Special Session: Challenges and Opportunities for Sustainable Multi-Scale Computing Systems

Xavier Ouvrard
EcoCloud Center, EPFL
Lausanne, Switzerland
xavier.ouvrard@epfl.ch

Miguel Peón-Quirós
EcoCloud Center, EPFL
Lausanne, Switzerland
miguel.peon@epfl.ch

David Atienza
EcoCloud Center, EPFL
Lausanne, Switzerland
david.atienza@epfl.ch

ABSTRACT

Multi-Scale computing systems aim at bringing the computing as close as possible to the data sources, to optimize both computation and networking. These systems are composed of at least three computing layers: the terminal layer, the edge layer, and the cloud layer. Enhancing the sustainability of the whole system requires a transversal approach to solving challenges such as energy efficiency or resource consumption. The offload of computing tasks between the data center, edge, and terminal, taking into account all parameters, is then the kingpin to enhance sustainability, through transversal concerns among the different scales, including parameters related to the three sustainability pillars: environment, society, and economy.

KEYWORDS

Digital Technologies, Climate Neutrality Policies, Sustainability, Green Transition

1 INTRODUCTION

Information technology (IT) has emerged as a pillar for the digital economy, enabling national security and international competitiveness. At its core is cloud computing, which centralizes global IT services with ubiquitous protected access at a minimal cost. Cloud computing is an enabler for sustainable technologies in all sectors of the digital economy. However, the growth of IT services requires sustainable IT building blocks. The infrastructure underlying the cloud is the data center, a collection of low-cost volume servers hosting data with commodity hardware and software. However, traditional data centers did not include sustainability in IT as a core requirement. Moreover, artificial intelligence (AI) and Internet of Things (IoT) exacerbate the growth rate and energy consumption of IT.

IT has the potential to evolve from being part of the problem to becoming an enabler of improved sustainability in all areas of the digital economy. Pop et al. [16] emphasize how key digital technologies represented by edge computing, artificial intelligence, architectures, design processes, methods and tools, connectivity, and smart components and systems can help to achieve greenhouse gases (GHG) reduction. The authors give various examples in different domains from energy generation, delivery and consumption to sustainable transportation and agriculture, as well as in circular economy. However, future post-Moore platforms require to be developed holistically with sustainability as a core expectation. Therefore, new co-design approaches covering multiple abstraction levels—algorithms, software, and hardware—and going beyond the clustering of commodity components—as it is done in current cloud servers—are essential to improve data center density while keeping energy use constant. One key trend to improving sustainability is the design of multi-scale computing systems (MSCS) that are distributed and can dynamically adapt to process data closer to where IoT sensors acquire them. This new architecture will need to combine the latest edge infrastructure concepts to complement today's cloud computing to drive sustainability in IT services.

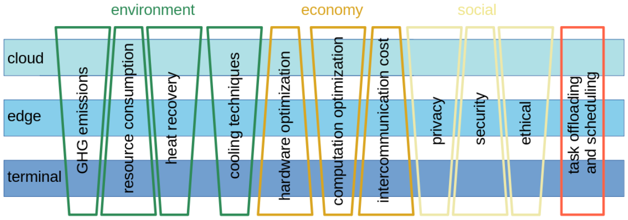

Figure 1: Transversal concerns between the three sustainability pillars and multi-scale computing systems.

MSCS aim to optimize the overall sustainability by moving the computing barycenter to the network edge to reduce the resources required for computation, to reduce the network burden, and the amount of centrally stored data by computing closer to data sources. Three layers are considered in MSCS: data acquisition and terminal computing, edge (or boundary) computing, and cloud computing [5]. As computation takes place the closest possible to data sources, possibly in devices with limited power capabilities, offloading, orchestration, and forecasting are key to optimizing the workflow between terminals, edge computing devices, and/or cloud data centers. A holistic view that considers all parameters, from power efficiency to computation power and availability, storage capacities, and possibly security and privacy of data, is needed to achieve global sustainability. Fig. 1 shows how the MSCS layers interact with the three pillars of sustainability [17], i.e. social, environmental and economic sustainability, balancing advantages and drawbacks, and their importance.

However, MSCS present their own challenges. For example, there is a potential for higher energy consumption due to the need for strict security measures (e.g. encryption of data) and an increased consumption of raw materials with the increasing number of devices. In fact, the entire life cycle of MSCS components has an impact on GHG emissions and all steps need to be taken into account.

In this paper, we present an overview of the challenges in the different MSCS layers and identify promising avenues to overcome them and ensure that IT becomes a net contributor to sustainability in the digital economy.

2 CHALLENGES AT THE TERMINAL

In IoT, the terminal layer is composed of sensors, potentially grouped in networks. The number of sensor devices is in constant augmentation, with some analyses forecasting a trillion sensors by 2035 [3]. Those sensors often digitize analog data and perform some form of in-device processing such as denoising, filtering, and event detection, before sending the resulting information to the cloud for further processing or storage. Finding the right trade-off between in-device processing, device-to-device communication, and communication with the cloud enables a reduction of network traffic while extending the device battery lifetime. Furthermore, future research will have to address the division between on-device and cloud computing to avoid spikes in energy consumption in data centers when there is insufficient renewable energy available, to avoid net GHG emissions. Security and privacy in sensor networks constitute another key challenge, as both short- and long-range communications have to be secured against a wide range of possible attacks.

Energy harvesting in IoT nodes can also play a relevant role in reducing GHG emissions, as well as the use of new nanomaterials such as metal nitrides and carbides (MXenes), borophene, and advanced 2D materials [6]. However, the carbon footprint of sensor networks due to the heterogeneity of materials requires careful evaluation. Pirson and Bol [15] propose a methodology to decompose into functional blocks—actuator, casing, connectivity, memory, electromechanical components, power supply, processing, security, transport, UI—the evaluation of the carbon footprint of different types of IoT devices. Similar methodologies could also be useful for the other MSCS layers.

3 CHALLENGES AT THE EDGE

In MSCS, layer interactions play an important role since communication is a potential bottleneck. At the extreme, sensors in large physics and astrophysics experiments, such as the LHC high luminosity project at CERN[1] or the Square Kilometer Array[2] produce hundreds of petabytes per year. The orchestration of all elements that make up the system is the kingpin for achieving sustainability. Forecasting future demands based on past ones enables a task scheduling that considers the task placement and partitioning (i.e., computing at the sensor, edge, or cloud), and the availability of energy and compute resources at each level.

Strategies to distribute tasks between different layers are an active field of research [11] that must include considerations of environmental and power to achieve sustainability. However, these strategies need access to all the parameters involved. Therefore, FAIR (Findable, Accessible, Inter-operable and Reusable) principles [22] have to be followed to ensure that relevant data is accessible in real time, especially with regard to energy consumption and GHG emissions, not only for data centers and network infrastructure suppliers, but also for device-based manufacturing.

Security and privacy are often linked, and the developed measures come with energy costs. Active research—for instance, in medical applications [8]—is being done on securing edge computing devices that can be vulnerable to attacks, including authentication and authorization attacks, and malware injection both on the device and cloud sides [23]. Therefore, it is crucial to consider these aspects from the beginning to achieve full sustainability.

4 CHALLENGES IN THE CLOUD

In MSCS, the upper layer of computation resides in the cloud. Unfortunately, the cloud data centers (DCs) are built using commodity servers as basic components. Without the exponential growth in server density enabled by Moore's law and Dennard's scaling during previous decades, the path of least resistance to increase DC capacity is growth in size and numbers. Since commodity servers were never built with sustainability as a design goal, DCs (and thus cloud computing) risk becoming even less sustainable.

To improve its energy efficiency, the next generation DC infrastructure requires innovation in integrated and co-designed technologies for electricity generation, storage, and management in order to maximize the use of renewable energies as a terminal in the electricity grid. It also requires cooling infrastructure with co-designed technologies that both maximize cooling efficiency and enhance synergy between infrastructure cooling requirements and IT load management, while ensuring that these technologies are best integrated to efficiently recycle heat.

The efficiency of a DC is commonly measured through its power usage efficiency (PUE), i.e., the ratio between the overall energy consumed by the complete DC versus the energy consumed by the IT equipment itself. PUE is nowadays complemented by carbon usage effectiveness (CUE) [2], which is obtained by multiplying PUE by carbon emission factor (CEF), which depends on the location of DC in the world and can have a variability of a factor of 10.[3] Offloading computation to a DC located in a low CEF country helps to improve sustainability. Dedicated tools [20] could help in this evaluation, but require sufficient information on the physical location on which the computation is performed.

The design and location of the DCs are crucial, not only for CEF and PUE: multiple parameters to evaluate the sustainability of buildings must be taken into account [4]. Overall indicators include among others: energy and water usage, indoor environment quality (IEQ), materials, waste and pollution management. Proposals of green labels such as the SDEA label developed by the Swiss Datacenter Efficiency Association are also part of a holistic approach to sustainability in DC.

DC energy consumption is affected, in addition to how energy was produced, by multiple factors such as the intrinsic IT power consumption of the equipment—doing more with less energy—and the cooling efficiency—for example, water has a heat conductivity that is 4000 times that of air, which enhances heat exchange. Research on IT equipment power consumption is a well established research line, with several labels such as the US Energy Star program[4] or the European energy label[5] introduced long ago. These efforts, together with the rise of battery-powered mobile computing and efforts to prevent dark silicon [9], have favored intense progress on the topic of IT equipment energy consumption. The introduction of dedicated architectures to perform specific computation tasks, from ASICs to GPUs and FPGAs, can have a significant impact on overall efficiency. This potential has been particularly well studied for deep learning applications [7, 18].

[1] https://home.cern/science/accelerators/high-luminosity-lhc

[2] https://www.skao.int

[3] https://ourworldindata.org/grapher/carbon-intensity-electricity

[4] USA Energy Start program

[5] European Commission energy label

Cooling is energy-consuming in itself and therefore affects DC's PUE. Historically, the cooling vector was air, using large AC in data centers. To provide more efficient heat exchange, rear-door passive liquid/air exchangers were introduced, which are nowadays a mature technology. Nevertheless, there are still ample opportunities for research on cooling efficiency. Two proposals under research are microcooling, which takes place at server hot spots and starts to be developed by different companies—one of them, Corintis,[6] which is an Ecole Polytechnique Fédérale de Lausanne (EPFL) startup—and full immersion cooling, which implies the use of some dielectric liquids [14].

One way of reducing PUE is to reuse waste heat to warm buildings during cold times, or, less seasonally, to produce domestic hot water. Waste heat can also be used in absorption refrigeration to cool buildings or in energy-intensive industries that require heat in their processes [19, 21]. Furthermore, waste heat can also be used to produce electricity through thermodynamic cycles such as organic Rankine cycle (ORC)—although these options are often limited to low yields due to the low waste heat quality—i.e., its low temperature—and high investment costs [12]. Heat pumps are often an efficient way to recycle waste energy that can be coupled with solar panels and some form of heat storage. Finally, waste heat can also be used during preheating phases, such as in power plants. With the increasing cost of energy, the reuse of waste heat is an avenue for new innovative research.

Sustainability in DCs also requires work on standardization. For example, defining open standards for network transceivers, both optical and direct-attached copper cables, which currently require specific devices for each supplier brand. This standardization would improve compatibility between equipment from different vendors and prevent waste in incompatible cables and transceivers. More generally, a systematic approach to the lifespan of IT equipment has to be considered in order to make a trade-off between the possible performance gains of renewing equipment versus the initial GHG production cost. Different estimations place these costs between 15 % to 40 % for servers, and up to 80 % in the case of user devices, of the GHG emitted during the entire life cycle of the equipment. A simple workaround is to systematically extend the life span of equipment. Working on GHG amortization schemes is also a hot research topic.

Resource optimization is also key for sustainability, and constitutes a challenging research avenue. Load balancing leverages resource availability to increase occupation and diminish the final exploitation cost, while enabling a reduction in both energy consumption and GHG emissions [1, 13]. Load balancing algorithms are hot research topics, where exact strategies are often NP-hard and require heuristics to solve them; multiple parameters such as server load, network traffic, migration overhead, and reliability have to be taken into account in the decision-making process.

Computation optimization is another pathway to sustainability, as choosing the right languages and libraries, and optimizing the code for energy efficiency often leads to significant gains. Another important factor is the development of efficient libraries to execute specific kernels on dedicated GPUs or FPGAs in a more efficient way than on general-purpose CPUs. The development of tools to evaluate computing GHG emissions and their accuracy is also an important requirement [10].

5 THE CASE FOR NEW RESEARCH FACILITIES IN MSCS

Inter-disciplinary approaches are essential to convert IT into a net contributor of sustainability in the digital economy. EPFL's EcoCloud[7] is an inter-disciplinary research center created in 2011 with the goal of sponsoring research on IT sustainability in Switzerland and pioneering an industrial/academic alliance. EcoCloud brings together 29 EPFL laboratories and 12 industrial affiliates to develop new sustainable smart cloud infrastructures through resilient, efficient, secure, and trustworthy data platforms and technologies. The researchers affiliated with EcoCloud cover topics from algorithms and information theory down to infrastructure for cooling and electricity distribution, generation, storage and recovery.

Following the importance of counting with advanced facilities for research on IT and DC sustainability, EcoCloud has started a project to build an experimental facility at the new EPFL DC ("CCT-DC2020"), which will provide the required infrastructure to undertake these multidisciplinary projects. EPFL's new DC has been designed following the most advanced considerations in energy efficiency. Servers are cooled using water running through exchangers located in rack rear doors using passive air flow produced by server and switch fans—additional optional external door fans are always possible when equipment fans are not sufficiently scaled in high density setups. The heat extracted by this water is recovered in ammonia-based heat exchangers and injected into the heating circuit of the campus. In fact, the DC and the heating central of the campus are built one on top of the other in the same building for efficient heat recovery. The whole building is covered with photovoltaic panels (PVs) that generate renewable energy for the building and DC operation. More importantly, the DC counts with a large surface devoted entirely to experimentation. That area counts with the same infrastructure as the main DC room, including liquid cooling with heat recovery and access to a part of the PVs to simulate real-world working conditions with access to renewable energy production.

This type of new research facilities integrated directly in production-level DCs will be indispensable to support experiments that integrate energy generation through PVs and storage in a dedicated Uninterruptible Power Supply (UPS), cloud-level server infrastructure, liquid-cooling, network topologies, algorithmic-level energy optimization, etc.

These new generation of experimental facilities will enhance experimentation in setups similar to those of real-life DCs. Multidisciplinary experiments are foreseen, with a large spectrum, including without being exhaustive: novel microfluidic cooling mechanisms able to efficiently extract heat at several kilowatts per square centimeters magnitude, disaggregated memory architectures with challenges related to microsecond Input/Output (I/O) communication, decentralized machine learning systems computation and communication cost reduction requiring decentralized large-scale network simulations with thousands of nodes, power/performance characteristics exploration of methods for automatic bug-finding

[6] https://www.corintis.com/

[7] https://ecocloud.epfl.ch

on different Central Processing Unit (CPU) and accelerator architectures, DC power supply architectures with direct power injection from multiple sources—PVs, energy recovery systems and UPS among others—, urban districts heating systems integrating DCs high-temperature microfluidic cooling heat waste reuse.

6 CONCLUSION

IT has the potential to enable sustainability in multiple areas of the digital economy. However, to become part of the solution, rather than part of the problem, new algorithm/software/hardware co-design approaches far beyond the clustering of commodity components in current cloud servers are required to improve data center density while keeping energy use constant in the absence of Moore's law and Dennard's scaling.

In this paper, we have revised the challenges that sustainable multi-scale computing requires at the different layers. We have also analyzed the implied optimization of transversal topics where offloading and task scheduling, effective resource usage, power consumption efficiency, and waste heat recovery are key challenges. Moreover, we have highlighted that inter-disciplinary approaches are essential to solving these problems. These new interdisciplinary approaches need to be able to cover research projects that span from algorithms and information theory down to infrastructure or cooling, electricity distribution, generation, usage, and recovery. Therefore, new interdisciplinary research centers and validation infrastructures, such as EcoCloud's new experimental facility—integrated inside new EPFL data center and central heating production centre—are necessary for the validation of sustainable multi-scale IT systems.

REFERENCES

[1] Mohammed Ala'Anzy and Mohamed Othman. 2019. Load balancing and server consolidation in cloud computing environments: a meta-study. *IEEE Access* 7 (2019), 141868–141887.

[2] Dan Azevedo, Michael Patterson, Jack Pouchet, and Roger Tipley. 2010. Carbon usage effectiveness (CUE): A green grid data center sustainability metric. *The green grid* 32 (2010).

[3] Janusz Bryzek. 2013. Roadmap for the trillion sensor universe. *Berkeley, CA, April* 2 (2013).

[4] Senhong Cai and Zhonghua Gou. 2023. A comprehensive analysis of green building rating systems for data centers. *Energy and Buildings* 284 (2023), 112874.

[5] Keyan Cao, Yefan Liu, Gongjie Meng, and Qimeng Sun. 2020. An overview on edge computing research. *IEEE Access* 8 (2020), 85714–85728.

[6] Vishal Chaudhary, Ajeet Kaushik, Hidemitsu Furukawa, and Ajit Khosla. 2022. Towards 5th generation AI and IoT driven sustainable intelligent sensors based on 2D MXenes and borophene. *ECS Sensors Plus* 1, 1 (2022), 013601.

[7] Yunji Chen, Tianshi Chen, Zhiwei Xu, Ninghui Sun, and Olivier Temam. 2016. DianNao family: energy-efficient hardware accelerators for machine learning. *Commun. ACM* 59, 11 (Oct. 2016), 105–112.

[8] Bakkiam David Deebak, Fida Hussain Memon, Kapal Dev, Sunder Ali Khowaja, and Nawab Muhammad Faseeh Qureshi. 2022. AI-enabled privacy-preservation phrase with multi-keyword ranked searching for sustainable edge-cloud networks in the era of industrial IoT. *Ad Hoc Networks* 125 (2022), 102740.

[9] Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, and Doug Burger. 2011. Dark Silicon and the End of Multicore Scaling. In *Proc. of ISCA*. ACM, 365–376. https://doi.org/10.1145/2000064.2000108

[10] Mathilde Jay, Vladimir Ostapenco, Laurent Lefèvre, Denis Trystram, Anne-Cécile Orgerie, and Benjamin Fichel. 2023. An experimental comparison of software-based power meters: focus on CPU and GPU. In *IEEE/ACM CCGrid*. 1–13.

[11] Hai Lin, Sherali Zeadally, Zhihong Chen, Houda Labiod, and Lusheng Wang. 2020. A survey on computation offloading modeling for edge computing. *Journal of Network and Computer Applications* 169 (2020), 102781.

[12] S Nadaf and P Gangavati. 2014. A review on waste heat recovery and utilization from diesel engines. *Int. J. Adv. Eng. Technol* 31 (2014), 39–45.

[13] Ali Pahlevan, Marina Zapater, Ayse K. Coskun, and David Atienza. 2021. ECOGreen: Electricity Cost Optimization for Green Datacenters in Emerging Power Markets. *IEEE Transactions on Sustainable Computing* 6, 2 (2021), 289–305. https://doi.org/10.1109/TSUSC.2020.2983571

[14] Nugroho Agung Pambudi, Alfan Sarifudin, Ridho Alfan Firdaus, Desita Kamila Ulfa, Indra Mamad Gandidi, and Rahmat Romadhon. 2022. The immersion cooling technology: Current and future development in energy saving. *Alexandria Engineering Journal* 61, 12 (2022), 9509–9527.

[15] Thibault Pirson and David Bol. 2021. Assessing the embodied carbon footprint of IoT edge devices with a bottom-up life-cycle approach. *Journal of Cleaner Production* 322 (2021), 128966.

[16] Paul Pop, Christian Graulund, Sonia Yeh, and Martin Törngren. 2023. Digital Technologies for Sustainability: Research Challenges and Opportunities. In *Submitted to CODES+ISSS*.

[17] Ben Purvis, Yong Mao, and Darren Robinson. 2019. Three pillars of sustainability: in search of conceptual origins. *Sustainability science* 14 (2019), 681–695.

[18] Aidin Shiri, Arnab Neelim Mazumder, Bharat Prakash, Nitheesh Kumar Manjunath, Houman Homayoun, Avesta Sasan, Nicholas R Waytowich, and Tinoosh Mohsenin. 2020. Energy-efficient hardware for language guided reinforcement learning. In *GLVLSI*. ACM, 131–136. https://doi.org/10.1145/3386263.3407652

[19] Mirko Z Stijepovic and Patrick Linke. 2011. Optimal waste heat recovery and reuse in industrial zones. *Energy* 36, 7 (2011), 4019–4031.

[20] Tristan Trébaol. 2020. *CUMULATOR—a tool to quantify and report the carbon footprint of machine learning computations and communication in academia and healthcare.* Technical Report.

[21] Mikko Wahlroos, Matti Pärssinen, Samuli Rinne, Sanna Syri, and Jukka Manner. 2018. Future views on waste heat utilization–Case of data centers in Northern Europe. *Renewable and Sustainable Energy Reviews* 82 (2018), 1749–1764.

[22] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[23] Yinhao Xiao, Yizhen Jia, Chunchi Liu, Xiuzhen Cheng, Jiguo Yu, and Weifeng Lv. 2019. Edge computing security: State of the art and challenges. *Proc. IEEE* 107, 8 (2019), 1608–1631.