Sustainable Computing & the MLPerf Project

Carole-Jean Wu

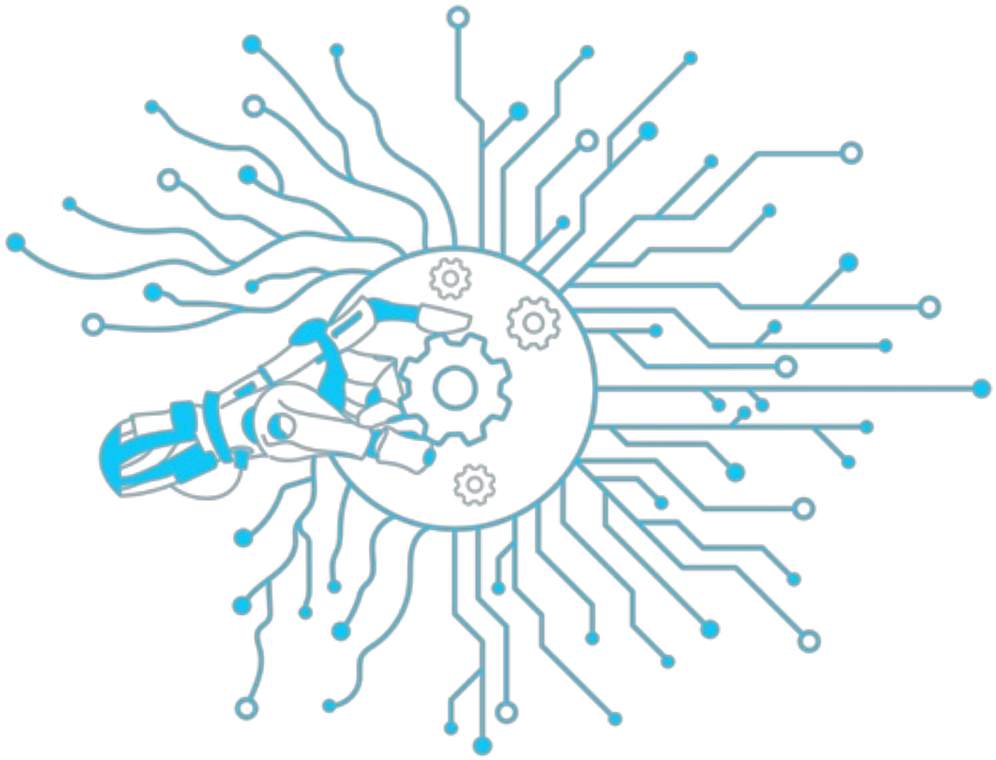
Director
FAIR at Meta

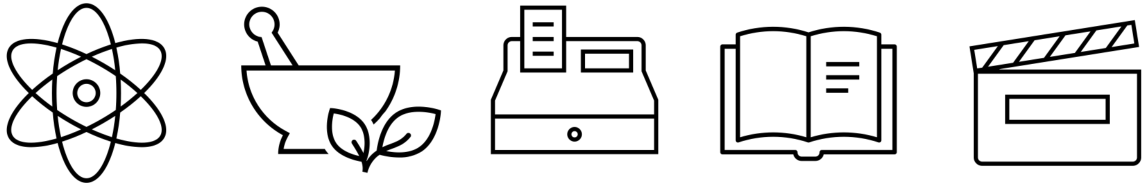
Invited Plenary Talk

International Workshop on Advanced Computing and Analysis Techniques in Physics Research

Computing Industry Faces an Unprecedented Growth

Artificial Intelligence



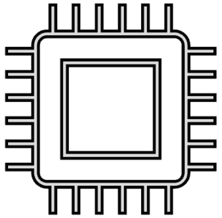


Open Catalyst

AlphaFold

FarmBeats

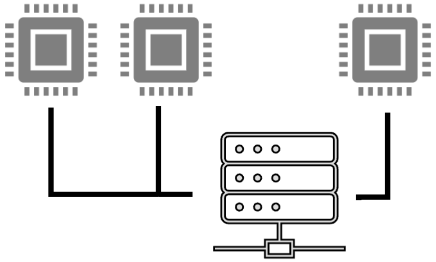
Decades of Innovations in Computer Systems



>1, 000,000x+ Transistors

750x Faster

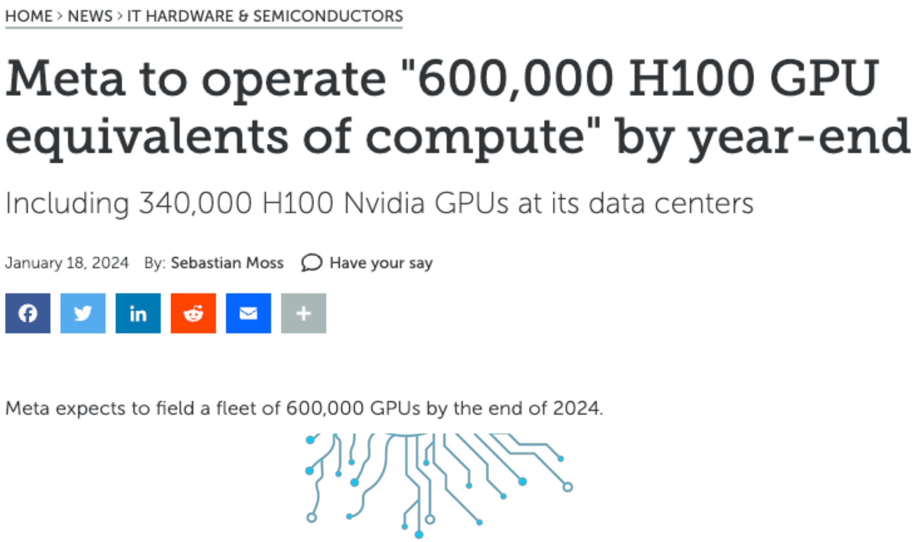
Vertical Scaling



15,000+ GPUs

20x Faster

Horizontal Scaling



HOME > NEWS > IT HARDWARE & SEMICONDUCTORS

Meta to operate "600,000 H100 GPU equivalents of compute" by year-end

Including 340,000 H100 Nvidia GPUs at its data centers

January 18, 2024 By: Sebastian Moss Have your say

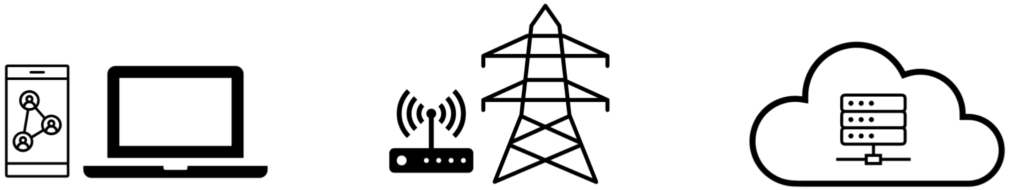
Meta expects to field a fleet of 600,000 GPUs by the end of 2024.

Domain-Specific Acceleration

Specialization

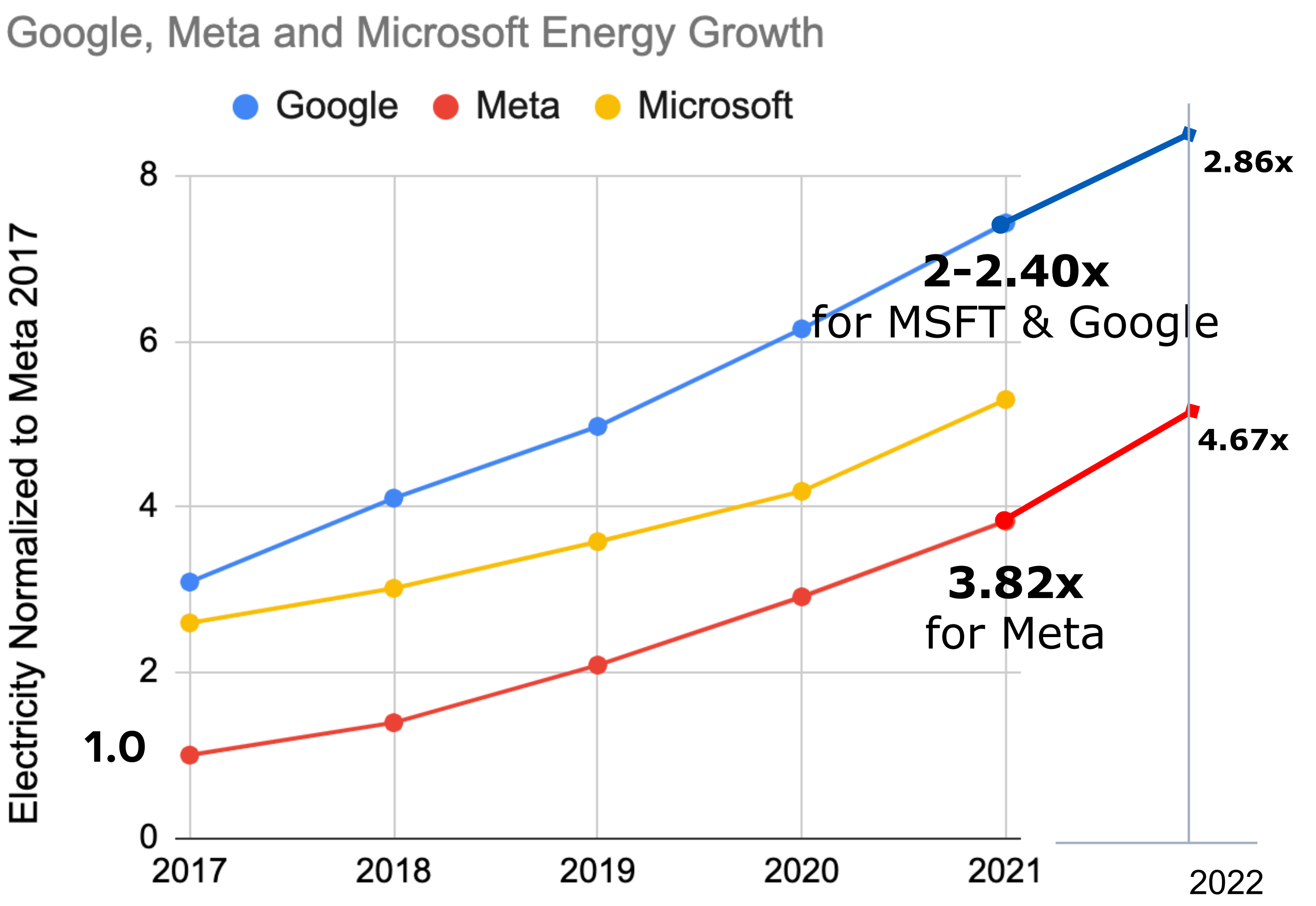
Computing's Energy Footprint

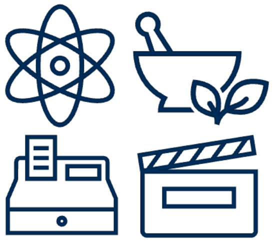
700 million tons of CO_2e

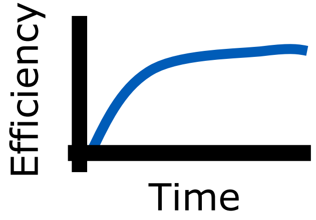


Half of the aviation industry's emissions

Computing’s Energy Footprint







Computing’s Footprint Projected to Double over the Decade

Corporate Climate Pledges

Google The Keyword

Latest stories Product updates Company news

A MESSAGE FROM OUR CEO

Our third decade of climate action: Realizing a carbon-free future

Microsoft

Official Microsoft Blog Microsoft On the Issues The AI Blog Transform

Microsoft will be carbon negative by 2030

Jan 16, 2020 | Brad Smith - President

Amazon

Sustainability in the Cloud

Amazon Web Services (AWS) is committed to running our business in the most environmentally friendly way possible and achieving 100% renewable energy usage for our global infrastructure.

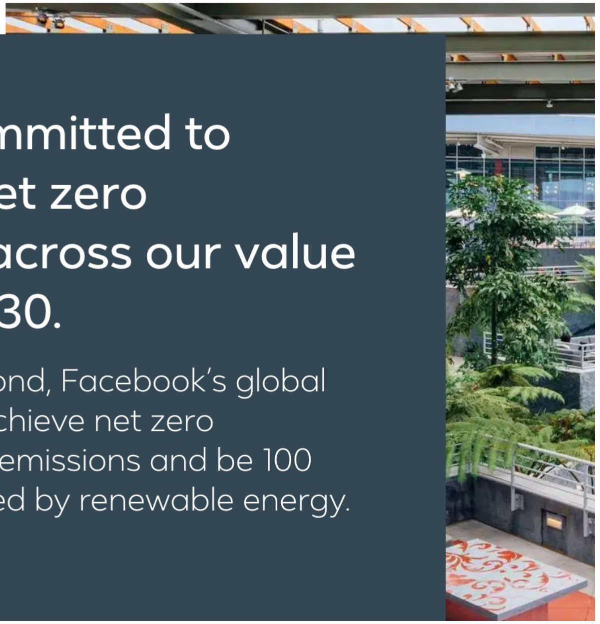
https://sustainability.aboutamazon.com/environment/the-cloud

Meta Sustainability

Innovation for our world Collaboration for good

We are committed to reaching net zero emissions across our value chain in 2030.

In 2020 and beyond, Facebook's global operations will achieve net zero greenhouse gas emissions and be 100 percent supported by renewable energy.



PRESS RELEASE
July 21, 2020

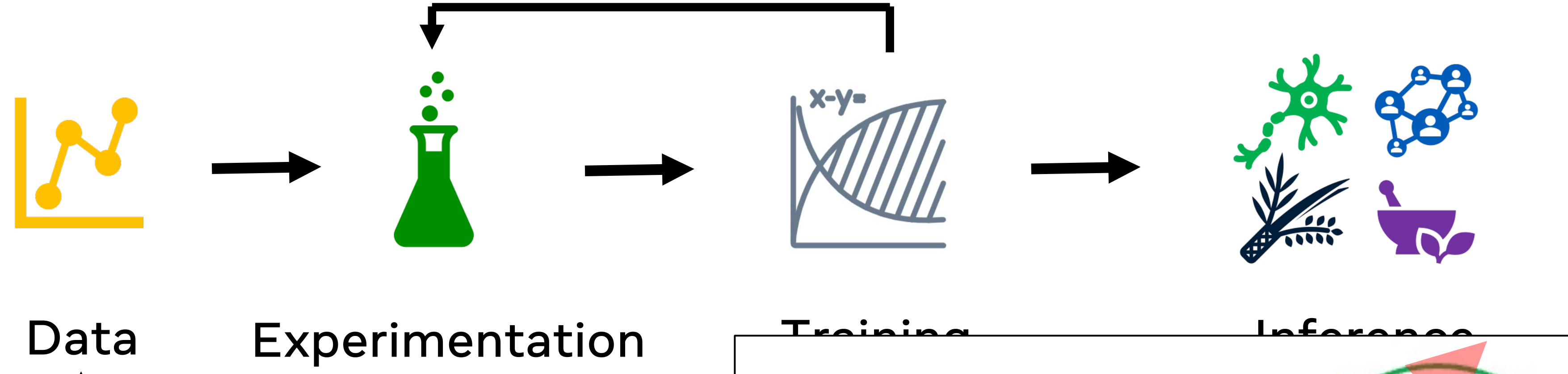
Apple commits to be 100 percent carbon neutral for its supply chain and products by 2030

Outline

- Introduction
- Landscape of AI
- Future of AI: A Sustainable Development Cycle

Exponential Growth Trend of AI

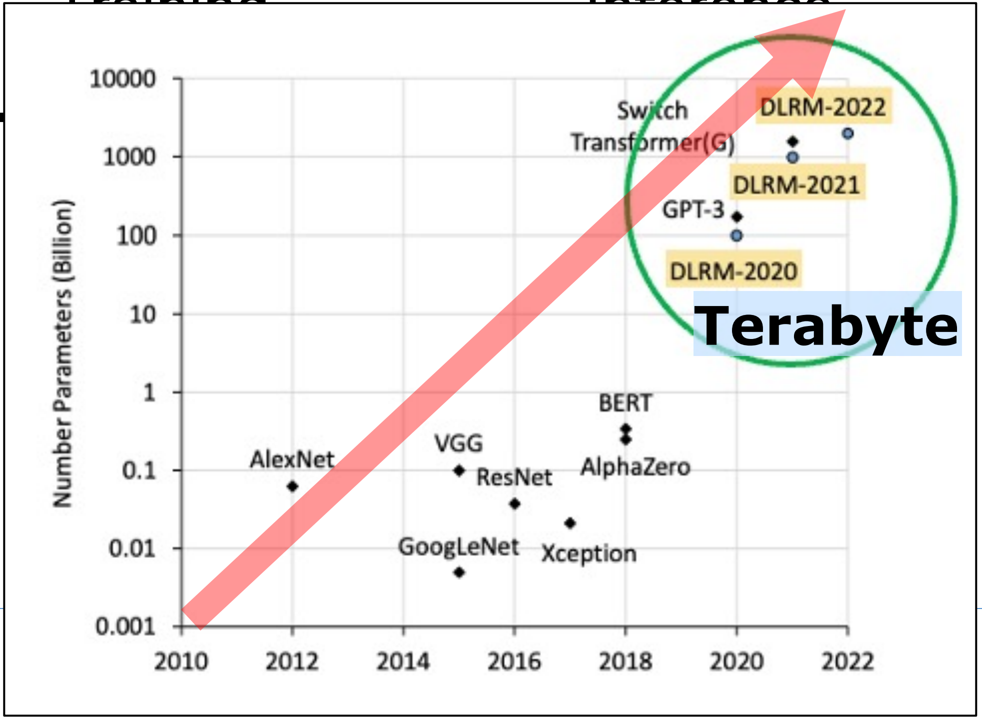
Data, Model Sizes, System Infrastructures



Production Data Size
2+ times / 2019-21

Recommendation Model Size
20+ times / 2019-21

Training and Inference Infrastructure
2+ times / 2019-21



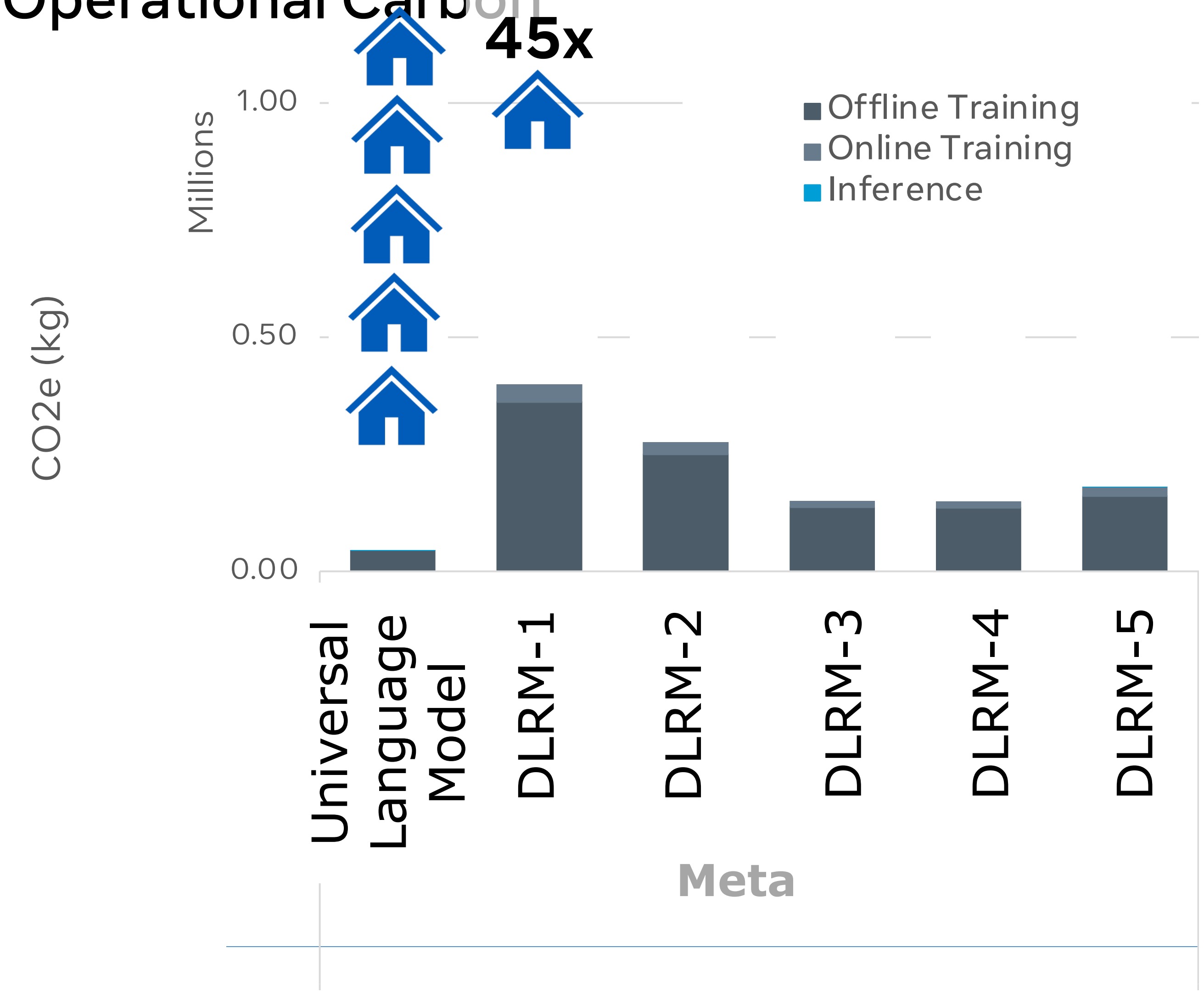
AI's Carbon Footprint

Operational Carbon

Operational tCO2e =
training/inference time *
# of processors *
power consumption per processor *
PUE *
kg CO2e per KWh

AI's Carbon Footprint

Operational Carbon



Universal Language Model Training

≈5 Home's Annual

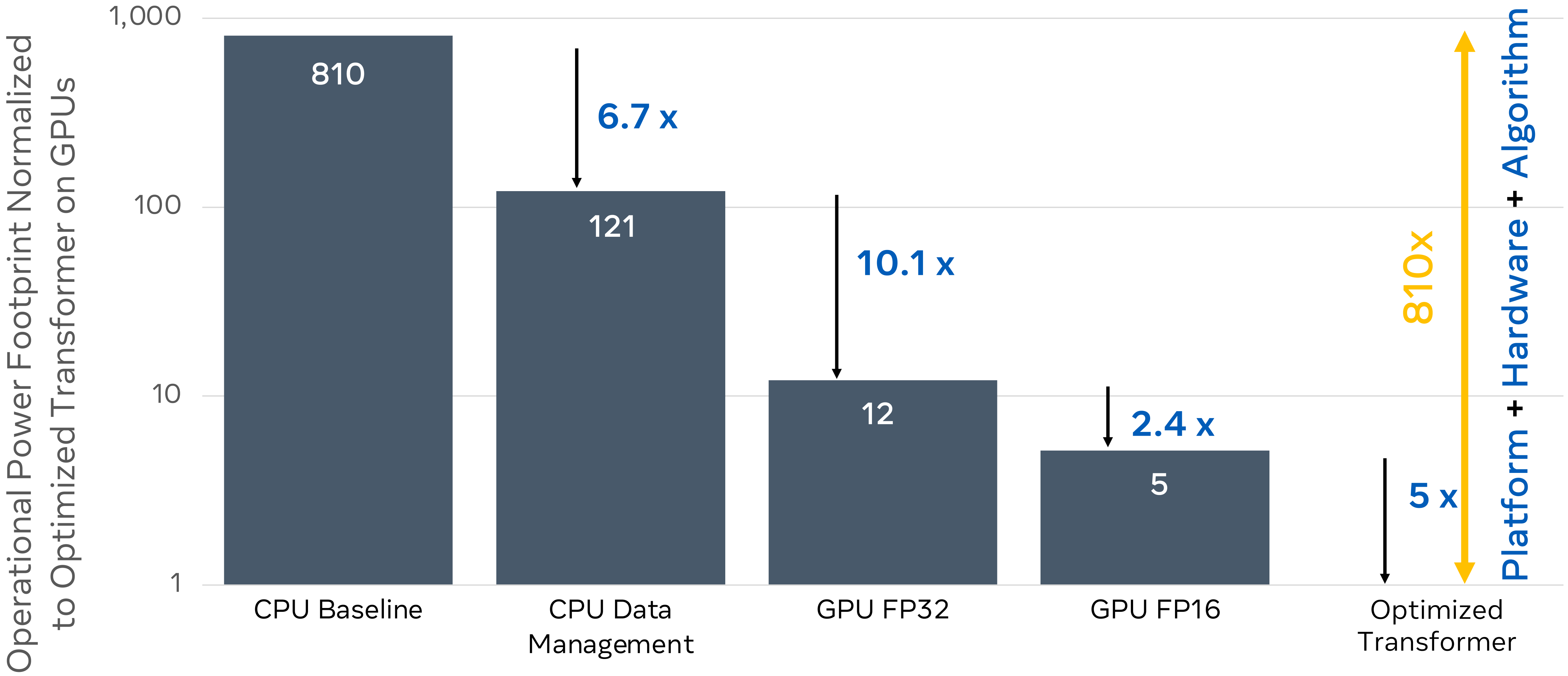
Recommendation Model Training

≈45 Home's Annual

Sustainable AI: Environmental Implications, Challenges and Opportunities. Wu et al. MLSys-2022.

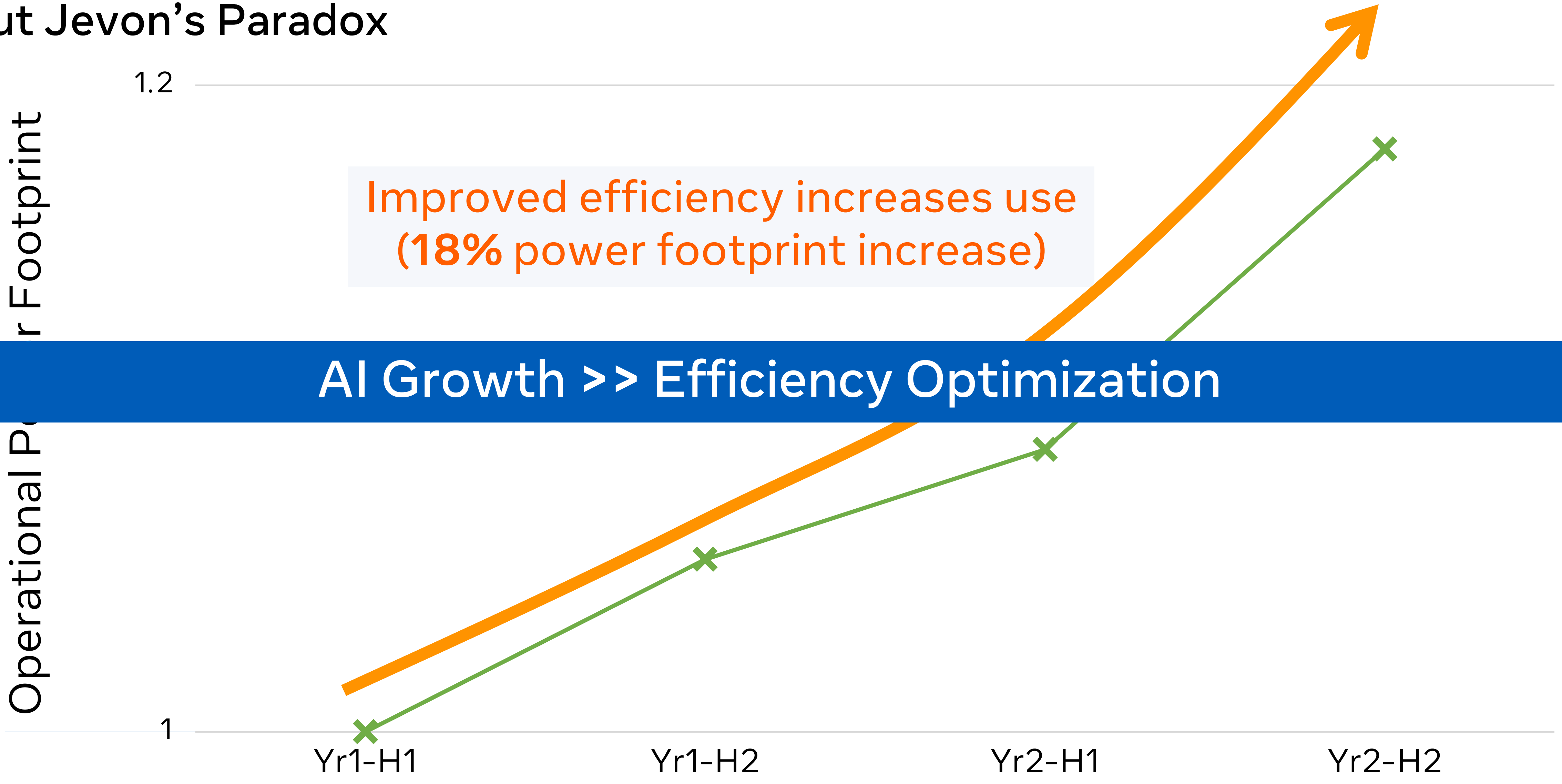
Carbon Optimization via HW-SW Co-Design

Universal Language Translation



Efficiency Optimization 👍

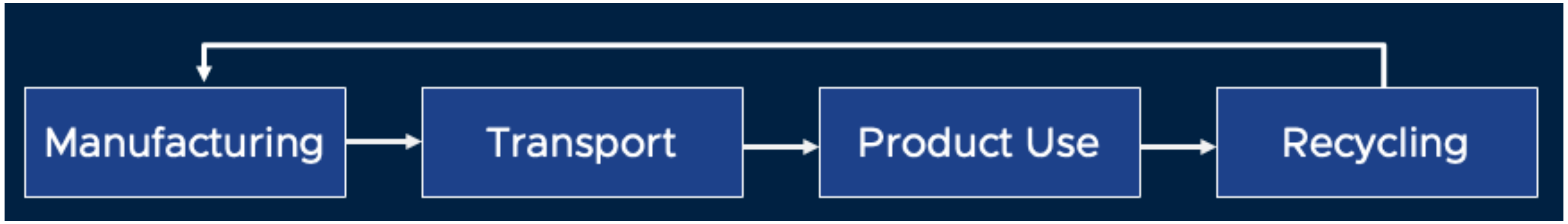
But Jevon's Paradox



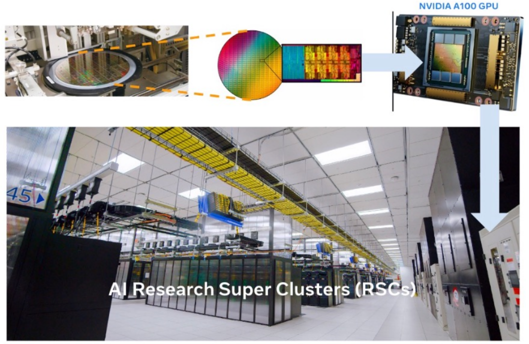
AI Growth >> Efficiency Optimization

Lifecycle Carbon Emissions

Embodied CO_2 | Operational CO_2



Emissions from fabs building chips



Emissions of system use (Software and Hardware)

MAGAZINE · SEMICONDUCTORS

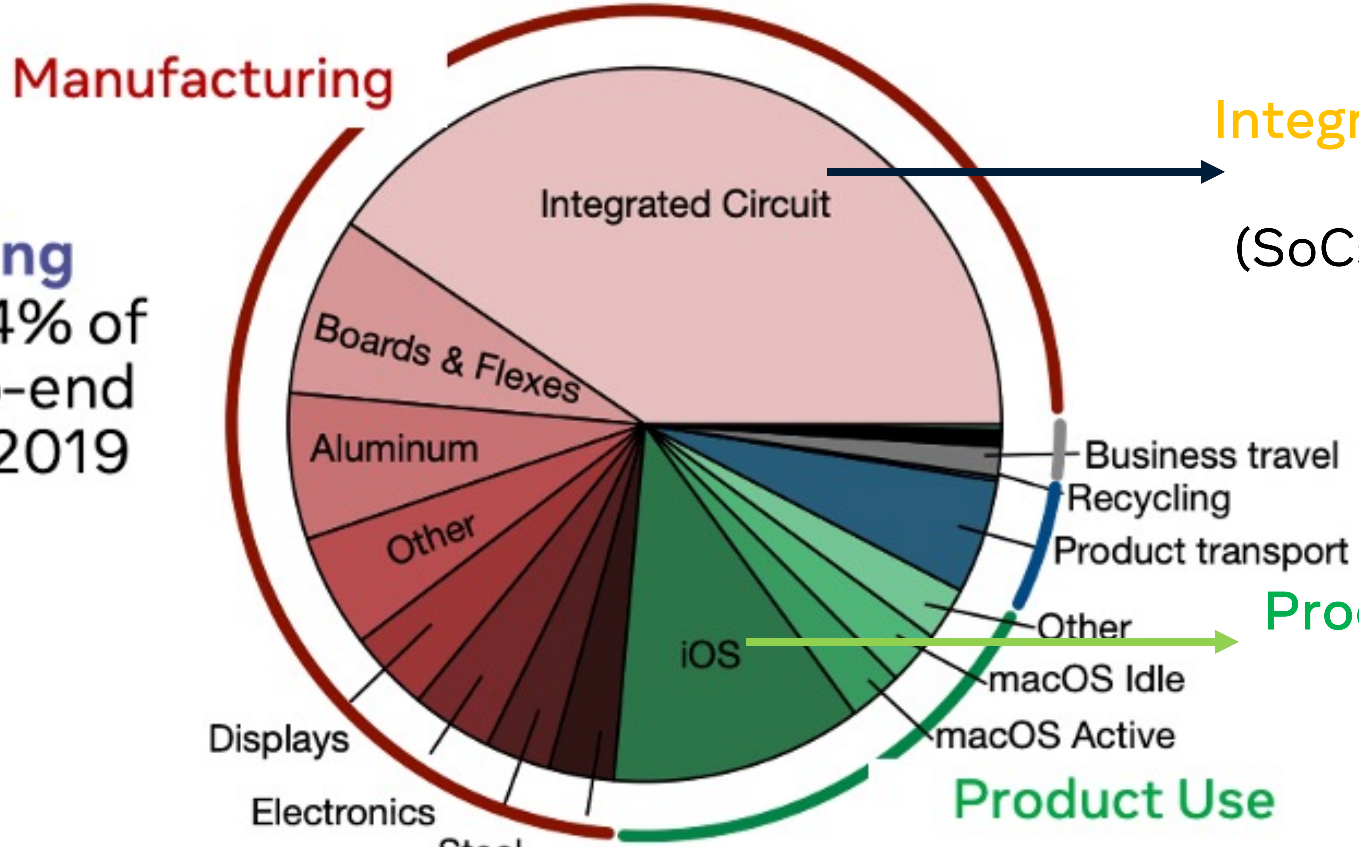
The chip industry's dirty little secret: It's very dirty

BY MICHAL LEV-RAM
January 29, 2024 at 6:00 AM EST

An Under-Explored Aspect of Computing's Carbon Footprint

Embodied Carbon

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



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

Product use account for **19%** of emissions

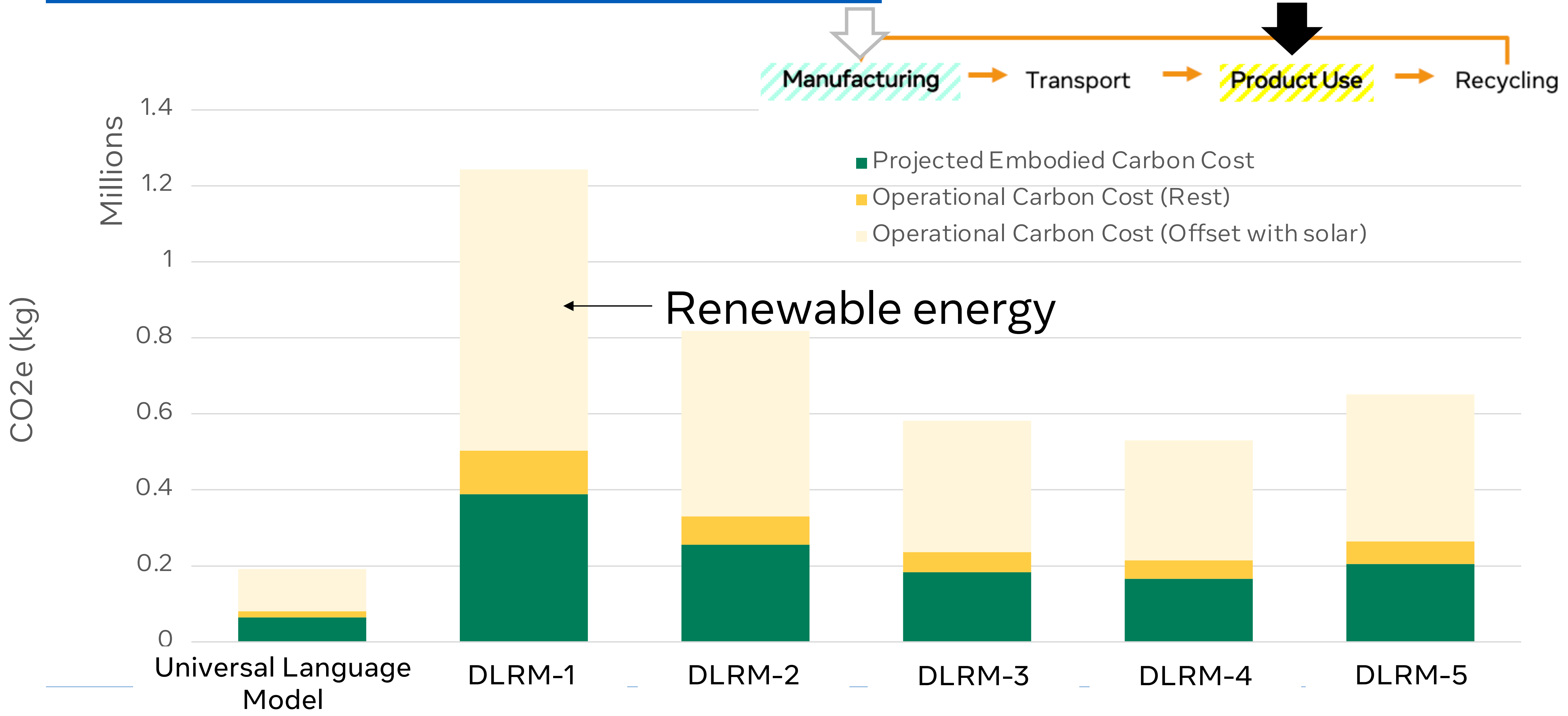
Manufacturing and operational carbon footprint (location-based) is roughly equal for cloud infrastructure.

AI's Carbon Footprint

Embodied Carbon

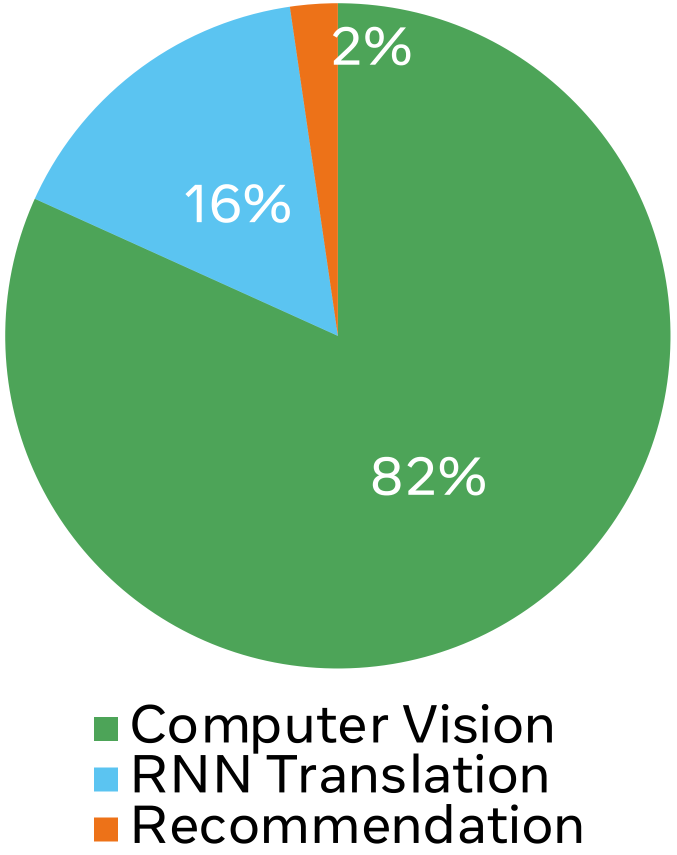
Embodied tCO2e $= \sum_{i=1}^{xPUs,DRAM,SSD,HDD} \frac{Time_{application}}{Lifetime_{hardware}} \boxed{CO_2^{embodied}(i)}$

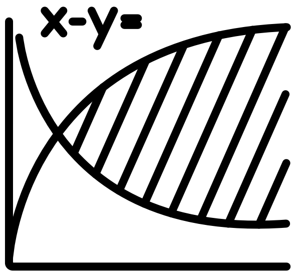
AI's (Operational & Embodied) Carbon Footprint



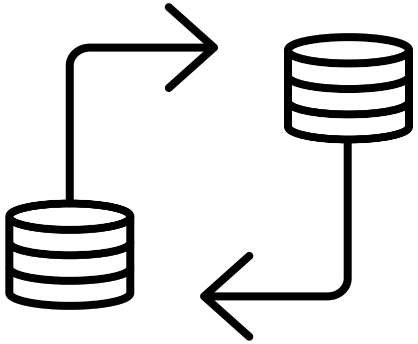
Sustainable AI: Environmental Implications, Challenges and Opportunities. Wu et al. MLSys-2022.

Open Research Fosters Innovations





Representative Benchmarks



Datasets

CV: **ImageNet**

NLP: **LibriSpeech**

Recommendation?

MLPerf includes DLRM + Criteo Ads Dataset

MLPerf

A machine learning performance benchmark suite with broad industry and academic support

MLPerf Includes DLRM + Criteo Ads Dataset

Recommendation Benchmark Advisory Board

Recommendation Model

- Cover a diverse set of use cases with the goal to optimize for both *click-through-rate* and *conversion-rate*, as well as to improve *long-term values*

Recommendation Datasets

- *Capture the degree of sparsity found in industry-scale problems*
- *Cover user- and item-features as well as user-item interactions*







DEVELOPING A RECOMMENDATION BENCHMARK FOR MLPERF TRAINING AND INFERENCE

Carole-Jean Wu [1] Robin Burke [2] Ed H. Chi [3] Joseph Konstan [4] Julian McAuley [5] Yves Raimond [6] Hao Zhang [7]

1 INTRODUCTION

Deep learning-based recommendation models are used pervasively and broadly, for example, to recommend movies, products, or other information most relevant to users, in order to enhance the user experience. Among various application domains which have received significant industry and academia research attention, such as image classification, object detection, language and speech translation, the performance of deep learning-based recommendation models is less well explored, even though recommendation tasks unarguably represent significant AI inference cycles at large-scale datacenter fleets (Jouppi et al., 2017; Wu et al., 2019a; Gupta et al., 2019).

To advance the state of understanding and enable machine learning system development and optimization for the e-commerce domain, we aim to define an industry-relevant recommendation benchmark for the MLPerf Training and Inference suites. We will refine the recommendation benchmark specification annually to stay up to date to the current academic and industrial landscape. The benchmark will reflect standard practice to help customers choose among hardware solutions today, while also being forward looking enough to drive development of hardware for the future.

The goal of this white paper is twofold:

- We present the desirable modeling strategies for personalized recommendation systems. We lay out desirable characteristics of recommendation model architectures and data sets.
- We then summarize the discussions and advice from the MLPerf Recommendation Advisory Board.

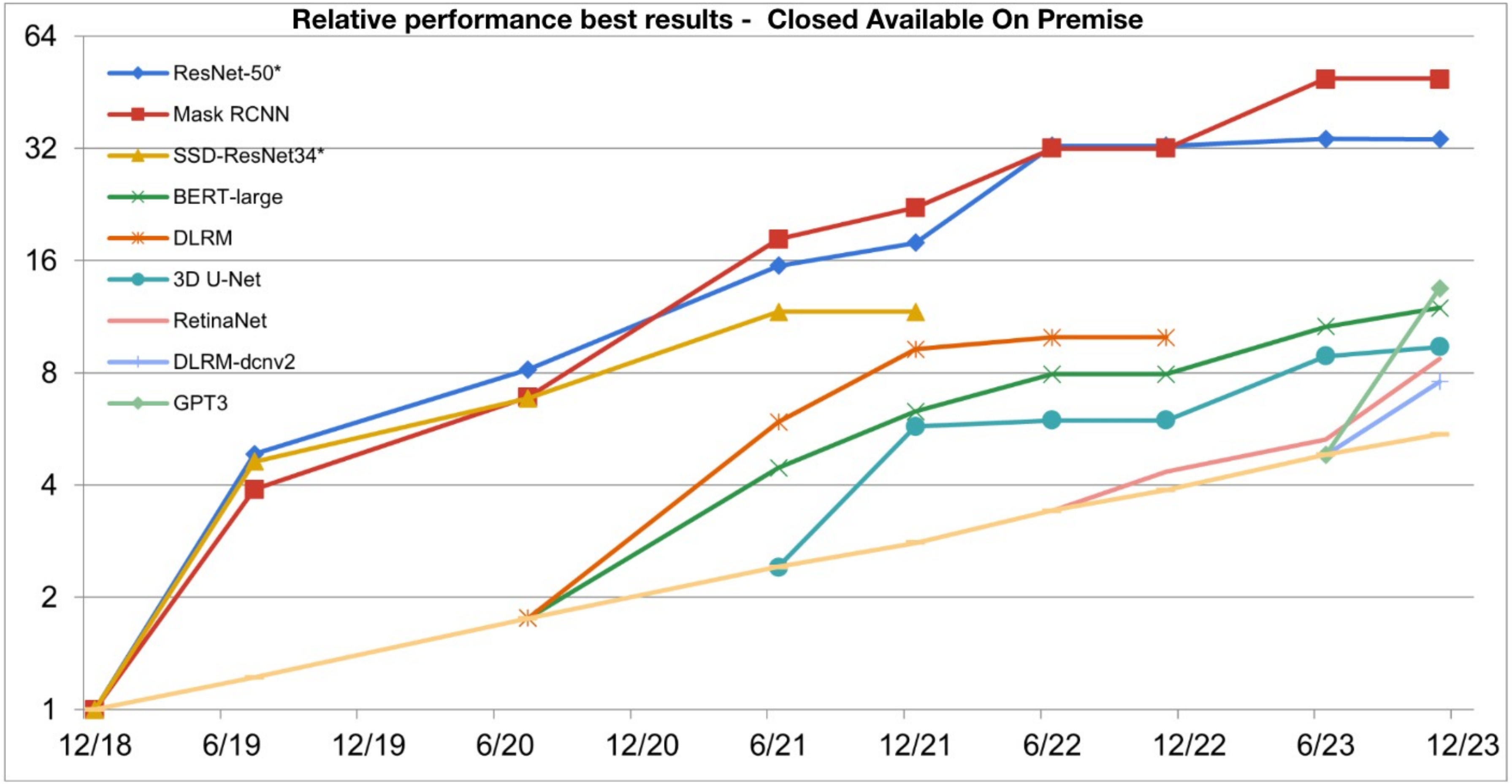
Desirable characteristics for ideal recommendation benchmark models should represent a diverse set of use cases, covering a long tail. For example, most recommendation tasks with large candidate sets have both a candidate generation model and a ranking model working together. The candidate generation model tends to be latency-sensitive with a dot-product or softmax on top, while a ranking model tends to have a lot of interactions being considered. The end-to-end model should ideally produce predictions for both *click-through rate* and *conversion rate*. *To enable a representative coverage of the recommendation task diversity and different scales of recommendation tasks (that are often dependent on the scale of the available data), wed want to consider recommendation benchmarks of different scales.*

Recommendation models are tasked to produce novel, non-obvious, diverse recommendations. This is really at the heart of the recommendation problem – we learn from patterns in the data that generalize to the tail items, even if the items only occur a few times, despite the temporal changes in the data sets. Thus, from the system development and optimization perspective, *even though less-frequently indexed items can consume significant memory capacity in a system and it can be challenging to select an optimizer to determine meaningful weights for the embedding entries in a few epochs, we must retain all user and item categories in a feature to capture representative system requirement.*

Many enhancement techniques have been explored to improve recommendation prediction quality. For example, variations of RNNs (e.g. attention layers, Transformer/LSTM styles) are under active investigation for at-scale industrial practice. It is not clear yet how to best exploit the temporal sequence in DNN-based recommendation models. In addition, dense-matrix multiplication with very sparse vectors is an interesting case as well. This could be thought of as embeddings where input vectors are not just indices but also carry numerical value, to, say, be multiplied with the corresponding embedding row. *We should keep an eye on the development of the aforementioned enhancement techniques and refine the recommendation model architecture when it is proven to improve inference quality for practical use cases.*

[1]Facebook/ASU [2]University of Colorado, Boulder [3]Google Research [4]University of Minnesota [5]University of California, San Diego [6]Netflix [7]Facebook. Send correspondence to *carole-jeanwu@fb.com*

A ML System Performance Benchmark Suite on Speed



2.8X performance gains in **5 months** for LLM benchmark!

ML Commons

Benchmarks | Datasets | Working Groups | Research | About Us | Blog | Join Us

MLPerf Training

The MLPerf Training benchmark suite measures how fast systems can train models to a target quality metric.

Learn more →

MLPerf Training: HPC

The MLPerf HPC benchmark suite measures how fast systems can train models to a target quality metric.

Learn more →

MLPerf Inference: Datacenter

The MLPerf Inference: Datacenter benchmark suite measures how fast systems can process inputs and produce results using a trained model.

Learn more →

MLPerf Inference: Edge

The MLPerf Edge benchmark suite measures how fast systems

MLPerf Inferen

The MLPerf Mobile benchmark suite measures how fast systems can process inputs a

Learn more →

MLPerf Inferen

The MLPerf Tiny benchmark suite measures how fast systems can process inputs a

Learn more →

MLPerf Storag

The MLPerf Storage benchmark suite measures how fast storage systems can supply t

Learn more →

AI Safety

Data

- Datasets
- Best Practices
- Medical
- Croissant

Research

- Algorithms
- Data-centric ML
- Chakra
- Science

Outline

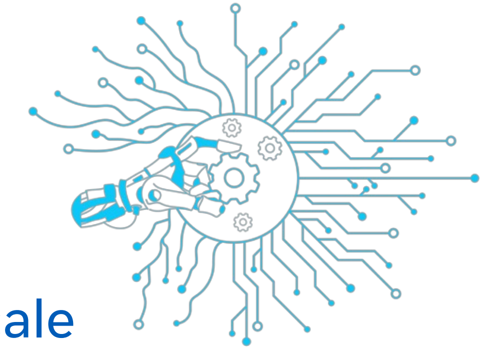
- Introduction
- Landscape of AI and Its Carbon Footprint
- Future of AI: A Sustainable Development Cycle

2000s

Computer Architecture

2020s

Domain-Specific Architecture



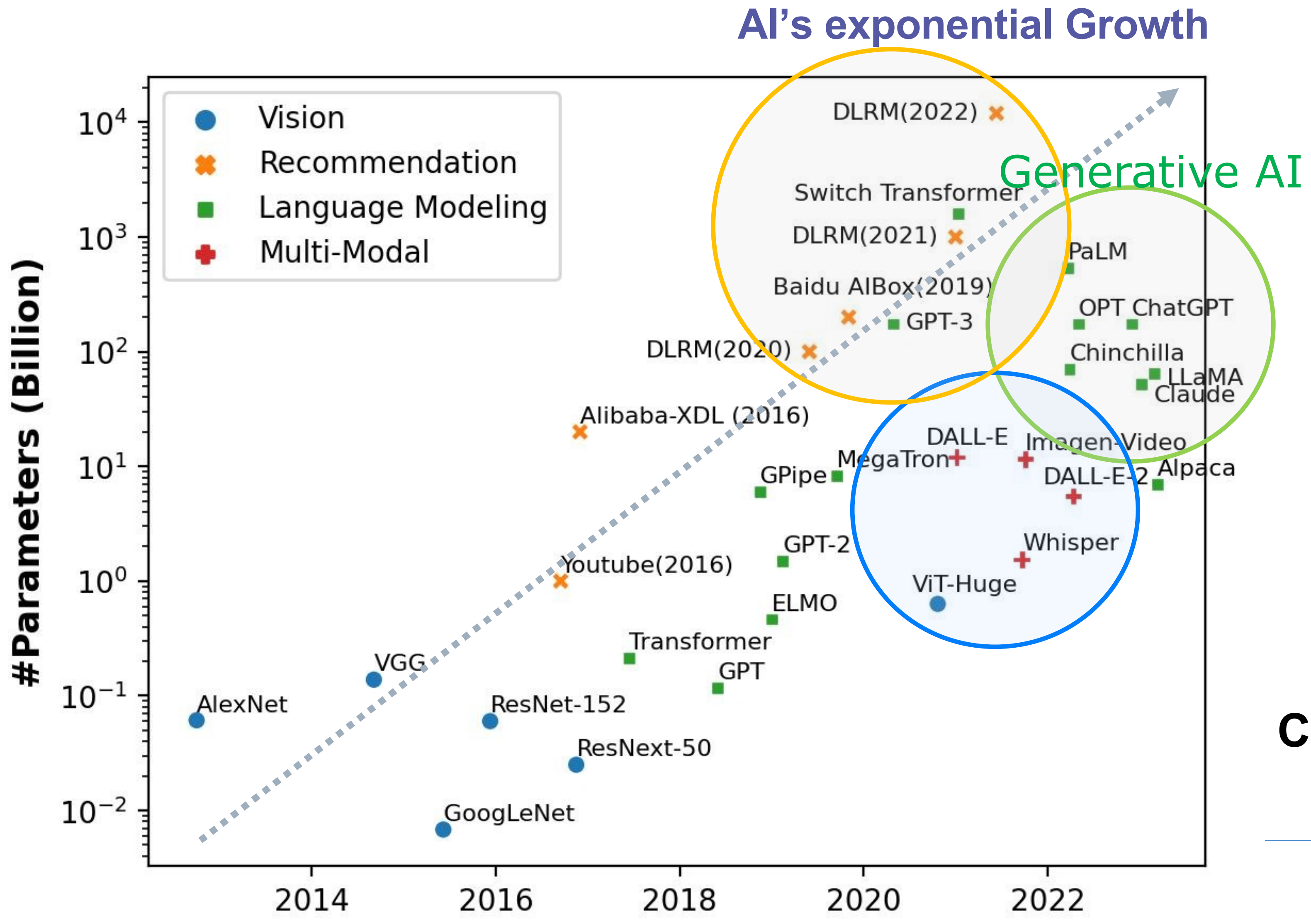
2040+?

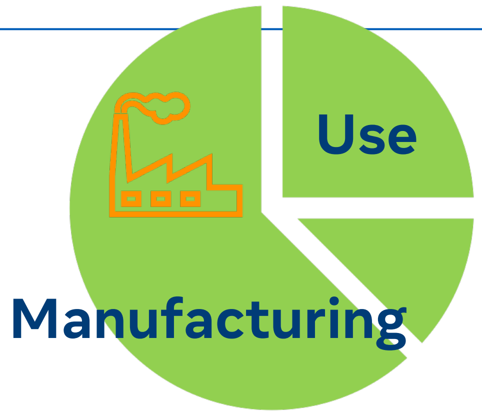
Moore's Law Scaling

Dennard Scaling

Efficiency @ Scale

Scaling Limit of AI?



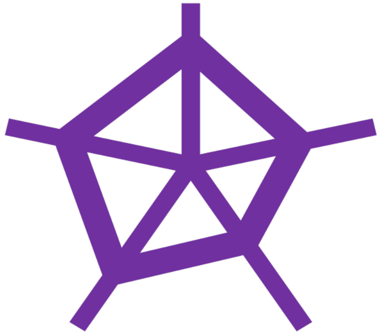


Holistic Lifecycle Approaches

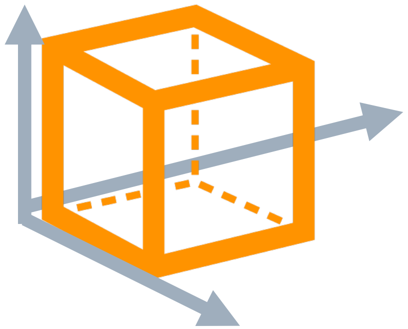
Efficiency Optimization
Data, model, system hardware
Infrastructure at-scale

Carbon Efficiency Optimization
Embodied vs. Operational CO_2

Scaling Computing Sustainably: Paths Forward



Metrics & Accounting



AI Design & Optimization Space with CO_2

Sustainable Development

MLPerf & OCP Standard

ACT

https://github.com/facebookresearch/ACT

Carbon Explorer

https://github.com/facebookresearch/CarbonExplorer

Cross-Stack System Design

Programming Language

Runtime Management

System Architecture

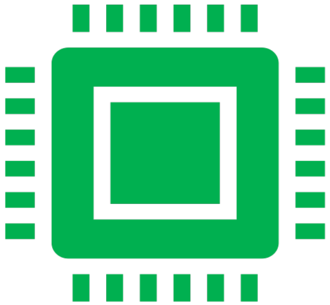
IC Hardware Design

Semiconductor Manufacturing

Computing & Sustainability

Circular Economy

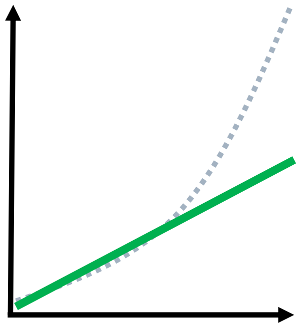
Scaling AI and Computing Sustainably



Environmentally Sustainable Systems

ACT
[Gupta et al.; ISCA 2022]

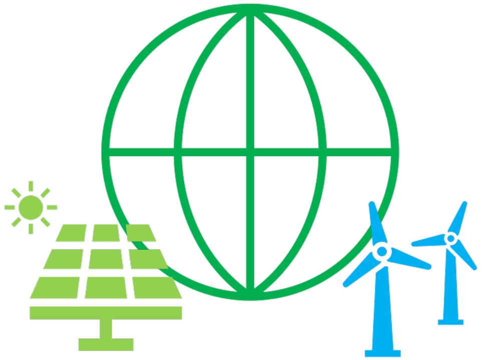
Carbon-Efficient XR Systems
[Elgamal et al.; arXiv 2023]



Carbon-Efficient AI Data/Models/Algorithms

TT-Rec
[Ying et al.; MLSys 2021]

Carbon-Efficient AI Models
[Gupta et al.; ICLR Climate Change AI 2023]



Optimization at Scale

Carbon Explorer
[Acun et al.; ASPLOS 2023]

AI Anytime Anywhere

AutoScale / AutoFL
[Kim et al.; MICRO 2020]

GreenScale
[Kim et al.; arXiv 2023]

Computing & Sustainability

Read more about Scaling AI Computing Sustainably

Sustainable AI: Environmental Implications, Challenges and Opportunities

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

Facebook AI

***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, considering the life cycle of system hardware. Taking a step further, we capture the operational and manufacturing carbon footprint of AI computing and present an end-to-end analysis for *what* and *how* hardware-software design and at-scale optimization can help reduce the overall carbon footprint of AI. Based on the industry experience and lessons learned, we share the key challenges and chart out important development directions across the many dimensions of AI. We hope the key messages and insights presented in this paper can inspire the community to advance the field of AI in an environmentally-responsible manner.**

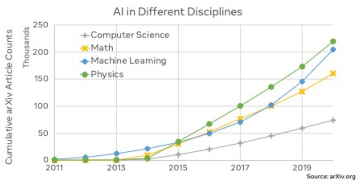


Fig. 1. The growth of ML is exceeding that of many other scientific disciplines. Significant research growth in machine learning is observed in recent years as illustrated by the increasing cumulative number of papers published in machine learning with respect to other scientific disciplines based on the monthly count (y-axis measures the cumulative number of articles on arXiv).

I. Introduction

Artificial Intelligence (AI) is one of the fastest growing domains spanning research and product development and significant investment in AI is taking place across nearly every industry, policy, and academic research. This investment in AI has also stimulated novel applications in domains such as science, medicine, finance, and education. Figure 1 analyzes the number of papers published within the scientific disciplines, illustrating the growth trend in recent years[1].

AI plays an instrumental role to push the boundaries of knowledge and sparks novel, more efficient approaches to conventional tasks. AI is applied to predict protein structures radically better than previous methods. It has the potential to revolutionize biological sciences by providing in-silico methods for tasks only possible in a physical laboratory setting [1]. AI is demonstrated to achieve human-level conversation tasks, such as the Blender Bot [2], and play games at superhuman levels, such as AlphaZero [3]. AI is used to discover new electrocatalysts for efficient and scalable ways to store and utilize renewable energy [4], predicting renewable energy availability in advance to improve energy utilization [5], operating hyperscale data centers efficiently [6], growing plants using less natural resources [7], and, at the same time, being used to tackle climate changes [8], [9]. It is projected that, in the next five years, the market for AI will increase by 10× into hundreds of billions of dollars [10]. All of these investments in research, development, and deployment have led to a super-linear growth in AI data, models, and infrastructure capacity. With the dramatic growth of AI, it is imperative to understand the environmental implications, challenges, and opportunities of this nascent technology. This is because technologies tend to create a self-accelerating growth cycle, putting new demands on the environment.

This work explores the environmental impact of AI from a *holistic* perspective. More specifically, we present the challenges and opportunities to designing sustainable AI computing across the key phases of the machine learning (ML) development process — *Data*, *Experimentation*, *Training*, and *Inference* — for a variety of AI use cases at Facebook, such as vision, language, speech, recommendation and ranking. The solution space spans across our fleet of datacenters and on-device computing. Given particular use cases, we consider the impact of AI *data*, *algorithms*, and *system hardware*. Finally, we consider emissions across the life cycle of hardware systems, from manufacturing to operational use.

AI Data Growth. In the past decade, we have seen an exponential increase in AI training data and model capacity. Figure 2(b) illustrates that the amount of training data at Facebook for two recommendation use cases — one of the fastest growing areas of ML usage at Facebook— has increased by 2.4× and 1.9× in the last two years, reaching exabyte scale. The increase in data size has led to a 3.2× increase in data ingestion bandwidth demand. Given this increase, data storage and the ingestion pipeline accounts for a significant portion of

[1]Based on monthly counts, Figure 1 estimates the cumulative number of papers published per category on the arXiv database.

Think Globally, Design Deliberately: Taking an Inclusive Approach to Innovation.
Bobbie Manne, Carole-Jean Wu, Partha Ranganathan, Sarah Bird, Shane Greenstein. ACM SIGARCH Blog 2021.

Chasing Carbon: The Elusive Environmental Footprint of Computing.
Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin Lee, Gu-Yeon Wei, David Brooks, Carole-Jean Wu. HPCA-2021.

Carbon Dependences in Datacenter Design and Management.
Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Aditya Sundarrajan, Kiwan Maeng, Manoj Chakkaravarthy, David Brooks, Carole-Jean Wu. HotCarbon-2022.

ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool.
Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin Lee, David Brooks, Carole-Jean Wu. ISCA-2022.

Carbon Explorer: A Holistic Approach for Designing Carbon Aware Datacenters.
Bilge Acun, Benjamin Lee, Kiwan Maeng, Manoj Chakkaravarthy, Udit Gupta, David Brooks, Carole-Jean Wu. ASPLOS-2023.

Work Done by Many