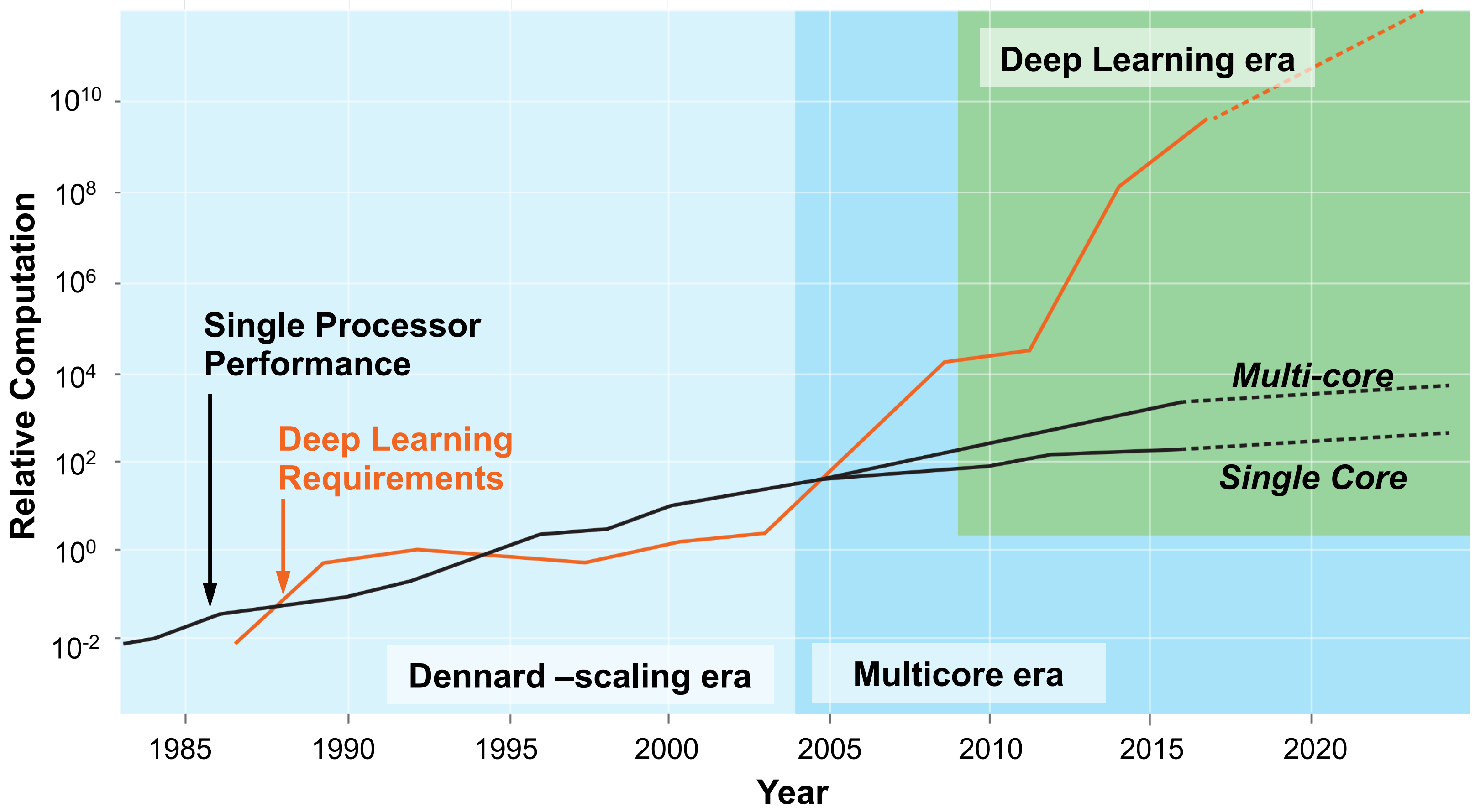
Understanding (and reducing) the Energy Impact of AI

Vijay Gadepally
vijayg@mit.edu

Siddharth Samsi, Joseph McDonald, Daniel Edelman
Baolin Li (Northeastern University), Devesh Tiwari (Northeastern University)

MIT LINCOLN LABORATORY
SUPERCOMPUTING CENTER

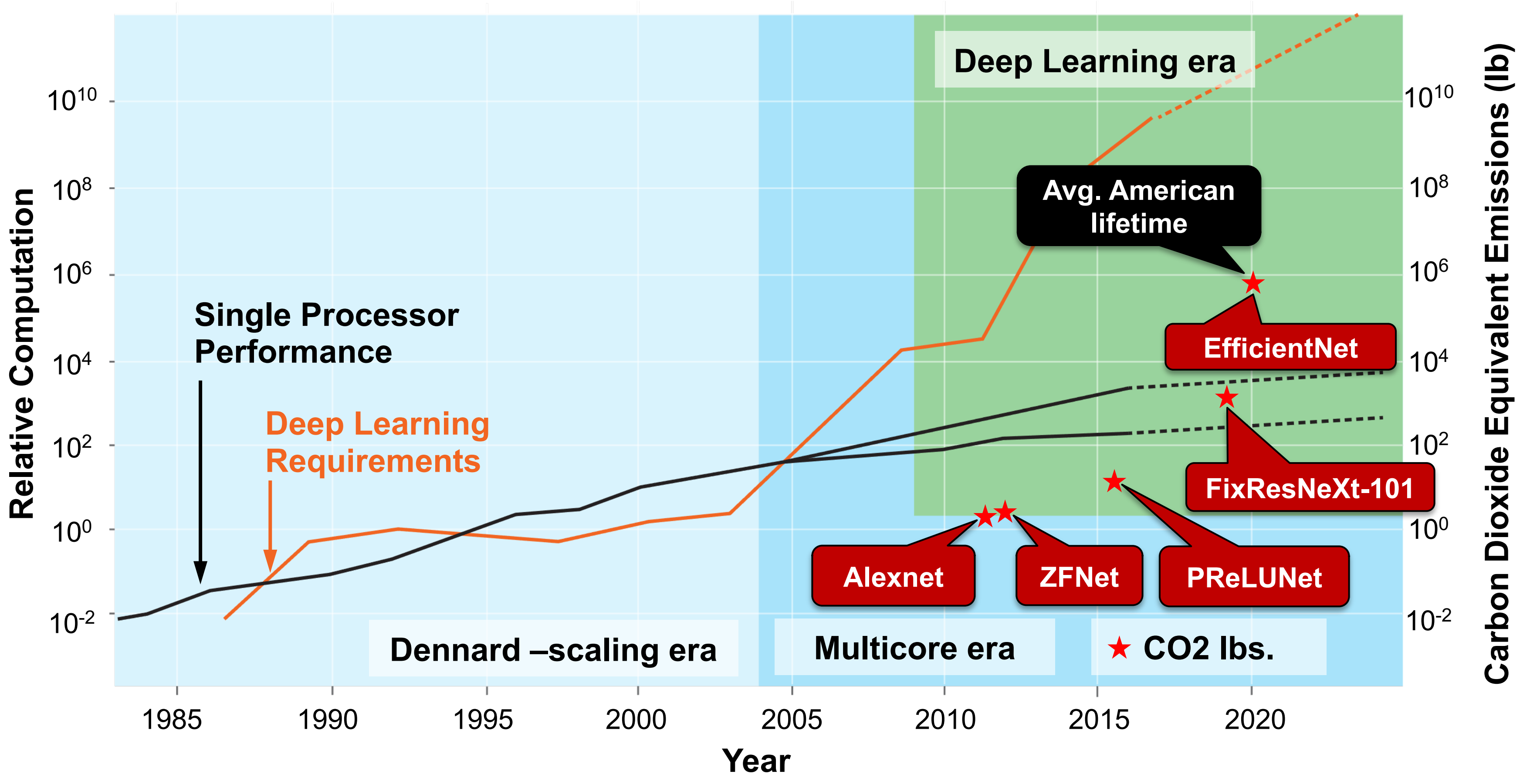
Growth of AI Computing Requirements



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

Other Facets

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

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

- Environmental footprint goes beyond “operational” energy usage
  – E.g., carbon costs of hardware manufacturing

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

Zhao, et. al. “A Greener World for A.I.”, IPDPS ADOPT 2022 (Accepted)

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

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:

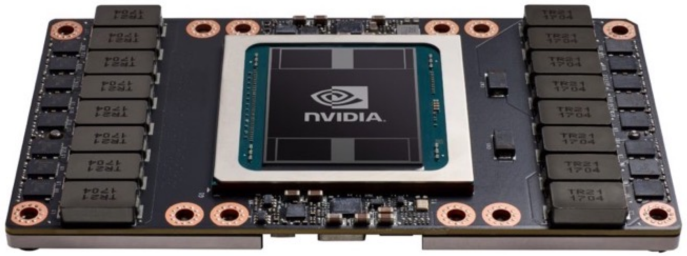
- Prioritizing energy-efficient models
- More experiments run, more computation, more energy consumed
- Awareness of environmental footprint of AI research, applications

How can we make A.I. research and practice more sustainable?

Our Testbed

Low Carbon Emission

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



- **<u>Operates on renewable energy</u>**

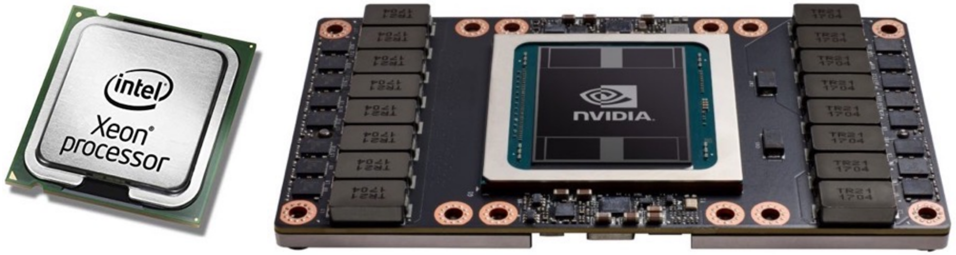
	Capability
Processor	Intel Xeon & Nvidia Volta
Total Cores	737,000
Peak	7.4 Petaflops
Top500	5.2 Petaflops
Memory	172 Terabytes
Peak AI Flops	100+ Petaflops
Network Link	Intel OmniPath 25 GB/s

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

Sources of Carbon Emissions in HPC

Scope 2: Data Center Operations

Scope 3: Embodied (Manufacturing)



Outline

- Introduction
- Reducing Operational Footprint
- Modelling Embodied Footprint
- Next Steps

Reducing Operational Carbon Emissions

<u>Challenge</u>:

- **Decrease the footprint of operational AI applications *without making large structural changes to infrastructure or code*?**

<u>Solution approaches</u>:

- **More efficient code, training practices**

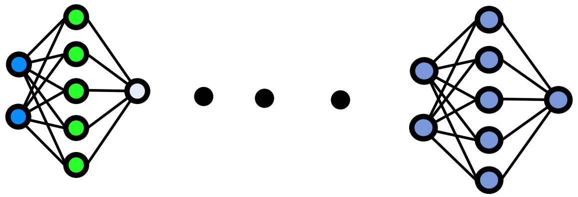
- **Tuning hardware on individual nodes**

- **Improving datacenter operations**

Reducing Development Environment Computing Demands

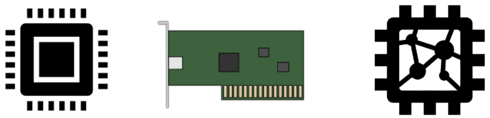
Challenge

Model Development



- Model design, testing, and development
- AI training & inference

Hardware Usage Strategies



- Hardware variety
- Matching workload to hardware capabilities

Datacenter Operations

- Hardware power modulation

Proposed Approach

Model Development

- AI-enabled Model Discovery[1]
- Knowledge Informed Models

Hardware Usage Strategies

- Hardware-based interventions
- ML-based hardware selection[2]

Datacenter Operations

- Power limiting[3]
- Clock frequency scaling[3]
- Auto-tuning complex applications[4]

[1]Neural Scaling of Deep Chemical Models – Frey, et. al, *Nature Machine Intelligence (submitted)*

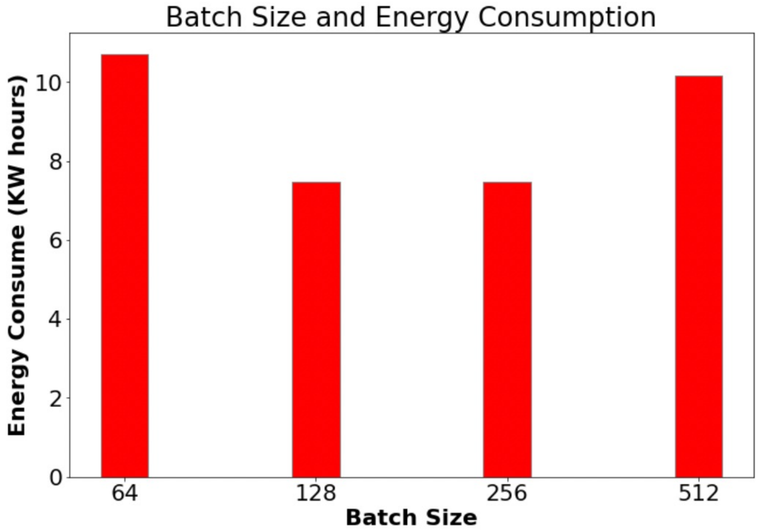
[2] DASH: Scheduling Deep Learning Workloads on Multi-Generational GPU-Accelerated Clusters – Li, et. al., IEEE HPEC 2022

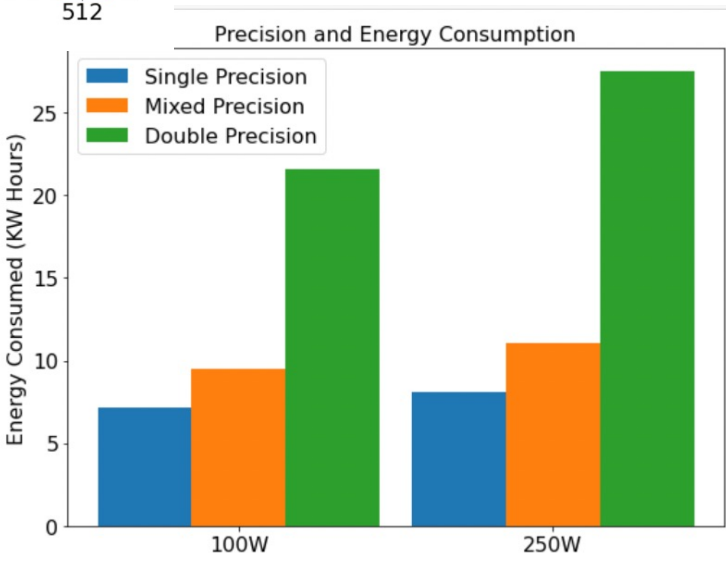
[3] Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models – McDonald, et. al., NAACL 2022

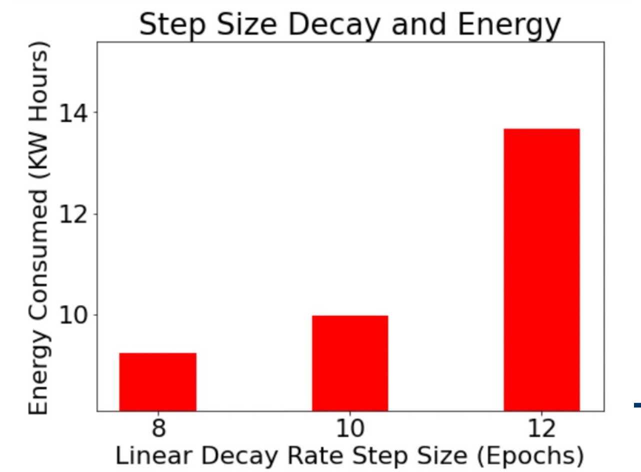
[4] Bliss: auto-tuning complex applications using a pool of diverse lightweight learning models – Roy, et. al., *PLDI 2021*

Example: Energy Optimizing Hyper Parameters

- **Hyper-parameter and training settings can have significant impact tot both training time and energy consumed**
- **Early results when training a ResNet on ImageNet based on MLPerf Challenge**
- **Example settings**
  - **Batch Size (~20% savings possible)**
  - **Precision (going from mixed->single: 25% savings)**
  - **Step Size Linear Decay**

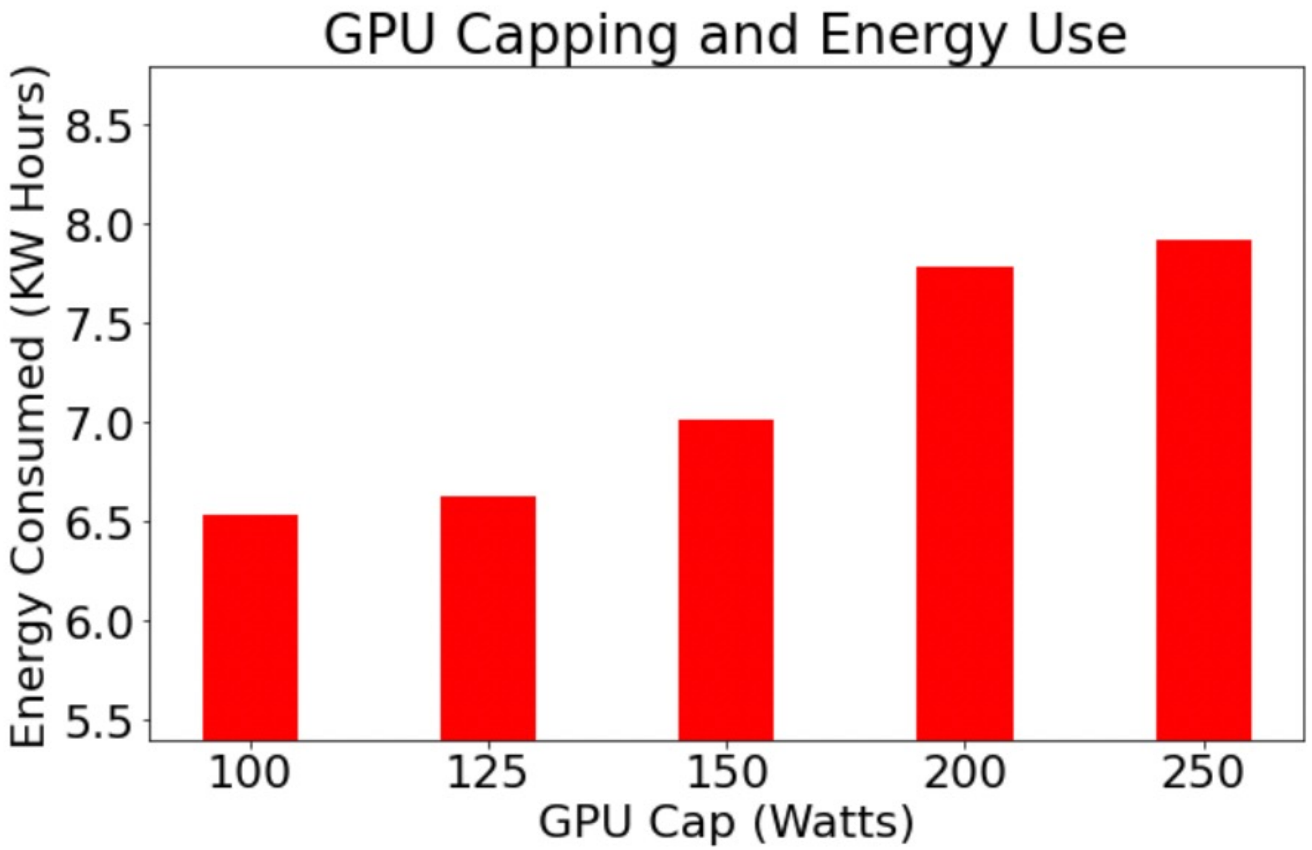






Example: Hardware Tuning

- **Compared energy usage of different power-caps for training, inference with ImageNet**
- **Power-cap choices: 250W (default) versus 200W, 150W, 125W, 100W**
- **Caps typically reduced energy usage with no statistically significant change in runtime**
- **Lower Power Cap seems optimal reducing energy use with an insignificant change in job runtime**



Simple hardware interventions provide ~10-15% energy savings with minimal impact to performance

Example: Energy-aware scheduling

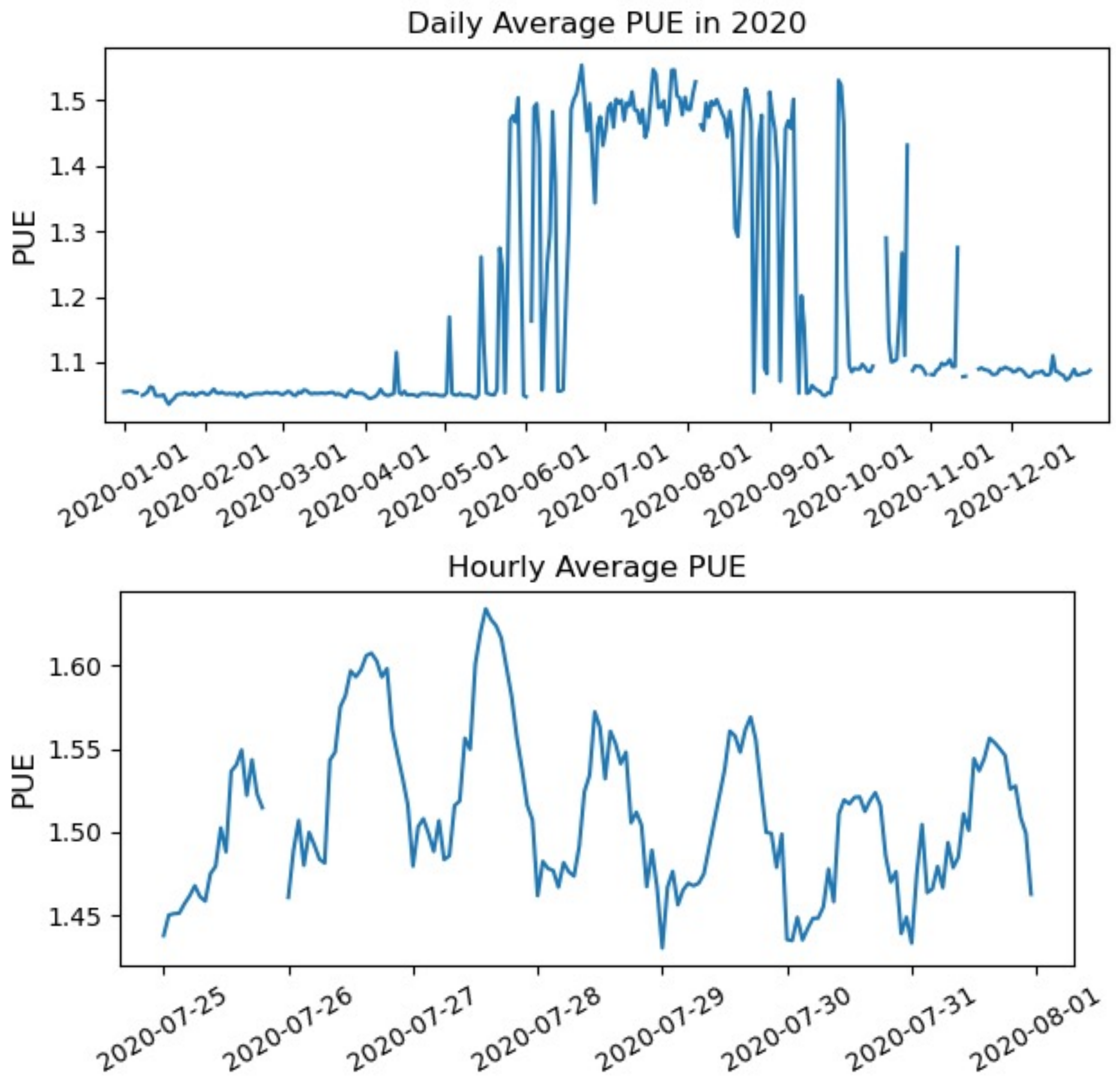
- **Datacenter PUE varies continuously, depending on compute workloads and cooling power**

- **Daily variation in PUE computed as percent difference between max hourly average PUE and min:**

$$\frac{\max(PUE) - \min(PUE)}{\min(PUE)}$$

- **Average daily variation is 7.3% over all 2020**

Time-shifting compute-intensive jobs could save up to 20% energy



Outline

- Introduction
- Reducing Operational Footprint
- Estimating Embodied Footprint
- Next Steps

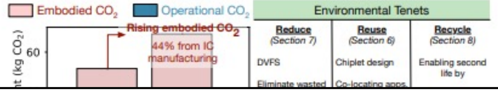
Carbon footprint has become an important topic in systems research

ACT: Designing Sustainable Computer Systems With An Architectural Carbon Modeling Tool

Udit Gupta, ugupta@g.harvard.edu, Harvard University/Meta, USA
Mariam Elgamal, mariamelgamal@g.harvard.edu, Harvard University, USA
Gage Hills, ghills@g.harvard.edu, Harvard University, USA
Gu-Yeon Wei, guyeon@seas.harvard.edu, Harvard University, USA
Hsien-Hsin S. Lee, leehs@fb.com, Meta, USA
David Brooks, dbrooks@eecs.harvard.edu, Harvard University/Meta, USA
Carole-Jean Wu, carolejeanwu@fb.com, Meta, USA

ABSTRACT

Given the performance and efficiency optimizations realized by the computer systems and architecture community over the last decades, the dominating source of computing's carbon footprint



Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters

Bilge Acun, acun@meta.com, Meta, USA
Benjamin Lee, leebcc@seas.upenn.edu, University of Pennsylvania, Meta, USA
Fiodar Kazhamiaka, fiodar@stanford.edu, Stanford University, USA
Kiwan Maeng, kwmaeng@meta.com, Meta, USA
Udit Gupta, uditg@meta.com, Harvard University, Meta, USA
Manoj Chakkaravarthy, mchakkar@meta.com, Meta, USA
David Brooks, dbrooks@eecs.harvard.edu
Carole-Jean Wu, carolejeanwu@meta.com

Sustainable AI: Environmental Implications, Challenges and Opportunities

Carole-Jean Wu [1] Ramya Raghavendra [1] Udit Gupta [1 2] Bilge Acun [1] Newsha Ardalani [1] Kiwan Maeng [1] Gloria Chang [1] Fiona Aga Behram [1] James Huang [1] Charles Bai [1] Michael Gschwind [1] Anurag Gupta [1] Myle Ott [1] Anastasia Melnikov [1] Salvatore Candido [1] David Brooks [1 2] Geeta Chauhan [1] Benjamin Lee [1 3] Hsien-Hsin S. Lee [1] Bugra Akyildiz [1] Max Balandat [1] Joe Spisak [1] Ravi Jain [1] Mike Rabbat [1] Kim Hazelwood [1]

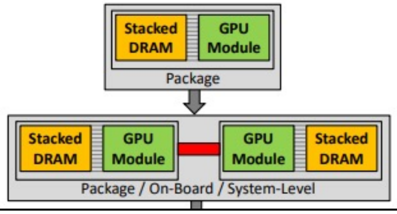
ABSTRACT

This paper explores the environmental impact of the super-linear growth trends for AI from a holistic perspective, spanning *Data*, *Algorithms*, and *System Hardware*. We characterize the carbon footprint of AI computing by

Understanding the Future of Energy Efficiency in Multi-Module GPUs

Akhil Arunkumar*, Evgeny Bolotin†, David Nellans†, and Carole-Jean Wu*
*Arizona State University, † NVIDIA
Email: {akhil.arunkumar, carole-jean.wu}@asu.edu, {ebolotin, dnellans}@nvidia.com

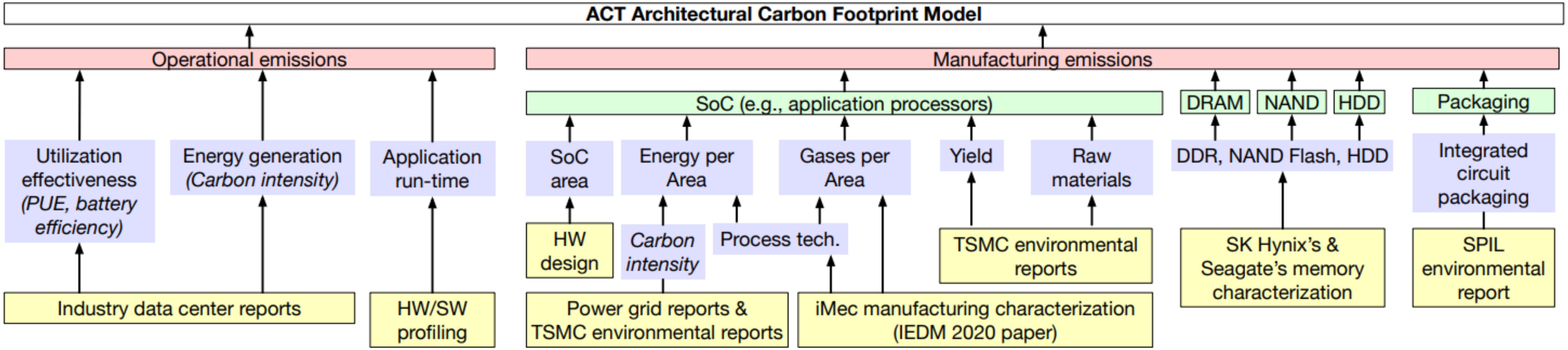
Abstract—As Moore's law slows down, GPUs must pivot towards multi-module designs to continue scaling performance at historical rates. Prior work on multi-module GPUs has focused on performance, while largely ignoring the issue of energy efficiency. In this work, we propose a new metric for GPU efficiency called EDP Scaling Efficiency that quantifies the effects of both strong performance scaling and overall energy efficiency in these designs. To enable this analysis, we develop a novel top-down GPU energy estimation framework that is accurate within 10% of a recent GPU design. Being



Section Credit: Baolin Li

Carbon footprint modeling: the ACT approach

- **ACT (Gupta et. Al., ISCA'22) is a carbon footprint modeling tool. It organizes the carbon emission of a system into two categories**
  - **Embodied carbon**
  - **Operational carbon**



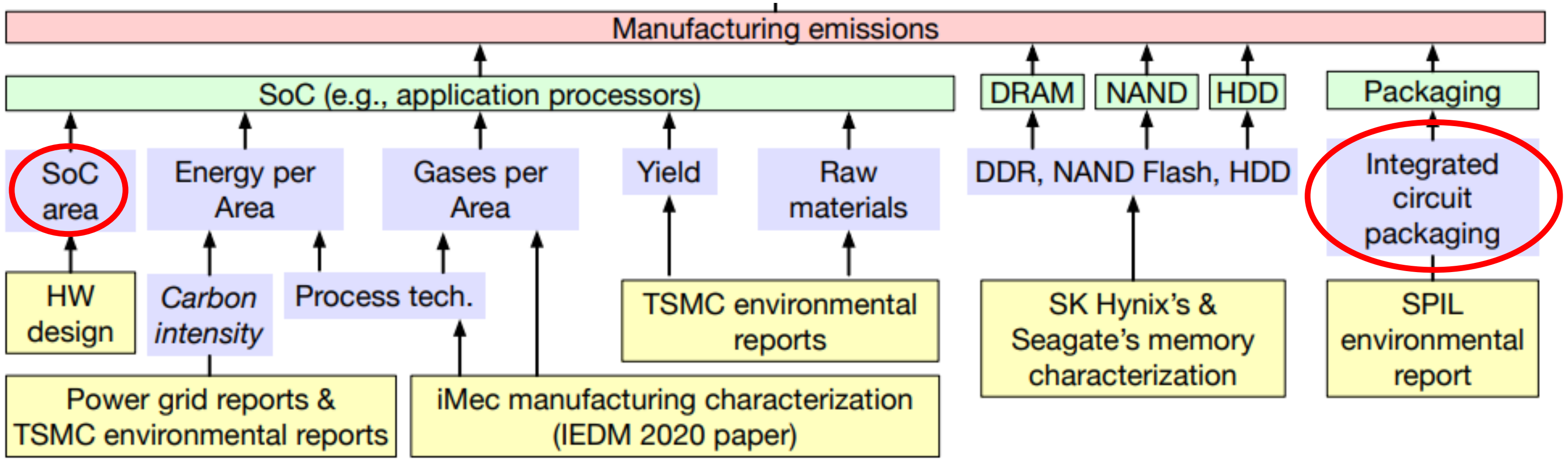
Gupta, Udit, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. "ACT: Designing sustainable computer systems with an architectural carbon modeling tool." In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pp. 784-799. 2022.

Goal of this presentation

Share our experience and the challenges we encountered while using the ACT tool to model the carbon footprint of a large-scale GPU-accelerated HPC system

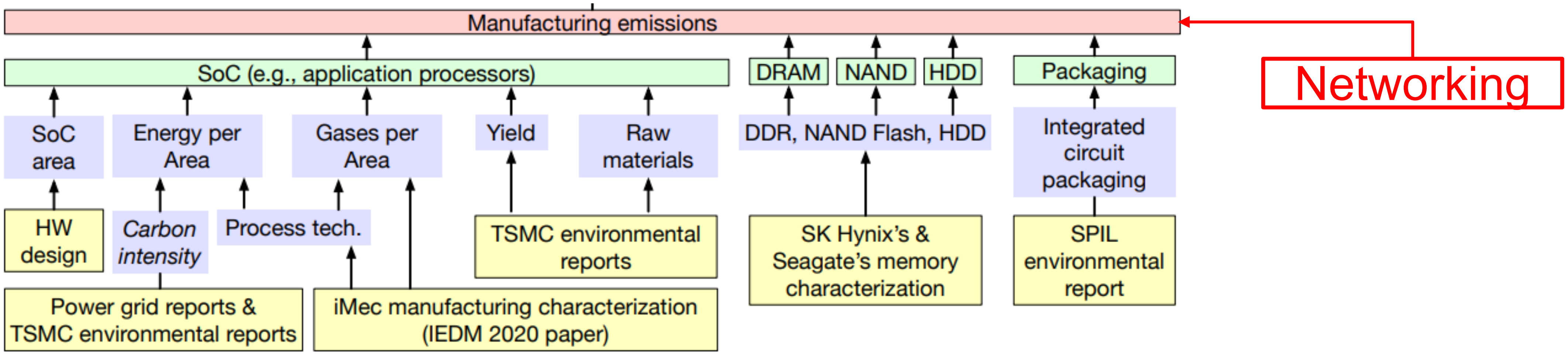
Embodied Footprint Modeling Challenge 1

- **Difficult to obtain information related to carbon footprint modeling from vendors' product datasheet, for example**
  - **Number of ICs packaged on a NVIDIA GPU card**
  - **Die area of Intel Xeon processors**



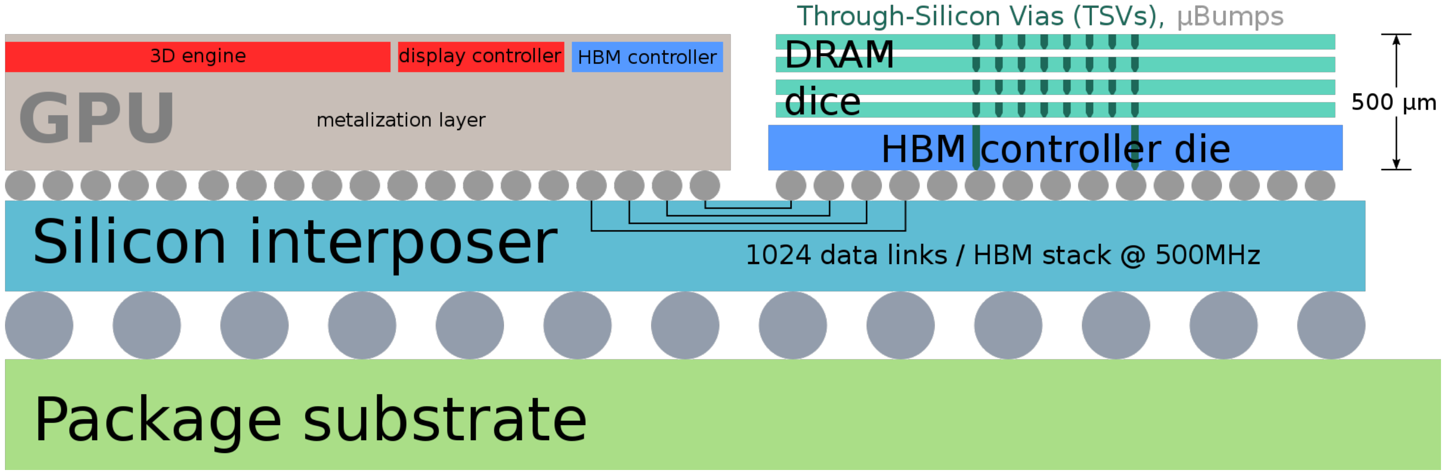
Embodied footprint modeling challenge 2

- **ACT's model works well for a single device, e.g., desktop, phone**
- **But lacks extensibility to large scale distributed systems**
  - **For example, the network fabrics for inter-node communication**



Embodied footprint modeling challenge 3

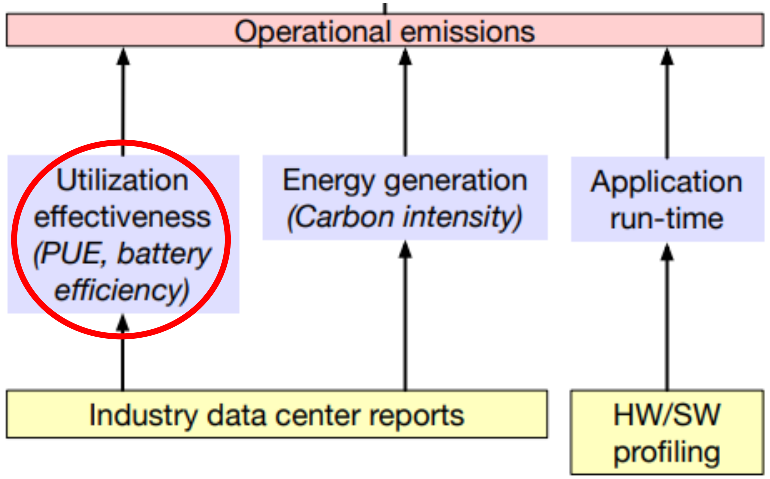
- **Need for GPU-specific features to model GPU-accelerated systems**
  - **ACT models GPUs like CPUs – based on the processor's die area**
  - **Modern GPUs use FinFET technology compared to traditional CMOS**
  - **GPUs such as NVIDIA V100 use HBM2 memory that is stacked vertically and integrated into the same package with the GPU cores**
    - **Unlike CPUs that use DDR4/DDR5 discrete memory chips**



https://en.wikipedia.org/wiki/High_Bandwidth_Memory

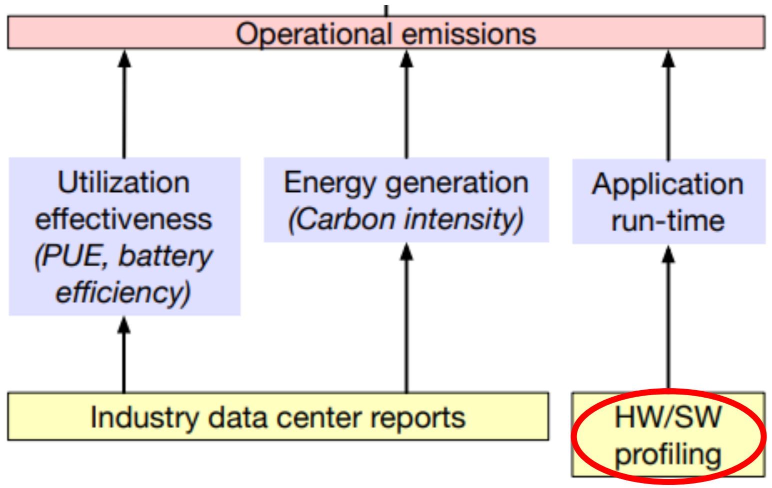
Operational carbon footprint challenge 1

- **Need for systematic power monitoring tool**
  - **We need to monitor CPU/GPU power at node level**
  - **Use this to estimate operational energy**
  - **Then convert to emitted carbon using real-time carbon intensity**
  - **Good to have a universal software suite that can be used in any datacenter in any location**



Operational carbon footprint challenge 2

- Difficult to estimate operational carbon emission on the next-generational system
  - When making system upgrade decisions, need to build carbon footprint model for the next generational system
  - But the HW/SW profiling for operational carbon is difficult to obtain from new hardware in the future
  - System operators also usually do not have information about the user workload



Summary and recommendations

- Hardware manufacturers
  - Provide more data to customers from the carbon perspective
- Embodied carbon modeling
  - Extension to audiences from HPC and distributed system field is needed
- Operational carbon modeling
  - Need for universal and systematic monitoring tool
  - Would be helpful for system operators to record history of previous hardware upgrades for reference

Section Credit:
Baolin's email: li.baol@northeastern.edu
Baolin's website: https://baolin-li.netlify.app/

Outline

- Introduction
- Reducing Operational Footprint
- Estimating Embodied Footprint
- Next Steps

Better Understanding Embodied Footprint: Provide Standardized Data!

- Difficult to estimate embodied footprint due to lack of data from manufacturers providers
- Opportunity to help vendors by making a standardized datasheet that can be filled out
- Calling on OCP community to help develop these guidelines
  – Important to make them easy to collect for vendors
  – Opportunities for third-party auditing in certain cases

Goal: Create standardized (and easy-to-implement) datasheets to better understand manufacturer carbon emissions

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

Make Energy Efficiency a First Order Priority

- No benchmark for training/testing machine learning models focusing on energy usage

- Common AI benchmarks
  - MLPerf gives suite of training benchmarks for hardware optimizations and time-to-completion for variety of research areas (Image Classification/NLP/Reinforcement)

- 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

Energize Research into Reducing Operational footprint through smarter computing technique and algorithms

Conclusions

- Growing energy impact of AI and machine learning
- Many low to no overhead changes that can be made to give big energy savings
  - Some can be done without user intervention
  - Minimal code changes needed
  - Starting point for much more user-in-the-loop feedback
- A best practice can save and reduce energy use before training larger and more complicated systems