Frugal Computing

On the need for low-carbon and sustainable computing and the path towards zero-carbon computing

Wim Vanderbauwhede

***Abstract*—The current emissions from computing are almost 4% of the world total. This is already more than emissions from the airline industry and are projected to rise steeply over the next two decades. By 2040 emissions from computing alone will account for more than half of the emissions budget to keep global warming below 1.5°C. Consequently, this growth in computing emissions is unsustainable. The emissions from production of computing devices exceed the emissions from operating them, so even if devices are more energy efficient producing more of them will make the emissions problem worse. Therefore we must extend the useful life of our computing devices.**

As a society we need to start treating computational resources as finite and precious, to be utilised only when necessary, and as effectively as possible.

We need *frugal computing*: achieving our aims with less energy and material.

I. Defining computational resources

Computational resources are all resources of energy and material that are involved in any given task that requires computing. For example, when you perform a web search on your phone or participate in a video conference on your laptop, the computational resources involved are those for production and running of your phone or laptop, the mobile network or WiFi you are connected to, the fixed network it connects to, the data centres that perform the search or video delivery operations. If you are a scientist running a simulator in a supercomputer, then the computational resources involved are your desktop computer, the network and the supercomputer. For an industrial process control system, it is the production and operation of the Programmable Logic Controllers, IoT devices etc.

II. Computational resources are finite

Since the start of general purpose computing in the 1970s, our society has been using increasing amounts of computational resources.

For a long time the growth in computational capability as a function of device power consumption has literally been exponential, a trend expressed by Moore's law.

With this growth in computational capability, increasing use of computational resources has become pervasive in today's society. Until recently, the total energy budget and carbon footprint resulting from the use of computational resources has been small compared to the world total. As a result, computational resources have until recently effectively been treated as unlimited.

Because of this, the economics of hardware and software development have been built on the assumption that with every generation, performance would double for free. Now, this unlimited growth is no longer sustainable because of a combination of technological limitations and the climate emergency. Therefore, we need to do more with less.

Moore's law has effectively come to an end as integrated circuits can't be scaled down any more. As a result, the improvement in performance per Watt is slowing down continuously. On the other hand, the demand for computational resources is set to increase considerably.

The consequence is that at least for the next decades, growth in demand for computational resources will not be offset by increased power efficiency. Therefore with business as usual, the total energy budget and carbon footprint resulting from the use of computational resources will grow dramatically to become a major contributor to the world total.

Furthermore, the resources required to create the compute devices and infrastructure are also finite, and the total energy budget and carbon footprint of production of compute devices is huge. Moore's Law has conditioned us to doubling of performance ever two years, which has led to very short effective lifetimes of compute hardware. This rate of obsolescence of compute devices and software is entirely unsustainable.

Therefore, as a society we need to start treating computational resources as finite and precious, to be utilised only when necessary, and as frugally as possible. And as computing scientists, we need to ensure that computing has the lowest possible energy consumption. And we should achieve this with the currently available technologies because the lifetimes of compute devices needs to be extended

dramatically.

I would like to call this "frugal computing": achieving the same results for less energy by being more frugal with our computing resources.

III. The scale of the problem

A. Meeting the climate targets

To limit global warming to 1.5°C, within the next decade a global reduction from 55 gigatonnes C_2 equivalent ($GtCO_2e$) by 32 $GtCO_2e$ to 23 $GtCO_2e$ per year is needed [6]. So by 2030 that would mean a necessary reduction in overall CO_2 emissions of more than 50%. By 2040, a further reduction to 13 $GtCO_2e$ per year is necessary. According to the International Energy Agency [11], emissions from electricity are currently estimated at about 10 $GtCO_2e$.

The global proportion of electricity from renewables is projected to rise from the current figure of 22% to slightly more than 30% by 2040 [16]. A more optimistic scenario by the International Energy Agency [18] projects 70% of electricity from renewables, but even in that scenario, generation from fossil fuels reduces only slightly, so there is only a slight reduction in emissions as a result.

In other words, we cannot count on renewables to eliminate CO_2 emissions from electricity in time to meet the climate targets. The same is true for nuclear, offsetting and CCS: they will all come too late. Reducing the energy consumption is the only option.

B. Emissions from consumption of computational resources

The consequence of the end of Moore's law was expressed most dramatically in a 2015 report by the Semiconductor Industry Association (SIA) "Rebooting the IT Revolution: a call to action" [1], which calculated that, based on projected growth rates and on the 2015 ITRS roadmap for CMOS chip engineering technologies [17]

> computing will not be sustainable by 2040, when the energy required for computing will exceed the estimated world's energy production.

It must be noted that this is purely the energy of the computing device, as explained in the report. The energy required by e.g. the data centre infrastructure and the network is not included.

The SIA has reiterated this in their 2020 "Decadal Plan for Semiconductors" [2], although they have revised the projection based on a "market dynamics argument":

> If the exponential growth in compute energy is left unchecked, market dynamics will limit the growth of the computational capacity which would cause a flattening out the energy curve.

This is merely an acknowledgement of the reality that the world's energy production is not set to rise dramatically, and therefore increased demand will result in higher prices which will damp the demand. So computation is not actually going to exceed the world's energy production.

> Ever-rising energy demand for computing vs. global energy production is creating new risk, and new computing paradigms offer opportunities to dramatically improve energy efficiency.

In the countries where most of the computational resources are consumed (US and EU), electricity production accounts currently for 25% of the total emissions [5]. According to the SIA's estimates, computation accounts currently for a little less than 10% of the total electricity production but is set to rise to about 30% by 2040. This would mean that, with business as usual, computational resources would be responsible for at least 10% of all global CO_2 emissions by 2040.

The independent study "Assessing ICT global emissions footprint: Trends to 2040 & recommendations" [3] corroborates the SIA figures: they estimate the computing greenhouse gas emissions for 2020 between 3.0% and 3.5% of the total, which is a bit higher than the SIA estimate of 2.5% because it does take into account networks and datacentres. Their projection for 2040 is 14% rather than 10%, which means a growth of 4x rather than 3x.

To put it in absolute values, based on the above estimate, by 2040 energy consumption of compute devices would be responsible for 5 $GtCO_2e$, whereas the target for world total emissions from all sources is 13 $GtCO_2e$.

Other projections are of the same order, and the key message is that "computing's carbon footprint growing at a rate unimaginable in other sectors." [4].

C. Emissions from production of computational resources

To make matters worse, the carbon emissions resulting from the production of computing devices exceeds those incurred during operation. This is a crucial point, because it means that we can't rely on next-generation hardware technologies to save energy: the production of this next generation of devices will create more emissions than any operational gains can offset. It does not mean research into more efficient technologies should stop. But their deployment cycles should be much slower. Extending the useful life of compute technologies must become our priority.

The report about the cost of planned obsolescence by the European Environmental Bureau [8] makes the scale of the problem very clear. For laptops and similar computers, manufacturing, distribution and disposal account for 52% of their Global Warming

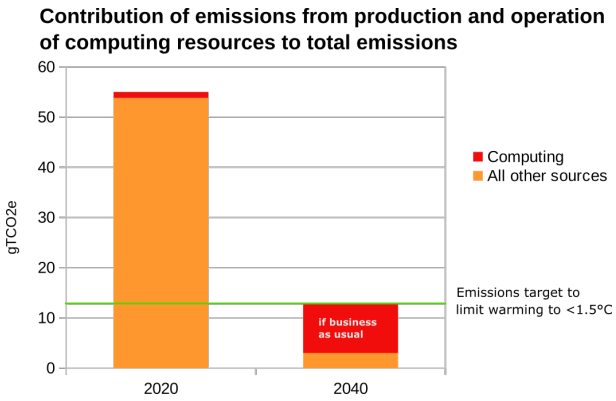

Fig. 1. Actual and projected emissions from computing (production+operation), and 2040 emission target to limit warming to $\leq$1.5°C

Device type	2020 sales
Phones	3000M
Servers	13M
Tablets	160M
Displays	40M
Laptops	280M
Desktops	80M
TVs	220M
IoT devices	2000M

TABLE I
NUMBER OF DEVICES SOLD WORLDWIDE IN 2020

Device type	TWh
TVs	560
Other Consumer Devices	240
Fixed network (wired+wifi)	1400
Mobile network	100
Data centres	700
Total	3000

TABLE II
YEARLY ENERGY CONSUMPTION ESTIMATES IN TWH

Potential (i.e. the amount of CO_2-equivalent emissions caused). For mobile phones, this is 72%. The report calculates that the lifetime of these devices should be at least 25 years to limit their Global Warming Potential. Currently, for laptops it is about 5 years and for mobile phones 3 years. According to [9], the typical lifetime for servers in data centres is also 3-5 years, which again falls short of these minimal requirements. According to this paper, the impact of manufacturing of the servers is 20% of the total, which would require an extension of the useful life to 11-18 years.

D. The total emissions cost from computing

Taking into account the carbon cost of both operation and production, computing would be responsible for 10 $GtCO_2e$ by 2040, almost 80% of the acceptable CO_2 emissions budget [2], [3], [15], as illustrated in Fig. 1.

E. A breakdown per device type

To decide on the required actions to reduce emissions, it is important to look at the numbers of different types of devices and their energy usage. If we consider mobile phones as one category, laptops and desktops as another and servers as a third category, the questions are: how many devices are there in each category, and what is their energy consumption. The absolute numbers of devices in use are quite difficult to estimate, but the yearly sales figures [11] and estimates for the energy consumption for each category [12], [13], [14], [15] are readily available from various sources. The tables below show the 2020 sales (Table I) and yearly energy consumption estimates (Table II) for each category of devices. A detailed analysis is presented in [15].

The energy consumption of all communication and computation technology currently in use in the world is currently around 3,000 TWh, about 11% of the world's electricity consumption, projected to rise by 3-4 times by 2040 with business as usual according to [2]. This is a conservative estimate: the study in [15] includes a worst-case projection of a rise to 30,000 TWh (exceeding the current world electricity consumption) by 2030.

The above data make it clear which actions are necessary: the main carbon cost of phones, tablets and IoT devices is their production and the use of the mobile network, so we must extend their useful life very considerably and reduce network utilisation. Extending the life time is also the key action for datacentres and desktop computers, but their energy consumption also needs to be reduced considerably, as does the energy consumption of the wired, WiFi and mobile networks.

IV. A VISION FOR LOW CARBON AND SUSTAINABLE COMPUTING

It is clear that urgent action is needed: in less than two decades, the global use of computational resources needs to be transformed radically. Otherwise, the world will fail to meet its climate targets, even with significant reductions in other emission areas. The carbon cost of both production and operation of the devices must be considerably reduced.

To use devices for longer, a change in business models as well as consumer attitudes is needed. This requires raising awareness and education but also providing incentives for behavioural change. And to support devices for a long time, an infrastructure for repair and maintenance is needed, with long-term availability of parts, open repair manuals and training. To make all this happen, economic incentives and policies will be needed (e.g. taxation, regulation). Therefore we need to convince key decision makers in society, politics and business.

A vision for zero-carbon computing

Imagine that we can extend the useful life of our devices and even increase their capabilities, purely

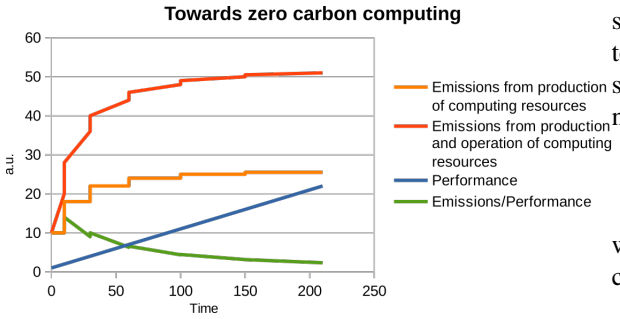

Fig. 2. Towards zero carbon computing: increasing performance and lifetime and reducing emissions. Illustration with following assumptions: every new generation lasts twice as long as the previous one and cost half as much energy to produce; energy efficiency improves linearly with 5% per year.

through better computing science. With every improvement, the computational capacity will in effect increase without any increase in energy consumption. Meanwhile, we will develop the technologies for the next generation of devices, designed for energy efficiency as well as long life. Every subsequent cycle will last longer, until finally the world will have computing resources that last forever and hardly use any energy.

V. Research challenges

This is a very challenging vision, spanning all aspects of computing science. To name just a few of the challenges:

A. Cloud computing

Saving energy during operation, optimising for energy consumption (e.g. DVFS, accelerators, scheduling and placement), more energy-efficient software on all layers; better use of renewables through smarter scheduling. Increasing the useful life, e.g reliability monitoring and early-warning systems, degradation-aware operation

B. Ultra-HD video & VR/AR

Roll-out could lead to order of magnitude increases in video/3D traffic. To mitigate this, we need better compression (e.g. tailored, AI), local rendering (e.g. using FPGAs), better caching, energy-efficient edge computing

C. IoT

The projected growth in IoT devices is huge, resulting in huge increase in network traffic as well as in emissions from production. To mitigate this , we need to increase the device lifespan; reducing energy consumption helps primarily with this; and use edge computing to reduce network traffic.

D. Mobile devices

The projected growth in mobile devices is still very large, and current lifespans much too short. We mainly need longer-term software support, so better software engineering practices, in particular relating to security. Apps should be designed to minimise full-system energy consumption; user interfaces should nudge users towards energy efficient behaviour.

VI. Research directions

We can also consider the research directions that would contribute to this vision. Again, this is not a comprehensive list.

As a necessary condition, our systems must be as energy-efficient and long-lived as possible. This requires advances in many areas:

- Operating systems: energy-aware resource allocation and scheduling
- Networking: Energy consumption as QoS criterion
- Software engineering: better processes and sustainable practices will play a key role in extending lifetimes of systems.
- Data centre/cloud: energy efficient heterogeneous systems, resource allocation, scheduling.

Sustainable systems need to be data-driven. Large systems produce huge amounts of system data. Making sense of this data is crucial for whole-system energy optimisation.

HCI has a key role to play in achieving low carbon computing

- Make users aware of energy/carbon costs of their actions
- Nudge user behaviour towards more sustainable practice
- Human-computer interfaces influence both energy consumption and useful life of devices

Finally, formal methods also have a key role to play in many aspects of low-carbon computing:

- Programming languages need to become full-system energy aware.
- Algorithms need to focus on minimising overall minimal energy consumption
- Compilers need to compile for overall minimal energy consumption: not just CPU, also RAM, DMA, I/O wait etc; compilers need to be much better at selecting the right algorithms for given architectures, as algorithms strongly influence performance.

VII. Conclusion

My position on the direction computing should take is clear: we should reduce emissions from computing by using less energy and less materials. From the computing science perspective, this provides us with the challenges of radical optimisation of energy efficiency and useful life of our computing resources. And this change needs to happen within the next two decades.

REFERENCES

[1] Rebooting the IT revolution: a call to action, Semiconductor Industry Association/Semiconductor Research Corporation, Sept 2015

[2] Full Report for the Decadal Plan for Semiconductors, Semiconductor Industry Association/Semiconductor Research Corporation, Jan 2021

[3] Assessing ICT global emissions footprint: Trends to 2040 & recommendations, Lotfi Belkhir, Ahmed Elmeligi, Journal of Cleaner Production 177 (2018) 448–463

[4] Our House is on Fire: The Climate Emergency and Computing's Responsibility, B. Knowles, K. Widdicks, G. Blair, M. Berners-Lee, A. Friday, Communications of the ACM, Vol 6(6), June (2022) 38-–40

[5] Sources of Greenhouse Gas Emissions, United States Environmental Protection Agency, Last updated on April 14, 2021

[6] Emissions Gap Report 2020, UN Environment Programme, December 2020

[7] The link between product service lifetime and GHG emissions: A comparative study for different consumer products, Simon Glöser-Chahoud, Matthias Pfaff, Frank Schultmann, Journal of Industrial Ecology, 25 (2), pp 465-478, March 2021

[8] Cool products don't cost the Earth – Report, European Environmental Bureau, September 2019

[9] The life cycle assessment of a UK data centre, Beth Whitehead, Deborah Andrews, Amip Shah, Graeme Maidment, Building and Environment 93 (2015) 395–405, January 2015

[10] Statista, retrieved June 2021

[11] Global Energy & CO_2 Status Report, International Energy Agency, March 2019

[12] Redefining scope: the true environmental impact of smartphones?, James Suckling, Jacquetta Lee, The International Journal of Life Cycle Assessment volume 20, pages 1181–1196 (2015)

[13] Server Rack Power Consumption Calculator, Rack Solutions, Inc., July 2019

[14] Analysis of energy consumption and potential energy savings of an institutional building in Malaysia, Siti Birkha Mohd Ali, M.Hasanuzzaman, N.A.Rahim, M.A.A.Mamun, U.H.Obaidellah, Alexandria Engineering Journal, Volume 60, Issue 1, February 2021, Pages 805-820

[15] On Global Electricity Usage of Communication Technology: Trends to 2030, Anders S. G. Andrae, Tomas Edler, Challenges 2015, 6(1), 117-157

[16] BP Energy Outlook: 2020 Edition, BP plc

[17] 2015 International Technology Roadmap for Semiconductors (ITRS), Semiconductor Industry Association, June 2015

[18] Net Zero by 2050 – A Roadmap for the Global Energy Sector, International Energy Agency, October 2021