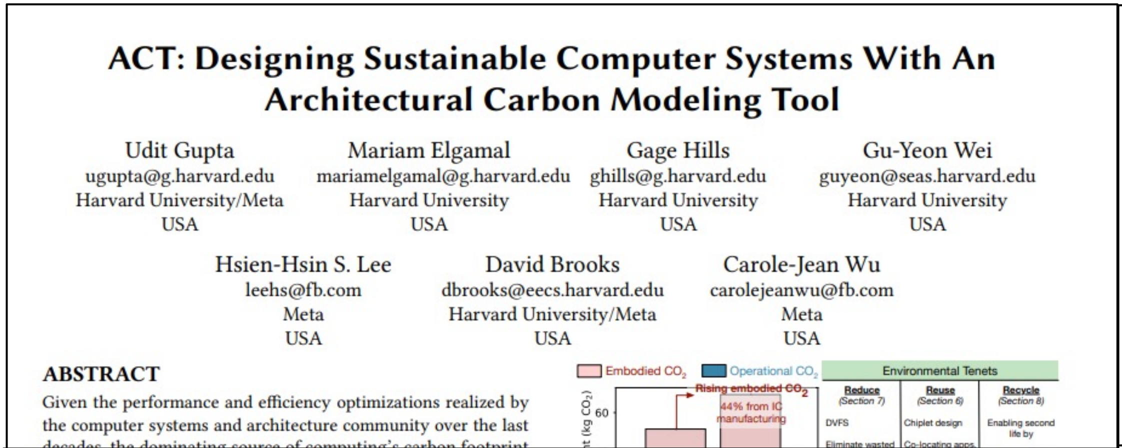
Challenges in Building the Carbon Footprint Model for Large-Scale GPU Systems

Baolin Li, Vijay Gadepally,
Siddharth Samsi, Devesh Tiwari

Northeastern University

MIT LINCOLN LABORATORY

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

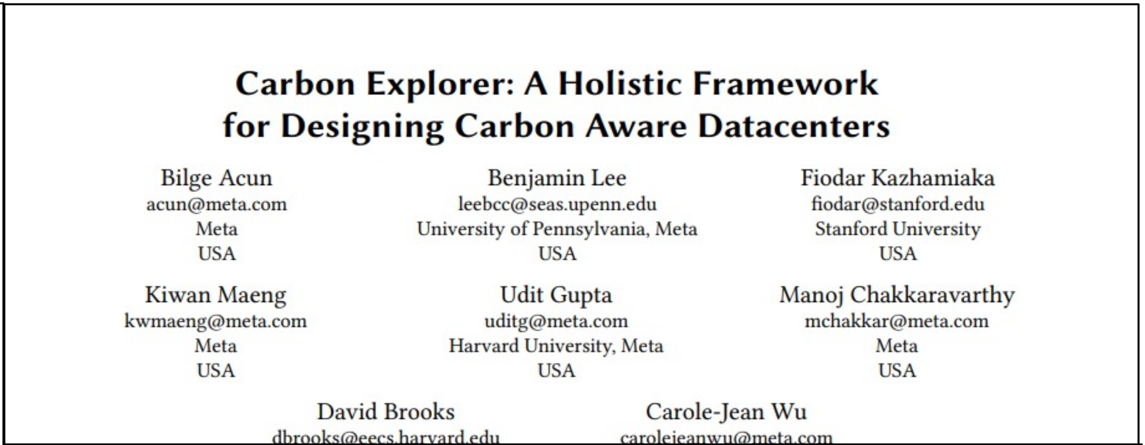
Hsien-Hsin S. Lee
leehs@fb.com
Meta
USA

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

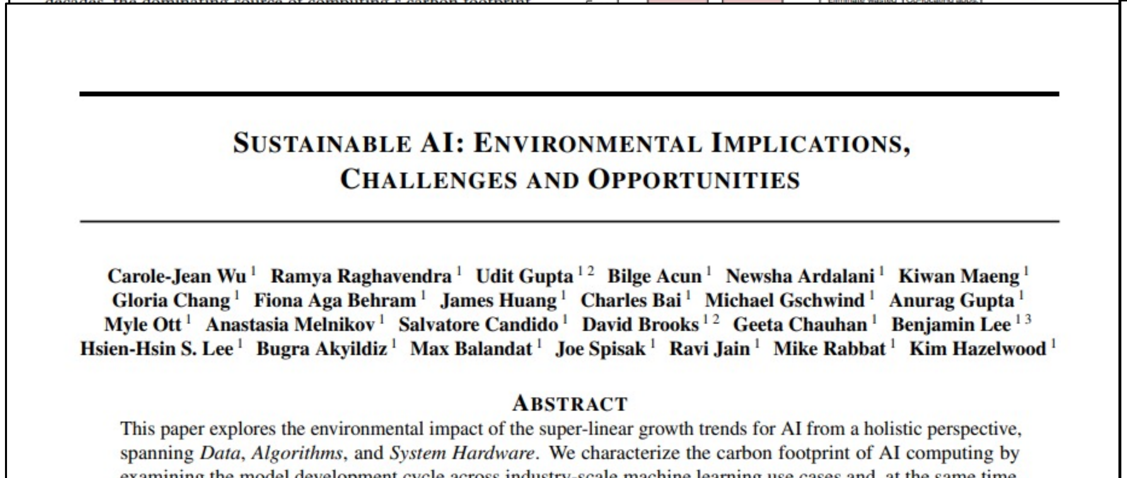
Kiwan Maeng
kwmaeng@meta.com
Meta
USA

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

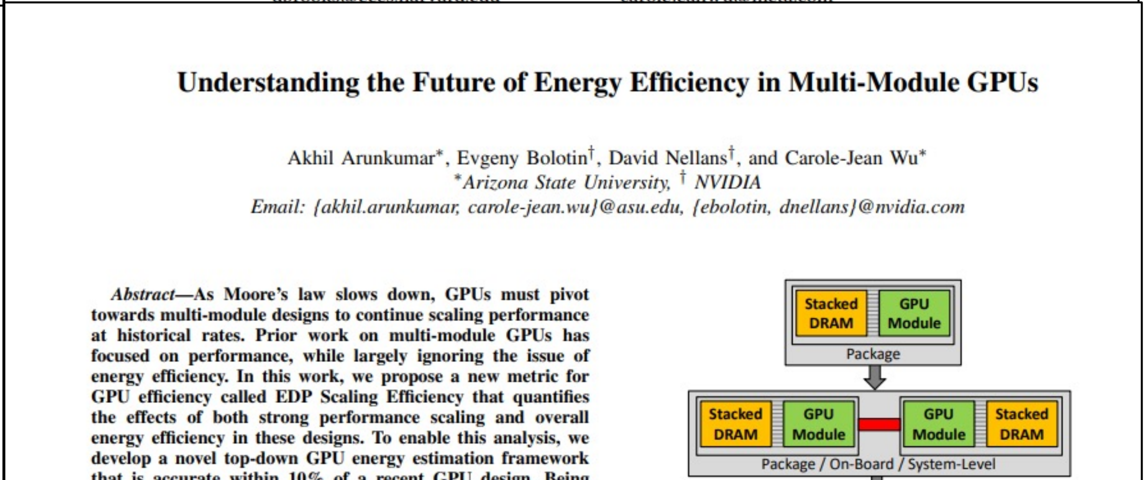


SUSTAINABLE AI: ENVIRONMENTAL IMPLICATIONS, CHALLENGES AND OPPORTUNITIES

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 examining the model development cycle across industry-scale machine learning use cases and, at the same time



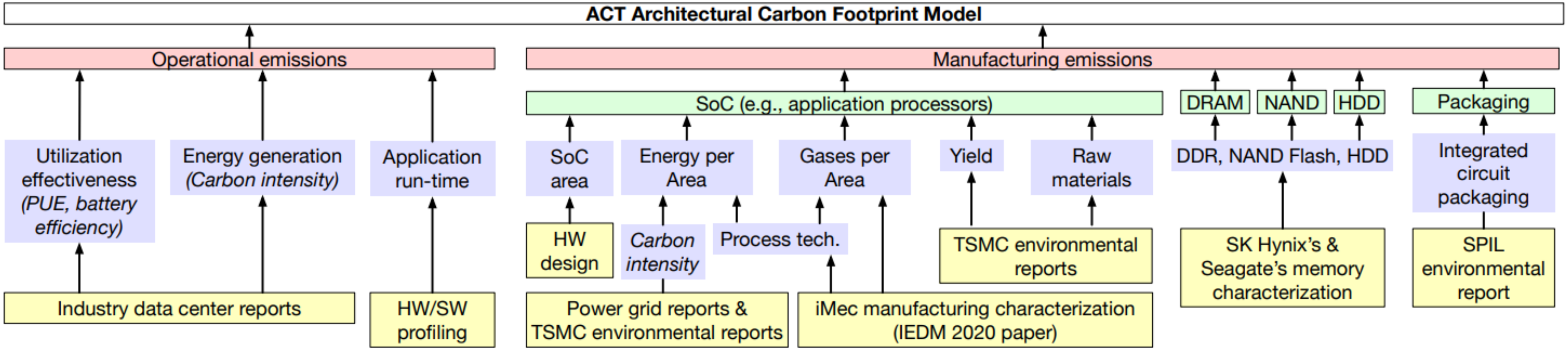
Understanding the Future of Energy Efficiency in Multi-Module GPUs

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

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**



ACT Architectural Carbon Footprint Model
Operational emissions
Manufacturing emissions
SoC (e.g., application processors)
DRAM
NAND
HDD
Packaging
Utilization effectiveness (PUE, battery efficiency)
Energy generation (Carbon intensity)
Application run-time
SoC area
Energy per Area
Gases per Area
Yield
Raw materials
DDR, NAND Flash, HDD
Integrated circuit packaging
HW design
Carbon intensity
Process tech.
TSMC environmental reports
SK Hynix’s & Seagate’s memory characterization
SPIL environmental report
Industry data center reports
HW/SW profiling
Power grid reports & TSMC environmental reports
iMec manufacturing characterization (IEDM 2020 paper)

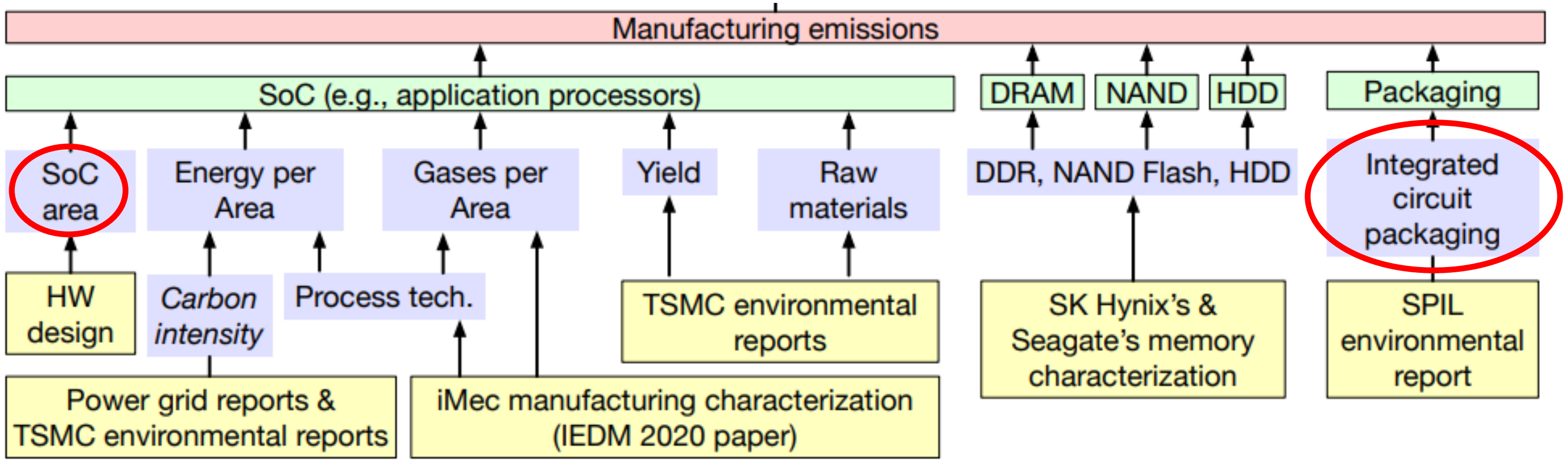
Gupta, Udit, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. "ACT: Designing sustainable computer systems with an architectural carbon modeling tool." In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pp. 784-799. 2022.

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

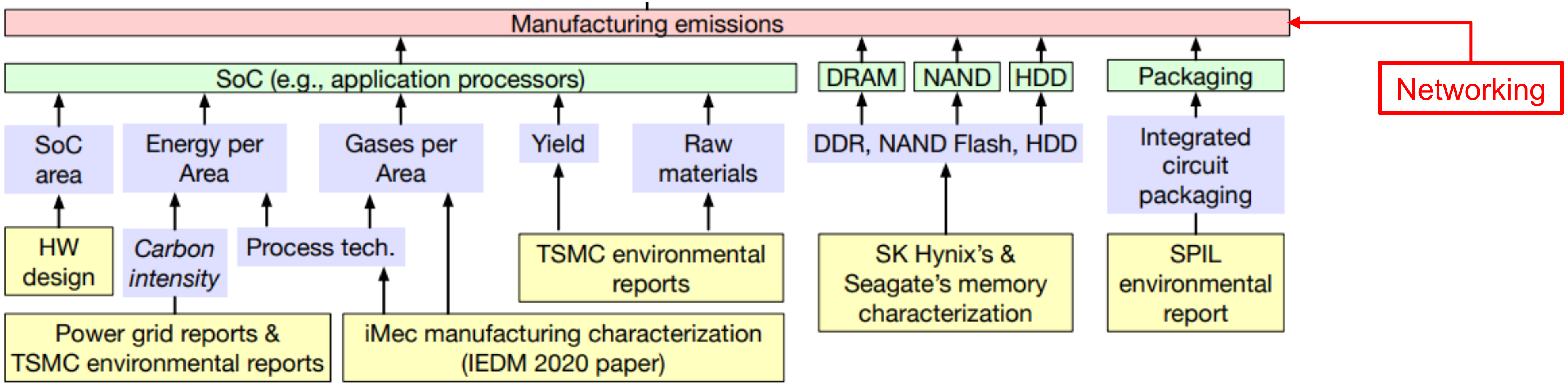
Embodied footprint modeling challenge I

- **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


Manufacturing emissions
SoC (e.g., application processors)
DRAM
NAND
HDD
Packaging
SoC area
Energy per Area
Gases per Area
Yield
Raw materials
DDR, NAND Flash, HDD
Integrated circuit packaging
HW design
Carbon intensity
Process tech.
TSMC environmental reports
SK Hynix's & Seagate's memory characterization
SPIL environmental report
Power grid reports & TSMC environmental reports
iMec manufacturing characterization (IEDM 2020 paper)

Embodied footprint modeling challenge 2

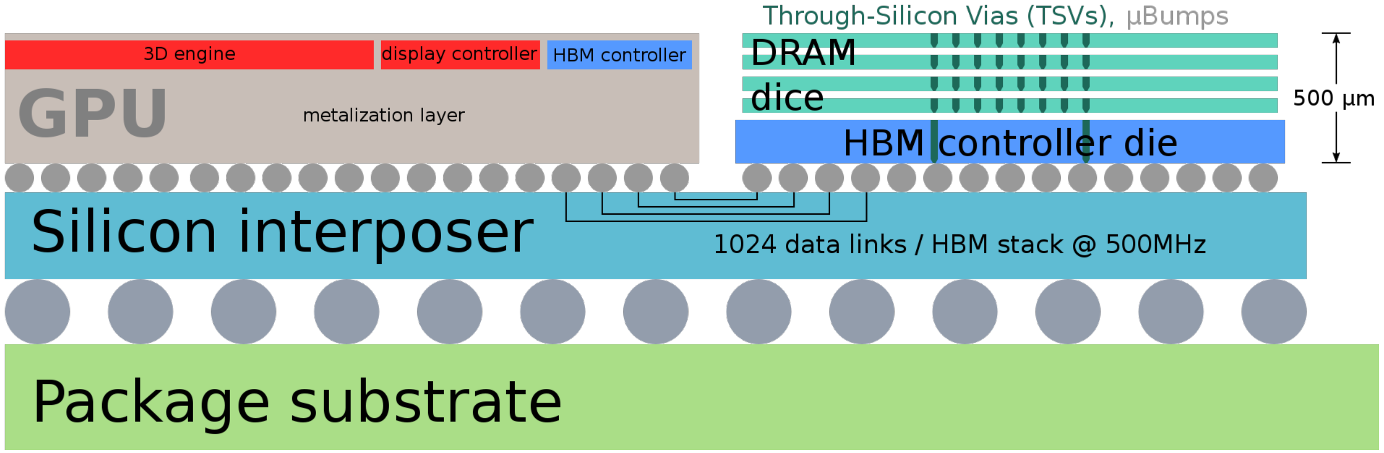
- ACT's model works well for a single device, e.g., desktop, phone
- But **lacks extensibility to large scale distributed systems**
  - For example, the network fabrics for inter-node communication



Manufacturing emissions
SoC (e.g., application processors)
DRAM
NAND
HDD
Packaging
SoC area
Energy per Area
Gases per Area
Yield
Raw materials
DDR, NAND Flash, HDD
Integrated circuit packaging
HW design
Carbon intensity
Process tech.
TSMC environmental reports
SK Hynix's & Seagate's memory characterization
SPIL environmental report
Power grid reports & TSMC environmental reports
iMec manufacturing characterization (IEDM 2020 paper)
Networking

Embodied footprint modeling challenge 3

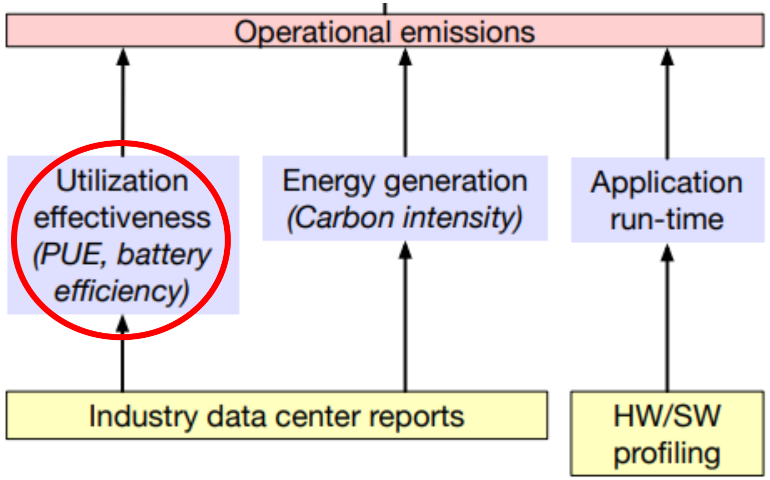
- **Need for GPU-specific features to model GPU-accelerated systems**
  - ACT models GPUs like CPUs – based on the processor’s die area
  - Modern GPUs use FinFET technology compared to traditional CMOS
  - GPUs such as NVIDIA V100 use HBM2 memory that is stacked vertically and integrated into the same 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 I

- **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 emissions

Utilization effectiveness *(PUE, battery efficiency)*

Energy generation *(Carbon intensity)*

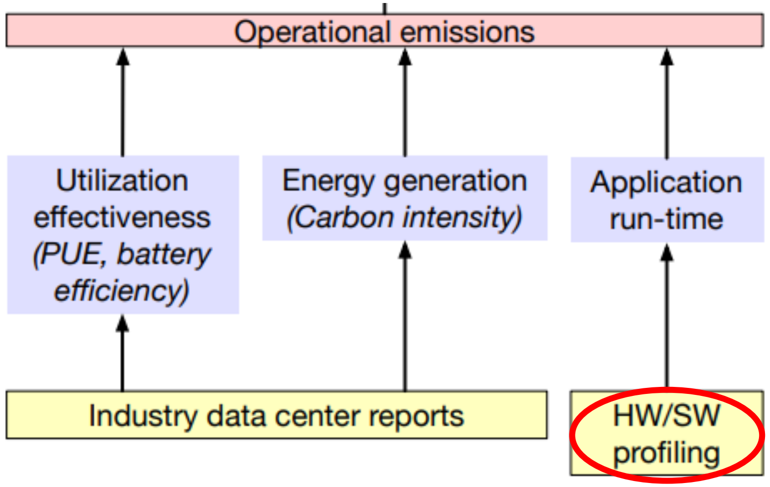
Application run-time

Industry data center reports

HW/SW profiling

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


Operational emissions
Utilization effectiveness *(PUE, battery efficiency)*
Energy generation *(Carbon intensity)*
Application run-time
Industry data center reports
HW/SW profiling

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

Feel free to reach out!
Baolin's email: li.baol@northeastern.edu
Baolin's website: https://baolin-li.netlify.app/

Acknowledgements

Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Thank you!