Ensuring that computing systems are sustainable is a highly complex area that requires a holistic approach. Given the importance of sustainability, though, the computing systems community must rise to this challenge.

Towards sustainable computer architecture: A holistic approach

By LIEVEN EECKHOUT

Sustainability and climate change represent a major challenge for our generation. This article argues that sustainable development requires a holistic approach and involves multi-perspective thinking.

Applied to computing, sustainable development means that we need to consider the entire environmental impact of computing, including raw-material extraction, component manufacturing, product assembly, transportation, use, repair/maintenance, and end-of-life processing (disassembly and recycling/reuse). Analysing current trends reveals that the embodied footprint is, or will soon be, more significant compared to the operational footprint.

The article summarizes what computer architects and engineers can and should do to better understand the sustainability impact of computing, and design sustainable computer systems.

Key insights

- Improving computing-system sustainability is **more involved than minimizing carbon emissions during production and usage.** Material use (including rare-earth elements and/or minerals from politically unstable regions in the world) and ultra-pure water consumption are significant sustainability concerns related to chip production. **Even if all the energy consumed during production and use were green, the environmental impact of computing would still be significant, and growing**.
- Sustainable development requires multi-perspective thinking along at least six **dimensions: materials, energy, environment, regulation, society, and economics**.
- The environmental footprint of computing continues to grow under current scaling trends. **When focusing on carbon emissions, embodied emissions are, or will soon become, the biggest contributor compared to operational emissions** across the broad range of computing devices.
- **Embodied emissions are** growing at a fast pace because of increasing demand for chips and **increasing energy intensity of semiconductor manufacturing.** Perhaps contradictory to popular belief, improving the energy efficiency of computing systems does not necessarily make them more sustainable.
- Improving the energy and power efficiency of computing systems may lead to a rebound effect (Jevons' paradox) which may be counterproductive to the environmental impact if the resulting **increase in demand outweighs the efficiency improvement.**
- Improving computing-system sustainability requires a **holistic approach to computer architecture design and development**, requiring **multi-dimensional optimization including chip area, power, energy, performance, reliability, and fault tolerance**.

Key recommendations

- Computer architects should take a **holistic approach when designing sustainable computer systems, and not solely focus on carbon emissions**.
- Computer architects and engineers should primarily focus on **reducing the embodied footprint of computer systems.** Reducing the operational footprint is of secondary importance, although still significant.
- Reducing the **embodied footprint of computing** can be achieved through a variety of options:
  - producing **fewer chips** (e.g. by consolidating functionality)
  - **extending the lifetime of chips** (e.g. by deploying fault tolerance and/or reconfigurability)
  - designing **smaller chips** (i.e. using additional transistors in accordance with Moore's Law, as long as it holds, in a frugal way)
  - manufacturing chips in **older technology nodes**
- **Decarbonizing the manufacturing process is not a panacea** as it does not affect other sustainability concerns related to material use and extraction, chemicals and gases emitted, and ultra-pure water consumed during production.

- Computer scientists and engineers should be wary of **Jevons' paradox. Efficiency improvements most often lead to a significant rebound effect.** Collaborating with entrepreneurs may yield new, **more sustainable business models** for computing.
- Computer architects should collaborate with various partners along the supply chain, user groups, and end-of-life recyclers to obtain **high-quality data to assess the environmental impact of raw material extraction, manufacturing, production, assembly, transportation, product use, maintenance, recycling**, etc.
- **Sustainability modelling tools** (both detailed models and high-abstraction analytical models) need to be developed, finetuned and validated to be able to **holistically balance the embodied and operational footprint** of computing devices.
- **Existing and emerging architecture paradigms** (multicore processing, hardware specialization, chiplet-based integration, etc.) need to be **assessed and re-evaluated from a sustainability perspective**.

Sustainability versus climate change

Sustainability is one of the grand challenges of our generation. Climate change is happening. A recent United Nations Climate Change report [1] in preparation for COP 27, the Sharm el-Sheikh Climate Change Conference, in November 2022 alerts that, while countries are making progress to trend down global greenhouse gas (GHG) emissions, these efforts are insufficient to limit global temperature rise to 1.5 °C by the end of the century. Much more effort is needed to keep this threshold within reach.

Virtually all economic sectors contribute to global emissions. The five economic sectors that contribute most to GHG emissions are industry, electricity, agriculture, transportation and buildings, accounting for nearly 90% of emissions, according to the Organization for Economic Co-operation and Development (OECD) [2]. Freitag et al. [3] recently reported that information and communication technology (ICT) is estimated to contribute 2.1% to 3.9% of worldwide GHG emissions, and this contribution is rising. As computer scientists and engineers, it is our responsibility to limit ICT's contribution to global warming, and, if possible, even decrease it.

While climate change is receiving increasingly wide attention – rightfully so! – it is important that we keep the broader picture in mind when reasoning about potential solutions. The broader picture relates to sustainability. To give just one concrete example: the transition towards green energy sources relies heavily on battery technology, which should be produced with the lowest possible environmental impact, using materials obtained in a safe, responsible, social, and ecological way. Moreover, at the end of their life, batteries should be repurposed, remanufactured, or recycled. In other words, and put more bluntly, solving the climate problem should not create an environmental problem. For similar points, see also Patrick Blouet's article on sustainability in this HiPEAC Vision.

Sustainability is about the efficient use of materials and energy, environmental impact, emissions, (socio-)economics, impact on human health, human rights, politics, policy, legislation, etc. Global warming is an important aspect of sustainable development, but it should not be the only focus. The extraction and use of raw materials and energy sources is another important aspect of sustainability. What to do when devices reach their end of life and how to repurpose e-waste is equally critical. Sustainability also affects and may require new legislation and business models to reduce pressure on raw-material extraction and to incentivize a circular economy with a reduced environmental footprint.

The Brundtland report of the World Council on Economic Development from 1987 provides a broad, yet useful, definition of sustainable development. It states the following: *"Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs."* This definition is an extremely powerful and unambiguous statement. It is a call to action for our generation: when generating economic activity and developing new devices and services, we should be wary of the impact this may have on future generations.

Sustainable development

Sustainable development is often narrowed down to reducing energy consumption and/or transitioning towards green energy sources. However, sustainable development is much more involved than providing and using green energy. If we are solely aiming for carbon-neutral computing, this may not necessarily bring us to a more sustainable future.

The fundamental reason is that sustainable development is extremely complicated and multi-faceted. It requires multi-perspective thinking and reasoning, involving many stakeholders with (often) conflicting interests. Moreover, the problem statement is often poorly defined. Overall, in many cases, there is no "right" answer to questions of sustainable development. The question then is how to assess sustainable development. We hence need a framework for critical thinking that recognizes the complexity and the interdependences between the various goals, interests, and constraints.

There are at least six dimensions to consider when reasoning about and assessing sustainable development [4]:

1. **Materials**: What materials do we need? How many materials do we need? How efficiently are we using these materials? Is there a supply chain? Is the supply chain reliable and secure? Important

concerns to account for include the growing demand for materials (as an example, the European Union (EU) would need 60 times more lithium by 2050 to be climate neutral [5]), the availability of critical materials on earth (e.g. rare-earth materials), price volatility, monopoly of supply, supply-chain risks, geopolitics, import and export regulations, etc.

2. **Energy**: How much energy is needed for material extraction, transportation, production, use, repair, and end-of-life processing? What is the energy source? Is the energy source reliable and secure? Some materials require substantial amounts of energy to extract, and if material extraction is done mostly using brown energy sources, there is a non-negligible carbon footprint associated with material extraction. For example, the extraction of 1 kg gold requires around 250 billion joules of energy and leads to around 15 tonnes of CO_2 emissions [6].

3. Environment: What is the carbon footprint of a device throughout the entire lifetime of a device from production to use to end-of-life processing? Are the carbon emissions during production offset by reducing carbon emissions during a device's lifetime? What is the environmental impact to the air, water, and land? How many land resources are needed? What is the impact on biodiversity?

4. **Regulation**: What are the national and international regulations regarding material use? Are there export and import rules for materials and/or components? What do legislation and directives stipulate regarding the collection and recycling of devices at the end of their lifetimes? Legislation and subsidies to stimulate the green energy transition may have significant impact on how the economy invests in its decarbonization.

5. **Society**: Will the development create jobs and welfare? Will the affected communities along the entire chain (from material supply to end-of-life recycling) benefit from the development? Are there potential concerns regarding health during the production process, usage, and end-of-life processing?

6. **Economics**: Is the development economically viable? What is the cost-benefit balance? Is the upfront investment going to generate revenue and financial benefits?

To illustrate the inherent complexity and multi-dimensionality of sustainable development, let us provide a concrete real-life example: Ireland has decided to limit data-centre construction until 2028 [7]. The reason is that allowing more data centres to be deployed would compromise the country's commitment that 80 percent of the nation's electricity grid should come from renewables by 2030, i.e. Ireland is unable to build renewable capacity fast enough to meet all demands and at the same decarbonize the grid. This example illustrates how sustainable development is a multi-objective optimization problem affecting all six dimensions:

- **materials** (to build renewable energy capacity)
- **energy provision** (to decarbonize the grid)
- **environment** (to reduce air pollution),
- **regulation** (the moratorium on data-centre construction)
- **society** and **economics** (the decision affects employment and the nation's welfare)

The Ireland example is not an isolated case; in fact, several countries are pushing for European legislation for tighter control over the instalment of data centres that consume vast amounts of electricity [8]. Furthermore, following newly adopted legislation, data-centre operators (like any other large company in Europe) will be required to report how their business activities affect sustainability [9].

Another example illustrates that carbon-free operation does not necessarily imply the most carbon-efficient solution. Acun et al. [10] point out that a data centre that operates solely on renewable energy does not minimize the total carbon footprint because of the large number of solar panels, wind farms, and batteries needed to enable carbon-free operation. The reason is that the embodied carbon emissions to produce and manufacture the renewable-energy devices (solar panels, wind farms and batteries) outweigh the operational carbon emissions saved during the lifetime of the data centre. This implies that, to minimize the total carbon footprint of a data centre, a more holistic approach is needed that accounts for both the embodied and operational emissions, rather than just focusing on the operational side.

Understanding trends in environmental impact

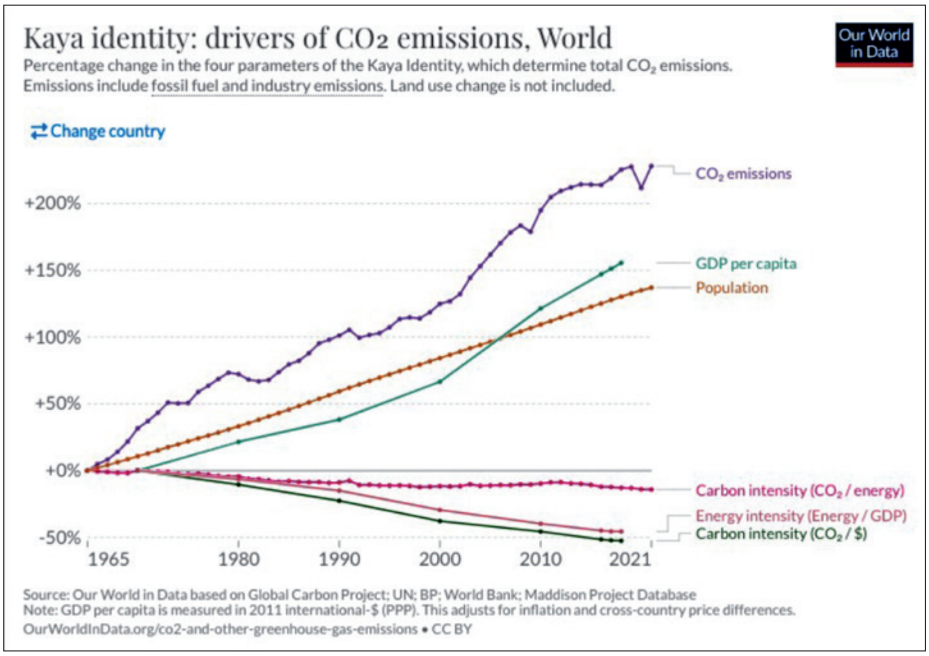
To understand the overall environmental impact of humankind, it is enlightening to go back to a simple yet informative formula developed by the biologist Paul Ehrlich and environmental scientist John Holdren in 1971: $I=P{\cdot}A{\cdot}T$. This formula quantifies the impact I of human activity on the environment as a function of the population P, the affluence per person A, and the impact technology T has on the environment per unit of affluence.

The world's population is growing and so is the average affluence per person. If the growth rate of the world's population and the average per-person affluence exceeds the reduction by technology, the environmental impact increases. This is happening today: the earth overshoot day – the date when the world's population has used all the biological resources that the Earth regenerates during the entire year – moved from a day towards the end of December in 1971 to end of July in 2022 [11].

The environmental impact in the IPAT equation can be measured along several dimensions, including materials used, GHG emissions, water pollution, biodiversity, etc. Yoichi Kaya, an energy economist, reformulates the IPAT equation to specifically focus on carbon dioxide (CO_2) emissions:

$$F=P \cdot G/P \cdot E/G \cdot F/E$$

where P represents the world's population, G/P the gross domestic product (GDP) per capita, E/G the energy intensity or the amount of energy consumed per unit



Kaya identity: drivers of CO2 emissions, World

Percentage change in the four parameters of the Kaya Identity, which determine total CO₂ emissions. Emissions include fossil fuel and industry emissions. Land use change is not included.

Source: Our World in Data based on Global Carbon Project; UN; BP; World Bank; Maddison Project Database
Note: GDP per capita is measured in 2011 international-$ (PPP). This adjusts for inflation and cross-country price differences.
OurWorldInData.org/co2-and-other-greenhouse-gas-emissions • CC BY

Figure 1: Kaya identity: population growth and affluence per capita growth outpace the decrease in energy and carbon intensity, leading to a net overall increase in carbon emissions. Taken from [12].

of GDP, and F/E the carbon intensity per unit of energy. The growth rate in in the global population and GPD per capita is currently outpacing the decrease in energy intensity and carbon intensity, leading to a net annual increase in carbon emissions, as illustrated in Figure 1.

While the IPAT and Kaya equations are insightful and widely used, they should be interpreted with care. In particular, the equations suggest that the different variables are independent of each other. However, they are not. For example, reducing the energy intensity of a device or service typically leads to a price reduction, which in turn may stimulate consumption. If the increased consumption outweighs the energy intensity reduction, we end up with a net increase in environmental impact – exactly opposite of what we had envisioned!

This is the well-known Jevons' Paradox, named after Williams Stanley Jevons, who was the first to observe the rebound effect of the steam engine's improved coal efficiency leading to an overall increase in coal consumption [13]. This rebound effect is also (part of) the reason why improving energy or power efficiency of a computing device does not necessarily lead to a net reduction in environmental impact. Most often, an energy- or power-efficiency gain leads to increased usage and deployment, effectively increasing the environmental impact of computing – as we will discuss further.

ICT's environmental impact

We will now leverage the above equations to analyse the environmental impact of computing. To do so, it is important to make a distinction between embodied versus operational emissions [14]. For the discussion that follows, we will mostly focus on GHG emissions, but several aspects also pertain to other environmental concerns, as we will point out.

Embodied emissions relate to raw-material extraction, manufacturing, assembly, transportation, repair, maintenance, and end-of-life processing. Operational emissions relate to product use during a device's lifetime.

Embodied emissions can be further categorized in scope-1, scope-2, and scope-3. Scope-1 refers to the chemicals and gases used during manufacturing – this includes fluorinated greenhouse gases with orders of magnitude higher global warming potential than CO_2. Scope-2 refers to the energy consumption during chip manufacturing – this includes empowering the extensive production facilities with hundreds of manufacturing tools and requiring climate and humidity control. Scope-3 pertains to the energy consumption for the extraction and production of materials used for integrated circuit manufacturing. For the purposes of this discussion, we will focus on scope 1 and scope-2, because scope-3 follows a similar pattern. See [15] for more details.

The embodied scope-2 emissions of computing can be modelled as follows:

$$F_{(scope\text{-}2)} = C \cdot W/C \cdot E/W \cdot F/E$$

where C represents the number of chips produced, W/C the number of wafers needed per chip, E/W the amount of energy needed per wafer, and F/E the carbon intensity of chip manufacturing. The number of wafers needed per chip W/C tends to stagnate as we approach the reticle limit of chip manufacturing, i.e. the maximum size that lithography machines can process with a single mask. In contrast, the number of chips C tends to increase at a growth rate of 9% per year [16], and the amount of energy needed for manufacturing E/W increases at a rate of 11.9% per year as we transition to new technology nodes, according to recent data provided by imec [17]. The carbon intensity of chip manufacturing is not improving fast enough to compensate for the increase in chip demand and energy intensity of manufacturing, which as a result leads to an overall annual increase in scope-2 emissions due to chip manufacturing.

The embodied scope-1 emissions can be modelled similarly:

$$F_{(scope\text{-}1)} = C \cdot W/C \cdot F/W$$

where F/W represents the carbon dioxide equivalents due to fluorinated compounds per wafer. Imec data reports that this factor is increasing by slightly more than 9.3% per year [17]. With the number of chips increasing by 9% per year and chip die size being constant, embodied scope-1 emissions are hence trending up.

Operational emissions can be modelled as follows:

$$F_{operational} = C \cdot E/C \cdot F/E$$

where *E/C* represents the total electricity usage of a chip over its entire lifetime and *F/E* the carbon intensity during device use. When there is work to be done, a chip's operational emissions are proportional to its energy consumption. When the chip is idle, operational emissions are proportional to the chip's idle power. A variety of optimizations across the system stack improve the energy and power efficiency of individual devices: from transistor tuning to dynamic voltage and frequency scaling, clock gating, power gating, etc. The question is whether the per-device energy and power efficiency improvements are outweighed, or even worse, outpaced, by the increase in the number of chips deployed.

What do the trends look like?

Gupta et al. [14] performed a survey of consumer devices from vendors including Apple, Google, Huawei and Microsoft. They conclude that embodied emissions dominate for battery-operated devices such as wearables, smartphones, tablets, and laptops, while operational emissions dominate for always-connected devices such as speakers, desktop computers and gaming consoles. For data centres, most emissions are related to construction, infrastructure, and hardware manufacturing: interestingly, while total energy usage is trending up – presumably because of increased server count and/or higher degree of consolidation (cf. Jevons' paradox) – total operational emissions are decreasing for Facebook and Google, thanks to their policy of contracting and securing green energy sources to power their hyperscale data centres.

Making decisive conclusions about the environmental impact of specific computing devices is inherently difficult because of the variety of use cases in which computer systems are manufactured, deployed, and used. For example, the use of green energy sources during chip manufacturing may shift the contribution from embodied to operational emissions. Nevertheless, the overall conclusion that can be reached based on the above equations [15] indicates that embodied emissions are continuing to grow under current scaling trends, and that embodied emissions already are, or will soon be, the biggest contributor; see also Figure 2. The fundamental reason is the increasing demand for chips (because of economic dynamics based on selling products) and the growing energy intensity of semiconductor manufacturing (because of advancements in chip technology), which do not seem to be counterbalanced by the transition to green energy sources and improvements in per-device energy and power efficiency.

Looking forward

There are several important conclusions to be taken from the above analysis.

First and foremost, to reduce both the embodied and operational emissions of computing, we could reduce the number of chips that we produce and sell. This could possibly be achieved by integrating more functionality within individual chips. Modern-day heterogeneous system-on-chip (SoC) designs integrate a couple of dozen accelerators in addition to central processing unit (CPU) and graphics processing unit (GPU) cores, yet this has not led to a reduction in carbon footprint, on the contrary (yet another example of Jevons' paradox).

Current business models are based on selling devices and hence stand in the way of reducing the number of chips that we produce. The number of connected devices is rapidly increasing: Cisco estimates that the internet of things (IoT) was born between 2008 and 2009 when there started to be more connected devices than people; today there are more than seven connected devices per person – this number is even higher in the Western world (up to 12.9 and 8.9 devices per person in North America and Western Europe, respectively) [18].

Service-model based business models, such as leasing a smartphone as recently offered by Fairphone [19], may incentiv-

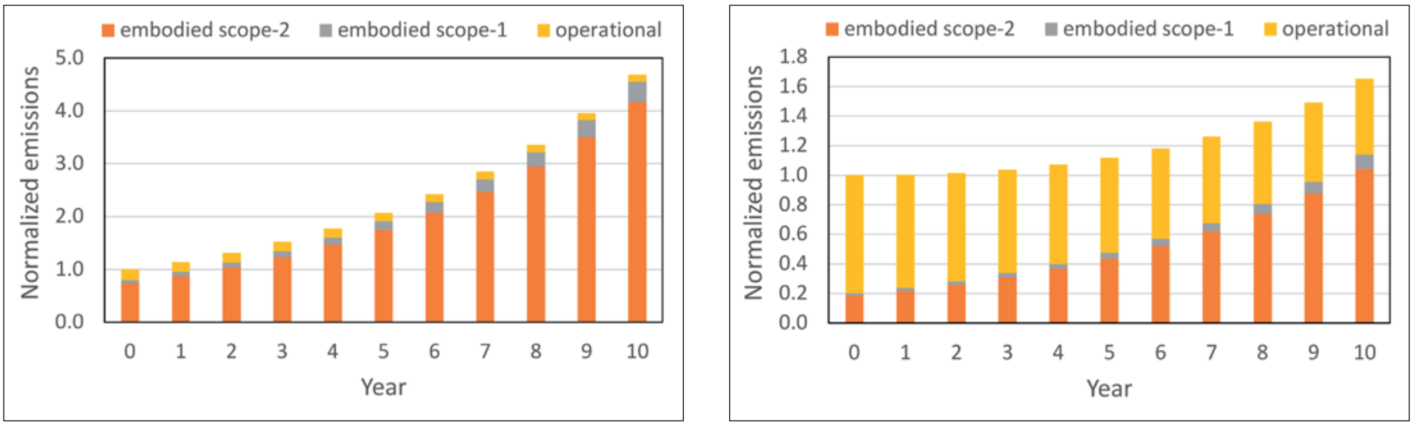


Figure 2: Projection for total emissions over the next decade given current scaling trends for two scenarios: (left) embodied emissions dominate initially (80% of total emissions in year zero), and (right) operational emissions dominate initially (80% of total emissions in year zero). Total emissions are increasing dramatically, and embodied emissions are, or will start, dominating. This analysis assumes the following annual growth rates: +9% number of chips C, +11.9% wafer energy intensity (E/W), +9.3% wafer chemical/gas intensity (F/W), -2.5% carbon intensity (F/E), -10% operational energy intensity (E/C), and 0% die size (W/C) – note the assumed negative growth rates for carbon and operational energy intensity, and the zero growth rate for chip die size. Taken from [15].

ize manufacturers to design systems with a longer lifetime that can be repaired, remanufactured, reassembled, reused, recycled, etc. (See the article "Everything as a service" in this HiPEAC Vision for more examples of service-based offerings and their contribution to sustainability.)

Another (complementary) approach to reduce the demand for more chips could be to prolong the lifetime of a device by deploying fault-tolerance techniques to fix errors or to enable graceful degradation (e.g. disabling a faulty core in multicore CPUs or GPUs). Reprogrammable hardware, e.g. field-programmable gate arrays (FPGAs), and hardware reconfigurability could also be viable enablers to provide hardware acceleration in a more sustainable way.

Indeed, it is a fact that electronic devices have a short lifespan and e-waste is a major problem. In particular, the average lifetime of a smartphone is around four years, but young people tend to replace their smartphones more quickly: 40% of 18- to 24-year-olds keep their phones for less than a year [20]. As a result, e-waste is a major issue, with more than 7 kg of e-waste per person on average and more than 16 kg of e-waste per person in Europe [21].

We could design smaller chips to reduce the embodied footprint per chip. There is some leeway in fact because of Moore's Law. Moving to a new chip technology node offers twice the number of transistors for the same chip area, but because of the increased energy intensity of new technology nodes, the total embodied footprint per chip in fact increases. There is a middle ground, though in which computer architects use only a fraction of the additional transistors (for as long as Moore's Law continues to hold) to add functionality such that the total embodied footprint does not increase or, even better, decreases.

In fact, this implies that we need to design smaller chips, albeit with a (slightly) higher transistor count than the previous generation. As projected in [15], reducing die size by 25% each year still allows 6% more transistors each year to add new functionality, and yet embodied scope-2 emissions would decrease by 12% each year. (Note that smaller chips not only help reduce the embodied footprint; they also improve manufacturing yield.) This path, unfortunately, is not what industry is currently pursuing, in part presumably because of market dynamics and competition.

Manufacturing chips in older, less energy-demanding chip technologies could also help reduce embodied emissions. That potentially comes at the cost of higher operational emissions because of a less energy-efficient chip technology. Whether the trade-off balances towards more versus less sustainable system design remains to be seen. In any case, if this pathway turns out to be promising, now could be the right time given the recent emergence of chiplet-based integration in which different chiplets may be manufactured in different chip technologies, thereby reducing the overall embodied footprint of chiplet-integrated devices.

Rapidly transitioning to green energy sources for manufacturing will drastically reduce carbon emissions. In fact, TSMC has engaged itself to supply 25% of its fabrication plant power supply from renewable energy and be carbon-neutral by 2050 [22]. While this greatly affects embodied scope-2 emissions, it does not affect scope-1 or scope-3 emissions. In addition, it does not reduce other sustainability concerns such as those around raw materials and ultrapure water supply needed for manufacturing, e-waste, etc.

Moreover, if the green energy supply is appropriated from the global market through green energy contracts, this does not fundamentally reduce the carbon footprint at the global societal scale because other users are hence deprived from green energy. Finally, green energy sources are not carbon free either – solar panels, wind turbines, etc. also incur an embodied carbon footprint for manufacturing, maintenance, transportation, and end-of-life handling.

Reducing a chip's operational emissions, while less important than reducing its embodied emissions, is still an important optimization criterion. Lowering energy consumption when there is work to be done reduces a chip's operational emissions. Lowering idle power consumption when there is no work to be done, also reduces a chip's operational emissions. Note though that efficiency improvements are possibly (in practice, frequently) subject to Jevons' paradox. For example, energy optimization due to higher performance enables more jobs per unit of time to be executed, thereby increasing overall energy consumption. Likewise, a power-saving optimization may enable more concurrent jobs within the available power envelope, which may lead to an overall increase in energy consumption.

Sustainable design and inherent data uncertainty

It is clear from the above discussion that improving the sustainability of computing systems is complicated and requires a holistic design approach that touches upon a variety of design criteria including design complexity and chip area, performance, energy and power efficiency, reliability, and fault tolerance.

Computer architects are well versed in optimizing along a single design criterion while taking other design criteria into consideration, for example, optimizing performance with limited impact on power consumption and design complexity, or improving reliability with limited impact on performance and energy consumption. Optimizing for sustainability, on the other hand, requires a more holistic approach, considering all design criteria and stakeholders at the same time while optimizing the impact on the overall embodied and operational footprint and being subject to significant degrees of uncertainty.

For example, while a fault-tolerance technique that fixes hard errors or enables graceful degradation may prolong the lifetime of a device, thereby damping the quest for more chips, it comes at the cost of increased embodied footprint (to provide the fault-tolerance hardware circuitry) and operational footprint (to dynamically monitor the operation during the device's lifetime). Whether a fault-tolerance technique leads to net overall reduction in environmental footprint depends on the relative importance of the embodied versus operational footprint, the likelihood of an error, and the typical use case of the device under design. Optimizing sustainable computer systems is without any doubt a challenging design problem.

This design challenge is further complicated by the large degree of uncertainty in a variety of dimensions. While companies' sustainability reports and product lifecycle-assessment (LCA) reports provide a wealth of data, there remain many unknowns and data limitations, in part because of industry secretiveness, or simply because of lack of reliable data. For example, a recent study by imec [17], which attempts to quantify the environmental footprint of modern-day chip manufacturing, makes assumptions regarding the energy consumption of a fab's facility equipment (i.e. it is *"assumed to contribute to 40% of the total energy"*); furthermore, the degree of abatement of fluorinated GHGs (scope-1) is unknown, as well as the use of materials and the energy needed for material extraction (scope-3). As another example, the Apple iPhone12 LCA report [23] uses industry averages when parameters are unknown for the production process, i.e. a company may not know the sustainability impact of its suppliers.

The operational footprint and its importance relative to the embodied footprint is even harder to assess, as it depends on typical user behaviour, product lifetime, and the geographic location of the user (which determines the carbon intensity of the user's power grid mix). Historical data could be insightful, but it only provides a hint. Note further that product use may be subject to the infamous rebound effect, which may significantly shift the relative importance of the operational versus embodied footprint.

Overall, it is safe to conclude that there is inherent data uncertainty. Gupta et al. [24] recently proposed the ACT model to analyse a computer system's sustainability at design time. This model relies on detailed numbers from production processes in industry. This is an important step for our community at large (both in industry and academia). Nevertheless, the authors note that there is *"lack of up-to-date carbon emission data for the latest compute, memory, and storage technologies"*. Furthermore, they hope to *"encourage industry to publish more detailed carbon characterizations to standardize carbon footprint accounting"*. Imec's sustainable semiconductor technology and systems (SSTS) program aims at addressing exactly this issue by collaborating with major industry players to quantify the environmental impact of integrated circuit manufacturing [25]. While significant progress is being made regarding the embodied footprint of computing, more is needed. Moreover, an equally substantial effort needs to be made to quantify the operational footprint of computing, which may turn out to be even more challenging.

Sustainable design based on first principles

And yet, despite the large degrees of uncertainty and the multi-faceted design problem, computer architects need to make design decisions to make computer systems more sustainable. One option may be to revert to first principles and guide sustainable design decisions using a first-order model. First-order modelling should not be viewed as a replacement for, but rather as a useful complement to, detailed models like ACT and others. In fact, a detailed sustainability accounting method can provide initial data for a first-order model, and

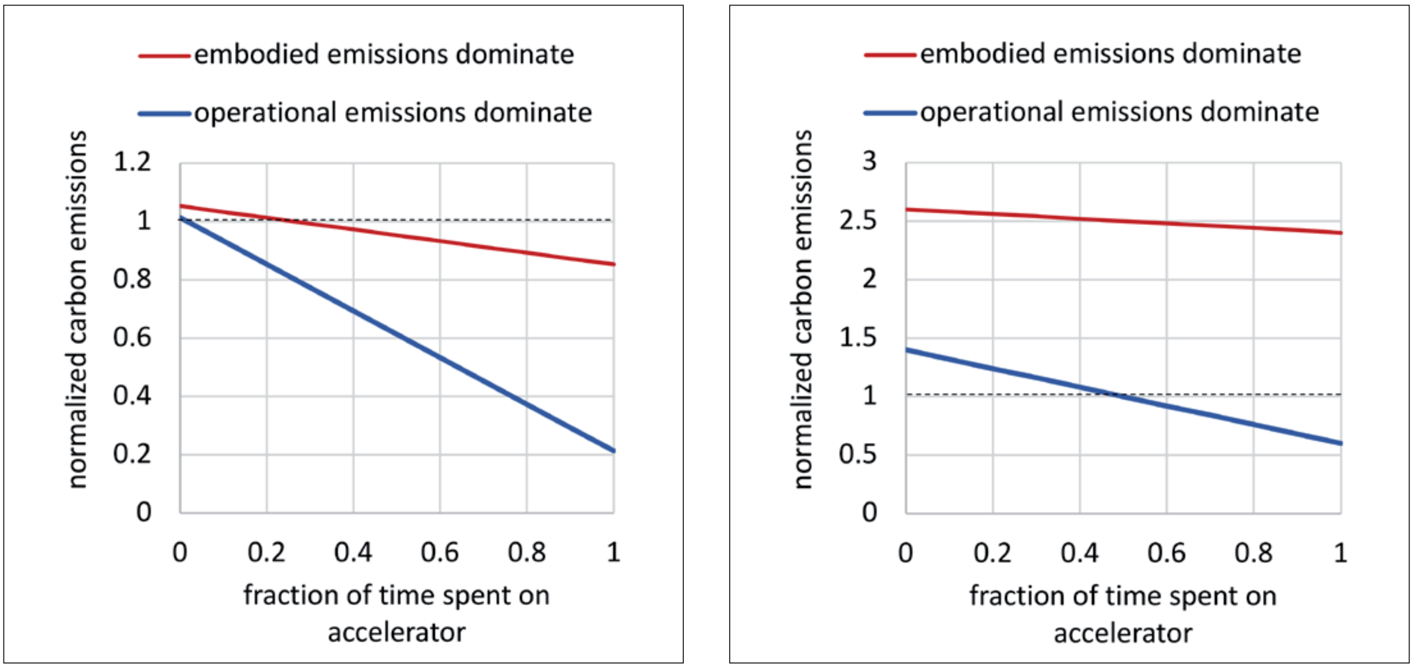


Figure 3: Total carbon footprint of a general-purpose CPU plus accelerator as a function of its degree of use, assuming that the accelerator takes up 6.5% extra chip area (left) versus 2x extra chip area (right), normalized to a general-purpose CPU without an accelerator. The accelerator is assumed to consume 500x less energy than the general-purpose CPU for performing the same work. Two scenarios are considered: embodied emissions account for 80% of total emissions versus 20% of total emissions. The larger the chip area of the accelerator, the more frequently it needs to be used and the higher the relative weight of the operational emissions need to be for the accelerator to be sustainable. Taken from [10].

vice versa, a first-order model can provide directions where the detailed model should be further refined.

A first-order model uses proxies for the embodied and operational footprint that computer architects have control over, see for example [10] for more details about a first-order model for computer chips. A useful, first-order proxy for the embodied footprint of a chip is its die size, i.e. the larger the chip, the higher the embodied footprint for a given chip technology in terms of the energy and materials needed and the chemicals and gases emitted during production. A useful proxy for the operational footprint of a chip is energy consumption assuming a fixed-work scenario (i.e. a device performs a fixed amount of work during its entire lifetime) and power consumption assuming a fixed-time scenario (i.e. a device is used for the same amount of time, and hence performs more work). The relative importance of embodied versus operational emissions can be captured via a parameter the architect can vary to explore different use case scenarios.

Although (deliberately) simple, a first-order sustainability model can reveal a variety of interesting insights which computer architects can take forward to design more sustainable computer systems. There is a fruitful avenue of future work to explore how computer architectures can be made more sustainable. Analysing to what extent archetypical CPU and GPU design paradigms and solutions (e.g. caching, speculation, microarchitecture, acceleration, etc.) affect computer system sustainability would be extremely valuable.

For example, as reported in [10], the first-order model can be used to assess whether hardware specialization is sustainable. Integrating a hardware accelerator next to a general-purpose processor incurs a cost in terms of embodied footprint (because of a larger chip) which may be compensated for by the reduced operational footprint (because of lower energy consumption when using the special-purpose accelerator rather than a general-purpose CPU). In other words, the reduced operational footprint amortizes the increased embodied footprint.

The question is where the tipping point is. The larger the accelerator, the more frequently the accelerator needs to be used and the higher the relative weight of the operational emissions needs to be for the accelerator design to be sustainable, as illustrated in Figure 3; if the accelerator is taking up significant chip area, and the embodied emissions dominate, the reduction in operational emissions does not compensate for the increased embodied emissions.

This suggests that the current trend towards large system-on-chip (SoC) designs with dozens of accelerators that occupy a significant fraction of the chip and that are not powered on all the time due to dark-silicon constraints, may not be a sustainable design paradigm. A more fruitful, sustainable design paradigm might be to consolidate accelerator designs to a common-denominator accelerator that can serve multiple critical applications while incurring less chip area, thereby reducing the embodied footprint at the expense of an increased operational footprint, with a net improvement in sustainability. Investigating these (and other) architecture trade-offs in more detail is a promising research avenue for computer architects in industry and academia.

Conclusion

Improving computing-system sustainability is a challenging and multi-faceted problem. The embodied footprint is, or will soon be, a more important contributor than the operational footprint, primarily due to an increasing demand for chips and increased energy intensity of integrated circuit manufacturing. Decarbonizing the production process and use phase of compute devices is not a panacea, though, because it does not address other sustainability concerns including raw material extraction, chemicals and gases emitted, and ultra-pure water used during production.

What makes sustainable computer system design unique compared to traditional optimization criteria is that it requires a holistic approach considering chip area, power, energy, performance, lifetime, reliability, etc. The field of computer architecture specifically, and computer science and engineering in general, has only recently embarked on this endeavour.

Computer architects should continue to (1) collect high-quality data to assess the sustainability impact across the entire lifetime of a computing device, from raw-material extraction, transportation, manufacturing, assembly, use, repair, end-of-life processing, etc., (2) develop detailed and high-abstraction models to help designers evaluate the impact on sustainability at design time, and (3) analyse and revisit architecture design paradigms considering their sustainability impact. Overall, sustainable system design is an extremely timely and societally important topic where substantial innovation is to be achieved and expected in the following years.

References

[1] "Climate Plans Remain Insufficient: More Ambitious Action Needed Now," United Nations, 26 October 2022. [Online]. Available: https://unfccc.int/news/climate-plans-remain-insufficient-more-ambitious-action-needed-now. [Accessed 28 November 2022].

[2] "Climate Action Explore policy solutions by key economic sector," OECD, [Online]. Available: https://www.oecd.org/stories/climate-action/key-sectors/. [Accessed 28 November 2022].

[3] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, A. Friday, "The Real Climate and Transformative Impact of ICT: A Critique of Estimates, Trends, and Regulations," Patterns, vol. 2, no. 9, pp. 100340, https://doi.org/10.1016/j.patter.2021.100340, 2021.

[4] M. Ashby, Materials and Sustainable Development, 1st edition, Elsevier, 2015.

[5] "Critical Raw Materials Resilience: Charting a Path towards greater Security and Sustainability," 2 September 2020. [Online]. Available: https://ec.europa.eu/docsroom/documents/42849. [Accessed 28 November 2022].

[6] L. Eeckhout, "A First-Order Model to Assess Computer Architecture Sustainability," IEEE Computer Architecture Letters, vol. 21, no. 2, pp. 137-40, July-Dec 2022.

[7] P. Judge, "EirGrid pulls plug on 30 Irish data center projects," Datacenter dynamics , 24 May 2022. [Online]. Available: https://www.datacenterdynamics.com/en/news/eirgrid-pulls-plug-on-30-irish-data-center-projects/. [Accessed 28 November 2022].

[8] A. Roach and E. Krukowska, "Big Tech Gets Caught Up in Europe's Energy Politics," Bloomberg, 23 June 2022. [Online]. Available: https://www.bloomberg.com/news/articles/2022-06-23/google-facebook-data-centers-face-europe-political-snags-over-in-energy-crisis?sref=FwE94DUF. [Accessed 28 November 2022].

[9] "Corporate Sustainability Reporting Directive (CSRD) in "A European Green Deal"," European Parliament, 20 October 2022. [Online]. Available: https://www.europarl.europa.eu/legislative-train/theme-a-european-green-deal/file-review-of-the-non-financial-reporting-directive. [Accessed 28 November 2022].

[10] B. Acun, B. Lee, K. Maeng, M. Chakkaravarthy, U. Gupta, D. Brooks, C.-J. W, "Carbon Explorer: A Holistic Approach for Designing Carbon-Aware Datacenters," in ACM International Conference on Architecture Support for Programming Languages and Operating Systems (ASPLOS), Vancouver, 2023.

[11] "Earth Overshoot Day," [Online]. Available: https://www.overshootday.org/. [Accessed 28 November 2022].

[12] H. Ritchie and M. Roser, "Emissions drivers," Our World in Data, [Online]. Available: https://ourworldindata.org/emissions-drivers. [Accessed 28 November 2022].

[13] "W. Stanley Jevons, "The Coal Question," 1865," Yale University, [Online]. Available: https://energyhistory.yale.edu/library-item/w-stanley-jevons-coal-question-1865. [Accessed 28 November 2022].

[14] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, C.-J. Wu, "Chasing Carbon: The Elusive Environmental Footprint of Computing," in IEEE International Symposium on High-Performance Computer Architecture (HPCA), Virtual, 2021.

[15] L. Eeckhout, "Kaya for Computer Architects: Towards Sustainable Computer Systems," IEEE Micro, pp. 1-8, https://ieeexplore.ieee.org/document/9932869, 2022.

[16] "The McClean Report 2022," IC Insights, January 2022. [Online]. Available: https://www.icinsights.com/services/mcclean-report/. [Accessed 28 November 2022].

[17] M. Garcia Bardon, P. Wuytens, L.-A. Ragnarsson, G. Mirabelli, D. Jang, G. Willens, A. Mallik, S. Spessot, J. Ryckaert, B. Parvais, "DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies," in 2020 IEEE International Electron Devices Meeting (IEDM), Virtual, 2020.

[18] "Cisco Annual Internet Report (2018–2023) White Paper," Cisco, 9 March 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html. [Accessed 28 November 2022].

[19] "Fairphone Easy: The sustainable smartphone solution," Fairphone, [Online]. Available: https://shop.fairphone.com/en/fairphone-easy. [Accessed 28 November 2022].

[20] Z. Muhammad, "People have started changing their smartphone less often," Digital Information World, 28 August 2019. [Online]. Available: https://www.digitalinformationworld.com/2019/08/how-often-should-you-upgrade-your-phone-infographic.html. [Accessed 28 November 2022].

[21] I. Tiseo, "Per capita electronic waste generation worldwide from 2010 to 2019," Statista, 4 October 2022. [Online]. Available: https://www.statista.com/statistics/499904/projection-ewaste-generation-per-capita-worldwide/ . [Accessed 28 November 2022].

[22] P. Lin and L. Sun, "TSMC Becomes the World's First Semiconductor Company to Join RE100, Committed to 100% Renewable Energy Usage," TSMC ESG, 27 July 2020. [Online]. Available: https://esg.tsmc.com/en/update/greenManufacturing/caseStudy/37/index.html. [Accessed 28 November 2022].

[23] "Product Environmental Report: iPhone 12," 13 October 2020. [Online]. Available: https://www.apple.com/environment/pdf/products/iphone/iPhone_12_PER_Oct2020.pdf.

[24] 8. U. Gupta, M. Elgamal, G. Hills, G.-Y. Wei, H.-H. S. Lee, D. Brooks, C.-J. Wu, "ACT: Designing Sustainable Computeer Systems with an Architectural Carbon Modeling Tool," in ISCA '22: Proceedings of the 49th Annual International Symposium on Computer Architecture, New York, 2022.

[25] L.-A. Ragnarsson, C. Rolin, S. Shamuilia, E. Parton, "The green transition of the IC industry," Imec, [Online]. Available: https://www.imec-int.com/en/expertise/cmos-advanced/sustainable-semiconductor-technologies-and-systems-ssts/stss-white-paper. [Accessed 28 November 2022].

Lieven Eeckhout is a senior full professor in the department of electronics and information systems at Ghent University, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2023.

Cite as: Lieven Eeckhout. Towards sustainable computer architecture: A holistic approach. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 216-224, Jan 2023.

DOI: 10.5281/zenodo.7461989

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.