FOCAL: A First-Order Carbon Model to Assess Processor Sustainability

Lieven Eeckhout
Ghent University

Abstract

Sustainability in general and global warming in particular are grand societal challenges. Computer systems demand substantial materials and energy resources throughout their entire lifetime. A key question is how computer engineers and scientists can reduce the environmental impact of computing. To overcome the inherent data uncertainty, this paper proposes FOCAL, a parameterized first-order carbon model to assess processor sustainability using first principles. FOCAL's normalized carbon footprint (NCF) metric guides computer architects to holistically optimize chip area, energy and power consumption to reduce a processor's environmental footprint. We use FOCAL to analyze and categorize a broad set of archetypal processor mechanisms into strongly, weakly or less sustainable design choices, providing insight and intuition into how to reduce a processor's environmental footprint with implications to both hardware and software. A case study illustrates a pathway for designing strongly sustainable multicore processors delivering high performance while at the same time reducing their environmental footprint.

CCS Concepts: • **Computer systems organization** → **Superscalar architectures**; **Multicore architectures**; • **Computing methodologies** → **Modeling methodologies**.

Keywords: Computer Architecture, Sustainability, Modeling

ACM Reference Format:
Lieven Eeckhout. 2024. FOCAL: A First-Order Carbon Model to Assess Processor Sustainability. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24), April 27-May 1, 2024, La Jolla, CA, USA*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3620665.3640415

1 Introduction

As the world's population and average affluence per capita continues to increase, the world-wide demand for resources, both materials and energy, continues to grow. The extraction of raw materials, the manufacturing of products, transportation, usage, and finally depletion and recycling requires huge amounts of energy, most often provided by fossil fuels. This in turn leads to global warming and climate change as a result of increased greenhouse gas (GHG) emissions, which has now reached such proportions that we need to act.

Freitag et al. [14] report that information and communication technology (ICT) contributes for 2.1 to 3.9% of the world's GHG emissions — currently on par with the aviation industry — and is likely to increase (substantially) in the near future. To understand the environmental footprint of computing, Gupta et al. [20] provide a comprehensive carbon characterization of mobile devices, always-connected computers, and hyperscale datacenters. They conclude that most emissions related to personal mobile (battery-operated) devices and datacenter equipment come from hardware manufacturing and infrastructure, the so-called *embodied carbon footprint*. In contrast, for always-connected personal devices, most emissions come from the energy usage during a device's lifetime, the so-called *operational footprint*.

Understanding the carbon footprint of computing raises the question how to reduce it. This is a non-trivial question because of inherent data uncertainty. There are many unknowns due to lack of accurate and reliable data sources, industry secrecy, non-transparent supply chains, hard-to-predict usage patterns, rebound effects, changes in power grid mix, etc.

This work embraces the inherent data uncertainty by proposing FOCAL, a First-Order analytical CArbon modeL to assess processor sustainability while being deliberately simple and flexible [11]. FOCAL's primary goal is to provide insight, intuition and guidance for computer architects in research and early-stage development of sustainable processors. FOCAL is based on first principles, using proxies for the embodied and operational footprints that relate to what computer architects have control over at design time. In particular, the proxy for embodied emissions is chip area, while the proxy for operational emissions is energy and power consumption assuming a fixed-work and fixed-time scenario, respectively. The model further includes a parameter to weigh the relative importance of the embodied versus

ASPLOS '24, April 27-May 1, 2024, La Jolla, CA, USA

ACM ISBN 979-8-4007-0385-0/24/04
https://doi.org/10.1145/3620665.3640415

operational footprint to account for variation in product use and lifetime, and to anticipate the infamous rebound effect of increased usage.

FOCAL computes the *normalized carbon footprint (NCF)* metric enabling computer architects to holistically optimize chip area, energy and power consumption for improving overall processor sustainability. Making a distinction between the fixed-work and fixed-time scenarios enables assessing whether a design is strongly, weakly or less sustainable depending on whether it reduces the environmental impact under all circumstances, specific circumstances, or under no circumstances, respectively. We use FOCAL to assess to what extent archetypal processor design choices are sustainable, with implications to both hardware and software. We conclude that while some processor mechanisms are strongly sustainable (e.g., low-complexity core microarchitecture, multicore, voltage scaling), others are only weakly sustainable (e.g., heterogeneity, speculation), or not sustainable (e.g., turboboosting, dark silicon). Combining FOCAL with chip manufacturing carbon footprint data, we demonstrate a pathway for designing strongly sustainable multicore processors across technology nodes delivering higher performance while at the same time reducing total carbon footprint.

2 Inherent Data Uncertainty

This work is motivated by the observation that developing a detailed sustainability model for computer architects to steer the design process is extremely complex and involved, if at all possible. There is inherent uncertainty in modeling the environmental footprint due to data limitations and various unknowns [19]. While some contributing factors are known and can be accounted for to some extent, such as the use of materials and energy as well as the amount of chemicals and gases emitted during manufacturing, others are unknown, or at least, there is substantial uncertainty about the specific values [16]. Furthermore, the operational footprint depends on the user, the intensity of use, product lifetime, and geographic location of the user which determines the power grid mix. Operational footprint hence needs to be estimated using historical data of similar products [3].

To make things even worse, improving the efficiency of a device often has the unintended side-effect of a rebound effect, also known as Jevons' paradox [2], which essentially means that an improvement in efficiency leads to an increase in demand (thereby increasing the embodied footprint) and/or usage (thereby increasing the operational footprint), which ultimately leads to a net increase in the environmental impact which the designer originally intended to reduce.

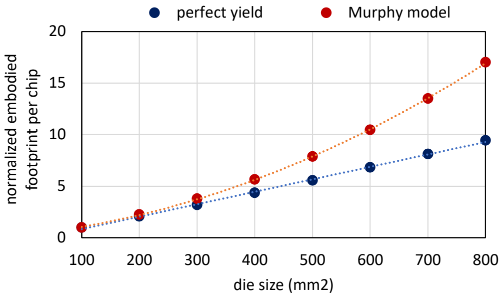


Figure 1. Embodied footprint per chip as a function of die size for a 300 mm wafer assuming perfect yield and the Murphy yield model; the trendlines show linear and second-degree polynomial approximations, respectively.

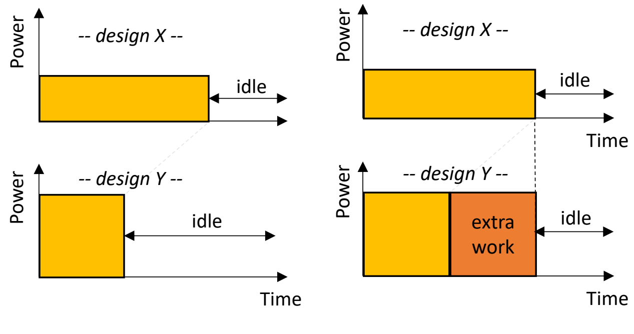
3 FOCAL: First-Order Carbon Modeling

In light of the inherent data uncertainty, we propose FOCAL, a model built upon first principles for computer architects to gain insight and reason about the sustainability implications of design choices without being tampered with inaccurate and missing data, and unknown and unintended side-effects and parameters. FOCAL uses first-order proxies for the embodied and operational footprints, and a parameter to weigh their relative importance.

3.1 Embodied Footprint

The embodied footprint of chip manufacturing consists of three major components, following the GHG Protocol [41]. Scope-2 refers to the amount of energy needed during production. Imec [16] recently analyzed the amount of energy needed for a range of CMOS technology nodes from 28 nm to 3 nm, and reports that the annual growth rate in energy per wafer is estimated to be around 11.9% as a result for increased complexity with increasing number of process steps, increasing number of metal layers, new extreme ultraviolet lithography (EUV) equipment, etc. Scope-1 refers to the emissions of chemicals and gases including fluorinated compounds (e.g., SF_6, NF_3 and CF_4, among others), which is estimated to increase by 9.3% per year. Scope-3 refers to the carbon emissions during raw material extraction and processing both upstream and downstream along the manufacturing process.

The unit of production in a semiconductor fabrication plant is a wafer, and what we, computer architects, have control over when it comes to embodied footprint is chip size and design complexity. A big chip (large die size) means fewer chips per wafer, and thus a larger embodied footprint (and also cost) per chip. In contrast, a small chip implies more chips per wafer, and thus a smaller per-chip embodied footprint. The number of chips per wafer, and the embodied footprint per chip, thus depends, to the first order, on the chip's die size. de Vries [10] provides a formula that empirically derives the number of *chips per wafer (CPW)* as a



(a) fixed-work **(b)** fixed-time

Figure 2. Operational footprint is proportional to (a) energy consumption under a fixed-work scenario, i.e., the highlighted area which is the product of power and execution time, and (b) power consumption under a fixed-time scenario, i.e., the highlighted areas are proportional to power consumption because execution time is constant.

function of die size A:

$$CPW = \frac{\pi d^2}{4A} - 0.58\frac{\pi d}{\sqrt{A}},$$

with d the wafer's diameter (e.g., 300 mm). The embodied footprint per chip is inversely proportional to CPW.

Figure 1 shows the embodied footprint per chip for a 300 mm wafer as a function of die size in the region of practical concern, up to 800 mm^2 (close to the reticle limit) and normalized to 100 mm^2. The two curves represent perfect yield (i.e., all dies are good dies) versus a Murphy yield model [30] assuming a 0.09 defect density per cm^2 which is achievable for volume production processes according to TSMC [44]. Indeed, the larger a chip's die size and the larger its complexity, the lower the yield. In practice, to maximize profit, industry increases the effective yield by turning off or bypassing defective circuit blocks in large chips, selling those chips as lower-performance, lower-power products. In fact, profit is maximized when all defective chips can be sold as alternative products, thereby approaching the perfect yield model curve. FOCAL therefore uses die size as a proxy for a processor's embodied footprint.

3.2 Operational Footprint

The operational footprint relates to the total energy consumed by a processor over its entire lifetime. To anticipate different use cases, we consider fixed-work and fixed-time scenarios.

The *fixed-work* scenario assumes that a processor performs a fixed amount of work during its entire lifetime. As illustrated in Figure 2(a), design X takes more time to get the same amount of work done than design Y but the latter consumes more power. Which design incurs the smallest operational footprint is determined by the amount of energy consumed or the product of power consumption and execution time, i.e., the highlighted areas in Figure 2(a). Under the fixed-work scenario, the proxy for the operational footprint thus equals energy consumption. Examples of a fixed-work scenario are strong-scaling workloads on a supercomputer, or a video decoder on a mobile device that decodes a fixed number of frames per unit of time.

The *fixed-time* scenario overcomes the (simplifying) assumption by the fixed-work scenario that the amount of work done by a device is fixed for its entire lifetime. Instead, the fixed-time scenario assumes that a more efficient device performs more work and is used for the same amount of time as a less efficient device. As illustrated in Figure 2(b), design Y achieves higher performance than design X at the cost of consuming more power. This implies that Y can perform 'extra work' within the same amount of time as X. Because time is constant, total energy consumption — and thus the operational footprint — is proportional to power consumption, and thus the height of the highlighted areas in Figure 2(b). As a result, under the fixed-time scenario, the proxy for operational footprint is power consumption. Examples of a fixed-time scenario are weak-scaling workloads on a supercomputer, an always-on network interface, or a data center system where improved performance leads to the deployment of new applications filling the freed up idle time. (Note that the lifetime operational footprint of an idle device is also proportional to its idle (leakage) power.)

Which of the two scenarios is most representative depends on the anticipated use case, for which the appropriate proxy should be used. However, many designs do not strictly fall under only a single scenario and/or the typical use case may be (largely) unknown at design time. In such cases, both scenarios can and should be considered. In particular, one may expect that a device's operational carbon footprint is in line with the fixed-work scenario at the bare minimum, and, because of increased usage due to rebound effects, with the fixed-time scenario in many practical use cases.

3.3 Embodied versus Operational Footprint

To assess a processor's total footprint one has to weigh the embodied and operational footprint. The relative importance of the embodied versus operational footprint is hard to assess though at design time as it depends on a number of factors. For one, and as aforementioned, the ratio varies across devices [20]. Second, the ratio also depends on the lifetime of the device, i.e., the longer the lifetime of the device, the higher weight the operational footprint carries in the total footprint, and the less significant the embodied footprint is. Third, and as alluded to before, the rebound effect may increase the usage of more efficient devices, possibly increasing the relative importance of operational emissions. Finally, whether green energy sources are used during product manufacturing and/or product lifetime also affects the relative ratio of the embodied versus operational footprint. Note that even if manufacturing would be done using only green energy, the embodied footprint still incurs a substantial environmental impact as a result of the materials used (scope-3) and the chemicals and gases emitted during manufacturing (scope-1).

FOCAL weighs the relative importance of the embodied versus operational footprint using a parameter called the *embodied-to-operational (E2O) weight* α_{E2O}. How to set the α_{E2O} parameter depends on the anticipated use case, which is known approximately at best, and yet we need to make a holistic assessment considering both the embodied and operational footprints and potential rebound effects. Moreover, the relative weight may even change when comparing different designs due to increased usage, shift in energy mix, etc. It is hence advised to consider multiple scenarios and ranges of the weighing factor α_{E2O} to understand the sustainability impact of a particular design despite the inherent data uncertainty. Based on Gupta et al. [20], we consider a scenario where the embodied footprint dominates ($\alpha_{E2O} = 0.8 \pm 0.1$) as well as a scenario where the operational footprint dominates ($\alpha_{E2O} = 0.2 \pm 0.1$), see also Section 5.

3.4 Normalized Carbon Footprint (NCF) Metric

To compare the environmental footprint of alternative designs, FOCAL computes the *Normalized Carbon Footprint (NCF)* metric as a weighted sum of the normalized embodied footprint and operational footprint, with the weight α_{E2O} determined by the anticipated use case. This leads to the NCF metrics $NCF_{fw,\alpha_{E2O}}$ and $NCF_{ft,\alpha_{E2O}}$ for the fixed-work and fixed-time scenarios, respectively, and a given embodied-to-operational weight α_{E2O}. FOCAL uses the proxies for embodied and operational footprint as previously derived. This leads to the definition for NCF when comparing designs X versus Y under a fixed-work scenario:

$$NCF_{fw,\alpha_{E2O}}(X,Y) = \alpha_{E2O}\frac{A_X}{A_Y} + (1-\alpha_{E2O})\frac{E_X}{E_Y},$$

and under a fixed-time scenario:

$$NCF_{ft,\alpha_{E2O}}(X,Y) = \alpha_{E2O}\frac{A_X}{A_Y} + (1-\alpha_{E2O})\frac{P_X}{P_Y}.$$

The normalized embodied footprint is computed as the ratio of chip areas A for X and Y, while the normalized operational footprint is computed as the ratio of the energy consumption E and power consumption P under the fixed-work and fixed-time scenarios, respectively. An NCF value below one implies that X incurs a lower footprint than Y; an NCF value above one implies that X incurs a higher footprint than Y.

What sets sustainability apart from other optimization objectives is that computer architects should consider chip area, energy consumption and power consumption holistically. Indeed, computer architects commonly take these optimization targets into account, but the objective and focus has not been to minimize the overall environmental impact. Moreover, computer architects do not (typically) trade off embodied versus operational footprint, e.g., incur larger embodied footprint to reduce operational footprint, or vice versa. FOCAL makes these design trade-offs explicit, and provides an intuitive tool to gain insight into how design decisions impact sustainability.

3.5 Comparison against ACT

The ACT model [19] is a recently proposed empirically-based, data-driven carbon model that quantifies the carbon footprint of a computing device in absolute terms using a variety of data sources including semiconductor fabs and industrial sustainability reports. While FOCAL is built on the same principled foundations as ACT, i.e., considering both the embodied and operational footprint, FOCAL takes a different approach by focusing on first principles in light of the inherent data uncertainty. FOCAL should hence not be considered an alternative to ACT, but rather as a useful complement to gain insight, steer decisions in early stages of the design process, and assess the sustainability of new research and development proposals.

The key difference is that FOCAL is a top-down, parameterized model in contrast to ACT which is a bottom-up, data-driven approach. As acknowledged by prior work — including the ACT work — there is inherent data uncertainty, putting a fundamental limit on what can be achieved with a data-driven approach. This is why the FOCAL model intentionally starts from first principles while considering (1) different use-case scenarios (fixed-work versus fixed-time), (2) ranges of embodied-versus-operational weights, and (3) potential rebound effects. This enables powerful analyses despite inherent data uncertainty: if we are reaching similar conclusions across a range of scenarios and embodied-to-operational footprint weights, we can be confident that the conclusions hold true despite the unknowns. If different conclusions are reached across scenarios (e.g., with and without rebound effects), the overall conclusion is that we need to be more cautious when formulating overall conclusions.

3.6 Model Validation

Validating a sustainability model is extremely challenging, if at all possible, because of the lack of detailed and accurate data — again, this is exactly why FOCAL is a parameterized first-order analytical model. To be able to validate the FOCAL model, one would need precise data regarding the carbon footprint of individual processor chips, which is currently not available unfortunately. It would be greatly beneficial if processor manufacturers would publish detailed carbon characterizations of the processors they bring to market, which would enable precise validation of architecture carbon models pushing research and development towards more sustainable processors. The closest available data includes

Life Cycle Assessment (LCA) reports of existing systems, aggregating the total footprint of the entire system into a single number, making it impossible to assess the carbon footprint of individual components including processor chips.

The ACT model [19] was validated but the authors acknowledge "*a non-negligible gap*" between LCA-based carbon cost estimates and ACT. They further hypothesize that the difference comes from the "*lack of up-to-date carbon emission data for the latest compute [...] technologies*". In contrast, by building upon first principles and by considering ranges of model parameters and use-case scenarios, the FOCAL model enables comprehensive sustainability analyses despite the inherent data uncertainty.

3.7 Limitations

While it is important to understand a model's strengths, it is equally important to understand its limitations. First, as mentioned above, FOCAL is less detailed than ACT and should therefore not be used to make fine-grained design trade-offs when detailed numbers are available. Instead, it can be used to gain insight and understand major design choices. The proxies used by FOCAL are first-order approximations. Regarding the embodied footprint, while a small chip die area may emit less carbon during fabrication, it may lead to more carbon being emitted during other steps in the production process (such as assembly, testing, packaging, etc.) — prior work indicates though that semiconductor manufacturing dominates the embodied footprint [47], hence the choice for chip area as a first-order proxy. Regarding the operational footprint, the actual operational footprint of a device is a complex function of its idle time, the workloads it runs, its lifetime, its usage (possibly subject to rebound effects), etc. Processor performance (possibly) degrades over time, also affecting the operational footprint. Anyhow, a chip's operational footprint is proportional to its power and energy consumption, hence the choice for power and energy consumption as first-order proxies under the fixed-time and fixed-work scenarios, respectively.

Second, one has to understand the broader context. Reducing the environmental footprint of computing is not just an engineering challenge, it also touches upon market dynamics and economic business models, as well as policy and legislation. While FOCAL allows for understanding the environmental footprint of an individual processor, it does not capture market dynamics which, due to Jevons' paradox, may lead to increased deployment. As mentioned before, rebound effects can happen in various ways: efficiency improvements can lead to increased usage of existing devices (thereby increasing the operational footprint of individual devices) and/or can lead to increased deployment (thereby increasing the overall embodied footprint of producing more devices). In the FOCAL model, the former is captured via the fixed-time scenario, while the latter can be modeled by changing the embodied-to-operational weight. Estimating the magnitude and form of the rebound effect is hard, if at all possible. In a market mostly driven by sales, companies are not incentivized to reduce the environmental footprint of their processors if that may hurt their profit margin. This is where new business models, legislation and policy come in to challenge the processor industry to reduce their environmental footprint. Similarly, investors and electronics makers keen on reporting green supply chains to end customers, may further push processor manufacturers to ramp up action and tackle their climate footprint. Understanding the impact of a processor's reduced environmental footprint on market dynamics falls beyond the scope of this work though.

4 Strong versus Weak Sustainability

FOCAL's key strength is to explore the design space and understand how gross design choices affect a processor's environmental footprint under a variety of scenarios despite the inherent data uncertainty. In fact, the notion of the fixed-work versus fixed-time scenarios provides a unique opportunity to assess whether a processor design choice is strongly, weakly, or less sustainable. Or in other words, whether a design choice is sustainable under all circumstances (including the rebound effect of increased usage), specific circumstances (assuming constant work only), or no circumstances, respectively.

More specifically, we define a system X to be *strongly sustainable* compared to system Y if it yields a lower total footprint under both the fixed-work and fixed-time scenarios, i.e., $NCF_{fw}(X,Y) < 1$ $\underline{\text{and}}$ $NCF_{ft}(X,Y) < 1$. The intuitive meaning of strong sustainability is that design X is always more sustainable than system Y, under all circumstances, even under a rebound effect where system X performs more work than Y. We define a system X to be *weakly sustainable* compared to system Y if it yields a lower carbon footprint under the fixed-work scenario but not the fixed-time scenario, i.e., $NCF_{fw} < 1$ $\underline{\text{and}}$ $NCF_{ft} > 1$, or vice versa, i.e., $NCF_{fw}(X,Y) > 1$ $\underline{\text{and}}$ $NCF_{ft}(X,Y) < 1$. Weakly sustainable implies that X is more sustainable than Y under specific circumstances, but definitely not all, i.e., it is subject to a rebound effect. Finally, we define a system X to be *less sustainable* than Y if its carbon footprint is larger under both the fixed-work and fixed-time scenarios, i.e., $NCF_{fw}(X,Y) > 1$ $\underline{\text{and}}$ $NCF_{ft}(X,Y) > 1$.

5 Archetypal Processor Design Choices

We now use FOCAL to assess the environmental footprint of archetypal processor design choices. This analysis is done using a variety of analytical performance and power models, first-order approximations, and previously published results.[1]

[1] A welcome corollary is that this methodology minimizes the environmental footprint of this research compared to a simulation-based setup typically seen in architecture papers.

We consider the following processor design choices: multicore, heterogeneity, hardware acceleration, dark silicon, caching, core microarchitecture, speculation, frequency/voltage scaling, and power/energy saving. In the discussion to follow, we make a distinction between a scenario where the embodied footprint dominates (i.e., $\alpha_{E2O} = 0.8$) versus a scenario where the operational footprint dominates (i.e., $\alpha_{E2O} = 0.2$). To further account for modeling error and data uncertainty, we also report error bar ranges for $\alpha_{E2O} \in [0.7, 0.9]$ and $\alpha_{E2O} \in [0.1, 0.3]$, respectively. These scenarios are based on the empirical findings by Gupta et al. [20] who report that the total carbon footprint of mobile battery-operated devices and hyperscale-datacenter servers is mostly dominated by the embodied footprint while the carbon footprint of always-connected devices is mostly dominated by the operational footprint.

5.1 Multicore

We rely on Amdahl's Law and its extensions to evaluate the impact of multicore on processor sustainability. Following Hill and Marty [23], we assume a chip of N base core equivalents (BCEs). A multi-core processor consisting of N cores, i.e., one BCE per core, yields the following speedup over a one-BCE single-core processor, assuming that a fraction f of the sequential execution can be parallelized:

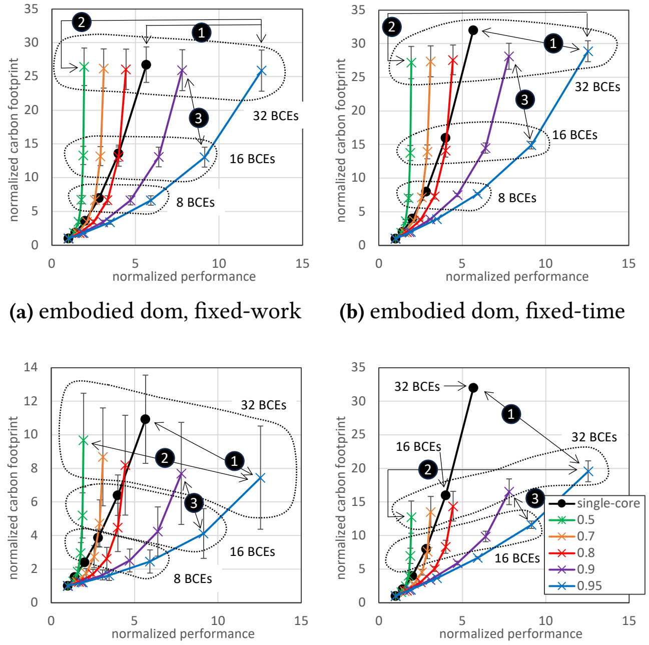
$$S = \frac{1}{(1-f) + \frac{f}{N}}. \tag{1}$$

To compute multicore power and energy consumption, we use the extensions to Amdahl's Law by Woo and Lee [50]. During serial execution, a single core is active and consumes one unit of power, while the other $N - 1$ cores consume γ power each due to leakage ($0 < \gamma < 1$). Hence, during the serial execution phase, which takes a fraction $(1 - f)$ of the execution time, the multicore processor consumes $1 + (N - 1)\gamma$ units of power, i.e., one active core consumes one unit of power and $N - 1$ cores consume γ leakage power. During the parallel execution phase, which takes a fraction f/N of the execution time, all cores are active and consume one unit of power. A multicore's average power consumption hence equals the power consumed during serial execution plus the power consumed during parallel execution divided by the total execution time, i.e., serial plus parallel execution phases:

$$\begin{aligned} P &= \frac{(1-f)(1+(N-1)\gamma) + N\frac{f}{N}}{(1-f)+\frac{f}{N}} \\ &= \frac{1+(1-f)(N-1)\gamma}{(1-f)+\frac{f}{N}}. \end{aligned} \tag{2}$$

Energy consumption is power consumption times execution time, i.e., Equation 2 divided by Equation 1, and hence equals

$$E = 1 + (1-f)(N-1)\gamma. \tag{3}$$



(a) embodied dom, fixed-work **(b)** embodied dom, fixed-time

(c) operational dom, fixed-work **(d)** operational dom, fixed-time

Figure 3. Comparing (symmetric) multicore against single-core with varying number of BCEs (from 1 to 32 in powers of two) and varying degrees of parallelism f from 0.5 to 0.95. Carbon footprint and performance are normalized to a one-BCE single-core processor. *Multicore is strongly sustainable, and is beneficial for performance and sustainability when software is highly parallel.*

In contrast, a big single-core processor consisting of N BCEs (same chip area as the multicore we compare against) delivers a speedup of $\sqrt{N}$ over a one-BCE single-core processor, following Pollack's rule [7]. If we assume that one BCE consumes one unit of power, this N-BCE single-core processor consumes N units of power, and its energy consumption equals $\sqrt{N}$, i.e., power consumption divided by speedup.

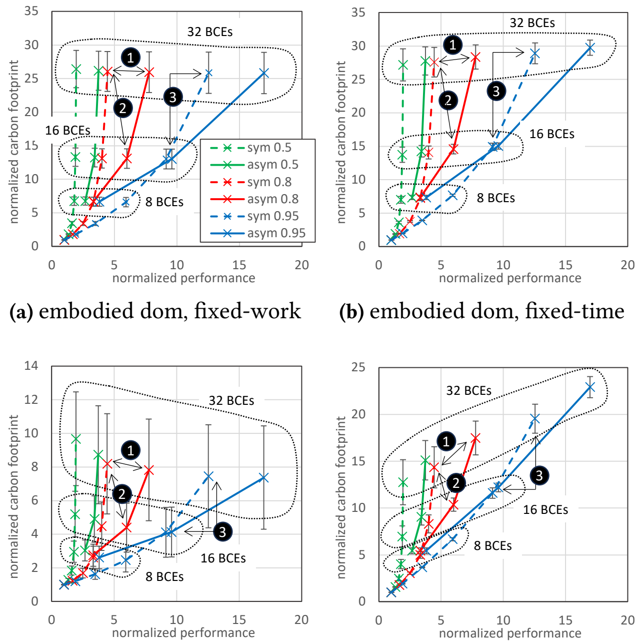
Figure 3 reports normalized carbon footprint as a function of performance for four scenarios: when the embodied and operational footprint dominate (subfigures (a) and (b) versus subfigures (c) and (d), respectively) and under the fixed-work and fixed-time scenarios (subfigures (a) and (c) versus subfigures (b) and (d), respectively); all results are normalized to a one-BCE single-core processor, while assuming $\gamma = 0.2$. We vary the number of BCEs from 1 to 32 in powers of two; the multi-core curves hence show results for 1, 2, 4, 8, 16 and 32 cores, and we consider different values of f ranging from 0.5 to 0.95 to denote varying degrees of parallel software.

Finding #1: Multicore is strongly sustainable, especially when the operational footprint dominates. Multicore reduces the carbon footprint under both the fixed-work and fixed-time scenarios compared to a single-core processor with the same chip area. The reduction in carbon footprint increases as the operational footprint increases in importance, see ❶ in Figure 3. For example, assuming 32 BCEs and $f = 0.95$ under the fixed-time scenario, multicore reduces the environmental footprint by 10% when the embodied footprint dominates, see Figure 3(b), and by 39% when the operational footprint dominates, see Figure 3(d).

Finding #2: Parallelizing software is weakly sustainable. Parallelizing software reduces the overall footprint under a fixed-work scenario while increasing it under a fixed-time scenario, see ❷ in Figure 3. This is particularly the case when the operational footprint dominates, see Figures 3(c) and (d): increasing the degree of parallelism from $f = 0.5$ to $f = 0.95$ reduces the carbon footprint by 23% under a fixed-work scenario while increasing it by 53% under a fixed-time scenario. Software parallelization is a two-edged sword: it reduces the footprint for a given processor iff this does not lead to increased usage.

Finding #3: Parallelizing software is a more sustainable way to improve performance than increasing the number of cores. Parallelizing software has the potential not only to improve performance — as is well known — but also to improve sustainability by enabling the processor to feature fewer cores (and thus be smaller in size), thereby reducing its embodied and overall footprint, see ❸ in Figure 3. For example, a multicore with 16 BCEs and highly parallel software ($f = 0.95$) yields 17% higher performance compared to a multicore that is twice as big (32 BCEs) and (slightly) less parallel software ($f = 0.9$). *At the same time,* the environmental footprint is reduced by 30%, see Figure 3(d), and up to 50%, see Figure 3(a). This insight implies that from a system's sustainability perspective, software has a critical role to play: parallelizing software should be preferred over adding more cores.

Discussion. Industry embraced multicore for power efficiency reasons, and this turns out to be a sustainable design choice as well. However, parallelizing software (e.g., using high-performance parallel programming frameworks, if possible) is more sustainable than adding cores. It seems though that, at least for general-purpose desktop computing, industry has pushed harder for adding cores than for parallelizing software [6], which turns out not to be the most sustainable design choice.



(a) embodied dom, fixed-work **(b)** embodied dom, fixed-time

(c) operational dom, fixed-work **(d)** operational dom, fixed-time

Figure 4. Comparing asymmetric multicores (one 4-BCE big core plus $N - 4$ one-BCE small cores) versus symmetric multicores with N one-BCE small cores. Carbon footprint and performance are normalized to a one-BCE single-core processor. *Multicore heterogeneity is weakly sustainable, and is beneficial for performance and sustainability when software lacks high parallelism.*

5.2 Asymmetric Multicore

Performance asymmetry or heterogeneity [18, 28] has been introduced in multicore designs by integrating one (or few) high-performance 'big' core(s) alongside low-power 'small' cores to improve performance in an energy-efficient way. Hill and Marty [23] extend Amdahl's Law for asymmetric multicores with N BCEs. Assuming that the one big core takes M BCEs, and hence achieves a level of performance of $\sqrt{M}$ (i.e., Pollack's rule), alongside $N - M$ small one-BCE cores with a performance of one, asymmetric multicore speedup over a one-BCE single-core processor equals:

$$S = \frac{1}{\frac{1-f}{\sqrt{M}} + \frac{f}{N-M}}. \tag{4}$$

Woo and Lee [50] estimate power consumption of an asymmetric multicore as follows. During serial execution, which takes $(1 - f)/\sqrt{M}$ time units, the big core consumes M units of power while the $N - M$ small cores are idle (consuming γ leakage power, $0 < \gamma < 1$). During parallel execution, which takes $f/(N - M)$ time units, the $N - M$ small cores consume one unit of power, while the big remains idle, hence consuming $M\gamma$ leakage power. Average power consumption thus equals:

$$P = \frac{\frac{1-f}{\sqrt{M}}(M + (N - M)\gamma) + \frac{f}{N-M}(M\gamma + (N - M))}{\frac{1-f}{\sqrt{M}} + \frac{f}{N-M}}. \tag{5}$$

Energy is obtained by dividing power consumption (Equation 5) with speedup (Equation 4):

$$E = \frac{1 - f}{\sqrt{M}}(M + (N - M)\gamma) + \frac{f}{N - M}(M\gamma + (N - M)). \tag{6}$$

Figure 4 reports normalized carbon footprint as a function of performance assuming that the asymmetric core consists of one 4-BCE big core (i.e., $M = 4$) alongside $N - 4$ small one-BCE cores. We consider three asymmetric multicore configurations with 8, 16 and 32 BCEs, respectively. We further consider three values for f: 0.5, 0.8 and 0.95.

Finding #4: Heterogeneity is weakly sustainable. Performance asymmetry reduces the carbon footprint under a fixed-work scenario while increasing the carbon footprint under a fixed-time scenario compared to a symmetric multicore of the same chip area, see ❶ in Figure 4. This is most notable when the operational footprint dominates: for 32 BCEs and $f = 0.8$, heterogeneity reduces the footprint by 4% under a fixed-work scenario, see Figure 4(c), while increasing the footprint by 22% under a fixed-time scenario, see Figure 4(d). Heterogeneity is hence a two-edged sword: it only reduces the overall footprint iff it does not lead to increased usage.

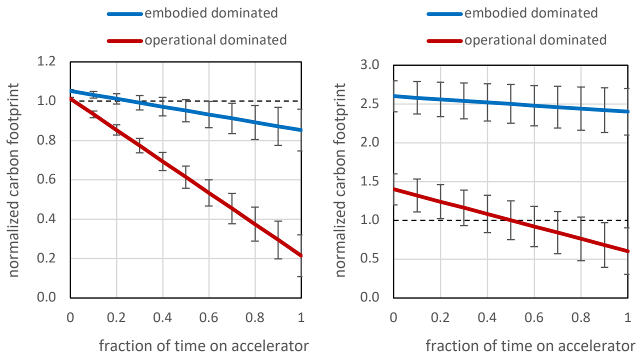
Finding #5: Heterogeneity is a sustainable way to improve performance only when software lacks high degrees of parallelism. Heterogeneity reduces the environmental footprint while *at the same time* improving performance if software is modestly parallel ($f \leq 0.8$), see ❷ in Figure 4. For example, an asymmetric multicore with 16 BCEs and $f = 0.8$ yields 35% higher performance compared to a symmetric multicore that is twice as big (32 BCEs) *while at the same time* reducing the environmental footprint ranging from 28%, see Figure 4(d), to 50%, see Figure 4(a).

In contrast, when software is highly parallel ($f = 0.95$), while the 16-BCE asymmetric multicore reduces the carbon footprint between 38%, see Figure 4(d), and 50%, see Figure 4(a), it also degrades performance by 23.5% compared to a 32-BCE symmetric multicore, see ❸ in Figure 4. In other words, heterogeneity is beneficial from a sustainability perspective iff software lacks high degrees of parallelism.

Discussion. Industry embraced heterogeneity for energy efficiency reasons, see for example ARM's big.LITTLE system [18] and Intel's Alder Lake CPU [43]. Heterogeneous multicores turn out to be a sustainable design choice when software offers limited parallelism, which appears to be the case for some classes of workloads, e.g., mobile [15], desktop [6].

5.3 Hardware Acceleration

Hardware acceleration is widely seen as a way to continue to improve performance in a power- and energy-efficient way in the post-Dennard era. As an example, Hameed et al. [21] propose an H.264 accelerator that incurs 6.5% extra chip area when delivering similar performance and consuming 500× less energy compared to an out-of-order (OoO) core. Figure 5(a) reports the total footprint for the OoO core plus accelerator, normalized to the OoO core without the accelerator, as a function of the fraction of time spent on the accelerator when embodied emissions dominate ($\alpha_{E2O} = 0.8$) and operational emissions dominate ($\alpha_{E2O} = 0.2$).



(a) 6.5% extra chip area **(b)** 200% extra chip area

Figure 5. Total footprint of hardware specialization (normalized to OoO core) for an accelerator that incurs (a) 6.5% extra chip area and (b) twice as much chip area. *Hardware acceleration is strongly sustainable if the operational footprint dominates; if the embodied footprint dominates, the accelerator needs to be extensively used for it to be sustainable. Dark silicon is not sustainable.*

Finding #6: Hardware acceleration is strongly sustainable if the operational footprint dominates. If the embodied footprint dominates, acceleration only improves sustainability if the accelerator is extensively used. The total footprint of the OoO core plus accelerator reduces relative to the OoO core without the accelerator as the accelerator is used more intensively, i.e., the increased embodied footprint gets offset by the decrease in operational footprint. When the operational emissions dominate, the hardware accelerator reduces the overall footprint even when used for only a small fraction of the time and significant savings can be obtained when extensively used. For example, if the accelerator is used 50% of the time, the environmental footprint reduces by 60%, see Figure 5(a). However, when the embodied emissions dominate, the accelerator needs to be used intensively to amortize the increased embodied footprint. For this particular example, the accelerator needs to be used for more than 30% of the time for it to lead to a net saving in overall footprint.

Discussion. Because the embodied footprint dominates for mobile devices [20], it seems unlikely that hardware acceleration leads to a net reduction in environmental footprint. Hardware acceleration drastically reduces the operational footprint, which is a sustainable choice only if it does offset the increase in embodied emissions.

5.4 Dark Silicon

A modern-day processor often times is a system-on-chip (SoC) featuring several tens of accelerators [24]. Not all accelerators can be powered on all the time due to power constraints — a phenomenon called dark silicon [13, 49]. To evaluate the impact of dark silicon on sustainability, we now assume that the tens of accelerators occupy two thirds of the entire chip, and that — same assumption as in previous section — when in use, each accelerator consumes 500× less energy for the same level of performance and, when not in use, an accelerator does not incur any leakage power, see Figure 5(b).

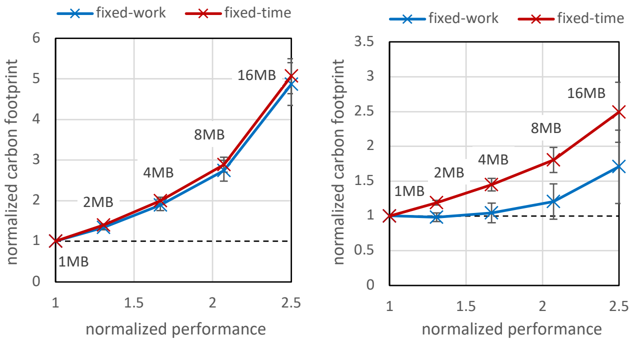
Finding #7: Dark silicon is not sustainable. If the embodied footprint dominates (again, the likely case today [20]), it is clear that dark silicon leads to a substantial ∼ 2.5× increase in total carbon footprint. If the operational footprint dominates, dark silicon should be used (very) intensively (more than 50% of the time) to amortize the embodied footprint to reduce the overall footprint. This might not be feasible, simply because it is *dark* silicon and cannot be powered on all at the same time within the available power and thermal budget.

Discussion. Industry has embraced dark silicon and hardware specialization, however, this is not a sustainable design choice. Instead of having many fixed-function accelerators, it might be more sustainable to design reconfigurable accelerators to amortize the embodied footprint across multiple applications.

5.5 Caching

Caches take up a significant fraction of the total chip area today. The larger the cache, the larger the embodied footprint. A larger cache leads to a reduced miss rate, hence higher performance and fewer accesses to the next level in the hierarchy. If the increase in energy consumption of a larger cache is offset by the reduction in energy consumption in the next level of the hierarchy (due to fewer misses), this leads to a net reduction in energy consumption. If the reduction in energy consumption outweighs the improvement in performance, this also leads to a reduction in power consumption.

To evaluate the impact of caching on sustainability, we use area and energy results from CACTI 5.1 [48] for a last-level caches (LLCs) ranging from 1 MB to 16 MB in powers of two assuming a 65 nm technology node. We assume that the 1 MB LLC occupies 25% of the core chip area; according to CACTI, the LLC size increases by a factor 20.7× from 1 MB to 16 MB. A cache access consumes between 0.55 nJ (1 MB) to 2.9 nJ (16 MB). We further assume a memory-intensive workload that spends 80% of its energy and execution time waiting for memory with a 1 MB LLC, and we follow the empirical rule that cache miss rate scales following a square-root of its size [22]. A reduction in miss rate leads to a proportional reduction of the memory stall time. As a sanity check, LLC chip area of a 2 MB LLC is approximately equally large as the entire core as is the case for the AMD Renoir CPU [4], and power consumption in memory takes up between 25% to 40% of total server power [5, 31, 34].



(a) embodied dominated **(b)** operational dominated

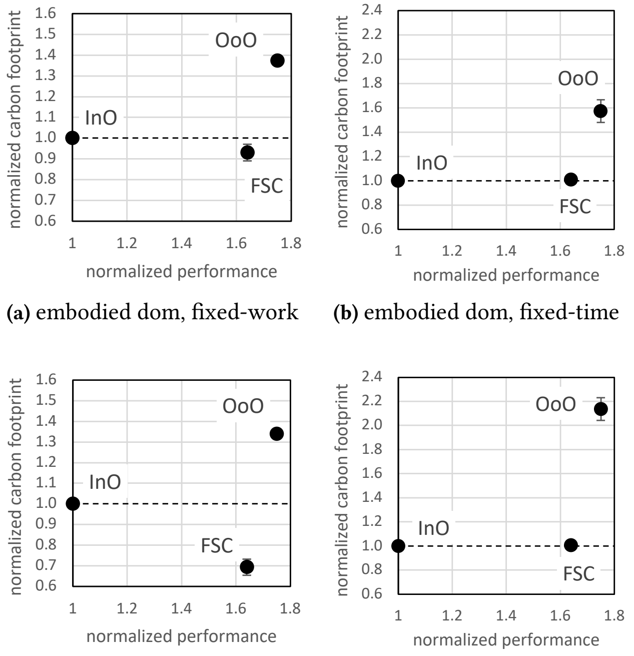
Figure 6. Sustainability impact of last-level caches: NCP as a function of cache size. *Caching is not sustainable or (marginally) weakly sustainable if the operational footprint dominates.*

Finding #8: Caching is not sustainable when the embodied footprint dominates. Caching is (marginally) weakly sustainable if the operational footprint dominates and when the reduction in energy consumption offsets the area increase in terms of its carbon impact. As noted in Figure 6, the increase in embodied footprint is substantial for larger caches, which makes caching not sustainable when the embodied footprint dominates. When the operational footprint dominates on the other hand, caching may lead to a net reduction in overall carbon footprint under a fixed-work scenario for relatively small cache sizes for which the reduction in energy consumption offsets the increase in chip area.

Discussion. Caching is critical to performance, but not sustainable, unfortunately. Paradigms that bring computation where the data is, i.e., processing-in-memory, may reduce the need for large caches and therefore be more sustainable.

5.6 Core Microarchitecture

We now evaluate how core microarchitecture affects sustainability by considering three microarchitectures: (1) an in-order (InO) core, (2) a Forward Slice Core (FSC) [29], and (3) an out-of-order (OoO) core. FSC is a state-of-the-art slice-out-of-order core that achieves a level of performance that is comparable to OoO at a small area and power overhead over InO by featuring in-order issue queues that operate out-of-order with respect to each other. We take the chip area, power, energy and performance numbers from [29] for comparing InO, FSC and OoO: 64% and 75% higher performance is achieved for FSC and OoO compared to InO for 1% and 39% extra chip area, and 1% and 2.32× higher power consumption, respectively. All three microarchitectures operate at the same 2 GHz clock frequency, feature the same cache hierarchy and superscalar width (2-wide). Area and power numbers were obtained using McPAT [32] and CACTI v6.5 [35] assuming 22 nm.



(a) embodied dom, fixed-work **(b)** embodied dom, fixed-time

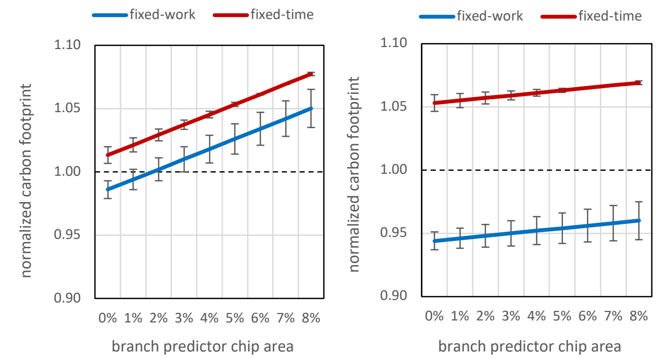
(c) operational dom, fixed-work **(d)** operational dom, fixed-time

Figure 7. Comparing InO, FSC and OoO microarchitectures in terms of environmental footprint versus performance. *An OoO core is less sustainable than an InO core. FSC is (close to being) strongly sustainable compared to InO.*

Figure 7 reports the NCF (normalized carbon footprint) for InO, FSC and OoO as a function of normalized performance assuming a fixed-work scenario, when (a) the embodied emissions dominate ($\alpha_{E2O} = 0.8$), and (b) the operational emissions dominate ($\alpha_{E2O} = 0.2$). Subfigures (c) and (d) report similar results under a fixed-time scenario. Design points towards the bottom-right are optimal, i.e., highest performance and lowest environmental footprint.

Finding #9: OoO cores are less sustainable than InO cores. Inversely, InO is strongly sustainable compared to OoO. OoO cores are less sustainable than InO cores under both the fixed-work and fixed-time scenarios, i.e., the carbon footprint of an OoO core is larger than an InO core. Obviously, OoO cores yield higher performance. As a result, OoO and InO cores represent different trade-offs in the design space: an OoO core yields higher performance at the cost of a higher carbon footprint; in contrast, an InO core delivers lower performance but also incurs a smaller carbon footprint.

Finding #10: A low-complexity core such as FSC is (very close to being) strongly sustainable compared to InO. FSC achieves a lower total footprint than InO under a fixed-work scenario. Under a fixed-time scenario, FSC's footprint is slightly higher than InO, but only barely so.



(a) embodied dominated **(b)** operational dominated

Figure 8. Sustainability impact of branch prediction: NCF as a function of branch predictor chip area. *Branch prediction is weakly sustainable if the operational footprint dominates, and not sustainable when embodied emissions dominate (assuming that the branch predictor takes up more than 2% of core chip area).*

Finding #11: FSC is strongly sustainable compared to OoO. FSC offers an interesting sustainability-performance trade-off compared to OoO: the environmental footprint is 32% to 53% smaller (depending on the scenario) at a relatively small 6.3% degradation in performance.

Discussion. Industry opts for either InO or OoO cores. However, a complexity-effective core like FSC yields higher performance than InO for a similar environmental footprint, and a drastic reduction in environmental footprint at the cost of a small performance degradation compared to OoO.

5.7 Speculation

Branch Prediction. Although a branch predictor consumes additional energy, it leads to a net reduction in the total energy consumed by reducing the amount of useless work done by the processor for fetching, decoding, renaming, issuing and executing wrong-path instructions. Parikh et al. [39] report that the largest hybrid branch predictor considered in their study reduces total CPU energy consumption by 7% compared to a small bimodal branch predictor, while improving performance by 14%. This implies that CPU power consumption increases by 6.6%. Unfortunately, Parikh et al. do not report the impact on chip area. Modern-day branch predictors incur multiple tens and even several hundreds of KBs of hardware overhead [17, 45]. This translates into a couple percent of the total core's die area devoted to the branch predictor [4];[2] a 64 KB TAGE-SC-L branch predictor is reported to take up 4.4% of a CPU's total chip area [40]. Figure 8 reports the normalized carbon footprint as a function of chip area.

[2] Based on annotated die shot photos, see for example https://www.techpowerup.com/268747/amd-renoir-die-annotation-raises-hopes-of-desktop-chips-featuring-x16-peg

Finding #12: Branch prediction is weakly sustainable when operational emissions dominate, and leads to a less sustainable system when embodied emissions dominate. Branch prediction reduces the overall footprint irrespective of its size only under a fixed-work scenario and when the operational footprint dominates. The branch predictor needs to be small when the embodied footprint dominates under a fixed-work scenario. Under a fixed-time scenario, dynamic branch prediction increases the overall footprint irrespective of its size.

Runahead Execution. Runahead prefetches future memory requests when the core stalls on a long-latency load [36]. Precise Runahead Execution (PRE) [37] is a state-of-the-art runahead technique improving performance by 38.2% compared to an out-of-order baseline core while at the same time reducing energy consumption by 6.8%; as a result, power consumption increases by 29.8%. PRE is reported to incur 1.24 KB of extra hardware (assumed area increase of 0.5%).

Finding #13: Runahead execution is weakly sustainable. Because the hardware overhead is small, the normalized carbon footprint of runahead execution is primarily a function of its energy and power consumption. Runahead execution reduces energy consumption but also increases power consumption, hence it is weakly sustainable. When the operational footprint dominates, PRE reduces the carbon footprint under a fixed-work scenario ($NCF_{fw,0.2} = 0.95$) while increasing the footprint under a fixed-time scenario ($NCF_{ft,0.2} = 1.23$). The same conclusion holds true when the embodied footprint dominates: $NCF_{fw,0.8} = 0.99$ and $NCF_{ft,0.8} = 1.06$.

Discussion. Speculation is a key technique to boost performance. However, it is weakly sustainable when the hardware overhead is small and operational emissions dominate, and less sustainable when the overhead is high and embodied emissions dominate. Making speculation more sustainable requires minimizing its hardware and power overheads.

5.8 Frequency and Voltage Scaling

Dynamically scaling frequency and voltage is a widely deployed mechanism to either save energy and power (by scaling down voltage and frequency), or boost performance (by scaling up) [25]. Dynamic power consumption scales cubically with voltage and frequency, while dynamic energy consumption scales quadratically. Leakage power scales linearly. On-chip voltage regulators incur a small chip area (no more than a couple percent over a core) [26].

Finding #14: DVFS is strongly sustainable. The increase in chip area is (most likely) offset by the reduction in energy and power consumption, leading to a net reduction in carbon footprint. DVFS might lead to a net increase in carbon footprint if the reduction in energy and power does not offset the increase in chip area (though unlikely).

Finding #15: Turboboosting leads to a less sustainable system. Boosting the clock frequency (and voltage) when there is thermal headroom [42] increases the carbon footprint by increasing energy and power consumption, on top of the extra chip area needed to implement the turboboost circuitry.

5.9 Power/Energy Saving

Manne et al. [33] proposed pipeline gating to reduce the amount of useless work due to branch mispredictions. A confidence predictor steers the number of instructions into the pipeline: when several low-confidence branches have been dispatched, the pipeline is gated to save power and energy by limiting the number of wrong-path instructions being fetched, decoded, renamed and executed. Parikh et al. [39] report that pipeline gating reduces energy consumption by 3.5% while degrading performance by 6.6%; power hence reduces by almost 10%. The confidence estimator incurs no additional hardware overhead when relying on the values of the saturating counters in a hybrid predictor.

Finding #16: Pipeline gating is strongly sustainable. By reducing both power and energy consumption at no extra hardware cost, the net carbon footprint reduces both when the embodied footprint dominates ($NCF_{fw,0.8} = 0.99$ and $NCF_{ft,0.8} = 0.98$) and when the operational footprint dominates ($NCF_{fw,0.2} = 0.97$ and $NCF_{ft,0.2} = 0.92$).

6 Die Shrink

The previous analysis implicitly assumed the same technology node. We now consider implementing an existing processor in a new chip technology node.

Finding #17: A die shrink is strongly sustainable. The 50% reduction in chip area across consecutive technology nodes is partially offset by the increase in energy consumption due to manufacturing in a new technology node — Imec [16] reports that the amount of energy consumed to manufacture a wafer (i.e., scope-2) increases by 25.2% between two consecutive tech nodes, while increasing the amount of chemicals and gases emitted (i.e., scope-1) by 19.5%. The increase in energy consumption and chemicals/gases is offset by the decrease in chip area. A die shrink hence leads to a net reduction in embodied footprint.

For the operational footprint, we make a distinction between classical scaling versus post-Dennard scaling [49]. Assuming classical scaling, power consumption reduces by a factor 2×, and because the circuit can be clocked at 1.41× higher frequency, energy consumption is reduced by a factor 2.82×. In other words, the operational footprint reduces under classical scaling, under both the fixed-work and the fixed-time scenarios. In contrast, under post-Dennard scaling, power consumption remains constant, while clock frequency is 1.41× higher and energy reduces by a factor 1.41×. This implies that the operational footprint reduces under a fixed-work scenario while remaining unchanged under a fixed-time scenario.

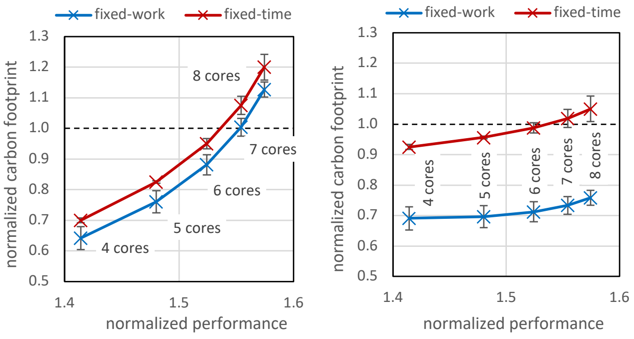
Discussion. This analysis implies that microprocessor chips would have become more sustainable over time if we would have leveraged Moore's Law to make our chips smaller. This is not what we have seen though in practice. Architects have used the additional transistors, which have become exponentially cheaper thanks to Moore's Law, when moving from one technology node to the next to add functionality (e.g., more cores, larger caches, accelerators, etc.), which has led to an overall increase in environmental footprint — this is yet another example of Jevons' paradox.

7 Sustainable Multicore Design

With these insights, we now consider a case study in which we explore the trade-offs in sustainability versus performance when designing a next-generation multicore processor in a new technology node. Ideally, a next-generation processor should deliver higher performance at a lower carbon footprint. One option is to keep the number of cores constant, i.e., implement a die shrink which will halve the chip area. Another option is to increase the number of cores; if we assume the same core microarchitecture, this implies that we can integrate twice as many cores in the next technology node for the same chip area. It appears that current practice is aligned with the latter option, more so than with the former. We now explore the impact on sustainability and performance for these two options (constant-core and constant-area) and intermediate options.

We consider a quad-core processor with 4 BCEs, i.e., one core occupies one BCE, in a current technology node. Moving to the next-generation technology node, we consider the options of integrating 4, 5, 6, 7 or 8 cores while preserving the core's microarchitecture. We further assume that the parallel workload is modestly parallel ($f = 0.75$); an idle core consumes $\gamma = 0.2$ leakage power while an active core consumes one unit of power. Because modern-day processors are power-constrained, we assume that power consumption in the new technology node is the same as in the old technology node. This implies that the achievable clock frequency reduces with the number of cores, i.e., clock frequency in the new technology node reduces from being 1.41× higher for 4 cores compared to the same architecture in the old technology node to being 1.24× higher for 8 cores.

Relative to the embodied carbon footprint of the 4-core option in the old technology node, the embodied carbon footprint of the 4-core option in the new technology node equals 0.625, i.e., chip area halves but the manufacturing footprint increases by 25.2%. In contrast, the 8-core option leads to a normalized embodied carbon footprint of 1.25, i.e., incurring the full increase in manufacturing footprint by maintaining the same chip area. Because power consumption is constant across the multicore options, the operational footprint is the same in the new technology node compared to the old one under a fixed-time scenario. However, under a fixed-work scenario, the operational energy consumption decreases due to the improved clock frequency and performance. As a result, the operational carbon footprint either remains constant or declines when transitioning to a new technology node.



(a) embodied dominated **(b)** operational dominated

Figure 9. Exploring the sustainability versus performance trade-off when designing a multicore processor in a new technology node; the different data points correspond to 4, 5, 6, 7 and 8 cores. *The design points with 4, 5 and 6 cores are strongly sustainable, while the 7- and 8-core design points are weakly (or not) sustainable.*

Figure 9 reports the impact on total carbon footprint. When the operational footprint dominates, see Figure 9(b), there is a net decrease in footprint for all multicore options under a fixed-work scenario. However, under a fixed-time scenario, the overall footprint for 7 and 8 cores exceeds the footprint of the 4-core processor in the old technology node; limiting the number of cores to 4, 5 or 6 leads to a net overall footprint reduction (or stagnation for 6 cores). In other words, 4, 5 or 6 cores are a strongly sustainable design options, while the 7- or 8-core designs are weakly sustainable. When the embodied footprint dominates, see Figure 9(a), there is a net reduction in total footprint when limiting the number of cores to at most 6; implementing 7 or 8 cores leads to a net footprint increase, even under a fixed-work scenario. This implies that the designs with 4, 5 and 6 cores are strongly sustainable, while the designs with 7 and 8 cores are not sustainable.

Discussion. It is interesting to note that the sober design options with 4, 5 or 6 cores are strongly sustainable under both scenarios. Note further that they also offer a substantial performance improvement in the 1.41× to 1.52× range for the specific example in Figure 9. In contrast, the more aggressive design options with 7 or 8 cores are not (or only weakly) sustainable, i.e., they may potentially lead to an increased overall carbon footprint depending on the scenario. The question is whether the relatively small performance improvement for 7 and 8 cores outweighs the increase in overall carbon footprint. *A microprocessor market that primarily focuses on performance gears the industry towards the non-sustainable design options with more cores. This case study* *suggests though that there is pathway to continue to improve multicore processor performance in a strongly sustainable way.*

8 Related Work

A couple studies have proposed methodologies and tools to explore the computing system design space from a sustainability perspective, but none has proposed a first-order and parameterized carbon model like FOCAL that embraces uncertainty and accounts for a variety of use-case scenarios — prior work implicitly considers a fixed-work scenario only, i.e., does not account for rebound effects, and assumed a fixed embodied-to-operational footprint ratio. The work closest related to FOCAL is the ACT model [19], previously discussed and compared against in Section 3.5. GreenChip [27] is a tool for evaluating the total carbon footprint of computing systems, which computes the indifference point or the point in time when a new device will reach the same total carbon footprint as the system it will replace, indicating when it is beneficial from a sustainability perspective to upgrade the system. Using GreenChip, Brunvand et al. [8] analyze the sustainability impact of processor configurations (increasing core, count and cache size) and dark silicon, and conclude that upgrading to larger core counts and cache sizes as well as dark silicon is worthwhile from a sustainability perspective only for high-performance computing systems for which the total carbon footprint is highly dominated by the operational footprint — this aligns with the conclusions reached in our analysis. CarbonExplorer [1] is a tool to explore the design space of sustainable datacenters while considering the deployment of renewable energy sources, batteries, and scheduling.

Other prior work analyzed the carbon footprint of existing designs from either a general scope or specific application-specific scope. Gupta et al. [20] quantify the carbon footprint for a range of computing devices, from personal devices to datacenter servers, by analyzing their LCAs. The embodied footprint of mobile personal devices and datacenter servers appears to outweigh their operational footprint. Personal always-connected devices appear to have a larger operational footprint. Wu et al. [51] analyze the operational and manufacturing carbon footprint of Artificial Intelligence (AI) workloads in hyperscale datacenters. Ollivier et al. [38] explore the design space encompassing GPUs, FPGAs, and processing-in-memory (PIM) for AI processing at the edge, concluding that while GPUs offer higher energy efficiency, their high embodied footprint makes them less sustainable than PIM-based solutions. Chang et al. [9] introduce the thermodynamic metric of exergy consumption, which is essentially equivalent to the embodied footprint, and which they use to analyze and optimize server design for reduced overall (embodied plus operational) environmental impact. Eeckhout [12] analyzes how current trends in microprocessor chip demand and manufacturing energy and carbon footprint affect the overall computer chip carbon footprint, and concludes that computer architects should primarily focus on designing smaller chips to reduce the embodied carbon footprint; reducing the operational footprint is of secondary importance, yet still significant. Switzer et al. [46] extend the lifetime of discarded smartphones by repurposing them into so-called 'junkyard' cloudlets; this amortizes the smartphones' embodied footprint over an extended lifetime. Zhang et al. [52] propose the performance per wafer metric to balance and trade off performance against cost and sustainability in multi-chip-module GPUs.

9 Conclusion

This paper proposed FOCAL, a parameterized carbon model based on first principles to drive processor design decisions. FOCAL offers valuable insight despite the inherent data uncertainty regarding sustainability. FOCAL uses proxies for embodied and operational emissions, considers fixed-work and fixed-time scenarios to account for rebound effects, and parameterizes the embodied-to-operational footprint ratio. FOCAL computes the normalized carbon footprint (NCF) metric to holistically optimize chip area, energy and power consumption for improving overall computer system sustainability. A variety of archetypal processor mechanisms were analyzed and categorized in strongly, weakly and less sustainable design choices. A case study illustrated how FOCAL can guide the design of future processors that deliver both higher performance and incur a smaller environmental impact by leveraging technology innovation in a sober way.

Acknowledgements

Thanks to the anonymous reviewers and shepherd, Benjamin C. Lee, for the valuable feedback. Thanks also to Udit Gupta, Hsien-Hsin (Sean) Lee, and Trevor E. Carlson for discussions about some of the aspects in this work. Finally, this work was supported in part by UGent-BOF-GOA grant No. 01G01421, Research Foundation Flanders (FWO) grant No. G018722N, and European Research Council (ERC) Advanced Grant agreement No. 741097.

A Artifact Appendix

A.1 Abstract

The artifact consists of an Excel data sheet containing the analytical models, first-order approximations, and previously published results based on which the graphs in the paper were generated.

A.2 Artifact check-list (meta-information)

- **Model:** formulas are implemented in Excel data sheet
- **Data set:** data set is provided in Excel data sheet
- **Publicly available:** yes

A.3 Description

The Excel data sheet can be downloaded from https://doi.org/10.6084/m9.figshare.25103501.v1 and contains several work sheets (accessible via the tabs at the bottom) for the various analyses in Sections 3, 5 and 7.

References

[1] B. Acun, B. Lee, F. Kazhamiaka, K. Maeng, U. Gupta, M. Chakkaravarthy, D. Brooks, and C.-J. Wu. Carbon explorer: A holistic framework for designing carbon aware datacenters. In *Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, volume 2, pages 118–132, 2023.

[2] B. Alcott. Jevons' paradox. *Ecological Economics*, 54(1), 2005.

[3] Apple. iPhone 12 product environmental report, 2020.

[4] S. Arora, D. Bouvier, and C. Weaver. AMD next generation 7nm Ryzen 4000 APU 'Renoir'. In *HotChips*, Aug. 2020.

[5] 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. Morgan and Claypool Publishers, 2009.

[6] G. Blake, R. G. Dreslinski, T. N. Mudge, and K. Flautner. Evolution of thread-level parallelism in desktop applications. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*, pages 302–313, June 2010.

[7] S. Borkar. Thousand core chips — a technology perspective. In *Proceedings of the Design Automation Conference (DAC)*, pages 746–749, June 2007.

[8] E. Brunvand, D. Kline, and A. K. Jones. Dark silicon considered harmful: A case for truly green computing. In *Proceedings of the International Green and Sustainable Computing Conference (IGSC)*, pages 1–8, June 2019.

[9] J. Chang, J. Meza, P. Ranganathan, A. Shah, R. Shih, and C. E. Bash. Totally green: Evaluating and designing servers for lifecycle environmental impact. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 25–35, Mar. 2012.

[10] D. K. de Vries. Investigation of gross die per wafer formulas. *IEEE Transactions on Semiconductor Manufacturing*, 18(1):136–139, 2005.

[11] L. Eeckhout. A first-order model to assess computer architecture sustainability. *IEEE Computer Architecture Letters (CAL)*, 21(2):137–140, July–Dec 2022.

[12] L. Eeckhout. Kaya for computer architects: Toward sustainable computer systems. *IEEE Micro*, 43:9–18, Jan/Feb 2023.

[13] H. Esmaeilzadeh, E. M. Blem, R. S. Amant, K. Sankaralingam, and D. Burger. Dark silicon and the end of multicore scaling. In *Proceedings of the IEEE/ACM International Symposium on Computer Architecture (ISCA)*, pages 365–376, June 2011.

[14] C. Freitag, M. Berbers-Lee, K. Widdicks, B. Knowles, G. S. Blair, and A. Friday. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns*, 2(9), 2021.

[15] C. Gao, A. Gutierrez, R. G. Dreslinski, T. N. Mudge, K. Flautner, and G. Blake. A study of thread level parallelism on mobile devices. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 126–127, Mar. 2014.

[16] M. Garcia Bardon, P. Wuytens, L.-A. Ragnarsson, G. Mirabelli, D. Jang, G. Willems, A. Mallik, A. Spessot, J. Ryckaert, and B. Parvais. DTCO including sustainability: Power-performance-area-cost-environmental score (PPACE) analysis for logic technologies. In *IEEE International Electron Devices Meeting (IEDM)*, 2020.

[17] B. Grayson, J. Rupley, G. D. Zuraski, E. Quinnell, D. A. Jiménez, T. Nakra, P. Kitchin, R. Hensley, E. Brekelbaum, V. Sinha, and A. Ghiya. Evolution of the Samsung Exynos CPU microarchitecture. In *Proceedings of the IEEE/ACM International Symposium on Computer Architecture (ISCA)*, pages 40–51, June 2020.

[18] P. Greenhalgh. Big.LITTLE processing with ARM Cortex-A15 & Cortex-A7: Improving energy efficiency in high-performance mobile platforms. http://www.arm.com/files/downloads/big_LITTLE_Final_Final.pdf, Sept. 2011.

[19] U. Gupta, M. Elgamal, G.-Y. W. G. Hills, H.-H. S. Lee, D. Brooks, and C.-J. Wu. ACT: Designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the ACM/IEEE Inernational Symposium on Computer Architecture (ISCA)*, pages 784–799, 2022.

[20] 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 *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 854–867, 2021.

[21] 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 IEEE/ACM Symposium on Computer Architecture (ISCA)*, pages 37–47, 2010.

[22] A. Hartstein, V. Srinivasan, T. R. Puzak, and P. G. Emma. On the nature of cache miss behavior: Is it $\sqrt{2}$? *Journal of Instruction-Level Parallelism (JILP)*, 10, Jan/Feb 2008.

[23] M. D. Hill and M. R. Marty. Amdahl's law in the multicore era. *IEEE Computer*, 41(7):33–38, July 2008.

[24] M. D. Hill and V. J. Reddi. Accelerator-level parallelism. *Communications of the ACM*, 64(12), 2021.

[25] S. Kaxiras and M. Martonosi. *Computer Architecture Techniques for Power-Efficiency*. Morgan & Claypool Publishers, 2008.

[26] 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 *Proceedings of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 123–134, Feb. 2008.

[27] D. Kline, N. Parshook, X. Ge, E. Brunvand, R. G. Melhem, P. K. Chrysanthis, and A. K. Jones. GreenChip: A tool for evaluating holistic sustainability of modern computing systems. *Sustainable Computing: Informatics and Systems*, 22:322–332, June 2019.

[28] R. Kumar, K. I. Farkas, N. P. Jouppi, P. Ranganathan, and D. M. Tullsen. Single-ISA heterogeneous multi-core architectures: The potential for processor power reduction. In *Proceedings of the ACM/IEEE Annual International Symposium on Microarchitecture (MICRO)*, pages 81–92, Dec. 2003.

[29] K. Lakshminarasimhan, A. Naithani, J. Feliu, and L. Eeckhout. The forward slice core microarchitecture. In *Proceedings of the IEEE International Conference on Parallel Architectures and Compilation Techniques (PACT)*, pages 361–372, 2020.

[30] R. C. Leachman. Yield modeling and analysis, 2014.

[31] C. Lefurgy, K. Rajamani, F. L. R. III, W. M. Felter, M. Kistler, and T. W. Keller. Energy management for commercial servers. *IEEE Computer*, 36(12):39–48, Dec. 2003.

[32] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi. McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures. In *Proceedings of the IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 469–480, Dec. 2009.

[33] S. Manne, A. Klauser, and D. Grunwald. Pipeline gating: Speculation control for energy reduction. In *Proceedings of the IEEE/ACM International Symposium on Computer Architecture (ISCA)*, pages 132–141, June 1998.

[34] D. Meisner, B. T. Gold, and T. Wenisch. PowerNap: Eliminating server idle power. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 205–216, Mar. 2009.

[35] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi. CACTI 6.0: A tool to model large caches. Technical report, Hewlett-Packard Laboratories, Apr. 2009.

[36] O. Mutlu, J. Stark, C. Wilkerson, and Y. N. Patt. Runahead execution: An alternative to very large instruction windows for out-of-order processors. In *Proceedings of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 129–140, 2003.

[37] A. Naithani, J. Feliu, A. Adileh, and L. Eeckhout. Precise runahead execution. In *Proceedings of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 397–410, 2020.

[38] S. Ollivier, S. Li, Y. Tang, S. Cahoon, R. Caginalp, C. Chaudhuri, P. Zhou, X. Tang, J. Hu, and A. K. Jones. Sustainable AI processing at the edge. *IEEE Micro*, 43:19–28, Jan/Feb 2023.

[39] D. Parikh, K. Skadron, Y. Zhang, M. Barcella, and M. R. Stan. Power issues related to branch prediction. In *Proceedings of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 233–244, Feb. 2002.

[40] S. Pruett and Y. Patt. Branch runahead: An alternative to branch prediction for impossible to predict branches. In *Proceedings of the IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 804–815, Oct. 2021.

[41] J. Ranganathan and et al. The greenhouse gas protocol: A corporate accounting and reporting standard – revised edition, 2004.

[42] E. Rotem, A. Naveh, D. Rajwan, A. Ananthakrishnan, and E. Weissmann. Power-management architecture of the intel microarchitecture code-named sandy bridge. *IEEE Micro*, 32:20–27, March/April 2012.

[43] E. Rotem, A. Yoaz, L. Rappoport, S. J. Robinson, J. Y. Mandelblat, A. Gihon, E. Weissmann, R. Chabukswar, V. Basin, R. Fenger, M. Gupta, and A. Yasin. Intel Alder Lake CPU architectures. *IEEE Micro*, 42:13–19, May 2022.

[44] D. Schor. TSMC 5-nanometer update, 2019.

[45] A. Seznec, S. Felix, V. Krishnan, and Y. Sazeides. Design tradeoffs for the Alpha EV8 conditional branch predictor. In *Proceedings of the IEEE/ACM International Symposium on Computer Architecture (ISCA)*, pages 295–306, June 2002.

[46] J. Switzer, G. Marcano, R. Kastner, and P. Pannuto. Junkyard computing: Repurposing discarded smartphones to minimize carbon. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, volume 2, pages 400–412, Mar. 2023.

[47] P. Teehan and M. Kandlikar. Comparing embodied greenhouse gas emissions of modern computing and electronics products. *Environmental Science and Technology*, 43(9):3997–4003, May 2013.

[48] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi. CACTI 5.1. Technical report, Hewlett-Packard Laboratories, Apr. 2008.

[49] G. Venkatesh, J. Sampson, N. Goulding, S. Garcia, V. Bryksin, J. Lugo-Martinez, S. Swanson, and M. B. Taylor. Conservation cores: Reducing the energy of mature computations. In *Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 205–218, 2010.

[50] D. H. Woo and H.-H. S. Lee. Extending Amdahl's law for energy-efficient computing in the many-core era. *IEEE Computer*, 41(12):24–31, Dec. 2008.

[51] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. A. Behram, J. Huang, C. Bai, M. Gschwind, A. Gupta, M. Ott, A. Melnikov, S. Candido, D. Brooks, G. Chauhan, B. Lee, H.-H. S. Lee, B. Akyildiz, M. Balandat, J. Spisak, R. Jain, M. Rabbat, and K. M. Hazelwood. Sustainable AI: Environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems (MLSys)*, Aug. 2022.

[52] S. Zhang, M. Naderan-Tahan, M. Jahre, and L. Eeckhout. Balancing performance against cost and sustainability in multi-chip-module GPUs. *IEEE Computer Architecture Letters (CAL)*, 22(2):145–148, July–Dec 2023.