Rethinking Computing Systems in the Era of Climate Crisis: A Call for a Sustainable Computing Continuum

Ella Peltonen, *University of Oulu, 90014, Oulu, Finland*
Suzan Bayhan, *University of Twente, Enschede, The Netherlands*
David Bermbach, *TU Berlin, Berlin, Germany*
Sebastian Buschjäger, *Lamarr Institute for ML and AI, Dortmund, Germany*
Victoria Degeler, *University of Amsterdam, Amsterdam, The Netherlands*
Aaron Yi Ding, *Technische Universiteit Delft, 2600GA, Delft, The Netherlands*
Özlem Durmaz Incel, *University of Twente, Enschede, The Netherlands*
Dewant Katare, *TU Delft, 2628 BX, Delft, The Netherlands*
Mikkel Baun Kjærgaard, *University of Southern Denmark, Odense, Denmark*
Sam Leroux, *Ghent University - imec, 9052, Ghent, Belgium*
Toktam Mahmoodi, *King's College London, London, United Kingdom*
Zoltán Ádám Mann, *University of Halle-Wittenberg, Halle, Germany*
Nirvana Meratnia, *Eindhoven University of Technology, Eindhoven, The Netherlands*
Andy D. Pimentel, *University of Amsterdam, Amsterdam, The Netherlands*
Jan S. Rellermeyer, *Leibniz University Hannover, Hannover, Germany*
Etienne Rivière, *UCLouvain, 1348, Louvain-la-Neuve, Belgium*
Dolly Sapra, *University of Amsterdam, Amsterdam, The Netherlands*
Gürkan Solmaz, *NEC Laboratories Europe, 69124, Heidelberg, Germany*
Bram van der Waaij, *TNO, Groningen, The Netherlands*

The advancement and widespread adoption of computing technology has yielded services that could help mitigate the climate crisis. However, the retirement of obsolete equipment, the consumption of rare earth materials, and the escalating energy demands associated with massive data processing and cloud infrastructures have raised new environmental dilemmas. Existing design and development methodologies primarily focus on fulfilling functional requirements and improving performance. In this article, we argue that these methodologies must be augmented with sustainability considerations encompassing energy efficiency, material usage, longevity, and upgradability. Solutions at different

Digital Object Identifier 10.1109/MIC.2025.3566642
Date of publication 5 May 2025; date of current version 30 June 2025.

layers of the system stack, from the physical to the application layer, must be integrated. Moreover, there should be a strong focus on the transparency of sustainability metrics across the whole computing continuum. Building on fruitful discussions at the International Lorentz Workshop on Future Computing for Digital Infrastructures, we advocate novel approaches in the design, development, and operation of the computing continuum.

The climate crisis is a pressing challenge for societies worldwide. Information and communication technology (ICT), omnipresent in daily life and businesses, has long been foreseen to play an essential role in addressing it. Computing and communication technologies are pivotal for a low-carbon economy, as they facilitate digitization and optimization of otherwise carbon-heavy processes across various domains, e.g., by enabling efficient collection and dissemination of critical environmental data. This facilitates better decision-making for resource management, energy conservation, and waste reduction by processing the sensory data collected in the compute continuum for smarter operation, as shown in Figure 1. As ICT supports sustainable everyday practices such as active mobility or remote work, thereby reducing the carbon footprint of physical commuting, they are considered crucial for the United Nations Sustainable Development Goals efforts. At the same time, ICT contributes to energy and natural resource consumption by producing and operating physical infrastructure and equipment, which must be disposed of when reaching its end-of-life. ICT equipment production requires using hazardous substances or rare earth materials and is responsible for 78% of its carbon footprint. Meanwhile, ICT device usage accounts for 21%, and their distribution for 1%.

Based on estimates from the European Commission, the ICT sector accounts for 5%–9% of the electricity consumption and more than 2% of the emissions worldwide. In 2018, data centers in the EU accounted for 2.7% of the total regional electricity demand, projected to reach 3.2% by 2030 if there are no sustainability interventions. Many sectors rely on advanced ICT solutions, requiring real-time and data-intensive machine learning (ML), distributed edge intelligence, energy-hungry generative artificial intelligence (AI) and large language model (LLM) approaches, and sophisticated communication architectures and Internet of Things (IoT) networks for digitizing their processes. With the constant evolution of mobile technologies (e.g., the transition from 4G to 5G in operational networks in recent years and the research on 6G already ongoing) and the desire for orders of magnitude higher data rates, new data-intensive applications such as holographic media might emerge, resulting in even higher data processing and communication needs.

Computing in the *continuum* from the far edge to the cloud is a crucial enabler for more sustainable practices while also causing significant environmental costs. Ensuring that the resulting sustainability costs do not offset the sustainability benefits of using ICT is vital. Thus, research should explore two main questions:

1) How to use ICT to reduce carbon emissions and environmental impacts in other domains?
2) How to reduce the environmental impact of ICT, considering the entire equipment lifetime?

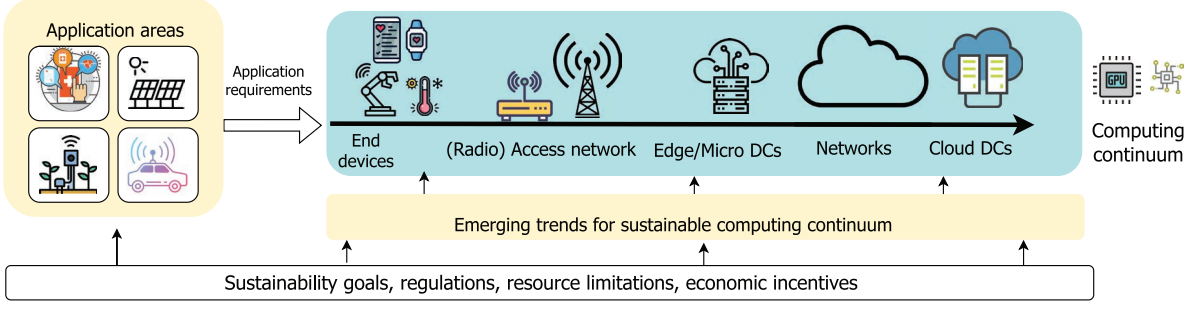

FIGURE 1. The sustainable continuum design is driven by sustainability goals in various sectors, regulations, resource limitations, and economic incentives. DC: data center.

While both questions are important, this article focuses on the second: designing and operating a sustainable compute continuum from end devices to cloud data centers, as shown in Figure 1. We define a *sustainable computing continuum* as a *collection of computing and communication technologies spanning from end devices to cloud resources that operate respecting the environmental boundaries, e.g., water, energy, and materials, and fulfill functional and nonfunctional requirements with minimum environmental cost.* Since this vision requires sustainable components and the interworking of different segments, we discuss the emerging techniques that aim at sustainability at various system layers. While our key focus is on ICT energy consumption, we stress that other environmental impacts should also be considered in the design and operation of the computing continuum.

Sustainable computing has recently become a very active area of research. There are now conferences [e.g., IEEE International Green and Sustainable Computing Conference (IGSC), International Conference on Sustainable Computing and Smart Systems (ICSCSS), International Conference on Green Computing and Engineering Technologies (ICGCET)] and journals (e.g., *IEEE Transactions on Sustainable Computing, Sustainable Computing: Informatics and Systems*) dedicated to this topic. Also, papers on sustainable computing regularly appear in major computing conferences, such as the ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), IEEE International Conference on Software Architecture (ICSA), and ACM Symposium on Cloud Computing (SoCC).

ICT's ENVIRONMENTAL COST

ICT systems span multiple tiers, including data centers, network infrastructure, and devices. The Agency for Ecological Transition in France estimates[a] that their digital environmental impact, including carbon, energy/resource, and metal/mineral consumption, is split as 79% for devices and 16% and 5% for data centers and networks. Improvements have been made in each tier. For example, the global energy consumption of data centers increased by only 6% compared with 2010, despite computing capacity increasing by 550%.[1] Similarly, the carbon footprint of computing devices typically follows a decreasing trend.[2] However, given the growth of the population as well as of the number of devices per person, such efficiency improvements cannot alone suffice to decrease ICT's environmental footprint significantly, which is essential for reaching the climate targets set for 2030, especially with increasing demand for certain data or compute-intensive applications. In this context, two use cases are worth highlighting, namely video streaming and AI.

Video streaming is known as the most energy-consuming application of the Internet, contributing to 82% of data transmission[b] in 2023. Estimates of CO_2 equivalent (CO_2e) emissions of video streaming per hour range widely[c,d] from 0.04 kg/h to 0.85 kg/h. In total, video streaming contributes to around 1% of global greenhouse gas (GHG) emissions. Estimating the resource consumption of video streaming is complex due to the variability across different tiers. Small to medium data centers consume around 50 to 200 kW, while large data centers use over 2500 kW. Edge devices, including smartphones and tablets, typically use only a few watts, while desktop PCs demand upwards of 250 W. The estimations of energy consumption of transmitting 1 GB of data vary widely, with a meta-study determining the best estimate at around 0.06 kWh/GB in 2015.[3]

AI and, in particular, ML approaches have the potential to grow as the biggest resource consumer in the future, even if they lead to innovative applications across different digital areas, as seen, for example, with the rapid development of LLMs. ML applications typically undergo three phases: data collection and initial experimentation, large-scale training on GPU or specialized accelerators, and final deployment running inference requests. Meta estimates that the energy footprint of these phases is distributed as 39%:29%:40% in their data centers[4] and that the operational CO_2e footprint for a single AI application can reach 1000 t CO_2e, excluding the embodied carbon in the hardware, and up to 1300 t CO_2e when considering it. Meta reports achieving a 20% power footprint reduction every six months through iterative AI optimizations in their data centers.

Common to these examples is that they require functionality over multiple tiers of the computing continuum. To become truly sustainable, applications must be seen from the viewpoint of all of the tiers in the ICT ecosystem, from client devices to cloud services.

[a]https://en.arcep.fr/uploads/tx_gspublication/press-kit-study-Ademe-Arcep-lot3_march2023.pdf.

[b]https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[c]https://www.carbonbrief.org/factcheck-what-is-the-carbon-footprint-of-streaming-video-on-netflix/

[d]https://theshiftproject.org/en/article/shift-project-really-overestimate-carbon-footprint-video-analysis/

REGULATORY APPROACHES AND SOCIETAL CHALLENGES

Regulatory efforts for sustainability have intensified globally due to the growing urgency. The SG5, an *environment, climate change, and circular economy* study group of the International Telecommunication Union (ITU-T),[e] the UN body for telecommunications harmonization and standardization worldwide, published several recommendations toward standardization on analyzing and improving ICT carbon emissions. L.1450 develops tools for stakeholders to calculate GHG emissions, considering energy consumption and embodied emissions related to the product lifecycle. L.1410 extends the lifecycle assessment approach standardized by ISO 14040:2006 and ISO 14044:2006. L.1470 describes guidelines for calculating and decreasing emissions for mobile and fixed network operators and data centers.

The EU has a leading role in energy efficiency regulations at the regional level. The European Telecommunications Standards Institute (ETSI) has three sustainability-related technical committees[f]: Environmental Engineering Committee (TC EE), Access, Terminals, Transmission and Multiplexing Committee (TC ATTM), and Industry Specification Group on Operational energy Efficiency for Users (OEU). They develop key performance indicators (KPIs) and methods for assessing ICT goods' lifecycle and communication networks' energy efficiency. The EU adopted several directives to reach its 2030 target of 55% reduction of its GHG emissions compared to 1990. The Ecodesign Directive[g] obligates manufacturers to ensure that products are designed to minimize their energy consumption and other negative environmental impacts. The regulation for servers and data storage products[h] defines consumption restrictions for servers' power supply units. The Energy Efficiency Directive[i] promotes guidelines for energy efficiency as a core design and operation principle and requires monitoring and disclosing data center sustainability indicators such as energy consumption, power and waste heat utilization, water usage, and use of renewable energy.

The Corporate Sustainability Reporting Directive creates awareness of the energy footprint of service provisioning and product development by enforcing information disclosure about the impact on the environment and its habitat and transparency about adopted measures to reach sustainability goals. These efforts influenced similarly targeted legislation or proposals outside of Europe. In the United States, the Securities and Exchange Commisssion (SEC) attempts to establish climate-related disclosure standards as part of the Environmental, Societal, and Government impacts in the nonfinancial reporting that publicly listed companies are obliged to perform. Like the EU directive, the proposed SEC regulation is built around the GHG Protocol.

There are also environmental compliance certifications, e.g., EU WEEE[j] to minimize e-waste and to maximize recycling and reuse, Clean Air Act[k] to phase out ozone-depleting substances, or ISO 14001 for environmental management systems. The EC published Code of Conduct documents, e.g., Data Center Energy Efficiency[l] on sustainable practices for data centers and broadband equipment, including Wi-Fi hotspots. ENERGY STAR[m] in the United States lists the recommended consumption levels for networking equipment, from data centers to home gateways, for energy-efficient equipment label eligibility.

We argue that regulations covering mainly hardware production and power efficiency metrics fall short of preventing the environmental impact of ICT in three ways. First, we need a better understanding of the impact of software on the total environmental cost and to promote best practices for environmentally sustainable software, including handling the impact of continuous updates. Second, embodied carbon already represents a significant fraction of carbon emissions for battery-powered devices such as smartphones and laptops, and it might also dominate for always-connected devices as we move toward greener energy sources.[5] Thus, repurposing/refurbishing the hardware rather than producing new equipment is preferable. Methods that achieve the desired quality levels with less hardware (e.g., infrastructure sharing) should be developed and promoted. Third, awareness of users and experts (e.g., software developers) should be enhanced for behavioral change, and sustainable practices should be adopted by making the environmental cost of technology more visible.

[e]https://www.itu.int
[f]fhttps://www.etsi.org/technologies/energy-efficiency
[g]EUDirective2009/125/EC;https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009L0125
[h]EURegulation2019/424;https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019R0424
[i]EUDirective2023/1791;https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules
[j]https://environment.ec.europa.eu/topics/waste-and-recycling/waste-electrical-and-electronic-equipment-weee_en
[k]https://www.epa.gov/ods-phaseout
[l]https://e3p.jrc.ec.europa.eu/communities/data-centers-code-conduct
[m]https://www.energystar.gov

EMERGING TRENDS AND TECHNOLOGY PERSPECTIVES FOR SUSTAINABILITY IN COMPUTING

Sustainable computing requires maintaining the ability to perform large-scale computing while avoiding resource depletion. It includes reducing the need for more hardware and limiting claims on resources as much as possible. Since renewable energy sources cannot deliver continuously, it is also about being adaptive in resource usage. As shown in Figure 2, sustainable computing trends emerge in all layers of the system stack. A sustainable computing continuum leverages opportunities across layers and considers various tradeoffs which we present next. Please note that some approaches require cross-layer design but are discussed only once for the sake of brevity.

Physical Layer

The physical layer concerns physical phenomena related to hardware, such as heat dissipation. Thus, this layer offers opportunities to directly influence the sources from which energy is drawn and what happens with the produced heat.

Moving Computation Toward Renewable Energy Resources

As renewable energy sources are significantly lower in their carbon intensity,[5] when local renewable power is insufficient, or grid congestion prevents its transfer, delay-tolerant computations can be moved in the computing continuum to areas with renewable energy or energy with lower carbon intensity.[6] Alternatively, computing systems may wait and queue processing jobs until the shortage of local renewable energy ends or switch to batteries that were charged with renewable energy in times of abundance. When there is a local surplus of electricity from renewable sources, queued jobs can be run, batteries recharged, and the capacity to run jobs from other locations sustainably can be advertised.

Reusing Heat

Data centers produce large amounts of heat and need cooling, leading to reusing the produced heat somewhere where heating is needed. Unfortunately, heat consumers such as industry and residential buildings are often not located near large data centers. Transporting heat over longer distances leads to significant losses. Therefore, it is better to deploy data centers near the heat consumers. This drives an evolution from big data centers in rural areas toward many small/micro data centers in or near cities. Thus, heat reuse strengthens the drive toward edge computing. Many heat consumers require high temperatures (around 70–80° C). This makes modern cooling techniques, such as oil and microfluidic chip cooling, preferred over traditional, low-temperature air cooling systems, avoiding the need for electricity-hungry heat pumps to increase the temperature. However, optimizing the power usage efficiency (PUE), defined as the ratio between the total consumed energy (including cooling and other overheads) and that of computing equipment, of small data centers is challenging.

Energy Harvesting, Battery-Free Devices

To improve sustainability at the edge of the continuum, energy should be leveraged from various ambient renewable sources such as light, heat, and RF signals at the devices. Given that wireless signals are ubiquitous, repurposing them as an energy source is gaining

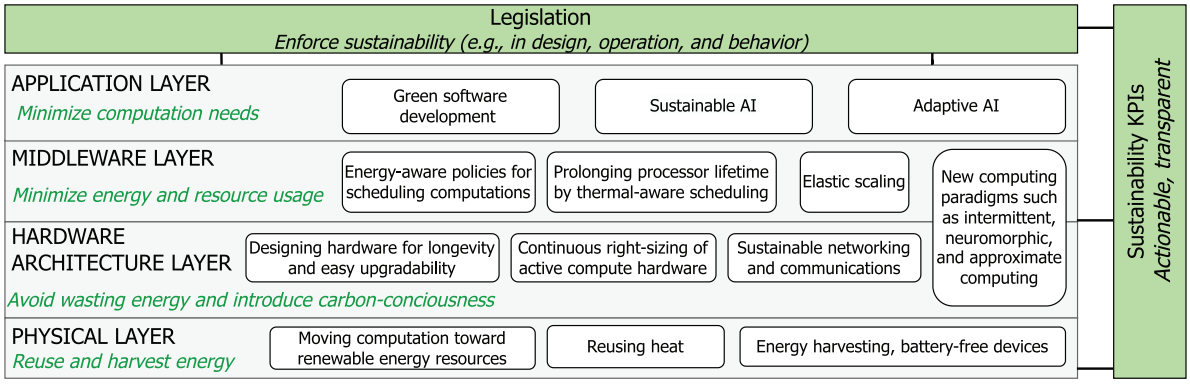

FIGURE 2. Trends toward sustainable computing emerge in all layers of the system stack.

attention.[7] This, however, requires adapting the software and protocols to account for the intermittent nature of this energy source.

Hardware Architecture Layer

Hardware design significantly impacts the sustainability of computing. For example, making it possible to deactivate unused parts of the hardware helps reduce energy consumption, just as devising new and more efficient hardware architectures.

Designing Hardware for Longevity and Easy Upgradability

As device production is a significant source of emissions and becoming more prominent with operational emissions decreasing due to efficiency improvements and increase in renewable energy use,[5] it is vital that hardware is designed for longevity and facilitates upgrading, for example, through modular architectures.

Continuous Right-Sizing of Active Compute Hardware

While edge computing provides benefits in terms of latency and lower network traffic, cloud computing supports on-demand resources and scaling, an efficient way to reduce the energy consumption of compute nodes. Depending on the hardware and application types, unused hardware could be deactivated and reactivated only when needed.[8] This is also relevant for different IoT and cyberphysical systems, e.g., collaborative robots assembling or packaging products.

Sustainable Networking and Communications

In mobile networks, infrastructure sharing can mitigate total emissions. It is typically only on passive network elements (e.g., cell towers), but active sharing can significantly decrease energy consumption and emissions. Similarly, virtualized network functions and programmable networks can improve sustainability as devices are updated with software rather than being replaced. A typical approach to decrease energy consumption is to put some network equipment in lower energy states, e.g., sleep, when the network load is low. Since the radio access network segment accounts for around 80% of the electricity consumption of a mobile network,[n] such strategies and user association schemes can be very useful. Also, traffic engineering approaches that utilize network elements with less carbon-intensive energy sources can help make communication more sustainable.

New Computing Paradigms Such as Intermittent, Neuromorphic, and Approximate Computing

Conventional computing performance will reach a plateau, and novel computing paradigms beyond the standard von Neumann computer architectures, such as quantum or optical computing, might unlock future applications. Application-specific processors such as tensor processing units and recent ML-optimized GPUs are more power-efficient than general-purpose processors and can improve computation performance per watt by factors of 2–5,[9] so we may see a more diverse landscape of specialized computing platforms. *Chiplets,*[10] smaller, specialized semiconductor components developed independently and combined to form a larger integrated circuit offer a scalable and flexible approach to meet the growing demands of various applications. *Approximate computing*[11] emphasizes trading off precision for efficiency, challenges the reliance on exact computations, and opens new avenues for energy-efficient processing. *Analog computing,*[12] leveraging continuous physical variables for computation, offers an alternative to discrete digital systems, improving the efficiency for specific computations. *Intermittent computing*[13] systems operate with sporadic energy availability, where operations might terminate unpredictably.

Middleware Layer

Middleware—including operating systems, hypervisors, and distributed software execution frameworks—is responsible for mapping applications to hardware resources, impacting computation and communication efficiency.

Energy-Aware Policies for Scheduling Computations

Flexible scheduling is essential for utilizing local energy sources. However, scheduling must capture critical computational jobs' timing constraints, e.g., a guaranteed completion time frame. Scheduling policies can include "complete by," as proposed for serverless computing,[14] or specify an execution time frame or periodicity, e.g., "once per day" or "no more than 12 hours between runs." In a geographically distributed computing continuum, scheduling can consider the type and location of energy sources, as the same energy consumption can lead to very different carbon emissions depending on where the computation is executed.[15]

[n]https://www.etsi.org/images/files/Magazine/ETSI_Enjoy_MAG_2021_N01_January.pdf

Prolonging Processor Lifetime by Thermal-Aware Scheduling

Historically, progress in microprocessor efficiency often justified replacing still-functional processors to save energy. These advancements have now slowed, highlighting the importance of extending processors' lifespan and reducing their manufacturing and disposal carbon footprint. Operational lifetime can be prolonged by improving processors' reliability. The aging of the underlying transistors often causes processor faults due to the generated heat and necessitates its replacement since they either cause unreliable results or render it unusable. Thermal-aware scheduling can prolong the lifetime of processors. Such techniques schedule tasks/threads in time and space (i.e., mapping to a particular processor core) to reduce thermal stress. If processor cores fail due to permanent faults, tasks/threads are migrated to still operational counterparts.[16]

Elastic Scaling

Instead of enabling peak performance at all times, performance can be calibrated to be good enough by scaling down applications' storage and computational needs, e.g., during nighttime.[17] The need for sustainability requires aggressive scaling and designing applications that can adapt to the scaled-down infrastructure.

Application Layer

Careful development of software can significantly improve sustainability. Some relevant techniques pertain to any application, while others are specific to certain application types with high resource demand.

Green Software Development Practices

The Green Software Foundation[o] emphasizes several guidelines of sustainable software development including carbon efficiency, energy efficiency, and hardware efficiency. First, carbon efficiency focuses on minimizing the carbon footprint of software, while energy efficiency targets energy consumption in software processes, reducing it throughout the software lifecycle. Last, carbon awareness encourages adaptive software activity based on the eco-friendliness of available electricity. In light of these principles, software developers can pay attention to resource usage and associated carbon footprint, e.g., reducing data transmission by leveraging caching or data storage by compression techniques, removing unused features, or opting for programming languages that are more resource efficient or cloud providers with lower carbon footprint.[p]

Sustainable AI

A 2024 U.S. Data Center Energy Usage Report shows a rapid increase in electricity consumption by data centers, from representing 1.9% of total annual U.S. consumption in 2018 to 4.4% in 2023, and forecasted 6.7% to 12.0% in 2028.[q] The report highlights the rapid emergence of AI hardware and GPU-accelerated servers as culprits. Breakthrough AI advancements, particularly LLMs, largely drive this increase. Stressing environmental costs and regulatory transparency is critical to managing AI's environmental impact. The focus on the environmental impact of AI has led to energy-reducing strategies like sparse models.[18] A framework for assessing ML models' energy efficiency,[19] inspired by the EU's energy label system, enables the comparison of energy consumption of different ML models, including MobileNet, EfficientNet, VGG, and DenseNet, using ML frameworks like ONNx, PyTorch, and TensorFlow. Besides sparse models, other approaches to improve the sustainability of AI include application-specific processors, increasing data center's PUE, and efficient utilization of computing locations by picking those with the cleanest energy sources.[9]

Adaptive AI Models

Recently developed adaptive AI models can modify their executable partition based on the available infrastructure[20] and use more or less resources as needed. Adaptive AI models contribute to sustainability by optimizing and adapting their performance over time, thus extending the usable life of deployed models and reducing the frequency of creating and training entirely new models from scratch. The extent of scalability of adaptive AI models to large models such as LLMs is still under investigation.

RESEARCH CHALLENGES FOR A SUSTAINABLE COMPUTING CONTINUUM

To achieve the previously described advances, we outline cross-cutting challenges in observability, measurability, awareness, adaptability, and reusability. Also, avoiding rebound effects and other unintended impacts of efficiency is crucial.

[o]https://greensoftware.foundation/

[p]https://greensoftware.foundation/articles/10-recommendations-for-green-software-development

[q]https://eta.lbl.gov/publications/2024-lbnl-data-center-energy-usage-report

Quantifiability and Observability

The lack of reliable, quantifiable measures for energy and carbon-equivalent costs of hardware and software components hinders the development of sustainability-aware systems. Guidelines, monitoring infrastructures, and benchmarks for sustainability metrics must be developed. For example, data centers have established metrics for energy usage effectiveness (e.g., PUE). Similar metrics for telecommunications devices, software, and applications are also required. Quantifying the embodied carbon cost of producing, transporting, and disposing of components is challenging, especially for software, as the footprint is often linked to infrastructure and development. The availability of such metrics also allows design guidelines for future systems to ensure the move forward is toward more sustainable ICT. Transparency requires public disclosure of metrics and energy sources. Moreover, for ICT to adapt its operation in near-real-time according to the carbon intensity of the electricity grid, timely information flow about sustainability indicators to and from the electricity grid is essential. These measures collectively contribute to a system that is well-informed about energy dynamics, capable of optimizing computation scheduling, and compliant with the regulatory frameworks. Computation hubs benefit from middleware that uses appropriate metrics and prediction models to manage computational requests and explain pricing strategies and policy specifications.

Footprint Awareness

Making users and developers aware of the costs through simple yet grounded models is vital to comprehending the environmental costs associated with their digital interactions. The concept of energy labels for software, akin to those used for household devices, is a potential concrete step in this direction. However, developing and adopting such labels also entails political and sociological aspects. Striking a balance between simplicity and accuracy in these labels is critical to ensuring widespread understanding and acceptance.

Adapting Applications and Programming Models

Applications should be adapted to more intermittent and specialized hardware, integrating efficiency and supporting mixed-mode functioning. Current methodologies scale the available infrastructure, but the applications must also be adaptive and dynamic. Such continuum-native applications require handling dynamic environments, managing data complexity, addressing dynamic resource constraints on the underlying hardware, balancing specialization and complexity, and ensuring ease of use to gain adoption among programmers. This requires a multidisciplinary approach and expertise in various software development domains.

Reusability and Repurposability

Fostering a culture of sustainability in both hardware and software development requires a shift in current practices. Ensuring compatibility and interoperability between different generations of components in the age of rapid advancements will always be challenging. Legacy software systems may present challenges regarding reusability and adaptability to new technologies and can introduce vulnerabilities if data privacy and security considerations are not adequately addressed. Designing hardware with high repurposability involves creating devices adapted for different uses or upgraded with new functionalities, extending their relevance.

Avoiding the Rebound Effect

The rebound effect refers to the unintended consequence where efficiency improvements lead to increased deployment, more demanding applications, and higher usage, offsetting the intended gains in sustainability. In this sense, quantifiability and observability metrics, and the derived awareness solutions, must consider volume, for example, by promoting global efficiency measures under emission constraints.

CONCLUSION

We presented multiple challenges in reducing ICT's environmental impact and how a sustainable computing continuum could help. The solutions span beyond individual continuum tiers and system layers and are cross-cutting, requiring a holistic approach. To raise awareness of developers, infrastructure managers, and end users, we need standardized reporting and metrics of environmental footprints at all levels. These should include software and hardware, and the whole lifecycle of equipment and applications should be considered. We need to expand regulations beyond the existing efforts and promote good practices such as hardware reuse, tolerance to degraded or intermittent resource availability, or software development tailored to limited resources while involving all stakeholders and policymakers, not only developers or hardware designers. Continuous efforts are necessary to promote these with practitioners on all seniority levels. Educational resources play a crucial role and should be promoted actively by the IEEE, ACM, and other scientific collectives.

ACKNOWLEDGMENTS

The authors acknowledge the fruitful discussions with the participants of the Lorentz workshop on Future Computing for Digital Infrastructures, 2023.

REFERENCES

1. E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020, doi: 10.1126/science.aba3758.
2. L. Eeckhout, "The sustainability gap for computing: Quo vadis?" *Commun. ACM*, vol. 68, no. 3, pp. 70–79, 2025, doi: 10.1145/3699595.
3. J. Aslan, K. Mayers, J. G. Koomey, and C. France, "Electricity intensity of internet data transmission: Untangling the estimates," *J. Ind. Ecol.*, vol. 22, no. 4, pp. 785–798, 2018, doi: 10.1111/jiec.12630. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jiec.12630
4. C.-J. Wu et al., "Sustainable AI: Environmental implications, challenges and opportunities," *Proc. Mach. Learn. Syst.*, vol. 4, pp. 795–813, 2022.
5. U. Gupta et al., "Chasing carbon: The elusive environmental footprint of computing," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archit. (HPCA)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 854–867, doi: 10.1109/HPCA51647.2021.00076.
6. L. Wu et al., "CarbonEdge: Leveraging mesoscale spatial carbon-intensity variations for low carbon edge computing," 2025, *arXiv:2502.14076*.
7. M. Kodali, L. N. Nguyen, and S. Sigg, "Towards battery-less RF sensing," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops Other Affiliated Events (PerCom Workshops)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 352–355.
8. G. Halácsy and Z. Á. Mann, "Optimal energy-efficient placement of virtual machines with divisible sizes," *Inf. Process. Lett.*, vol. 138, pp. 51–56, Oct. 2018, doi: 10.1016/j.ipl.2018.06.003.
9. D. Patterson et al., "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, vol. 55, no. 7, pp. 18–28, Jul. 2022, doi: 10.1109/MC.2022.3148714.
10. X. Ma, Y. Wang, Y. Wang, X. Cai, and Y. Han, "Survey on chiplets: interface, interconnect and integration methodology," *CCF Trans. High Perform. Comput.*, vol. 4, no. 1, pp. 43–52, 2022, doi: 10.1007/s42514-022-00093-0.
11. J. Henkel et al., "Approximate computing and the efficient machine learning expedition," in *Proc. 41st IEEE/ACM Int. Conf. Computer-Aided Des.*, 2022, pp. 1–9.
12. A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "AI and ML accelerator survey and trends," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 1–10, doi: 10.1109/HPEC55821.2022.9926331.
13. F. Bambusi, F. Cerizzi, Y. Lee, and L. Mottola, "The case for approximate intermittent computing," in *Proc. 21st ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 463–476, doi: 10.1109/IPSN54338.2022.00044.
14. T. Schirmer, N. Japke, S. Greten, T. Pfandzelter, and D. Bermbach, "The night shift: Understanding performance variability of cloud serverless platforms," in *Proc. 1st Workshop Serverless Syst., Appl. Methodologies*, New York, NY, USA: ACM, May 2023, pp. 27–33, doi: 10.1145/3592533.3592808.
15. E. Ahvar, S. Ahvar, Z. Á. Mann, N. Crespi, R. Glitho, and J. Garcia-Alfaro, "DECA: A dynamic energy cost and carbon emission-efficient application placement method for edge clouds," *IEEE Access*, vol. 9, pp. 70,192–70,213, 2021, doi: 10.1109/ACCESS.2021.3075973.
16. D. Sapra and A. D. Pimentel, "Exploring multi-core systems with lifetime reliability and power consumption trade-offs," in *Embedded Computer Systems: Architectures, Modeling, and Simulation*, C. Silvano, C. Pilato, and M. Reichenbach, Eds., Cham, Switzerland: Springer-Verlag, 2023, pp. 72–87.
17. N. Chaurasia, M. Kumar, R. Chaudhry, and O. P. Verma, "Comprehensive survey on energy-aware server consolidation techniques in cloud computing," *J. Supercomputing*, vol. 77, no. 10, pp. 11,682–11,737, 2021, doi: 10.1007/s11227-021-03760-1.
18. M. Gutierrez, M. Á. Moraga, F. Garcia, and C. Calero, "Green-IN machine learning at a glance," *Computer*, vol. 56, no. 6, pp. 35–43, Jun. 2023, doi: 10.1109/MC.2023.3254646.
19. R. Fischer, M. Jakobs, S. Mücke, and K. Morik, "A unified framework for assessing energy efficiency of machine learning," in *Proc. ECML Workshop Data Sci. Social Good*, 2022, pp. 39–54.
20. S. Minakova, D. Sapra, T. Stefanov, and A. D. Pimentel, "Scenario based run-time switching for adaptive CNN-based applications at the edge," *ACM Trans. Embedded Comput. Syst.*, vol. 21, no. 2, pp. 1–33, 2022, doi: 10.1145/3488718.

ELLA PELTONEN is an assistant professor (tenure track) at the University of Oulu, 90014, Oulu, Finland. Her research interests include distributed machine learning/artificial intelligence, large-scale sensor data analytics, and intelligent

Internet of Things systems. Peltonen received her Ph.D. degree in computer science from the Huniversity of Helsinki, Finland. She is a Senior Member of IEEE. Contact her at ella.peltonen@oulu.fi.

SUZAN BAYHAN is an associate professor at the University of Twente, 7522NB, Enschede, The Netherlands. Her research interests include sustainable wireless networks and computing. Bayhan received her Ph.D. degree in computer engineering from Bogazici University. She is a member of ACM. Contact her at suzan.bayhan@utwente.nl.

DAVID BERMBACH is a full professor at TU Berlin and Einstein Center Digital Future, 10587, Berlin, Germany. His research interests include scalable, often data-intensive edge-to-cloud software systems and their interdisciplinary applications, and systems benchmarking. Bermbach received his Ph.D. degree in computer science from Karlsruhe Institute of Technology, Germany. Contact him at david.bermbach@tu-berlin.de.

SEBASTIAN BUSCHJÄGER is a postdoctoral researcher and the coordinator for resource-aware ML and AI at the Lamarr Institute, 44227, Dortmund, Germany. His research interests include efficient machine learning for embedded systems, resource-constrained environments, and ensemble methods. Buschjäger received his Ph.D. degree from TU Dormund University, Germany. Contact him at sebastian.buschjaeger@tu-dortmund.de.

VICTORIA DEGELER is an assistant professor at the University of Amsterdam, 1098XH, Amsterdam, The Netherlands. Her research interests include smart environments, digital twins, pervasive systems, and context modeling and representation, with particular interest in sustainable applications, such as energy and water. Degeler received her Ph.D. degree in computer science from the University of Groningen, The Netherlands. Contact her at v.o.degeler@uva.nl.

AARON YI DING is a senior associate professor with tenure at Technische Universiteit Delft, 2600GA, Delft, The Netherlands. His research interests include interconnected artificial intelligence systems. Ding received his Ph.D. degree in computer science from the Universtiy of Helsinki, Finland. He is a member of IEEE and ACM. Contact him at aaron.ding@tudelft.nl.

ÖZLEM DURMAZ INCEL is an associate professor at the University of Twente, 7522NB, Enschede, The Netherlands. Her research interests include applied machine learning for building context-aware, resource-efficient, intelligent systems, Internet of Things, wearable computing, and edge artificial intelligence. Incel received her Ph.D. degree from the University of Twente, The Netherlands. Contact her at ozlem.durmaz@utwente.nl.

DEWANT KATARE is a Ph.D. candidate in the Department of Engineering Systems and Services, Delft University of Technology, 2628 BX, Delft, The Netherlands. His research interests include energy-efficient computing and edge artificial intelligence. Katare received his M.Sc. degree in electrical and computer engineering from Purdue University. He is a Graduate Student Member of IEEE. Contact him at d.katare@tudelft.nl.

MIKKEL BAUN KJÆRGAARD is a professor at the University of Southern Denmark, DK-5230 Odense, Denmark. His research interests include software engineering and energy-efficient systems. Kjærgaard received his Ph.D. degree from Aarhus University. Contact him at mbkj@mmmi.sdu.dk.

SAM LEROUX is an assistant professor at the IDLab, an IMEC research group at Ghent University, 9052, Ghent, Belgium. His research interests include efficient neural network architectures and edge computing. Leroux received his Ph.D. degree from Ghent University. Contact him at sam.leroux@ugent.be.

TOKTAM MAHMOODI is a professor of communication engineering and head of the Center for Telecommunications Research, Department of Engineering, King's College London, London, WC2R 2LS, U.K. Her research interests include network intelligence and mission-critical networking, impacting health care, automotive, smart cities, and emergency services. Mahmoodi received her Ph.D. degree in telecommunications from King's College London. Contact her at toktam.mahmoodi@kcl.ac.uk.

ZOLTÁN ÁDÁM MANN is a professor at the University of Halle-Wittenberg, 06120, Halle, Germany. His research interests include security, privacy, and distributed systems. Mann received his Ph.D. degree in computer science from Budapest University of Technology and Economics. Contact him at zoltan.mann@gmail.com.

NIRVANA MERATNIA is a full professor at Eindhoven University of Technology, 5600 MB, Eindhoven, The Netherlands. Her research interests include artificial intelligence (AI) for systems and systems for AI, pervasive and sustainable computing, and cyberphysical systems. Meratnia received her Ph.D. degree from the University of Twente, The Netherlands. Contact her at n.meratnia@tue.nl.

ANDY D. PIMENTEL is a full professor at the University of Amsterdam, 1098XH, Amsterdam, The Netherlands. His research interests include the design, programming, and run-time management of parallel and distributed embedded computer systems. Pimentel received his Ph.D. degree in computer science from the University of Amsterdam. Contact him at a.d.pimentel@uva.nl.

JAN S. RELLERMEYER is a full professor at Leibniz University Hannover, 30167, Hannover, Germany. His research interests include scalable and dependable software systems, including cloud and edge computing. Rellermeyer received his Ph.D. degree from ETH Zurich, Switzerland. Contact him at rellermeyer@vss.uni-hannover.de.

ETIENNE RIVIÈRE is a professor at UCLouvain, 1348, Louvain-la-Neuve, Belgium. His research interests include distributed systems, cloud and edge computing, middleware, and security and privacy. Rivière received his Ph.D. degree in computer science from the University of Rennes and Inria, France. Contact him at etienne.riviere@uclouvain.be.

DOLLY SAPRA is an assistant professor at the University of Amsterdam, 1098XH, Amsterdam, The Netherlands. Her research interests include the efficient deployment of artificial intelligence on edge devices, with emphasis on adaptive neural architectures, privacy preservation, and optimal resource utilization. Sapra received her Ph.D. from the University of Amsterdam. Contact her at d.sapra@uva.nl.

GÜRKAN SOLMAZ is a senior researcher at NEC Laboratories Europe, 69124, Heidelberg, Germany. His research interests include machine learning and cloud/edge computing. Solmaz received his Ph.D. in computer science from the University of Central Florida. Contact him at gurkan.solmaz@neclab.eu.

BRAM VAN DER WAAIJ is a senior researcher at TNO, 9747AA, Groningen, The Netherlands. His research interests include sustainable computing, energy-flexible computing, and cloud federation. Van der Waaij received his Master's degree in computer science from the University of Twente, The Netherlands. Contact him at bram.vanderwaaij@tno.nl.