2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA) | 978-1-6654-2235-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/HPCA51647.2021.00076

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.**

***Index Terms*—Data center, mobile, energy, carbon footprint**

I. Introduction

The world has seen a dramatic advancement of information and communication technology (ICT). The rise in ICT has resulting in a proliferation of consumer devices (e.g., PCs, mobile phones, TVs, and home entertainment systems), networking technologies (e.g., wired networks and 3G/4G LTE), and data centers. Although ICT has enabled applications including cryptocurrencies, artificial intelligence (AI), e-commerce, online entertainment, social networking, and cloud storage, it has incurred tremendous environmental impacts.

Figure 1 shows that ICT is consuming more and more electricity worldwide. The data shows both optimistic (top) and expected (bottom) trends across mobile, networking, and data-center energy consumption [1], [2]. On the basis of even optimistic estimates in 2015, ICT accounted for up to 5% of global energy demand [1], [2]. In fact, data centers alone accounted for 1% of this demand, eclipsing the total energy consumption of many nations. By 2030, ICT is projected to account for 7% of global energy demand. Anticipating the ubiquity of computing, researchers must rethink how to design and build sustainable computer systems.

Given the growing energy demand of computing technology, software and hardware researchers have invested heavily in maximizing the energy efficiency of systems and workloads.

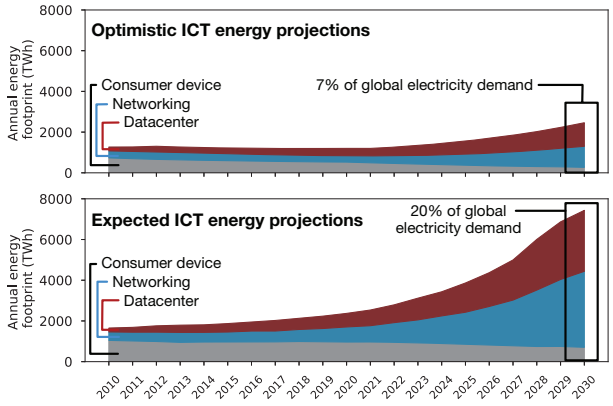

Fig. 1. Projected growth of global energy consumption by information and computing technology (ICT). On the basis of optimistic (top) and expected (bottom) estimates, ICT will by 2030 account for 7% and 20% of global demand, respectively [1].

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 [3]. For salient applications, such as AI [4]–[9], molecular dynamics [10], video encoding [11], and cryptography [12], 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 improve power usage effectiveness (PUE) [13].

The aforementioned energy-efficiency improvement reduces the operational energy consumption of computing equipment, in turn mitigating environmental effects [14], [15]. In addition, using renewable energy further reduces operational carbon emissions. Figure 2 (left) shows the energy consumption (black) and net carbon footprint from purchased energy (red) for Facebook's Prineville data center. Between 2013 and 2019, as the facility expanded, the energy consumption monotonically increased. On the other hand, the carbon emissions started decreasing in 2017 [16]. By 2019, the data center's net operational carbon output reached nearly zero [16], a direct result of powering it with renewable energy such as

978-1-6654-2235-2/21/$31.00 ©2021 IEEE
DOI 10.1109/HPCA51647.2021.00076

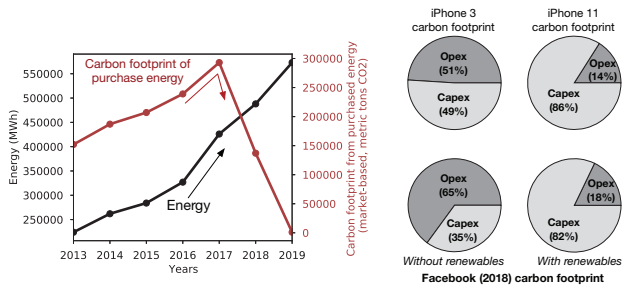

Fig. 2. Carbon footprint depends on more than just energy consumption (left). Although the energy consumption of Facebook's Prineville data center increased between 2013 and 2019, its net operational carbon output decreased because of renewable-energy purchases. The carbon-emission breakdown has shifted from primarily opex-related activities to overwhelmingly capex-related activities (right). The top two pie charts show the breakdown for the iPhone 3 (2008) versus the iPhone 11 (2019); the bottom two show the breakdown for Facebook's data centers with and without renewable energy.

wind and solar. Note, the decrease in net operational carbon emissions from using renewable energy is not the same as the data center being powered exclusively with renewable energy at all times [17]–[19]. The distinction between energy consumption and carbon footprint highlights the need for computer systems to directly minimize directly optimize for environmental impact.

Given the advancements in system energy efficiency and the increasing use of renewable energy, most carbon emissions now come from infrastructure and the hardware. Similar to dividing data-center infrastructure-efficiency optimization into opex (recurring operations) and capex (one-time infrastructure and hardware), we must divide carbon emissions into opex- and capex-related activities. For the purposes of this work, we define opex-related emissions as emissions from hardware use and operational energy consumption; we define capex-related emissions as emissions from facility-infrastructure construction and chip manufacturing (e.g., raw-material procurement, fabrication, packaging, and assembly. Figure 2 (top, right) shows that between 2009 (iPhone 3GS) and 2019 (iPhone 11), most carbon emissions attributable to mobile devices shifted from opex related to capex related [20], [21]. Similarly, Figure 2 (bottom, right) shows that in 2018, after having converted its data centers to renewable energy, most of Facebook's remaining emissions are capex related [16].

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 [22], [23]. 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 [24], 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 lays the foundation for characterizing 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 industry partners and to study the carbon footprint of mobile and data-center hardware (Section II). On the basis of publicly available sustainability reports from AMD, Apple, Facebook, Google, Huawei, Intel, Microsoft, and TSMC, we show that the hardware-manufacturing process, rather than system operation, is the primary source of carbon emissions (Section III and IV). Despite the growing use of renewable energy to power semiconductor manufacturing, hardware manufacturing and capex-related activities will continue to dominate the carbon output (Section V). Finally, we outline future research and design directions across the computing stack that should enable the industry to realize environmentally sustainable systems and to reduce the carbon footprint from technology (Section VI).

The important contributions of this work are:

1) We show that given the considerable efforts over the past two decades to increase energy efficiency, the dominant factor behind the overall carbon output of computing has shifted from operational activities to hardware manufacturing and system infrastructure. Over the past decade, the fraction of life-cycle carbon emissions due to hardware manufacturing increased from 49% for the iPhone 3GS to 86% for the iPhone 11.
2) Our smartphone-based measurement shows that efficiently amortizing the manufacturing carbon footprint of a Google Pixel 3 smartphone requires continuously running MobileNet image-classification inference for three years—beyond the typical smartphone lifetime. This result highlights the environmental impact of system manufacturing and motivates leaner systems as well as longer system lifetimes where possible.
3) 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, for instance, capex- and supply-chain-related activities accounted for 23× more carbon emissions than opex-related activities at Facebook.
4) We chart future paths for software and hardware researchers to characterize and minimize computing technology's environmental impact. Sustainable computing will require interdisciplinary efforts across the computing stack.

II. Quantifying Environmental Impact

The environmental impact of ICT is complex and multifaceted. For instance, technology companies consider many environmental matters including consumption of energy, water, and materials such as aluminum, cobalt, copper, glass, gold,

Technology company	Scope 1	Scope 2	Scope 3
Chip manufacturer	Burning PFCs, chemicals, gases	Energy for fabrication	Raw materials, hardware use
Mobile-device vendor	Natural gas, diesel	Energy for offices	Chip manufacturing, hardware use
Data-center operator	Natural gas, diesel	Energy for data centers	Server-hardware manufacturing, construction

TABLE I
IMPORTANT FEATURES OF SCOPE 1, SCOPE 2, AND SCOPE 3 EMISSIONS, FOLLOWING THE GREENHOUSE GAS (GHG) PROTOCOL, FOR SEMICONDUCTOR MANUFACTURERS, MOBILE-DEVICE VENDORS, AND DATA-CENTER OPERATORS.

tin, lithium, zinc, and plastic. In this paper we focus on a single important environmental issue: carbon emissions, which represents total greenhouse-gas (GHG) emissions.

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 [16], [25]–[29]. 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, including data-center and mobile platforms. The remainder of the paper builds on these methods to quantify the carbon output of computer systems.

A. Industry-level carbon-emission analysis

A common method for quantifying organization-level carbon output is the Greenhouse Gas (GHG) Protocol [30], an accounting standard by which many companies report their carbon emissions. For example, AMD, Apple, Facebook, Google, Huawei, Intel, and Microsoft publish annual sustainability reports using the GHG Protocol [30]. Our analysis builds on such publicly available reports. As Figure 3 shows, the GHG Protocol categorizes emissions into three scopes: Scope 1 (direct emissions), Scope 2 (indirect emissions from purchased energy), and Scope 3 (upstream and downstream supply-chain emissions). We define each one in the context of technology companies, as follows.

Scope 1 emissions come from fuel combustion (e.g., diesel, natural gas, and gasoline), 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 a large fraction for chip manufacturers. Overall, it accounts for over half the operational carbon output from Global Foundries, Intel, and TSMC [25], [26], [31]. Much of these emissions come from burning perfluorocarbons (PFCs), chemicals, and gases. TSMC reports that nearly 30% of emissions from manufacturing 12-inch wafers are due to PFCs, chemicals, and gases [31]. In this paper we show that chip manufacturing, as opposed to hardware use and energy consumption, accounts for most of the carbon output attributable to hardware systems.

Scope 2 emissions come from purchased energy and heat powering semiconductor fabs, offices, and data-center operation. They depend on two parameters: the energy that operations consume and the GHG output from generating the consumed energy (in terms of carbon intensity—i.e., grams of CO_2 emitted per kilowatt-hour of energy). Scope 2 emissions are especially important in semiconductor fabs and data centers.

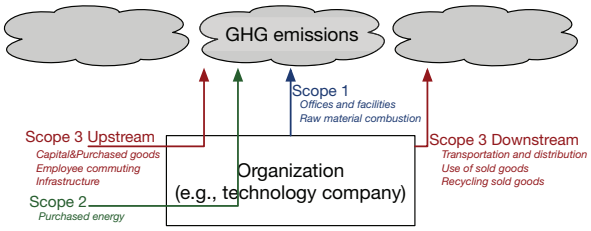

Fig. 3. Many organizations follow the Greenhouse Gas (GHG) Protocol to quantify their environmental impact. This protocol categorizes emissions into Scope 1 (direct), Scope 2 (indirect), and Scope 3 (upstream and downstream supply chain).

Semiconductor companies need copious energy to manufacture chips. Energy consumption, for instance, produces over 63% of the emissions from manufacturing 12-inch wafers at TSMC [31]. And energy demand is expected to rise, with next-generation manufacturing in a 3nm fab predicted to consume up to 7.7 billion kilowatt-hours annually [31], [32]. TSMC's renewable-energy target will account for 20% of its fabs' annual electricity consumption, reducing its average carbon intensity. Despite these improvements, this work shows hardware manufacturing will constitute a large portion of computing's carbon footprint.

Scope 2 emissions are also especially important for data centers. The operational footprint of a data center has two parameters: the overall energy consumption from the many servers and the carbon intensity of that energy. Note that the carbon intensity varies with energy source and grid efficiency. Compared with "brown" energy from coal or gas, "green" energy from solar, wind, nuclear, or hydropower produces up to 30× fewer GHG emissions [33]–[35]. Scope 2 emissions for a data center therefore depend on the geographic location and energy grid. In fact, warehouse-scale data centers are purchasing renewable energy (e.g., solar and wind) to reduce GHG emissions.

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

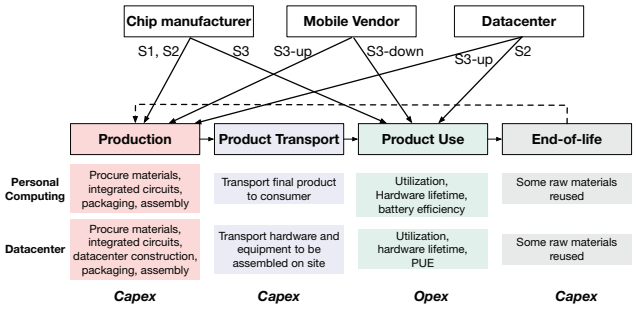

Fig. 4. The hardware life cycle includes production, transport, use, and end-of-life processing. Opex-related (operational) carbon emissions are based on use; capex-related emissions results are from aggregating production/manufacturing, transport, and end-of-life processing.

chain). Accurate accounting of Scope 3 emissions requires in-depth analysis of GHGs resulting from construction, hardware manufacturing, and the devices' frequency of use as well as mixture of workloads. These characteristics must consider the lifetime of systems; for example, data centers typically maintain server-class CPUs for three to four years [13].

Table I summarizes the salient emissions from each category for chip manufacturers, mobile vendors, and data-center operators.

B. System-level carbon-output analysis

In addition to industry- and organization-level analysis using the GHG Protocol, carbon output can be computed for individual hardware systems and components. Knowing the carbon footprint of individual hardware systems (e.g., server-class CPUs, mobile phones, wearable devices, and desktop PCs) not only enables consumers to understand personal carbon impact but also enables designers to characterize and optimize their systems for environmental sustainability. Typically, evaluating the carbon footprint of an individual hardware system involves life-cycle analyses (LCAs) [36], including production/manufacturing, transport, use, and end-of-life processing, as Figure 4 shows.

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 tens of 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; their construction is yet another factor. Quantifying individual systems requires quantifying GHG emissions across chip manufacturers, mobile vendors, and data-center operators. Figure 4 ties the Scope 1 (S1), Scope 2 (S2), Scope 3 upstream (S3-up), and Scope 3 downstream (S3-down) of technology companies to hardware manufacturing and operational use. Note that although LCAs can help determine system-level emissions, they are lengthy and incur high effort to perform.

Computer systems have four LCA phases:

- **Production:** carbon emissions from procuring or extracting raw materials, manufacturing, assembly, and packaging.

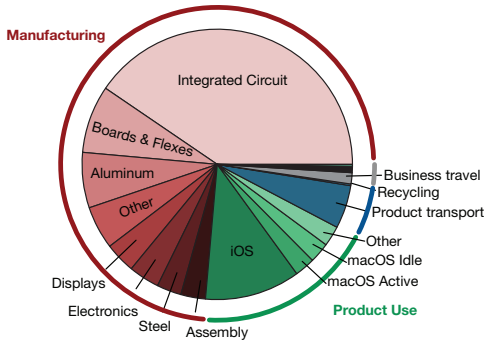

Fig. 5. Apple's carbon-emission breakdown. In aggregate, the hardware life cycle (i.e., manufacturing, transport, use, and recycling) comprises over 98% of Apple's total emissions. Manufacturing accounts for 74% of total emissions, and hardware use accounts for 19%. Carbon output from manufacturing integrated circuits (i.e., SoCs, DRAM, and NAND flash memory) is higher than that from hardware use.

- **Transport:** carbon emissions from moving the hardware to its point of use, including consumers and data centers.
- **Use:** carbon emissions from the hardware's operation, including static and dynamic power consumption, PUE overhead in the data center, and battery-efficiency overhead in mobile platforms.
- **End-of-life:** carbon emissions from end-of-life processing and recycling of hardware. Some materials, such as cobalt in mobile devices, are recyclable for use in future systems.

For this paper we chose accredited and publicly reported LCAs from various industry sources, including AMD, Apple, Google, Huawei, Intel, Microsoft, and TSMC, to analyze the carbon output of computer systems.

III. Environmental Impact of Personal Computing

Using publicly reported carbon-emission data from industry, this section studies the environmental impact of consumer (personal) computing devices. First, we characterize the overall carbon emissions of mobile-device developers, such as Apple, and find that hardware manufacturing dominates their environmental impact. Next, we examine in detail various platforms (e.g., mobile phones, wearable devices, personal assistants, tablets, laptops, and desktop PCs) as well as historical trends. Our analysis considers more than 30 products from Apple, Google, Huawei, and Microsoft. Finally, 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 more efficient and environmentally sustainable.

A. Overall breakdown of mobile vendors

Takeaway 1: *Hardware manufacture and use dominate the carbon output of personal-computing companies (e.g., Apple). More emissions come from designing and manufacturing integrated circuits (e.g., SoCs, DRAM, and storage) than from hardware use and mobile energy consumption.*

Figure 5 shows the breakdown of Apple's annual carbon footprint for 2019 [27]. It separates the company's total emissions—25 million metric tons of CO_2—into manufacturing (red), product use (green), product transport (blue), corporate facilities (grey), and product recycling. Manufacturing, which includes integrated circuits, boards and flexes, displays, electronics, steel, and assembly, accounts for over 74% of all emissions. By comparison, emissions from product use—energy consumption from applications running on hardware[1]—account for only 19% of Apple's overall output.

Among the salient hardware-manufacturing components are integrated circuits, boards and flexes, aluminum, electronics, steel, and assembly. Integrated circuits, comprising roughly 33% of Apple's total carbon output, consist of CPUs, DRAMs, SoCs, and NAND flash storage [27]. In fact, capex-related carbon emissions from manufacturing integrated circuits alone eclipse opex-related carbon emissions from device energy consumption. Additional capacitors, resistors, transistors, and diodes soldered to bare boards and flexes constitute "electronics"; battery cells, plastic, and glass constitute "other." The role of integrated circuits illustrates the potential impact computer-architecture and circuit researchers can have on sustainable-hardware design.

B. Personal-computing life-cycle analyses

Apple's overall carbon footprint aggregates the emissions from all of its mobile phones, tablets, wearable devices, and desktops. Here we detail the carbon footprint of each type. The analysis includes devices from Apple, Google, Huawei, and Microsoft.

Takeaway 2: *The breakdown of carbon footprint between manufacturing and use varies by consumer device. Manufacturing dominates emissions for battery-powered devices, whereas operational energy consumption dominates emissions from always-connected devices.*

Figure 6 (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 [20], [37]–[63]. 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 [27], [29].

To reduce the carbon footprints of personal-computing devices, hardware and software designers must consider the carbon impact of both hardware manufacturing (capex) and energy consumption (opex). For instance, Figure 6 (top) shows that manufacturing (capex) accounts for roughly 75% of the emissions for battery-powered devices. Energy consumed (opex) by these devices accounts for approximately

[1]Apple reports GHG emissions from product use on the basis of the amount of time device is in active operation and the geographically specific energy-grid efficiency.

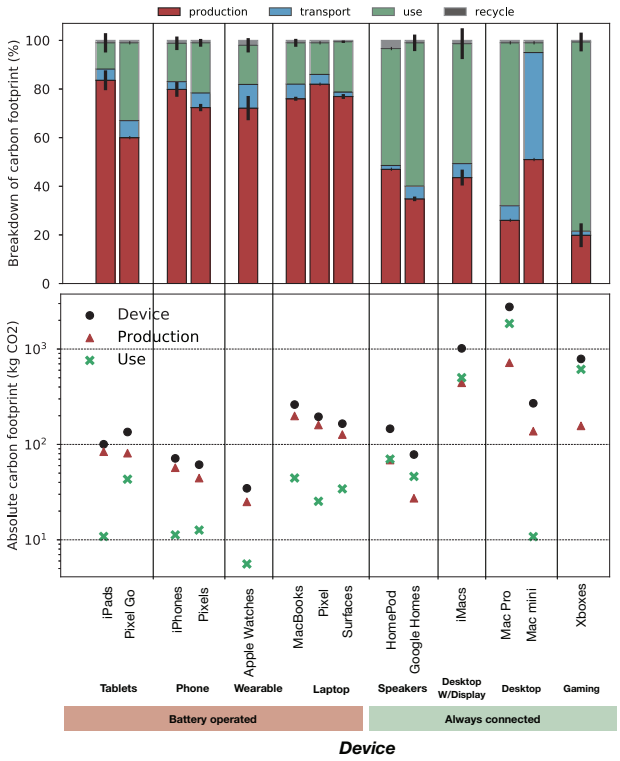

Fig. 6. Breakdown of carbon emissions for various Apple, Google, and Microsoft personal-computing platforms. As the top chart shows, hardware manufacturing dominates the carbon output for battery-powered devices (e.g., phones, wearables, and tables); most emissions for always connected devices (e.g., 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.

20% of emissions. By comparison, most emissions for always connected devices are from their energy consumption. Nonetheless, even for these devices, hardware manufacturing accounts for 40% of carbon output from personal assistants (e.g., Google Home) and 50% from desktops.

Takeaway 3: *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 6 (bottom) shows the absolute carbon emissions for manufacturing (▲), operation (X), and the overall device total (●). Results are based on the average footprint for each device type.

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 capex-related output depends on the platform design and scale rather than being a static overhead.

Takeaway 4: *As energy efficiency improves and hardware*

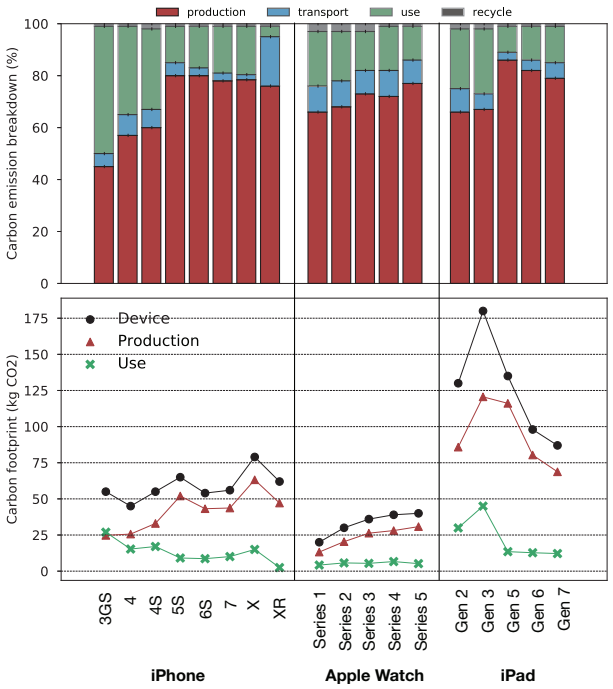

Fig. 7. Carbon emissions and breakdown of emissions across generations for Apple iPhones, Watches, and iPads. Across all devices (top), the fraction from production and manufacturing increased from generation to generation. The absolute carbon output (●) for iPads decreased over time, while for iPhones and Watches it increased (bottom). The rising carbon emissions are largely due to a growing contribution from manufacturing. For iPhones, as carbon from operational use (X) decreased, the manufacturing contribution (▲) increased.

capability increases from one device generation to another, a rising percentage of hardware life-cycle carbon emissions comes from manufacturing.

Figure 7 (top) shows the carbon breakdown over several generations of battery-powered devices: iPhones (from 2008's 3GS to 2018's XR), Apple Watches (2016's Series 1 to 2019's Series 5), and iPads (2012's Gen 2 to 2019's Gen 7). In all three cases, the fraction of carbon emissions devoted to hardware manufacturing increased over time. For iPhones, manufacturing accounts for 40% of emissions in the 3GS and 75% in the XR; for Apple Watches, it accounts for 60% in Series 1 and 75% in Series 5; and for iPads, 60% in Gen2 and 75% in Gen 7.

Figure 7 (bottom) shows the absolute carbon output across generations for the same devices. As performance and energy efficiency of both software and hardware have improved over the past few years, the opex-related carbon output from energy consumption (X) has decreased. Despite the energy-efficiency increases over iPhone and Apple Watch generations, however, total carbon emissions (●) grew steadily. The increasing ouputs owe to a rising contribution from manufacturing (▲) as hardware provides more flops, memory bandwidth, storage, application support, and sensors. The opposing energy-efficiency and carbon-emission trends underscore the inequality of these two factors. Reducing carbon output for the hardware life cycle requires design for lower manufacturing emissions or engagement with hardware suppliers.

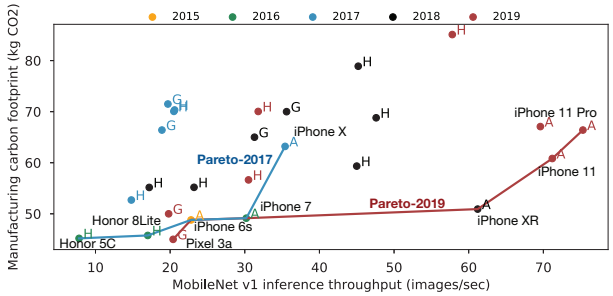

Fig. 8. Performance (MobileNet v1 inference throughput) versus carbon footprint Pareto frontier by mobile-phone generation. ("A" represents Apple, "G" Google, and "H" Huawei.) The Pareto frontier shifted primarily to the right between 2017 (blue line) and 2019 (red line), highlighting the focus on increasing system performance as opposed to decreasing carbon emissions.

C. 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 5: *From 2017 to 2019, software and hardware optimizations primarily focused on maximizing performance, overlooking the growth trend of carbon footprint.*

Figure 8 illustrates the tradeoff between performance, measured as MobileNet v1 throughput (i.e., inference images per second) [64], [65], and the manufacturing carbon footprint. The analysis categorizes devices by their release year (color) and vendor ("G" for Google, "H" for Huawei, and "A" for Apple) [20], [21], [37], [40], [54], [56], [57], [59], [66]. Finally, we highlight two performance/carbon footprint Pareto frontiers for devices made in 2017 and earlier (blue) and for devices made in 2019 and earlier (red).

The Pareto frontiers illustrate a tradeoff between AI performance and carbon footprint. Across the 2019 performance/carbon-footprint frontier, the iPhone 11 Pro achieves a MobileNet v1 throughput of 75 images per second at a manufacturing output of 66 kg of CO_2; in comparison, the Pixel 3a achieves an inference throughput of 20 images per second with 45 kg of CO_2. In addition to this tradeoff, the Pareto frontier between 2017 and 2019 shifts to the right, prioritizing higher performance through more-sophisticated SoCs and specialized hardware. In fact, although the iPhone X (2017) achieved a throughput of 35 images per second at 63 kg of CO_2, the iPhone 11 (2019) doubled that performance at a slightly lower 60 kg of CO_2. While greater performance is important to enabling new applications and improving the user experience, moving the Pareto frontier down is also important—that is, to mitigate the environmental impact of emerging platforms and applications, it is crucial to design workloads and systems with similar performance but lower environmental impact.

Takeaway 6: *Given the energy-efficiency improvements from software and hardware innovation over the last decade,*

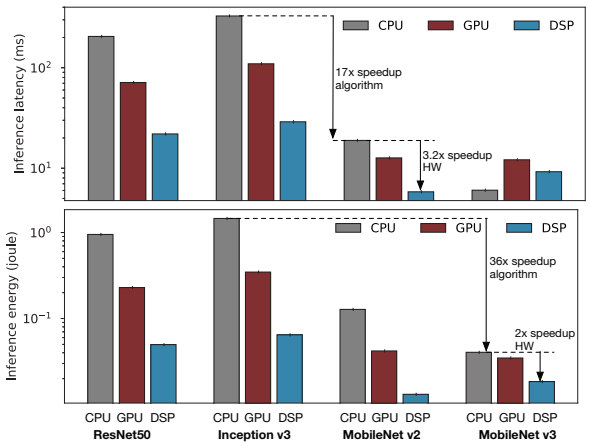

Fig. 9. Evaluating the improvement of inference throughput (top) and energy efficiency (bottom) for different convolutional-neural-network and hardware generations. Algorithmic and hardware advances have considerably increased the performance and operational energy consumption.

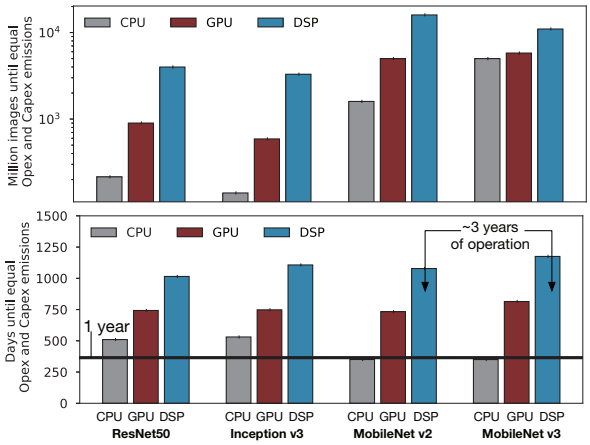

Fig. 10. Evaluating carbon footprint between manufacturing- and operational-related activities for Google Pixel 3 smartphone. Algorithmic AI and hardware advances dramatically shifted carbon emissions toward manufacturing overhead. The top chart shows the number of inference images necessary for operational output to equal the integrated-circuit-manufacturing output. The bottom chart shows how many days of image processing is necessary for operational output to equal integrated-circuit-manufacturing output.

amortizing the manufacturing carbon output requires continuously operating mobile devices for three years—beyond their typical lifetime.

Figure 9 illustrates the inference latency (top) and energy (bottom) 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 [67]. We measured energy consumption on a Monsoon power monitor [68], [69]. As expected, algorithmic and hardware innovation has improved both performance and energy efficiency. For instance, when running on a CPU, MobileNet v2 is 17× faster than Inception v3 [70]. Moreover, it is an additional 3× faster when running on a DSP than on a CPU. Similarly, algorithmic and hardware innovation has increased energy efficiency by 36× and 2×, respectively. The performance and energy optimizations have also affected AI carbon footprint 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 10 shows this breakeven in terms of the number of inferences (top) and days of operation (bottom) on a Google Pixel 3 phone. We converted our measured power consumption (using a Monsoon power monitor [68]) to operational carbon emissions by assuming the average US energy-grid output: 380 g of CO_2 per kilowatt-hour [14]. Finally, the manufacturing carbon footprint considers the overhead of building the SoC alone—assuming half the production carbon emissions are due to integrated circuits (see Figure 5).

Algorithmic and architectural innovation has boosted energy efficiency, lengthening the amortization time. For instance, Figure 10 (top) shows that to bring the carbon output from energy consumption (opex) to parity with that from hardware manufacturing (capex), ResNet-50 requires 200 million images and Inception v3 requires 150 million images when running on a CPU [71], [72]. MobileNet v3 on a CPU takes five billion [70] images; the 25× increase owes to algorithmic advances reducing mobile AI's memory and compute requirements. In addition, hardware enhancements also reduce operational emissions. For example, running MobileNet v3 on a DSP rather than a CPU reduces the operational footprint by 2×, requiring 10 billion images for operational- and hardware-manufacturing-related carbon emissions to balance. In comparison, the ImageNet training set consists of 14 million images [67].

Furthermore, Figure 10 illustrates how many days of continual AI inference are necessary for the operational carbon footprint to equal the hardware-manufacturing footprint. MobileNet v3 running on a CPU, for example, takes 350 days of continuous operation. DSPs increase the duration to nearly 1,200 days due to 1.5× and 2.2× improvements in performance and power efficiency, respectively. By comparison, the device's expected lifetime is three years (about 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, considering the environmental implications of both edge and data-center systems becomes important. 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-

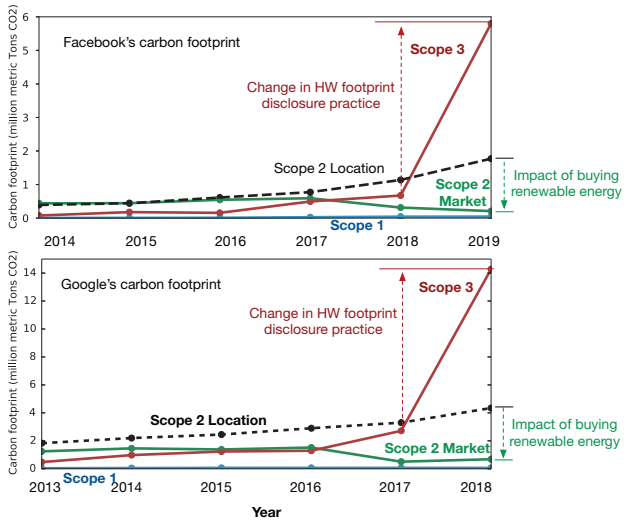

Fig. 11. Carbon footprint of Facebook and Google (two large data center operators). As data centers increasingly rely on renewable energy, carbon emissions originate more from Scope 3, or supply-chain emissions (e.g., hardware manufacturing and construction).

detailed accounting and reporting. Finally, we present a case study of renewable energy's effect on data-center footprint.

A. Breakdown of warehouse-scale data centers

Takeaway 7: *For data-center operators and cloud providers, most emissions are capex-related—for example, construction, infrastructure, and hardware manufacturing.*

Figure 11 illustrates the carbon footprint of Google (2013 to 2018) and Facebook (2014 to 2019) [16], [29]. Following the GHG Protocol, we split emissions into Scope 1 (blue), Scope 2 (green), and Scope 3 (red). Recall that Scope 1 (opex) emissions come from facility use of refrigerants, natural gas, and diesel; Scope 2 (opex) emissions come from purchased electricity; and Scope 3 (capex) emissions come from the supply chain, including employee travel, construction, and hardware manufacturing (see Section II for details).

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. Figure 12 illustrates Facebook's breakdown of Scope 3 emissions in 2019. Here, construction and hardware manufacturing (capital goods) account for up to 49% of the company's Scope 3 emissions.

Similarly, we anticipate most of Google's Scope 3 emissions are from construction and hardware manufacturing. Figure 11 shows that between 2017 and 2018, the company reported a 5× increase in that output. We attribute the large increase

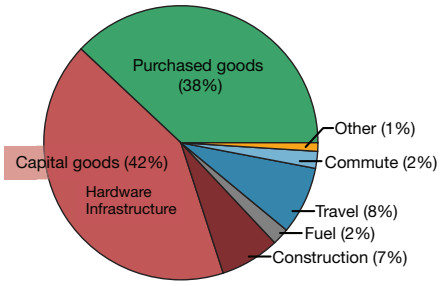

Fig. 12. Breakdown of Facebook's 2019 Scope 3 carbon emissions. Capital goods (e.g., hardware, infrastructure) and construction account for up to 49% of the annual total [16]

to the additional accounting and disclosure of hardware-manufacturing emissions—during that time, data-center energy consumption only increased by 30% [29]. Given the additional disclosure, the proportion of capex-related emissions increases compared with opex-related emissions. Note that because industry disclosure practices are evolving, publicly reported Scope 3 carbon output should be interpreted as a lower bound. The varying guidelines highlight the importance of better carbon accounting and reporting.

B. Impact of renewable energy

To decrease operational carbon emissions, data centers are increasingly employing renewable energy. Here we detail the impact of renewable energy on overall emissions and the breakdown between opex- and capex-related factors.

Takeaway 8: *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 11 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 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.

Takeaway 9: *For hardware, the carbon footprint between hardware manufacturing and use depends on the energy source. Powering hardware with renewable energy reduces emissions from operational energy consumption; consequently,*

Source	Carbon intensity (g CO_2/kWh)	Energy-payback time (months)
Coal	820	2 [33]
Gas	490	1 [33]
Biomass	230	∼12 [73]
Solar	41	∼36 [34]
Geothermal	38	72 [74]
Hydropower	24	∼12–36 [33], [75]
Nuclear	12	2 [33]
Wind	11	≤12 [35]

TABLE II
CARBON EFFICIENCY OF VARIOUS RENEWABLE-ENERGY SOURCES.

Geographic average	Carbon intensity (g CO_2 / kWh)	Dominant energy source
World	301	–
India	725	Coal/gas
Australia	597	Coal
Taiwan	583	Coal/gas
Singapore	495	Gas
United States	380	Coal/gas
Europe	295	–
Brazil	82	Wind/hydropower
Iceland	28	Hydropower

TABLE III
GLOBAL CARBON EFFICIENCY OF ENERGY PRODUCTION [14], [76], [77].

hardware manufacturing begins to dominate the carbon footprint.

Figure 13 illustrates the impact of renewable energy on opex- and capex-related emissions on the basis of reported carbon data from Intel (top) and AMD (bottom) [26], [28]. The format mimics hardware life cycles. Carbon emissions from device use over a three-year lifetime appear in green; emission from hardware manufacturing appear in red. Although Intel and AMD both assume the average US energy-grid mix, we scaled the hardware-use emissions in accordance with variable energy sources. Recall that warehouse-scale data centers often employ solar, hydropower, wind, and other renewable-energy sources.

The use of renewable energy dramatically changes the breakdown of carbon emissions across the hardware life cycle. For the baseline, assuming the US energy grid, roughly 60% of Intel's carbon emissions and 45% of AMD's come from hardware use and energy consumption. With renewable energy, however, emissions from operational consumption decrease. This decline is because renewable energy is orders of magnitude more efficient in grams of CO_2 emitted per kilowatt-hour of energy generated, as Table II shows. Table III lists the carbon intensity of energy production globally. The baseline case assumes the US energy grid (301 g CO_2 per kilowatt-hour); solar and wind emit 41 g and 11 g of CO_2 per kilowatt-hour, respectively. Figure 13 shows that when using solar and wind, which frequently power data centers, 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.

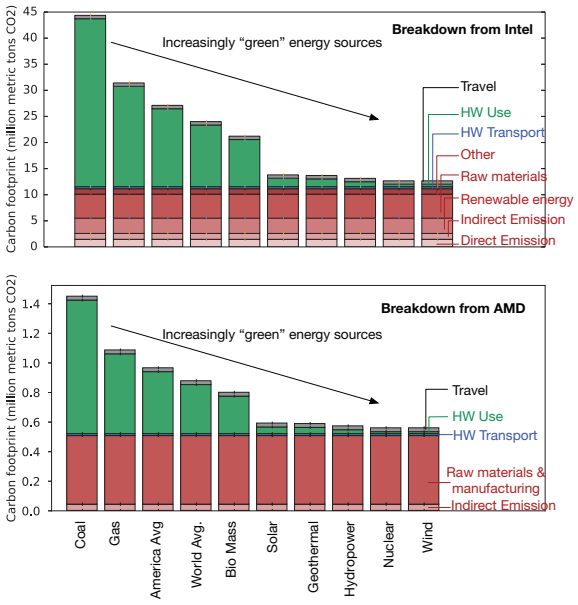

Fig. 13. Reported carbon-footprint breakdown for Intel (top) and AMD (bottom) as renewable energy increasingly (from left to right) powers hardware operation. The use of renewable energy reduces carbon emissions dramatically; most of the remaining emissions are from manufacturing.

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 10: *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 14 shows the carbon breakdown for wafer manufacturing at TSMC [31]. 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 (see Table II). 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 14 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 [31]. Intel already uses renewable energy to meet much of its demand; only 9.7% of the energy consumed by Intel fabs comes from nonrenewable sources (see Figure 13). Recall

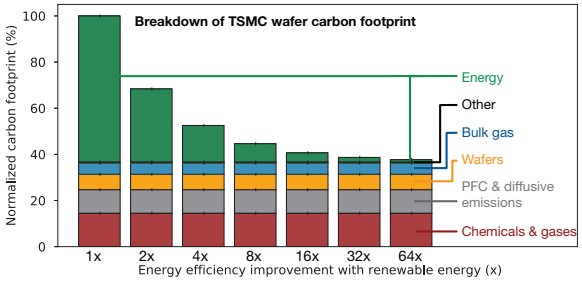

Fig. 14. Carbon-emissions breakdown for TSMC wafer manufacturing. Renewable energy provides up to a 64× reduction in emissions from electricity, and overall emissions for wafers drops by 2.7×. Although the reduction will reduce the carbon output of manufacturing, consideration of capex-related emissions for mobile and data-center hardware will remain important.

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.

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 (opex) and hardware manufacturing (capex). Given its immediate importance and scale, we must adopt vertically integrated research methods to minimize the emissions associated with computing. Figure 15 illustrates some of the important research directions and their corresponding layers in the computing stack. This section outlines future directions across the computing stack in light of the state-of-the-art and prior work. We use AI training and inference as an example application, but the tradeoffs extend to other domains.

Applications and algorithms. Application- and algorithm-level optimizations can reduce both opex- and capex-related carbon emissions. As a motivating example, consider AI training. The parameters of the energy footprint for training AI models are threefold: the footprint of processing one example (E), the data-set size (D), and the hyperparameter search (H) [15], [78]. Even though some data centers employ renewable energy (see Section IV), researchers must still consider the hardware's carbon footprint.

Table IV presents the compute, memory, and carbon footprint characteristics for two Apple Mac Pro desktops. Compared with the first configuration, the second has dual AMD Radeon Vega GPUs, providing 4×, 8×, and 16× more flops, memory bandwidth, and capacity, respectively. It represents a data-center-scale server [79], [80], yielding a 2.6× greater manufacturing carbon footprint.

Reducing carbon emissions requires training on systems with fewer resources. Given the faster–than–Moore's Law AI improvement, novel methods to train models given lesser compute and storage capabilities can directly reduce associated carbon emissions (i.e., E, D). Similarly, reducing

Parameter	Mac Pro 1	Mac Pro 2
CPU (cores × threads)	8×2	28×2
DRAM (GB)	32	1,536
Storage (GB)	256	4,096
GPU performance (teraflops)	6.2	28.4
GPU-memory BW (GB/s)	256	2,048
System TDP (W)	310	730
Manufacturing CO_2 (kg)	700	1,900

TABLE IV
Comparing Apple Mac Pro desktops. The high-performance configuration—more cores, memory, storage, GPU flops, and GPU memory bandwidth—has a 2.7× higher manufacturing-related CO_2 [82].

the hyperparameter-search factor (H) reduces the necessary number of parallel training nodes and the AI-training carbon footprint [78], [81]. Generally, algorithmic optimizations for scale-down systems will drastically cut emissions.

Run-time systems. Run-time systems, including schedulers, load-balancing services, and operating systems, can reduce both opex- and capex-related carbon footprints. Optimizing for opex-related emissions, cloud providers use machine learning to improve the efficiency of data-center cooling infrastructure [83]. Furthermore, recent work proposes scheduling batch-processing workloads during periods when renewable energy is readily available [84]–[88]. Doing so decreases the average carbon intensity (see Table II) of energy consumed by data-center services.

Optimizing for capex-related emissions requires reducing hardware resources. Recently proposed schedulers optimize infrastructure efficiency in terms of total power consumption while balancing performance for latency-critical and batch-processing workloads [79], [80], [89]–[91]. Others have proposed novel scheduler designs to enable energy-efficient, collaborative cloud and edge execution [69], [92]. Our analysis enables future studies to co-optimize for tail-latency-, throughput-, power-, and infrastructure-related carbon emissions.

Systems. Systems researchers can guide overall mobile- and data-center-scale system provisioning to directly reduce capex-related emissions. Recently, systems have scaled up and out to boost performance and power efficiency [93], [94]. Figure 8 and Figure 9 show more-sophisticated hardware for improving mobile-AI performance and power efficiency; simultaneously, scale-up hardware increases manufacturing- and device-level carbon emissions. Instead, as Table IV shows, scaling systems down can reduce the environmental impact of hardware—assuming they still provide sufficient performance.

Data centers often comprise heterogeneous platforms, with custom hardware for important applications, to maximize infrastructure efficiency (e.g., performance and power). For instance, Facebook data centers have custom servers for AI inference and training [95]. Our work enables systems researchers to consider how heterogeneity can reduce carbon footprint by reducing overall hardware resources in the data center. But researchers must balance the environmental benefits with the challenges of heterogeneity: programmability, resource management, cross-ISA execution, and debugging [96].

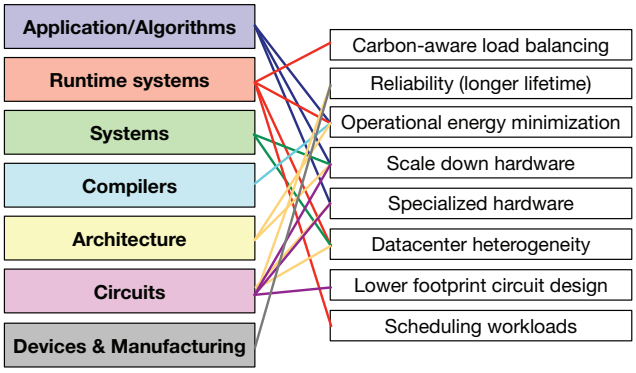

Fig. 15. Reducing the carbon output of computer systems requires cross-layer optimization across the computing stack. The potential opportunities (right) overlap with multiple stack layers (left).

Compilers and programming languages. Recent work has proposed new programming languages to enable more-energy-efficient code [97]. Others propose compiler-level optimizations to increase the code's energy efficiency [98]–[100]. Although optimizing nonfunctional properties, such as minimizing energy consumption through instruction-level energy annotation [101]–[103], can indirectly mitigate environmental impact, future compiler and programming-language technologies can optimize for CO_2 emissions directly by applying our emission analysis.

Architecture. Over the past 20 years, computer-architecture researchers and designers have devoted substantial effort to energy-efficient mobile and data-center systems. Their work includes system-performance increases, power management and energy-efficiency optimization through dynamic voltage and frequency scaling (DVFS) [99], [100], [104]–[114], and specialized hardware [4]–[9]. As Figure 9 and Figure 10 show, architectural improvements have considerably reduced the operational carbon footprint of mobile AI inference.

Future architecture research can also minimize capex-related carbon emissions. For instance, as Table IV shows, higher-performance hardware incurs higher manufacturing-related carbon emissions. More generally, as billion-transistor devices experience low utilization, systems must balance dark silicon with manufacturing emissions [24]. Similarly, architectural optimizations can directly reduce CO_2 output by judiciously provisioning resources, scaling down hardware, and incorporating specialized circuits.

Circuits. In addition to architectural innovations, circuit designers have enabled high-performance and low-power hardware through efforts that include clock/power gating [115], DVFS [116], and circuit-level timing speculation [117]. These efforts indirectly minimize opex-related carbon emissions.

Future circuit research can also reduce capex-related carbon emissions. First, it may consider circuit-level resource provisioning to balance performance, area, energy efficiency, and carbon footprint. Next, DRAM and NAND-flash-memory research should investigate low-carbon technologies and higher reliability to lengthen hardware lifetimes. Finally, vertically integrated research into specializing low-carbon circuits for salient applications will also decrease capex-related emissions. For example, in the case of AI, co-designing neural-network fault tolerance and compression for specialized circuits and memories can allow hardware consolidation and, therefore, smaller carbon footprints [4], [5], [7], [118], [119].

Semiconductor devices and manufacturing. Finally, capex emissions must be addressed through device modeling, characterization, design, and fab manufacturing. For instance, hardening a device's reliability and endurance extends its lifetime, cutting capex-related carbon emissions. Moreover, research into 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 should approach the environmental consequences of computing, from mobile to data-center-scale systems. First, we demonstrated that reducing energy consumption alone fails to reduce 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 becomes 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. We conclude by outlining future steps toward that goal.

Better accounting practices. Although many organizations publicly report their carbon emissions, improved accounting (e.g., broader participation as well as standardized accounting and disclosures) will provide further guidance on tackling salient challenges in realizing environmentally sustainable systems.

Carbon footprint as a first-order optimization target. Researchers and developers across the computing stack should consider carbon footprint to be a first-class design metric alongside increased workload and system characterization, and they should incorporate optimizations for lower environmental impact.

Beyond carbon emissions. Although this work focuses on carbon emissions, the environmental impact of computing systems is multifaceted, spanning water consumption as well as use of other natural resources, including aluminum, cobalt, copper, glass, gold, tin, lithium, zinc, and plastic.

VIII. Acknowledgements

We would like to thank Urvi Parekh, Stephanie Savage, and Julia Rogers for their valuable insights and many discussions on the environmental sustainability of technology companies. The collaboration leads to the deeper understanding of the challenges technology companies face in enabling environmentally sustainable operation presented in this work. We would also like to thank Kim Hazelwood for supporting and encouraging this work. The support helped foster new interdisciplinary collaborations on understanding and tackling the environmental impact of computing. Additionally, we would like to thank Srilatha Manne for the insightful discussions and valuable context for sustainable computing efforts across industry. Finally we would like to thank the anonymous reviewers for their detailed feedback on this work.

References

[1] A. S. Andrae and T. Edler, "On global electricity usage of communication technology: trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015.

[2] N. Jones, "How to stop data centres from gobbling up the world's electricity," *Nature*, vol. 561, no. 7722, pp. 163–167, 2018.

[3] P. J. Denning and T. G. Lewis, "Exponential laws of computing growth," *Communications of the ACM*, vol. 60, p. 54–65, Dec. 2016.

[4] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: Efficient inference engine on compressed deep neural network," in *Proceedings of the ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pp. 243–254, IEEE, 2016.

[5] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 52, no. 1, pp. 127–138, 2017.

[6] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1–12, IEEE, 2017.

[7] B. Reagen, P. Whatmough, R. Adolf, S. Rama, H. Lee, S. K. Lee, J. M. Hernández-Lobato, G.-Y. Wei, and D. Brooks, "Minerva: Enabling low-power, highly-accurate deep neural network accelerators," in *Proceedings of the ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pp. 267–278, IEEE, 2016.

[8] S. Zhang, Z. Du, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, and Y. Chen, "Cambricon-x: An accelerator for sparse neural networks," in *Proceedings of the 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 1–12, Oct 2016.

[9] S. Zhang, Z. Du, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, and Y. Chen, "Cambricon-x: An accelerator for sparse neural networks," in *Proceedings of the 49th Annual IEEE/ACM International Symposium on Microarchitecture*, p. 20, IEEE Press, 2016.

[10] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, *et al.*, "Anton, a special-purpose machine for molecular dynamics simulation," *Communications of the ACM*, vol. 51, no. 7, pp. 91–97, 2008.

[11] R. Hameed, W. Qadeer, M. Wachs, O. Azizi, A. Solomatnikov, B. C. Lee, S. Richardson, C. Kozyrakis, and M. Horowitz, "Understanding sources of inefficiency in general-purpose chips," in *Proceedings of the 37th annual International Symposium on Computer Architecture (ISCA)*, pp. 37–47, 2010.

[12] A. M. Caulfield, E. S. Chung, A. Putnam, H. Angepat, J. Fowers, M. Haselman, S. Heil, M. Humphrey, P. Kaur, J.-Y. Kim, *et al.*, "A cloud-scale acceleration architecture," in *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 1–13, IEEE, 2016.

[13] L. A. Barroso and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 4, no. 1, pp. 1–108, 2009.

[14] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *arXiv preprint arXiv:2002.05651*, 2020.

[15] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

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

[17] A. A. Chien, R. Wolski, and F. Yang, "The zero-carbon cloud: High-value, dispatchable demand for renewable power generators," *The Electricity Journal*, vol. 28, no. 8, pp. 110–118, 2015.

[18] F. Yang and A. A. Chien, "Zccloud: Exploring wasted green power for high-performance computing," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 1051–1060, 2016.

[19] F. Yang and A. A. Chien, "Large-scale and extreme-scale computing with stranded green power: Opportunities and costs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 5, pp. 1103–1116, 2018.

[20] Apple, "Product environmental report iphone 11," 2019.

[21] Apple, "iphone 3gs environmental report," 2009.

[22] J. Park, M. Naumov, P. Basu, S. Deng, A. Kalaiah, D. Khudia, J. Law, P. Malani, A. Malevich, S. Nadathur, *et al.*, "Deep learning inference in facebook data centers: Characterization, performance optimizations and hardware implications," *arXiv preprint arXiv:1811.09886*, 2018.

[23] M. Naumov, J. Kim, D. Mudigere, S. Sridharan, X. Wang, W. Zhao, S. Yilmaz, C. Kim, H. Yuen, M. Ozdal, *et al.*, "Deep learning training in facebook data centers: Design of scale-up and scale-out systems," *arXiv preprint arXiv:2003.09518*, 2020.

[24] 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.

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

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

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

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

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

[30] G. G. PROTOCOL, "The greenhouse gas protocol: A corporate accounting and reporting standard."

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

[32] J. Lee, "Tsmc's 3nm fab passed the environmental impact assessment," 2018.

[33] D. Weißbach, G. Ruprecht, A. Huke, K. Czerski, S. Gottlieb, and A. Hussein, "Energy intensities, erois (energy returned on invested), and energy payback times of electricity generating power plants," *Energy*, vol. 52, pp. 210–221, 2013.

[34] T. U. N. R. E. Laboratory, "Photovoltalics faqs," 2004.

[35] A. Bonou, A. Laurent, and S. I. Olsen, "Life cycle assessment of onshore and offshore wind energy-from theory to application," *Applied Energy*, vol. 180, pp. 327–337, 2016.

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

[37] Apple, "Product environmental report iphone 11 pro," 2019.

[38] Apple, "Product environmental report iphone 11 pro max," 2019.

[39] Apple, "Product environmental report iphone se," 2020.

[40] Apple, "iphone xr environmental report," 2018.

[41] Apple, "Product environmental report ipad air," 2019.

[42] Apple, "Product environmental report ipad," 2019.

[43] Apple, "Product environmental report ipad mini," 2019.

[44] Apple, "Product environmental report ipad pro (11 inch)," 2020.

[45] Apple, "Product environmental report ipad pro (12.9 inch)," 2020.

[46] Apple, "Apple watch series 3 (gps+cellular) environmental report," 2017.

[47] Apple, "Apple watch series 3 (gps) environmental report," 2017.

[48] Apple, "Apple watch series 5 environmental report," 2019.

[49] Apple, "Product environmental report 13-inch macbook air with retina display," 2020.

[50] Apple, "Product environmental report 13-inch macbook pro," 2020.

[51] Apple, "Product environmental report 16-inch macbook pro," 2019.

[52] Apple, "Product environmental report mac mini," 2018.

[53] Apple, "Product environmental report mac pro," 2019.

[54] Google, "Google pixel 2 product environmental report," 2017.

[55] Google, "Google pixel 2xl product environmental report," 2017.

[56] Google, "Google pixel 3a product environmental report," 2019.

[57] Google, "Google pixel 3 xl product environmental report," 2018.

[58] Google, "Google pixel 3a xl product environmental report," 2019.
[59] Google, "Google pixel 3 product environmental report," 2018.
[60] Google, "Google home product environmental report," 2016.
[61] Google, "Google home hub product environmental report," 2018.
[62] Google, "Google home mini product environmental report," 2017.
[63] Google, "Google pixelbook go product environmental report," 2019.
[64] Geekbench, "Geekbench 5," 2019.
[65] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
[66] Huawei, "Product environmental information," 2020.
[67] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
[68] M. S. Inc., "High voltage power monitor," 2020.
[69] Y. G. Kim and C.-J. Wu, "Autoscale: Optimizing energy efficiency of end-to-end edge inference under stochastic variance," 2020.
[70] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.
[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.
[72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1–9, 2015.
[73] K. Madsen and N. S. Bentsen, "Carbon debt payback time for a biomass fired chp plant—a case study from northern europe," *Energies*, vol. 11, no. 4, p. 807, 2018.
[74] K. Li, H. Bian, C. Liu, D. Zhang, and Y. Yang, "Comparison of geothermal with solar and wind power generation systems," *Renewable and Sustainable Energy Reviews*, vol. 42, pp. 1464–1474, 2015.
[75] Varun, R. Prakash, and I. K. Bhat, "Life cycle energy and ghg analysis of hydroelectric power development in india," *International Journal of Green Energy*, vol. 7, no. 4, pp. 361–375, 2010.
[76] "electricitymap," 2020.
[77] S. Bhawan and R. Puram, "Co2 baseline database for the indian power sector," 2018.
[78] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *arXiv preprint arXiv:1907.10597*, 2019.
[79] U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, *et al.*, "The architectural implications of facebook's dnn-based personalized recommendation," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 488–501, IEEE, 2020.
[80] U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-H. S. Lee, D. Brooks, and C.-J. Wu, "Deeprecsys: A system for optimizing end-to-end at-scale neural recommendation inference," *arXiv preprint arXiv:2001.02772*, 2020.
[81] D. Hernandez and T. B. Brown, "Measuring the algorithmic efficiency of neural networks," *arXiv preprint arXiv:2005.04305*, 2020.
[82] Apple, "Product environmental report: Mac pro," 2019.
[83] "DeepMind AI reduces google data centre cooling bill by 40%," 2020.
[84] A. Radovanovic, "Our data centers now work harder when the sun shines and wind blows," 2020.
[85] K. Le, R. Bianchini, T. D. Nguyen, O. Bilgir, and M. Martonosi, "Capping the brown energy consumption of internet services at low cost," in *International Conference on Green Computing*, pp. 3–14, IEEE, 2010.
[86] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, (New York, NY, USA), p. 175–186, Association for Computing Machinery, 2012.
[87] C. Li, A. Qouneh, and T. Li, "Characterizing and analyzing renewable energy driven data centers," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pp. 131–132, 2011.
[88] A. Gujarati, S. Elnikety, Y. He, K. S. McKinley, and B. B. Brandenburg, "Swayam: distributed autoscaling to meet slas of machine learning inference services with resource efficiency," in *Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference*, pp. 109–120, 2017.
[89] S. Kanev, K. Hazelwood, G.-Y. Wei, and D. Brooks, "Tradeoffs between power management and tail latency in warehouse-scale applications," in *2014 IEEE International Symposium on Workload Characterization (IISWC)*, pp. 31–40, IEEE, 2014.
[90] Q. Wu, Q. Deng, L. Ganesh, C.-H. Hsu, Y. Jin, S. Kumar, B. Li, J. Meza, and Y. J. Song, "Dynamo: facebook's data center-wide power management system," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 469–480, 2016.
[91] A. Sriraman, A. Dhanotia, and T. F. Wenisch, "Softsku: optimizing server architectures for microservice diversity@ scale," in *Proceedings of the 46th International Symposium on Computer Architecture (ISCA)*, pp. 513–526, 2019.
[92] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," *ACM SIGARCH Computer Architecture News*, vol. 45, no. 1, pp. 615–629, 2017.
[93] I. Magaki, M. Khazraee, L. V. Gutierrez, and M. B. Taylor, "Asic clouds: Specializing the datacenter," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pp. 178–190, IEEE, 2016.
[94] B. Grot, D. Hardy, P. Lotfi-Kamran, B. Falsafi, C. Nicopoulos, and Y. Sazeides, "Optimizing data-center tco with scale-out processors," *IEEE Micro*, vol. 32, no. 5, pp. 52–63, 2012.
[95] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro, *et al.*, "Applied machine learning at facebook: A datacenter infrastructure perspective," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 620–629, IEEE, 2018.
[96] C. Delimitrou, "The increasing heterogeneity of cloud hardware and what it means for systems," 2020.
[97] A. Sampson, W. Dietl, E. Fortuna, D. Gnanapragasam, L. Ceze, and D. Grossman, "Enerj: Approximate data types for safe and general low-power computation," in *Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI, (New York, NY, USA), p. 164–174, Association for Computing Machinery, 2011.
[98] E. Schulte, J. Dorn, S. Harding, S. Forrest, and W. Weimer, "Post-compiler software optimization for reducing energy," in *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS, (New York, NY, USA), p. 639–652, Association for Computing Machinery, 2014.
[99] Q. Wu, M. Martonosi, D. W. Clark, V. J. Reddi, D. Connors, Y. Wu, J. Lee, and D. Brooks, "A dynamic compilation framework for controlling microprocessor energy and performance," in *Proceedings of the 38th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO 38, (USA), p. 271–282, IEEE Computer Society, 2005.
[100] Qiang Wu, M. Martonosi, D. W. Clark, V. J. Reddi, D. Connors, Youfeng Wu, Jin Lee, and D. Brooks, "Dynamic-compiler-driven control for microprocessor energy and performance," *IEEE Micro*, vol. 26, no. 1, pp. 119–129, 2006.
[101] Y. S. Shao and D. Brooks, "Energy characterization and instruction-level energy model of intel's xeon phi processor," in *Proceedings of the 2013 International Symposium on Low Power Electronics and Design (ISLPED)*, 2013.
[102] D. Pandiyan and C. J. Wu, "Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms," in *2014 IEEE International Symposium on Workload Characterization (IISWC)*, 2014.
[103] A. Arunkumar, E. Bolotin, D. Nellans, and C. Wu, "Understanding the future of energy efficiency in multi-module gpus," in *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 519–532, 2019.
[104] D. Shingari, A. Arunkumar, B. Gaudette, S. Vrudhula, and C. Wu, "Dora: Optimizing smartphone energy efficiency and web browser performance under interference," in *2018 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 64–75, 2018.
[105] Y. Wang, Q. Xie, A. Ammari, and M. Pedram, "Deriving a near-optimal power management policy using model-free reinforcement learning and

bayesian classification," in *Proceedings of the 48th Design Automation Conference*, DAC, (New York, NY, USA), p. 41–46, Association for Computing Machinery, 2011.
[106] Q. Qiu and M. Pedram, "Dynamic power management based on continuous-time markov decision processes," in *Proceedings of the 36th Annual ACM/IEEE Design Automation Conference*, DAC, (New York, NY, USA), p. 555–561, Association for Computing Machinery, 1999.
[107] B. Gaudette, C.-J. Wu, and S. Vrudhula, "Optimizing user satisfaction of mobile workloads subject to various sources of uncertainties," *IEEE Transactions on Mobile Computing (TMC)*, vol. 18, no. 12, pp. 2941–2953, 2019.
[108] B. Gaudette, C. Wu, and S. Vrudhula, "Improving smartphone user experience by balancing performance and energy with probabilistic qos guarantee," in *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 52–63, 2016.
[109] C. Isci, A. Buyuktosunoglu, C. Cher, P. Bose, and M. Martonosi, "An analysis of efficient multi-core global power management policies: Maximizing performance for a given power budget," in *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 347–358, 2006.
[110] C. Isci, G. Contreras, and M. Martonosi, "Live, runtime phase monitoring and prediction on real systems with application to dynamic power management," in *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 359–370, 2006.
[111] A. Buyuktosunoglu, S. Schuster, D. Brooks, P. Bose, P. W. Cook, and D. H. Albonesi, "An adaptive issue queue for reduced power at high performance," in *Proceedings of the First International Workshop on Power-Aware Computer Systems-Revised Papers*, PACS, (Berlin, Heidelberg), p. 25–39, Springer-Verlag, 2000.
[112] V. Srinivasan, D. Brooks, M. Gschwind, P. Bose, V. Zyuban, P. N. Strenski, and P. G. Emma, "Optimizing pipelines for power and performance," in *35th Annual IEEE/ACM International Symposium on Microarchitecture, 2002. (MICRO). Proceedings.*, pp. 333–344, 2002.
[113] K. K. Rangan, G.-Y. Wei, and D. Brooks, "Thread motion: Fine-grained power management for multi-core systems," in *Proceedings of the 36th Annual International Symposium on Computer Architecture*, ISCA, (New York, NY, USA), p. 302–313, Association for Computing Machinery, 2009.
[114] R. Teodorescu and J. Torrellas, "Variation-aware application scheduling and power management for chip multiprocessors," in *2008 International Symposium on Computer Architecture*, pp. 363–374, 2008.
[115] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, and P. Bose, "Microarchitectural techniques for power gating of execution units," in *Proceedings of the 2004 International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 32–37, 2004.
[116] W. Kim, M. S. Gupta, G.-Y. Wei, and D. Brooks, "System level analysis of fast, per-core dvfs using on-chip switching regulators," in *2008 IEEE 14th International Symposium on High Performance Computer Architecture (HPCA)*, pp. 123–134, IEEE, 2008.
[117] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, *et al.*, "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proceedings. 36th Annual IEEE/ACM International Symposium on Microarchitecture, 2003 (MICRO).*, pp. 7–18, IEEE, 2003.
[118] L. Pentecost, M. Donato, B. Reagen, U. Gupta, S. Ma, G.-Y. Wei, and D. Brooks, "Maxnvm: Maximizing dnn storage density and inference efficiency with sparse encoding and error mitigation," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 769–781, 2019.
[119] U. Gupta, B. Reagen, L. Pentecost, M. Donato, T. Tambe, A. M. Rush, G.-Y. Wei, and D. Brooks, "Masr: A modular accelerator for sparse rnns," in *2019 28th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, pp. 1–14, IEEE, 2019.