Improving the Energy Efficiency of AI and HPC

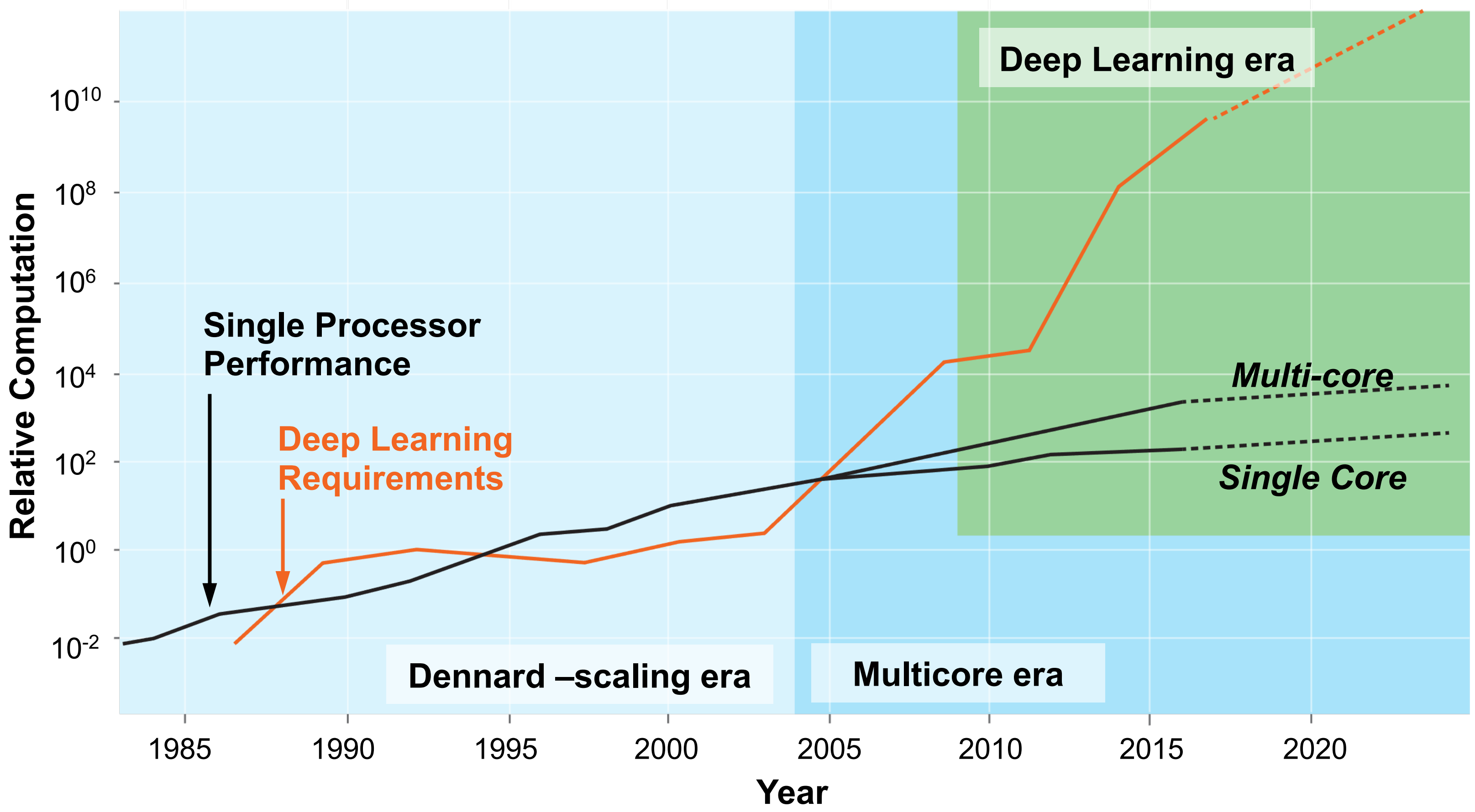
Vijay Gadepally

MIT Lincoln Laboratory Supercomputing Center
MIT Connection Sciences

MIT LINCOLN LABORATORY
SUPERCOMPUTING CENTER

Slide Contributions from Siddharth Samsi (MIT LL), Neil Thompson (MIT CSAIL),
Joseph McDonald, Matthew Weiss (MIT LL)

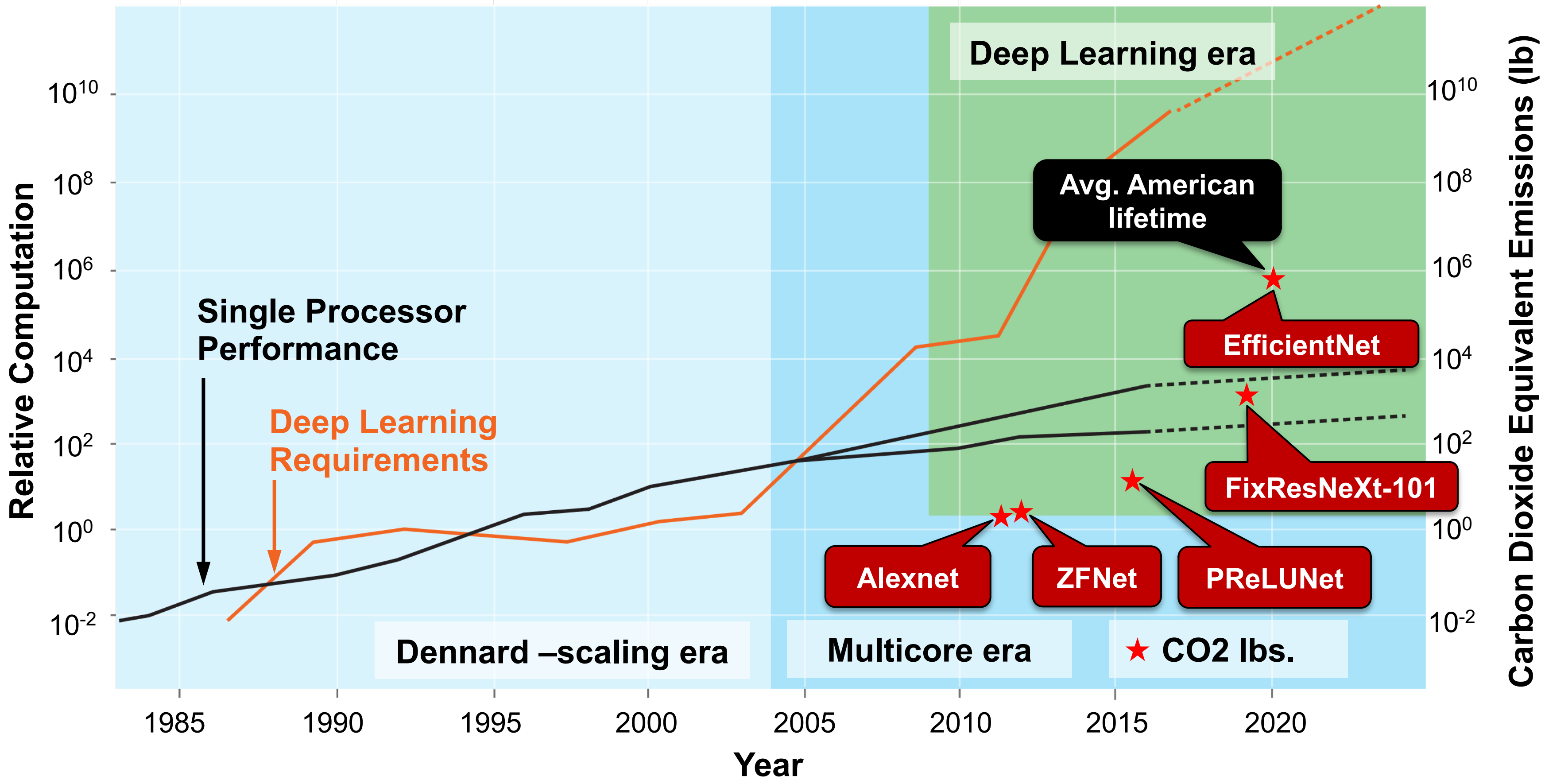
Growth of AI Computing Requirements



Deep learning compute requirements are growing faster than hardware performance

[1] Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2021. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. IEEE Spectrum.

AI Computing Carbon Emissions

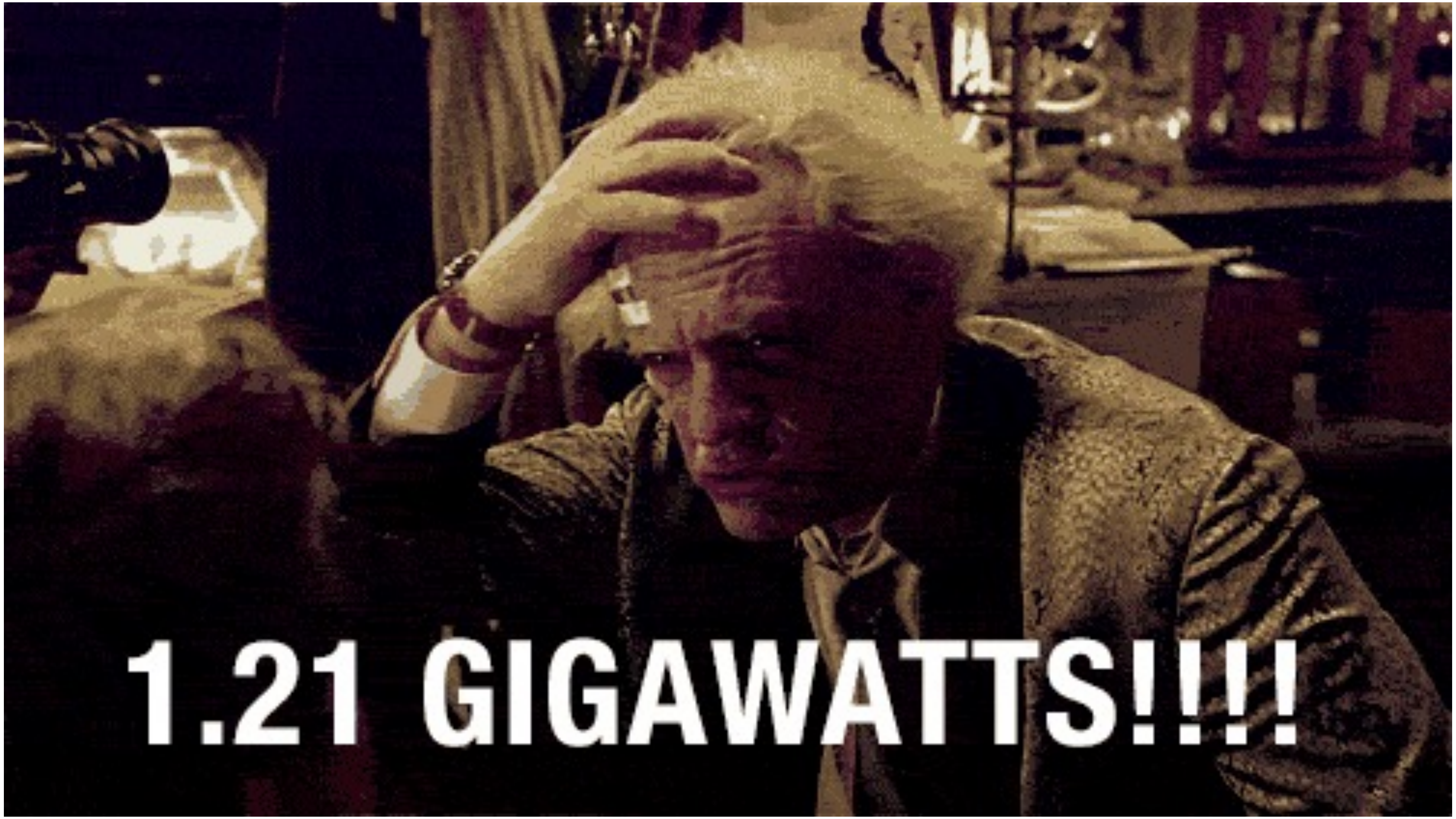


Deep learning energy requirements are growing unsustainably

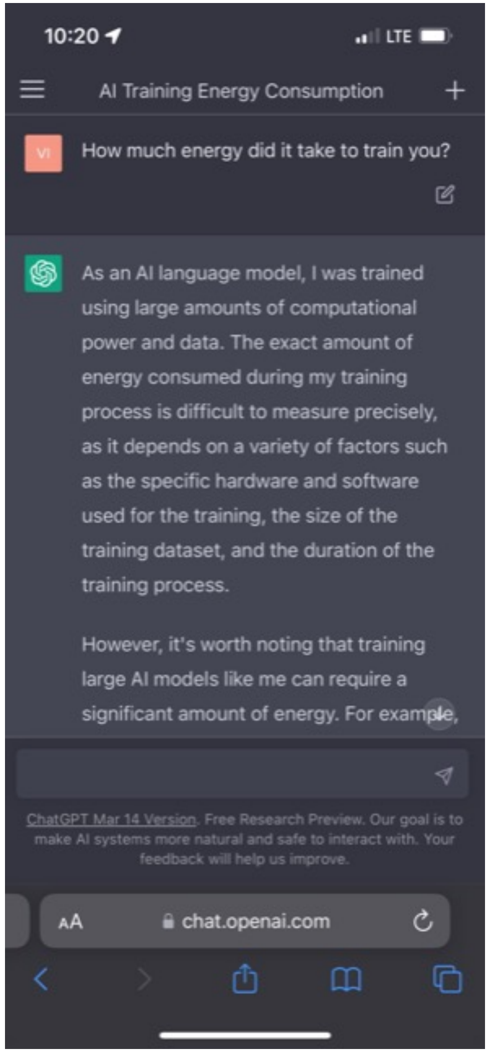
[1] Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2021. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. IEEE Spectrum.

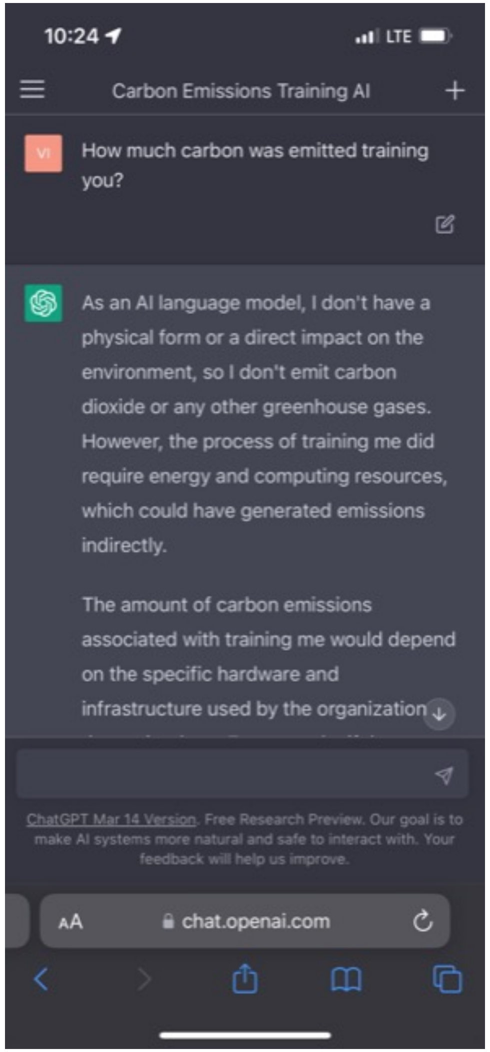
[2] The Energy and Carbon Footprint of Training End-to-End Speech Recognizers - Parcollet, T., & Ravanelli, M., 2021

…to put it in perspective



How about ChatGPT?





Other Facets

- Current datacenter energy consumption ~ 1-2% global energy demand
  – Estimated to increase to 8-21% by 2030

- Significant water usage
  – 20% of water from stressed watersheds
  – 50% of servers supplied by power plants in water stressed areas

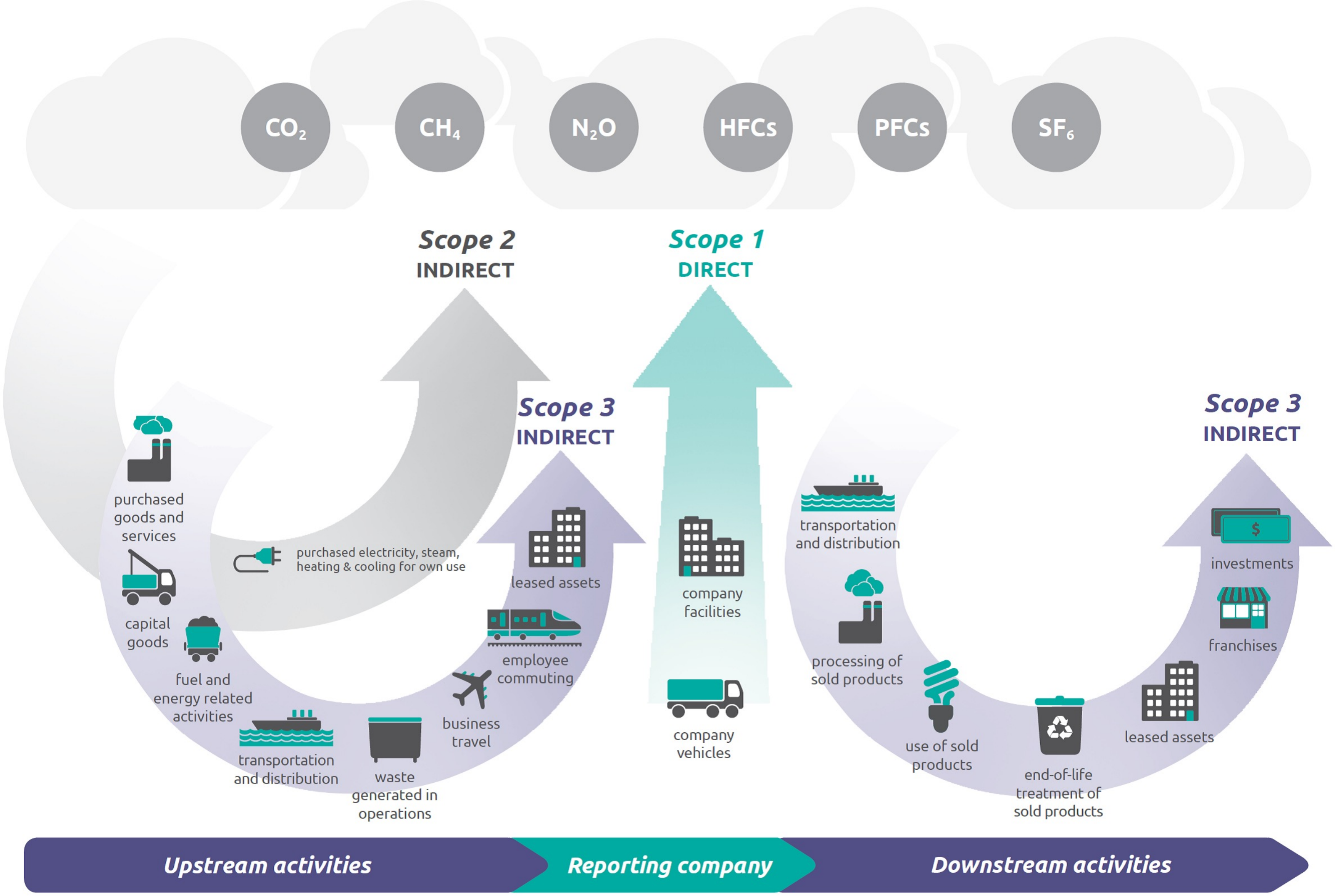
- Environmental footprint of AI goes beyond just datacenter usage
  – E.g., carbon costs of hardware manufacturing (embodied carbon)

Opportunity to reduce ~1-2% of global electricity demand

Zhao, et. al. "A Greener World for A.I.", IPDPS ADOPT 2022

Frey, et. al. "Benchmarking Resource Usage for Efficient Distributed Deep Learning", SuperComputing 2022 (Submitted)

Greenhouse Gas Scopes and Emissions



For a datacenter:

- **Scope 1 Examples: Backup generators**
- **Scope 2 Examples: Emissions from energy used for cooling, computing, building management,…**
- **Scope 3 Examples: Hardware manufacturing emissions**

Image Source: https://www.epa.gov/climateleadership/scope-1-and-scope-2-inventory-guidance

Sustainability Challenges in AI

Current incentives for A.I. research, applications:

- Prioritizing best-performing models (accuracy)
- Faster run-times, more experimentation, faster results
- Publications in high-visibility journals and conferences

What gets missed:

- Prioritizing energy-efficient models
- More experiments run, more computation, more energy consumed
- Awareness of environmental footprint of AI research, applications

Research Theme: How can we make AI research and practice more sustainable?

Understanding a Datacenter's Carbon Footprint

How can you calculate a datacenter's carbon footprint?

*From the perspective of a datacenter operator

Operational Footprint

- Due to operating a datacenter
- Includes, energy to IT gear as well as facilities operations (e.g., cooling)
- Power Usage Effectiveness (PUE), datacenter efficiency metric

$$PUE = \frac{FE + IT}{IT}$$

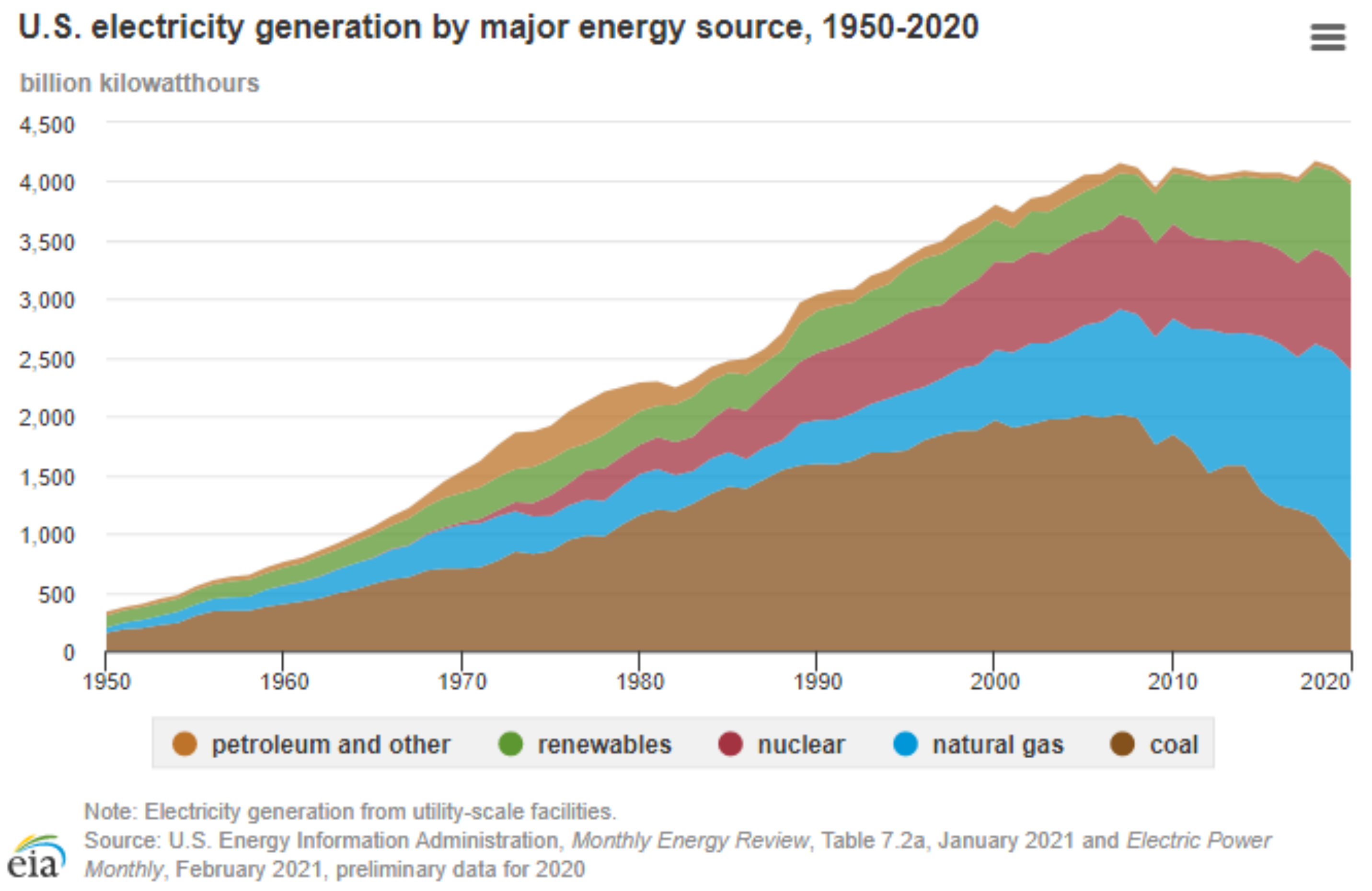
IT – Information technology energy
FE – Facility energy

- Global average is 1.58 (2018), efficient datacenters are close to 1

Common strategy: leverage renewable energy sources for your datacenter

Moving to renewables can be a Zero-sum game

(at any given time)



Renewables are a worthy investment; Also need ways to be more energy efficient

Image source: https://www.epa.gov/green-power-markets/us-electricity-grid-markets

Embodied Footprint

- **With renewable, operational carbon footprint may dramatically reduce**
  - **Increasing proportion of carbon coming from manufacturing**
- **Embodied carbon includes manufacturing**
  - **Energy**
  - **Chemicals involved (e.g., for etching)**
- **Some estimates: 80+% of datacenter footprint due to embodied carbon**

(when leveraging renewables to reduce operational footprint)

- **Difficult to estimate Embodied Carbon Footprint => Opportunities to improve!**

Processor Data Sheet Vendor: XYZ
System On Chip
SoC Area
Energy / Area
Carbon Intensity
…
Packaging
Chemical Footprint
Environmental Report
…
Memory Modules
DRAM
HDD
…

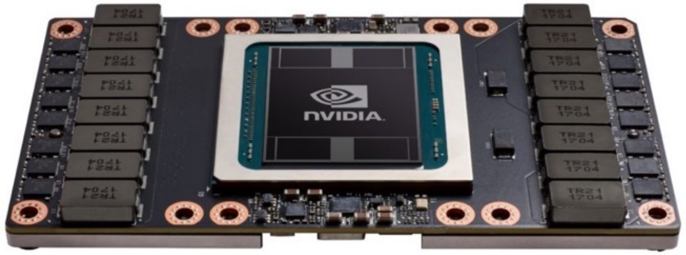
Proposed Data Sheet

Reducing the Operational Footprint of a Real Datacenter

Our Testbed: MIT SuperCloud

Low Carbon Emission

- **Significant increase in computing power for simulation, data analysis, and machine learning**
- **Leverages power of 900 Nvidia Volta GPUs**



- **Operates on renewable energy**

	Capability
Processor	Intel Xeon & Nvidia Volta
Total Cores	737,000
Peak	7.4 Petaflops
Top500	5.2 Petaflops
Memory	172 Terabytes
Peak AI Flops	100+ Petaflops
Network Link	Intel OmniPath 25 GB/s

*Based on 2020 Top500.org
AI Flops = 4x4 matrix multiply half precision in, single precision out (mixed precision training)

Research Goals

Challenge:

- **Improve energy efficiency of AI applications *without making large structural changes to infrastructure or code*?**

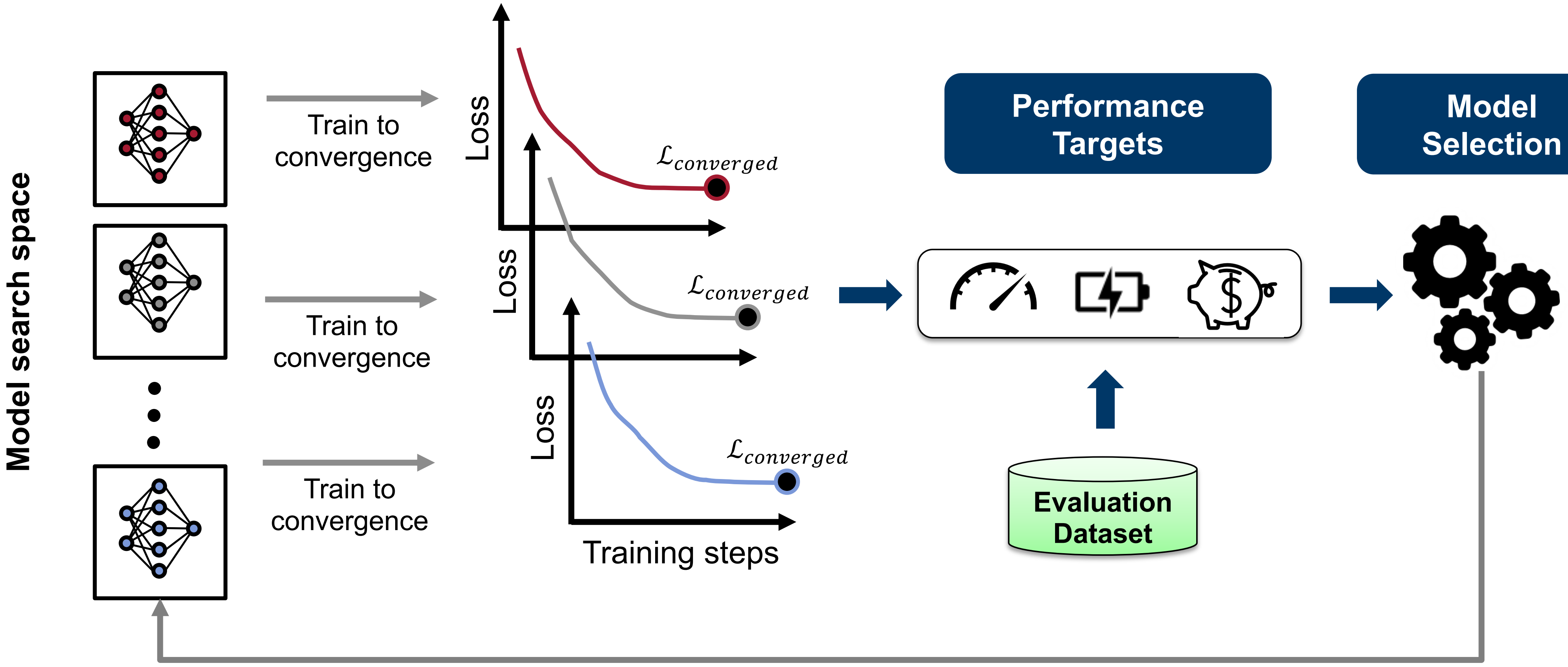
Approaches – and example results:

- **Better application usage - More efficient AI development**

- **Improve datacenter efficiency - Reduce hardware energy usage**

- **Reduce carbon intensity - Shifting computations for efficiency**

Efficient AI Model Development

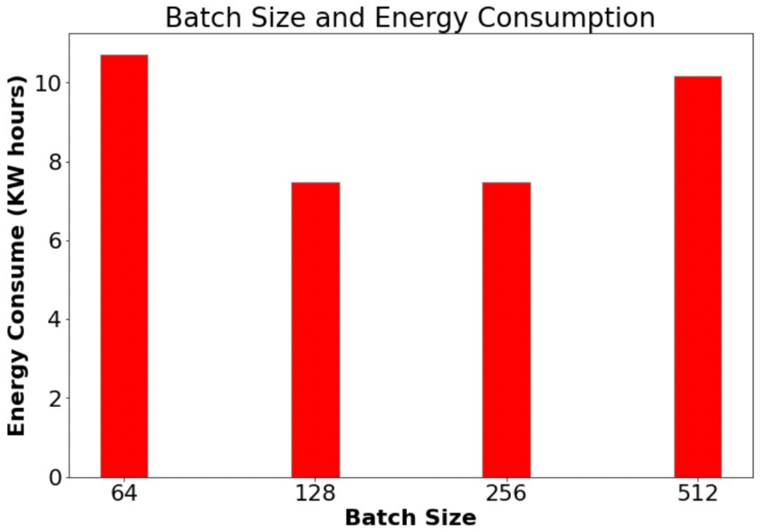


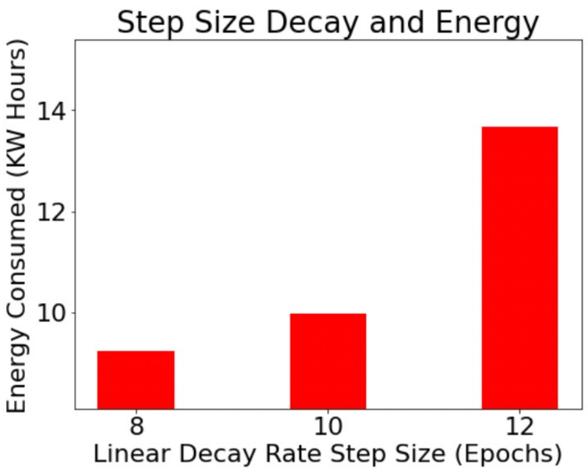
Architecture searches and parameter optimization have significant compute requirements

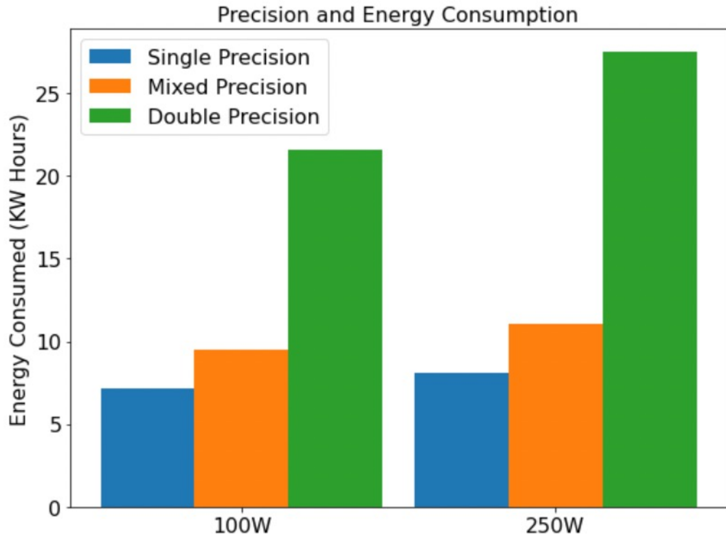
[1] Energy-aware neural architecture selection and hyperparameter optimization – Frey, et. al,, *IEEE IPDPS ADOPT 2022*
[2] Neural Scaling of Deep Chemical Models – Frey et. al, Nature Machine Intelligence (under review)

Why do hyper-parameter searches?

- **Hyper-parameter and training settings have significant impact to training time and energy consumed**
- **For example, ResNet on ImageNet based on MLPerf Challenge**
- **Example tuning settings**
  - **Batch Size (~20% savings possible)**
  - **Precision (going from mixed->single; 25% savings)**
  - **Step Size Linear Decay**







Modeling performance: training speed estimation (TSE)

How do we speed up *time to performance* for new models and datasets?

Area under training loss curve

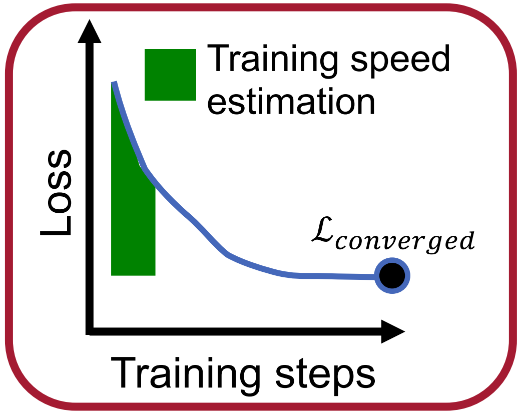
Features and Associated Labels

$$\mathrm{TSE} = \sum_{t=1}^{T} \left[\frac{1}{B} \sum_{i=1}^{B} \ell\left(f_{\theta_{t,i}}(\mathbf{X}_i), \mathbf{y}_i\right) \right]$$

Neural Network

Number of Completed Epochs

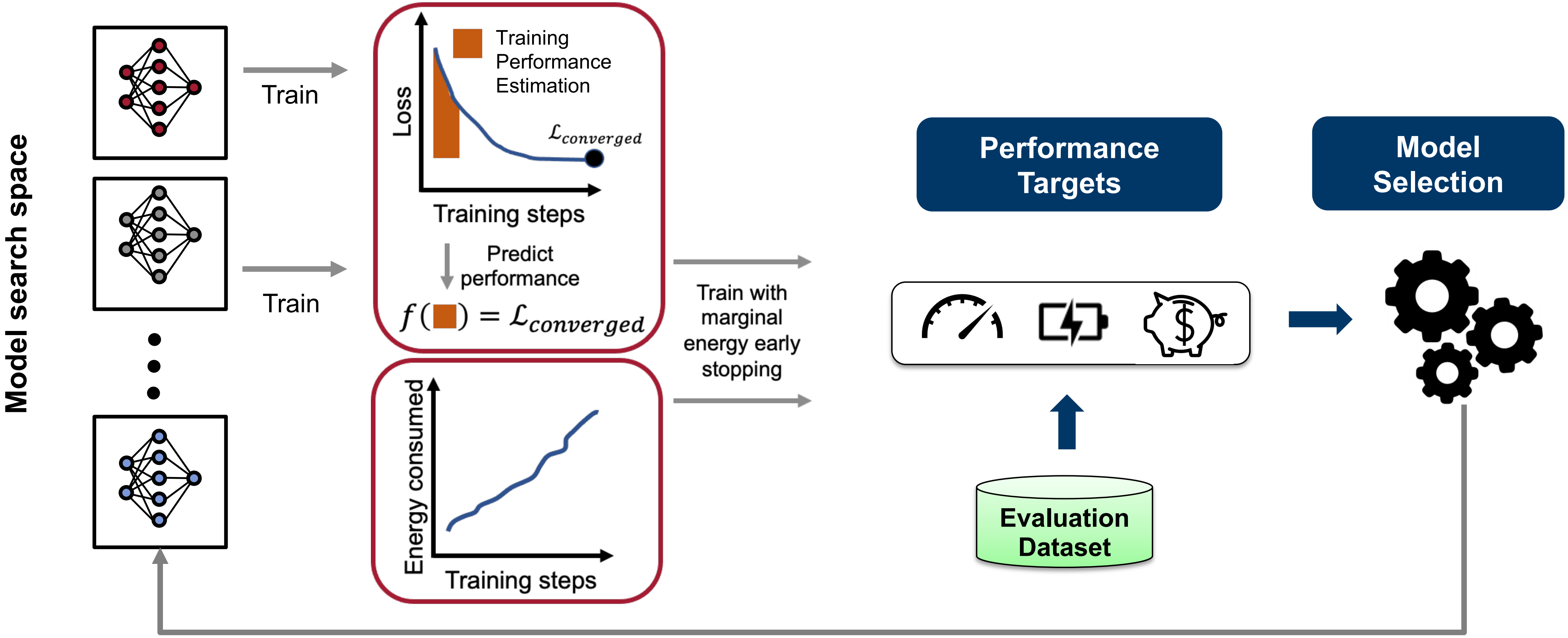
Loss Function



- **TSE is a simple, efficient, computationally cheap method for neural architecture search**

(Ru, Robin, et al. "Speedy Performance Estimation for Neural Architecture Search." *Advances in Neural Information Processing Systems* 34 (2021).)

Intervention for Efficient Neural Architecture Search and Hyperparameter Optimization

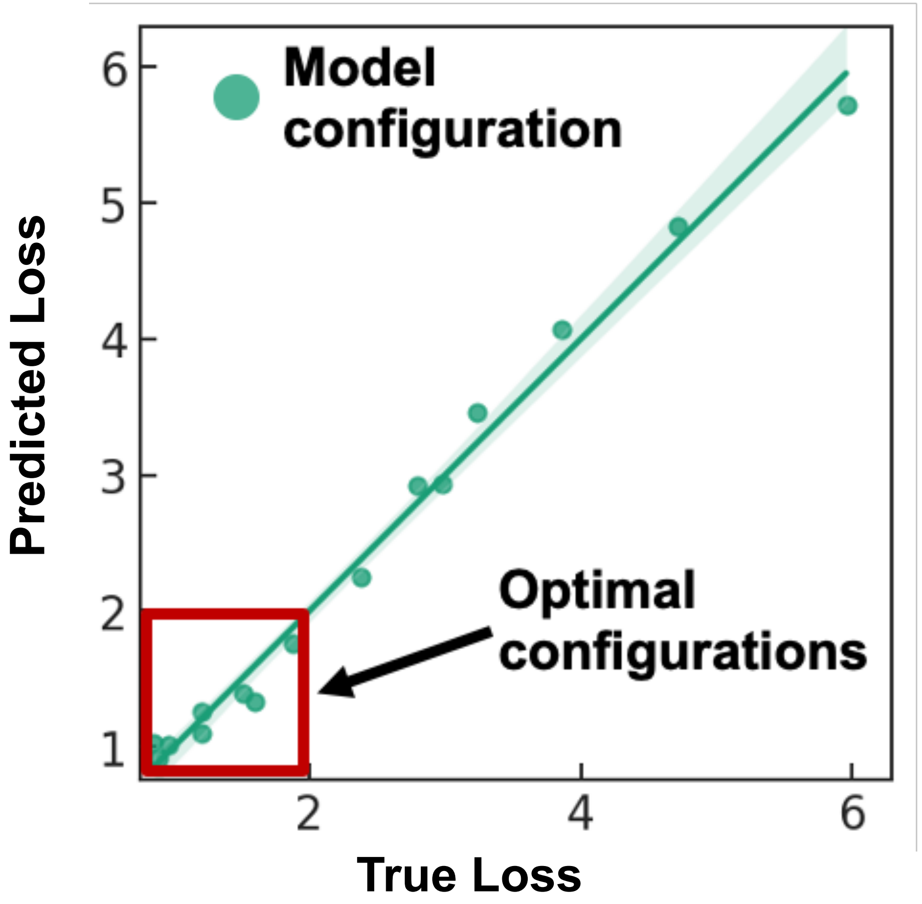


Training performance estimation (TPE) combines training speed estimation and energy consumption tracking to minimize energy expenditure

[1] Energy-aware neural architecture selection and hyperparameter optimization – Frey, et. al,, *IEEE IPDPS ADOPT 2022*
[2] Neural Scaling of Deep Chemical Models – Frey et. al, Nature Machine Intelligence (under review)

Energy-Efficient Neural Architecture Optimization for Graph Neural Networks

Predicted Model Performance for SchNet[2]



80% **total energy savings with early identification of optimal training configurations**

[1] Neural Scaling of Deep Chemical Models - Frey, et. al. Nature Machine Intelligence (under review) 2022

[2] Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, Schutt, et. al, *NeurIPS 2017*

Reducing Hardware Energy Usage

- **Hardware mechanisms to reduce energy:**
  - **Power Capping**
  - **Clock frequencies scaling**

- **Experimental setup for Natural Language Processing, Computer Vision Models:**
  - **Model architecture choices: BERT, DistilBERT, BigBird, ResNet, …**

- **GPU architectures: V100, A100, K80, T4**
  - **Varied outcomes when testing newer (A100) and older (T4, K80) NVIDIA devices**

Initial experiments indicate significant power savings, lower operating temperatures with only modest impact to computational performance

[1]Average US household ~29kWh/day (https://www.eia.gov/tools/faqs/faq.php?id=97&t=3)

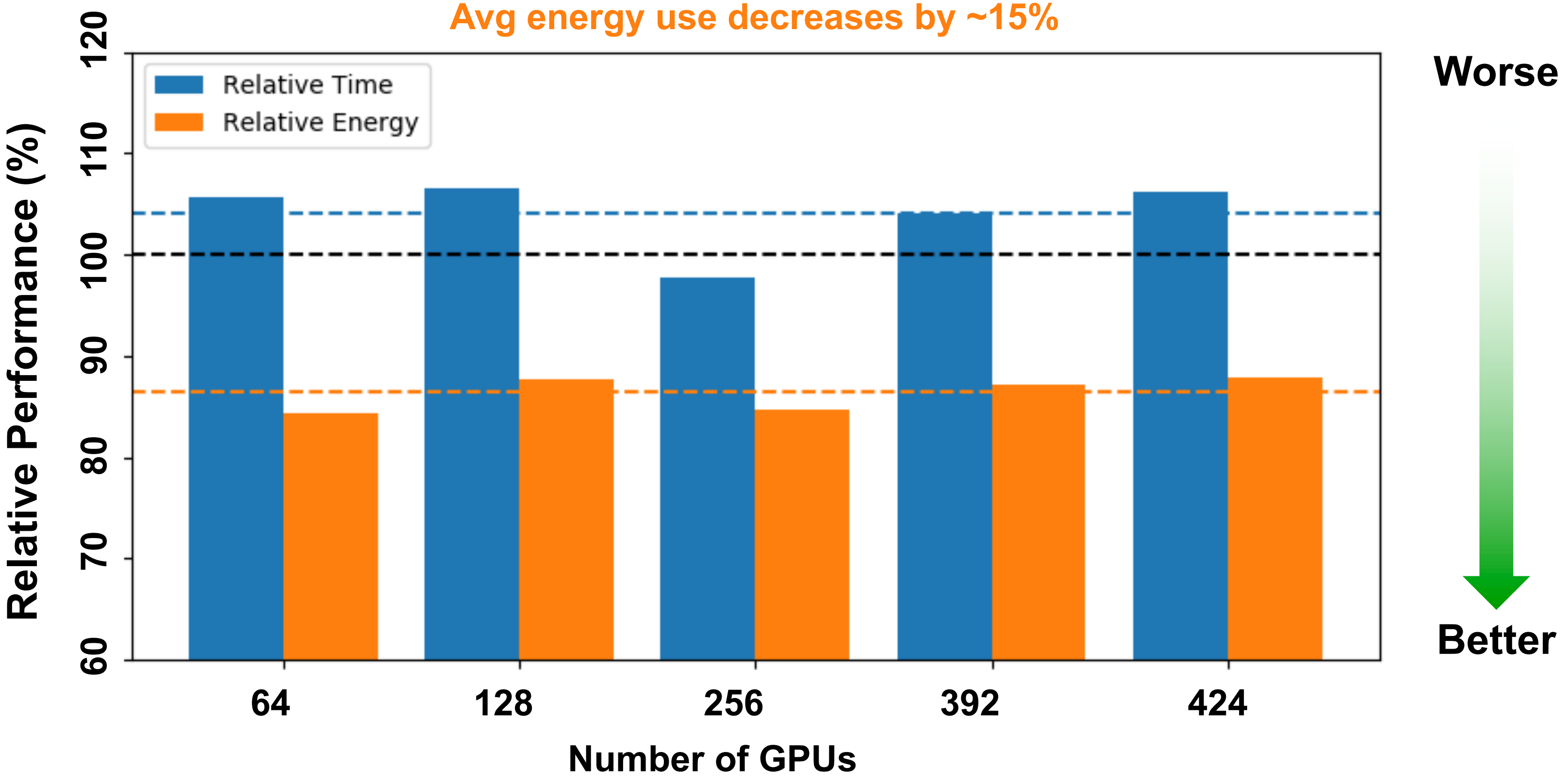
[2]Full BERT training estimates from Strubell, et. al., Energy and policy considerations for deep learning in NLP. ACL 2019

Energy Tuning on Existing Hardware

BERT training on V100 GPU with 60% power limit

Avg training time increase < 5%

Avg energy use decreases by ~15%

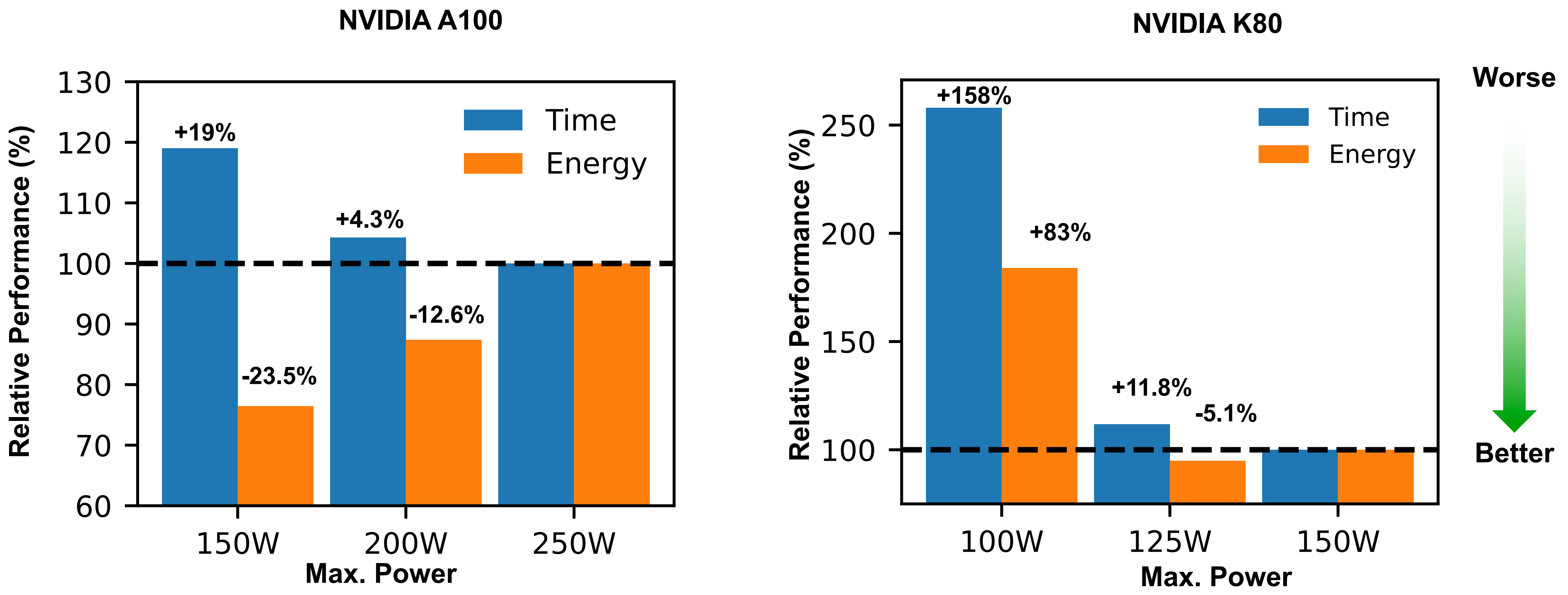


For a modest ~3-hour increase in training time, this intervention can save over a week's[1,2] worth of household energy usage.

[1]Average US household ~29kWh/day (https://www.eia.gov/tools/faqs/faq.php?id=97&t=3)

[2]Full BERT training estimates from Strubell, et. al., Energy and policy considerations for deep learning in NLP. ACL 2019

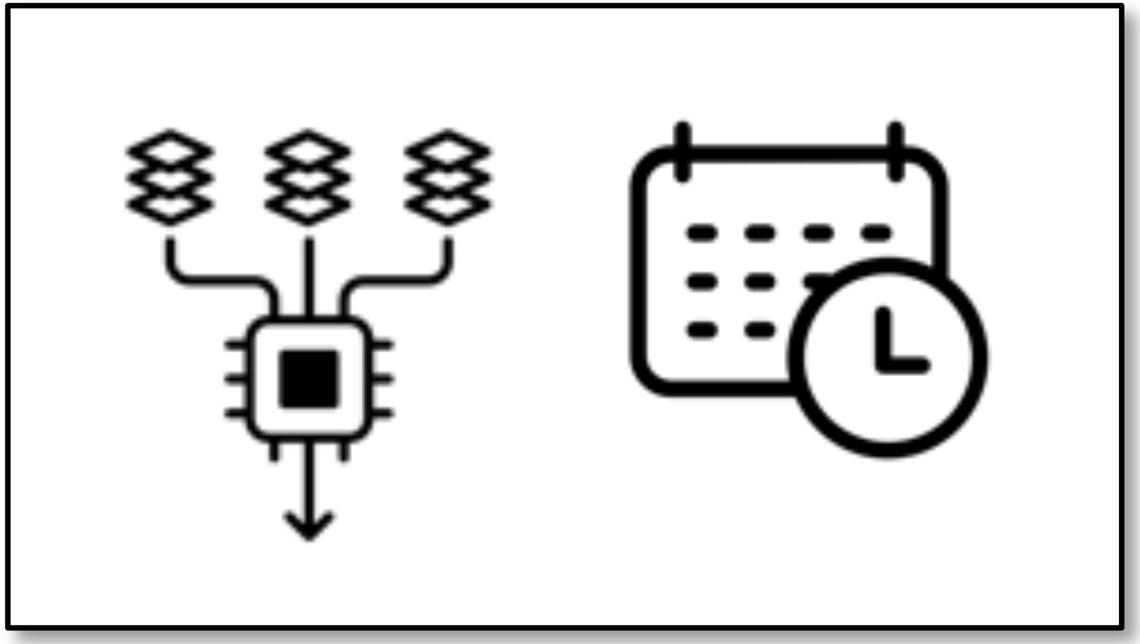
Energy Tuning on Hardware



Power-capping effective across GPU architectures

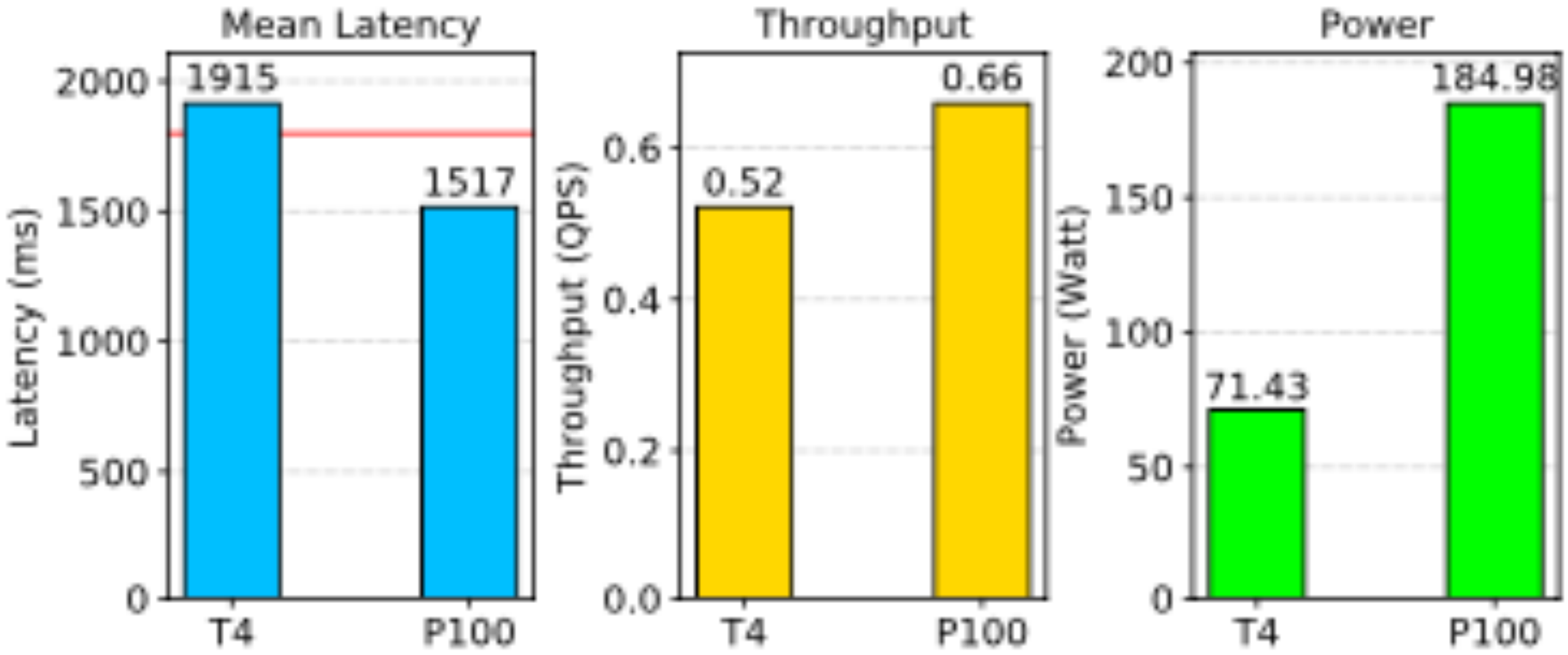
 [1] Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models – Findings of the NAACL 2022

Energy-Aware Scheduling



- Schedule jobs on efficient hardware
- Carbon-aware scheduling

Scheduling on Efficient Hardware



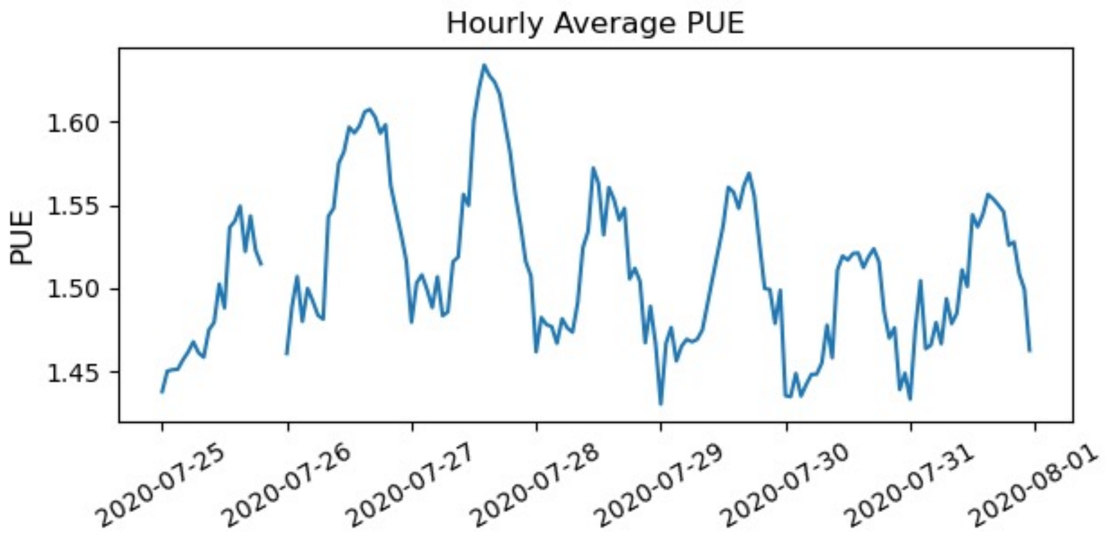
Idea: Pick the hardware platform best suited to solve the problem given application constraints (e.g., lowest latency, fastest throughput, lowest energy,…)

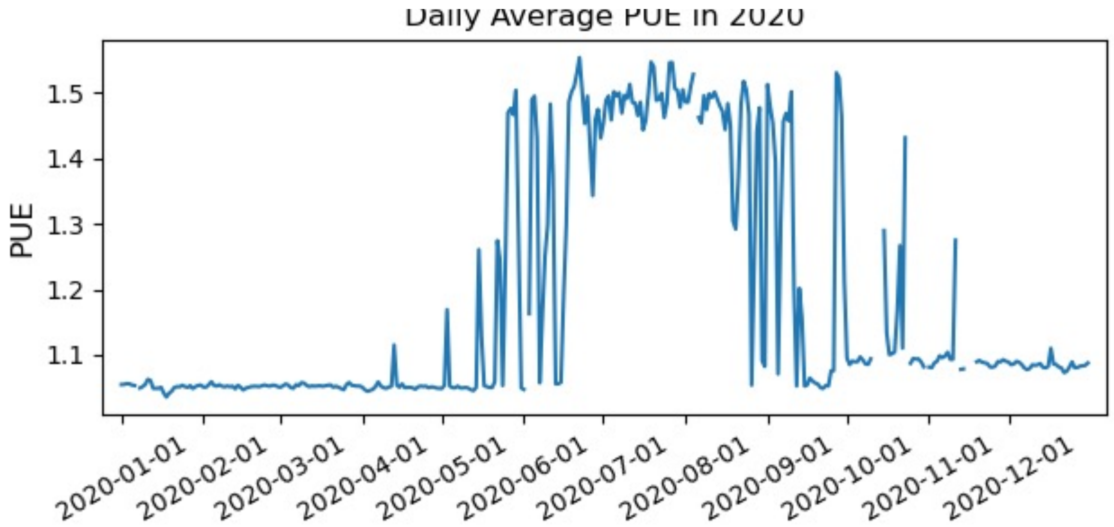
RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

Carbon-Aware Scheduling

- **Datacenter efficiency varies based on compute workloads, environmental factors,…**
  - **Correlated with carbon intensity**
- **Moving a workload from day->night:**
  - **~7.5% energy savings (annual average)**
  - **~20% energy savings (hot days)**
- **Moving from a hot days-> a cold day:**
  - **Nearly 30-40%! (e.g., summer->winter)**
  - **Geographically distribute datacenters?**

Idea: Leverage less carbon intense days, times and locations to run heavy workloads



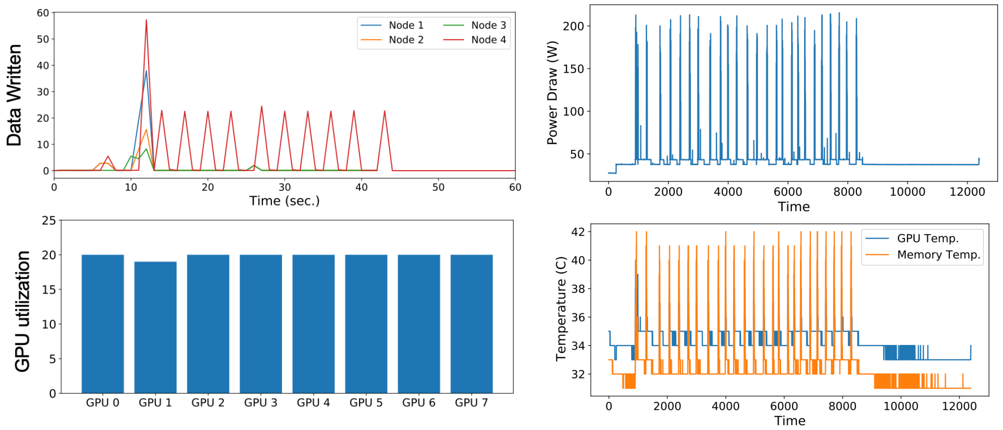


PUE = Power Usage Effectiveness

Collaboration Opportunities?

Green AI Challenge

- **No benchmark for training/testing machine learning models focusing on energy usage**
- **Green AI Benchmarks: tasks similar to existing benchmarks with energy baselines:**
  - **Problem definition and metrics**
  - **Model categories/constraints, training/validation datasets**
  - **Reasonable target accuracy**
  - **Baseline implementations with associated energy stats**
- **Open sourcing data from our datacenter**



Energize research into reducing operational footprint with smarter computing technique and algorithms

https://dcc.mit.edu/
https://news.mit.edu/2022/taking-magnifying-glass-data-center-operations-0824

Understanding Opportunities with your Organization

Acknowledgements


Siddharth Samsi


David Bestor


Nathan Frey


Mike Jones


Joseph McDonald


Matthew Weiss

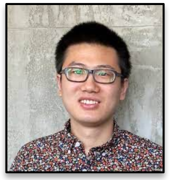

Daniel Edelman

U.S. AIR FORCE


Lt. Col. Andrew Bowne


Capt. Lindsey McEvoy


Prof. Devesh Tewari


Baolin Li

And many others…

- Charles Leiserson (CSAIL)
- Tim Kraska (CSAIL)
- Manya Ghobadi (CSAIL)
- Sam Madden (CSAIL)
- Mike Stonebraker (CSAIL)
- T.B. Schardl (CSAIL)
- Anson Cheng (USAF)
- Allan Vanterpool (USAF)
- Andrew Kirby (Alum)
- Emily Do (Alum)
- Matthew Hutchinson (Alum)
- William Arcand
- William Bergeron
- Chansup Byun
- Matthew Hubbell
- Michael Houle
- Hayden Jananthan
- Jeremy Kepner
- Anna Klein
- Peter Michaleas
- Lauren Milechin
- Julie Mullen
- Andrew Prout
- Albert Reuther
- Antonio Rosa
- Pat Ross
- Charles Yee
- Stephen Rejto
- Jeff Gottschalk
- Marc Zissman
- Dave Martinez
- Mark Veillette
- Bob Bond
- Jason Williams
- Brad Dillman
- Daniela Rus
- Col. Garry Floyd
- CK Prothmann

Summary

- Compute and energy requirements of AI are growing at an unsustainable rate.

- Tradeoffs between AI performance and energy consumption can offer significant opportunities for carbon reduction.

- Numerous approaches to reducing footprint
  - Technological, behavioral, economic, environmental, social implications

Looking for partners!
Email: vijayg@ll.mit.edu