This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3163226, IEEE Micro

Chasing Carbon: The Elusive Environmental Footprint of Computing

Udit Gupta[1,2], Young Geun Kim[3], Sylvia Lee[2], Jordan Tse[2],
Hsien-Hsin S. Lee[2], Gu-Yeon Wei[1], David Brooks[1], Carole-Jean Wu[2]

[1]Harvard University, [2]Facebook Inc., [3]Arizona State University

ugupta@g.harvard.edu carolejeanwu@fb.com

***Abstract*—Given recent algorithm, software, and hardware innovation, computing has enabled a plethora of new applications. As computing becomes increasingly ubiquitous, however, so does its environmental impact. This paper brings the issue to the attention of computer-systems researchers. Our analysis, built on industry-reported characterization, quantifies the environmental effects of computing in terms of carbon emissions. Broadly, carbon emissions have two sources: operational energy consumption, and hardware manufacturing and infrastructure. Although carbon emissions from the former are decreasing thanks to algorithmic, software, and hardware innovations that boost performance and power efficiency, the overall carbon footprint of computer systems continues to grow. This work quantifies the carbon output of computer systems to show that most emissions related to modern mobile and data-center equipment come from hardware manufacturing and infrastructure. We therefore outline future directions for minimizing the environmental impact of computing systems.**

I. Introduction

The world has seen a dramatic advancement of information and communication technology (ICT) in the last two decades. The rise in ICT has resulted in a proliferation of consumer devices, networking technologies, and data centers. Despite its myriad societal benefits, ICT has incurred tremendous environmental footprint. ICT accounted for up to 3% of the global energy demand as of 2015. In fact, data centers alone accounted for 1% of the global demand, eclipsing the total energy consumption of many nations. By 2030, ICT is projected to account for 7% of the global energy demand. Anticipating the ubiquity of computing, researchers and developers must enable the design and deployment of sustainable computer systems.

To curb the growing energy demand of computing technology, software and hardware researchers have invested heavily in maximizing the energy efficiency of systems and workloads. For instance, between the late twentieth and early twenty-first centuries, Moore's Law has enabled fabrication of systems that have billions of transistors and 1,000× higher energy efficiency. For salient applications, such as AI, molecular dynamics, video encoding, and cryptography, systems now comprise specialized hardware accelerators that provide orders-of-magnitude higher performance and energy efficiency. Moreover, data centers have become more efficient by consolidating equipment into large, warehouse-scale systems and by reducing cooling and facility overhead to operate at near optimal power usage effectiveness (PUE of 1.1).

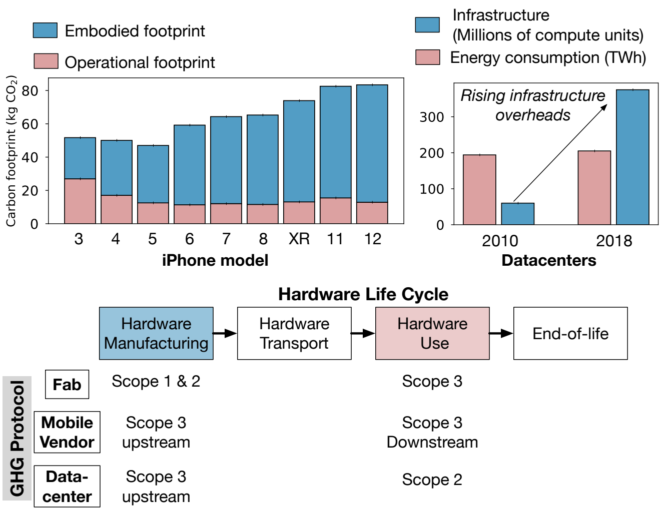

Fig. 1. The majority of computing's carbon footprint comes from embodied emissions not operational energy use. (Top) From iPhone 3 to iPhone 12 , the operational footprint reduced by 2.09× given efficiency optimizations while the embodied emissions rose by 2.85× due to higher manufacturing overheads. Between 2010 and 2018, the operational energy consumption of datacenters increased by only 6% collectively while the number of compute instances, i.e., infrastructure overheads, increased by 6×. (Bottom) Moving forward, enabling sustainable computing requires optimizing emissions across hardware life cycles. Each life cycle phase corresponds to different categories of the GHG protocol for fabs, mobile vendors, and datacenter operators.

Given the advancements in energy efficiency, this paper shows computer system and architecture researchers must go beyond energy and consider the carbon footprint of platforms end-to-end. Similar to infrastructure efficiency optimization targeting operational expenditures (opex, recurring operations) and capital expenditures (capex, one-time infrastructure and hardware), we can categorize carbon emissions into opex- and capex-related activities. We define opex-related emissions as emissions from hardware use and energy consumption (**operational footprint**) and capex-related emissions as emissions from facility-infrastructure construction and chip manufacturing (**embodied footprint**), such as, procuring raw materials, fabrication, packaging, and assembly. Figure 1 (left) shows that, between iPhone 3 (2009) and iPhone 11 (2019), the operational footprint has decreased due to improved energy efficiency while embodied footprint has increased due to more sophisticated hardware. For client devices (e.g., iPhones) the dominating source of carbon footprint has shifted from being opex-related to capex-related. In contrast, Figure 1 (right) shows the carbon footprint breakdown for data centers. At the data-center scale we find operational emissions between 2010 and 2018 have increased by 6% while infrastructure capacity, which correspond to embodied hardware emissions, has increased by 6× [1].

If left unchecked, we anticipate the gap between opex- and capex-related carbon output will widen in coming years. As energy efficiency rises along with the use of renewable energy, opex-related emissions will become a less significant part of computing's environmental impact. Increasing application demand will exacerbate capex-related emissions, however. In less than two years, Facebook hardware devoted to AI training and inference has grown by 4× and 3.5×, respectively. Likewise, to support emerging applications (e.g., AI and AR/VR) on mobile devices, smartphones today have more transistors and specialized circuits than their predecessors; limited by dark silicon [2], the additional hardware exacerbates capex-related carbon footprints. Addressing both opex- and capex-related emissions requires fundamentally rethinking designs across the entire computing stack.

This paper takes a data-driven approach to studying the carbon breakdown of hardware life cycle—including manufacturing, transport, use, and recycling—for consumer devices and data-center systems. It serves as a *call-to-action* for computer system and architecture researchers to tackle computing's growing environmental crisis by laying the foundation for understanding and creating more-sustainable designs. First, we present the state of industry practice using the Greenhouse Gas (GHG) Protocol to quantify the environmental impact of commercial mobile and data-center supply chains (Section II). On the basis of publicly available industry sustainability reports from, we show that the hardware-manufacturing process, rather than system operation, is the primary source of carbon emissions (Section III and IV). Moreover, we demonstrate that the use of renewable energy to power fabs is no panacea; hardware manufacturing and infrastructure-related activities will continue to dominate the carbon output (Section V). Finally, we outline future research and design directions across the computing stack to realize environmentally sustainable systems (Section VI).

The important contributions of this work are:

1) We show that given the considerable efforts over the past two decades to improve energy efficiency, the Amdahl's law bottleneck for computing's carbon output has shifted from operational activities to capex-related activities such as hardware manufacturing and system infrastructure.
2) We take a data-driven approach to quantify computing's end-to-end carbon footprint in mobile and data-center scale systems. For instance, the fraction of life-cycle carbon emissions due to hardware manufacturing increased from 49% for the iPhone 3GS to 86% for the iPhone 11 Similarly, we show that because an increasing fraction of warehouse-scale data centers employ renewable energy (e.g., solar and wind), data-center carbon output is also shifting from operation to hardware design/manufacturing and infrastructure construction. In 2019, capex- and supply-chain-related activities accounted for 23× more carbon emissions than opex-related activities at Facebook.
3) We chart future paths for software and hardware researchers to characterize and minimize computing technology's environmental impact.

II. Quantifying environmental impact

This section details the state-of-the-art industrial practices for quantifying carbon emissions. Our discussion presents carbon-footprint-accounting methods that serve broadly across technology companies, including AMD, Apple, Facebook, Google, Huawei, Intel, Microsoft, and TSMC [3]–[8]. First we review methods for analyzing organization-level emissions. Next, we analyze how to use the results of such analyses across the technology supply chain to develop models for individual computer systems at the data-center and mobile scale.

A. Industry-level carbon-emission analysis

A common method accounting standard for quantifying organization-level carbon output is the GHG Protocol. Many technology companies, including AMD, Apple, Facebook, Google, Huawei, Intel, and Microsoft publish annual sustainability reports using the GHG Protocol, which categorizes emissions into Scope 1 (direct emissions), Scope 2 (indirect emissions), and Scope 3 (upstream and downstream supply-chain emissions). Figure 1(bottom) summarizes the salient emissions from each category for chip manufacturers, mobile vendors, and data-center operators [9].

Scope 1 emissions come from fuel combustion, refrigerants in offices and data centers, transportation, and the use of chemicals and gases in semiconductor manufacturing. Although Scope 1 accounts for a small fraction of emissions for mobile-device vendors and data-center operators, it comprises over half the operational carbon output from chip manufacturers including Global Foundries, Intel, and TSMC [3], [4], [10]. Much of these emissions come from burning perfluorocarbons, chemicals, and gases.

Scope 2 emissions come from purchased energy powering semiconductor fabs, offices, and data-centers. They depend on two parameters: the amount of energy consumed and the GHG footprint from generating the consumed energy, i.e., grams of CO_2 emitted per kilowatt-hour of energy. Compared with "brown" energy (e.g., coal, gas), "green" energy (e.g., solar, wind, nuclear) produces up to 30× fewer GHG emissions. Scope 2 emissions are especially important in fabs and data centers.

Fabs need copious energy to manufacture chips. Energy consumption produces over 63% of the emissions from manufacturing 12-inch wafers at TSMC [10]. Fab energy consumption

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3163226, IEEE Micro

is expected to rise with next-generation semiconductor fabs and more advanced process technology nodes.

Scope 2 emissions are also crucially important for data centers whose operational footprint depends on the overall energy consumption from thousands of servers. To reduce their operational footprint, modern warehouse-scale data centers are procuring higher amounts of renewable energy.

Scope 3 emissions come from all other activities, including the full upstream and downstream supply chain. They often comprise business travel, logistics, and capital goods. For technology companies, a crucial and challenging aspect of Scope 3 analysis is accounting for hardware bought and sold. Data centers, for instance, contain thousands of server-class CPUs whose production releases GHGs from fabs. Constructing datacenters also produces GHG emissions. Similarly, mobile-device vendors must consider both the GHGs from upstream manufacturing and downstream use.

B. System-level carbon-output analysis

In addition to the organization-level analysis, carbon output can be computed for individual hardware systems and components. Knowing the carbon footprint of individual hardware systems is crucial for carbon optimization. The state-of-industry practice to evaluate the carbon footprint of individual systems is to conduct life-cycle analyses (LCAs) [11], to quantify emissions across production/manufacturing, transport, use, and end-of-life processing, as Figure 1(bottom) shows.

- **Production:** emissions from procuring or extracting raw materials, manufacturing, assembly, and packaging.
- **Transport:** emissions from moving the hardware to its point of use.
- **Use:** emissions from the hardware's operation, including static and dynamic power consumption.
- **End-of-life:** emissions from end-of-life processing and recycling of hardware.

Mobile and data-center devices integrate components and IP from various organizations. The design, testing, and manufacture of individual components (e.g., CPUs, SoCs, DRAM, and HDD/SSD storage) spreads across many technology companies. Furthermore, mobile devices comprise displays, batteries, sensors, and cases that contribute to their carbon footprint. Similarly, data centers comprise rack infrastructure, networking, and cooling systems; data center construction is yet another factor. Quantifying individual systems requires quantifying emissions across fabs, mobile vendors, and data-center operators. Figure 1(bottom) ties the Scope 1, Scope 2, and Scope 3 (upstream and downstream), of technology companies to hardware manufacturing and operational use. This paper uses accredited and publicly reported LCAs from industry, including Apple, Google, Microsoft, and TSMC.

III. Environmental impact of personal computing

Using publicly reported carbon-emission data from industry, this section studies the environmental impact of personal computing devices. We detail the carbon footprint of various platforms (e.g., mobile phones, wearable devices, personal assistants, tablets, laptops, and desktop PCs). Furthermore, we conduct a case study on tradeoffs between mobile performance, energy efficiency, and carbon emissions for an example AI inference workload. The results demonstrate that software and hardware researchers should revisit mobile design to build platforms that are not only efficient but also sustainable.

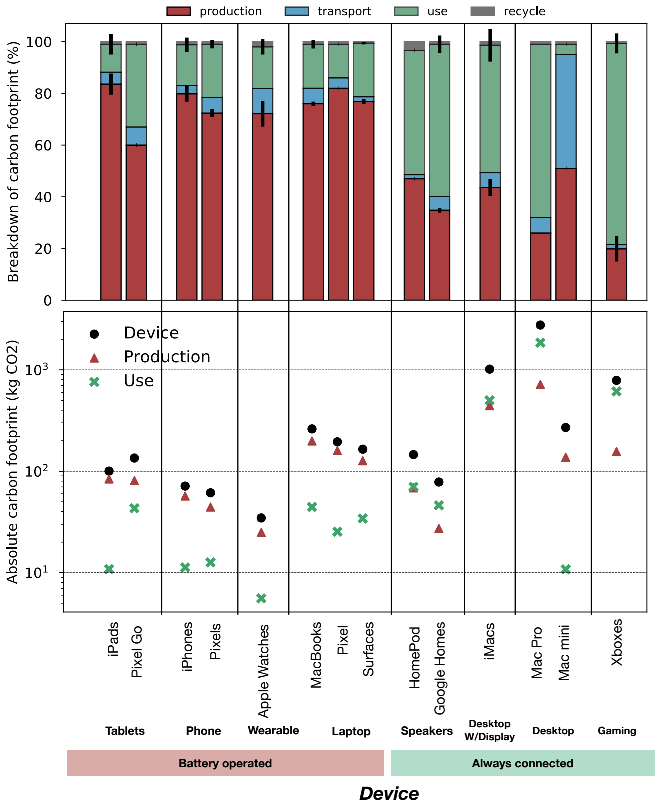

Fig. 2. Breakdown of carbon emissions for various Apple, Google, and Microsoft personal-computing platforms. As the top chart shows, hardware manufacturing dominates the carbon output for battery-powered devices (e.g., phones, wearables, and tables); most emissions for always connected devices (e.g., laptops, desktops, and game consoles) come from product use. The bottom chart shows the absolute carbon output of battery-powered and always connected devices. Overall, carbon footprint (total, manufacturing, and use) is variable and scales with the platform.

A. Personal-computing life-cycle analyses

While the computer systems community has devoted significant efforts to optimize the operational efficiency of personal-computing platforms, the majority of their carbon footprint owes to hardware manufacturing. For example, in 2019, Apple reported a company-wide carbon footprint of 25 million metric tons of CO_2. Hardware manufacturing of integrated circuits (IC's), boards and flexes, displays, electronics, steel, and assembly, accounts for over 74% of all emissions. By comparison, operational emissions from running Apple devices amounts to 19% of the company's total emissions [9].

Among the salient hardware-manufacturing components are integrated circuits, boards and flexes, aluminum, electronics, steel, and assembly. Manufacturing integrated circuits, comprises 33% of Apple's total carbon output, consist of CPUs, DRAMs, SoCs, and NAND flash storage [5]. In fact,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3163226, IEEE Micro

capex-related carbon emissions from manufacturing IC's alone eclipse opex-related carbon emissions from device energy consumption. The role of IC's illustrates the potential impact computer-architecture and circuit researchers can have on sustainable-hardware design.

Going deeper than the overall breakdown of personal computing devices, the breakdown of carbon footprint between manufacturing and use varies across devices.

Takeaway: *Manufacturing dominates emissions for battery-powered devices, whereas operational energy consumption dominates emissions from always-connected devices.*

Figure 2 (top) shows LCAs for different battery-powered devices (e.g., tablets, phones, wearables, and laptops) and always connected devices (e.g., personal assistants, desktops, and game consoles). The analysis aggregates LCAs from Apple, Google, and Microsoft products released after 2017. For devices with multiple models, such as the iPhone 11, iPhone XR, and iPhone SE, we show one standard deviation of manufacturing and operational-use breakdowns. For all devices, we aggregate each one's emissions across its lifetime, representing an average of three to four years for mobile phones, wearables, tablets, and desktops [5], [8].

To reduce the carbon footprints of personal-computing devices, hardware and software designers must consider the carbon impact of both hardware manufacturing (embodied) and energy consumption (operational). For instance, Figure 2 (top) shows that manufacturing accounts for roughly 75% of the emissions for battery-powered devices. Energy consumed by these devices accounts for approximately 20% of emissions. By comparison, most emissions for always connected devices are from operation. Nonetheless, hardware manufacturing accounts for 40% of carbon output from personal assistants and 50% from desktops.

Takeaway: *In addition to the carbon breakdown, the total output for device and hardware manufacturing varies by platform. The hardware-manufacturing footprint increases with increasing hardware capability (e.g., flops, memory bandwidth, and storage).*

Figure 2 (bottom) shows the average absolute carbon emissions for manufacturing (▲), operation (**X**), and the overall device total (●).

Across devices, the amount of total, manufacturing-related, and use-related emissions vary. For instance, always connected devices typically involve more emissions than battery-powered devices. To illustrate, the total and manufacturing footprint for an Apple MacBook laptop is typically 3× that of an iPhone. The varying total and manufacturing levels illustrate that the embodied emissions depend on the platform design and scale rather than being a static overhead.

In addition to varying emissions across devices, the carbon footprint across hardware generations varies as well. As an example, we consider generations of iPhones 3GS (2008) to XR (2018). While manufacturing accounted for 40% of iPhone 3GS' emissions, it is responsible for 75% of emissions in the iPhone XR. Furthermore, we find the absolute carbon output increased by nearly 50%, despite the increase in energy-efficiency, leading to lower operational emissions [9]. The rise in total footprint owes to a rising contribution from manufacturing owing to increased hardware sophistication (e.g., performance, efficiency, storage and memory capacity, application support). The opposing energy-efficiency and carbon-emission trends underscore the inequality of these two factors.

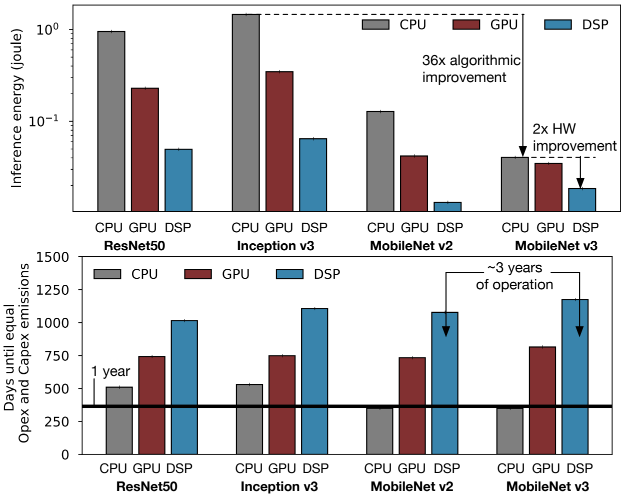

Fig. 3. Evaluating carbon footprint between manufacturing- and operational-related activities for Google Pixel 3 smartphone. Algorithmic AI and hardware advances dramatically shifted carbon emissions toward manufacturing overhead. The top shows the algorithm-level and hardware-level energy improvements for mobile AI inferences. The bottom chart shows how many days of image processing is necessary for operational output to equal integrated-circuit-manufacturing output.

B. Performance and energy versus carbon footprint

In addition to the overall carbon emissions from manufacturing and operational energy consumption, we also consider performance, energy, and carbon footprint tradeoffs for an example workload: mobile AI inference.

Takeaway: *Given the energy-efficiency improvements from software and hardware innovation over the last decade, amortizing the manufacturing carbon output requires continuously operating mobile devices for three years—beyond their typical lifetime.*

Figure 3 illustrates the energy (top) consumption of several well-known convolutional neural networks. Results are for a unit batch size and 224×224 images on a Google Pixel 3 phone with a Qualcomm Snapdragon 845 SoC. We measured energy consumption on a Monsoon power monitor. As expected, algorithmic and hardware innovation has improved energy efficiency by 36× and 2×, respectively. These energy optimizations are a result of improved performance and power efficiency; they also also affect carbon output of AI on mobile devices.

Carbon emissions from hardware manufacturing can be amortized by lengthening the hardware's operating time. Here, we define the starting point of this amortization when the carbon output from operational use equals that from hardware manufacturing (i.e., the ratio of opex emissions to capex emissions is 1). Figure 3 shows this breakeven in terms of the number of days of continuous operation (bottom) on a Google

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3163226, IEEE Micro

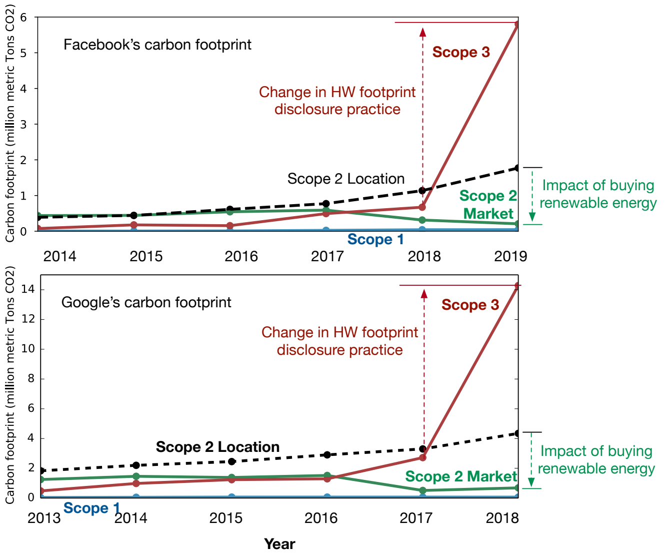

Fig. 4. Carbon footprint of Facebook and Google (two large data center operators). As data centers increasingly rely on renewable energy, carbon emissions originate more from Scope 3, or supply-chain emissions (e.g., hardware manufacturing and construction).

Pixel 3 phone. We converted our measured power consumption to operational carbon emissions by assuming the average US energy-grid output: 380 g of CO_2 per kilowatt-hour. Finally, the manufacturing carbon footprint considers the overhead of building the SoC alone—based on Apple's sustainability report that half of the device's manufacturing footprint come from integrated circuits.

Algorithmic and architectural innovation has boosted energy efficiency, lengthening the amortization time. Figure 3(bottom) illustrates how many days of continual AI inference are necessary for the operational carbon footprint to equal the manufacturing footprint. MobileNet v3 running on a CPU, for example, takes 5 billion images and 350 days of continuous operation. DSPs increase the duration to nearly 1,200 days—a total of 10 billion images—due to 1.5× and 2.2× improvements in performance and power efficiency, respectively. By comparison, the device's expected lifetime is three years, or 1,100 days. Generally, given algorithmic and architectural enhancements, amortizing carbon emissions from hardware manufacturing requires performing AI inference beyond the expected lifetime of most mobile devices.

IV. Environmental impact of data centers

As AI, autonomous driving, robotics, scientific computing, AR/VR, and other emerging applications become ubiquitous, it is vital to consider the environmental implications of both edge and data-center systems. In this section we explore the environmental impact of data centers. First we consider the carbon-emission breakdown of Facebook and Google facilities using industry-reported GHG Protocol data. Next, we discuss the historical trends of data-center carbon emissions. Our discussion highlights the positive impact of renewable energy on these emissions and the need for more-detailed accounting and reporting. Finally, we summarize the impact of renewable energy on data-center footprints.

A. Breakdown of warehouse-scale data centers

Takeaway: *For modern warehouse-scale data-center operators and cloud providers, most emissions are capex-related—for example, construction, infrastructure, and hardware manufacturing.*

Figure 4 illustrates the carbon footprint of Google (2013 to 2018) and Facebook (2014 to 2019). Following the GHG Protocol, we split emissions into Scope 1 (blue), Scope 2 (green), and Scope 3 (red). Recall, Scope 1 comprises emissions from burning refrigerants, gas, and diesel; Scope 2 emissions owe to purchased energy; and Scope 3 emissions come from the supply chain (see Section II).

Analyzing the most recent data, Scope 3 comprises the majority of emissions for both Google and Facebook. In 2018, Google reported 21× higher Scope 3 emissions than Scope 2 emissions—that is, 14,000,000 metric tons of CO_2versus 684,000. In 2019, Facebook reported 23× higher Scope 3 emissions than Scope 2 emissions—that is, 5,800,000 metric tons of CO_2versus 252,000.

Recall that Scope 3 emissions aggregate the entire supply chain; a large fraction of them are from data-center capex overhead such as construction and hardware manufacturing. In 2019, Facebook reported capital goods—a combination of construction and hardware manufacturing for servers—account for half of its Scope 3 emissions.

Similarly, we anticipate the majority of Google's Scope 3 emissions are from construction and hardware manufacturing. Figure 4 shows that between 2017 and 2018, the company reported a 5× increase in that output. In comparison, during the two years, Google's data-center energy consumption increased by 30%. The large increase in carbon emissions is a result of Google additionally accounting and disclosing emissions from capex-related activities (i.e., hardware manufacturing, construction). The additional disclosure results in capex-related emissions dominating opex-related ones. The varying guidelines for accounting and disclosing end-to-end supply chain emissions underscores the importance of improved and standardized measurement.

B. Impact of renewable energy

To decrease operational carbon emissions, data centers are increasingly employing renewable energy.

Takeaway: *Although overall data-center energy consumption has risen over the past five years, carbon emissions from operational energy consumption have fallen. The primary factor contributing to the growing gap between data-center energy consumption and carbon output is the use of renewable energy.*

Figure 4 illustrates the carbon footprint of Google and Facebook over six years. Although the figure divides these emissions into Scope 1, Scope 2, and Scope 3, Scope 2 comprises two types: location based and market based. Location-based emissions assume the local electricity grid

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3163226, IEEE Micro

produces the energy—often through a mix of brown (i.e., coal and gas) and green sources. Market-based emissions reflect energy that companies have purposefully chosen or contracted—typically solar, hydroelectric, wind, and other renewable sources. Around 2013, Facebook and Google began procuring renewable energy to reduce operational carbon emissions. These purchases decreased their operational carbon output even though their energy consumption continued to increase. Thus, minimizing the emissions related to data-center workloads and hardware must consider renewable energy and the tradeoffs between opex- and capex-related factors.

The importance of considering the impacts of renewable energy for data-center scale systems is further highlighted by sustainability analyses from Intel and AMD. The semiconductor manufacturers carbon data reports mimic the format of hardware life cycles categorizing emissions between manufacturing, transport, use, and recycling. Assuming a baseline US energy grid, roughly Intel and AMD report 60% and 45% of the hardware carbon footprint's owe to hardware use and energy consumption [9]. With renewable energy, however, emissions from operational consumption decrease. For instance, assuming the server-class systems are driven by solar or wind energy, over 80% of emissions come from hardware manufacturing.

Designing sustainable data centers should therefore consider the role of renewable energy, the effect of efficiency increases on opex-related emissions, and the effect of resource provisioning and leaner hardware on capex-related emissions.

V. Environmental impact from manufacturing

So far, our results show hardware manufacturing comprises a large portion of emissions in both mobile and data-center systems. In data centers, renewable energy is a significant contributor to the opex-related footprint. In this section, we consider the carbon footprint of chip manufacturing and the impact of powering fabs using renewable energy.

Takeaway: *Using renewable energy to power fabs will reduce the carbon emissions from hardware manufacturing. Even under optimistic renewable-energy projections, however, manufacturing will continue to represent a large portion of hardware-life-cycle carbon footprints.*

Figure 5 shows the carbon breakdown for wafer manufacturing at TSMC [10]. The breakdown is normalized to the baseline energy source. To model the impact of renewable energy, we vary the carbon intensity of the energy consumed. Although the precise energy-grid efficiency is unknown, our analysis considers a range of improvements, including the best case: replacing coal with 100% wind energy for a 70× improvement. Using greener energy directly reduces the fab's carbon output from consumed energy (green).

Even though using renewable energy can cut a fab's hardware-manufacturing carbon emissions, minimizing life-cycle and hardware-manufacturing emissions will remain important. As Figure 5 shows, a 64× boost in renewable energy reduces the overall carbon output by roughly 2.7×, an ambitious goal. By 2025, TSMC estimates renewable energy will produce 20% of the electricity that drives forthcoming 3nm fabs [10]. Recall that roughly 75% of the carbon footprint for battery-powered devices is from hardware manufacturing (see Section III), and opex is a small fraction for data centers. Even as fabs employ more renewable energy to reduce their environmental impact, hardware manufacturing will remain an important aspect of designing sustainable computers and workloads.

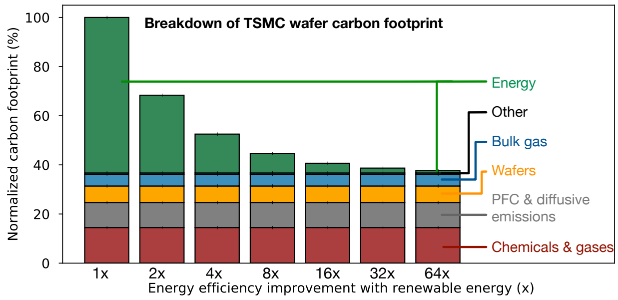

Fig. 5. Carbon-emissions breakdown for TSMC wafer manufacturing. Renewable energy provides up to a 64× reduction in emissions from electricity, and overall emissions for wafers drops by 2.7×. Although the reduction will reduce the carbon output of manufacturing, consideration of capex-related emissions for mobile and data-center hardware will remain important.

VI. Addressing carbon footprint of systems

Optimizing the environmental impact of mobile and data-center computing platforms requires addressing the carbon footprint from operational energy consumption and hardware manufacturing. Given its immediate importance and scale, we must adopt vertically integrated research methods to minimize the emissions associated with computing. This section outlines future directions to tackle the environmental footprint of computing.

A. The need for carbon accounting

The first step to design and optimization is the ability to quantify the carbon footprint of computing platforms. Although many organizations publicly report their carbon emissions, improved accounting and report is crucial to enable sustainability-aware optimizations. Similar to performance, power, energy, and area measurement and modeling tools, the systems and architecture needs commensurate carbon accounting methods. This includes a combination of detailed platform-level life-cycle analyses at the mobile and data-center scale, as well as component-level carbon footprint breakdown for processors, memory, storage, and networking IC's. Standardized accounting and disclosure as well as broader participation will provide further guidance on tackling salient challenges in realizing environmentally sustainable systems.

B. Carbon footprint as a first-order design optimization target

In addition to improve accounting and reporting, researchers and developers across the computing stack must consider carbon footprint as a first-order design metric alongside performance and efficiency. The analysis shown in this work demonstrates the distinct trends between efficiency optimization and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3163226, IEEE Micro

carbon emissions across end-to-end hardware life cycles. This requires researchers and developers to holistically consider both **operational** and **embodied** emissions.

Furthermore, efficiency optimization is insufficient to curb computing's rising carbon footprint. While improving efficiency can reduce both operational and embodied emissions, the benefits are typically overshadowed by rising application-level demands, as illustrated by Jevon's paradox. Enabling sustainable computing requires targeted carbon optimization.

C. Cross-stack carbon optimization

Applications and algorithms. Application- and algorithm-level optimizations can reduce both operational and embodied emissions. For example, at the datacenter scale, reducing application-level demands directly lowers power and infrastructure capacity devoted to services. As an example, consider AI training. The energy footprint of AI training is parameters based on the footprint of processing one example (E), the data-set size (D), and the hyperparameter search (H) [12]. Reducing carbon emissions requires training on systems with fewer resources. Novel methods to train models given lesser compute and storage capabilities can directly reduce carbon emissions (i.e., E, D). Similarly, reducing the hyperparameter-search factor (H) reduces the necessary number of parallel training nodes. Generally, algorithmic optimizations for scale-down systems will drastically cut emissions.

Systems and hardware. Systems researchers can guide overall mobile- and data-center-scale system provisioning to reduce both operational and embodied emissions. Recently, systems have scaled up and out to boost performance. Future sustainable systems must consider strict resource budgets based on application and service-level requirements. Furthermore, over the past 20 years, substantial effort has been devoted to energy-efficient mobile and data-center systems. As devices rise to the billion-transistor scale, workloads experience low utilization. Researchers must design hardware to balance under utilized dark silicon with the overhead of embodied emissions. By judiciously balancing general purpose and specialized circuits, hardware platforms can directly reduce carbon output.

Circuits and devices. Future circuit research can also reduce embodied carbon emissions. First, it may consider circuit-level resource provisioning to balance performance, area, energy efficiency, and carbon. Vertically integrated research into specializing low-carbon circuits for salient applications will also decrease embodied emissions. Finally, embodied emissions must be addressed through device modeling, characterization, design, and fab manufacturing. For instance, hardening a device's reliability and endurance extends its lifetime (e.g., DRAM, NAND-flash-memory), cutting embodied carbon emissions. Moreover, sustainable manufacturing processes via novel devices, yield enhancement, fabrication materials, renewable-energy sources, and maximum operating efficiency will directly reduce production overhead.

VII. Conclusion and future work

As computing technology becomes ubiquitous, so does its environmental impact. This work shows how developers and researchers can approach the environmental consequences of computing, from mobile to data-center-scale systems. First, we demonstrated that, going forward, reducing operational energy consumption alone fails to reduce end-to-end carbon emissions. Next, we described the industry's practice for quantifying the carbon output of organizations and individual systems. Finally, on the basis of our analysis, we characterized the carbon emissions of various hardware platforms. Our effort demonstrates that over the last decade, hardware manufacturing—as opposed to operational energy consumption—has increasingly dominated the carbon footprint of mobile systems. Similarly, as more data centers employ renewable energy, the dominant source of their total carbon footprint is shifted to hardware manufacturing.

We hope this work lays the foundation for future investigation of environmentally sustainable systems. Designing, building, and deploying such systems requires collective industry/academic collaboration.

VIII. Author Biographies

Udit Gupta is a PhD student in Computer Science at Harvard University and visiting research scientist at Meta AI Research. His research interests focus on improving the performance, efficiency, and sustainability of computer systems and architectures by co-designing solutions across the computing stack. He can be reached at ugupta@g.harvard.edu.

Young Geun Kim is currently an Assistant Professor of Computer Science at Korea University. His research focus lies in the domain of computer system architecture with particular emphasis on energy-efficient and low-power systems. He can be reached at younggeun.kim@korea.ac.kr.

Sylvia Lee Sylvia Lee currently leads Meta's climate sustainability programs. She served as the Associate Director at the World Economic Forum's Environmental Initiative in Switzerland, where she helped raise global awareness of climate and water as key global risks. She can be reached at sslee@fb.com.

Jordan Tse is a Sustainability Program Manager at Meta. focusing on supplier engagement and hardware circularity efforts across their data center and consumer product businesses. She is passionate about finding ways to leverage Meta's scale and innovation to build sustainable and resilient infrastructure and products. She can be reached at jordant@fb.com.

Hsien-Hsin S. Lee is an Area Research Lead at Meta AI. His work investigates inter-domain optimizations across computer architecture, machine learning, and semiconductor technologies. He is a Fellow of IEEE. Contact him at leehs@fb.com.

Gu-Yeon Wei is the Robert and Suzanne Case Professor of Electrical Engineering and Computer Science in the John A Paulson School of Engineer and Applied Sciences at Harvard University. He is also a Samsung Research Fellow. His recent research interests span a broad range of topics

from algorithm-hardware co-design for efficient computing systems to sustainable computing. He can be reached at guyeon@seas.harvard.edu.

David Brooks is a Professor of Computer Science at Harvard University. His research focuses on the interaction between the architecture and software of computer systems and underlying hardware implementation challenges.He is also a visiting Scientist at Meta. He can be reached at dbrooks@eecs.harvard.edu.

Carole-Jean Wu is currently a Research Scientist at Meta. Her research sits at the intersection of computer architecture and machine learning with emphasis on developing energy-efficient and sustainable systems, optimizing systems for machine learning execution, and designing learning-based approaches for system design and optimization. She can be reached at carolejeanwu@fb.com.

REFERENCES

[1] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020.

[2] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," *IEEE Micro*, vol. 32, no. 3, pp. 122–134, 2012.

[3] G. Foundries, "Global foundries corporate responsibility report," 2019.

[4] Intel, "Corporate responsibility at intel," 2020.

[5] Apple, "Apple environmental responsibility report," 2019.

[6] Facebook, "Facebook sustainability data 2019," 2020.

[7] A. M. Devices, "Advanced micro devices ("amd") corporate responsibility," 2020.

[8] Google, "Google environmental report 2019," 2020.

[9] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, and C.-J. Wu, "Chasing carbon: The elusive environmental footprint of computing," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 854–867, IEEE, 2021.

[10] TSMC, "Tsmc corporate social responsibility report," 2018.

[11] R. U. Ayres, "Life cycle analysis: A critique," *Resources, conservation and recycling*, vol. 14, no. 3-4, pp. 199–223, 1995.

[12] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *arXiv preprint arXiv:1907.10597*, 2019.