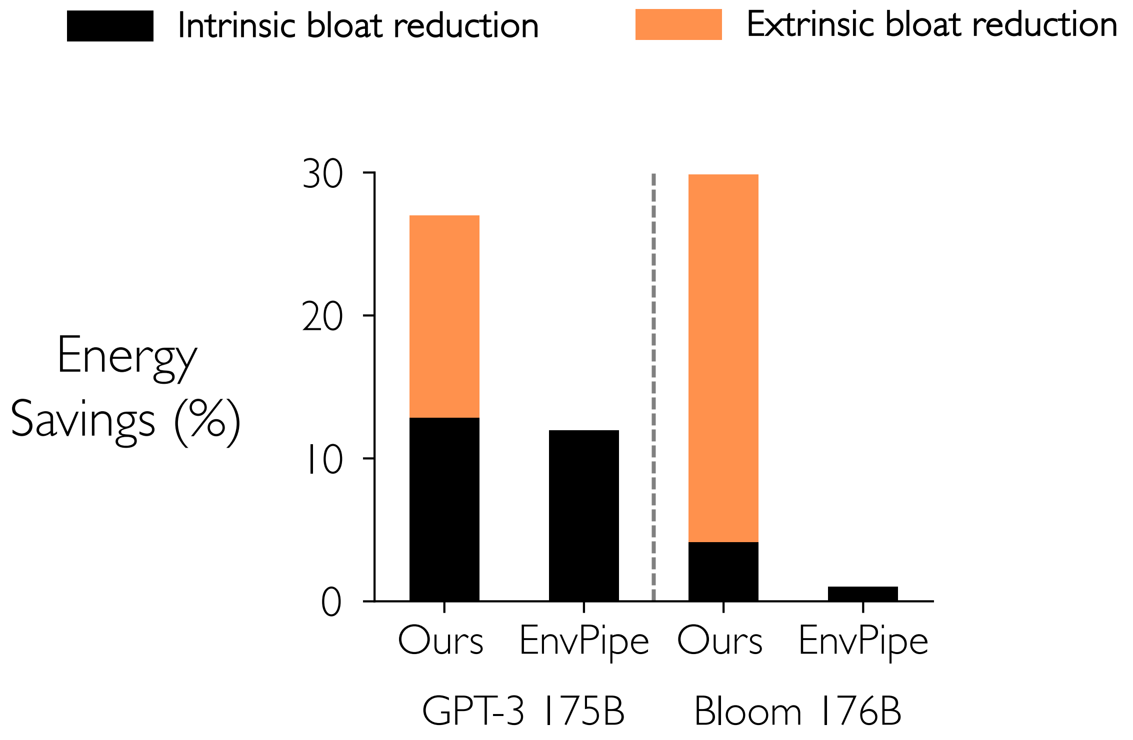
Significant Energy Bloat Reduction

Model	Energy Savings (%)	
	NVIDIA A100	NVIDIA A40
GPT-3	15.5	26.0
Bloom	15.6	26.4
BERT	16.9	24.1
T5	18.0	28.5
Wide-ResNet (scaled up)	12.7	26.3

13% to 29% energy reduction on real GPUs

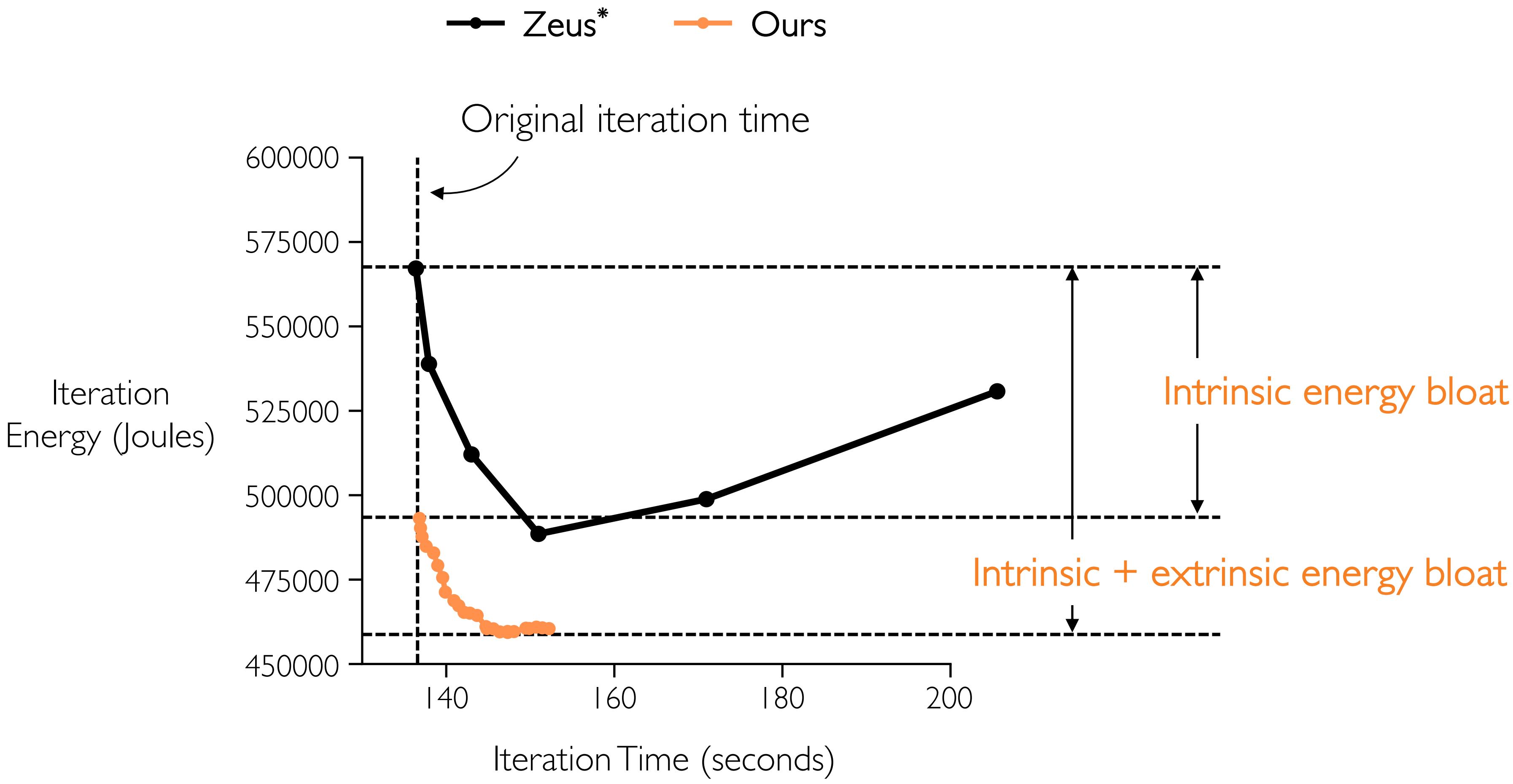
Experiment results on four A100 and eight A40 GPUs.
A100 savings are generally smaller because they are
PCIe models with lower TDP and small dynamic clock speed range.

Significant Energy Bloat Reduction



Emulation results for training each model on 1,024 A100 SXM GPUs.
Extrinsic energy bloat reduction is when the straggler pipeline is 20% slower.

Pushing the Frontier



Experiment results on NVIDIA A40 GPUs, training GPT-3 6.7B.

* *ZeusGlobal* baseline derived from Zeus, as Zeus does not support large model training.

Contributions

- Not all Joules contribute to E2E throughput
  - Some are energy bloat!
- An alternative framing for execution planning and stragglers
  - They can be cast into energy savings opportunities!
- Energy as a software-manageable ML systems resource
  - Carefully controlled and allocated, like time!

Towards an Energy-Optimal AI Stack

The ML.ENERGY Initiative
https://ml.energy

ZEUS Team

PhD Students

Jae-Won Chung
Insu Jang
Jiachen Liu
Dr. Jie You

Undergraduate & Master's Students

Yile Gu
Luoxi Meng
Zhenning Yang
Zhiyu Wu
Yong Seung Lee
Yuxuan Xia
Parth Raut
Wonbin Jin
Daniel Hou
Sharon Han
Oh Jun Kweon

ML.ENERGY Core PIs

Tom Anderson (UW)
Adam Belay (MIT)
Beidi Chen (CMU)
Mosharaf Chowdhury (UMich)
Asaf Cidon (Columbia)
Simon Peter (UW)