Power and Energy as First-Class AI Design Metrics

Jae-Won Chung

September 29th, 2025

Jae-Won Chung

Fifth-year PhD student at the University of Michigan, CSE

I build efficient software systems for AI/ML

Efficient management of time and energy as systems resources

Unprecedented Scale

Large

- Frontier models
- Kimi-K2 1T params
- Closed models could be larger

Numerous

- Specialized models
- Thousands of application use cases

Everywhere

- Mixture of models
- Deployed worldwide
- Serving millions

Demand for Compute

More accelerators

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

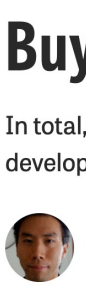
(David Paul Morris/Bloomberg via Getty Images)

Demand for Compute

More accelerators

Zuckerberg's Meta Is Spending Billions to Buy 350,000 refrigerators

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)

Demand for Compute

More accelerators

Gigawatt datacenters

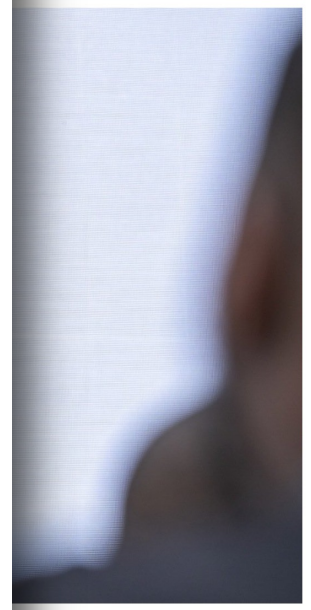
Zuckerberg says Meta will build data center the size of Manhattan in latest AI push

CEO says company plans to spend hundreds of billions on developing artificial intelligence products

Meta Founder and CEO Mark Zuckerberg speaks at LlamaCon 2025, an AI developer conference, in Menlo Park, California, on 29 April 2025. Photograph: Jeff Chiu/AP

Billions to

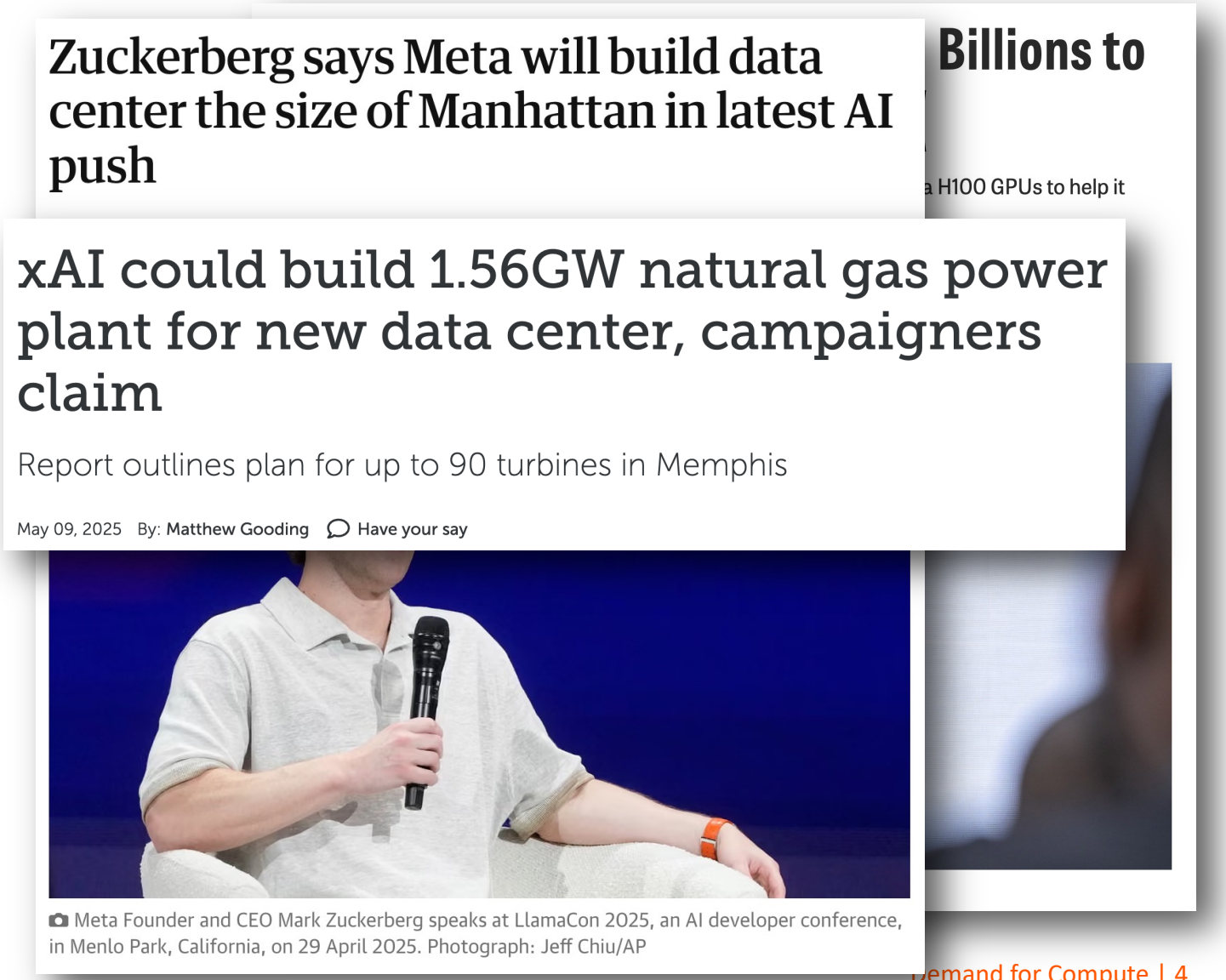
a H100 GPUs to help it



Demand for Compute

More accelerators

Gigawatt datacenters



Zuckerberg says Meta will build data center the size of Manhattan in latest AI push

Billions to

a H100 GPUs to help it

xAI could build 1.56GW natural gas power plant for new data center, campaigners claim

Report outlines plan for up to 90 turbines in Memphis

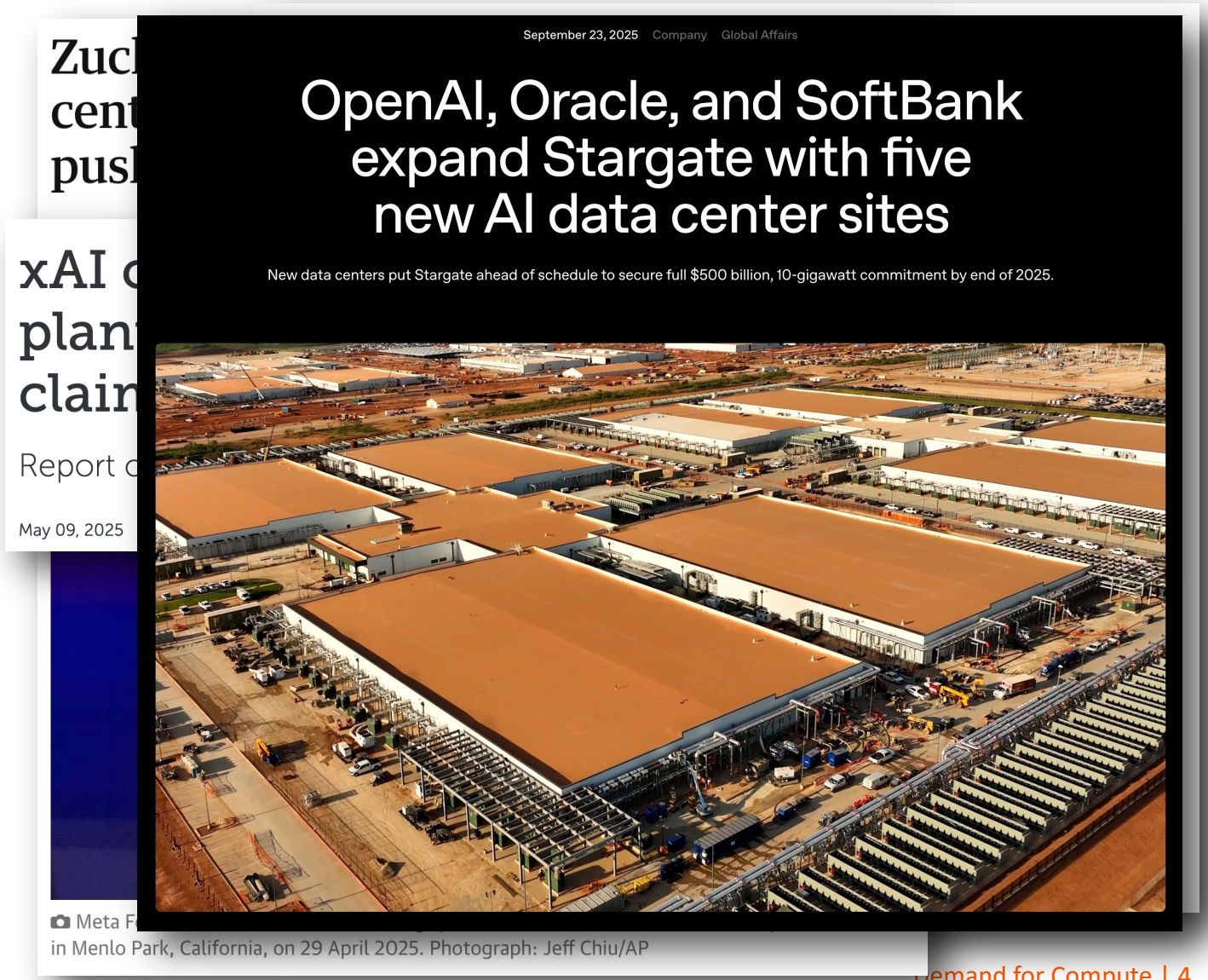
May 09, 2025 By: Matthew Gooding Have your say

Meta Founder and CEO Mark Zuckerberg speaks at LlamaCon 2025, an AI developer conference, in Menlo Park, California, on 29 April 2025. Photograph: Jeff Chiu/AP

Demand for Compute

More accelerators

Gigawatt datacenters



Challenge: Getting Power

We need the power now

- Next generation large frontier models were due yesterday
- It takes a whole datacenter to train one of them
  - E.g., Meta's Llama 3.1 405B used 16K H100 GPUs

Challenge: Getting Power

Getting power takes time

- Years of planning, approval, lead time, and construction

Challenge: Getting Power

Getting power takes time

- Years of planning, approval, lead time, and construction

Challenge: Getting Power

Getting power takes time

- Years of planning, approval, lead time, and construction

Challenge: Energy Costs and Constraints

Depends on the company & cluster setup

- Electricity costs become operational expenses
- Power source constraints (e.g., 24/7 carbon-free energy)
- Carbon offsetting costs (e.g., Net Zero)

Challenge: Managing Power

Packing within power budget

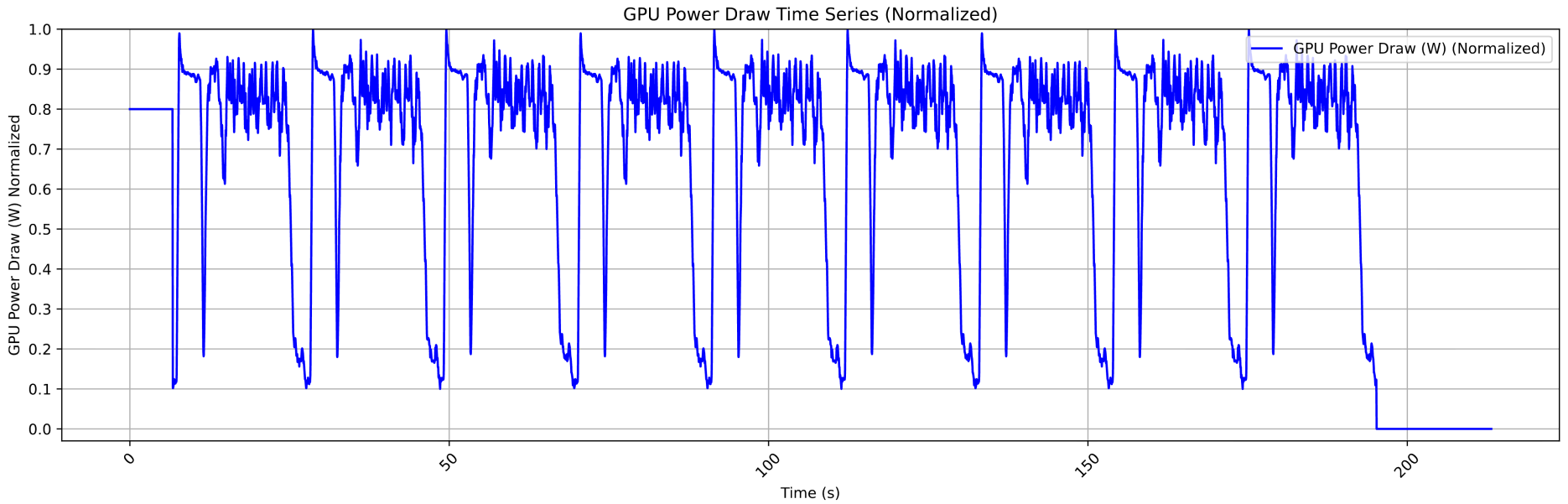
- Power is a hard ceiling; you can't draw beyond capacity
- You want to pack more accelerators within budget

Challenge: Managing Power

Packing within power budget *without blowing up the grid*

- Giant power fluctuations stress the grid
- May stress the resonance frequency of power plant turbines

Tens of megawatts



Power Stabilization for AI Training Datacenters, Technical report, 2025

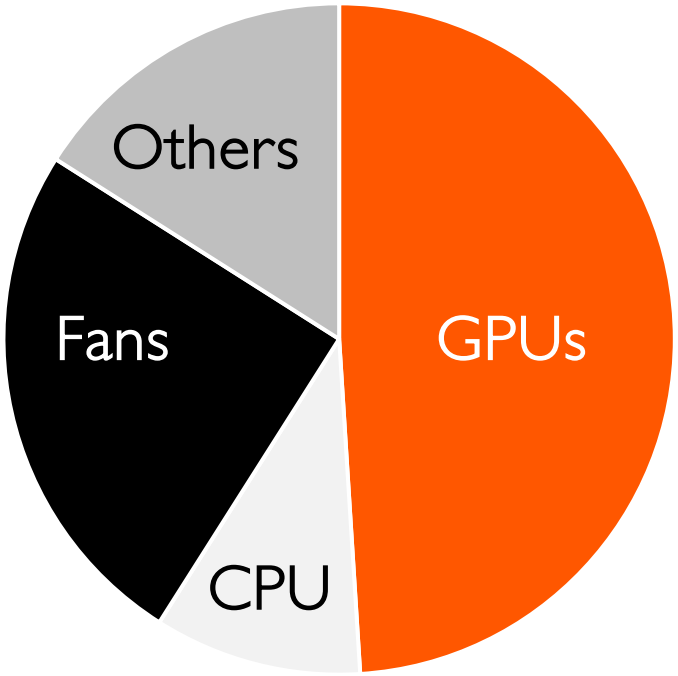
Energy as a Computing Resource

Every bit of computation consumes energy

- And energy is an increasingly scarce resource
- We don't understand it as well as time

Measuring, understanding, and optimizing the energy consumption of AI/ML workloads

"Where do the Joules go?"

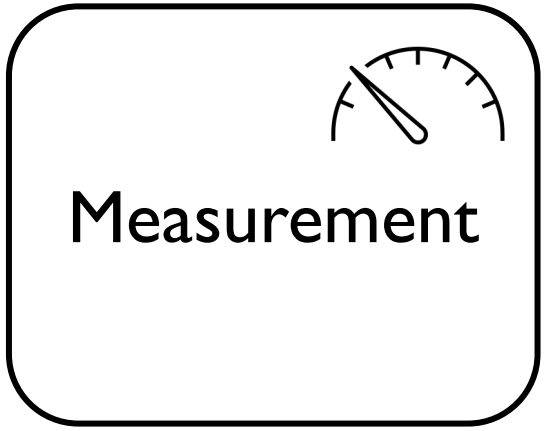


Provisioned Power
(8xA100-80GB AI Server)

Patel et al., ASPLOS'24

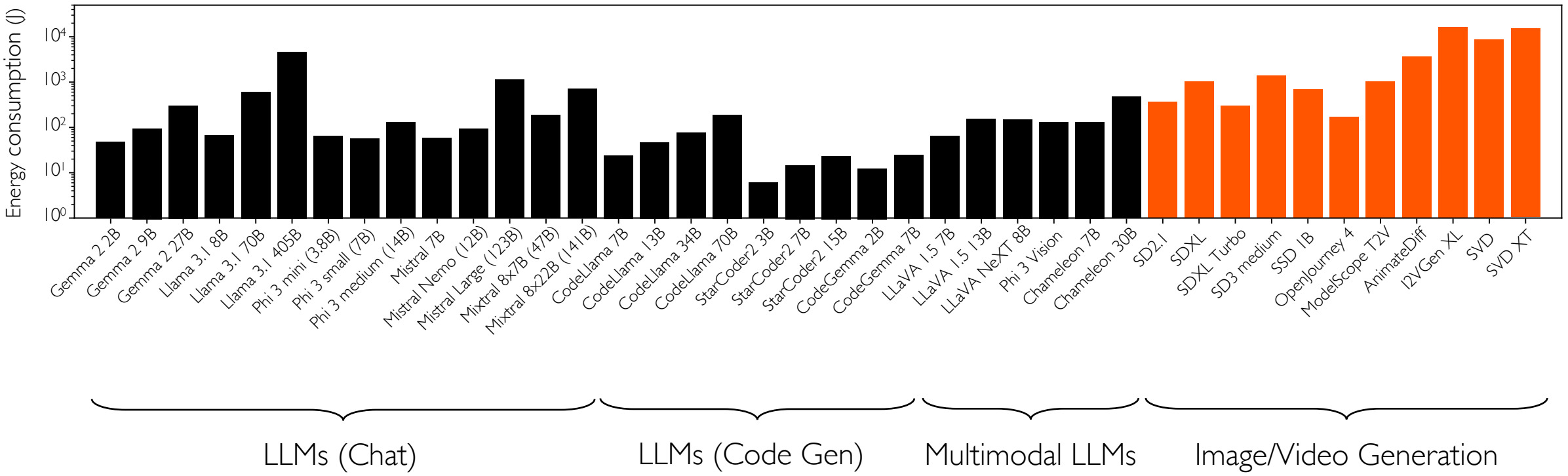
https://ml.energy/zeus

A PyTorch Ecosystem project

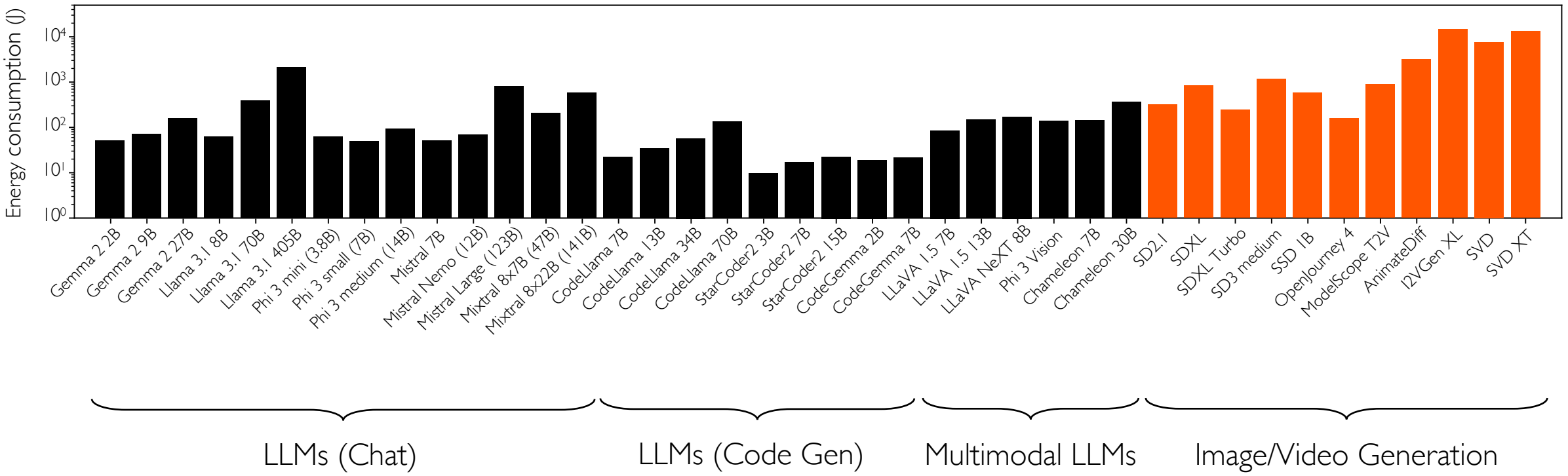



```python
from zeus.monitor import ZeusMonitor

monitor = ZeusMonitor()

monitor.begin_window("train")
# Much
# AI
res = monitor.end_window("train")

print(f"The computation consumed {res.total_energy} J.")
```

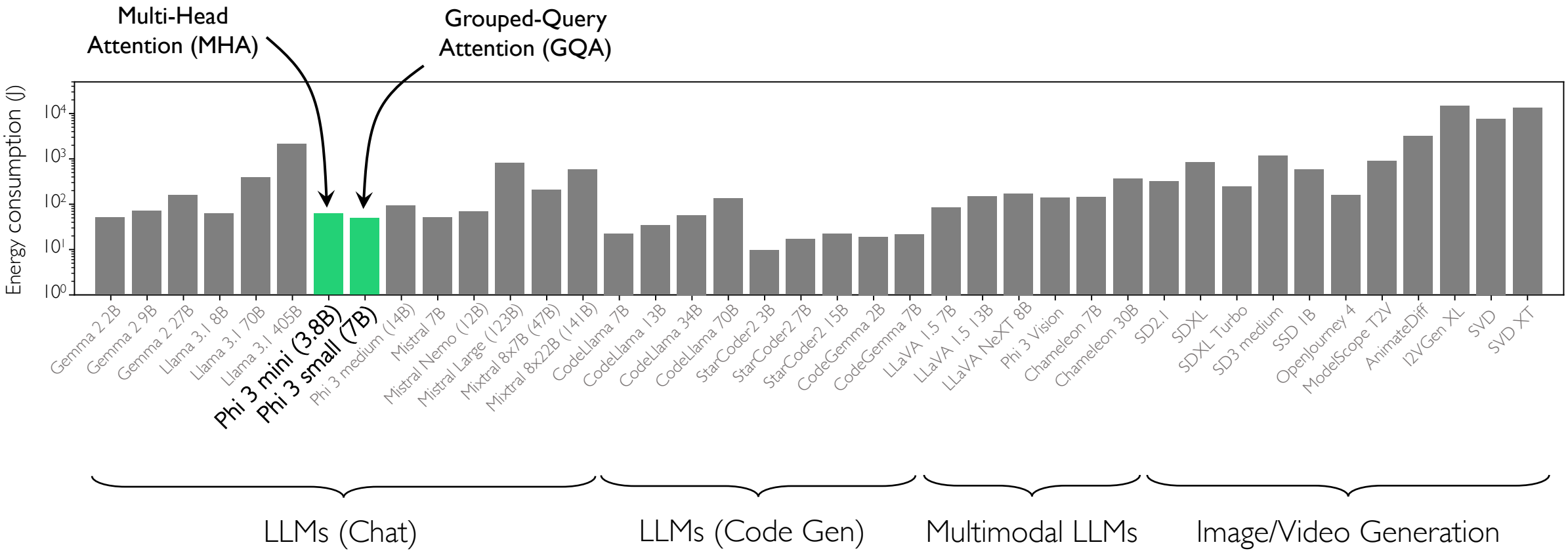
Inference Energy Consumption (A100)



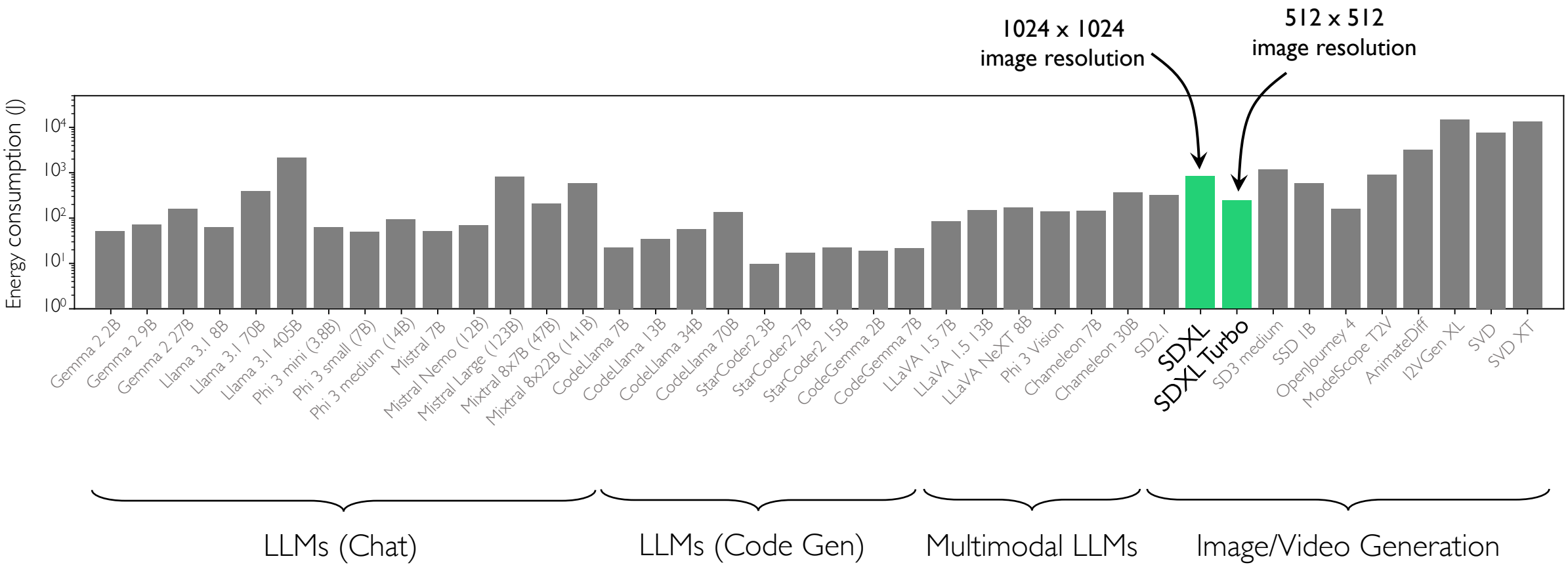
Hardware Choices Matter (H100)



ML Design Choices Matter



Deployment Choices Matter



The ML.ENERGY Benchmark & Leaderboard

https://ml.energy/leaderboard

- Open-source at https://github.com/ml-energy/leaderboard
- Upgraded V3.0 coming soon!

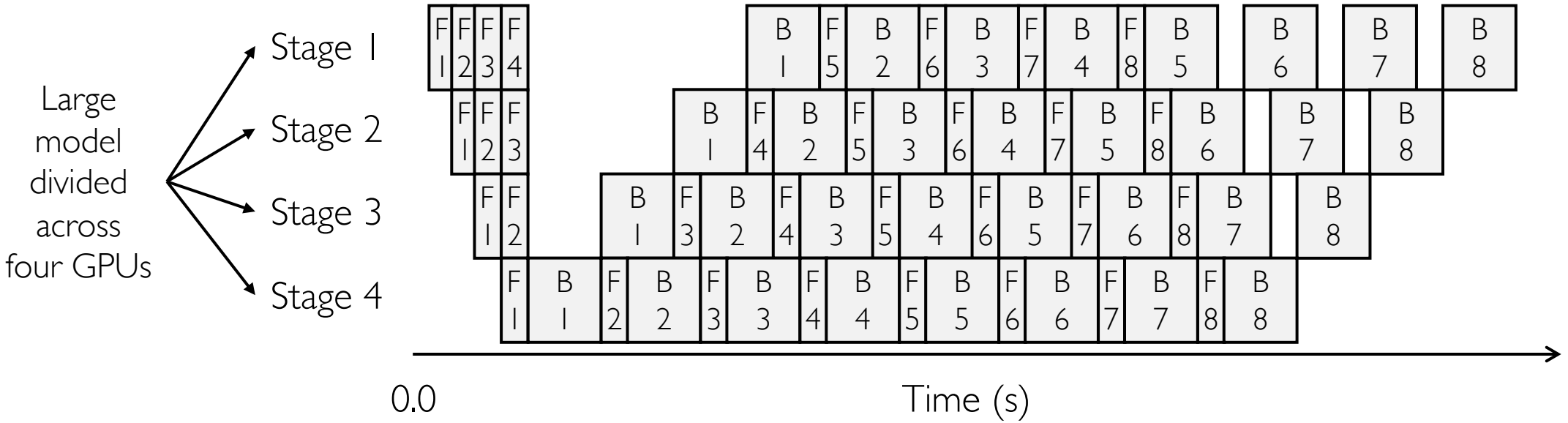
"Any way to reduce energy consumption?"

Estimated energy consumption of training large AI models

GPT-3[a]	2020	1,287 MWh
Amazon Q[b]	2022	11,900 MWh
Llama 3.1 405B[c]	2024	21,588 MWh
Llama 4 Scout[d]	2025	3,500 MWh

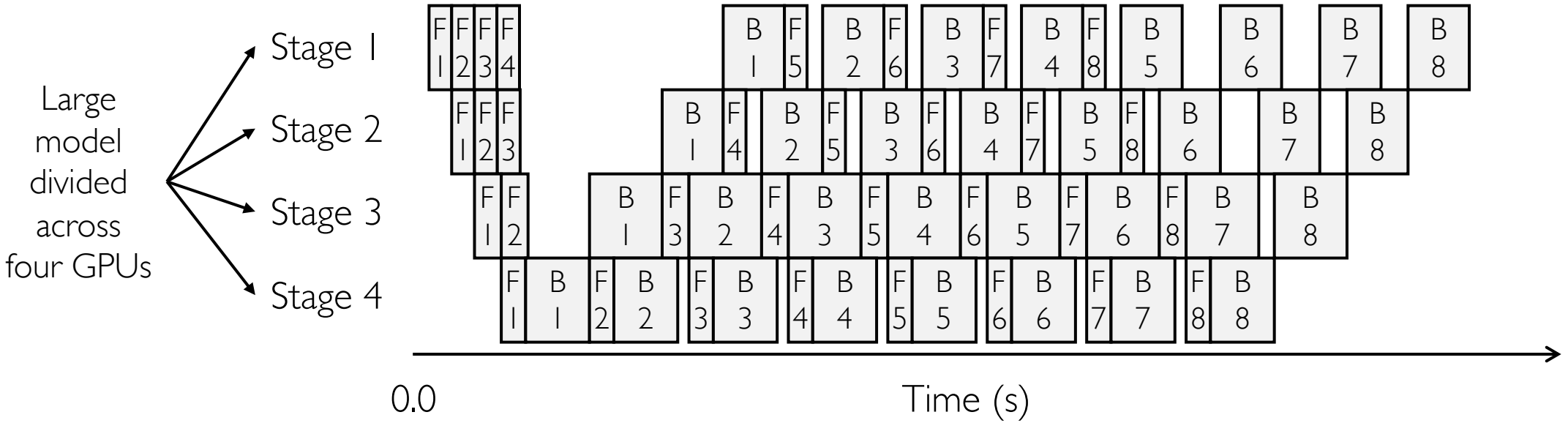
[a] Patterson et al., arXiv:2104.10350, 2021
[b] Hamilton, CIDR Keynote, 2024
[c] AI@Meta, Llama 3.1 405B Model Card, 2024
[d] AI@Meta, Llama 4 Scout Model Card, 2025

Where Do the Joules Go?



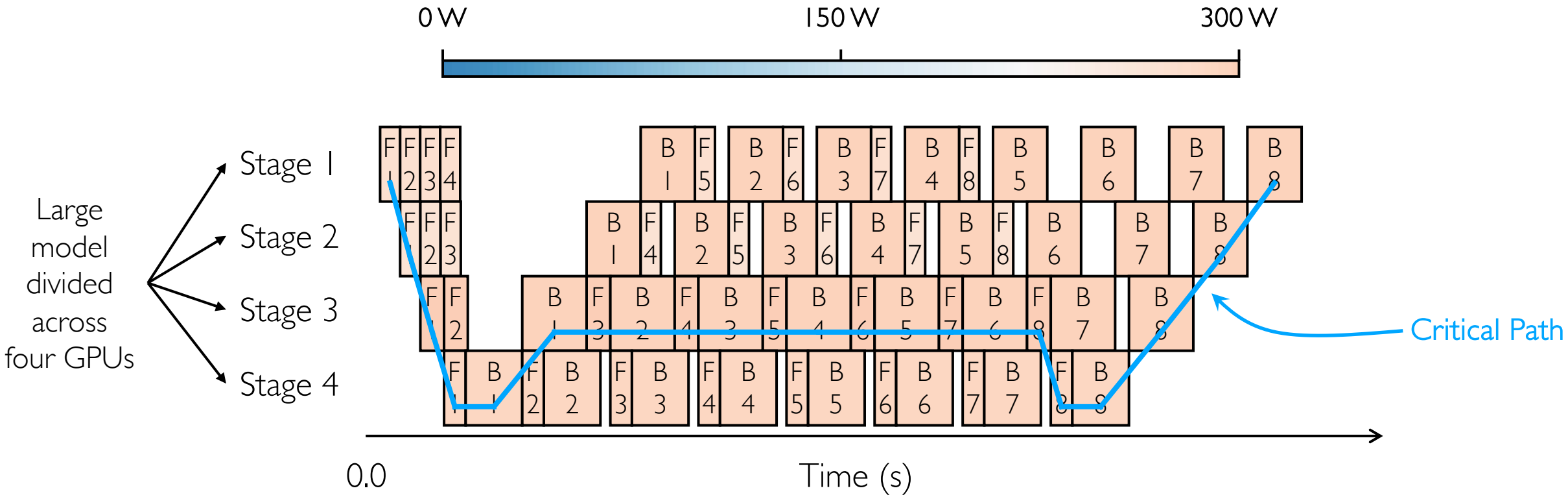
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).

Where Do the Joules Go?



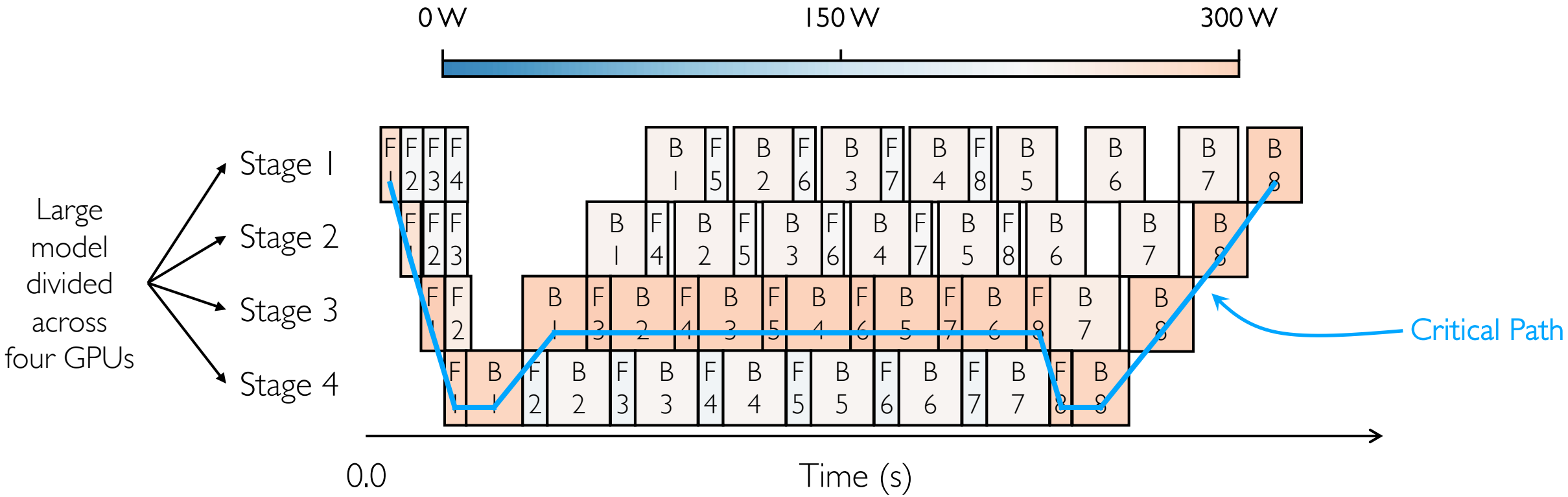
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).
Drawn to scale for GPT-3 1.3B on NVIDIA A100 GPUs.

Where Do the Joules Go?



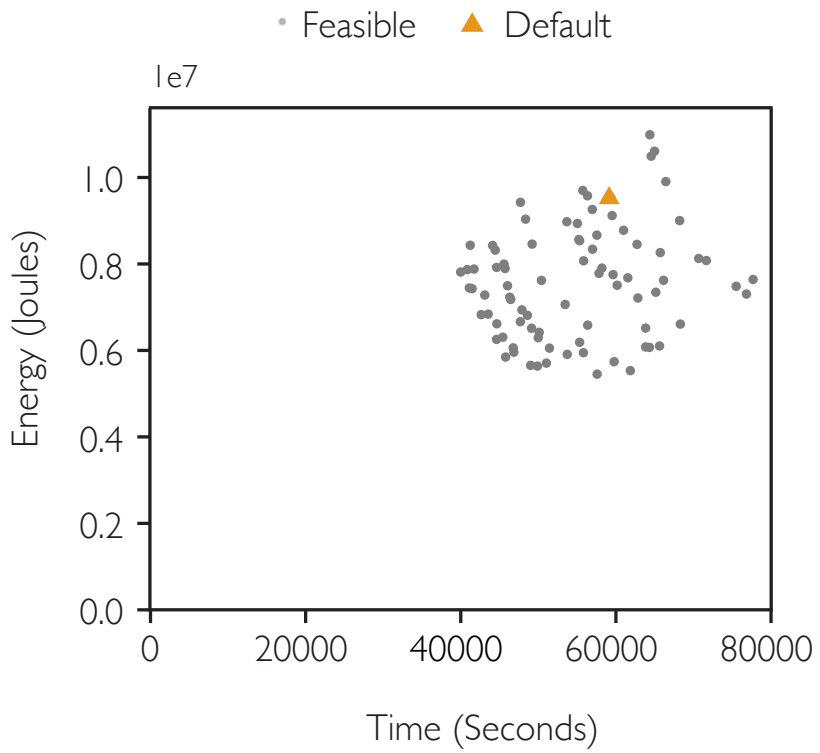
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).
Drawn to scale for GPT-3 1.3B on NVIDIA A100 GPUs.

Cutting 30% Energy Bloat



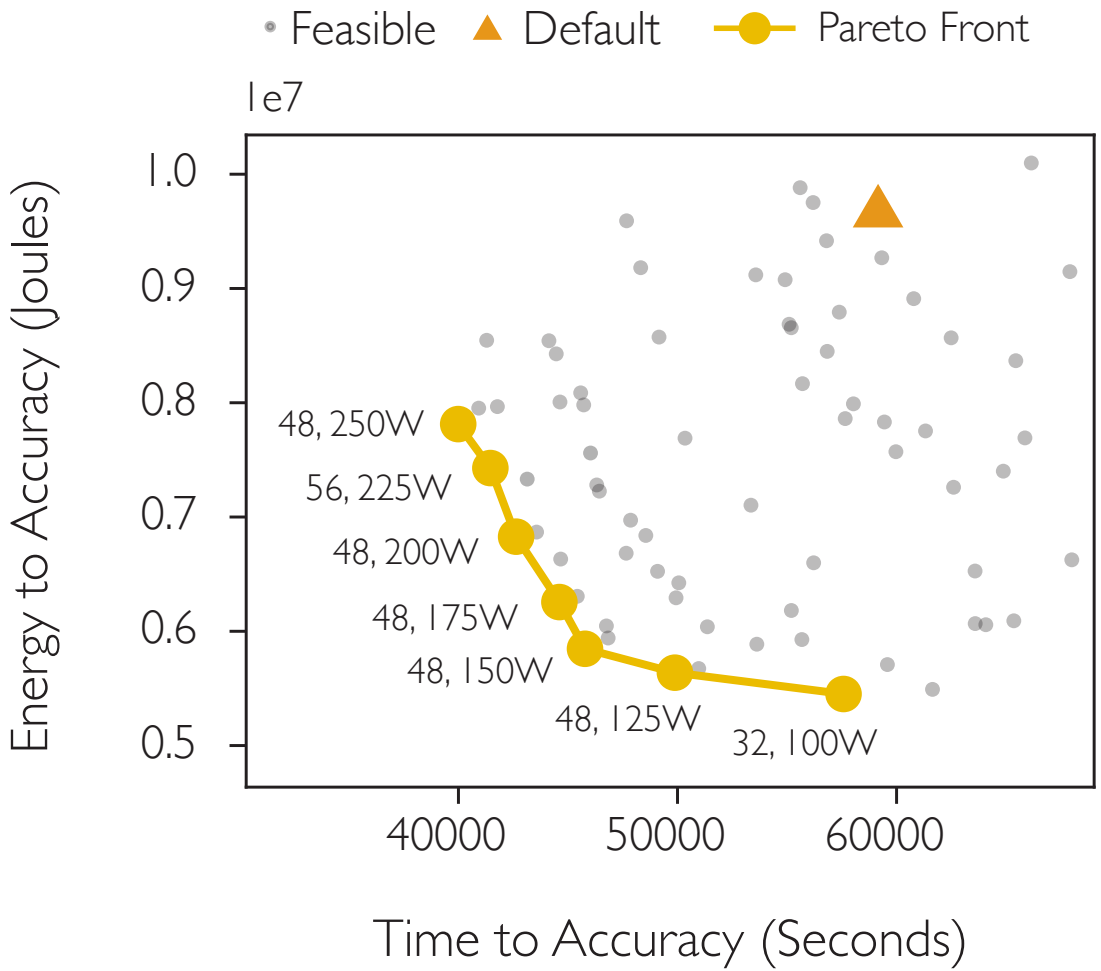
One training iteration with 4 pipeline stages and 8 microbatches (1F1B schedule).
Drawn to scale for GPT-3 1.3B on NVIDIA A100 GPUs.

"How does energy interact with time?"



Energy consumed vs. time taken to train DeepSpeech2

Time vs. Energy Trade-off

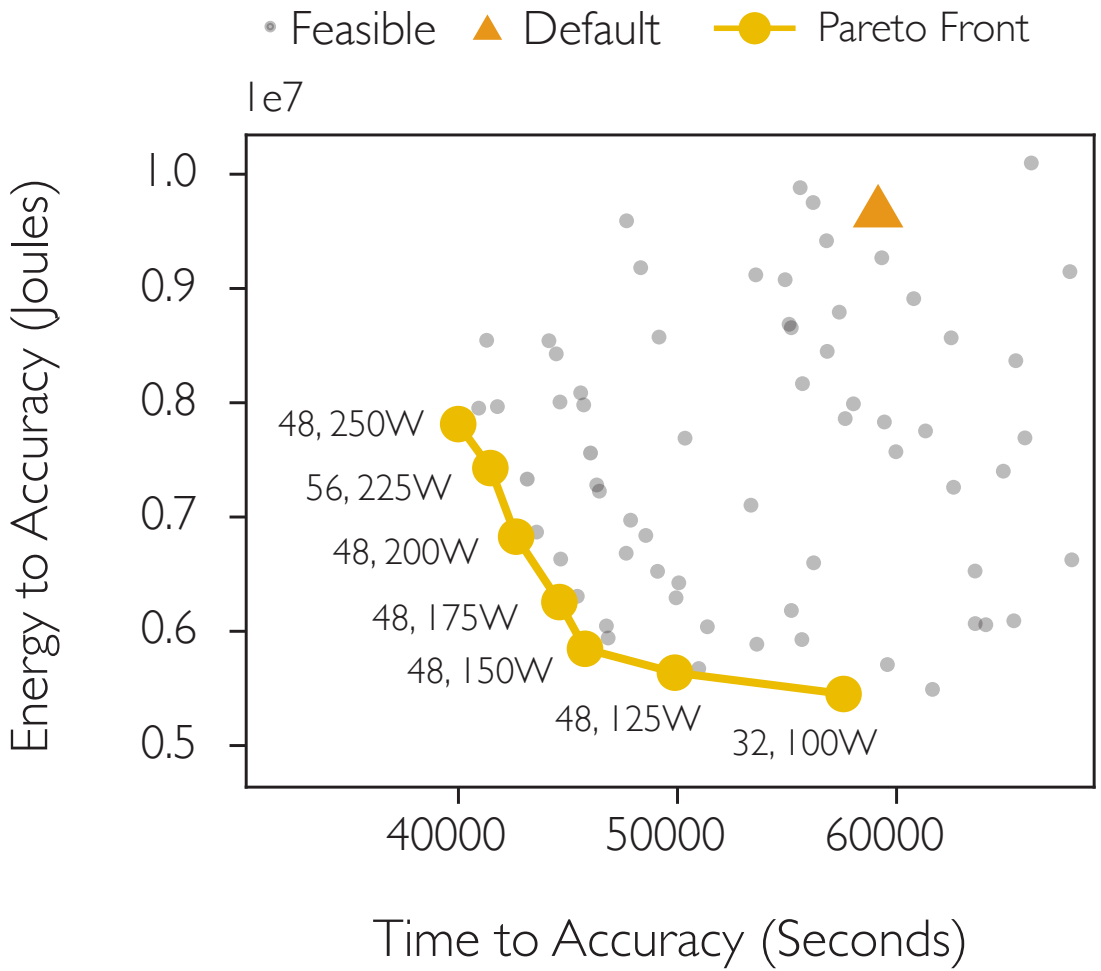


Training time and total energy affected by batch size and GPU power limit

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

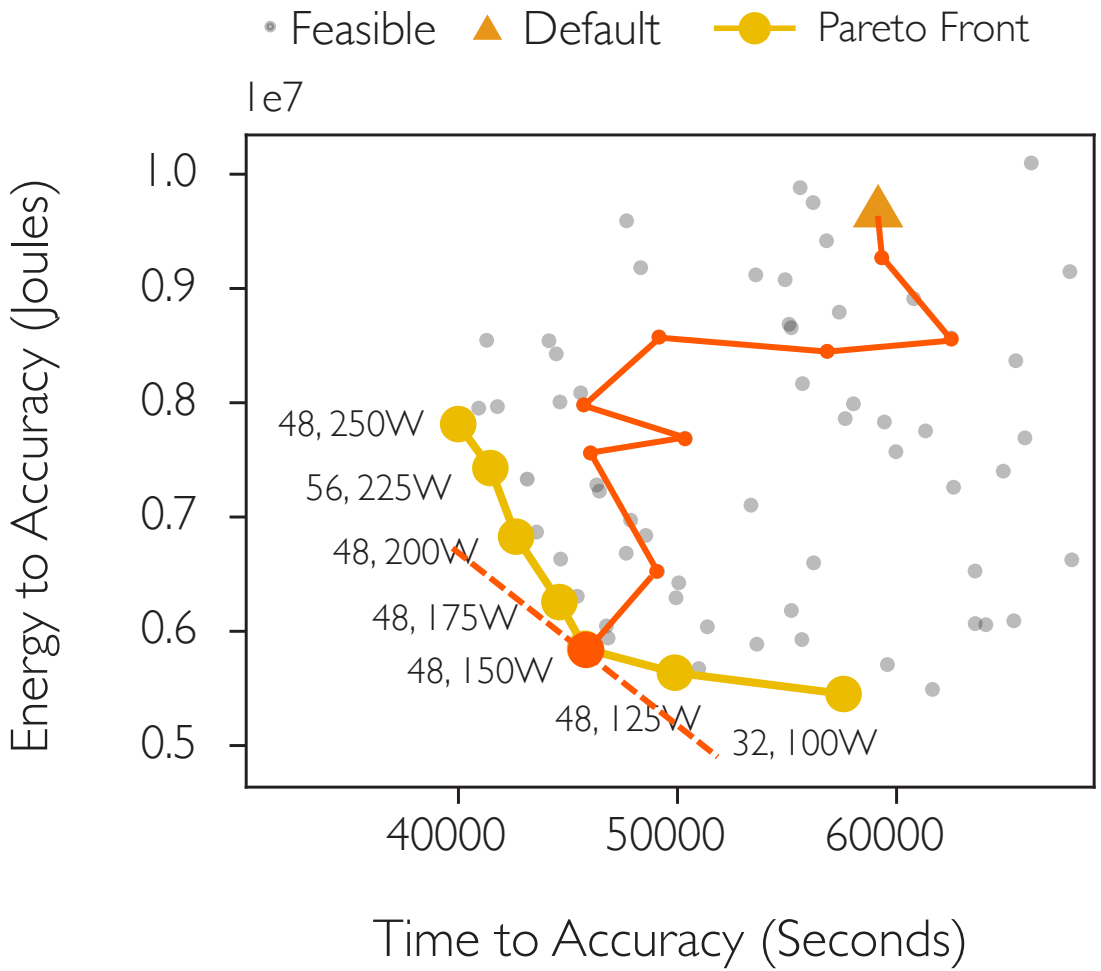


Which yellow point is the best?

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

Time vs. Energy Trade-off

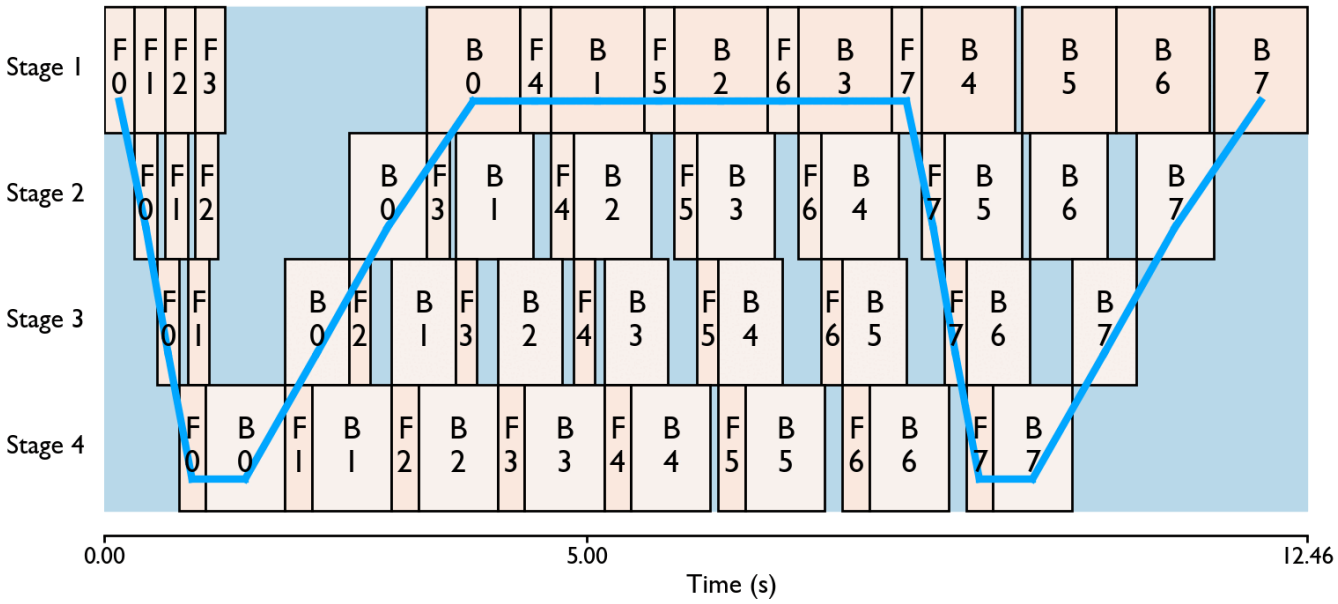


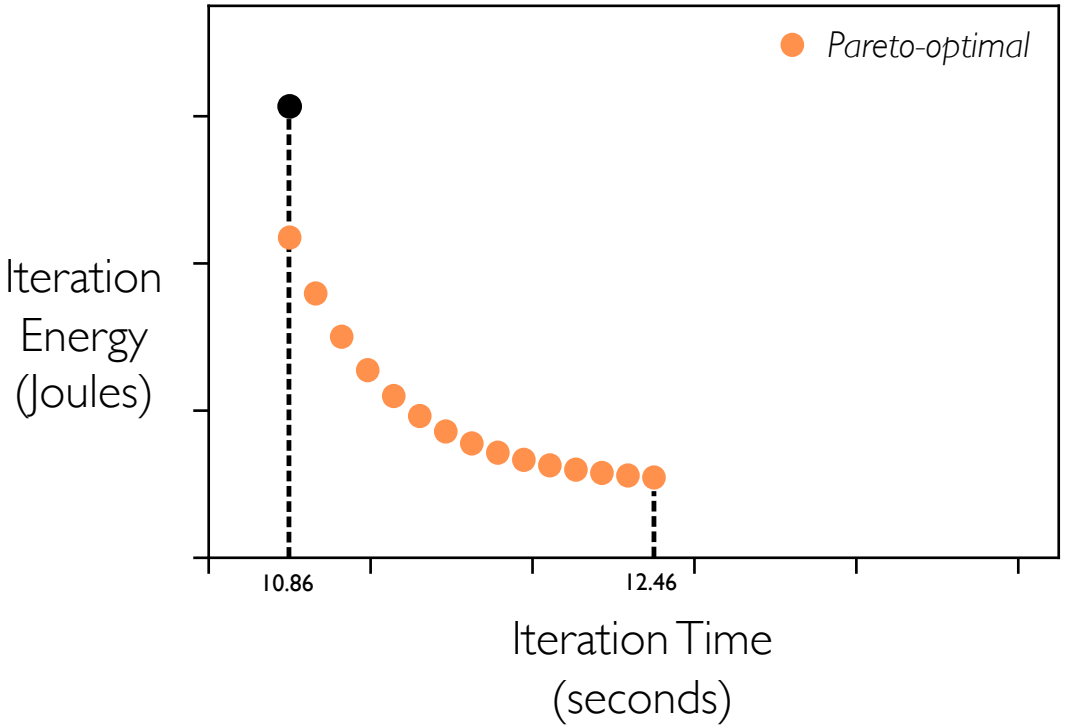
15% to 76% energy reduction
across diverse models
and multiple GPU generations

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

Time vs. Energy Trade-off

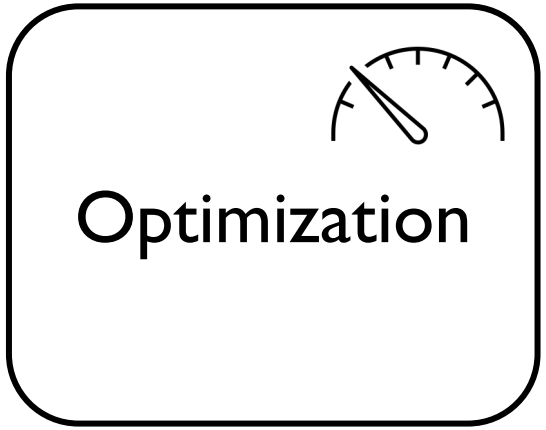




Reducing Energy Bloat in Large Model Training, SOSP, 2024

https://ml.energy/zeus

A PyTorch Ecosystem project




```python
from zeus.monitor import ZeusMonitor
from zeus.optimizer.power_limit import HFPowerLimitOptimizer
from transformers import Trainer

monitor = ZeusMonitor()
optimizer = HFPowerLimitOptimizer(monitor)

trainer = Trainer(
  ...,
  callbacks=[optimizer],
)
```

Towards an Energy-Optimal AI Stack

The ML.ENERGY Initiative

https://ml.energy

Team

PhD Students

Jae-Won Chung
Insu Jang
Ruofan Wu
Jeff J. Ma
Dr. 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

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