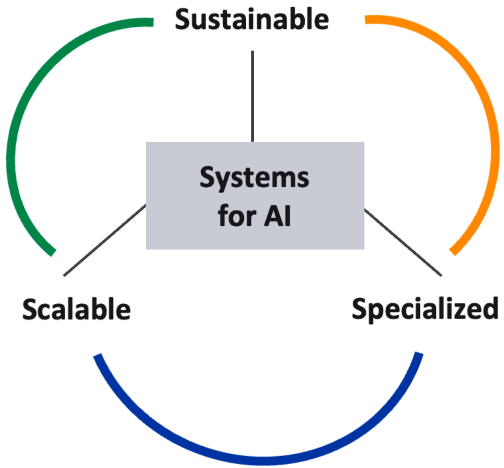
Quantifying the carbon footprint of AI and computing: Past, Present, and Future



Udit Gupta

Assistant Professor, ECE

1.2-2.2 Billion Metric-Tons CO_2

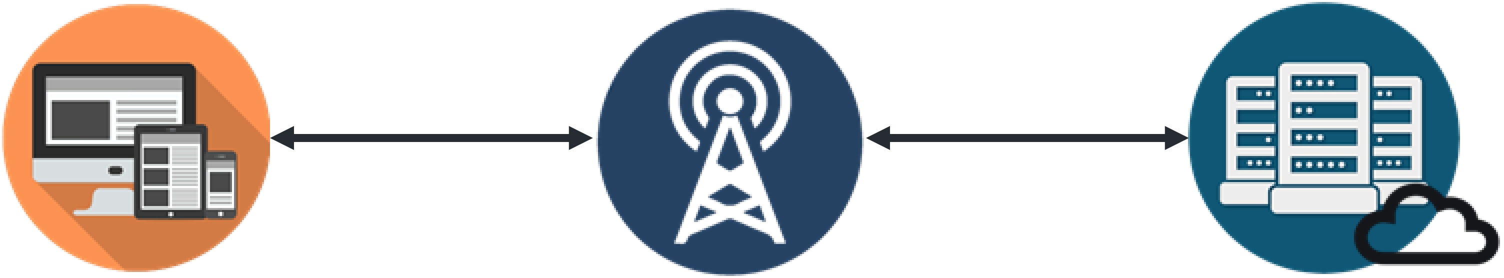
2.1 - 3.9% of worldwide emissions (Freitag'21)

On par with the aviation industry's footprint

Computing's emissions are rising given its growing demand!

1.2-2.2 Billion Metric-Tons CO_2

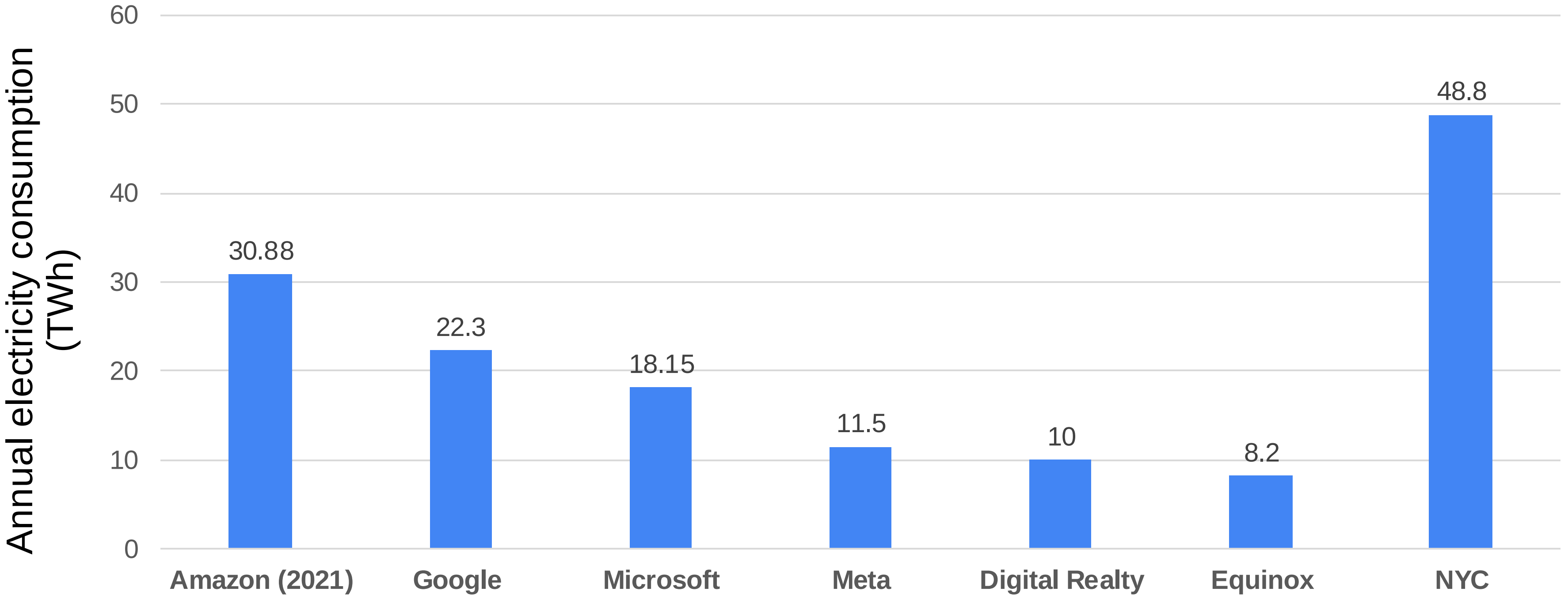
2.1 - 3.9% *of worldwide emissions (Freitag'21)*



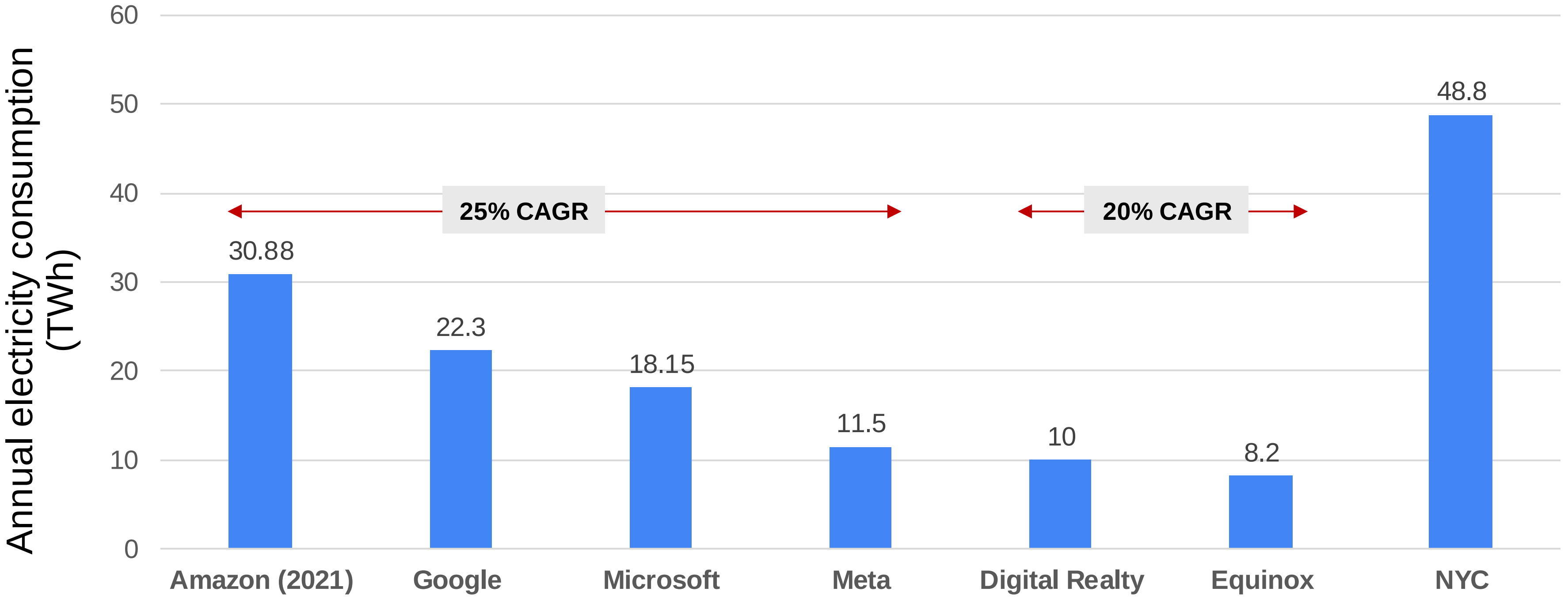
Mobile Communication Data center

On par with the aviation industry's footprint

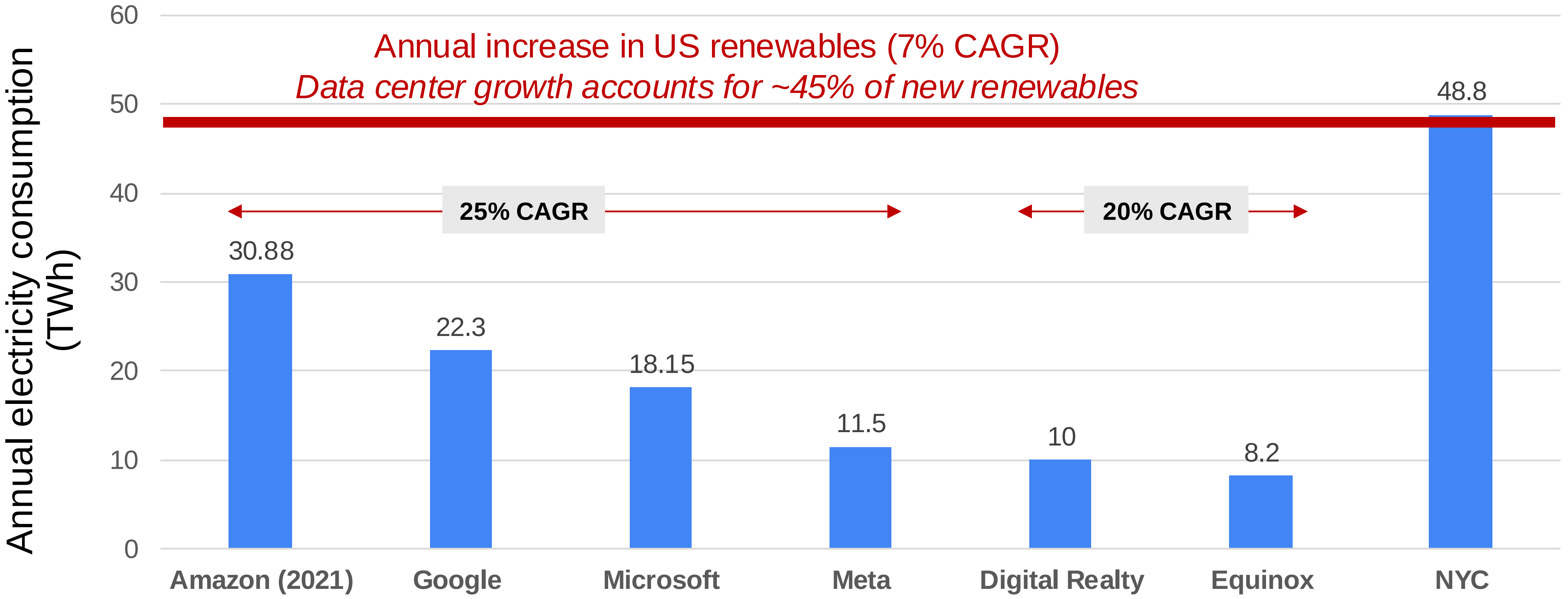
Growing rate of data center energy consumption



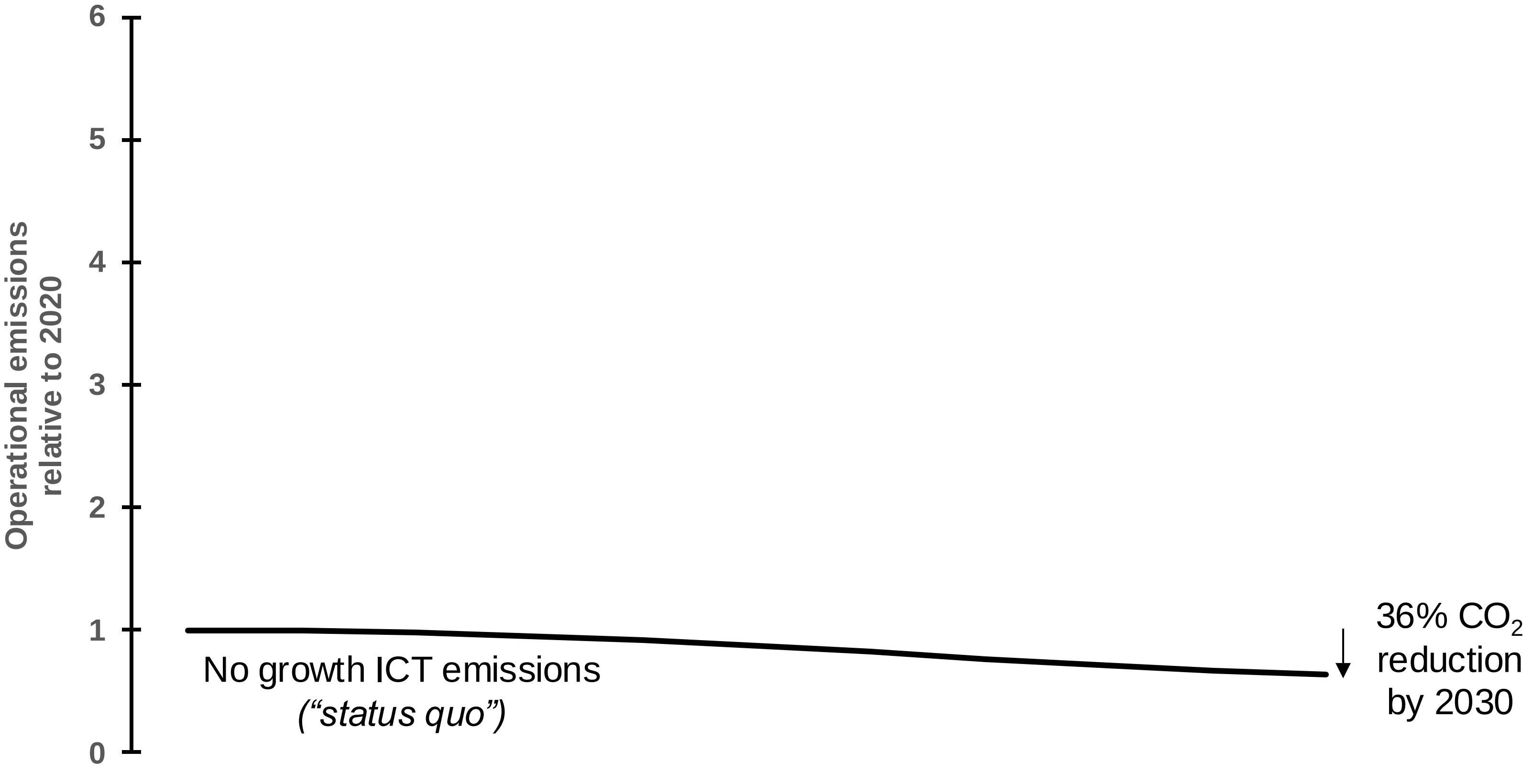
Growing rate of data center energy consumption



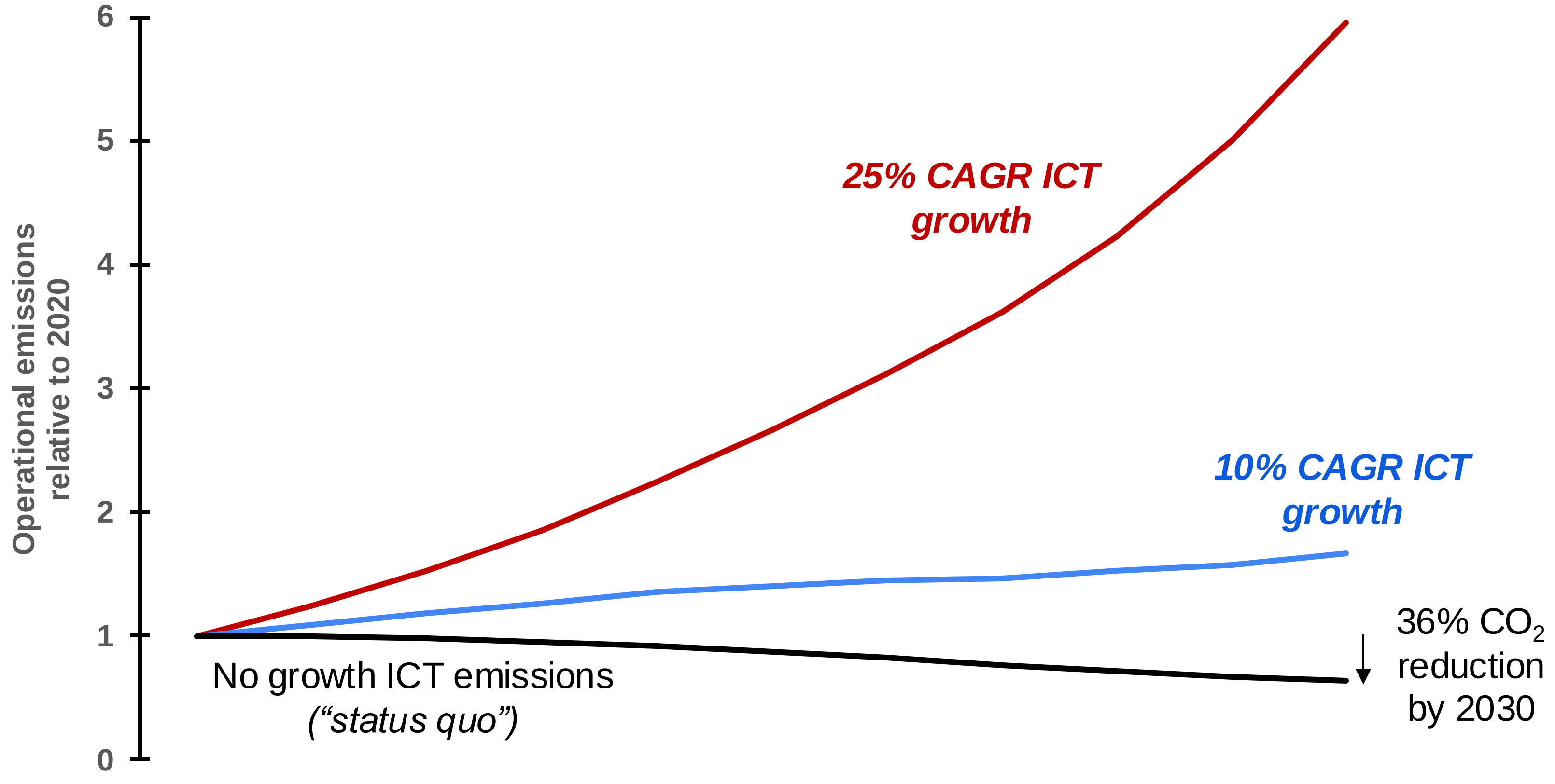
Growing rate of data center energy consumption



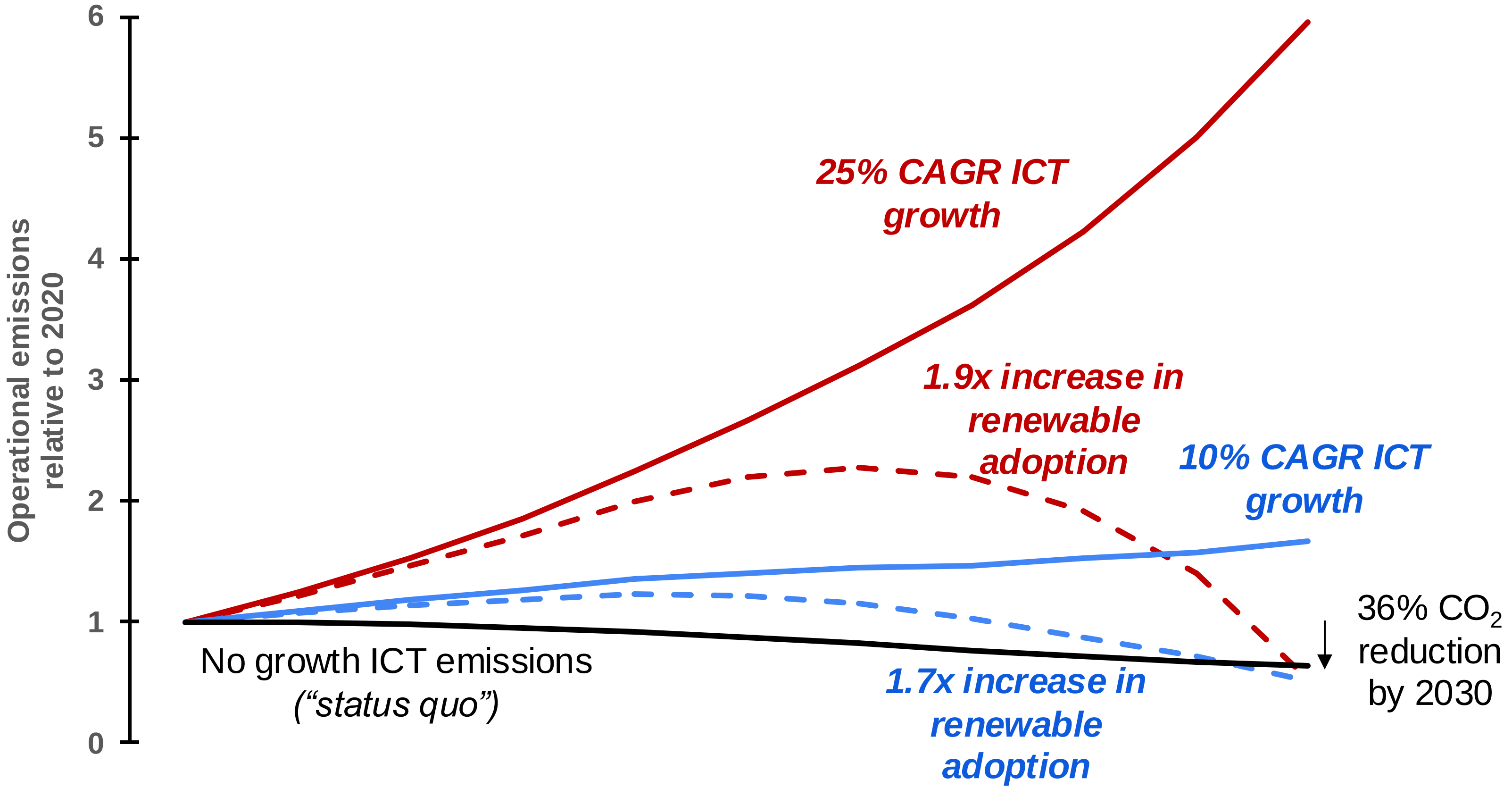
International Telecom. Union targets a 45% reduction in ICT emissions by 2030



International Telecom. Union targets a 45% reduction in ICT emissions by 2030



International Telecom. Union targets a 45% reduction in ICT emissions by 2030



Bloomberg

Live TV | Markets | Economics | Industries | Tech | Politics | Businessweek | Opinion | More

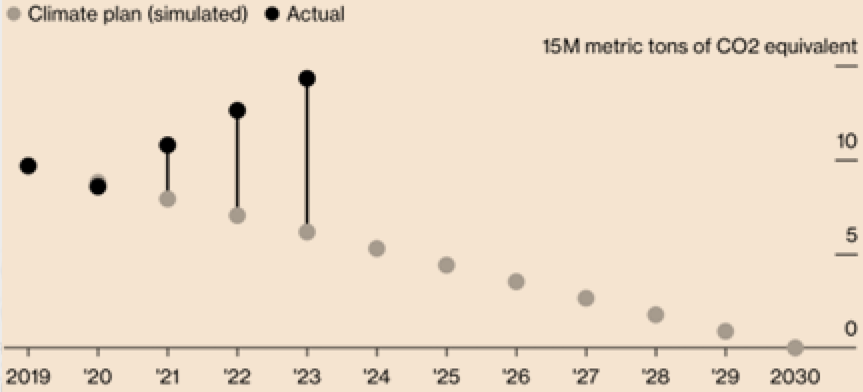
Green
ESG & Investing

Google Is No Longer Claiming to Be Carbon Neutral

The tech giant, which has seen its planet-warming emissions rise because of artificial intelligence, has stopped buying cheap offsets behind the neutrality claim. The company now aims to reach net-zero carbon by 2030.

Google's Emissions

Artificial intelligence is putting the tech giant's climate goals in peril



Source: Google (Scope 1, 2 and 3 data)
Note: Green dots represent linear decline to net-zero emissions goal.

Interface Inc
14.67 ▲+1.24%
Follow

company now aims to reach net-zero carbon emissions by 2030.

The Alphabet Inc. unit has claimed that it's been carbon neutral in its operations since 2007. The status was based on purchasing carbon offsets to match the volume of emissions that were

Bloomberg

Live TV | Markets | Economics | Industries | Tech | Politics | Businessweek | Opinion | More

The AI Race: Why It's So Expensive | Chip Arms Race | Global Energy Strain | DOJ Scrutiny | How Chatbots Work

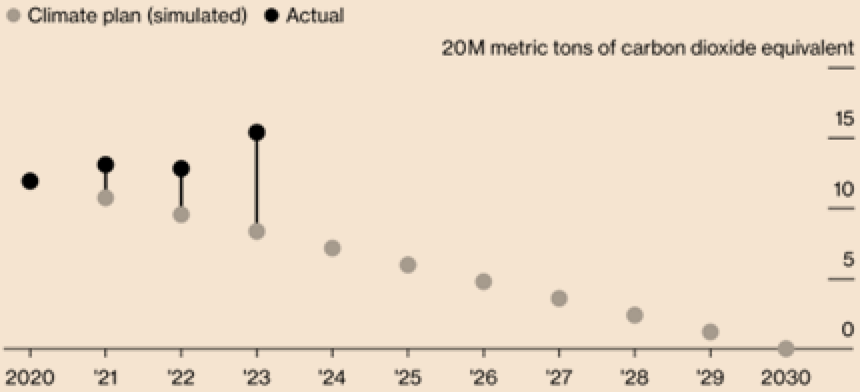
Green
Cleaner Tech

Microsoft's AI Push Imperils Climate Goal as Carbon Emissions Jump 30%

The company's goal to be carbon negative by 2030 is harder to reach, but President Brad Smith says the good AI can do for the world will outweigh its environmental impact.

Microsoft's Emissions

Artificial intelligence is putting the tech giant's climate goals in peril



Source: Microsoft (Scope 1, 2 and 3 "management criteria" data)
Note: Green dots represent linear decline to carbon negative goal.

Follow

Amazon.com Inc
199.29 ▼-0.35%
Follow

ambitious and comprehensive plans to tackle climate change. Now the software giant's relentless push to be the global leader in artificial intelligence is putting that goal in peril.

The Seattle-based company's total planet-warming impact is about

Goal:

Mitigate ICT carbon emissions by co-designing solutions across the stack

Goal:

Mitigate ICT carbon emissions by co-designing solutions across the stack

Economics and policy

Goal:

Mitigate ICT carbon emissions by co-designing solutions across the stack

Economics and policy

Education and workforce development

ECE 6960: Sustainable Computing (Spring 2024)

ECE 6960: Sustainable Computing (Spring 2024)

Description

This graduate level course provides an overview of the holistic environmental impact of computing platforms over the course of their lifetime. Topics include life cycle analyses of computing devices, carbon footprint of computing, computer architecture and systems, renewable energy driven data centers, intermittent computing, sustainable applications (e.g., AI), and emerging technologies. We will understand how to evaluate and consider the holistic environmental impact of computing platforms including carbon, water, e-waste, and materials used. Through reading, analyzing, and discussing papers, and an open-ended project students will develop a holistic understanding of the environmental impact of computing and designing sustainable platforms.

Logistics

- Room: Bloomberg Center 91 (Cornell Tech)
- Time: Tuesdays and Thursdays at 1:25pm - 2:40pm
- Please read the syllabus
- Gradescope link
- Zoom link

Course Staff

20 students
- 6 PhD, 11 Master’s, 3 Undergraduate

Surveyed a range of topics:
- Metrics, materials, tools, embedded devices, data center power and renewable energy integration, and AI

Roughly 10 final projects:
- 2 final projects exploring integration into **startups**
- At least 3 final projects looking to extend into follow-on **research**

Goal:

Mitigate ICT carbon emissions by co-designing solutions across the stack

Economics and policy

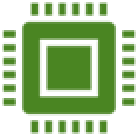
Education and workforce development

Carbon accounting and reporting

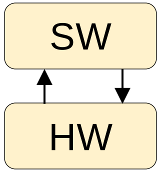
Today: Quantifying the carbon footprint of computing

with insights, opening new research and sustainable development opportunities

Understanding the source of computing's emissions

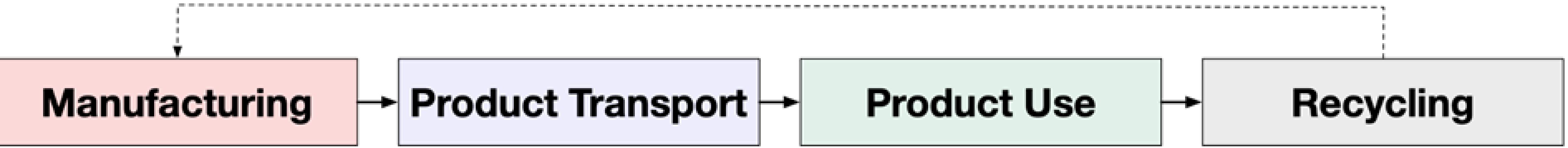


Deep dive: developing computer architectural models to estimate CO_2 emissions

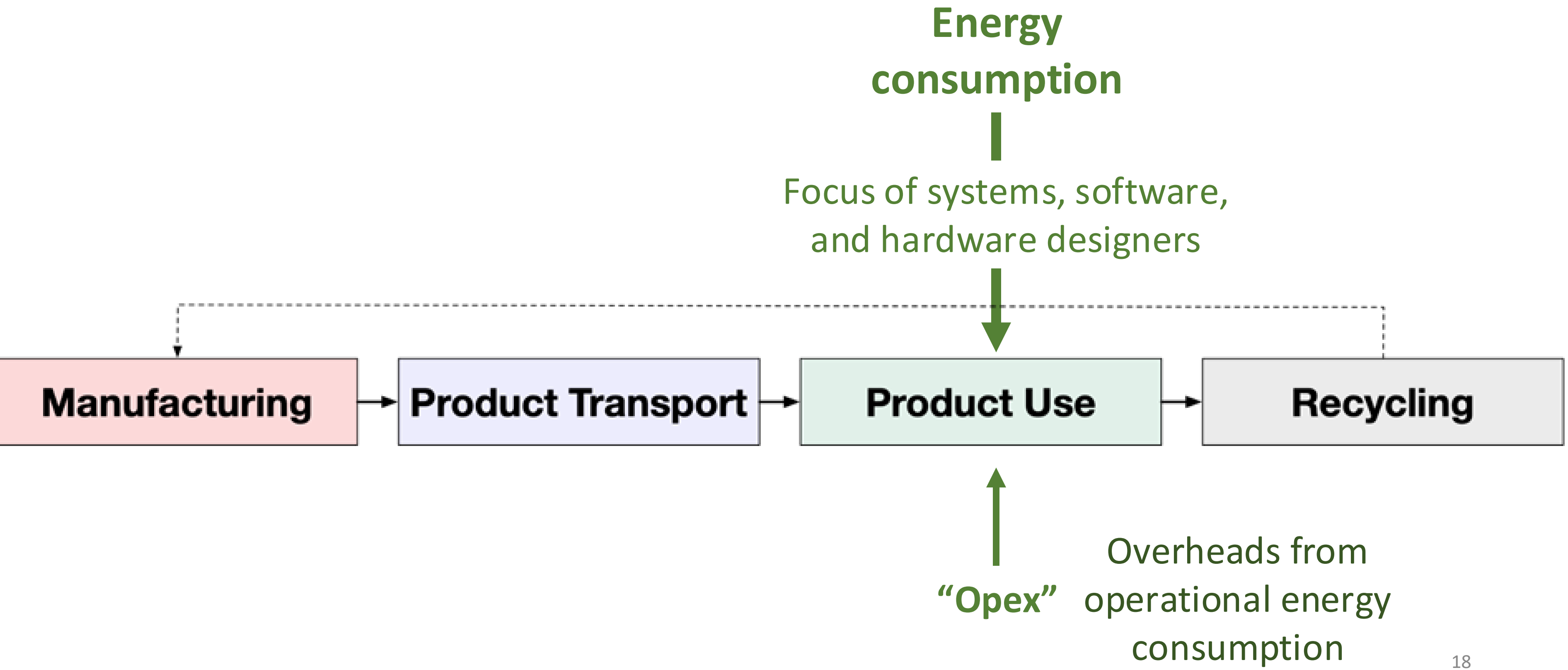


Cross stack: Developing modeling methods across the computing stack

Life **C**ycle **A**nalysis: key to understanding carbon emissions



Life Cycle Analysis: key to understanding carbon emissions



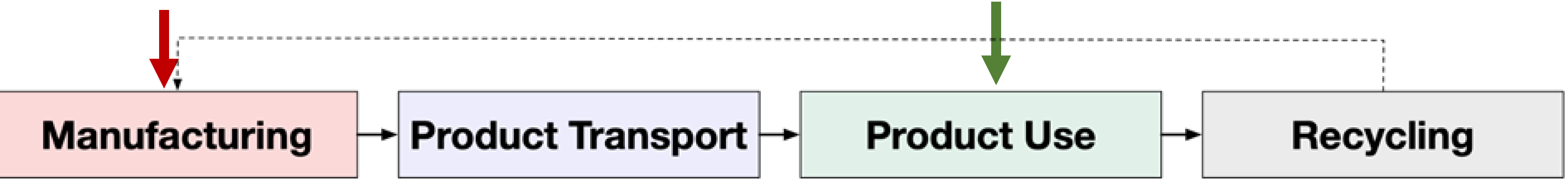
Life Cycle Analysis: key to understanding carbon emissions

Hardware manufacturing

Emissions from fabs building chips

Energy consumption

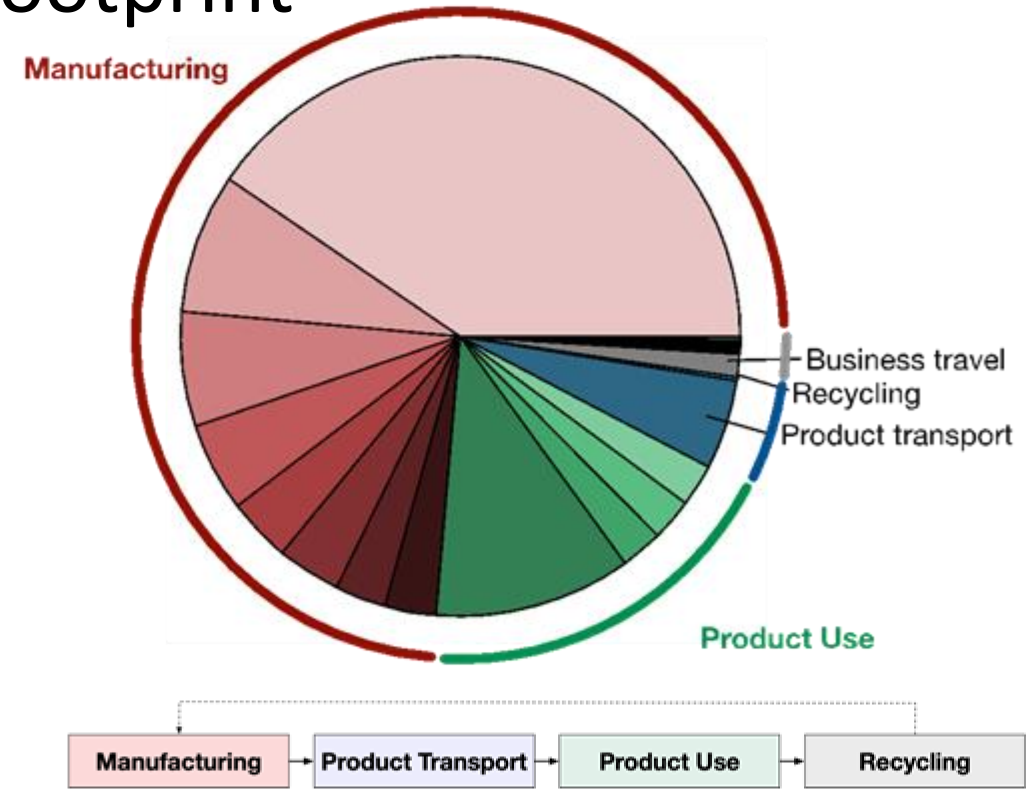
Focus of systems, software, and hardware designers



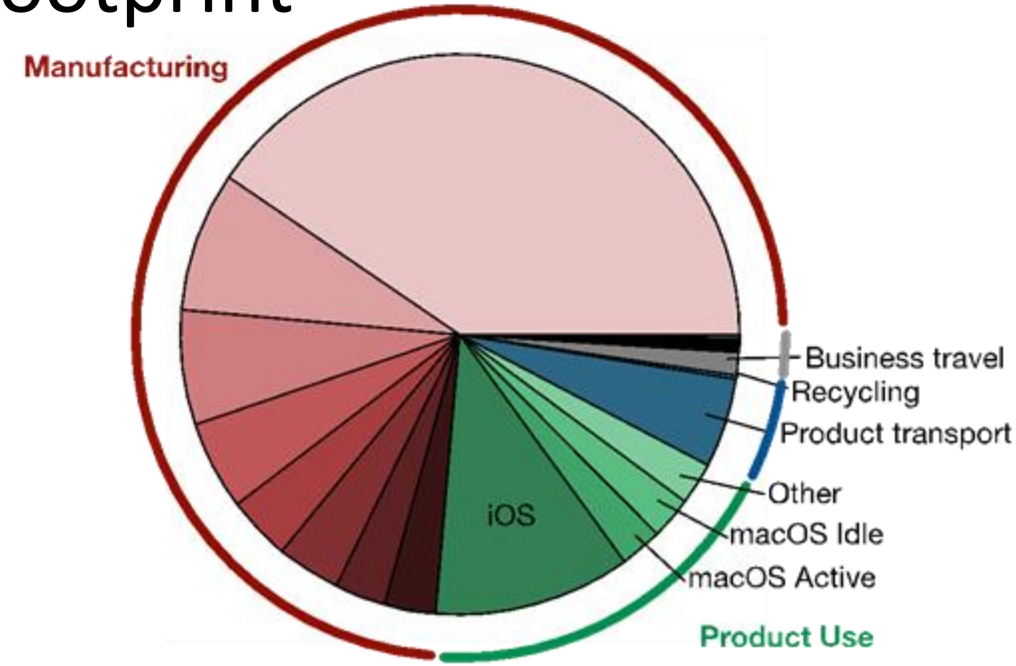
“Capex” Overheads from infrastructure related activities

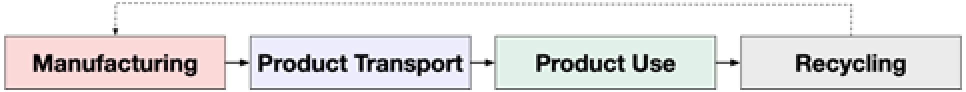
“Opex” Overheads from operational energy consumption

Manufacturing dominates Apple’s overall carbon footprint



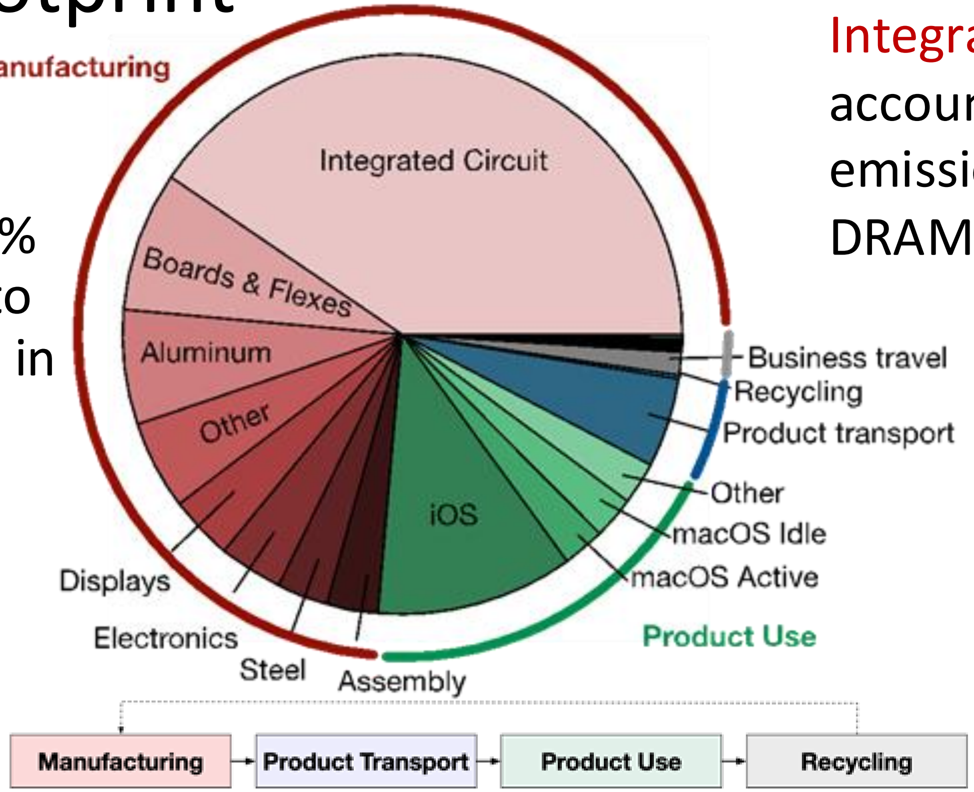
Manufacturing dominates Apple's overall carbon footprint





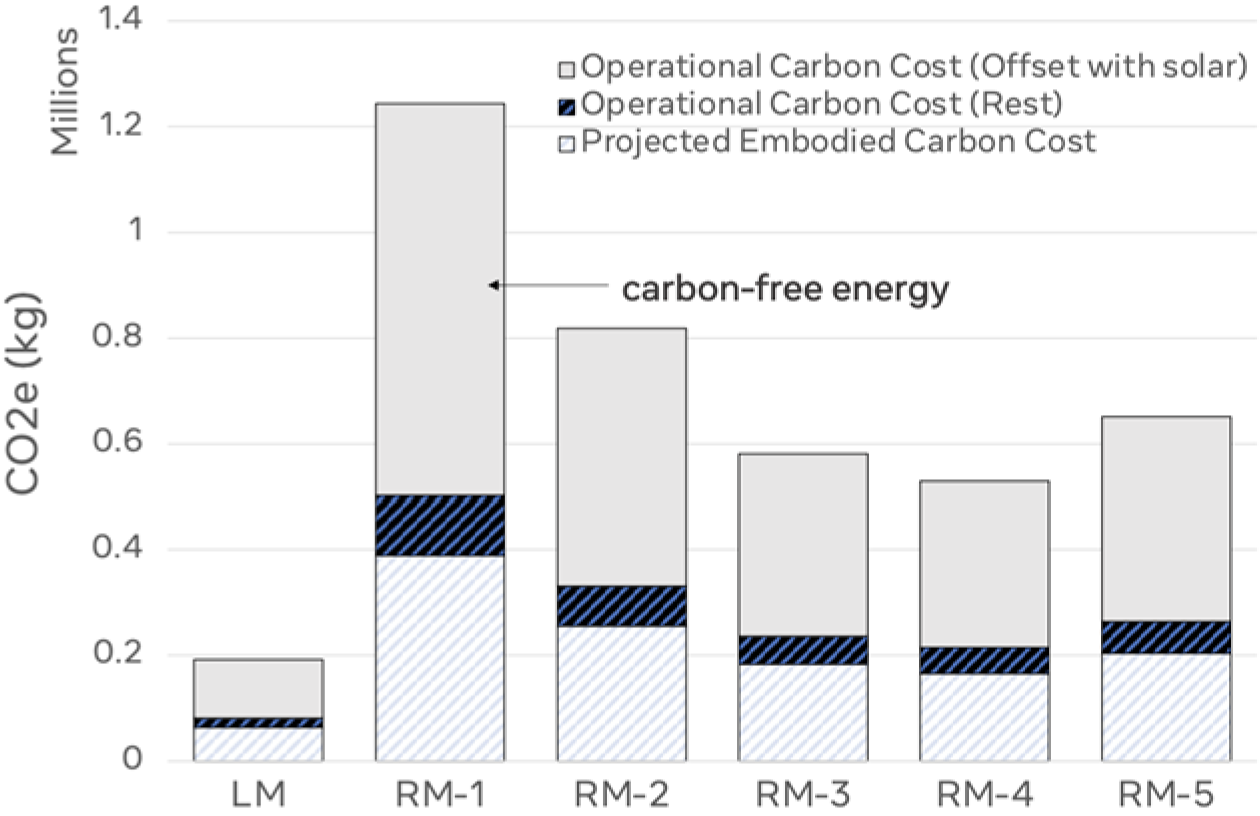
Manufacturing dominates Apple's overall carbon footprint

Manufacturing accounts for 74% of Apple's end to end breakdown in 2019

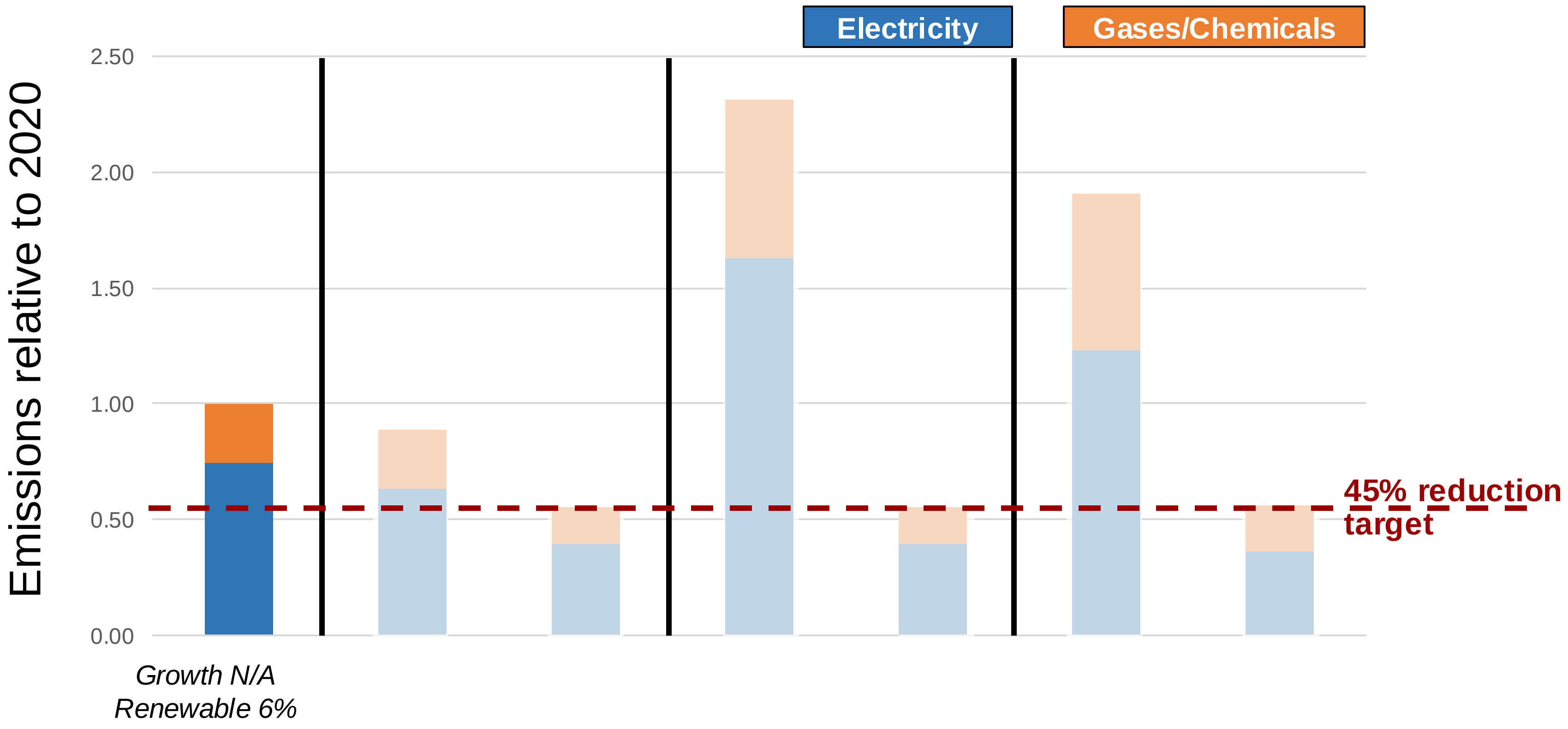


Integrated circuits account for 33% of emissions (SoCs, DRAMs, NAND Flash)

Crucial to look at emissions across HW cycle

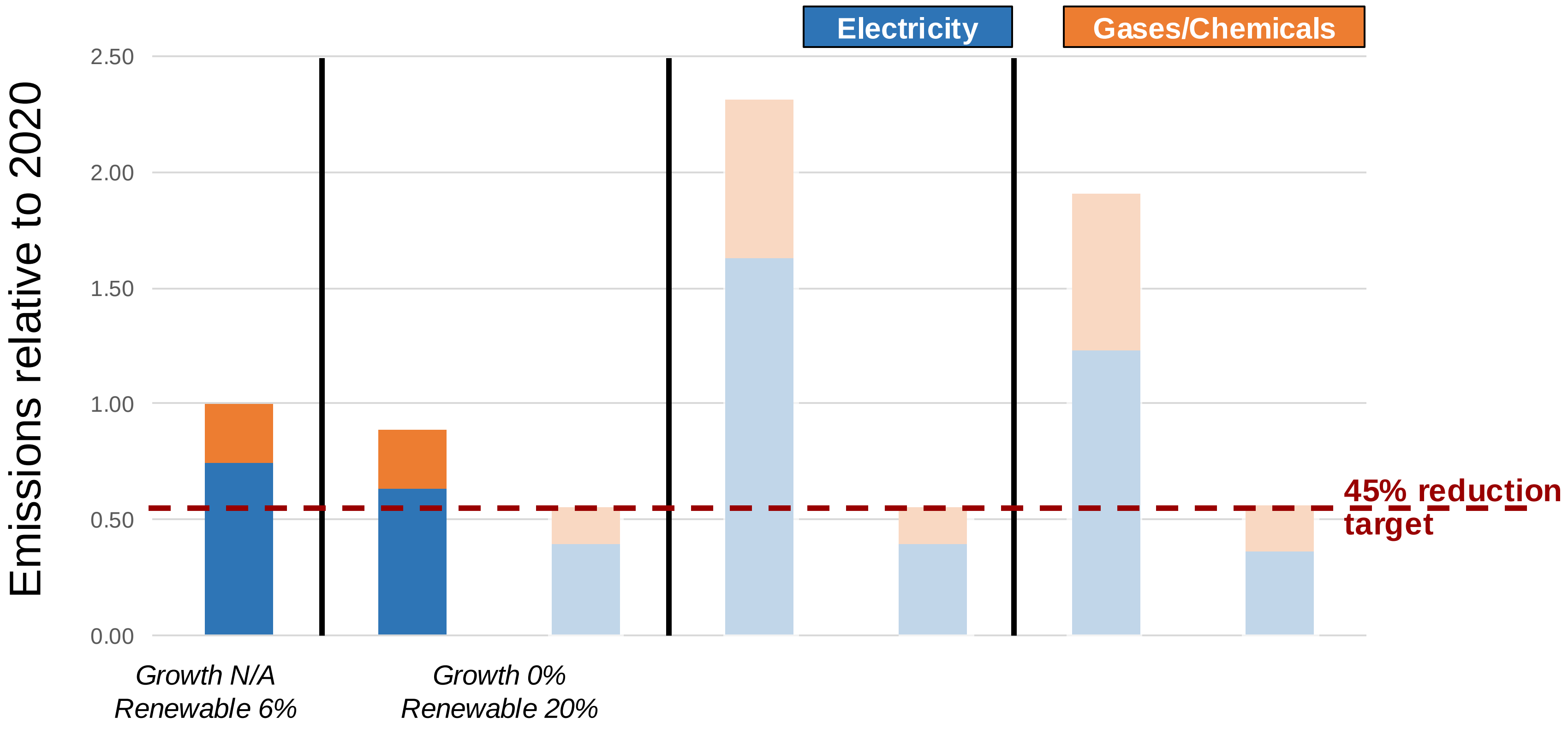


Chip manufacturing requires cross-stack optimization



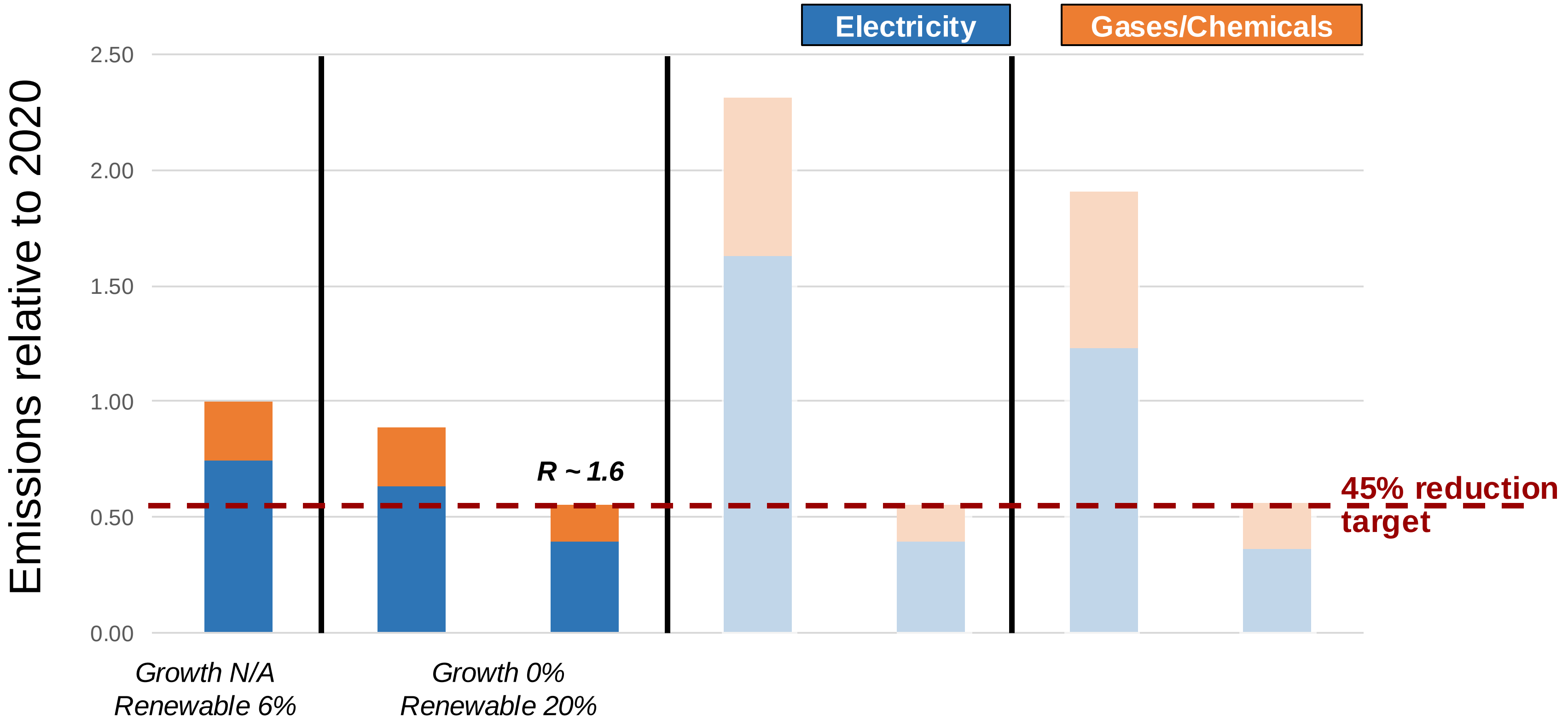
Source: 2021 corporate sustainability reports
“Carbon Connect: An Ecosystem for Sustainable Computing” Lee et. Al. (arxiv 2024)

Chip manufacturing requires cross-stack optimization



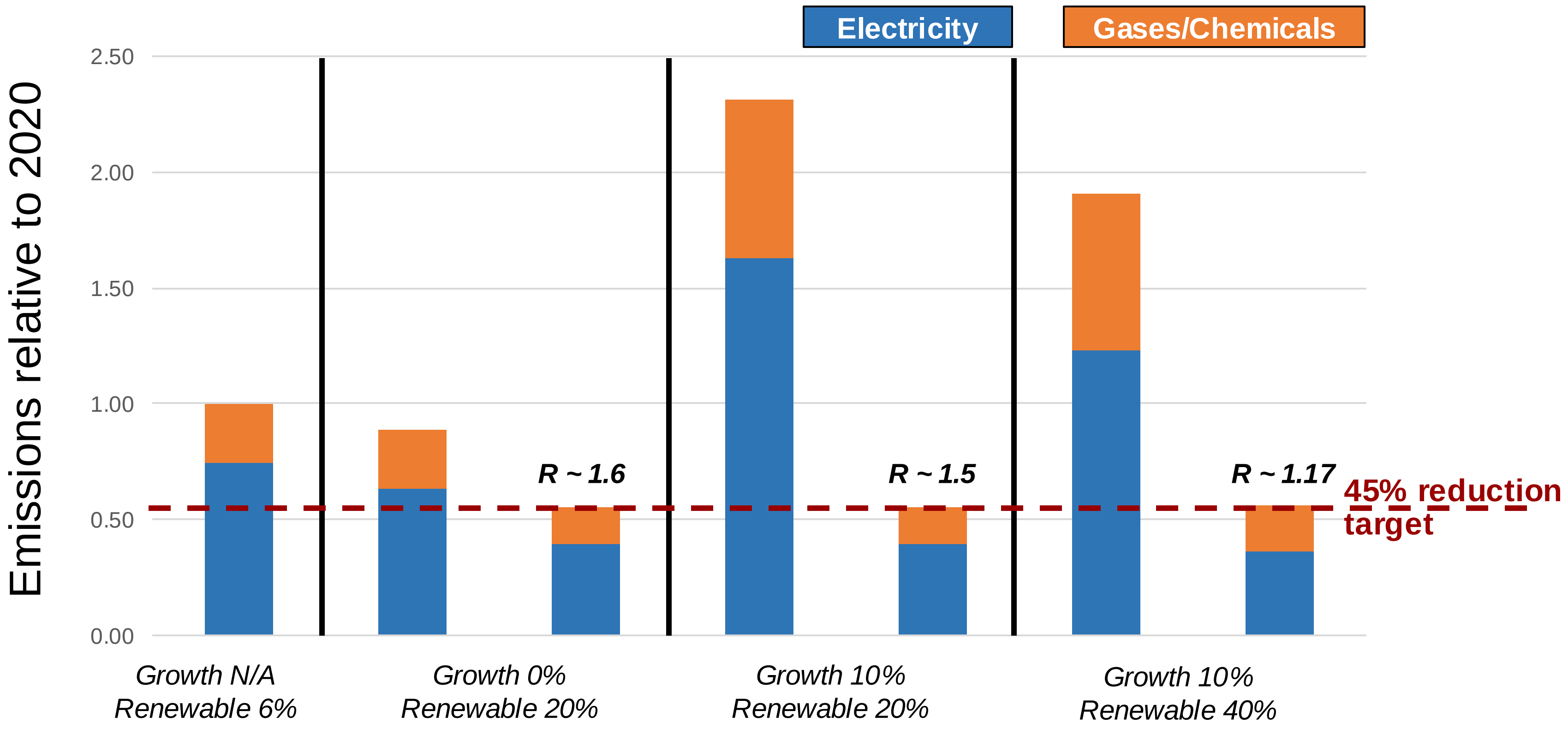
Source: 2021 corporate sustainability reports
“Carbon Connect: An Ecosystem for Sustainable Computing” Lee et. Al. (arxiv 2024)

Chip manufacturing requires cross-stack optimization



Source: 2021 corporate sustainability reports
“Carbon Connect: An Ecosystem for Sustainable Computing” Lee et. Al. (arxiv 2024)

Chip manufacturing requires cross-stack optimization



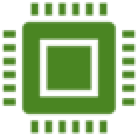
Source: 2021 corporate sustainability reports

"Carbon Connect: An Ecosystem for Sustainable Computing" Lee et. Al. (arxiv 2024)

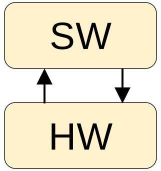
Today: Quantifying the carbon footprint of computing

with insights, opening new research and sustainable development opportunities

Understanding the source of computing's emissions



Deep dive: developing computer architectural models to estimate CO_2 emissions



Cross stack: Developing modeling methods across the computing stack

Current carbon accounting methodologies

Economic Input/Output (EIO)

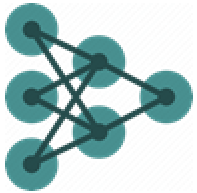
Carbon is tied directly to economic cost which is susceptible to market effects.

Life cycle analysis (LCA)

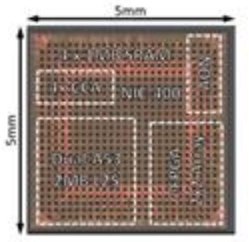
Current databases are out-of-date (45nm or older nodes).

LCA's take high $$ and time to conduct.

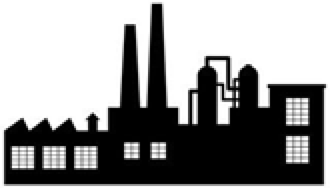
Architectural Carbon Modeling Tools (ACT)



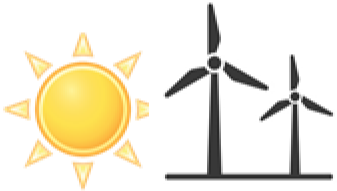
SW design & performance



HW design & power



Fab

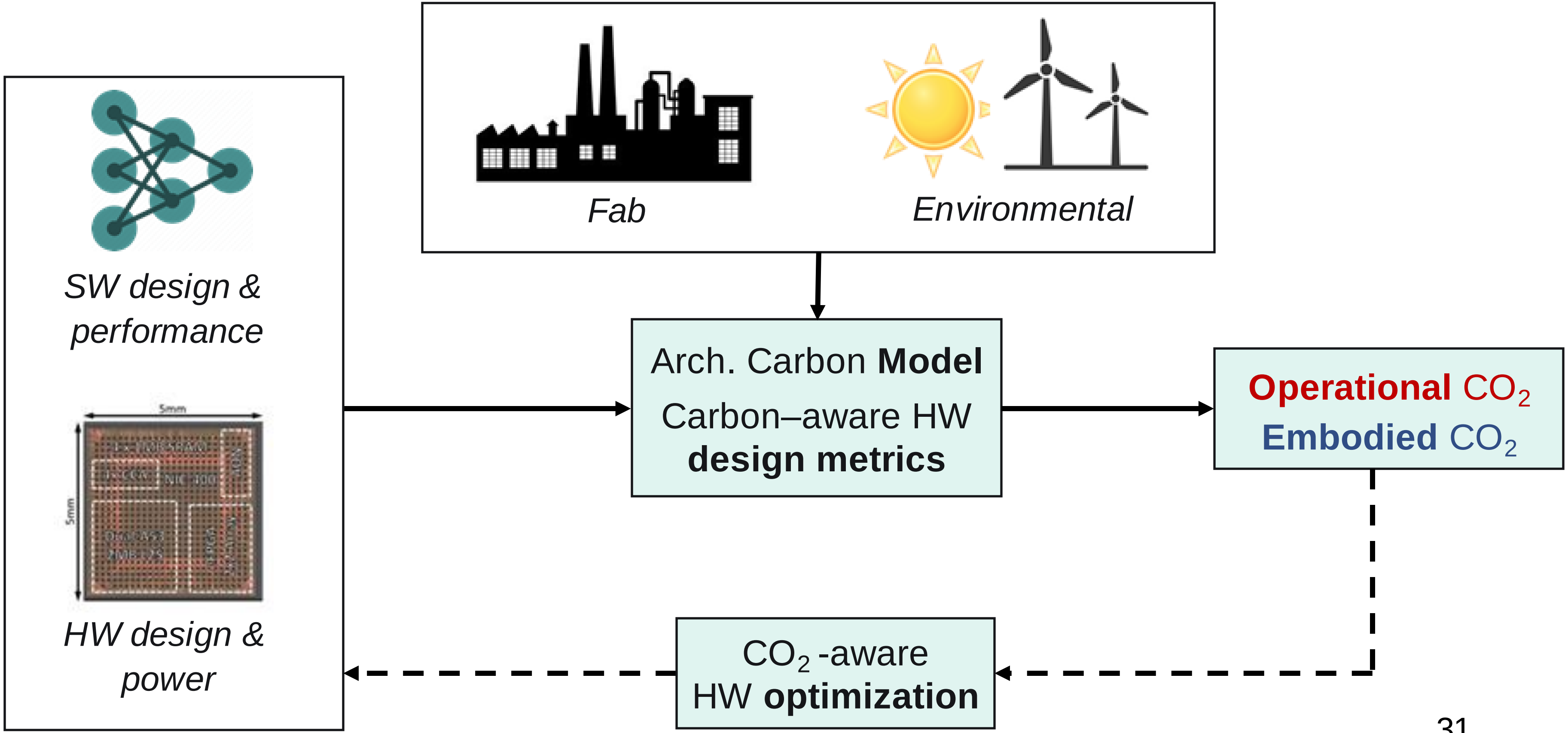


Environmental

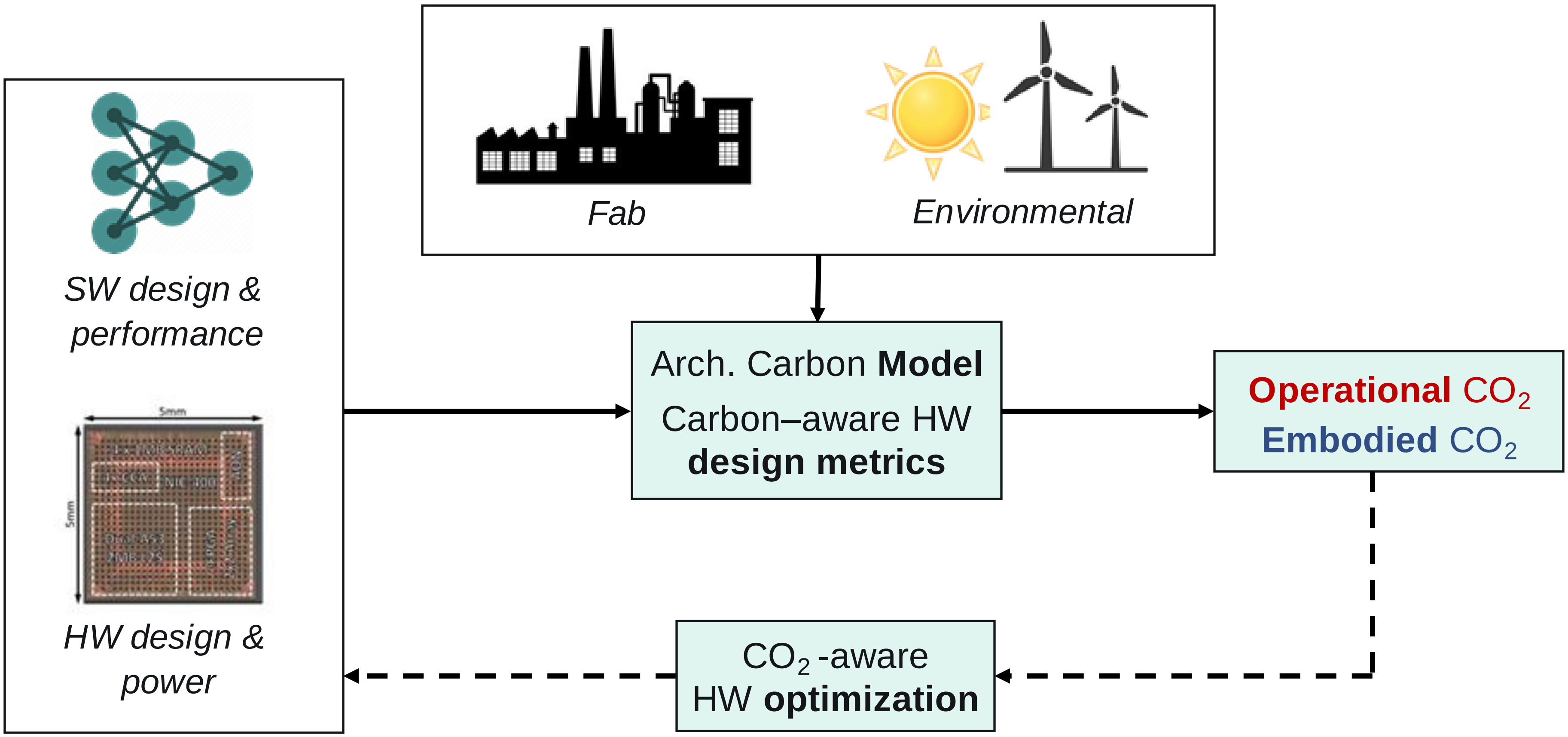
Arch. Carbon **Model**

Carbon–aware HW **design metrics**

Architectural Carbon Modeling Tools (ACT)



Architectural Carbon Modeling Tools (ACT)



Architectural Carbon Model

Model	Hardware/software input
$Carbon = OP_{CF} + \frac{Runtime}{Lifetime} Emb_{CF}$	**Performance/power/energy** and **lifetime** of hardware
$OP_{CF} = CI_{use} \times Energy$	**Energy efficiency** and environment (**carbon intensity**)
$Emb_{CF} = Packaging + \sum_{r}^{SoC,Memory,Storage} Emb_r$	Overhead of **hardware manufacturing**

Embodied carbon of application processors (SoC’s)

Energy source (grid)

Device and fab characterization (industry fabs, device data)

$$Emb_{SoC} = Area \times \frac{\left(CI_{fab} \times Fab_{energy}\right) + Fab_{chemicals} + Fab_{materials}}{Yield}$$

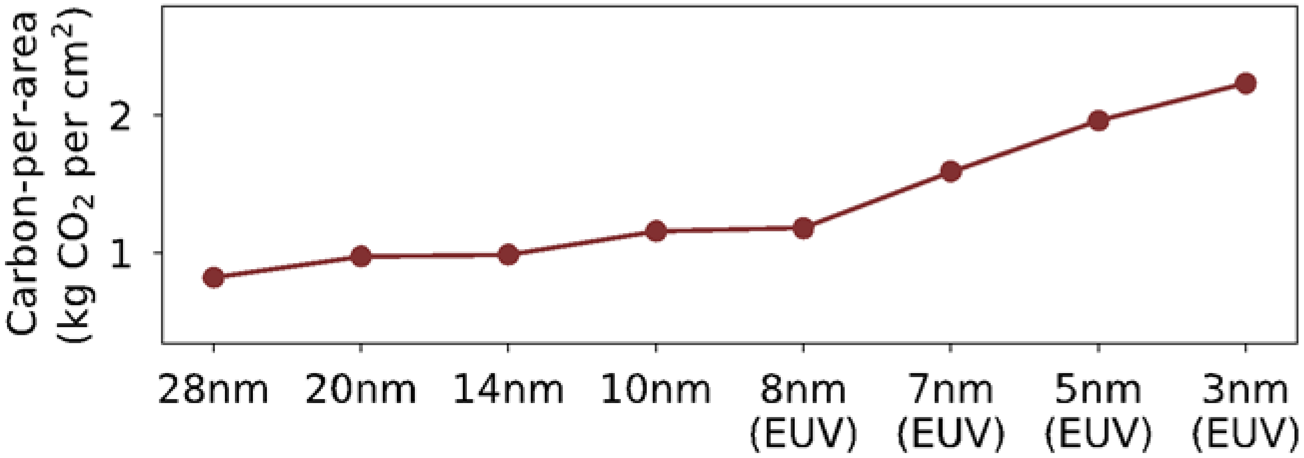
Hardware design

Embodied carbon of application processors (SoC's)

$$Emb_{SoC} = Area \times \mathbf{CPA}$$

Embodied carbon of application processors (SoC's)

$$Emb_{SoC} = Area \times \mathbf{CPA}$$

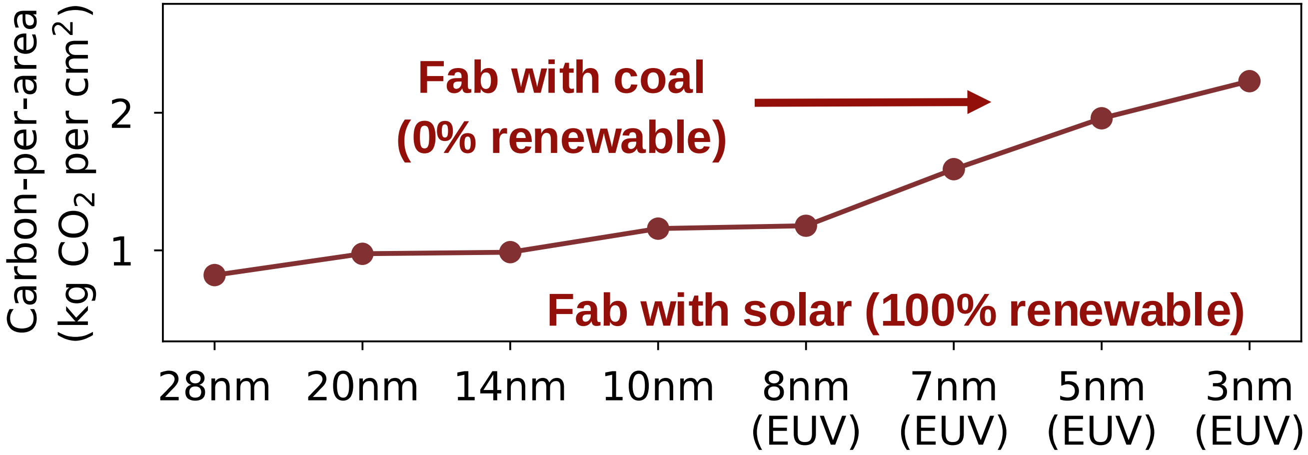


Data sources:

- *[IMEC] DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies. Bardon et. al (IEDM 2020)*
- *[TSMC] TSMC Sustainability Reports 2018-2020*

Embodied carbon of application processors (SoC's)

$$Emb_{SoC} = Area \times \mathbf{CPA}$$

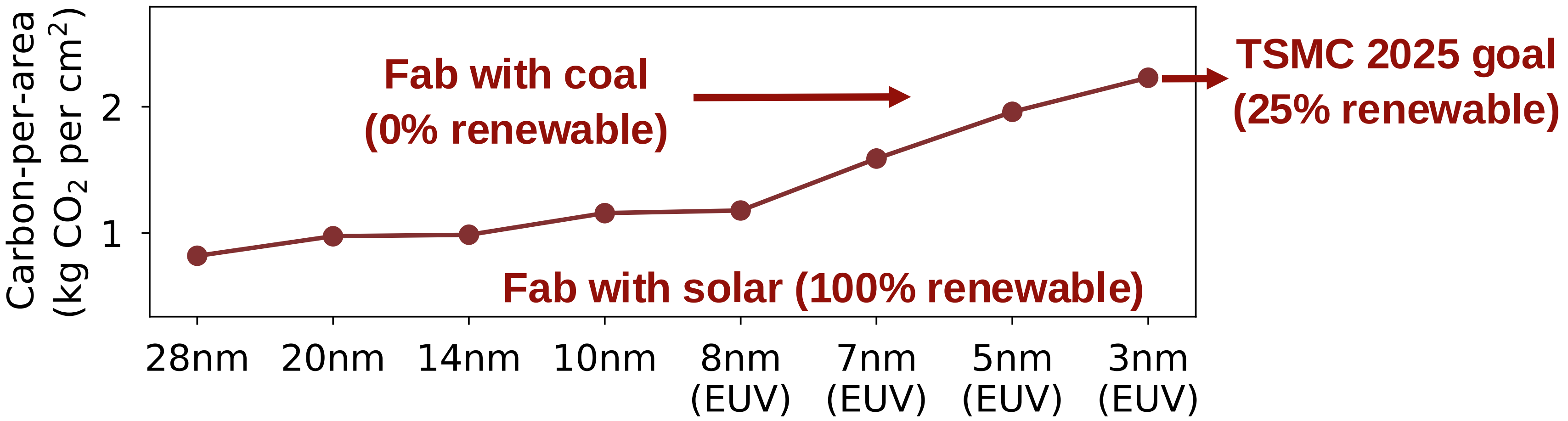


Data sources:

- *[IMEC] DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies. Bardon et. al (IEDM 2020)*
- *[TSMC] TSMC Sustainability Reports 2018-2020*

Embodied carbon of application processors (SoC's)

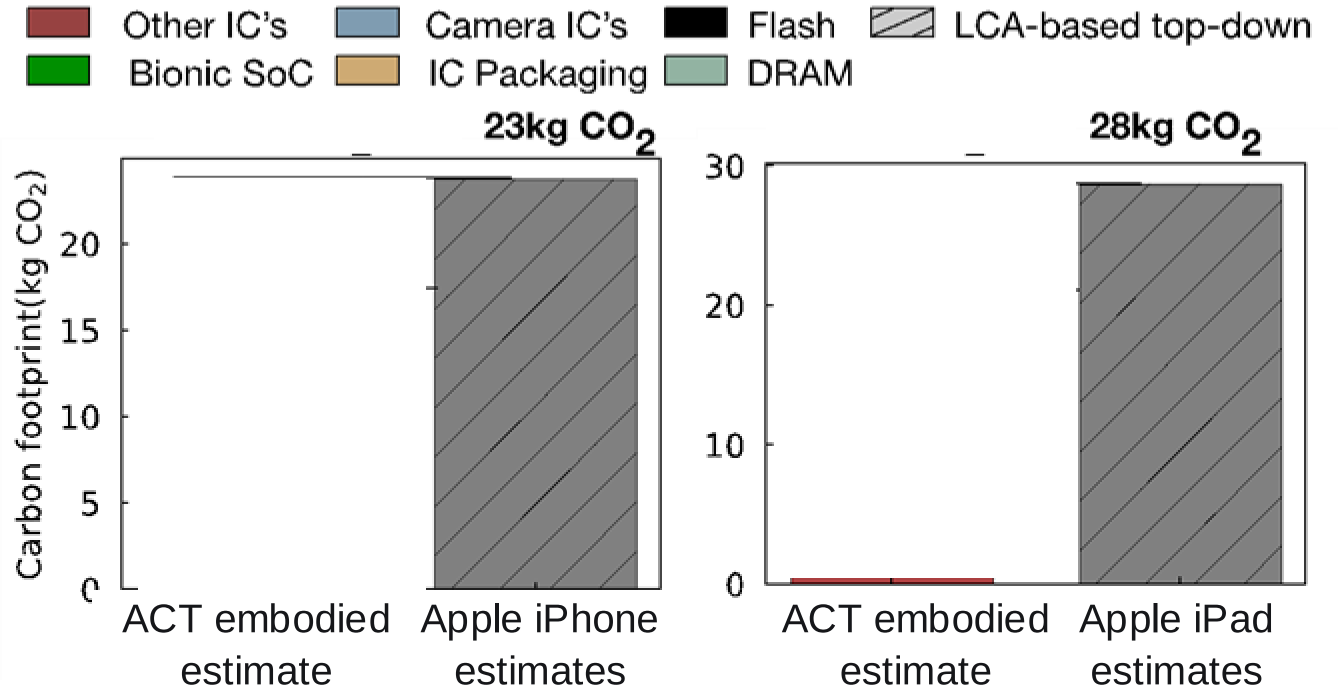
$$Emb_{SoC} = Area \times \boldsymbol{CPA}$$



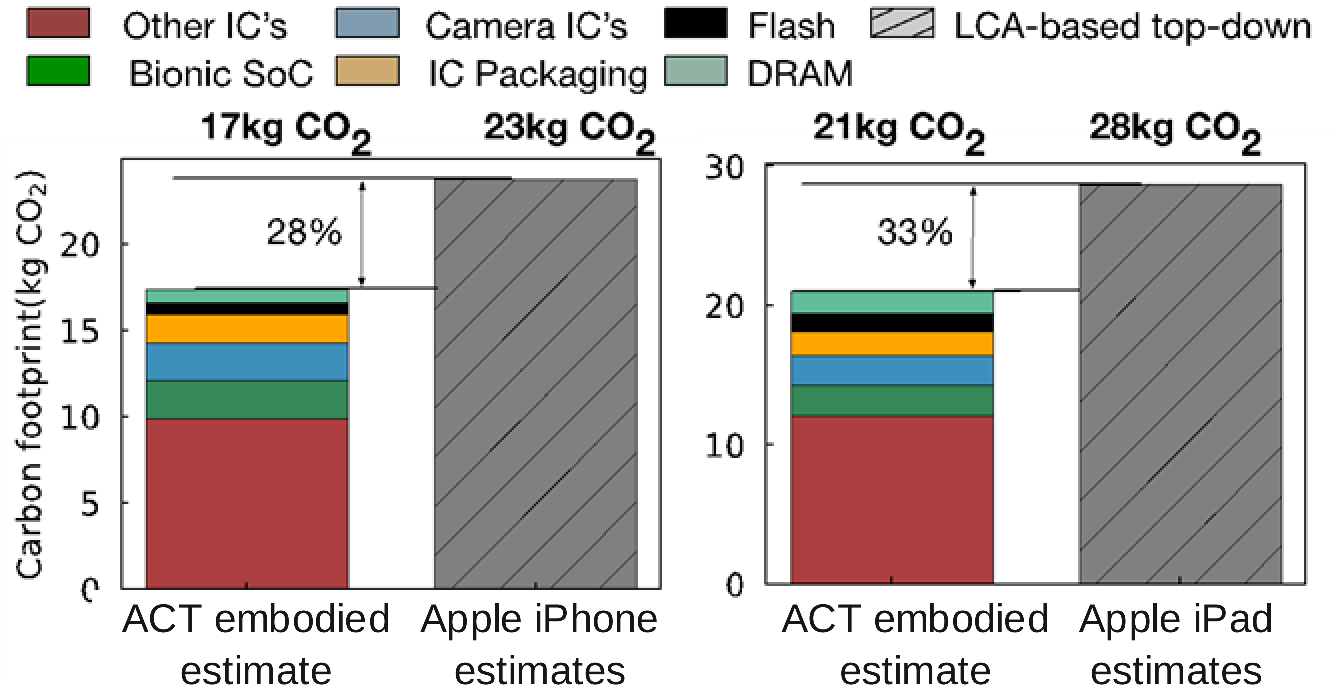
Data sources:

- *[IMEC] DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies. Bardon et. al (IEDM 2020)*
- *[TSMC] TSMC Sustainability Reports 2018-2020*

Comparing ACT with Apple's product environmental reports



Comparing ACT with Apple's product environmental reports



Setting the standard for data center sustainability

Understanding the life cycle impact of data center components

We cannot reduce what we do not measure. In 2022, we conducted Life Cycle Assessments (LCAs) on several data center hardware products and developed internal visualization tools to identify the highest carbon emitting components of each product.

At the data center fleet level, the Sustainability, Physical Modeling, and Meta AI Systems and Machine Learning teams have partnered on a large-scale project to develop and scale a dataset containing the best available embodied carbon estimates at the scale of the hundreds of millions of components in our data center hardware.

In 2022, the teams reached more than 90% coverage, meaning there is primary data, an LCA, or a modeled↗ value assigned to each asset. This dataset lays [illegible] carbon reductions by helping us use less or choose low-carbon options, engage suppliers, and drive value chain and system-level interventions in line with Meta's net zero strategy.

Tenets of Environmental Design

Reduce

Design leaner footprint software and hardware.

Recycle

Recover discarded systems and components.

Reuse

Repurpose systems already produce.

Tenets of Environmental Design

Reduce

Design leaner footprint software and hardware.

Recycle

Recover discarded systems and components.

Reuse

Repurpose systems already produce.

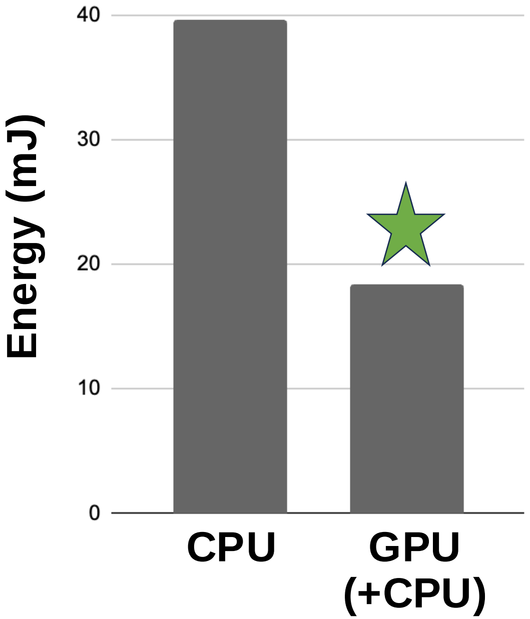
Reuse: General purpose versus custom mobile HW

Commodity (mobile CPU)
Specialized (mobile GPU)

→ **Architectural Carbon Model** → Carbon per inference

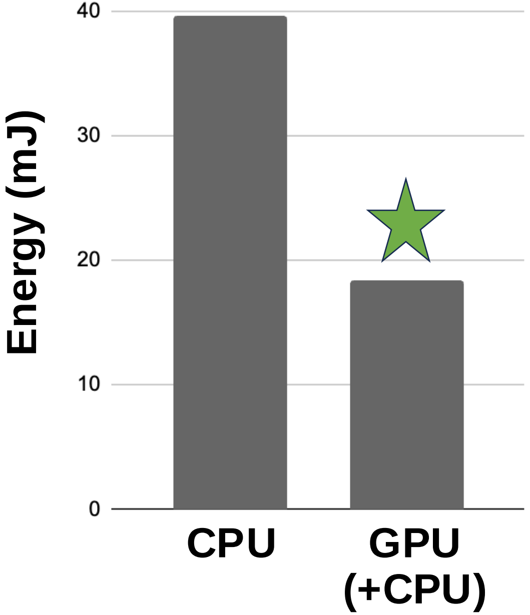
Reuse: General purpose versus custom mobile HW

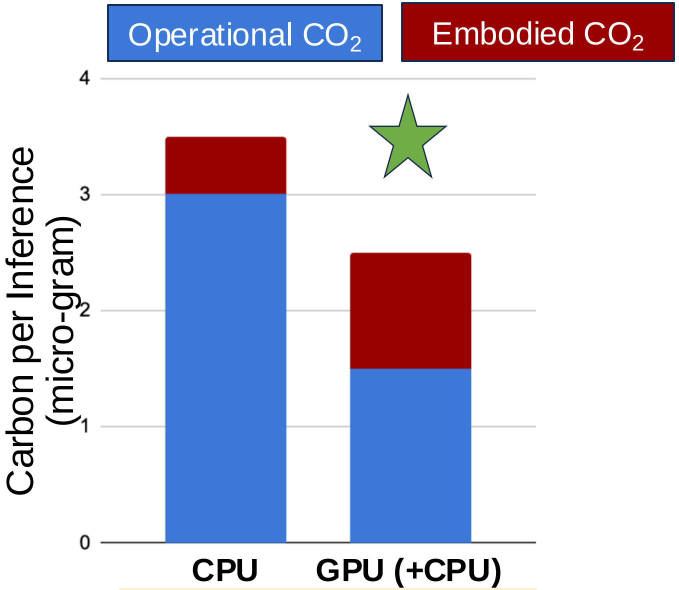
Commodity (mobile CPU)
Specialized (mobile GPU)
→ **Architectural Carbon Model** → Carbon per inference



Reuse: General purpose versus custom mobile HW

Commodity (mobile CPU)
Specialized (mobile GPU)
→ **Architectural Carbon Model** → Carbon per inference

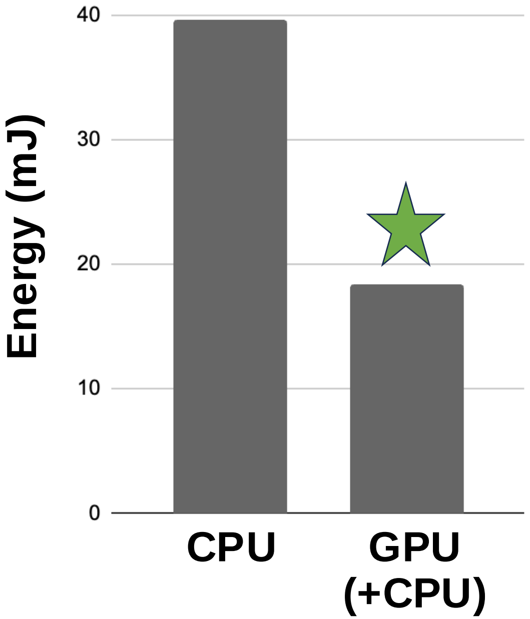


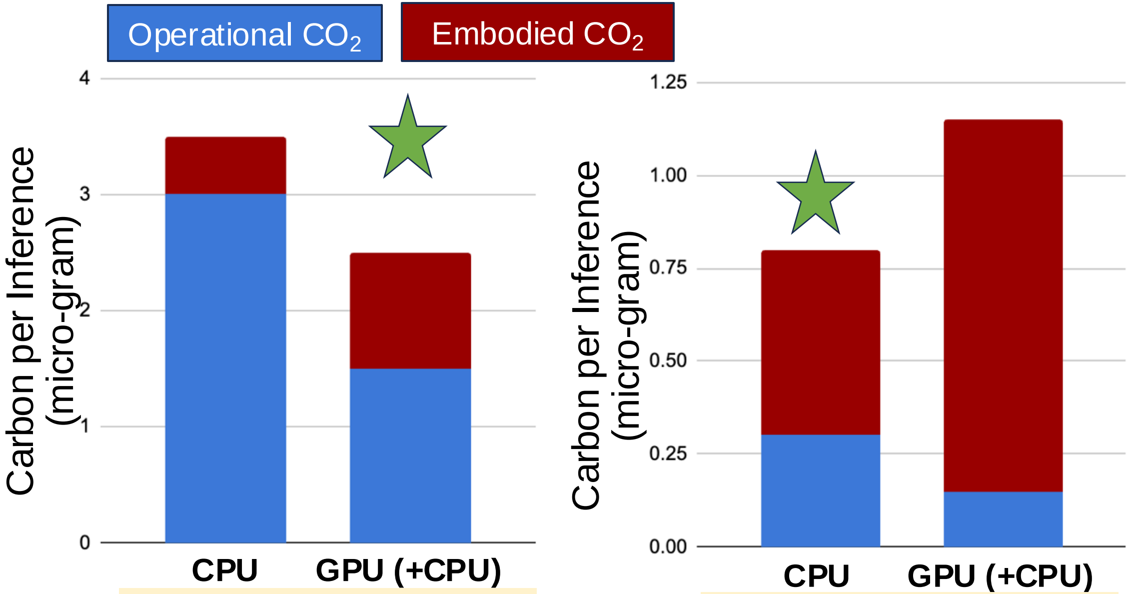


High utilization & carbon intensity

Reuse: General purpose versus custom mobile HW

Commodity (mobile CPU)
Specialized (mobile GPU)
→ **Architectural Carbon Model** → Carbon per inference





High utilization & carbon intensity

Low utilization & carbon intensity

ACT: Architectural Carbon Modeling Tool

ACT An Architectural Carbon Modeling Tool for Designing Sustainable Computer Systems

View on GitHub

ACT: Architectural Carbon Modeling Tool

Abstract

Motivation: Over the past two decades, the world has witnessed a dramatic rise in computing across data centers, mobile, and communication technologies. As of 2015 information computing technology (ICT) accounts for up to 3% of global carbon emissions. Unfortunately, as the demand for computing grows with new applications and platforms so will energy demand. Many technology companies, including Microsoft, Google, Facebook, and Amazon, have pledged to reduce their carbon footprints over the next decade. Meeting these pledges and enabling sustainable computing requires immediate action from the systems and architecture community.

Background: In addition to its importance, enabling environmentally sustainable computing introduces unique challenges to system and hardware design. First, carbon emissions are shifting from being dominated by operational energy consumption to hardware manufacturing. Traditionally the majority of emissions came from operational hardware use (i.e., energy consumption). However, given the energy efficiency optimizations and increasing fabrication complexity, the majority of carbon emissions have shifted to hardware manufacturing. Given these new challenges, enabling environmentally sustainable computing demands distinct solutions across the computing stack, hardware life cycles, and end-to-end systems.

Come join us at MICRO 2024 for the next ACT tutorial!

"Inherent Uncertainty: **Uncertainty is inherent to most GHG accounting methodologies and results**, and it increases when consideration enabled emissions reductions due to a lack of primary data and precise information about real-world actions and their effects. However, **understanding the sources, types, and magnitude of uncertainty is crucial to deploy conservative estimates**, inform improved data inputs, and properly interpret results."

The Environmental Insights Explorer makes actionable climate data available to more than 40,000 cities and regions worldwide.

Uncertainty is inherent in carbon accounting

Cradle-to-gate Life Cycle Assessment of CMOS Logic Technologies

L. Boakes, M. Garcia Bardon, V. Schellekens, I-Y. Liu, B. Vanhouche, G. Mirabelli, F. Sebaai, L. Van Winckel, E. Gallagher, C. Rolin, L.-Å. Ragnarsson
imec, Kapeldreef 75, 3001 Leuven, Belgium
lizzie.boakes@imec.be

Abstract— While concerted efforts have been made to promote greener IC manufacturing, achieving sustainable practices necessitates a comprehensive understanding of the environmental impacts associated with semiconductor fabrication. This paper presents a life cycle analysis of logic technology nodes N28 to A14 based on bottom-up modeling of a high-volume IC fabrication plant. This holistic approach provides granular results, enables sensitivity analysis, and highlights high-impact processes that could be improved to reduce environmental footprints in existing and pathfinding technologies.

I. Introduction

Digitalization enhances system efficiencies and consequently improves environmental sustainability and reutilizes material flows [1]. Nevertheless, the semiconductor manufacturing industry, critical for enabling digitalization through integrated circuit (IC) production, is inherently resource-intensive and has many environmental consequences.

Life Cycle Assessment (LCA) has been used to quantify these impacts in scientific literature. The pioneering study in [2] provides detailed primary data for CMOS logic chips from technology nodes N350 to N32. Subsequently, only a handful of studies have presented a bottom-up LCA model employing primary data [3]. Commercial LCA databases [4,5] assess technology nodes as low as N7 and N14, respectively, but lack process-level insight into environmental impacts.

This paper presents a cradle-to-gate LCA of modern CMOS logic chips through nine available and projected technology nodes, N28 to A14 (Table I). The analysis and

B. Functional unit definition

The results are provided relative to a functional unit, which is defined as the HVM of an industry average 10x10 mm² functional logic chip (representing a mobile system on chip) on a standard 300 mm silicon wafer in an ISO 14644-1 Class 5 clean room. This functional unit is expressed as per wafer, cm², or die, which considers the functional area (taking die yield and placement into account).

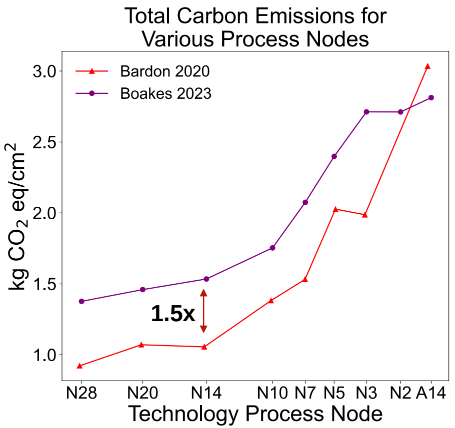
III. Life Cycle Inventory

C. Data classification and collection procedures

The LCA model includes three data streams: equipment level data, process flow data, and fab models. Equipment level data was collected from SSTS program partners and the imec fab. Process flow data and fab models were developed by the SSTS program and its partners. The data in this study is from V1.5.67 of imec.netzero. The output flows of foreground processes are limited to gaseous elementary flows that are not destroyed in the gas abatement system, characterized using [7]. Upstream material and energy flows were characterized using secondary sources and LCA databases [8-12].

D. Model assumptions

The 10x10 mm² die yield of 86% used in this study followed the Murphy model [13]. A line yield of 90% was assumed, which correlates to the line yield used in [5]. The same utilization, idle, and downtime for all Fab tools were assumed following the recommendation from SEMI S23 [14]. The SEMI S23 recommended energy conversion factors (ECFs) were assumed for the generation of fab utilities [14]. The total electrical consumption of the virtual fab facilities was



The Dirty Secret of SSDs: Embodied Carbon

SWAMIT TANNU, University of Wisconsin, Madison, USA
PRASHANT J. NAIR, University of British Columbia, Canada

Scalable Solid-State Drives (SSDs) have ushered in a transformative era in data storage and accessibility, spanning both data centers and portable devices. However, the strides made in scaling this technology can bear significant environmental consequences. On a global scale, a notable portion of semiconductor manufacturing relies on electricity derived from coal and natural gas sources. A striking example of this is the manufacturing process for a single Gigabyte of Flash memory, which emits approximately 0.16 Kg of CO_2 – a considerable fraction of the total carbon emissions attributed to the system. Remarkably, the manufacturing of storage devices alone contributed to an estimated 20 million metric tonnes of CO_2 emissions in the year 2021.

In light of these environmental concerns, this paper delves into an analysis of the sustainability trade-offs inherent in Solid-State Drives (SSDs) when compared to traditional Hard Disk Drives (HDDs). Moreover, this study proposes methodologies to gauge the embodied carbon costs associated with storage systems effectively. The research encompasses four key strategies to enhance the sustainability of storage systems.

Firstly, the paper offers insightful guidance for selecting the most suitable storage medium, be it SSDs or HDDs, considering the broader ecological impact. Secondly, the paper advocates for implementing techniques that extend the lifespan of SSDs, thereby mitigating premature replacements and their attendant environmental toll. Thirdly, the paper emphasizes the need for efficient recycling and reuse of high-density multi-level cell-based SSDs, underscoring the significance of minimizing electronic waste.

Lastly, for handheld devices, the paper underscores the potential of harnessing the elasticity offered by cloud storage solutions as a means to curtail the ecological repercussions of localized data storage. In summation, this study critically addresses the embodied carbon issues associated with SSDs, comparing them with HDDs, and proposes a comprehensive framework of strategies to enhance the sustainability of storage systems.

CCS Concepts: • **Social and professional topics → Sustainability**; • **Hardware → External storage**; • **Applied computing →** *Data centers; Microcomputers.*

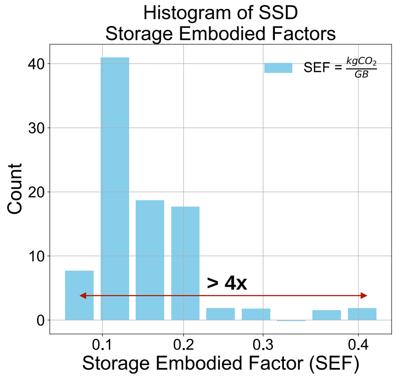
Additional Key Words and Phrases: Embodied Carbon, Solid State Drives, Hard Disk Drive, Sustainability

ACM Reference Format:
Swamit Tannu and Prashant J. Nair. 2023. The Dirty Secret of SSDs: Embodied Carbon. 1, 1 (October 2023), 6 pages. https://doi.org/XXXXXXX.XXXXXXX

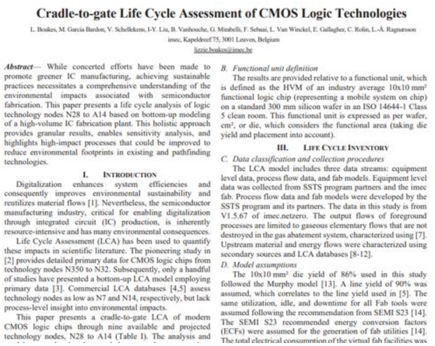
arXiv:2207.10793v2 [cs.AR] 28 Sep 2023

and operation to transportation and recycling. Of particular concern are the billions of hand-held devices, including smartphones, tablets, and web services, that have become integral to modern life. This proliferation of devices has contributed significantly to global warming, with the current combined carbon emissions from computing and networking devices already accounting for approximately 2% of the total carbon emissions [Freitag et al. 2021; Gelenbe and Caseau 2015]. Projections suggest that this percentage could double within the coming decade, underscoring the urgency of addressing these emissions. As digital data creation and consumption continue to surge worldwide, a comprehensive understanding of the carbon emissions from personal devices, data centers, and networking infrastructure—collectively known as the Information and Communication Technologies (ICT) sector—becomes paramount.

The majority of carbon emissions within the ICT sector stems from the utilization of "conventional" electricity sources [EPA 2022b], which play a pivotal role in both the manufacturing and operational phases of computing systems [Gupta et al. 2022]. The energy-intensive tasks of running and cooling computing and networking hardware translate to substantial electricity consumption. When this electricity is sourced from carbon-intensive fuels such as coal, natural gas, and crude oil, the resulting emissions contribute significantly to global warming. Conversely, electricity generated from renewable sources—such as wind, solar, nuclear, and hydroelectric—exhibits a considerably smaller Global Warming Potential (GWP). Nonetheless, a prevailing challenge persists: whether in the context of hand-held devices or server nodes, the manufacturing and operation of hardware invariably demand substantial electricity, often originating from carbon-intensive conventional sources.



Uncertainty is inherent in carbon accounting



Cradle-to-gate Life Cycle Assessment of CMOS Logic Technologies

L. Boakes, M. Garcia Bardon, V. Schellekens, I-Y. Liu, B. Vanhouche, G. Mirabelli, F. Sebaai, L. Van Winckel, E. Gallagher, C. Rolin, L.-Å. Ragnarsson
imec, Kapeldreef 75, 3001 Leuven, Belgium
lizzie.boakes@imec.be

Abstract— While concerted efforts have been made to promote greener IC manufacturing, achieving sustainable practices necessitates a comprehensive understanding of the environmental impacts associated with semiconductor fabrication. This paper presents a life cycle analysis of logic technology nodes N28 to A14 based on bottom-up modeling of a high-volume IC fabrication plant. This holistic approach provides granular results, enables sensitivity analysis, and highlights high-impact processes that could be improved to reduce environmental footprints in existing and pathfinding technologies.

I. INTRODUCTION

Digitalization enhances system efficiencies and consequently improves environmental sustainability and reutilizes material flows [1]. Nevertheless, the semiconductor manufacturing industry, critical for enabling digitalization through integrated circuit (IC) production, is inherently resource-intensive and has many environmental consequences.

Life Cycle Assessment (LCA) has been used to quantify these impacts in scientific literature. The pioneering study in [2] provides detailed primary data for CMOS logic chips from technology nodes N350 to N32. Subsequently, only a handful of studies have presented a bottom-up LCA model employing primary data [3]. Commercial LCA databases [4,5] assess technology nodes as low as N7 and N14, respectively, but lack process-level insight into environmental impacts.

This paper presents a cradle-to-gate LCA of modern CMOS logic chips through nine available and projected technology nodes, N28 to A14 (Table I). The analysis and

B. Functional unit definition

The results are provided relative to a functional unit, which is defined as the HVM of an industry average 10x10 mm^2 functional logic chip (representing a mobile system on chip) on a standard 300 mm silicon wafer in an ISO 14644-1 Class 5 clean room. This functional unit is expressed as per wafer, cm^2, or die, which considers the functional area (taking die yield and placement into account).

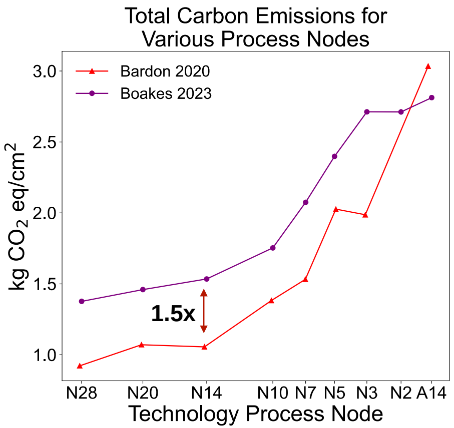
III. LIFE CYCLE INVENTORY

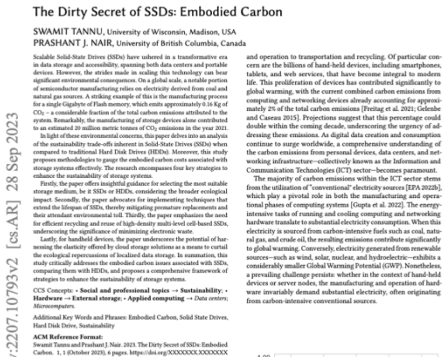
C. Data classification and collection procedures

The LCA model includes three data streams: equipment level data, process flow data, and fab models. Equipment level data was collected from SSTS program partners and the imec fab. Process flow data and fab models were developed by the SSTS program and its partners. The data in this study is from V1.5.67 of imec.netzero. The output flows of foreground processes are limited to gaseous elementary flows that are not destroyed in the gas abatement system, characterized using [7]. Upstream material and energy flows were characterized using secondary sources and LCA databases [8-12].

D. Model assumptions

The 10x10 mm^2 die yield of 86% used in this study followed the Murphy model [13]. A line yield of 90% was assumed, which correlates to the line yield used in [5]. The same utilization, idle, and downtime for all Fab tools were assumed following the recommendation from SEMI S23 [14]. The SEMI S23 recommended energy conversion factors (ECFs) were assumed for the generation of fab utilities [14]. The total electrical consumption of the virtual fab facilities was





The Dirty Secret of SSDs: Embodied Carbon

SWAMIT TANNU, University of Wisconsin, Madison, USA
PRASHANT J. NAIR, University of British Columbia, Canada

Scalable Solid-State Drives (SSDs) have ushered in a transformative era in data storage and accessibility, spanning both data centers and portable devices. However, the strides made in scaling this technology can bear significant environmental consequences. On a global scale, a notable portion of semiconductor manufacturing relies on electricity derived from coal and natural gas sources. A striking example of this is the manufacturing process for a single Gigabyte of Flash memory, which emits approximately 0.16 Kg of CO_2 – a considerable fraction of the total carbon emissions attributed to the system. Remarkably, the manufacturing of storage devices alone contributed to an estimated 20 million metric tonnes of CO_2 emissions in the year 2021.

In light of these environmental concerns, this paper delves into an analysis of the sustainability trade-offs inherent in Solid-State Drives (SSDs) when compared to traditional Hard Disk Drives (HDDs). Moreover, this study proposes methodologies to gauge the embodied carbon costs associated with storage systems effectively. The research encompasses four key strategies to enhance the sustainability of storage systems.

Firstly, the paper offers insightful guidance for selecting the most suitable storage medium, be it SSDs or HDDs, considering the broader ecological impact. Secondly, the paper advocates for implementing techniques that extend the lifespan of SSDs, thereby mitigating premature replacements and their attendant environmental toll. Thirdly, the paper emphasizes the need for efficient recycling and reuse of high-density multi-level cell-based SSDs, underscoring the significance of minimizing electronic waste.

Lastly, for handheld devices, the paper underscores the potential of harnessing the elasticity offered by cloud storage solutions as a means to curtail the ecological repercussions of localized data storage. In summation, this study critically addresses the embodied carbon issues associated with SSDs, comparing them with HDDs, and proposes a comprehensive framework of strategies to enhance the sustainability of storage systems.

CCS Concepts: • **Social and professional topics → Sustainability**; • **Hardware → External storage**; • **Applied computing** → *Data centers*; *Microcomputers*.

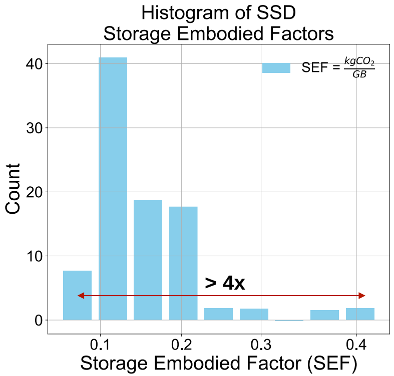
Additional Key Words and Phrases: Embodied Carbon, Solid State Drives, Hard Disk Drive, Sustainability

ACM Reference Format:
Swamit Tannu and Prashant J. Nair. 2023. The Dirty Secret of SSDs: Embodied Carbon. 1, 1 (October 2023), 6 pages. https://doi.org/XXXXXXX.XXXXXXX

and operation to transportation and recycling. Of particular concern are the billions of hand-held devices, including smartphones, tablets, and web services, that have become integral to modern life. This proliferation of devices has contributed significantly to global warming, with the current combined carbon emissions from computing and networking devices already accounting for approximately 2% of the total carbon emissions [Freitag et al. 2021; Gelenbe and Caseau 2015]. Projections suggest that this percentage could double within the coming decade, underscoring the urgency of addressing these emissions. As digital data creation and consumption continue to surge worldwide, a comprehensive understanding of the carbon emissions from personal devices, data centers, and networking infrastructure—collectively known as the Information and Communication Technologies (ICT) sector—becomes paramount.

The majority of carbon emissions within the ICT sector stems from the utilization of "conventional" electricity sources [EPA 2022b], which play a pivotal role in both the manufacturing and operational phases of computing systems [Gupta et al. 2022]. The energy-intensive tasks of running and cooling computing and networking hardware translate to substantial electricity consumption. When this electricity is sourced from carbon-intensive fuels such as coal, natural gas, and crude oil, the resulting emissions contribute significantly to global warming. Conversely, electricity generated from renewable sources—such as wind, solar, nuclear, and hydroelectric—exhibits a considerably smaller Global Warming Potential (GWP). Nonetheless, a prevailing challenge persists: whether in the context of hand-held devices or server nodes, the manufacturing and operation of hardware invariably demand substantial electricity, often originating from carbon-intensive conventional sources.

arXiv:2207.10793v2 [cs.AR] 28 Sep 2023



Open research questions:

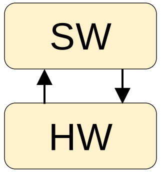
- What magnitude uncertainty exists across all IC components?
- What degree of uncertainty exists in embodied versus operational carbon?
- How do we consider uncertainty in carbon-aware hardware design to enable robust sustainable computing decisions?

Today: Quantifying the carbon footprint of computing

with insights, opening new research and sustainable development opportunities

Understanding the source of computing's emissions

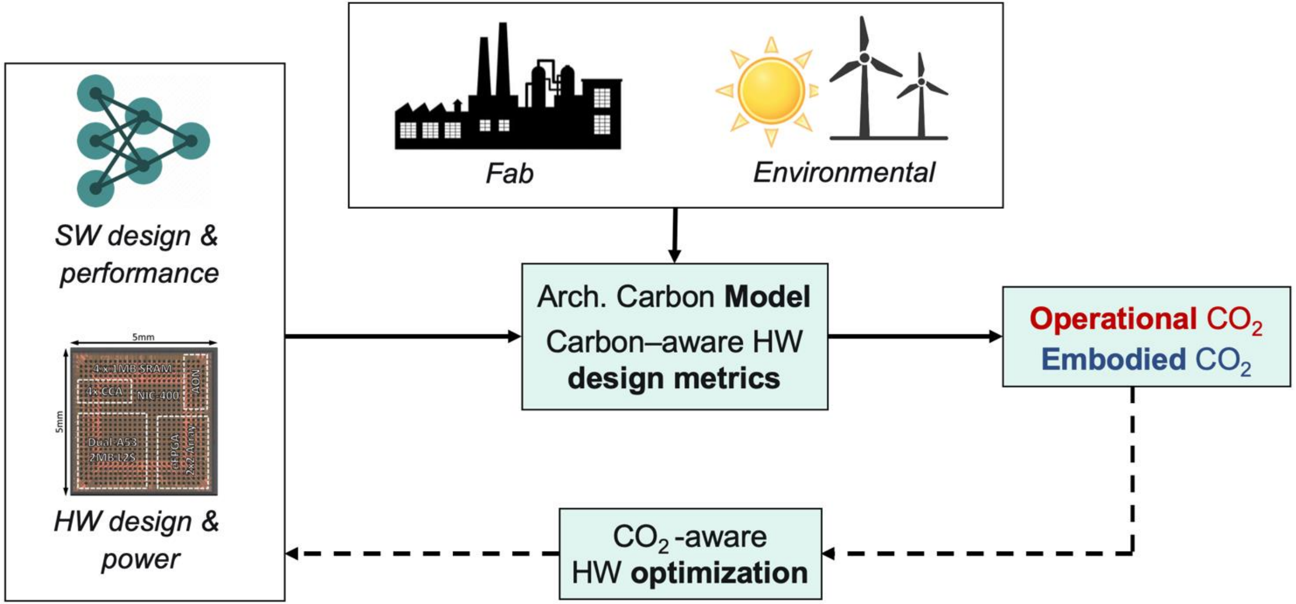
Deep dive: developing computer architectural models to estimate CO_2 emissions



Cross stack: Developing modeling methods across the computing stack

Need to go beyond architecture centric-view for cross-stack carbon accounting

- Application
- OS/Run-time
- Systems
- Compiler
- Architecture
- Circuits
- Devices & Technology



Application

OS/Run-time

Systems

Compiler

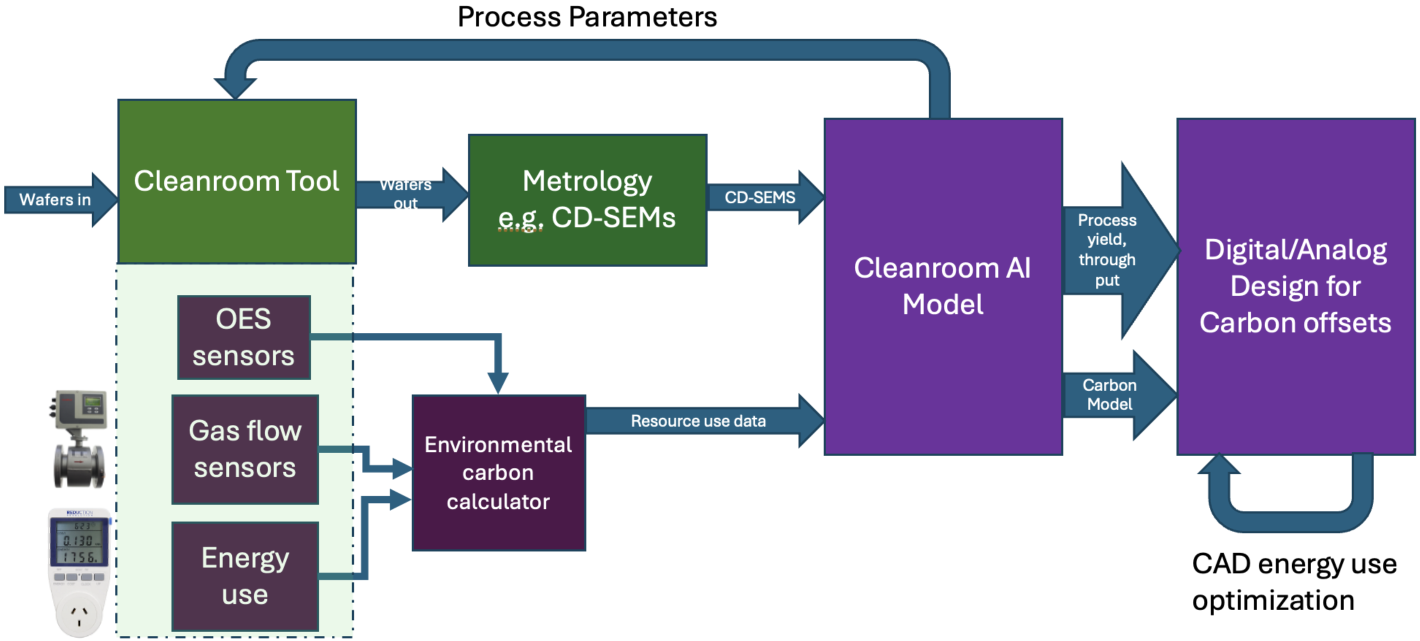
Architecture

Circuits

Devices & Technology

Instrumenting Cornell’s NanoScale Facility

- Application
- OS/Run-time
- Systems
- Compiler
- Architecture
- Circuits
- Devices & Technology

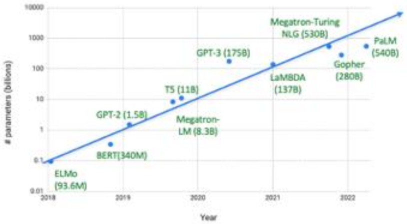


CNF Cornell NanoScale Science and Technology Facility

Rising emissions from AI Footprint

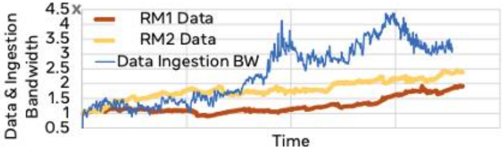
- Application
- OS/Run-time
- Systems
- Compiler
- Architecture
- Circuits
- Devices & Technology

Model Growth



- State-of-the-art DL models growing in capacity by **10x/year**

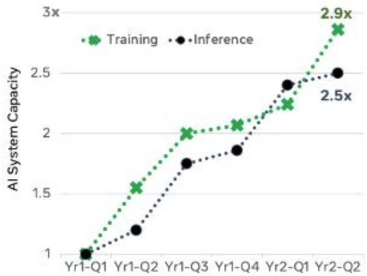
Data Growth



For recommendation models, in two years:

- Data stored and used **doubled**
- Storage bandwidth grew by **3.2x**

Systems Growth

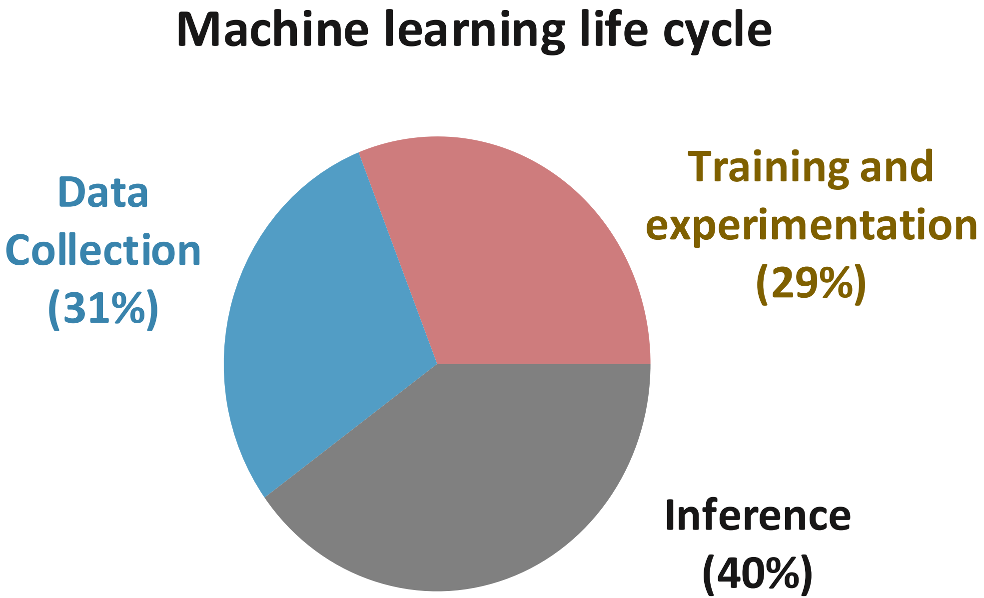


In 1.5 years:

- Training capacity grew by **2.9x**
- Inference capacity grew by 2.5x

Crucial to look at emissions across ML cycle

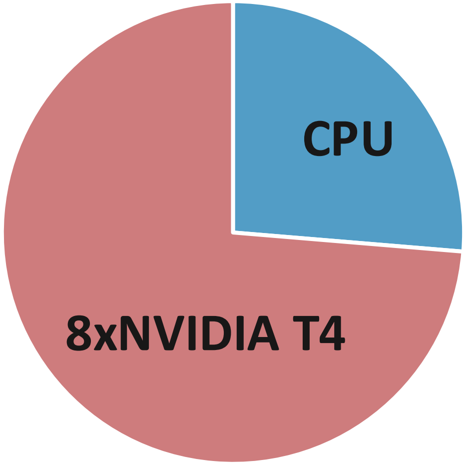
- Application
- OS/Run-time
- Systems
- Compiler
- Architecture
- Circuits
- Devices & Technology



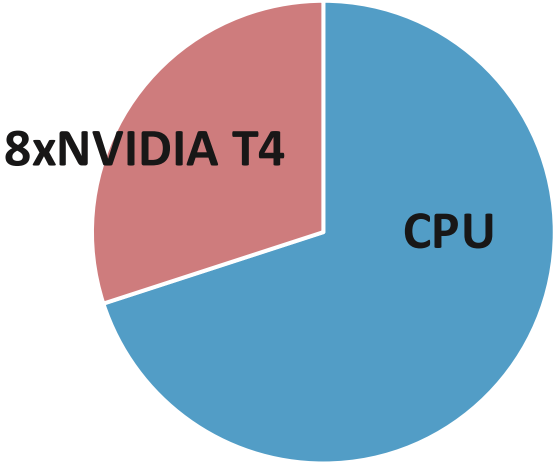
Crucial to look at emissions across ML cycle

- Application
- OS/Run-time
- Systems
- Compiler
- Architecture
- Circuits
- Devices & Technology

Inference Power



Inference Server Embodied Carbon



Crucial to look at emissions across ML cycle

Application

OS/Run-time

Systems

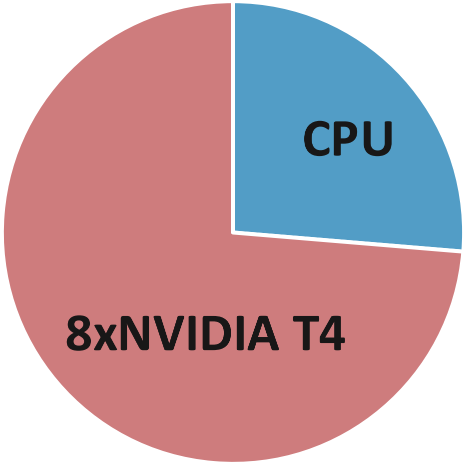
Compiler

Architecture

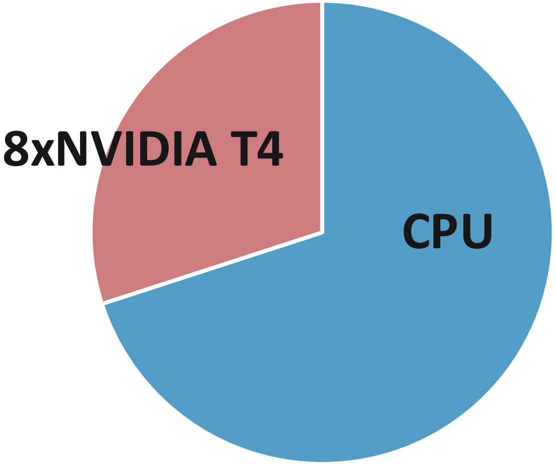
Circuits

Devices & Technology

Inference Power



Inference Server Embodied Carbon



Research questions:

- How do the breakdowns scale for different AI applications and hardware?
- How do we co-optimize AI systems for end-to-end carbon?

“Towards Carbon-efficient LLM Life Cycle” Yueying Li, Omer Graif, Udit Gupta (HotCarbon 2024)

Carbon Explorer: Carbon-Aware Datacenter Design

Application

OS/Run-time

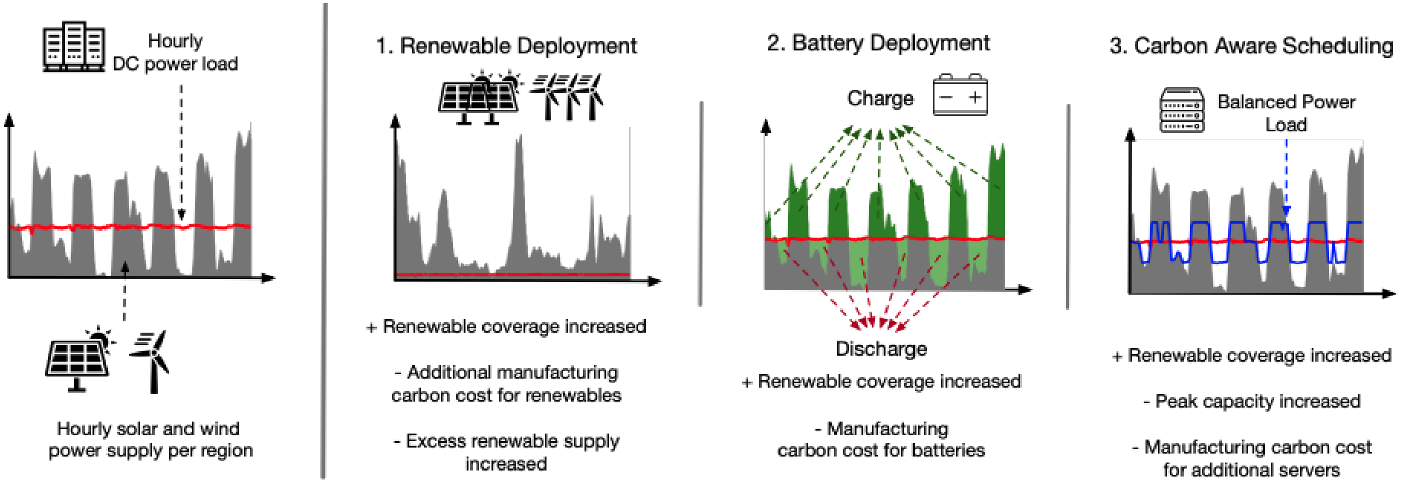
Systems

Compiler

Architecture

Circuits

Devices & Technology



Optimizing both *operational* and *embodied* carbon requires balancing (1) **renewable deployment**, (2) **battery deployment**, (3) **carbon-aware scheduling**

“Carbon Explorer: A Holistic Framework for Designing Carbon Aware Data centers” Acun et. al. (To appear in ASPLOS 2023)

Attributing carbon footprint of cloud usage

- Application
- OS/Run-time
- Systems
- Compiler
- Architecture
- Circuits
- Devices & Technology

Fairly attributing carbon

Application

OS/Run-time

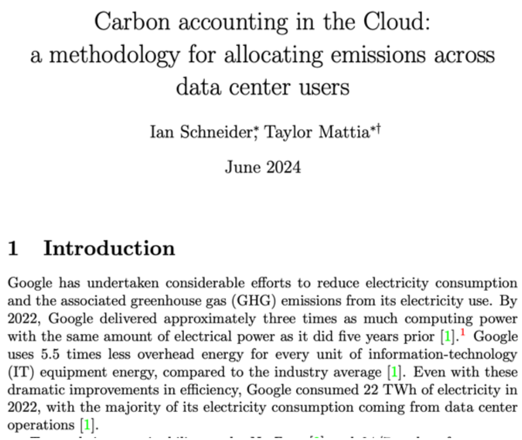
Systems

Compiler

Architecture

Circuits

Devices & Technology



Carbon accounting in the Cloud:
a methodology for allocating emissions across data center users

Ian Schneider*, Taylor Mattia*†

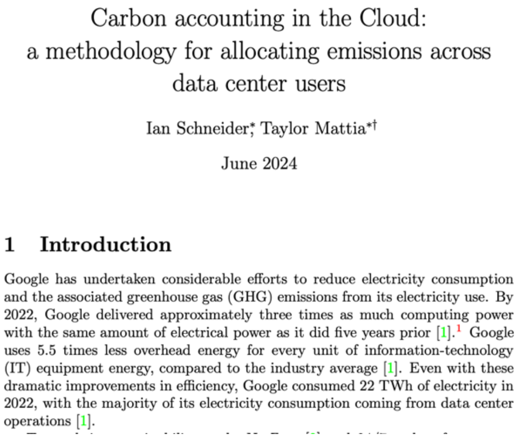
June 2024

1 Introduction

Google has undertaken considerable efforts to reduce electricity consumption and the associated greenhouse gas (GHG) emissions from its electricity use. By 2022, Google delivered approximately three times as much computing power with the same amount of electrical power as it did five years prior [1].[1] Google uses 5.5 times less overhead energy for every unit of information-technology (IT) equipment energy, compared to the industry average [1]. Even with these dramatic improvements in efficiency, Google consumed 22 TWh of electricity in 2022, with the majority of its electricity consumption coming from data center operations [1].

Fairly attributing carbon

- Application
- OS/Run-time
- Systems
- Compiler
- Architecture
- Circuits
- Devices & Technology



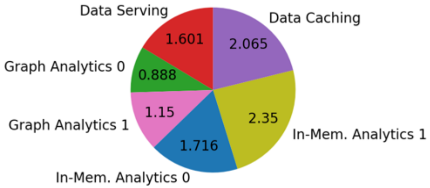
Carbon accounting in the Cloud:
a methodology for allocating emissions across data center users

Ian Schneider*, Taylor Mattia*†

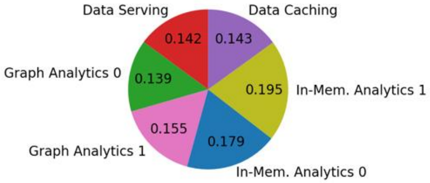
June 2024

1 Introduction

Google has undertaken considerable efforts to reduce electricity consumption and the associated greenhouse gas (GHG) emissions from its electricity use. By 2022, Google delivered approximately three times as much computing power with the same amount of electrical power as it did five years prior [1].[1] Google uses 5.5 times less overhead energy for every unit of information-technology (IT) equipment energy, compared to the industry average [1]. Even with these dramatic improvements in efficiency, Google consumed 22 TWh of electricity in 2022, with the majority of its electricity consumption coming from data center operations [1].



Operational Breakdown



Embodied Breakdown

Open research questions:

- How do we fairly attribute operational and embodied carbon to individual cloud services?
- How do we consider varying demand in data centers in attributing carbon responsibility?
- How do we scale attribution mechanisms to cloud-scale?

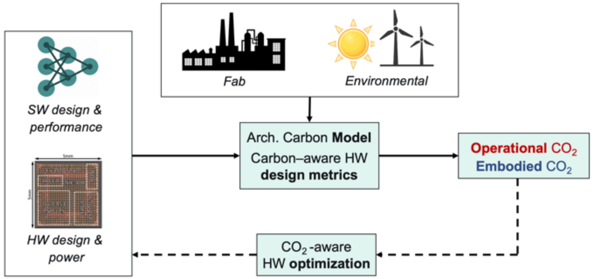
Key takeaways

Quantifying the carbon footprint of AI and computing: Past, Present, and Future

Pillars of Advancing Sustainable Computing

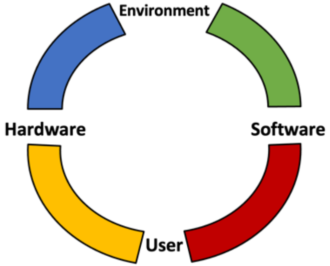
Economics & Policy
Education
Accounting and Reporting

Must analyze emissions across life cycles



Across hardware manufacturing to operational use

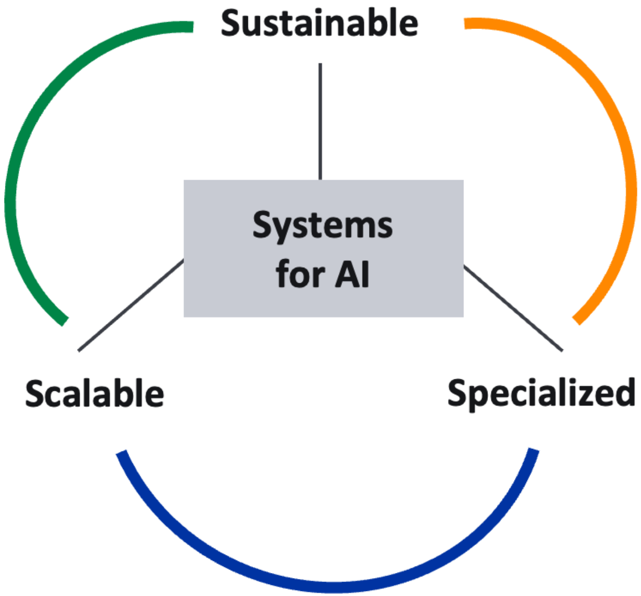
Dire need for cross-stack carbon optimization



From applications to devices
Across eco-HW/SW design loop

S^4AI: Specialized, Scalable, Sustainable Systems for AI

Collaborators



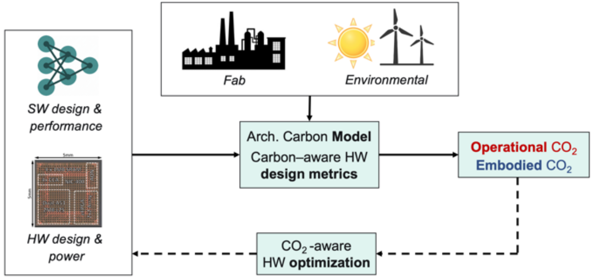
Key takeaways

Quantifying the carbon footprint of AI and computing: Past, Present, and Future

Pillars of Advancing Sustainable Computing

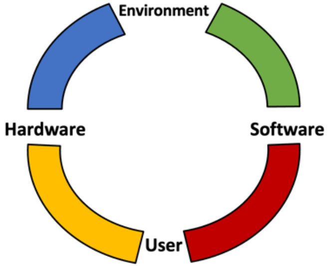
Economics & Policy
Education
Accounting and Reporting

Must analyze emissions across life cycles



Across hardware manufacturing to operational use

Dire need for cross-stack carbon optimization

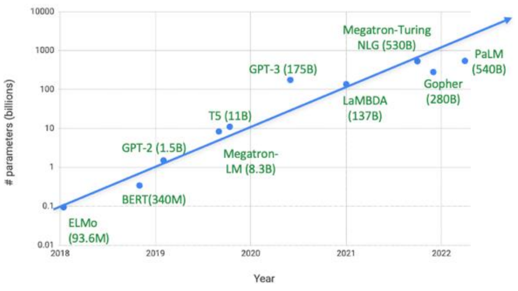


From applications to devices
Across eco-HW/SW design loop

AI Growth Driving Datacenter Energy Consumption

AI Growth Driving Datacenter Energy Consumption

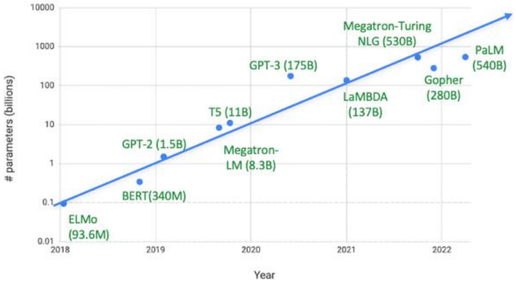
Model Growth



- State-of-the-art DL models growing in capacity by **10x/year**

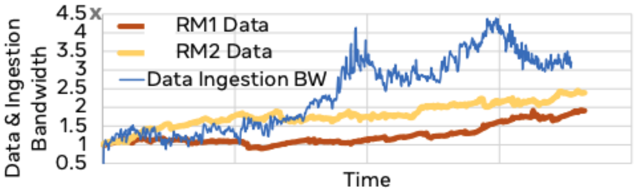
AI Growth Driving Datacenter Energy Consumption

Model Growth



- State-of-the-art DL models growing in capacity by **10x/year**

Data Growth

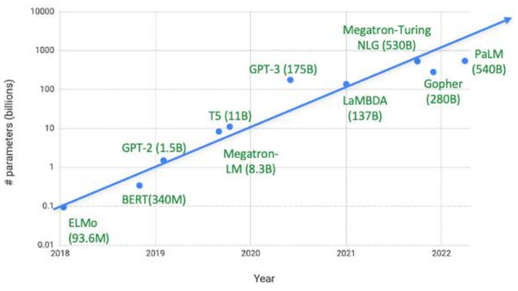


For recommendation models, in two years:

- Data stored and used **doubled**
- Storage bandwidth grew by **3.2x**

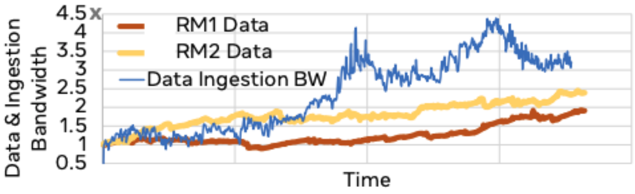
AI Growth Driving Datacenter Energy Consumption

Model Growth



- State-of-the-art DL models growing in capacity by **10x/year**

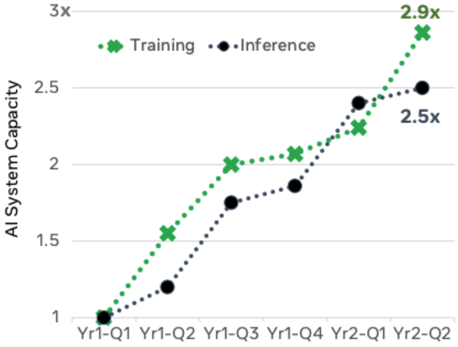
Data Growth



For recommendation models, in two years:

- Data stored and used **doubled**
- Storage bandwidth grew by **3.2x**

Systems Growth

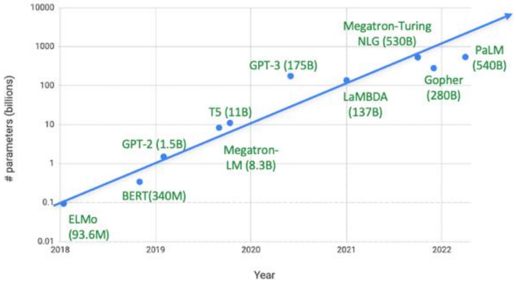


In 1.5 years:

- Training capacity grew by **2.9x**
- Inference capacity grew by 2.5x

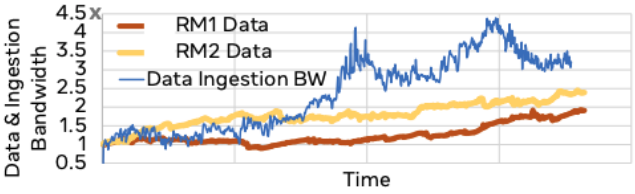
AI Growth Driving Datacenter Energy Consumption

Model Growth



- State-of-the-art DL models growing in capacity by **10x/year**

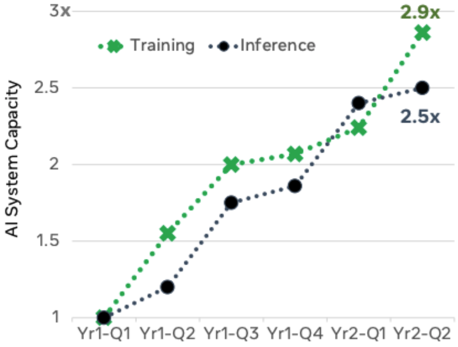
Data Growth



For recommendation models, in two years:

- Data stored and used **doubled**
- Storage bandwidth grew by **3.2x**

Systems Growth



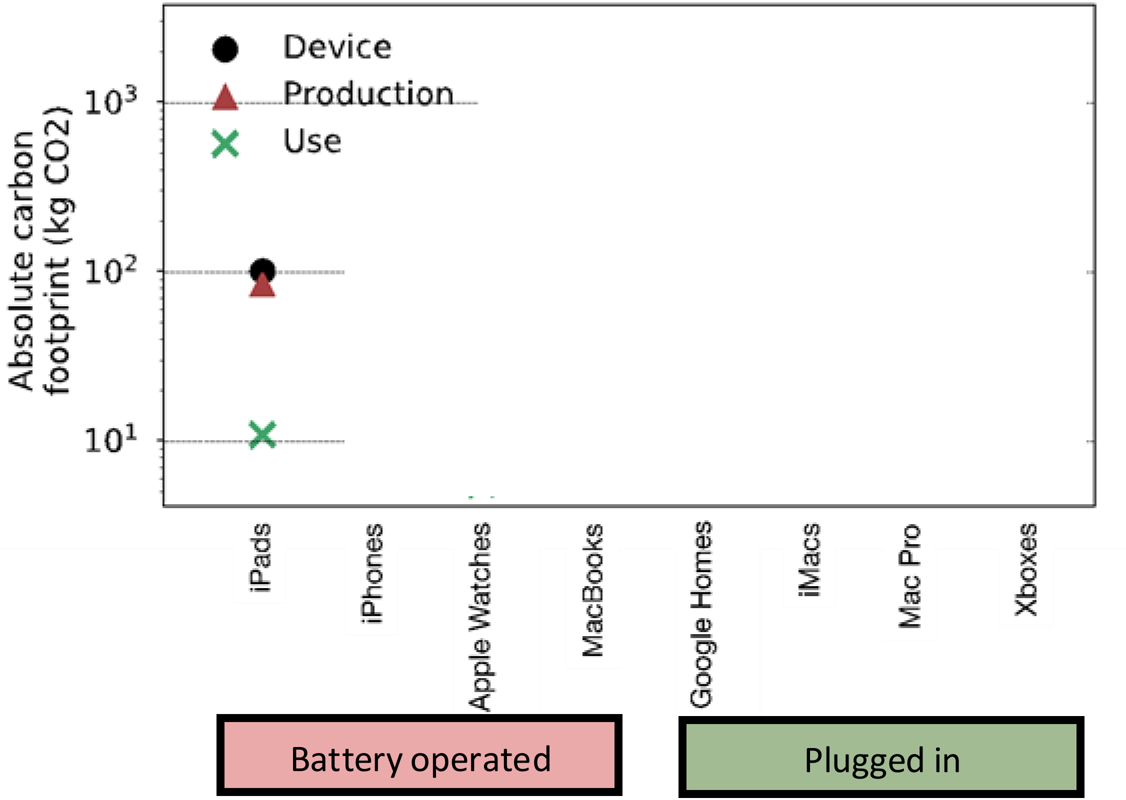
In 1.5 years:

- Training capacity grew by **2.9x**
- Inference capacity grew by 2.5x

Recent advances in LLMs further exacerbating model, data, and systems trends!

Carbon footprint characteristics vary across devices

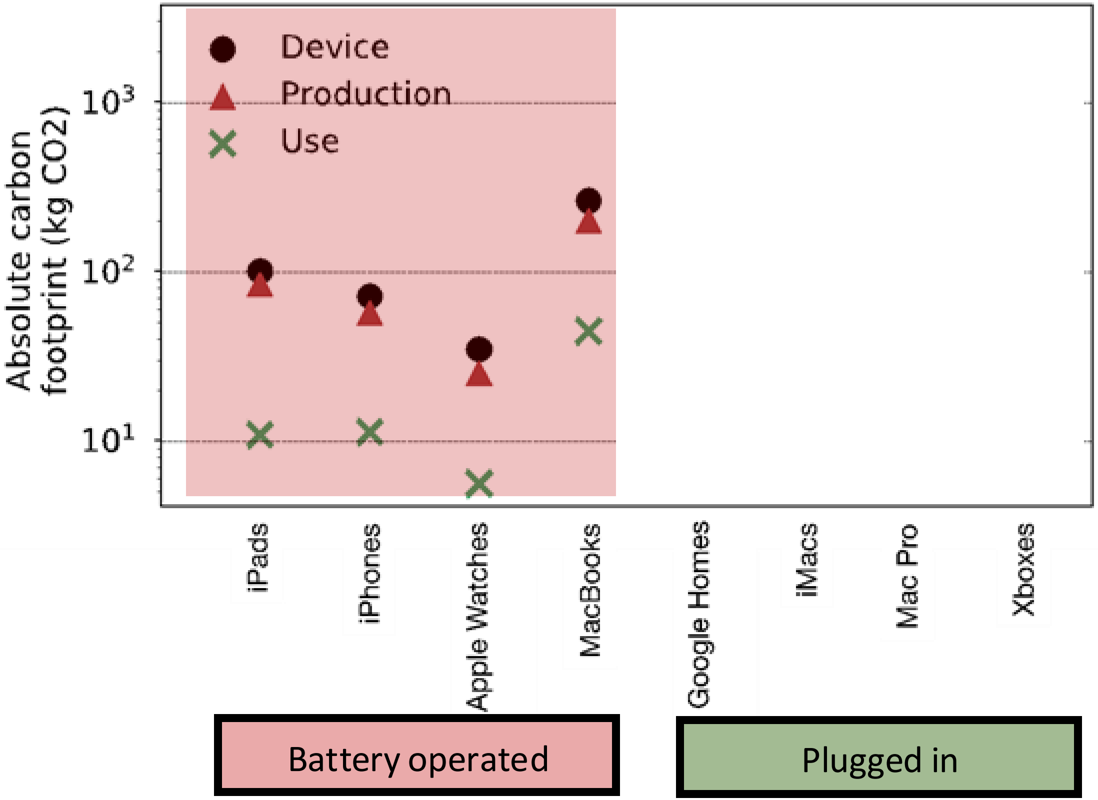
Data from public industry validated sustainability reports and life cycle analyses



Carbon footprint characteristics vary across devices

Data from public industry validated sustainability reports and life cycle analyses

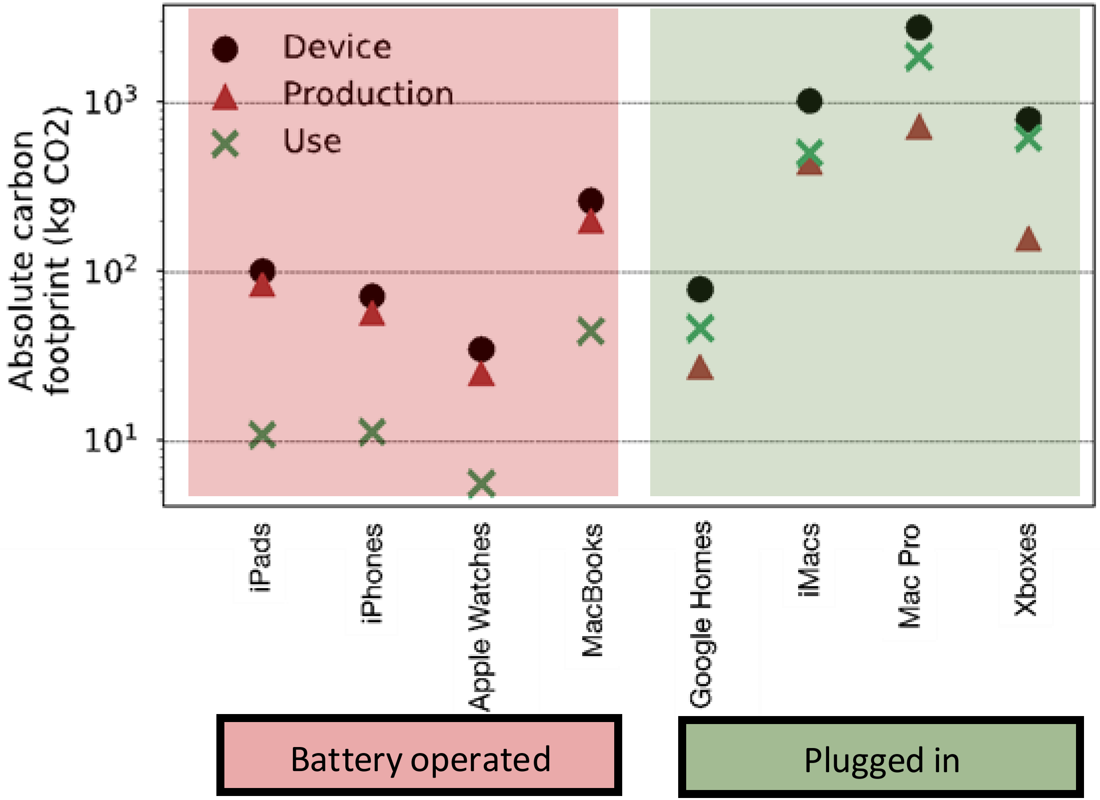
Roughly **75%** life cycle emissions for **battery operated devices** comes from hardware manufacturing.



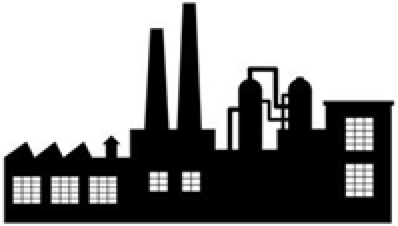
Carbon footprint characteristics vary across devices

Data from public industry validated sustainability reports and life cycle analyses

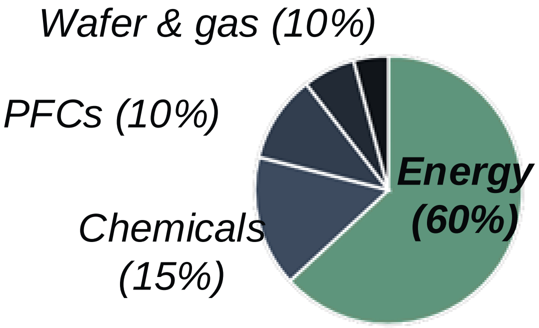
Roughly **75%** life cycle emissions for **battery operated devices** comes from hardware manufacturing.



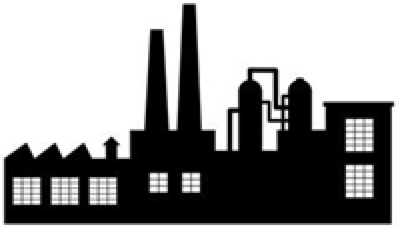
Emissions for **always-connected devices** come mainly from **energy consumption**



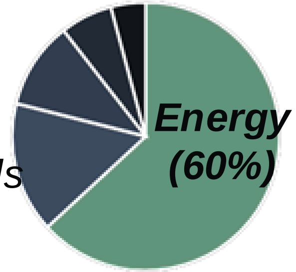
Semiconductor fab

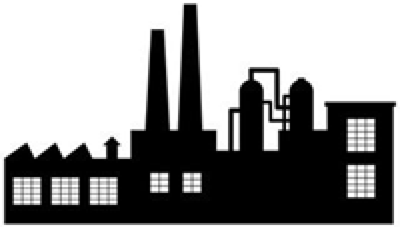


Gupta et. al. Chasing Carbon: The Elusive Environmental Footprint of Computing (HPCA 2021)

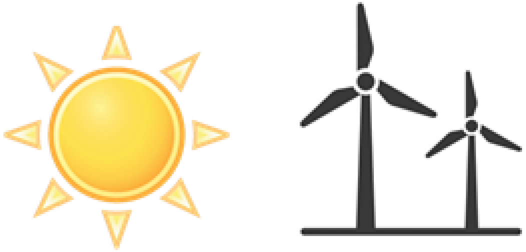


Semiconductor fab





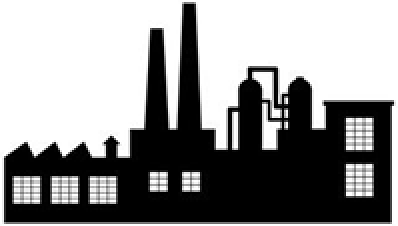
100% Renewable *powered semiconductor fab*



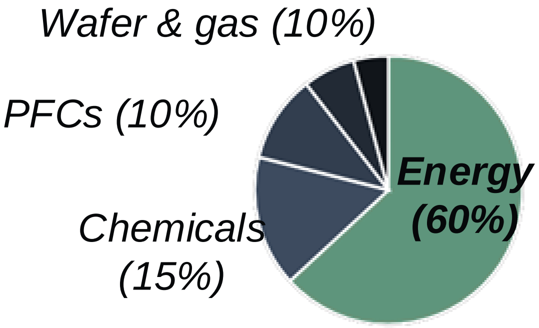
Reduces manufacturing footprint by ***2.5x***

Gupta et. al. Chasing Carbon: The Elusive Environmental Footprint of Computing (HPCA 2021)

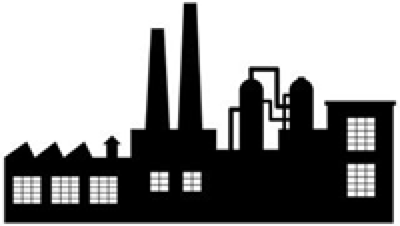
“Green” powered fabs are not enough



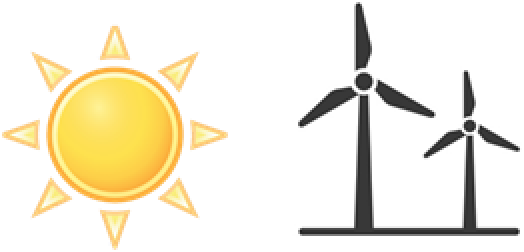
Semiconductor fab



TSMC plans for **25% renewable by 2025** and **100% renewable by 2050**.



100% Renewable *powered semiconductor fab*



Reduces manufacturing footprint by ***2.5x***

Gupta et. al. Chasing Carbon: The Elusive Environmental Footprint of Computing (HPCA 2021)

Architectural Carbon Model

Model	Hardware/software input

Architectural Carbon Model

Model	Hardware/software input
$Carbon = OP_{CF} + \frac{Runtime}{Lifetime} Emb_{CF}$	**Performance/power/energy** and **lifetime** of hardware

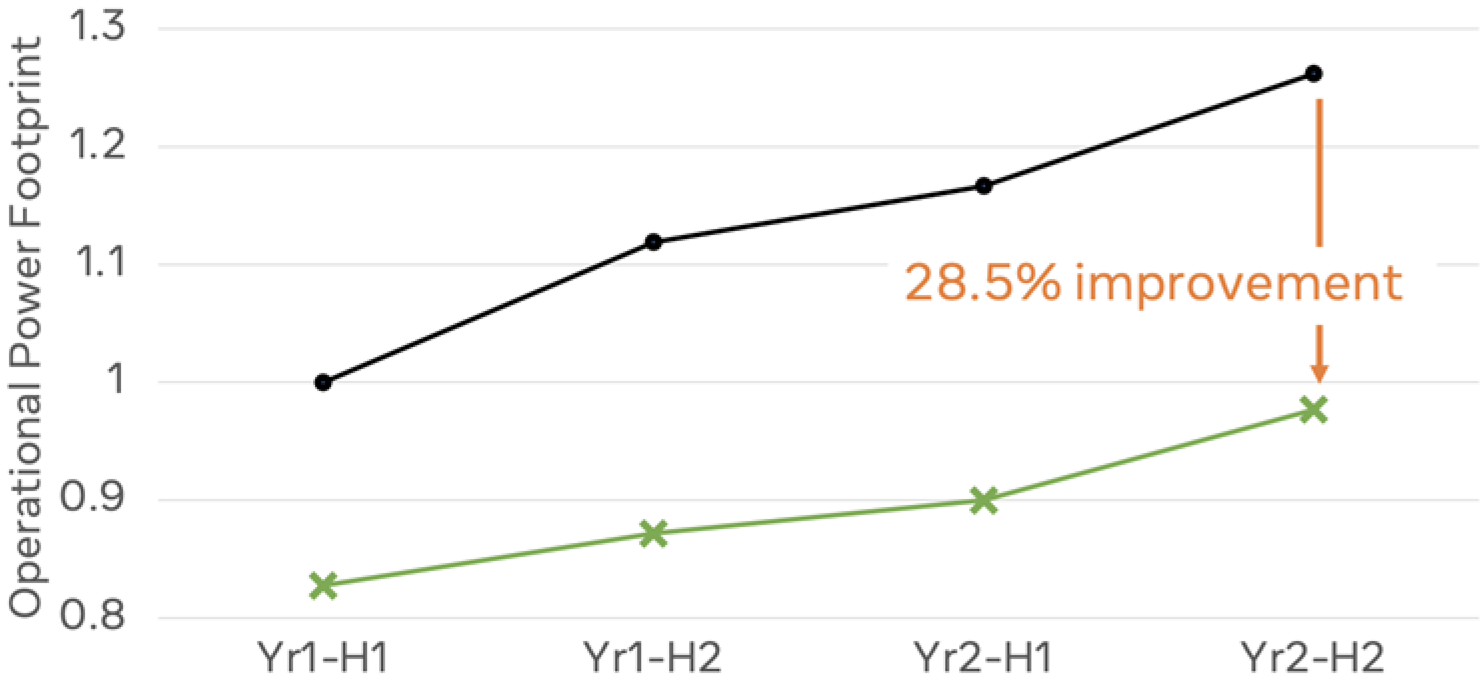
Architectural Carbon Model

Model	Hardware/software input
$Carbon = OP_{CF} + \frac{Runtime}{Lifetime} Emb_{CF}$	**Performance/power/energy** and **lifetime** of hardware
$OP_{CF} = CI_{use} \times Energy$	**Energy efficiency** and environment (**carbon intensity**)

Embodied carbon of application processors (SoC's)

$$Emb_{SoC} = Area \times \frac{(CI_{fab} \times Fab_{energy}) + Fab_{chemicals} + Fab_{materials}}{Yield}$$

Jevon’s paradox of AI at-scale

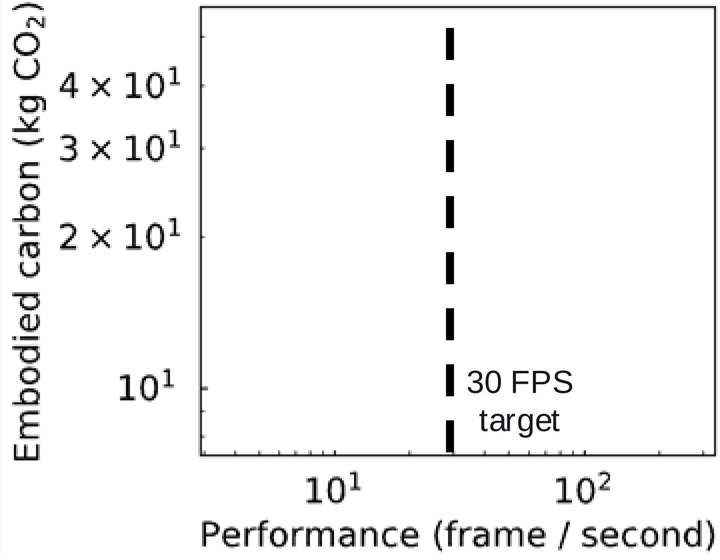


Reduce: Designing leaner hardware systems

{ Nvidia DL Accelerator (NVDLA) } → **Architectural Carbon Model** → Performance/energy vs. Carbon

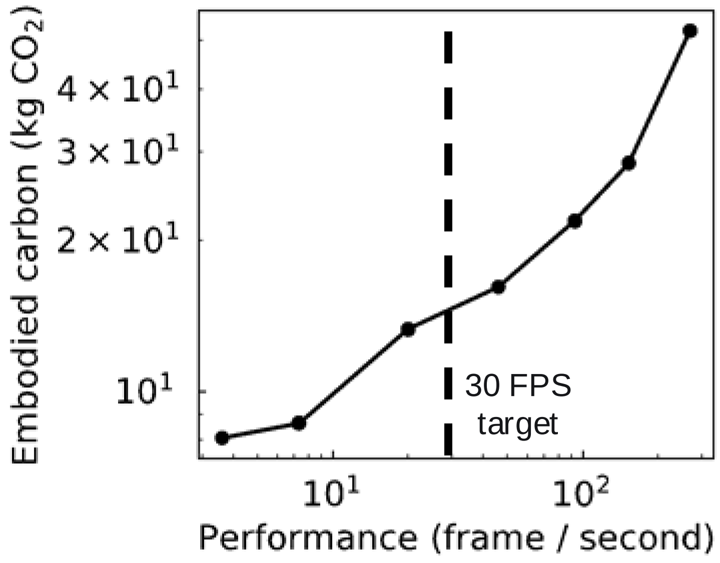
Reduce: Designing leaner hardware systems

{ Nvidia DL Accelerator (NVDLA) } → **Architectural Carbon Model** → Performance/energy vs. Carbon



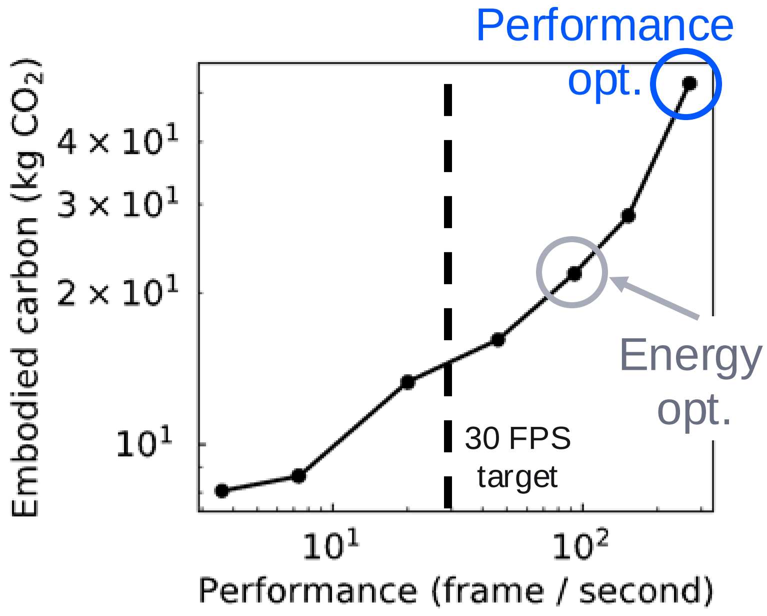
Reduce: Designing leaner hardware systems

Nvidia DL Accelerator (NVDLA) → **Architectural Carbon Model** → Performance/energy vs. Carbon



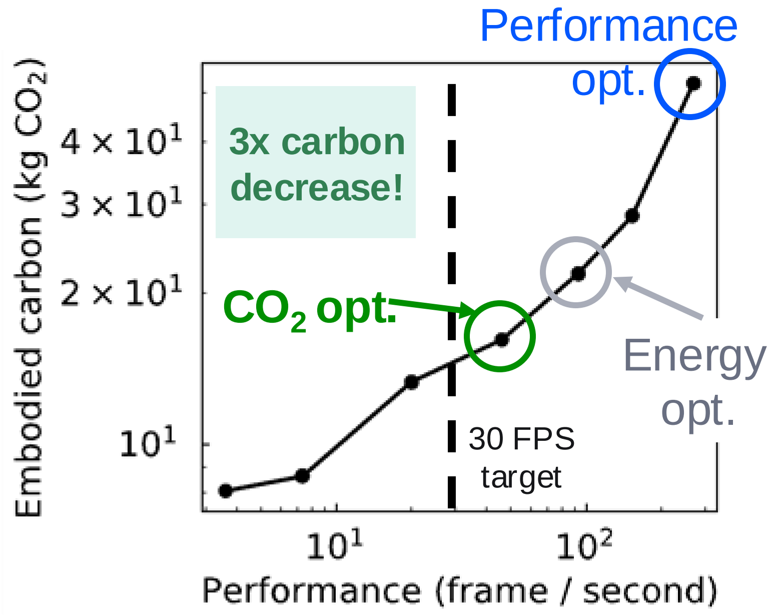
Reduce: Designing leaner hardware systems

{ Nvidia DL Accelerator (NVDLA) } → **Architectural Carbon Model** → Performance/energy vs. Carbon

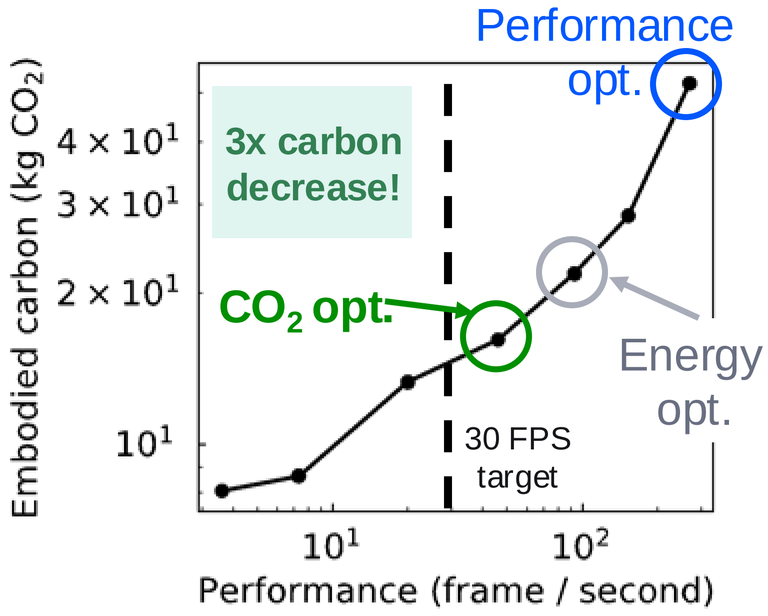


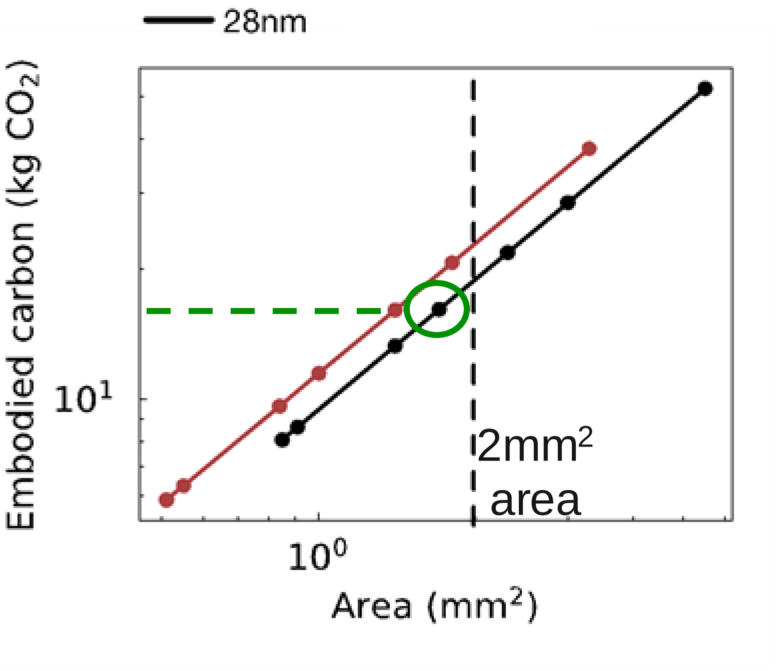
Reduce: Designing leaner hardware systems

Nvidia DL Accelerator (NVDLA) → **Architectural Carbon Model** → Performance/energy vs. Carbon



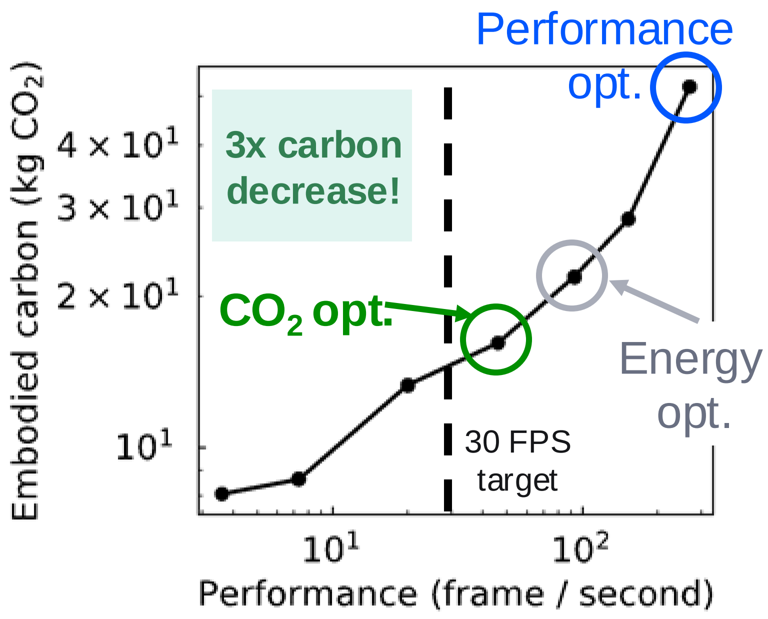
Reduce: Designing leaner hardware systems

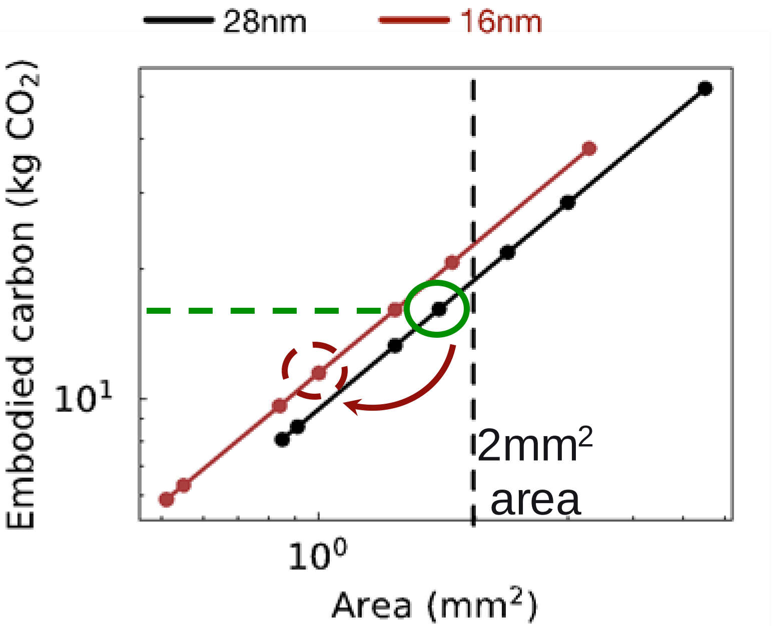




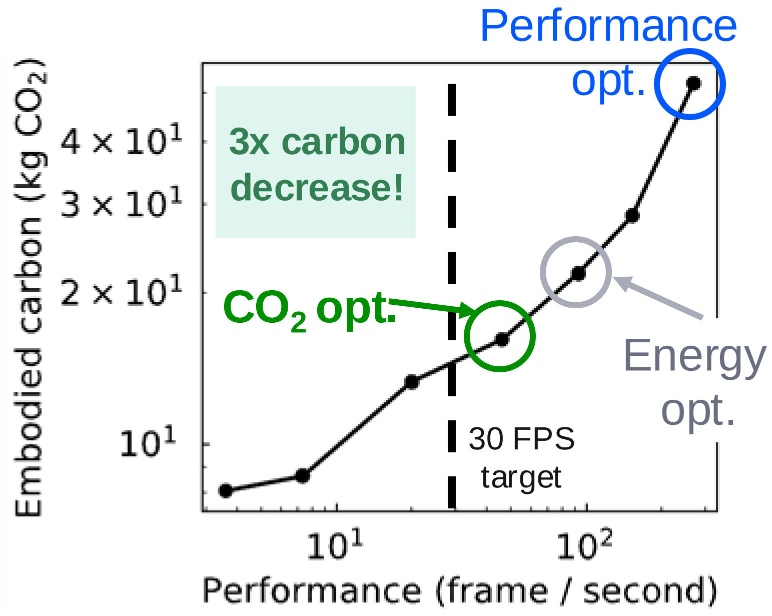
Reduce: Designing leaner hardware systems

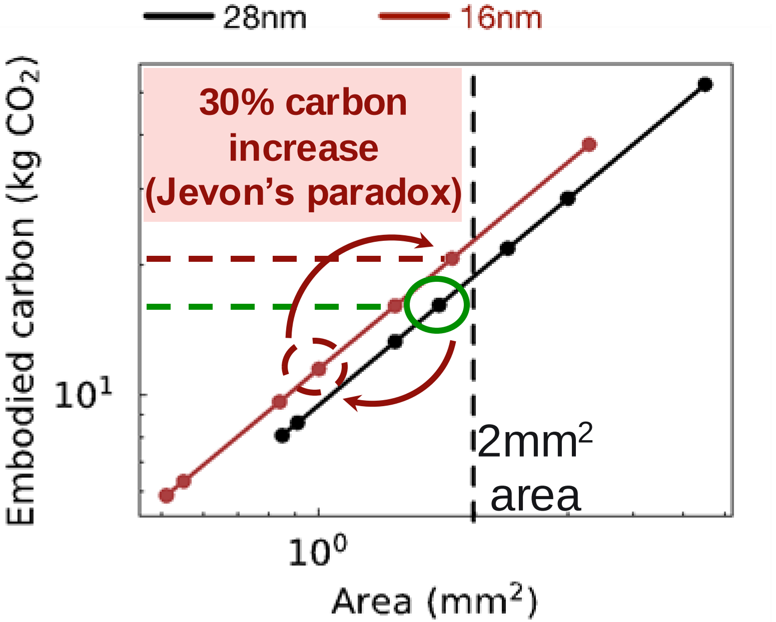
{Nvidia DL Accelerator (NVDLA)} → **Architectural Carbon Model** → Performance/energy vs. Carbon





Reduce: Designing leaner hardware systems





DRAM Memory Embodied Carbon Emissions