This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3218034

Kaya for Computer Architects: Towards Sustainable Computer Systems

Lieven Eeckhout (Ghent University)

***Abstract*—This paper reformulates the well-known Kaya identity to understand computer systems' impact on sustainability and its total carbon footprint. By making a distinction between embodied and operational carbon emissions, we are able to understand (1) how the global carbon footprint of computing is likely to scale in the future, and (2) what we, as computer architects, can do to reduce the environmental impact of computing. We conclude that computer architects should first and foremost design smaller chips; reducing lifetime energy consumption is of secondary importance, yet still significant.**

***Index Terms*—sustainability, computer systems, Kaya identity, Jevons' paradox**

I. Introduction

Sustainability is undeniably a grand challenge. As the world population and the average affluence per person continue to grow, we are eagerly consuming the earth's natural resources while at the same time inducing a climate change. Greenhouse gas (GHG) emissions are detrimental to global warming, and a recent study reports that the contribution of information and communication technology (ICT) to the world's global GHG emissions, currently between 2.1 and 3.9% [6], is growing at rapid pace.

To combat global warming, the Paris agreement under the United Nations (UN) auspices aims at limiting global warming to well below 2, preferably to 1.5 degrees Celsius, compared to pre-industrial levels. The UN recently stated that we need to cut global emissions by 7.6% each year over the next decade to meet the Paris agreement.[1]

Given the pressing need to act along with the significant and growing contribution of computer systems to global warming, it is imperative that we, as computer architects, ask ourselves the question what we can do to design sustainable computer systems. To do so, we first need to understand how the global carbon footprint of computing scales and what its the contributing factors are, and we then need to analyze what we, as computer architects, can do to tame or, even better, reduce the global carbon footprint of computing.

We do so in this paper by reformulating the well-known Kaya identity to understand how computer systems impact sustainability in general, and carbon emissions in particular. We make a distinction between embodied emissions (due to chip manufacturing) and operational emissions (due to computer system use during its lifetime), and we use recently published scaling numbers for each of the contributing factors to understand how the global carbon footprint of computing is likely to scale in the future.

Obtaining decisive conclusions is inherently difficult because how computer systems are manufactured, deployed and used affects sustainability. Nevertheless, our analysis suggests that the total carbon footprint of computing continues to grow under current scaling trends. Moreover, we find that embodied emissions are, or will soon become, the biggest contributor. The fundamental reason is that embodied emissions grow at fast pace because of increasing demands for chips, and increasing energy intensity of semiconductor manufacturing. Contradictory perhaps to common belief, improving sustainability does not equate improving computer system energy efficiency. Instead, to keep the overall carbon footprint of computing under control, computer architects should first and foremost design smaller chips (i.e., reduce die size). Reducing lifetime energy consumption is of secondary importance, yet still significant.

II. IPAT, Kaya and Jevons' Paradox

IPAT is the acronym of a well-known and widely used equation which quantifies the *impact* I of human activity on the environment as follows:

$$I = P \times A \times T. \tag{1}$$

P stands for *population* (i.e., the number of people on earth); A accounts for the *affluence* per person or the average consumption per person; and T quantifies the impact of the *technology* on the environment per product consumed. The impact on the environment due to human activity can be measured along a number of dimensions including the natural resources and materials (some of which may be critical and scarce) that are needed to produce affluence; greenhouse gases (GHG) emissions during the production, use and transportation of products; pollution of ecosystems and its impact on biodiversity; etc.

The Kaya identity, by the Japanese energy economist Yoichi Kaya, reformulates the IPAT equation by specifically focusing on carbon dioxide (CO2) emissions:

$$F = P \times \frac{G}{P} \times \frac{E}{G} \times \frac{F}{E}, \tag{2}$$

with F the global CO2 emissions from human sources, P the global population, G the world Gross Domestic Product (GDP), and E the global energy consumption. In other words, G/P is the GDP per capita, E/G is the energy intensity per unit of GDP, and F/E quantifies the CO2 emissions per unit

[1] https://unfccc.int/news/cut-global-emissions-by-76-percent-every-year-for-next-decade-to-meet-15degc-paris-target-un-report

This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3218034

of energy consumed. The Kaya identity is widely used for example by the Intergovernmental Panel for Climate Change (IPCC) in their annual reports.

The IPAT equation (as well as the Kaya identity) has been criticized for being too simplistic by assuming that the different variables in the equation are independent of each other. Indeed, in contrast to what the above formula may suggest, improving one of the variables does not necessarily lead to a corresponding reduction in overall impact. For example, reducing T in the IPAT model by 50% through innovations that reduce the environmental impact per product, does not necessarily reduce the overall environmental impact I by 50%. The fundamental reason is that a technological efficiency improvement typically leads to a price reduction, which in turn stimulates additional consumption of the resource that was supposed to be conserved. The end result may be an overall increase in impact rather than a reduction. This is the well-known *rebound effect* or *Jevons' paradox*, named after the English economist Williams Stanley Jevons who was the first to report the rebound effect as a result of improving the coal efficiency of the steam engine, which led to an overall increase in coal consumption.

The rebound effect can be (partly) accounted for in the IPAT/Kaya models by expressing each of the variables as a *Compound Annual Growth Rate (CAGR)*, defined as follows:

$$CAGR = \left(\frac{V_t}{V_0}\right)^{1/t} - 1, \tag{3}$$

with V_0 the variable's value at year 0 and V_t its value at year t. The IPAT/Kaya models can be expressed using CAGRs for the respective variables:

$$CAGR_{overall} = \prod_{i=1}^{N}(CAGR_i + 1) - 1. \tag{4}$$

This reformulation allows for computing the annual growth rate in overall environmental impact or CO2 emissions as a function of the growth rates of the individual contributing factors. If the growth rates incorporate the rebound effect, i.e., higher consumption rate as a result of higher technological efficiency, the model is able to make an educated guess about the expected growth rate in environmental impact.

Figure 1 visualizes how the different factors in the Kaya identity affect the overall carbon footprint (note the logarithmic scale along the vertical axis). World population continues to grow, and so does affluence (GDP) per capita. At the same time, the energy intensity per unit of GDP tends to decrease as we improve the energy efficiency of the technologies and goods that we use and produce. Similarly, the carbon intensity per unit of energy tends to decrease as we trade brown for green energy sources. The total carbon footprint decreases only if the positive growth rate in population and GDP per capita is overweighed by the negative growth rate in energy intensity per GDP and carbon intensity per unit of energy.

Bol et al. [3] decomposed the Kaya identity using CAGR factors for different ICT-subsectors, however, they did not make a distinction between embodied and operational emissions, and lacked a comprehensive conclusion for computer architects.

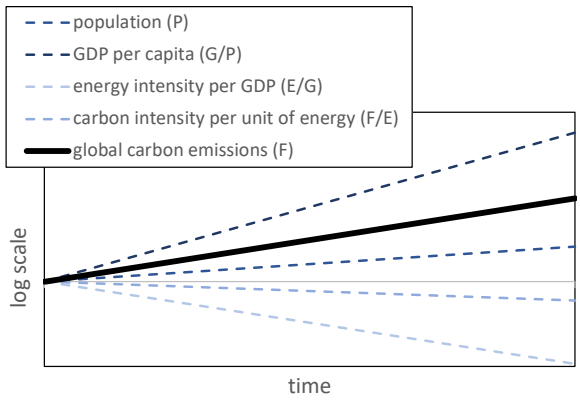


Fig. 1: Visualizing the Kaya model in terms of its contributing factors. *The increase in population and GDP per capita outweighs the decrease in energy and carbon intensity, leading to an overall increase in carbon emissions.*

III. Kaya for Computer Architects

We now reformulate the Kaya identity so it becomes useful and insightful for computer architects to reason about sustainability and the environmental impact of the computer systems they design. We focus on GHG emissions (this includes fluorinated greenhouse gases in addition to carbon dioxide), and make a distinction between operational and embodied emissions. Operational emissions are a result of product use during its lifetime, i.e., carbon emissions due to empowering electronic devices. Embodied emissions, in general, include the emissions due to raw material extraction, component manufacturing, product assembly, transportation, repair/maintenance during the product lifetime, and eventually end-of-life dismantling and disassembly. In this work, we specifically focus on GHG emissions during the manufacturing process, and following the GHG Protocol, we make a distinction between scope-1 and scope-2 embodied emissions. Scope-1 relates to chemicals and gases emitted, while scope-2 relates to the energy consumed during semiconductor manufacturing.

A. Embodied Scope-2 Emissions

We formulate the embodied scope-2 emissions as follows:

$$F_{scope\text{-}2} = C \times \frac{W}{C} \times \frac{E}{W} \times \frac{F}{E}, \tag{5}$$

with C the number of chips or dies that are produced, W/C the number of wafers produced per chip, E/W the energy needed to produce a wafer, and F/E the carbon intensity of the energy source during manufacturing. We now estimate the CAGRs for each of the factors in the above formula to understand how embodied scope-2 emissions are trending and what computer architects can do about them. Estimating CAGRs is challenging and different sources (may) report different numbers. The actual numbers are hence to be taken with a grain of salt, but at least they indicate current trends.

Source	Carbon intensity (g CO2e/kWh)
coal	820
gas	490
biomass	230
solar	41
geothermal	38
hydropower	24
nuclear	12
wind	11

TABLE I: Equivalent CO2 emission for different energy sources, taken from [8].

The number of chips C produced on an annual basis continues to grow. IC Insights in its 2021 McClean report [13] mentions a 30-year historical CAGR of $+9\%$, while forecasting a CAGR of $+11\%$ for the 2020–2025 timeframe (which is 5 percent points higher than the CAGR of $+6\%$ for the 2015–2020 timeframe).

The number of wafers needed per chip W/C depends on die size. de Vries [4] provides a formula that empirically derives the number of chips per wafer as a function of die size S:

$$\frac{C}{W} = \frac{\pi d^2}{4S} - 0.58\frac{\pi d}{\sqrt{S}}, \tag{6}$$

with d the wafer's diameter, which we assume to be 300 mm in this work. Kogge et al. [11] report a historical CAGR of $+16\%$ for die size until 1995, and a stagnation since then. Recently, we have been witnessing increasing die sizes for server CPUs and GPUs [14], with recent high-end CPU and GPU die sizes in the 700–800 mm^2 range. The above C/W formula does not account for yield, but the effective number of chips per wafer could easily be derated to incorporate lower yield for bigger die sizes.

The amount of energy needed to produce a wafer E/W increases with new chip technology nodes. The steady increase is a result of increased complexity: increasing number of process steps, increasing number of metal layers, new equipment such as extreme ultraviolet lithography (EUV), among others. Garcia Bardon et al. [7] recently published energy consumption numbers per wafer produced for a range of CMOS technology nodes from 28 nm (around year 2011) to 3 nm (year 2022). The CAGR for the amount of energy per wafer amounts to +11.9%.

The carbon intensity per unit of energy F/E depends on the brown versus green energy mix used during wafer production. As illustrated in Table I, the energy source determines the equivalent CO_2 emissions per kWh of electricity. Carbon intensity is trending down in Europe with a 30-year CAGR of around -2.5% according to the European Environment Agency [5]. Carbon intensity is decreasing at a (much) slower pace (if at all) in other parts of the world where semiconductor manufacturing takes place (e.g., US, Taiwan, China) [9]. If semiconductor manufacturers were to use green energy sources at a faster pace, carbon intensity for producing electronics could dramatically reduce.

The overall conclusion is that the current relatively slow transition towards green energy (CAGR of-2.5% for carbon intensity) does not compensate for the growing demand for chips (CAGR of $+9\%$) and the growing energy needed per wafer (CAGR of $+11.9\%$). Assuming that die size remains constant, the CAGR for overall carbon emissions amounts to $+18.9\%$, or a $5.7\times$ increase over the next decade. There appears to be no easy solution to reduce the embodied scope-2 emissions apart from quickly transitioning to renewable energy sources for chip manufacturing or drastically reducing die size. However, this is non-trivial. Even if semiconductor manufacturers were to transition to green energy sources at a (much) faster pace (e.g., CAGR of -10%), the total carbon emissions would still increase with a CAGR of $+9.8\%$. Even if we were to reduce die size at a fast and steady pace (e.g., CAGR of -10%), and assuming a CAGR of -2.5% for carbon intensity, the total embodied scope emissions would still increase with a CAGR of $+6.3\%$.

This brings us to the insight that the only way to reduce embodied scope-2 emissions is to drastically design smaller chips, and/or drastically shift towards green energy sources for chip manufacturing, and/or drastically sell/produce less chips. Since we, as computer architects, do not have direct control over the latter two options, we will focus on the former, namely to drastically design smaller chips. The good news is that this is possible and will increase yield. Recall that Moore's Law states that the number of transistors doubles with every new technology node, roughly every two years. Despite this becoming more and more challenging, industry has kept up with this empirical law [2], and predictions suggest that this trend will continue in the near future [10]. Moore's Law means that transistor density increases at a CAGR of $+41\%$. This implies that we would need approximately 30% fewer wafers each year to produce the same number of chips (i.e., CAGR of -30% for W/C). This in turn implies a reduction in embodied scope-2 emissions at a CAGR of -16.8% (assuming a CAGR of $+9\%$ for C, a CAGR of $+11.9\%$ for E/W, and a CAGR of -2.5% for F/E). This is not what is happening today and is in sharp contrast to the CAGR of $+18.9\%$ assuming a constant die size as discussed above. The reason is of course that we have been using the 41% additional transistors each year to add new functionality (i.e., more cores, larger caches, more complex microarchitectures, more accelerators, etc.) — this is a clear example of a rebound effect with an overall negative impact on embodied scope-2 emissions.

The large gap between leveraging Moore's Law to design chips that are 30% smaller each year with the same transistor count (total embodied scope-2 CAGR of -16.8%) versus continuing to design chips with the same constant die size and thus 41% more transistors each year (total embodied scope-2 CAGR of $+18.9\%$) provides an opportunity for computer architects to design chips in a more sustainable way. There is a middle ground that computer architects can explore to use some fraction of the additional transistors provided with each technology node for adding new functionality, but do so in a way that the total embodied scope-2 emissions stagnate, or

even better, decline. For example, under the above assumptions, if we were to use only half the additional transistors added each year, we would be able to reduce die size by 15% each year, and be (almost) neutral in terms of embodied scope-2 emissions compared to the current state. Reducing die size by 25% each year (still providing 6% more transistors each year for new functionality), embodied scope-2 emissions would decrease with a CAGR of -12%.

B. Embodied Scope-1 Emissions

Scope-1 emissions encompass chemicals and gases including fluorinated compounds used during manufacturing (e.g., SF_6, NF_3 and CF_4, among others). SF_6 and NF_3 are the two major contributors due to their high global warming potential (23,500$\times$ and 16,100$\times$, respectively, compared to CO_2 [7]). We formulate the embodied scope-1 emissions as follows:

$$F_{\textit{scope-1}} = C \times \frac{W}{C} \times \frac{F}{W}, \tag{7}$$

with F/W the carbon dioxide equivalents of fluorinated compounds per wafer. Garcia Bardon et al. [7] report equivalent CO_2 emissions for the various tech nodes. The CAGR for F/W amounts to +9.3%.

The conclusion for the embodied scope-1 emissions is somewhat similar to the scope-2 emissions: assuming a constant die size, scope-1 emissions increase with a CAGR of +19.2% as a result of the growing demand for chips and the increasing use of chemicals and gases during production. Hence, structurally reducing embodied scope-1 emissions can only be achieved by designing smaller chips. Note that a transition to green energy sources does not impact scope-1 emissions, so even if semiconductor manufacturers were to transition to renewables, scope-1 emissions would remain on a rising curve.

The relative importance of embodied scope-1 versus scope-2 emissions varies depending on the degree of abatement. Assuming an aggressive 99% abatement for the NF_3 compound, scope-2 is 12.1$\times$ more important than scope-1 [7], which is what we assume in the remainder of this work. Gupta et al. [8] report that for the world's leading chip manufacturer TSMC 63% of carbon emissions are due to scope-2 versus around 30% for scope-1 — a 2.1$\times$ ratio. Again, these numbers need to be taken with a grain of salt, nevertheless, the trends are clear, namely scope-2 emissions are more important than scope-1 emissions. The specific ratio of scope-2 versus scope-1 emissions does not affect the overall conclusions in this paper.

C. Operational Emissions

Operational emissions refer to carbon dioxide emissions during the lifetime of semiconductor chips. We formulate the operational emissions as follows:

$$F_{\textit{operational}} = C \times \frac{E}{C} \times \frac{F}{E}, \tag{8}$$

with E/C the total electricity usage (in kWh) over the entire lifetime of a chip. To understand how computer architects can affect operational carbon emissions, we make a distinction between two scenarios: (1) there is work to be done, and (2) the chip is idle.

Assuming that the chip is powered on when there is work to be done, and is turned off once the work is completed, the operational emissions are proportional to the amount of energy that it takes to get the work done. The less energy is consumed to get the work done, the lower the operational emissions. Note that a higher performance system could be more sustainable, even if it consumes more power: if the performance improvement outweighs the increase in power consumption, this leads to a net reduction in energy consumption, and thus a reduction in operational emissions. Likewise, a lower power system could be less sustainable: if the performance degradation that comes with the power saving outweighs the reduction in power consumption, this leads to an increase in energy consumption and thus an increase in operational emissions.

When there is no work to be done, and the chip is idle, the operational carbon emissions are proportional to the chip's idle power. It is hence critical to lower standby power to reduce operational emissions of idle chips.

Note that Jevons' paradox is always looming behind the corner as energy and power optimizations may lead a rebound effect in terms of total energy usage and thus operational emissions. For example, an energy saving optimization that decreases execution time while increasing power consumption may enable more jobs to be done per unit of work, which may lead to an overall increase in energy consumption. Likewise, a power saving optimization may enable more concurrent jobs to be completed within the available power envelope, which may lead to an overall increase in energy consumption. The fact that a computer system still consumes power when doing virtually no work (i.e., lack of energy-proportionality [1]) incentivizes high utilization degrees to maximize the achieved performance per Watt. Workload consolidation, i.e., running more concurrent jobs, achieves exactly this, thereby increasing the total energy usage.

It is enlightening to analyze current trends in operational carbon emissions for both consumer devices and datacenter infrastructures. Gupta et al. [8] performed a survey across a couple dozen mobile devices from vendors such as Apple, Google and Huawei, and they conclude that operational emissions per device tend to decrease over time (in contrast to embodied emissions) as a result of a variety of energy and power efficiency optimizations. Indeed, individual transistors become more energy-efficient across chip technology nodes [2], and various architecture-level optimizations such as Dynamic Voltage and Frequency Scaling (DVFS), clock gating, power gating, P-states, etc. have further reduced energy and power consumption. This suggests that the E/C factor from the above formula is trending down. If the increase in the number of chips C is outweighed by the decrease in energy intensity (E/C) and carbon intensity (F/E), the overall operational emissions for consumer devices may be trending down.

Gupta et al. [8] also analyzed the operational emissions in Facebook's and Google's datacenters. In spite of the fact that

total datacenter energy usage is trending up — presumably because of adding more servers and/or more jobs per server (Jevons' paradox) — the total operational carbon emissions are decreasing as a result of purposefully contracting and securing green energy sources. In terms of the above formula, $C \times E/C$ is trending up while F/E is trending down, which leads to an overall decrease in operational emissions in modern-day hyperscale datacenters.

IV. What Can We Do as Computer Architects?

Having described how embodied scope-1 and scope-2 emissions as well as operational emissions scale using the Kaya-inspired formulas, there are a number of important observations to be made. We are witnessing (i) a steady growth rate in the number of chips being produced (CAGR of $+9\%$), (ii) an increasing energy demand and chemical/gas emissions for new technology nodes (CAGR of $+11.9\%$ and $+9.4\%$, respectively), and (iii) a transition towards green energy sources at a rate that varies geographically (CAGR for carbon intensity of -2.5% in Europe and even less in other parts of the world) and across businesses (major shift towards green energy sources for hyperscale datacenters). Given the context, the key question is what we, as computer architects, can do to design chips in a more sustainable way taking into account both embodied as well as operational carbon emissions?

There are a number of aspects we have control over as computer architects. First, we can aim at tempering the growth rate of the number of chips that need to be produced. Reducing the need for ever more chips could possibly be achieved by integrating more functionality per chip, so we need fewer chips for providing the same overall functionality. This is already happening today as we are moving towards heterogeneous system-on-chip designs that integrate a variety of processing units, including CPUs, GPUs and accelerators, on a single chip. Unfortunately, so far, this trend has not led to an overall reduction in the number of chips. On the contrary, Jevons' paradox has led to an overall increase in the demand for chips. Another option to temper the demand for more chips is to deploy fault-tolerance techniques to fix hard errors or graceful degradation (e.g., disable a faulty core in a multi-core system) to extend the lifetime of existing chips so that customers are refrained from buying new devices. Whether the higher upfront embodied footprint to support fault tolerance leads to an overall reduction in environmental footprint by extending a device's lifetime is an interesting trade-off to be investigated. Moreover, drastically reducing the demand for new chips might need companies to move away from a business model that is based on selling new devices.

Second, we can design smaller chips to reduce embodied emissions. Designing smaller chips does not necessarily mean that we need to stop adding new functionality to the chips that we design. On the contrary, there is some leeway, as argued before. We can leverage Moore's Law (as long as it continues to last) to integrate new functionality using the newly available transistors, but we should do so in a sober way so that the embodied emissions stagnate or, even better, decrease. Given the relatively slow transition towards green energy sources, and the growing energy demands and chemical/gas emissions per technology generation, this implies that in practice die size should decrease. This seems to contradict the current trend of maintaining constant die size or, even worse, increasing die sizes with more processing units and accelerators with each chip generation.

Third, we need to continue to improve the energy and power efficiency of the chips that we design to reduce operational emissions. In particular, it is key to (1) reduce energy consumption when there is work to be done, and (2) reduce power consumption when idle. This seems to be happening today. Computer systems have become more energy and power-efficient, energy proportionality has improved, and idle power P-states have been added and optimized. Unfortunately, Jevons' paradox may counteract these energy and power efficiency improvements, unless operational energy is provided by increasingly green energy sources.

V. Putting It All Together

Now that we understand how a computer architect can impact embodied and operational emissions separately, the question is how to optimize total emissions including embodied and operational emissions. This obviously hinges on the relative importance of embodied versus operational emissions, and how the embodied and operational emissions scale over time. The relative importance between embodied versus operational emissions depends on a variety of factors, including manufacturing (fab location, technology node, die size, etc.) as well as operational usage (market segment, i.e., consumer versus server, as well as device usage and lifetime, i.e., a longer lifetime amortizes the embodied footprint over a longer period of time). Gupta et al. [8] analyzed the embodied versus operational footprint for a range of devices. Embodied emissions dominate operational emissions for battery-powered devices (e.g., smartphone, smart watches, tablets). On the contrary, for always-connected personal devices (e.g., desktop computers, game consoles, speakers), operational emissions dominate embodied emissions. In the datacenter, in part due to the transition towards green energy sources for empowering the IT and cooling equipment, embodied emissions dominate operational emissions.[2]

As a result, making decisive conclusions regarding sustainability is inherently difficult given the range of scenarios in which computer systems are manufactured, deployed and used. In particular, the degree to which renewable energy sources contribute to fabrication and use may shift the relative contribution of embodied emissions versus operational emissions. We hence consider a range of scenarios in what follows. It is worth keeping in mind that other sustainability concerns such as scope-1 emissions, raw materials needed, and ultra-pure water supply all relate to manufacturing, and are hence proportional to die size. Hence, even if computer

[2]If obtained through green energy contracts on the energy market, this effectively implies a preemption of green energy, not fundamentally reducing the carbon footprint at the global societal scale.

Factor	Unit	Symbol	CAGR
Number of chips	Number	C	$+9\%$
Wafer energy intensity	kWh/wafer	E/W	$+11.9\%$
Wafer chemical/gas intensity	CO2e/wafer	F/W	$+9.3\%$
Carbon intensity	CO2/kWh	F/E	-2.5%

TABLE II: CAGRs assumed throughout the paper, unless mentioned otherwise.

system manufacturing and use would be empowered by green energy sources only, the environmental impact of computer systems could still be significant; moreover, as small as their environmental impact may be, green energy sources are not environmentally neutral, see also Table I.

A. Current Scaling Trends

We now investigate how the embodied and operational emissions are likely to scale in the near future given current scaling trends. We consider two scenarios in which we change the ratio of embodied versus operational emissions in year zero (now): (1) *initially dominating embodied emissions*: we assume that embodied emissions are responsible for 80% of the initial total carbon emissions (and thus operational emissions account for 20%); and (2) *initially dominating operational emissions*: we assume that operational emissions account for 80% of the initial total emissions (and thus embodied emissions account for the remaining 80%). We assume that die size remains constant (*CAGR* $= 0\%$) and that operational energy intensity E/C improves with a CAGR of -10%. We further assume the default scaling trends for the other factors as mentioned in Table II.

Figure 2 reports the total emissions over the next decade relative to the present day, for the two scenarios. There are at least two important observations to be made. First, total emissions are increasing dramatically given current scaling trends, reaching a 4.7× and 1.65× increase over the next decade under the two scenarios, respectively. Second, the major culprit for the dramatic increase in total emissions is the rapid increase in embodied emissions. In fact, the embodied emissions grow in importance in both scenarios, even under the second scenario where operational emissions dominate initially: while embodied emissions were assumed to contribute to only 20% of the total emissions initially, they will contribute almost 69% of total emissions by the end of the next decade. The reason is the rapid annual growth rate for the underlying factors, namely the increasing demand for chips and the growing energy intensity of semiconductor manufacturing.

B. Reducing Die Size or Operational Energy

To understand how total emissions are likely to evolve over the next decade, we now analyze what the impact would be if we were to courageously optimize the most dominant initial emission contributor. We explore how total emissions scale as we leverage the two factors that computer architects have most control over, namely chip die size and lifetime operational energy consumption. In particular, for the scenario where the embodied emissions dominate, we assume that chip designers decrease chip area by 10% each year (CAGR of -10%) instead of keeping die size constant, see Figure 3(a). In the scenario where the operational emissions dominate, we assume that computer architects decrease lifetime operational energy consumption by 20% each year (CAGR of -20%) instead of decreasing by 10% per year, see Figure 3(b). Arguably, both efforts would require substantial and continued effort.

Total emissions would continue to increase over the next decade under both scenarios, in spite of the temporary decrease for the scenario where operational emissions dominate initially. The reason is the ever-increasing demand for chips and energy needed for manufacturing, which leads the embodied emissions to dominate. In other words, the bold efforts to reduce embodied and operational emissions are insufficient to keep total emissions in check. The underlying reason is the growing embodied emissions.

C. Reducing Both Die Size and Operational Energy

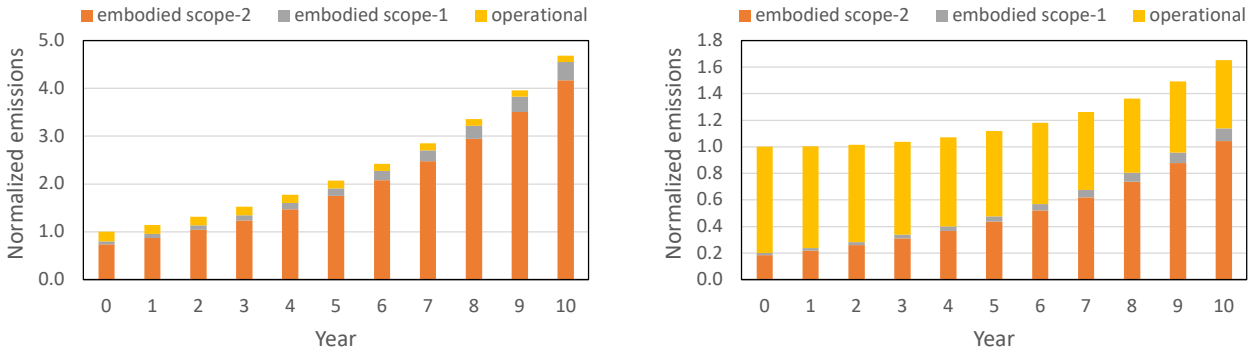
The analysis in the previous section suggests that only reducing embodied or operational emissions, whichever contributes most to the total emissions, is likely to be insufficient in the long term. We now explore how total emissions scale as we reduce both die size and lifetime operational energy consumption.

Figure 4 reports the overall 10-year annual growth rate as a function of how die size scales (CAGR along horizontal axis) and how operational energy intensity scales (CAGR in legend); again, we consider the two scenarios where embodied and operational emissions dominate. Several interesting observations can be made. First and foremost, to reduce overall carbon emissions, it is imperative to decrease die size. Indeed, an increasing die size does not lead to a reduction in overall emissions, even if we were to drastically reduce operational energy intensity. Decreasing die size at a fast enough pace is even more critical in case embodied emissions are high.

Second, in case the embodied emissions are more significant than the operational emissions, see Figure 4(a), reducing operational energy intensity is less critical than reducing embodied emissions, but this does not imply that it is unimportant to reduce embodied emissions. In case die size increases or does not decrease fast enough (CAGR larger than -10%), dramatically reducing operational energy intensity does not lead to a net reduction in global emissions. However, decreasing die size can be outweighed by increasing energy intensity. If die size decreases at a fast pace (e.g., CAGR of -20%), an increase in operational energy intensity (CAGR of $+20\%$) would still lead to an increase in overall emissions (CAGR of $+9.1\%$).

Third, in case the initial operational emissions are more significant than the embodied emissions, see Figure 4(b), it is critical to reduce operational energy intensity. Indeed, even if die size is reduced at fast pace (e.g., CAGR of -20%), operational energy intensity needs to (significantly) reduce to achieve an overall reduction in global emissions. Note though that reducing operational energy intensity by itself is not enough to reduce overall carbon emissions as embodied

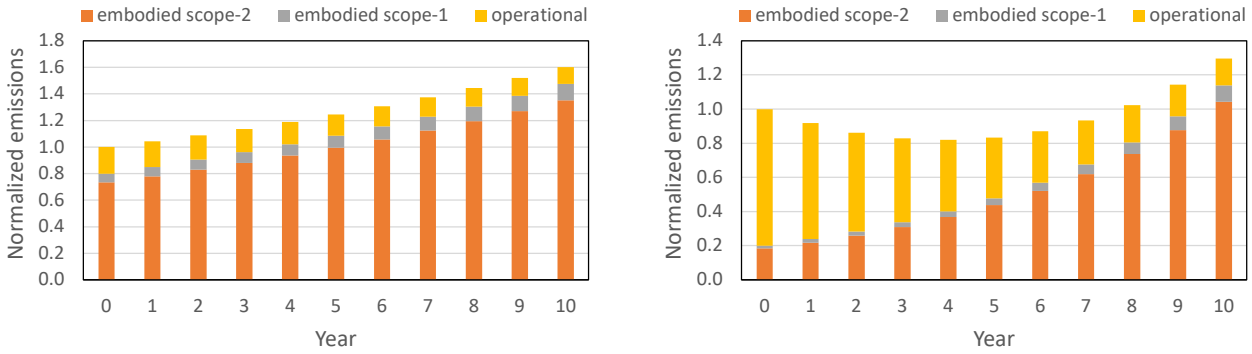
This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3218034



(a) initially dominating embodied emissions

(b) initially dominating operational emissions

Fig. 2: Projection for total emissions over the next decade given current scaling trends for two scenarios: (a) embodied emissions dominate initially (80% of total emissions in year zero), and (b) operational emissions dominate initially (80% of total emissions in year zero). *Total emissions are increasing dramatically, and embodied emissions are, or will start, dominating.*



(a) initially dominating embodied emissions

(b) initially dominating operational emissions

Fig. 3: Projection for total emissions over the next decade given current scaling trends for two bold scenarios: (a) decreasing die size by 10% each year for the scenario where embodied emissions dominate initially, and (b) decreasing operational energy intensity by 20% each year for the scenario where operational emissions dominate initially. *Total emissions continue to increase over the next decade (in spite of a temporary reduction in (b)) as a result for growing embodied emissions.*

emissions also need to be reduced. Assuming that die size remains constant, decreasing energy intensity at steady pace (e.g., CAGR of -10%) still leads to an increase in overall emissions (CAGR of $+5.1\%$).

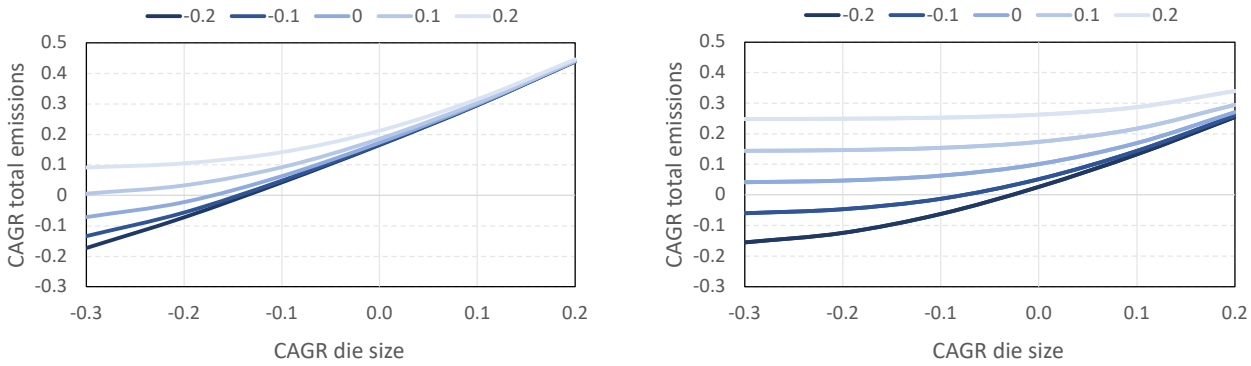
VI. Discussion

The overall conclusions are that (1) total carbon emissions continue to grow under current scaling trends; (2) embodied emissions already are, or will soon become, the biggest contributor; and (3) computer architects have the leverage to tame and even reduce total carbon footprint by reducing die size and reducing lifetime operational energy consumption, with reducing die size taking higher priority. It is worth noting that these conclusions hold true irrespective of the ratio of embodied versus operational emissions. They thus apply across devices (from mobile to server) and lifetimes (extending the lifetime of a device, and thus changing the relative contributions of the embodied versus operational emissions, does not fundamentally change the overall conclusions).

So far, we assumed that energy sources are transitioning towards renewables at a relatively slow pace (CAGR of -2.5%). However, a fast transition towards green energy sources could quickly reduce the total carbon emissions. This is already happening in hyperscale datacenters as they are rapidly adopting renewable energy sources [8], thereby dramatically reducing operational carbon emissions. Similarly, consumer electronics charged using a personal green energy source (e.g., solar panels at home) dramatically reduce operational carbon footprint. This does not imply that we should not reduce operational energy consumption: we should continue to reduce energy consumption because the amount of green energy available is bounded and the fraction of green energy left unused for ICT can be used elsewhere.

Nevertheless, while reducing operational carbon footprint is important, embodied emissions are (or will become) dominant. Hence, it is critical to dramatically reduce embodied emissions. This could be achieved if semiconductor manufacturers were to rapidly transition to renewable energy sources for chip production. TSMC's sustainable goal for 2030 is to supply 25% of power consumed by its fabrication plants from

This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2022.3218034



(a) initially dominating embodied emissions

(b) initially dominating operational emissions

Fig. 4: 10-year CAGR for total emissions as a function of CAGR for die size (horzontal axis) and CAGR for lifetime operational energy consumption (legend) for the two scenarios. *Reducing total emissions requires reducing die size, and, to a lesser albeit not unimportant extent, reducing lifetime operational energy consumption.*

renewable energy, and be carbon-neutral by 2050 [12].[3] Chip manufacturing is only one (yet, a significant) part of the problem: other components such as packages, printed circuit boards, power supplies, batteries, cases, etc., also contribute to the overall emissions during the production process of a device. The production processes for these components also needs to be empowered with green energy. While transitioning to green energy sources during device manufacturing dramatically reduces the embodied scope-2 emissions, it does not affect scope-1 emissions which will continue to increase given current trends, and – more importantly perhaps – it does not address other sustainability issues such as the increased use of raw materials (some of which are rare earth elements and/or require huge amounts of energy to extract) and ultra pure water [7], which are all related to chip production. This reinforces one of the take-aways from this work that computer architects can improve sustainability primarily by designing smaller chips.

VII. Conclusion

This paper reformulated the Kaya identity to understand how global carbon footprint, and more generally the environmental impact, of computer systems evolves over time. Current trends suggest that the carbon footprint is increasing and that embodied emissions are becoming increasingly more important, if not already the case today. We find that, given the rapidly growing energy intensity of semiconductor manufacturing, designing smaller chips is of critical importance to reduce the environmental impact of computing, while reducing lifetime operational energy usage is of secondary importance.

BIO

Lieven Eeckhout is a Senior Full Professor at Ghent University, Belgium. His research interests include computer architecture performance analysis and modeling, CPU/GPU microarchitecture and resource management, and sustainability. He received a Ph.D. degree in computer science engineering from Ghent University in 2002. He is an IEEE and ACM Fellow. Contact him at lieven.eeckhout@ugent.be.

[3]Same side-note applies here: purchasing agreements for green energy preempt green energy supply for others.

References

[1] L. Barroso and U. Hölzle. The case for energy-proportional computing. *IEEE Computer*, 40(12):33–37, 2007.

[2] M. T. Bohr and I. A. Young. CMOS scaling trends and beyond. *IEEE Micro*, 37(6):20–29, 2017.

[3] D. Bol, T. Pirson, and R. Dekimpe. Moore's law and ICT innovation in the anthropocene. In *IEEE Design, Automation and Test in Europe Conference (DATE)*, 2021.

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

[5] E. European Environment Agency. Greenhouse gas emission intensity of electricity generation in Europe, 2021. URL https://www.eea.europa.eu/ims/greenhouse-gas-emission-intensity-of-1.

[6] 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.

[7] 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.

[8] 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.

[9] M. R. Hannah Ritchie and P. Rosado. CO2 and greenhouse gas emissions. *Our World in Data*, 2020. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions.

[10] N. Horiguchi. Entering the nanosheet transistor era, 2021. URL https://www.eetimes.com/entering-the-nanosheet-transistor-era/.

[11] P. Kogge, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Keckler, D. Klein, and R. Lucas. 2008.

[12] P. Lin and L. Sun. TSMC becomes the worlds first semiconductor company to join RE100, committed to 100% renewable energy usage, 2020. URL https://esg.tsmc.com/en/update/greenManufacturing/caseStudy/37/index.html.

[13] B. McClean. The McClean report — a complete analysis and forecast of the integrated circuit industry. *IC Insights*, 2021. https://www.icinsights.com/services/mcclean-report/.

[14] L. Su. Delivering the future of high-performance computing, 2019. Keynote at Hot Chips Conference.