Renewable and Sustainable Energy Reviews 189 (2024) 114019

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: www.elsevier.com/locate/rser

Review article

A review on the decarbonization of high-performance computing centers

C.A. Silva [a,b,*], R. Vilaça [c], A. Pereira [c], R.J. Bessa [a]

[a] *INESC TEC, Centre for Power and Energy Systems, Porto, 4200-465, Portugal*
[b] *FEUP, Faculty of Engineering of University of Porto, Porto, 4200-465, Portugal*
[c] *MACC/HASLab University of Minho & INESC TEC, Department of Informatics, Gualtar Campus, Braga, 4710-070, Portugal*

ARTICLE INFO

Keywords:
High-performance computing
Sustainable computing
Energy efficiency
Distributed energy resources
Decarbonization
Resource management

ABSTRACT

High-performance computing relies on performance-oriented infrastructures with access to powerful computing resources to complete tasks that contribute to solve complex problems in society. The intensive use of resources and the increase in service demand due to emerging fields of science, combined with the exascale paradigm, climate change concerns, and rising energy costs, ultimately means that the decarbonization of these centers is key to improve their environmental and financial performance. Therefore, a review on the main opportunities and challenges for the decarbonization of high-performance computing centers is essential to help decision-makers, operators and users contribute to a more sustainable computing ecosystem. It was found that state-of-the-art supercomputers are growing in computing power, but are combining different measures to meet sustainability concerns, namely going beyond energy efficiency measures and evolving simultaneously in terms of energy and information technology infrastructure. It was also shown that policy and multiple entities are now targeting specifically HPC, and that identifying synergies with the energy sector can reveal new revenue streams, but also enable a smoother integration of these centers in energy systems. Computing-intensive users can continue to pursue their scientific research, but participating more actively in the decarbonization process, in cooperation with computing service providers. Overall, many opportunities, but also challenges, were identified, to decrease carbon emissions in a sector mostly concerned with improving hardware performance.

1. Introduction

High-performance computing (HPC) can be defined as the “field of endeavor that relates to all facets of technology, methodology and application associated with achieving the greatest computing capability possible at any point in time and technology” [1]. HPC relies on performance-oriented infrastructures with access to powerful computing resources to complete tasks that contribute to solving complex problems in society.

HPC systems are composed by compute and storage resources interconnected by a high-speed network. These systems may have thousands of compute nodes that are leveraged by its users to execute complex applications, known as jobs. The access to such resources is mediated by a job scheduler, which allocates compute nodes to jobs on a queue based on the user-defined conditions and availability of the resources in the system. The scheduler is also responsible for monitoring the jobs and controlling the contention to shared resources by managing a queue of pending work. The power drawn from the intensive use of computing, overhead equipment, and cooling resources makes HPC systems large-scale electricity consumers and contributors to climate change [2].

The demand for HPC is growing in both the public and private sectors. According to an analysis from MarketsandMarkets, the market for HPC is likely to grow from USD 36 billion in 2022 to USD 49.9 billion by 2027 [3]. The HPC industry, alongside similar ones such as traditional data centers, is largely driven by the increasing need for computing power, networking, and storage [4] of emerging fields like Artificial Intelligence (AI), Internet of Things (IoT), cryptocurrencies, 5G networks, and plays a vital role in smart-city infrastructures [5]. These fields significantly raised the demand for the Information and Communication Technology (ICT) industry [6], which is forced to scale and adapt, but also to become increasingly aware of its sustainability [7].

Additionally, HPC moving from petascale to exascale (systems capable of at least one exaflop) creates new challenges [8], such as a large amount of energy consumption, with operational costs getting closer to parity with capital costs. The TOP500 list [9] indicates that the current

[*] Corresponding author at: INESC TEC, Centre for Power and Energy Systems, Porto, 4200-465, Portugal.
E-mail address: carlos.silva@inesctec.pt (C.A. Silva).
URL: https://inesctec.pt (C.A. Silva).

https://doi.org/10.1016/j.rser.2023.114019
Received 6 April 2023; Received in revised form 26 September 2023; Accepted 23 October 2023
Available online 7 November 2023

Abbreviations

AI	Artificial Intelligence.
ANSI	American National Standards Institute.
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers.
DVFS	Dynamic Voltage and Frequency Scaling.
EDP	Energy Delay Product.
ETP4HPC	European Technology Platform for High Performance Computing.
HPC	High-performance computing.
HPCaaS	HPC as a service.
HPCG	High Performance Conjugate Gradients.
HPL	High-Performance Linpack.
ICT	Information and Communication Technology.
IoT	Internet of Things.
ISO	International Organization for Standardization.
IT	Information Technology.
KPI	Key Performance Indicator.
LEED	Leadership in Energy and Environmental Design.
LUMI	Large Unified Modern Infrastructure.
ODA	Operational data analytics.
PUE	Power Usage Effectiveness.
RAPL	Running Average Power Limit.
RES	Renewable Energy Sources.
RISC	Reduced Instruction Set Computer.
SLA	Service Level Agreement.
TACC	Texas Advanced Computing Center.
TES	Thermal energy storage.
UPS	Uninterruptible Power Supply.

fastest supercomputer, Frontier, the first exascale supercomputer with a performance of 1194 PFlops/s, consumes 22.7 MW of power while ranking 6th in the GREEN500 list [10] with a power efficiency of 52.2 GFlops/Watt. Further advances are needed to achieve the exascale HPC vision in a sustainable way. Current architectures seem to be distancing from traditional clusters of homogeneous nodes to clusters of heterogeneous nodes, due to the integration of specialized hardware optimized for specific computing purposes [11,12], as a way to accomplish more operations per second. This trend has led to the last ranking set with 7 of the top 10 supercomputers being heterogeneous clusters with GPUs.

Cutting-edge HPC systems use server-grade multicore CPUs, which rely in high core counts to achieve large computational throughput, often coupled with hardware accelerators, such as GPUs. The recent push to wider and faster processing devices has increased the power draw of both CPUs and accelerators. These factors, combined with the requirement to minimize the connection distances among CPUs and accelerators, result in high wattage densities at these systems' node- and rack-level. This imposes several challenges regarding the cooling of those supercomputers. Moreover, HPC centers often have unique characteristics, such as limitations in job scheduling strategies and management of electrical energy consumption, which stem from the requirement of high availability of most, if not all, computing resources.

The adoption of edge computing architectures and the IoT, which aim to process data close to its source, will be essential for time-critical and data-intensive applications. Those architectures also increase the potential of digital twins, virtual representations of objects that merge sensor data with surrogate models [13]. The combination of Cloud and HPC allows to apply large amounts of compute power by sharing the computational burden [14] and allows the collaboration among all stakeholders active in the digital continuum [15].

The previous arguments led to increased efforts by the industry and research institutions to enhance the sustainability of computing centers in their design and operation, with new technologies in the sector contributing to its decarbonization. Furthermore, it motivated regulation, standardization, and funding opportunities regarding implementing and reporting energy efficiency measures and energy management strategies to minimize energy consumption and carbon footprint. Recently, the concept of carbon-responsive computing was proposed as the range of techniques in which "energy sources and computing elements cooperate to prioritize energy usage with the least carbon intensity" [16]. This concept offers a holistic and broad perspective on the interdependencies of computing and energy systems, and the socio-technical implications of emerging techniques that aim to decarbonize the ICT industry as a whole.

There is a growing need for actions that enhance sustainability awareness and responsiveness in HPC centers due to increasing energy consumption, costs, and environmental impact concerns. A focus shift from costs to carbon emissions in HPC creates a research and industry gap, in which innovations related to hardware, software and applications, resource management systems, user interaction, and novel business models have an opportunity to thrive. These innovations must help reduce operational costs and carbon footprint but also support the overall growth of the industry [4].

This work aims to provide an overview of measures and opportunities related to sustainability in HPC centers. This objective is achieved with a thorough review of available literature, recent industry trends, and a critical view of HPC infrastructure, software, and resource management. The work also focuses on the roles of the users, digital service providers, and their interaction, to further contribute to the decarbonization of this sector. The main contribution of this review is exposing the current trends of the HPC sector from a decarbonization perspective, detailing available data for state-of-the-art supercomputers, and proposing future lines of research. While other reviews focused on renewable energy integration and decarbonization of data centers, this review highlights the specific challenges of decarbonizing the main components in HPC centers (infrastructure, hardware, software). Moreover, this work includes a new dimension in the discussion – the service provided by HPC centers – not adequately addressed in similar studies. This analysis covers the users and service providers, as well as recent trends in policy, funding, research, and business models. This work is expected to provide a critical analysis of the energy and IT sectors, with implications for the design of future HPC systems.

This document is organized as follows: Section 2 details the methodology used in the review and characterization of the work presented throughout this communication; Section 3 addresses the HPC infrastructure, and strategies for energy efficiency; Section 4 discusses hardware- and software-level tools to monitor and manage the energy usage of HPC resources; Section 5 presents the services provided by HPC centers, analyzing the impact of policies and business models; Section 6 presents the main key performance indicators available and some recent discussions around this topic; finally, Section 7 summarizes the most relevant topics of this review and proposes future research directions.

2. Literature review methodology

Previous surveys on sustainable and carbon-neutral data centers mostly focus on Cloud computing. While there are many surveys available on energy-aware computing in HPC, they may not cover all the aspects of such a fast-paced field, which include changes in technology, but also in the surrounding ecosystem. Therefore, this review aims to provide a comprehensive and multidisciplinary overview of the latest developments that may not have been covered in previous surveys.

Table 1
Comparison with other surveys.

Reference	Year	Target	Carbon	Renewable	Infrastructure	Hardware	Software	Service
[4]	2015	HPC, traditional		✓	✓		✓	
[17]	2019	HPC				✓	✓	
[18]	2021	HPC				✓	✓	
[7]	2021	traditional	✓	✓	✓			
[19]	2022	traditional	✓	✓	✓	✓	✓	
[20]	2022	HPC, traditional	✓	✓	✓		✓	
Current study		HPC	✓	✓	✓	✓	✓	✓

Table 1 summarizes the differences between this and other surveys on sustainable data centers. In detail, the columns in the table describe: if the target of the review is traditional data centers, HPC centers, or both; if carbon emissions are addressed; if the integration of RES is addressed; if the sustainability and energy usage at the facility/infrastructure level is addressed; if the sustainability and energy usage at the hardware level is addressed; if sustainability and energy usage at software level is addressed; and if the sustainability and energy usage of the digital service as a whole is addressed. The research in this study focuses on the decarbonization of HPC centers, not only on the infrastructure but going from the digital service to the surrounding ecosystem.

A thorough literature review on sustainability in the planning and operation of large-scale computing centers was conducted, with more focus given to HPC. To be able to provide a multidisciplinary view, we consider research from multiple domains, including electrical engineering, control systems, and computer science. The review included scientific publications, governmental and industry reports, as well as reports from scientific projects, and concrete examples of sustainability practices from research and industry (presentations, workshops, and news). The most relevant documents close to the topic addressed in this work were selected, and the remaining were discarded from the analysis. Although a large amount of literature is available for sustainability-related topics in traditional data centers, there is little work to reflect the current concerns and efforts on sustainability specific to the HPC industry. In some topics of this work, both systems share roughly the same issues, but priority was given to literature that was specific to HPC.

The data sources for the information in Tables 2–5 include the TOP500 and Green500 lists, as well as other academic and industry publications and presentations reporting on the energy efficiency and decarbonization measures of HPC centers. The main limitation of this study is the low availability of data regarding the performance of state-of-the-art HPC systems. Although there is some publicly available information on the mentioned lists related to hardware and some performance indicators, more specific and updated information on other performance metrics and each system's energy and IT infrastructure was scattered through multiple sources or even not found for some of the described HPC systems. Therefore, not all modern supercomputers were included in the analysis.

The four-pillar framework proposed by [21] divided energy efficiency in HPC centers into building infrastructure (reduction of energy losses, efficient cooling technologies, and practices, reuse of waste heat), hardware (reduction of hardware power consumption, acquisition of energy-efficient equipment), system software (workload management, tuning of software tools), applications (optimization of applications to the hardware in use, selection of efficient algorithms and libraries). The aim of this work was to promote cross-pillar measures that benefit from the interaction between different levels of the HPC center. Work developed by [22] extends this categorization and proposes a seven-pillar framework for energy efficiency in HPC: infrastructure, system hardware, system software, applications, network, policy, and usage.

The division in those previous works was done taking into account energy efficiency approaches, but not the broader sustainability concerns. While the latter work provides a more complete overview of the entities involved in the management and operation of HPC centers, its view of the systems and their connections presents limited opportunities to address sustainability concerns. This work adopts a more abstract view of this seven-pillar organization, focusing on the entities and relationships most relevant for a holistic view of the decarbonization concerns in HPC centers. This view is depicted in Fig. 1 where, for each highlighted topic (and most relevant subtopics), the main sustainability concerns are exposed and discussed in their respective sections throughout the study, unraveling recent progresses in the field and perspectives for future research. Furthermore, relevant information on current large-scale HPC systems is presented for better comprehension of the state of the art.

3. HPC energy assets and technologies

This section addresses the decarbonization of the HPC energy infrastructure, in terms of energy efficiency strategies and technologies, and new assets like RES and storage.

Fig. 2 presents an overview of the different components of an HPC center, such as the users, and the IT and energy infrastructure. Users interact with the HPC infrastructure using the frontend cluster nodes, which are dedicated to code compilation and job submission and interacts with the backend compute nodes through resource managers. A user submitted compute job is assessed by the resource manager, which combines a predicted computational profile of the job with information about the usage of the compute nodes to allocate a set of resources to compute the job.

The electrical energy required to power the whole infrastructure (the HPC center and its offices) can be obtained through on- or off-site generation, or directly from the grid. The heat output from supplying power to the hardware and HPC center can be reused to both office and district heating.

3.1. Energy efficiency strategies and technologies

The energy consumption of computing centers is mostly related to their computing resources (servers, communication equipment, and storage), and physical resources related to the infrastructure (cooling, power supply, lighting, and other systems). Cooling can account for up to 33%–40% of data center energy usage depending on location and power density and consumes hundreds of billions of gallons of water per year in the USA [23]. As a result, most efforts regarding energy efficiency are centered on the computing itself (hardware) and in new cooling equipment and technologies, their configuration, and possible applications for waste heat. Energy efficiency measures decrease the overall electrical energy consumption associated with the operation of an HPC system.

Multiple reviews of concerns on energy efficiency and usage in traditional data centers and the main measures for its improvement are available [4,7]. Regarding this topic, traditional data centers share roughly the same concerns as HPC, with a focus on infrastructure and the efficient management of energy and computing resources. Specifically for HPC, work by [24] gathers energy-saving techniques which include energy monitoring and control, site infrastructure, hardware, and software, while describing the main differences to traditional data centers.

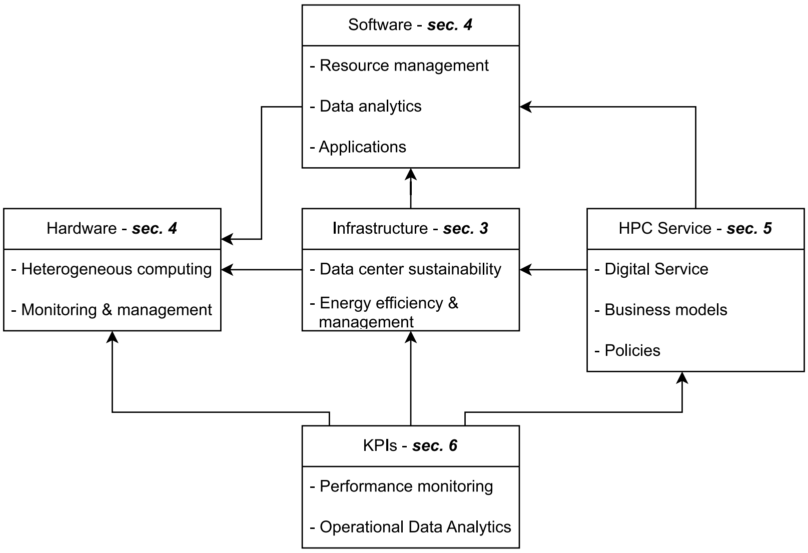

Fig. 1. Interaction among the entities in the proposed organization of this review.

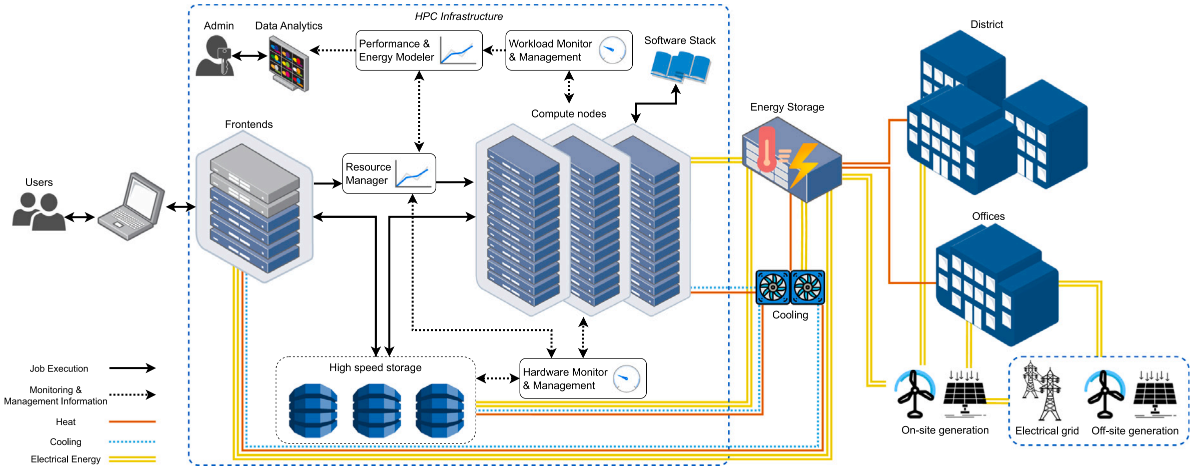

Fig. 2. Overview of the relations among the different components in an HPC infrastructure, its users and submitted jobs, and electrical and thermal energy flow.

The main concerns when designing a new HPC infrastructure are reported in [25]. In this work design solutions are mainly evaluated according to their support of future systems, sustainability, lifecycle, and initial costs. A brief overview of measures in different subsystems is presented in the following subsections.

3.1.1. Cooling

The cooling infrastructure is crucial to maintain the environmental conditions of computing rooms within a specified operational range, due to heat dissipation from IT equipment, a byproduct of the operation of HPC servers. HPC systems have demanding thermal stability specifications [26], thus optimized and well-designed cooling systems assure the safety of operation, and impact the overall efficiency and performance metrics. The selection of cooling technology is imperative and depends on factors such as server density, cost, total power consumption, available IT and building infrastructure, and how they are integrated. Cooling systems in computing centers are currently based on the following strategies: air-cooling, liquid-cooling, phase-change cooling, and immersion cooling.

Liquid-cooled systems are now the standard technology due to the high thermal loads that recent computing centers must support, replacing air-cooled systems. Usually, this technology is applied with the use of cold plates located near the equipment [27], and one of the main advantages is a more efficient heat transfer, which results in the possibility of using higher operating temperatures within the system, and consequently leads to a higher quality (higher temperature) waste heat. Another advantage is the potential of eliminating some system components. When server blades do not support water cooling, an alternative strategy is rear-door cooling [26], where water is distributed through a piped network to the back of the servers, transferring heat from hot air to the water, which is then cooled. As stated by [28], energy efficiency associated with the operation of the cooling equipment can be enhanced by using hot water, a technique that is already commonly adopted. Another technique is the free cooling method [4], to take advantage of natural resources available, although its effectiveness is greatly dependent on the location of the facility. This principle relies on the use of external low-temperature air, used directly or indirectly (by means of heat exchangers). It can also be done using a cold-water source. The system can, therefore, operate with lower or null cooling power by minimizing the use of mechanical active components for a certain number of hours per year.

Phase-change cooling is another solution that has been recently adopted by the computing industry, where liquid-vapor phase change occurs within the heat exchange process. A recent review [29] summarized the advantages of this solution and recent progress. Another type of cooling is immersion systems [30], which directly immerse hardware in a non-conductive liquid, although introducing some maintenance challenges and the need for compatible hardware. Several companies offer commercial immersive cooling solutions [31,32].

Airflow management covers a set of principles that computing center operators can rely on for cooling optimization. The usual server configuration takes advantage of the hot/cold aisle arrangement, a layout where the air enters the equipment through cold aisles and the hot air exits through hot aisles, preventing both streams from mixing and ensuring a higher energy efficiency [27]. A related measure is the prevention of hot spots, which appear due to recirculated airflow when airflow management is not well performed. Containment systems, layout optimization, calculation of ventilation efficiency indices for recirculation analysis, and computational fluid dynamics techniques can diagnose and act on suppressing hot spots, and better characterize the room's airflow and temperature field [33].

Another possibility for energy efficiency regarding cooling is to allow equipment to work in higher room temperatures [4]. This technique results in reduced energy demand, and increases the possibility of using free cooling for a greater percentage of the year. Reliability and operational conditions of each IT equipment are the main concern and drawbacks of this idea. Other strategies include the use of variable airflow, the use of equipment in partial load (fans, pumps, chillers), and the selection of the most efficient equipment in an early stage when designing a computing site.

3.1.2. Waste heat

Waste heat plays a critical role in energy efficiency, as most of the electrical energy consumed in a HPC center is converted into thermal energy. The powerful computing resources available in HPC result in large amounts of heat being removed, with the potential to be captured and reused for many applications [34]. The amount of waste heat (and its quality) depends on the selected cooling technology, which consequently defines the location and temperature at which the heat can be captured [35].

The main opportunity for waste heat reuse is in space and water heating, which includes local heating of the HPC facility. If feasible, waste heat can be distributed in nearby buildings and neighborhoods through district heating, which consists of using the heat generated from different energy sources, and connecting to consumers (residential, businesses, and industries) through a piped distribution network. A discussion is presented in [35] on the integration of residual heat specifically from data centers in district heating networks, and the main challenges. This type of solution for waste heat can be found in some state-of-the-art computing centers (Table 2). As an example, LUMI [36] is using the waste heat generated by their supercomputer to supply a local district heating network in Finland with up to 20% of that area's thermal needs.

A thorough review of other possible wasted heat applications was made by [27] while describing promising technologies. The authors addressed the issue of low-quality (i.e., low-temperature) residual heat and how to overcome this obstacle, including using heat pumps to increase fluid temperature. Options found by the authors for waste heat reuse include power plant co-location, absorption cooling, direct power generation, biomass co-location, and desalination/clean water.

3.1.3. Power supply

The power supply of an HPC infrastructure deals with ensuring a reliable power supply to the IT and remaining equipment, due to the critical nature of the service. As HPC scales in computing power, so do its power requirements.

HPC systems can be subjected to abrupt variations in their power consumption, due to the inherent variable behavior of computing-intensive applications scheduling and execution. Large peak demands (dozens of MW) are expected in the near future, which will impact power distribution requirements and enhance these concerns. Recently, load swings of over 7 MW have been reported in multiple large-scale HPC systems in timescales of seconds or less, which must be immediately addressed by the infrastructure's power supply [37]. A deeper understanding of the fluctuations and of local electrical grid stiffness [38], as well as the use of devices such as on-site electrical energy storage, can help support these critical periods, and ease the burden for the cooling system and the electrical grid, which must minimize voltage and frequency fluctuations.

Energy efficiency measures on the power supply architecture and equipment usually relate to efforts to reduce distribution inefficiencies/losses. The work of [4] indicates that to avoid AC/DC conversion inefficiencies, and direct current electrical networks were studied for power distribution architectures. DC-powered data centers have been studied in [39], where challenges related to power electronics were discussed at several levels (from building to individual racks and components). This work showed that projects with DC distribution at the building level have been tested, showing positive efficiency, availability, and reliability results. More recently, [40] thoroughly reviews the power conversion steps (AC/DC and DC/DC) of data center power supply systems, including high-voltage DC distribution, and reports that going from AC to DC-powered can lower the number of conversion stages and increase the overall efficiency (ranged from 93 to 96% for the reviewed architectures, from the grid to CPU-level). Moreover, the authors emphasize the importance of recent technological improvements (e.g., at the converter level), and argue that innovations related to circuit topology and control are important to enable renewable energy integration. The authors of [4] indicate that the usage of Uninterruptible Power Supply (UPS) units, which provide emergency power when the main power supply fails, can be bypassed to avoid conversion inefficiencies, and that modular UPS devices can be deployed (and activated/deactivated according to the workload). The use of UPS can also be considered only for critical infrastructure components, to reduce overall losses [41].

3.2. Distributed energy resources

3.2.1. Renewable energy sources

The integration of RES, such as wind or photovoltaic systems, in the power supply of computing centers is an obvious measure to decarbonize part of the HPC center electricity supply, diversify electrical energy sources, and achieve a certain level of independence (or resilience) from the main grid.

Besides on-site generation, off-site generation is also a viable opportunity if RES near the facility are deemed insufficient [35]. As power requirements continue to increase, on-site generation would also require a larger area, usually a limited resource. The authors of [35] also present details and examples of RES provided by third parties to offset carbon emissions, which include the acquisition of Renewable Energy Certificates (REC) that track the energy source of the consumed electricity, or Power Purchase Agreements (PPA), in which the facility purchases electricity from a RES producer via a contract that establishes the terms and prices. These solutions increase the RES share in the power supply while avoiding high upfront investments in RES systems.

In fact, currently, it is not practicable to power a data/HPC center only with RES at a reasonable cost and reliability, and thus energy storage and conventional generators are still necessary, as indicated by [20], which also provides many examples on investments for on-site RES systems from data center companies.

A good example of a mixed solution regarding the use of RES is the case of the Texas Advanced Computing Center (TACC), which purchases credits from wind power plants in West Texas, while using on-site generation through PV panels. Further examples of RES integration in HPC centers are presented in Table 2.

3.2.2. Energy storage

Battery energy storage systems are the standard solution in UPS units to assure backup power for short-time emergency operation. Nevertheless, large-scale on-site energy storage systems are also becoming the obvious solution to complement and replace, to some extent, diesel generators as backup power, and can be combined with on-site and nearby RES power plants [20].

Therefore, using batteries can increase the self-consumption levels associated with energy generated by on-site RES and decrease the dependence on the electrical grid. Moreover, it can also enable HPC infrastructures to act as providers of ancillary services to the power system or to participate in flexibility markets. Furthermore, it can help advanced energy management strategies to be deployed, in which charging and discharging operations are aligned with periods that are favorable to a specific objective (see Section 4 for a more detailed review).

Thermal energy storage (TES) is utilized to store captured heat removed from the IT equipment and provide local space and water heating. Within a district heating network, TES can act as a buffer when the network lacks demand. A recent review was done by [42] on thermal energy storage's current status and applications in data centers and its integration with the cooling system. TES can also consider the storage of coolant to use for IT equipment cooling, or to participate in district cooling networks. Work by [35] details concrete examples that used thermal storage, including ice storage systems.

3.2.3. Controllable loads

The electrical load of an HPC infrastructure is approximately the result of executing a set of workloads submitted by HPC users, plus the overhead consumption of the facility. Workloads can be scheduled to alter the overall electrical load, which relates to the concept of load shifting. Thus, the HPC load can be partially considered a controllable load. Furthermore, cooling is also a result of the workload execution, thus the cooling electrical load is also controllable to some extent, as long as the system operates in a safe range. Many restrictions limit the extent of this control, such as the low flexibility that results from the high utilization rates of HPC centers (when compared to traditional data centers), the resource management capabilities of the system, and the quality of service concerns and commitment to SLAs.

3.3. Use cases

A more robust approach to address sustainability concerns can consider the combination of different measures presented throughout the previous sections. Several HPC facilities are now following this approach, for example:

- The Super MUC-NG HPC at Leibniz Supercomputing Center has reportedly achieved around 30% savings in energy consumption using efficient measures such as low-power servers, reduced cooling power by using warm water cooling, energy-aware scheduling, and adsorption chiller as a strategy for heat reuse.
- LUMI, a consortium formed to apply to EuroHPC, is using the waste heat generated by their supercomputer to supply a local district heating network in Finland with up to 20% of that area's thermal needs, while integrating 100% renewable energy from hydropower, claiming to be the first carbon-negative HPC center.
- ESIF HPC center from the National Renewable Energy Laboratory considers a cooling system using warm water at the component level, and reusing the waste heat from the Peregrine and Eagle supercomputers in other applications within the surrounding building.
- A recent report provided by Los Alamos National Laboratory reveals the main concerns on how to design an energy-efficient HPC supercomputing center [25]. The institution is assessing design alternatives to support two exascale supercomputers arriving in 2026, with considerations in many levels of its energy infrastructure.
- Advanced Research on Integrated Energy Systems (ARIES) [43] is a research platform from the National Renewable Energy Laboratory that performs research related to the development of innovative energy-related technologies (at the 20 MW level), which is supported by an 8 PF supercomputer. Their Microgrid Infrastructure already demonstrated the capacity to support the associated campus during outages, when a recent and unexpected event made researchers successfully connect 430 kW of PV and a 1.5 MW wind turbine to charge their battery and repower the site.
- Massachusetts Green High Performance Computing Center relies on periods of low external air temperature to use free cooling techniques, claiming that, in certain conditions, chillers can be turned off for around 70% of the time, and are mainly powered by RES from near-by hydroelectric and photovoltaic generation.
- Minho Advanced Computing Center's Sustainable HPC project [44] is an ongoing project, focused on developing an innovative energy management solution that allows a more sustainable operation of the upcoming Deucalion supercomputer. The project is developing an energy management system with predictive capabilities leveraging controllable assets and on forecasted variables to actively shift the system to a low-carbon operation. It will integrate RES and multiple storage solutions (electrical and thermal).

3.4. Discussion

Table 2 provides a list of recent supercomputers and their HPC center information to support the discussion. A set of representative centers was chosen, which allows for analyzing the evolution of HPC energy infrastructures in recent years.

Liquid-cooled systems are now the standard in HPC centers due to the high thermal loads that recent computing centers must support, mostly replacing air-cooled systems. For example, the Riken HPC center previous supercomputer, the K computer, had a power ratio between supercomputer/liquid cooling/air cooling of 85/7/7, while its recent Fugaku supercomputer, currently 2nd in TOP500 [9], has a ratio of 86/11/1. While some systems only use rear-door cooling, the trend is to use 100% water-cooled components, not only for the CPU or GPU, but also dual in-line memory modules (DIMM) and network interface card (NIC), as on Frontier supercomputer racks. Table 2 also shows that temperatures used for liquid-cooling increased considerably in the last decade and that the current trend is using warm/hot water [45]. The use of warm/hot water decreases overall cooling power, due to a reduced use of the cooling units, and increases the potential of free cooling technology. Free cooling is now adopted in most centers, as shown in Table 2.

Many studies are also available on the topic of RES integration in data centers. There is a recent trend to increase the RES share in the electricity supply using local RES generation, off-site generation, or alternative means (power purchase agreements, credits, and certificates), with some examples presented in Table 2. Also, energy storage is being discussed beyond of the usual role of backup power.

The sustainability of supercomputers and HPC centers is a critical concern as these facilities continue to grow in scale and computational power. Taking into account the current trends and challenges we envision:

- Research into more efficient liquid cooling systems, including warm water cooling, to dissipate heat from increasing high-density computing components.
- Assessing the feasibility and effectiveness of deploying on-site renewable energy systems, such as solar panels and wind turbines, combined with microgrid solutions to power HPC centers. This also implies developing intelligent energy management strategies and algorithms that adapt computing tasks to match the availability of renewable energy resources.

Table 2
Trends in cooling and RES in high-performance computing centers.

Name	Year	RES	Waste heat reuse	Water Temp. (°C)	Free cooling
K	2012			7–9	✓
Titan	2012			5.5–8.8	
Summit	2017			17–21	✓
SuperMUC-NG	2018		✓	47–50	✓
ESIF	2019	✓	✓	18–23	✓
Frontera	2019	✓		20–25	✓
Fugaku	2019			15	✓
LUMI	2021	✓	✓	30	✓
CEA-HF	2021		✓	45	✓
Frontier	2022			32	✓
Leonardo	2022			37	✓

- Deploying advanced energy storage solutions, including high-capacity and fast-charging batteries, to store surplus renewable energy for use during periods of high demand or high electricity market prices, combined with thermal energy storage.

4. HPC software and hardware resources

Historically, new HPC systems focused on improving the raw computational performance, but recently, energy efficiency has emerged as an equally critical guideline. It is no longer possible to perceive software and hardware as isolated entities in the HPC ecosystem; instead, their intrinsic interdependence must be recognized to achieve energy-efficient computing. Energy expenditures not only strain budgets but also contribute to environmental challenges. Consequently, the HPC community is at an inflection point where the design and orchestration of the software stack, including the operating system, libraries, compilers, and applications, are intrinsically linked with the underlying hardware infrastructure, comprising heterogeneous servers, complex memory hierarchies, and specialized accelerators.

4.1. Resource management systems

The adequate management of the available computational resources in computing clusters is key to decrease data centers' energy consumption. Several strategies and tools have been designed to monitor and manage the hardware in such systems, but mostly targeting cloud computing. The hardware in HPC cannot be managed to the extent often required by frameworks that target cloud computing. The requirement for high availability of resources in an HPC cluster minimizes the potential for aggressive energy management, such as server sleep/shutdown, usually employed in cloud environments. However, other less intrusive approaches, such as dynamic frequency scaling, are still suitable, as long as workloads' behavior and energy consumption can be predicted with some accuracy.

4.1.1. Energy-aware computing

The reduction of energy consumption in cloud computing data centers is an issue that has been tackled since the beginning of the last decade. The work of [46] presents a framework for resource allocation that accounts for application performance and the power draw of the allocated hardware. However, this approach relies on self-optimization algorithms to predict the power draw of workloads, resulting in a non-ideal allocation of resources in several edge cases. The emergence of AI, which was not mature enough when this work was published, had the potential of significantly improving the predictions, leading to better resource management and less overall power draw.

Work by [47] has shown that AI can help improve resource allocation in cloud computing systems. The authors developed a framework combining a virtual machine resource-allocating scheduler with a dynamic power management policy, identifying an adequate configuration for the allocated hardware that minimizes power draw while maintaining reasonable performance. This work shows that reinforcement learning-based algorithms can aid in accurately predicting energy consumption and tuning the hardware resources when the state/action space is too large for other approaches.

The work of [48] takes this approach a step further, combining fine-grain tuning of the hardware resources at the software level with dedicated power monitoring hardware, proposing an open-source supercomputing architecture. The authors also use AI-based methods to predict the power draw of workloads, which impacts the hardware and its configuration allocated to a given job. However, the potential of this architecture is dependent on the adoption of custom monitor and management hardware to be integrated into the computing nodes, as the availability and accuracy of power sensors vary among chip manufacturers.

The operation of a supercomputer can be shifted by energy management strategies that contribute to a certain objective or target specific performance metrics. Author of [49] states that at exascale HPC, power is a scarce resource, thus wise decisions have to be made regarding power allocation. Currently, there is a greater emphasis on the interdependence of supercomputer and cooling infrastructure in HPC center designs [50]. Given their size and complexity, optimizing the operation of HPC systems requires a coordinated effort to plan and manage power consumption across the entire infrastructure.

Related surveys also discuss energy management in computing centers. For example, [20] divided the management of green data centers in energy source and storage management, and IT management (for different types of workload), and addressed both topics separately. A number of recent examples have been discussing how the management of energy and computing resources can be addressed in a more coordinated or cooperative way. In the current work, both energy and IT infrastructure management are discussed jointly.

In the literature, authors typically focus on the concept of energy-aware computing, defined as a set of techniques that enable managing computing and energy resources to adjust the real-time energy usage of the facility, based on the current workload and the available resources. A taxonomy study on energy-aware computing has been performed by [51], which characterize the different approaches on energy efficiency according to the scale of applicability (from compute servers to data centers and grid/clouds), the goal of the approach (direct or indirect energy savings), methodology (workload management, hardware configuration, programming, etc.), and viewpoint (different goals from a user, developer, or resource manager perspectives). While the related work mentioned in this study is outdated, the taxonomy is still relevant. A recent survey on the state of energy-aware HPC was developed by [17], and discusses the available optimization methods, categorizing them into monitoring and controlling.

For controlling methods such as scheduling, authors tend to rely on directly altering the workload, by shifting or migrating it (spatio-temporal workload shifting). Shifting is accomplished through job scheduling algorithms and policies. For example, the work in [52] matches the workload to the RES supply available and states that green-energy-aware scheduling can be crucial for developing a more sustainable IT ecosystem. Other works include [53,54]. A comprehensive survey in [20] lists existing RES-availability scheduling approaches on single and multiple data centers, with examples of how the demand from IT resources can be adapted to the energy or power availability.

The authors of [55] monitored the energy consumption of jobs using the resource and job management system to be reflected in the system's accounting. Such initiatives aim to sensibilize the users to the power draw of their applications, to make this a key consideration when submitting jobs to such systems. This approach was extended in [56,57] by implementing a resource and job management system that limits the overall energy consumption of HPC systems through several mechanisms: Dynamic Voltage and Frequency Scaling (DVFS), which adapts the processor voltage, lowering its frequency and consequently minimizing the power draw; Dynamic Concurrency Throttling (DCT), which reduces the amount of available resources in a processor, such as the number of hardware threads; and Power Capping, which is similar to DVFS but limits the power draw of the whole system according to an energy budget.

Moreover, the scheduling of jobs in a resource and job management system can consider the availability of local renewable energy sources or green energy from nearby power plants. GreenSlot [52], a parallel batch job scheduler designed for a data center with a power supply composed of photovoltaics and the electrical grid as backup, schedules the workloads in a way that optimizes green energy consumption, relying on the prediction of future availability of solar energy, and ensuring that jobs' deadlines are not compromised. Meanwhile, GreenSwitch [58] is a framework that enables scheduling workloads and choosing the optimal energy source (solar, battery, and/or grid) in different periods.

4.1.2. Thermal-aware computing

Thermal-aware computing is another approach that attempts to manage heat generation in computing hardware as a byproduct of computation. Such approaches monitor the heat generated by a system and manage the software and hardware to prevent excessive thermal output while maintaining adequate performance levels. Thermal-aware approaches can involve management of the cooling infrastructure, the configuration of the computing hardware, scheduling, and resource allocation. Mixed-integer linear programming formulations are explored by [59], in the context of task scheduling for HPC centers. To achieve this goal, thermal models that characterize the temperature profile (both in time and space) and different objective functions were considered, such as the minimization of makespan or energy consumption. A thermal-aware workload management problem is also formulated by [60] to maximize the RES self-consumption, by taking advantage of pre-cooling strategies. If there is RES surplus generated on-site, computer rooms can be cooled to a lower temperature (within the recommended range) and delay the need for cooling power. Different temperature control strategies and their impact on reducing carbon emissions are compared by [61], by developing an optimization problem and testing it with different workloads.

4.1.3. Carbon-aware computing

More recently, the concept of carbon-aware computing focused on the carbon intensity associated with the electricity mix supplied to computing facilities, and how different energy management and optimization strategies can reduce the overall carbon footprint of computing systems while maintaining the quality of service. Geographically distributed data centers were explored by [62], exploiting the spatio-temporal variability of grid carbon intensity. The authors compute the optimal operation of a cloud service under a carbon emission reduction budget, by developing a carbon-aware control framework. An emission-aware resource planning model considering day-ahead scheduling and uncertainty was introduced by [63] to reduce the carbon footprint of a microgrid that contained a data center. A two-stage optimization problem allocated workloads in different time slots within a job scheduler, by minimizing a weighted sum of electricity costs, carbon footprint, and operational risk.

Google published initial results for their Carbon-Intelligent Computing System [64], a data center carbon-aware workload shifting initiative. The project relied on creating virtual capacity curves (hourly resource usage limits) as a load-shaping mechanism for the next day's resource usage, based on forecasts of grid carbon intensity, and flexible/inflexible demand. The optimization problem minimized a weighted sum of carbon emissions and power peaks on a daily basis. Work by [65] analyzed workload temporal shifting to less carbon-intensive periods, by identifying delay-tolerant workloads and relying on carbon intensity forecasting. The authors considered carbon intensity data from different regions to better understand the potential of the approach, and relied on scheduling flexibility of the workload, although not accounting for resource constraints.

A review by [19] included carbon-aware computing as a key aspect to achieve carbon-neutrality in data centers, by job scheduling or migration to geographically distributed data centers. However, the authors state that the available literature does not reflect interactions between IT components and the surrounding infrastructure. Therefore, a digital twin approach (included in their work) could enhance scheduling policies by accounting for the impact on the remaining infrastructure. The Carbon Explorer framework was proposed by [66], with a holistic view of sustainable data centers given by a multi-dimensional solution space, according to investments in RES capacity (wind and solar), battery storage capacity, and carbon-aware workload scheduling. The solutions were analyzed according to the trade-off between their embodied and operational carbon footprint. It was concluded that only a combination of the different techniques could aid data centers in approaching a 24/7 RES operation.

Carbon-responsive computing [16] is a concept that explores the interdependencies of energy and computing in a holistic way, by prioritizing energy sources with the least carbon intensity. The term comprises techniques that exploit the collaboration between all kinds of computing elements and different energy sources while dealing with the increasingly distributed nature of RES and computing infrastructures. The concept also focuses on the sociotechnical implications that come from the implementation of carbon-aware techniques.

Usual performance metrics to be optimized with energy management include power and energy consumption, and job execution time. More recently, the focus shifted to maximizing the RES self-consumption or minimizing the indirect carbon emissions. Assessing the compromise between applying energy management techniques and the overall system's performance or degradation in the digital service is common.

4.1.4. Workload tracing and analysis

The wide range of applications, with varying characteristics and computational behaviors, contributes to the overall irregularity of the HPC workloads. Adequate tracing of the workloads is crucial to developing more accurate predictive models to mitigate this irregularity, which is shown to be a key limitation of the work presented in the previous section.

The work of [67] focuses on characterizing scientific workflows, the primary type of jobs executed in large-scale heterogeneous HPC clusters, and surveys current strategies for energy-aware job scheduling. A framework architecture for energy-aware workflow management and scheduling is proposed but addresses only the interaction between the workflow scheduler, energy prediction, and server management, rather than providing concrete solutions to tackle the issues at a lower level. The authors also propose several novel research opportunities of combining specific scheduling heuristics depending on the meta characteristics of the workflows. They emphasize the need for better and faster AI-based power draw prediction models, as their accuracy significantly impacts the quality of the scheduling solution.

This characterization of the energy properties is relevant at the workflow level, which combines user and third-party code (other tools, libraries, etc.), but also at a lower abstraction. Work by [68] addresses the energy efficiency of several constructs available in OpenMP, one of the most popular libraries in scientific computing, to develop parallel code. The authors developed benchmarks and assessed the performance of OpenMP directives and loop transformations across a variety of compilers. This work concludes that inadequate handling and scheduling of threads, which is usually under the responsibility of the developer or programmer, has a considerable effect on the overall power draw of an application. The use of parallelization libraries that have a greater focus on the scheduling of threads and irregular workloads over OpenMP has the potential to improve the power draw of an application.

The energy profile of several AI algorithms was analyzed by [69] to estimate trends in the power draw of such workloads on heterogeneous servers with hardware accelerators for AI tasks. The authors provide a comparative energy consumption analysis of AMD and NVIDIA GPUs, the most common solution for these workloads, but also include accelerators from Google, Cerebras, Qualcomm, and Intel, among others. From this analysis, the authors concluded that gains in energy efficiency cannot be expected only from improvements in hardware, as efficiency scaling is slowing between generations of hardware, and that the training of AI algorithms is becoming increasingly complex. The authors claim that energy efficiency should be enhanced by combining advancements in hardware architecture specifically designed for energy-intensive computations, such as the accelerators tested, with energy-aware fine-tuning of the code.

Simulating the overall operation of a computing center can save time, and avoid costly experiments and downtime of existing computing infrastructures while enhancing the reproducibility of studies regarding resource management systems and policies, and fostering innovation in resource management. Using simulations, multiple sustainable practices in HPC centers can be designed and tested at low cost and without disturbing their normal operation [70].

In [71], a complete machine learning framework has been created to identify performance anomalies at both the job and node levels for compute nodes operating within a production HPC system. The framework collects metrics related to resource usage and CPU performance counters of the applications running in the system to train and validate an artificial neural network. This network is then used to predict which anomalies are restricting both energy efficiency and performance of an application. This information is available through a web interface to the users of the HPC system.

Simulators can include only the scheduling part of the resource management system, to test different approaches to scheduling policies. However, other solutions exist that consider the whole computing resource management system, or even the infrastructure itself, with the interdependencies of computing and energy resources.

A review of the available tools for simulation of job scheduling and/or energy consumption in HPC systems was made by [17] and was then categorized according to its target system (grid, cluster, data center, among others). Some relevant tools found include: DCworms [72], a tool that assesses energy efficiency in distributed computing infrastructures, including HPC workloads [72], TracSim [73], a simulator designed for a common HPC cluster that works with a fixed power cap, SimGrid [74], a versatile and scalable discrete-event simulation framework for grid environments, extended by [75] to account for energy consumption of concurrent applications in HPC grids featuring DVFS technology for multicore processors. Additionally, the work of [76] evaluated the efficacy of two popular energy-aware workflow scheduling algorithms in producing effective schedules for I/O-intensive workflows.

A common issue with existing simulation studies is using only synthetic or random data for job traces. Other issues include unmaintained code, replication of simulators in different languages, and the risk of undetected errors and poor validation with real systems [77].

4.2. Computing resources

In the last years, CPU architectures have seen the adoption of RISC-based architectures, such as ARM, and ongoing with RISC-V, and big.LITTLE architecture, where a chip packs few high-performing cores with a large amount of slower energy-efficient cores (common in the mobile space for more than a decade). Despite energy consumption being a relevant topic for the silicon chips industry in the past years, CPUs have only seen incremental improvements in energy efficiency.

Large improvements in the energy efficiency of the hardware in HPC centers have revolved around the use of application-specific hardware, such as GPUs, which with lower power budgets enhance the performance of specific sorts of computations, common in a broad scope of applications that use these resources, over CPUs. A look at the Green500 list [10], which ranks the most energy-efficient HPC clusters worldwide, reveals that 9 systems in the top 10 rely on heterogeneous computing nodes. These servers combine the flexibility of multicore CPUs with hardware accelerators that are extremely efficient in executing specific tasks, resulting in an overall lower power draw of the system if used properly. Recently, RISC-V based hardware accelerators are becoming increasingly popular due to their reduced design cost, due to the open-source nature of the architecture, and increased power efficiency, and performance over GPUs for the tasks that these accelerators are designed for. The RISC-V consortium includes key players in the HPC industry, such as Nvidia, Google, and Intel [78]. For instance, Google and NASA are working with SiFive on the development of RISC-V based tensor processing units to accelerate AI workloads [79].

A reliable source of data for the energy consumption of a computing server is crucial for an accurate assessment of its carbon footprint, as this metric largely depends on the energy consumption profile. Work by [80] discusses the challenges of monitoring heterogeneous servers in large-scale HPC clusters. The authors compare the functionalities of four different frameworks for energy profiling of workloads (KWAPI, EML, PMLib, and ECTools) with a proposed framework (BEMOS), highlighting their strengths and shortcomings. This analysis emphasizes the fact that accurate power measurement, and consequently power prediction models, in such systems is challenging, as current hardware architectures provide limited access to these metrics. External factors, such as air temperature, humidity, and cooling systems of data centers also have a significant impact on power draw.

A study on the power consumption of the Summit high-performance computer at component-, node-, and system-level by [37] highlights the energy consumption irregularity of HPC applications. The authors identify that specific workloads led to large power swings, ranging between 4MW and 7MW in a period of just tens of seconds, which restricts the use of aggressive power management techniques since the factors affecting energy consumption cannot be easily monitored and/or predicted. According to this work, better energy-aware management of the computational resources should be achieved through more accurate prediction models built using a larger amount of server and system properties, which require more hardware instrumentation, higher polling rates, and monitoring frameworks that scale better while maintaining minimal operating overhead.

4.2.1. Monitoring and management

There have been several studies focusing on the dynamic power management of the two main components in heterogeneous servers: the CPU and the hardware accelerator. While historically most efforts tackled the energy consumption of multicore CPUs, recent work is applying similar heuristics to GPUs, which is the most common type of accelerator in high-performance computers.

The work by [81] proposes a workload queueing model that dynamically manages the energy consumption of the CPU, with the goal of minimizing the average power draw of a server in a given time period. The author compares two heuristics to control the clock frequency of the CPU, using DVFS to modify the CPU core voltage, one that updates the clock frequency between the execution of workloads, and another that changes this frequency during the workload execution. This analysis shows that using adequate power management heuristics, it is feasible to decrease the average energy consumption of a multicore CPU while minimizing the average task response time over a CPU with a constant clock speed. However, this analysis does not account for the power draw of the whole server, which, if considered, may mitigate the efficiency gains on the CPU.

Work by [82] combines a similar DVFS-based tuning approach with Uncore Frequency Scaling (UFS), which affects the speed and energy consumption of the memory hierarchy and CPU interconnections, for more fine-grained control of the system. The authors use a neural network to forecast the energy requirements of workloads and to adequately tune the core and uncore frequencies. While this study shows an efficiency improvement of 16%, and a 2x improvement over the static tuning of the system, the authors only used standard benchmarks to train the neural network and evaluate the proposed approach. An analysis is yet to be performed with real workloads. Additionally, the neural network requires a set of CPU hardware counters that is often only available on Intel chips.

As stated previously, most DVFS-based power management heuristics resort to CPU hardware counters, which vary significantly between chip manufacturers, that can be accessed through libraries such as RAPL. The effectiveness of these heuristics is limited by the accuracy and frequency of the energy measurements. Since most work focuses on x86 Intel CPUs, work by [83] attempts to provide an extensive description of the core frequency transition delays, workload-based frequency limitations, and impacts of I/O die P-states on the performance of memory on the AMD Zen2 microarchitecture. The authors provide a set of guidelines for developers and system administrators to utilize these chips in the most efficient way. However, a key takeaway of this study is that AMD's RAPL should be avoided as it provides inaccurate energy measurements of the CPU. DVFS libraries should be avoided in these CPUs if they rely on RAPL, as inaccurate energy measurements will lead to less-than-ideal tuning of the chips.

These heuristics were used by [84] on the power management of Intel and AMD CPUs, proposing a library that dynamically adjusts GPU hardware states to minimize the average power draw over a set of kernel executions. The authors use statistical methods to extract kernel execution patterns within an application, attempting to predict the behavior of these kernels based on a history of previous. While the authors use a comprehensive list of benchmarks, from memory- to compute-bound code, with best case energy savings up to 25% for a 2% performance drop, the effectiveness of this approach is yet to be tested with real applications. Additionally, as seen in other approaches, the quality of the GPU tuning is dependent on the accuracy of the kernel predictive model. AI-based models could be used to improve the accuracy of the predicted behavior of the kernels, leading to a more efficient hardware configuration.

A framework for multi-objective optimization is proposed in [85], in which the goal is to minimize the power consumption of CPU devices while maintaining reasonable performance. It operates during application runtime, where it attempts to build an energy model for the workload and resources by applying several energy constraints through power capping (limiting the resources power draw, which is managed using DVFS) in an initial phase of the application execution. The power capping is adjusted dynamically to maintain adequate performance, based on the data gathered in the exploratory phase, to achieve significant energy improvements with minimal performance losses (up to 50% with execution times increasing less than 10% for most test cases). This work was extended to tackle NVidia GPUs in [86], using a similar approach, with percentile energy savings very close to the performance losses. However, this work does not tackle applications with highly irregular workloads, as the energy-performance model is only updated in the initial phase of an application execution.

In [87], the authors demonstrate that Fugaku possesses several power control features, and by efficiently coordinating these features with the application's characteristics, it is feasible to achieve superior energy efficiency at the system level.

4.3. HPC applications

Although users are usually unaware of the effects of executing their compute jobs, they should be responsible for their workload submission, and its characteristics. Therefore, users must be conscious of the scientific applications they plan to use, for example, by taking full advantage of parallelism available for specific applications, or by choosing the most efficient programming languages [88], libraries, and algorithms available for their work, a crucial point for developers.

In his experiment [89], Portegies Zwart executed an algorithm using nearly a dozen programming languages, and he discovered that Python, one of the most commonly used languages, required significantly more time to execute, thereby generating greater carbon emissions compared to languages like C++ and Fortran. This is related to the two-language problem, where many scientific codes are prototyped in a slow but flexible language (to test an idea quickly) but then have to be moved to a faster but less flexible language for practical, scalable, and optimized applications. However, with the increasing effort in compilers and better-optimized codes [90] it is possible to have faster and more sustainable Python codes than C++ and Fortran while being simpler to understand and use. Another approach is the use of more modern languages that try to tackle the two-language problem, such as Julia [91].

The minimization of energy consumption in parallel code through adequate scheduling of workloads has recently been studied. Work by [92] proposes a scheduling policy for work-stealing intra-application schedulers that minimizes energy consumption with a limited negative impact on code performance. While traditional schedulers attempt to predict the execution time of tasks to properly assign them to compute units, this approach predicts the energy consumption per task after an initial set of measurements. This approach provides energy savings up to 40% with negligible impact on performance on select benchmarks, but its applicability to irregular workloads is debatable.

The matrix chain multiplication on GPUs is fundamental for various scientific fields, such as computer graphics, physics, and machine learning. Although its time performance has been widely investigated, optimizing its energy efficiency has received less attention. In [93], the authors introduce transformations for energy-efficient accelerated chain matrix multiplication (TEE-ACM2), which can save up to 10% energy.

The simulation of multi-scale flows in weather and climate modeling poses a significant challenge in meeting time-to-solution requirements while adhering to energy budgets, without compromising the application's accuracy and stability. The ESCAPE project [94] identified algorithmic motifs and developed prototype implementations on different hardware architectures with varying programming models. The project's resulting energy and time-to-solution measurements mean that there should be a focus on utilizing all accessible resources in hybrid CPU–GPU arrangements.

The improvement of the energy consumption of applications may often require a complete redesign of their architecture. The authors of [95] present such a case, where a stencil-based CFD code had to be modified significantly to adequately use the memory hierarchy while avoiding excessive synchronizations among CCDs in the AMD Rome microarchitecture. The authors show that a data-centric approach to the code design of memory-bound applications can significantly improve performance and reduce energy consumption, by factors of 9×-10× for the test case presented. A data-centric approach was also employed by [96], where transformer inference was optimized for ARM CPUs, resulting in performance improvements of up to 8× while maintaining the same energy consumption.

Table 3
Energy and computational efficiency metrics of recent supercomputers when executing synthetic benchmarks.

Name	Year	Center	GPUs per node	Top500 (MW)	HPL (PF/s)	HPL (GF/W)	HPCG (TF/s)	HPCG (GF/W)	Avg. Power (MW)
K	2012	Riken	0	12.7	10.5	0.8	602	0.05	12
Titan	2012	ORNL	1	9	27	3			
Summit	2017	ORNL	3	10.1	148.6	14.7	2926	0.29	5-6
SuperMUC-NG	2018	LRZ	0	3	19.5	6.5	207.8	0.07	2
Frontera	2019	TACC	0		23.5	3.9			6
Fugaku	2019	Riken	0	29.9	442	15.4	16 004	0.53	21
LUMI	2021	CSC	4	6	309	51.4	3408	0.57	
LUMI-C	2021	CSC	0	1.2	6.3	5.2	103	0.09	
CEA-HF	2021	CEA	0	4.9	23.2	4.7	341	0.07	
Frontier	2022	ORNL	4	22.7	1194	52.2	14 054	0.67	
Leonardo	2022	Cineca	4	5.6	175	32	2566	0.46	
Henri	2022	Flatiron Inst.	8	0.031	2	65.1			

4.4. Discussion

Table 3 lists metrics related to energy and computational efficiency of several recent supercomputers. The values align with the latest TOP500 list at the time of writing [9]. The latest column reports the average power of the machine in normal usage, running users' HPC applications.

Energy efficiency has been greatly improved (x10 over the last ten years) but at the expense of greater system heterogeneity with the generalization of accelerators (e.g., GPU) and the increasing ratio of GPUs per node (8 in the current Green 500 top supercomputer, Henri). A look at the LUMI supercomputer also presents this behavior, as shown in Table 3: LUMI-C, a subcluster using CPUs only, is around 10x less efficient in terms of floating point operations per watt in the High-Performance Linpack (HPL) and High Performance Conjugate Gradients (HPCG) benchmarks than LUMI, which uses 4 GPUs per server.

As previously stated, Fugaku has various power control features, and for the Top 500 list, they use their Boost mode, while Normal mode was used for Green 500. The result is that the HPL benchmark is 10% slower but consumes 4% less energy. All jobs executed in LRZ SuperMUC NG under BSC Energy Aware Runtime (EAR) [57] report that, when using a `MIN_ENERGY_TO_SOLUTION` policy while executing memory intensive applications, a reduction of energy consumption of up to 8% can be obtained. The use of the READEX Tool Suite [97] on a set of applications and benchmarks, which are parallelized with OpenMP and/or MPI, achieves energy savings of up to 34% by applying hardware and runtime parameter tuning.

As we can see in Table 3, the average power of the machine in normal usage, running users' HPC applications, is considerably less than the peak power running HPL benchmark for TOP 500. As such, there is an ongoing discussion on the real meaning of the Green500 metric, as it is based on running a synthetic benchmark, HPL, that is not representative of most modern HPC applications. Moreover, several modern supercomputers during normal usage, apart from periods to run benchmarks for acceptance and Top500, use energy restrictions tools to limit energy consumption with minimum performance degradation.

The HPL is a highly optimized library for linear algebra computation used as a benchmark by the Top500, as it is capable of reaching close to the theoretical performance cap of most computing resources. However, this library is so compute-bound that it does not represent the type of applications typically executed in HPC systems. Since discussions around the real-world value of HPL have become more prominent in the field [98], HPCG was introduced as a benchmark more representative of real HPC workloads due to its memory-bound behavior. The difference between these benchmarks is evident by looking at peak performance and flops/Watt in Table 3. The large ratio of flops/byte of HPL is translated into a high flop/watt when compared to HPCG. However, these peak values are not representative of the normal usage of a HPC center, with the real performance and energy consumption being more inline with the measurements from HPCG. Finally, these benchmarks consider utilizing all the computing resources of a supercomputer, leading to an inflated energy consumption of the system. Such high resource utilization very rarely happens in these systems.

Developers must consider the energy efficiency requirements of an application from its design phase. The power draw of an application depends on the programming languages used, the efficiency of third-party libraries, parallelization mechanisms, workload schedulers, and the know-how of the underlying hardware architecture. Authors of [68] showed that even within the same library, different parallelization mechanisms that achieve similar goals may have very different power draws.

Most job/task scheduling techniques presented in the previous subsections to minimize the system power draw, either at the hardware or software level, rely on measuring the energy consumption of workloads to model the scheduling of tasks. Measuring the power draw of the whole system is unfeasible as it requires specialized hardware, which means that these approaches rely on third-party libraries, being RAPL the most popular. However, as indicated in Section 4.2.1, RAPL does not have access to the same components across servers (CPU, memory, storage, etc.), and the information reported in several CPU architectures is inaccurate, reducing the precision of the predictions and leading to sub-optimal workload scheduling. The task scheduling is performed by optimizing a multi-objective function that combines a minimization of the execution time and energy consumption. Several multi-objective metrics have been proposed, from which the Energy Delay Product (EDP) group of metrics [99], originally used in energy optimization of electronic circuits, are the most popular. EDP combines the energy consumption E over the execution time D of a task by multiplying these two factors (ExD). This metric considers E and D to be equally important, but a generalization has been proposed to increase the weight of the execution time by n (ExD^n). Most energy-aware task schedulers, especially on in-application scheduling, use a variation of EDP, where the weight of E and D is adjusted according to the target use case.

In the topic of energy management, there is a trend to deal with energy and computing resource management in a more coordinated or cooperative way. Many authors attempt to use workload shifting through job scheduling policies and algorithms to reduce energy costs, and enhance the integration of RES or minimize the indirect carbon emissions, in which the recent carbon-aware computing concept stands out. Carbon intensity data is becoming more valuable, with clear applications in the computing sector. It is important, however, to be aware of specific HPC characteristics (high utilization rates, resource management capabilities, SLAs) that could limit the applicability or effectiveness of these techniques.

5. HPC services, business models and policy

This section addresses the digital services provided by HPC centers, from both the perspective of the service provider and the user. Furthermore, an overview of the recent initiatives on policy and standardization and its reflection in funding opportunities and research projects is made, as well as on standard and cross-sector business models that link computing and energy resources.

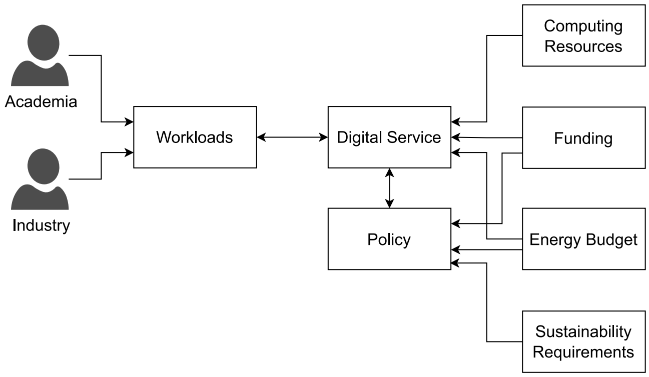

Fig. 3. The factors and relationships that affect the Digital Service provided by HPC centers.

5.1. Digital service

HPC centers generally provide access to high-performing computational resources and expertise to academia, industry, and government users, with the goal of accelerating their research activities. The services provided usually range from access to high-performance computing, storage, and efficient software, technical and consulting support, and training in the use of computing resources. The type of digital service provided varies among HPC centers and depends on their infrastructure, policies, and business models.

Fig. 3 illustrates the relations among the different factors, both user and infrastructure-driven, that influence the digital service provided by the HPC centers. The type and properties of the workloads, such as complexity, limitations, and requirements of specific hardware (such as GPU acceleration), directly impact their performance in HPC systems. The architecture and computing resources of supercomputers must be designed to accommodate the needs of the target workloads, and in turn also affect the service provided. The quality and quantity of these resources determine the capacity of the system to handle a large number of users and workloads, and, consequently, the availability and performance of the service. Funding is possibly the most influential factor in the digital service provided by HPC centers. Inadequate funding leads to less performing and/or specialized hardware and understaffing, which decreases the quality and availability of the provided service.

5.1.1. Service providers

The work in [100] concluded that although carbon emissions are already publicly reported by many entities, there is still room to improve on carbon accountability and reporting to develop fully sustainable computing systems. The authors also state that researchers and developers should make carbon footprint a first-class design metric.

Actions that aim to improve the sustainability of a computing center have to account for an important aspect: a computing center provides crucial digital services to its clients. This means that service providers usually have a commitment with their users, in which they can expect a certain quality and availability when using the requested service, by offering SLAs with legal implications [101]. By managing resources to assure that SLAs are fulfilled, strategies for sustainability can be somehow restricted in their implementation. Authors of [101] discuss the advantages of negotiating SLAs in the HPC environment and the effect it can have on both users and service providers, while also addressing the differences between per-job and long-term SLAs. Another example of the specificity of SLAs for HPC is long-running jobs, which could mean that users could require a policy for guaranteed job completion [102].

Regarding sustainability, the concept of Green Service-Level Agreements (Green SLAs) [103,104] is a way for users themselves to guarantee that their computing workload execution meets some kind of sustainability-related performance indicator. This can be achieved, for example, by establishing a minimum percentage of RES when executing a certain workload.

5.1.2. Users

Users of HPC resources are mostly composed of researchers and industry practitioners in certain fields of science. The research community is largely dependent on computing systems for the development of their own scientific work. Typically, researchers take advantage of large-scale resources available in HPC or traditional data centers, either on-premise or in cloud service providers. The common steps for an HPC user using the service are the following: Copy data (copying their data into the cluster storage), Preparation (compiling specific applications/modules, and creating a script to automate the execution), Execution (submitting the computational pipeline to the cluster for batch execution), Analysis (analyzing output data locally and, in most systems, interactive sessions for analysis and visualization), and Retrieval of data (retrieving the relevant results).

Researchers and other HPC users usually have little to no visibility or concerns on the impact that their work has on power consumption [105], and the consequent operational and environmental costs. Furthermore, their carbon footprint goes well beyond their computing needs and extends to their overall research and working practices. Ten simple rules to incorporate sustainable research in the agenda of the scientific community are indicated by [106] and include efforts for remote participation, avoiding duplicate work, and efficient carbon footprint reporting. Other contributions include the adoption of open science practices [107].

These facts are even more relevant in fields of science that show a high demand for computing resources for their core research, such as bioinformatics, computational fluid dynamics, and AI, among others. Many fields that intensively use resources and, therefore, contribute to climate change, are also active contributors to solving challenges in the energy sector through their scientific output. Fields related to weather forecasting and resource assessment for different energy sources, simulation, and development of advanced materials for innovative technologies, and modeling and optimization of energy systems are an example of this paradox [108].

The recent advances in AI led to important contributions to society, but with the cost of training increasingly large models that need specialized hardware, and large amounts of computing power for extended periods of time. The need for more powerful solutions than single-GPU setups has become evident. As a result, there has been an increasing convergence of AI and HPC, with the use of HPC systems for developing and deploying accelerated AI algorithms in both academia and industry settings [109]. This field has grown awareness of its carbon footprint, with emerging concepts such as Green AI [110]. The mentioned work focuses on computational cost and efficiency as an evaluation criterion in a field dominated by accuracy measures and recommends floating-point operations as a standard metric while assessing other possible metrics. Although the impact of AI in carbon footprint is indisputable, work by [111] recently proved that several authors have been overestimating carbon emissions of machine learning workloads, and described some best practices in the field. For machine learning projects, tools like CodeCarbon allow carbon accounting and reporting, improving reproducibility and avoiding duplicate work.

The carbon footprint of computing systems can be improved if users are aware of the impact of their work and if they feel motivated to participate in behavioral change through incentives. This depends on how computing and resource management systems are designed, how these systems interact with users, and how providers develop their computing ecosystems and billing processes. Awareness of environmental sustainability can be enhanced by informing users of how computing impacts power consumption in different periods. However, even if HPC users want to run their applications in favorable periods, they have very limited possibilities, as resource managers are the responsible entities for allocating resources and scheduling the execution of a specific workload.

Nonetheless, users can be targeted with incentive mechanisms that aim to optimize the use of resources or to minimize the impact of its utilization, by providing more details to the resource management system or by changing behaviors that favor the operation of the system. For example, the submission of workload could include the flexibility that users have regarding their deadline, and their runtime estimates. This would allow shifting the execution of workload to periods that contribute towards optimizing a specific objective for the computing center, and users could be compensated.

Incentive mechanisms can be implemented using credits or direct discounts when billing the use of resources in large computing systems. In [112], a parameterized model was designed to assess the changes in energy consumption as a result of frequency scaling techniques and to evaluate if the benefits for the HPC facility and its users were reasonable. The authors evaluated multiple pricing schemes, concluding that novel pricing schemes and energy accounting tools for users are needed, with promising preliminary results. In [113], the authors propose EnergyFairShare (EFS), which manages the energy budget of a supercomputer by prioritizing users with less energy-intensive computing jobs. This encourages users to target energy efficiency when developing code, ultimately leading them to reduced queuing and turn-around times.

Moreover, incentive mechanisms can contribute to the systematic reporting of sustainability metrics (such as carbon footprint) in scientific publishing, with journals partially waiving publishing fees, or granting some certification or badging for these publications. It is important to track the contribution of computational research to climate change to stimulate greener algorithms. Green Algorithms [114] is an open-source platform to calculate the estimated carbon footprint of scientific large-scale computation with a simple methodology considering the characteristics of different resources that require energy (processor, memory, overhead of computing facilities) and geographical location. The authors stated that scientists are usually unaware of their carbon footprint and indicated that the main challenge is to perceive the reporting of sustainability metrics as a prevailing practice, which could be included in scientific publications. Other tools and frameworks exist but are usually directed to individual computing.

These interaction mechanisms between providers and users can be limited by many factors. When implementing energy efficiency and management strategies, there is the possibility that the overall service provided is being affected via job scheduling, job-level management, or resource-limiting strategies, which can ultimately impact users' job deadlines or wall time. Therefore, sustainable practices can affect the overall performance, deteriorate the quality of service, or impact SLAs. If Green SLAs are deployed, then energy management strategies that aim to fulfill those SLAs can be designed. Furthermore, HPC infrastructures and their resources can be shared or preallocated for specific users, projects or entities. This means that there can be a part of the physical resources (e.g. a certain percentage of nodes) that are dedicated to an entity or that have to be available in specific periods or even different queues available for submission. Also, knowledge of the implemented scheduling policies is essential, as priorities on job scheduling can be established. This not only affects the flexibility and effectiveness of resource management strategies, but also the utilization rates and overall efficiency of the infrastructure.

5.2. Policy and recent initiatives

In recent years, multiple efforts for policies and standardization on sustainability measures in computing centers have been made by regulators and international standard organizations.

The European Green Deal stated that data center is one of the sectors where energy efficiency and circular economy measures will be implemented, while the European Union Digital Strategy mentioned the goal to achieve carbon-neutrality in this industry by 2030 [115]. The European Processor Initiative deals with implementing a roadmap for low-power European processors for extreme-scale computing, and other emerging applications.

The EuroHPC initiative, a joint initiative to develop the supercomputing ecosystem in Europe, demanded [116] that the design and operation of supercomputers supported by the Joint Undertaking must consider plans regarding energy efficiency and environmental sustainability. The Partnership for Advanced Computing in Europe (PRACE) [117] offers computing resources and services to foster scientific discovery and engineering research in Europe, and indicates that one of its goals is to reduce the energy consumption and environmental impact of the European HPC ecosystem, by developing tools and training users to adapt to technological changes in this field. The European Technology Platform for High Performance Computing (ETP4HPC) [118] includes in its structure a Working Group on Energy Efficiency that focuses on the global-level approach to the energy efficiency of HPC systems and regularly organizes workshops on this topic.

Other initiatives include the Energy Efficient HPC Working Group [119], with funding from the US Federal Energy Management Program, created to promote energy-efficient computing guidelines and to improve the design and operation of HPC systems regarding its energy performance, and a yearly document released by the Joint Research Center on the Best Practice Guidelines for the EU Code of Conduct for Energy Efficiency in Data Centers [120], to assist operators in implementing measures to improve energy efficiency in their facilities. The Master List of Energy Efficiency Actions, provided by the Center of Expertise for Energy Efficiency in Data Centers, is a document that also lists some best practice recommendations to address data center energy efficiency, mainly as a reference and for guiding operators [121].

A report [122] by the Center on Regulation in Europe focuses on regulation and policy recommendations regarding the participation of data centers in the European energy system, due to the rising computing needs in the context of the European Climate Law. According to the authors, energy efficiency improvements in recent years were able to partially mitigate the effects of the increasing computing demand on the electricity demand of this sector, and future energy consumption trends are uncertain. Regulations specifically for data centers are likely to occur with respect to energy efficiency and its overall integration in the energy system (providers of energy flexibility, integration with district heating networks), and current directives such as Ecodesign, Energy Efficiency, EU Green Public Procurement, and Energy Taxation Directive should be leveraged to encourage harmonization and be the basis for further regulation on sustainable practices that contribute to the European energy transition. The authors conclude that a dynamic regulatory approach should be preferred for data centers, with a combination of legal instruments, such as standards, guidance, legislation, and self-regulation mechanisms.

A recent report [123] argues that sustainability requirements for computing centers are mostly voluntary and that they will probably become mandatory in the near future. Also, the report indicates that operators should start preparing for net-zero commitments, while offsetting mechanisms and RES certificates are becoming less acceptable in favor of on-site RES generation.

Efforts for standardization are also increasing for computing centers. The series of standards ISO/IEC 30134 discusses the need for KPIs specifically for data centers and their standardization, while the ANSI/ASHRAE standard 90.4-2019 established the minimum energy efficiency requirements for the design and operation of data centers, including the use of RES. The ASHRAE Technical Committee 9.9 is also concerned with data centers and other technology spaces and facilities. Leadership in Energy and Environmental Design (LEED), established by the US Green Building Council, represents a set of rating systems (Silver, Gold, and Platinum) for the design, construction, operation, and maintenance of green buildings. Achieving LEED for data centers is difficult and thus LEED data centers are surprisingly rare, with fewer than 5% of all US data centers with LEED certification.

Table 4
Research projects.

Project (Year)	Scope	Outcomes	Ref.
HEROES (2021–2023)	Aims at developing an innovative platform to allow end users to submit their complex simulation and ML workflows to both HPC and Cloud data centers.	Possibility to choose the best option to achieve their goals in time, within budget, and with the best energy efficiency.	[127]
Montblanc (2017–2021)	Low-power processors for exascale HPC based on European technology.	Developed IP for low-power servers, methodologies for processor simulation, and virtual prototyping. The second generation of processors was planned for 2022 with a power efficiency of 50 GFlops/W.	[128]
ESCAPE (2015–2018)	Energy-efficient SCalable Algorithms for Weather Prediction at Exascale.	By modifying numerical algorithms and using new programming models, substantial improvements to weather and climate predictions were possible and affordable.	[129]
READEX (2015–2018)	Increasing energy efficiency in exascale HPC systems by optimizing available resources and adjusting system parameters to specific application requirements.	Built software packages and a run-time system using an automatic optimization approach that leverages the dynamic behavior of HPC applications to switch parameter configurations on exascale systems.	[97]
ANTAREX (2015–2018)	Solving challenges of exascale computers to increase energy efficiency by implementing a holistic approach spanning all decision layers of software stack management.	Provided a Domain Specific Language, libraries, and dynamic autotuning frameworks for runtime management of applications.	[130]
ECOSCALE (2015–2019)	Improving the performance of exascale computers to meet energy efficiency goals, relying on scalable programming environments and hardware architecture tailored to future HPC applications.	Provided an energy-efficient architecture, programming model, and runtime system, which resulted in a working prototype and simulator to run real-world HPC applications.	[131]
ADEPT (2013–2016)	Address the challenge of energy-efficient use of parallel technologies.	Integrating performance and energy consumption modeling for HPC and embedded systems. Specialized benchmarks have been developed to provide detailed insights into how systems utilize energy and power.	[132]
EXA2GREEN (2012–2015)	Energy-Aware Sustainable Computing on Future Technology – Paving the Road to Exascale Computing	A software tool able to trace and analyze the power and energy consumption of parallel scientific applications and energy-efficient algorithms. A new type of internal power meter for individual hardware components (CPU or memory).	[133]

5.3. Funding opportunities and research projects

The effect of recent policy and decision-making is also reflected in funding opportunities and research projects. For example, the European Union decided to address the challenges of exascale [124], and has been funding research initiatives in the last decade related to hardware design, software, and other key areas. Table 4 describes the scope and outcomes of funded research projects related to the review in the last decade. It is clear that most projects dealt with energy efficiency in the last decade, in particular, to address the many challenges posed by the exascale paradigm.

Forming part of the United States' Exascale Computing Project, the HPC PowerStack collaboration, joining experts from academia, research laboratories, and industry, aims to design a holistic, extensible power management framework [125]. The project aims to develop a cross-pillar system for power management within HPC that utilizes intelligent techniques to improve decisions related to scheduling, hardware and software.

The Strategic Research Agenda from ETP4HPC [15] contains the priorities of European research in HPC technology and identifies two new challenges for HPC, one of them being Sustainability. As the authors state, while HPC center capacities constitute only a subset of the entire data center, they are also growing. Therefore, actions are needed to reduce their carbon footprint, reduce their use of rare materials for the production of hardware components, increase the lifetime of systems, reduce electronic waste, and introduce the scheme of a circular economy.

EuroHPC JU has an open call, HORIZON-EUROHPC-JU-2022-TECH-03 [126], that aims to create a large-scale European initiative for the HPC ecosystem, with one of the expected outcomes being the development of energy-efficient high-end processors and accelerators utilizing RISC-V components.

5.4. Business models

Traditionally, HPC business models focus on offering services related to users accessing its computing resources. This section deals with innovative business model opportunities that can arise within the HPC ecosystem. A report by [134] on the future of supercomputing concluded that, although the public sector (universities, research centers) is the main driver for financing and usage of HPC, centers are creating new revenue streams by broadening their scope to commercial and industrial users. This recent approach accelerated access to HPC resources in many industries, contributing to the digitization of the economy, and fostering innovative business models that enhance the financial sustainability of HPC centers.

As previously stated, the energy and computing sectors are increasingly interdependent. In HPC, that dependency is visible because it directly impacts costs and carbon footprint, as available computing power and electricity costs rise. Moreover, the HPC industry deals with thermal energy and large volumes of data, which can offer new opportunities. Therefore, innovative business models that include both energy and computing resources can be relevant to the environmental and financial sustainability of computing infrastructures.

5.4.1. Computing

HPC cloud computing or HPC as a service (HPCaaS) uses cloud resources to execute HPC applications [105], and the adoption of these solutions is rising, as companies realize that they can have the same computing service without the CapEx and complexity associated with investing in and managing a whole infrastructure, with the ability to adapt their computing capacity to their needs easily over time. Work by [105] reported that the adoption of these services is more dependent on the financial sustainability of moving applications to the cloud (instead of maintaining a private HPC infrastructure) and the type of HPC workload to execute (due to possible scalability issues in the cloud). Multiple hyperscale providers (Google, and Amazon Web

Services, among others) exist and are robust solutions to institutions that need large computing resources [135,136].

HPC resources can also be traded in decentralized peer-to-peer networks, and even in marketplaces based on blockchain technology. The iExec project [137], for example, offers a platform where companies and individuals can make available IT resources in exchange for tokens for executing tasks for applications. Hypernet Labs [138] is implementing a decentralized computing marketplace where the first resource provider is an HPC system from Exaion, provided with low-carbon electricity.

Ultimately, HPC business models can be characterized by the way that its end users pay for access to computing resources. In the specific case of academia, a survey by [139] found roughly four (library, shareholder, cost center, and industry collaboration) that operate with different revenue streams.

5.4.2. Cross-sector opportunities

The location of a HPC center is key to the overall carbon footprint associated with the offered service, due to regional differences in weather conditions and energy mix available [140]. By taking advantage of sites with better conditions for free cooling and availability of RES, there is an increasing offer of sustainable HPC services in those sites, such as [141] and [142].

One of the most addressed business models for data centers in the literature is energy flexibility as a service. Extensive work is available in assessing data center energy flexibility and their potential for participating in demand-side management programs. The spatio-temporal workload shifting techniques can be seen as implementing demand response [16]. Work developed by [143] created typologies for data centers' business types to identify candidates for energy flexibility services. Demand-response models for HPC systems with job scheduling and resource provisioning schemes were developed by [144]. The authors of [145] proposed a framework to optimize power flexibility on an HPC system in different flexibility markets in Germany. As the proportion of renewable energy sources in electric power systems grows, it becomes more necessary for electrical grids to work together to balance supply and demand. To address this issue, the authors of [146] suggest using a straightforward site-wide power model, including both server and cooling, combined with QoS-aware demand response techniques, which increased cost savings by around ×1.3.

Revenue streams considering energy flexibility were indicated by [147], as data centers can simultaneously participate in electrical, thermal, and data networks. It stated that the potential of these facilities comes from the high energy consumption, high automation levels, and the nature of their hardware and workload. The authors considered the capture of waste heat, thermal storage, integration of RES, energy storage, workload shifting, and participation in demand response programs to provide optimization techniques that exploit a data center's energy flexibility.

Due to its high energy consumption and need for a reliable source of electricity, HPC systems can be partnered with their electricity service providers. In [148], the authors surveyed the top supercomputer facilities in the United States and concluded on the importance of engaging in demand response programs and providing demand forecasts to increase the reliability of the electrical grid. They identified job scheduling as the most interesting management strategy and summarized current challenges and opportunities regarding the potential interactions between HPC operators and their electricity service providers, and their participation in electricity markets. A qualitative study of service contracts between these two agents was done by [149] to identify imposed strategies in HPC (such as demand charges, demand response programs, and variable tariffs) and to conclude on current and future collaboration between entities. Lancium [150] provides services to the electrical grid and adopted the concept of carbon negative computing. By leveraging existing RES power plants, the company builds data centers that act as controllable loads near critical transmission system points usually overwhelmed by excess RES generation, to provide ancillary services to grid operators.

Another possibility is the integration of an HPC infrastructure within the microgrid concept. Microgrid capabilities include the management of on-site generation of electrical and thermal energy, engaging in demand response programs, selling electricity or ancillary services to the grid, and islanding to improve computing centers' vulnerability to power outages [151]. Deploying data centers within a microgrid can be the culmination of an integrated energy management strategy to reach low PUE values and create new revenue streams with enhanced reliability [144].

The revised Renewable Energy Directive (2018/2001/EU) and the Electricity Market Directive (2019/944/EU) required EU member states to provide frameworks that enable and promote the development of RES and citizen energy communities [152], which aim to provide social, environmental, and economic benefits to their participants. In the context of HPC, business models can be designed so that the infrastructure is an active member of a local energy community as a prosumer (with computing and energy resources).

The thermal energy available as waste heat, a byproduct from the cooling process of the HPC servers, can be interpreted as a possible source of revenue. In Telia Helsinki Data Center [153], the waste heat produced by the servers as a byproduct is processed with heat pumps, metered, and billed according to open district heating network pricing.

Lake Parime [154], a digital infrastructure company, provides a new solution (Powerbox) for both the HPC and RES sectors. The company works with RES producers to build HPC infrastructures on-site that take advantage of the surplus RES-based electricity to offer computing services for HPC application users. The company indicates that this solution avoids upfront costs of storage devices and ensures new revenue streams, as earnings per MWh of computing services are decoupled from electricity market prices. A similar solution is proposed by Soluna [155], which builds small-footprint data centers to buy and use surplus RES generation in data centers that specialize in performing "batchable" computing.

Another cross-sector opportunity is the actual use of HPC resources in the energy industry, contributing to society's overall decarbonization. The use of computational fluid dynamics in wind resource assessment, the modeling and optimization of energy systems, and the model training for RES forecasting are some computing-intensive applications and thus good candidates for using HPC as a backbone to generate revenue [156,157]. Furthermore, the processing burden from monitoring and controlling energy infrastructures, such as smart grids, can be a good match for edge HPC solutions. Recently, the U.S. Department of Energy developed the HPC4Energy Innovation initiative, which provides funding to industry partners to use HPC resources to advance their national energy agenda.

There are also opportunities in the data industry. Gaia-X [158] is an initiative to create the next generation of open-source data infrastructure in Europe concerned with reaching high standards of digital sovereignty, where different use cases are reported for the energy sector. A case study for edge data centers reports how the digital and energy sectors can be coupled by using surplus energy to generate computing power and a new revenue stream. Specifically for HPC, another case study proposes HPC as a service, as many potential users do not have the means to invest in such a system, with a sharing model. HPC services can thus be accessible to new users (academia, companies, temporary users) and face higher utilization rates, while users remain in control of their data.

A possible service resulting from integrating RES in HPC centers is the emission of certificates of origin for the energy used in computing tasks, an extension of the concept of certificates of origin. HPC users could be aware of or even publish the share of RES or carbon footprint associated with their scientific work. A scaled and automated version of a tool similar to Green Algorithms (discussed in Section 5.1.2) directed to HPC centers instead of individual computing could be a service provided to multiple HPC centers, to automatically send traceable certificates of origin after a computing job, to increase the awareness of the scientific community in this matter while providing an innovative service.

5.5. Discussion

A review of the HPC ecosystem shows recent trends for policy, regulation, standardization, and other efforts specifically targeting HPC and data centers, including new agents and entities. Policy-making processes led to research projects and funding in the last decade being dominated by concerns about energy efficiency within the exascale paradigm. The sector is broadening its user scope, with initiatives to establish partnerships that join academia and industry to drive the digitization of the economy. The sector is also broadening its scope of applications, for example, with an increasing trend of using HPC platforms to perform research related to AI, but also to incorporate initiatives such as Green AI.

Enhanced interaction between service providers and users of HPC systems can help operators optimize the energy and computing resources of the facility while making users aware of the impact of their workloads and being more active participants in the sector's decarbonization. However, these strategies must consider that HPC services are less flexible than those of traditional data centers, due to the nature of the workloads and of the service itself. Furthermore, innovative strategies could be developed to encourage the adoption of techniques that balance power consumption with performance since the current accounting and billing systems are unsuitable for this purpose.

Recent examples of business models beyond HPC as a service and cloud computing show a growing interdependency between the energy and computing sectors, and their progress and concerns, which can foster cross-sector opportunities that either leverage HPC to generate new revenue streams, or further integrate these centers in energy systems.

Overall, the HPC industry is moving towards offering a more environmentally sustainable service, by taking advantage of sites with better conditions for free cooling and availability of renewable energies, which also allows to reduce energy costs. Although not discussed in this section, the service provided by an HPC facility is limited to the system's lifetime. Current HPC systems are typically in service for five years [159], and a discussion on their lifetime could set the trend for its expansion (instead of consecutively replacing them with more recent and powerful machines, even if having an improved power efficiency), or for a second-life application. Recent work by [160] pursued this line of research, designing a framework for carbon footprint analysis in HPC systems, considering operational but embodied carbon footprint due to the production of different hardware components. Although very important to enhance performance, the authors found that hardware upgrades can increase the embodied carbon footprint and, depending on the center's conditions, can be hard to offset. Therefore, the authors argue that *"extending the hardware lifetime could be a worthy option"*. Furthermore, a lifecycle analysis in this sector could reflect the concerns of reducing electronic waste, and introduce the circular economy concept. For example, green procurement practices could be leveraged, to address the sustainability concerns of manufacturing processes, transport, and other steps of developing an HPC system from its design phase.

6. Key performance indicators

The performance of computing centers is ultimately assessed by a wide range of Key Performance Indicator (KPI). These metrics support the reporting of performance and overall use and management of resources at many levels. Furthermore, metrics allow to some extent a comparison between similar infrastructures and can serve as a goal when designing a new computing site, and the reporting effectiveness is largely dependent on the ability to accurately monitor multiple resources in the facility.

6.1. Overview of metrics

While KPIs must assess the performance of the digital service provided, these are usually associated with individual systems such as cooling, IT, and their interdependency. In the HPC ecosystem, power efficiency (GFLOPS/W) relates computing performance to its power consumption, and while relevant to ranking the most efficient computers in the Green500, fails to address the impact of the whole infrastructure with many shortcomings. Historically, PUE has been the standard metric reported when approaching energy efficiency, to account for overhead energy consumption. However, many limitations are pointed out, such as [161] lack of guidelines for calculation, lowering PUE has become a goal itself instead of lowering energy consumption, further improvements beyond 1.1 are difficult and disregard for the source of energy, on-site energy generation, or waste heat recovery. Some PUE values for state-of-the-art HPC centers are available in Table 5.

Therefore, there is an ongoing discussion on what KPIs best suit modern and sustainable computing centers [161,162] and efforts to improve their reporting [163]. Discussions are now centered on how to better reflect topics such as RES and energy storage integration, carbon footprint, use of resources beyond electrical energy (e.g. water), how waste heat is used, among others, and how to combine them for more robust performance metrics. Some KPIs that already address some of these concerns are, for example, Carbon Usage Effectiveness (CUE), Energy Reuse Effectiveness (ERE), and Water Usage Effectiveness (WUE).

A thorough review of available KPIs for sustainable computing centers is made by [164], where metrics were distinguished for different dimensions of the system: energy efficiency, cooling, performance, greenness, thermal/air management, network, storage, but also security and financial impact. An extensive gathering of KPIs available for data centers was also done by [165], which identified the name and corresponding promoter of each metric, and stated that existing metrics fail to have a holistic view of the data center operation. As a result, the authors presented a multidimensional approach considering productivity, efficiency, sustainability, and operations, alongside risk, normalized and weighted for efficient visualization by means of a scorecard. Work by [166] proposed a holistic performance assessment of data centers, where sustainability is seen not only as the result of environmental protection but also as a result of economic, operational, and social longevity.

Multiple entities are contributing towards a more standardized environment, including defining KPIs and detailing measurement points and techniques for a correct calculation. The series of standards ISO/IEC 30134 discusses the need for KPIs specifically for data centers and their standardization, regarding the effective use of resources and the reduction of CO_2 emissions, which covers the definitions of PUE, Renewable Energy Factor (REF) and other metrics. Recently the Energy Reuse Factor (ERF) was included (ISO/IEC 30134-6:2021), as the ratio between reused energy and the data center's total energy consumption. ISO 50001 is an international standard for managing an organization's energy performance. Lawrence Berkeley National Laboratory has implemented ISO 50001 as a way to ensure its energy and water management activities and efficiency savings are strategic, effective, and persistent.

Performance can be related to goals for which the system is being optimized, according to some energy management strategies. This includes energy, environment, financial, and computing-related metrics. For example, CO_2 emissions, self-consumption, or self-sufficiency rate can be important energy/environmental goals for HPC systems to focus on, while savings in electricity costs can be a financial one.

Computing-related metrics are key to assessing the trade-off between optimizing energy resources and the performance/utilization of the actual system or even the digital service provided (including user-centric performance metrics). Some examples are the system utilization rate and job wait time [167], job wall time, time-to-solution, energy-to-solution, throughput, fulfillment of SLAs, and availability, among others.

Table 5
Trends in PUE of high-performance computing centers.

	K	Titan	Summit	SuperMUC-NG	ESIF	Fugaku	LUMI	CEA-HF	Frontier	Leonardo
Year	2012	2012	2017	2018	2019	2019	2021	2021	2022	2022
PUE	1.20–1.27	1.29	1.1	1.08	1.04	1.1	1.04	1.1	1.03	1.08

6.2. KPI monitoring and optimization

Operational data analytics (ODA) systems enable the acquisition of real-time information in computing centers for enhanced decision-making. The infrastructure, as indicated by [50], facilitates the continuous monitoring, storage, and analysis of performance data from a machine and infrastructure level. This accumulated data is exploited to optimize system operations and, ultimately, improve KPIs.

In the following paragraph a brief description of successful applications of ODA is made. The use of ONNI in meeting organizational energy efficiency performance goals for the National Energy Research Scientific Computing Center's HPC cooling systems is summarized in [168]. The system collects and archives environmental and performance data from IT equipment, sensors, and devices on the HPC floor, allowing optimization of both hardware configuration and cooling. The system architecture design and current state of the operational data collection/monitoring platform being used in Fugaku is reported in [169]. Another example is the recent generic framework to facilitate real-time ODA on extensive HPC facilities described in [170], providing numerous configuration options to meet the diverse needs of ODA applications. This framework is built atop the holistic DCDB monitoring system [171], a scalable and modular monitoring system that can integrate data from all system levels in a distributed NoSQL data store.

6.3. Discussion

Table 5 provides a list of recent supercomputers and information on their KPIs to support the discussion. A set of representative centers was chosen, which allows analyzing the evolution of PUE in recent years.

The recent improvements in technology and the adoption of energy efficiency measures have led to a decrease in the PUE values of HPC centers. Recent values are near 1.0, which is a significant improvement from the PUE values of 1.7–2.0 that were common in HPC centers just a few years ago.

Typically reported KPIs (PUE, power efficiency) are now perceived as limited, as computing facilities grow in complexity and scale, and concerns on sustainability rise in an industry that aims to provide the greatest computing capability possible at any time. Metrics that better reflect the concerns of modern and sustainable computing centers (social, economic, technological, environmental) are being discussed and developed, with data centers heading towards a more standardized and accountable ecosystem on this topic.

7. Conclusions

The HPC ecosystem is expanding with the deployment of new large-scale supercomputers in recent years (with power draw in the dozens of MW), but also with new agents and entities that actively participate in innovating, financing, and scaling the usage of HPC for academia and different industry sectors, boosting the digitization of economy and society, preparing the exascale paradigm and solving its challenges (including energy-related). More entities are focused specifically in promoting best practices and guidelines, to assist and provide training for centers and operators in enhancing decarbonization in HPC design and operation. The approach considered in this review showed that multiple considerations and improvements on decarbonization are possible at each level of the HPC service, facility, and its basic components.

As ICT in general, and HPC in particular, gets more accessible and widens their user scope (namely from academia to industry), the usage of scientific applications within a HPC environment is set to increase, with researchers solving more and more complex problems with supercomputers. It was also shown that users, and not only service providers, actively affect the environmental impact of such systems, through the submission of jobs, or when designing and writing code for the executed applications. Therefore, there is a need for improved mechanisms that enable interaction between providers and users to optimize resource usage, and that target specifically HPC, due to additional constraints that traditional data centers do not experience.

The survey on industry trends revealed that, although computing power has been increasing, current concerns about energy usage go well beyond energy efficiency gains at the hardware level. The cooling technology and its operation have been improving considerably, enabling PUE values that are now trending towards unity. HPC centers are diversifying and decarbonizing their energy mix by deploying distributed energy resources, namely using RES, or using off-site generation and carbon offset mechanisms to ensure their power supply. Centers are also using their energy infrastructure to provide additional revenue streams, such as heat distribution through district heating networks. It was found that multiple opportunities exist for centers to explore their energy resources further, but also to improve their integration in energy systems and participation in the energy transition. Multiple business models can be designed to address these opportunities while dealing with both energy and computing resources, enhancing the financial viability of these centers in a scenario of rising energy costs. HPC is, therefore, an important pillar in the growing interdependency between energy and ICT infrastructures.

The highlights of the Supercomputing Conference 2022 reported by HPCWire [172] revealed several trends that are also reflected in this work. The speakers and panelists discussed how the sector is shifting from PUE and power efficiency metrics towards integrating RES, managing waste heat, and reducing energy costs and carbon emissions. HPC facilities can depict the electrical grid not as an obstacle, but as an asset, and workloads can be shifted in time and space, although with concerns on data sovereignty. Recently, Fugaku turned off 30% of its nodes for several months due to energy costs causing a financial crisis [172]. Ultimately, this means that energy usage and its associated costs can directly limit scientific research and progress in many fields of science, which are constrained not only by the availability of computing resources but also by the costs of keeping them operational.

The advent of exascale computing increases the complexity of operational challenges in HPC systems of large-scale and dynamic nature [50], and the coupling of power and cooling is driving the adoption of ODA frameworks, leading to the collection of extensive amounts of historical data not only from workload information and different usage metrics but also from the supercomputer and facilities' metrics. The analysis and exploration of these data, boosted by AI techniques, would allow for a more efficient operation and tackle the challenges of integrating local RES and energy storage technologies.

Currently, most HPC platforms, both academic and industrial, are on-premise. However, a report by Hyperian Research states that while the on-premise HPC market is growing at 7% a year until 2024, the HPC cloud market will grow at 17% a year. The recent increasing usage rate in HPCaaS, can help in having more sustainable high-performance computing. This will be possible not only by taking advantage of HPC Cloud sites with better conditions for free cooling and availability of RES but also by similar to what is already being done in traditional Cloud Computing, using spatio-temporal workload shifting with regional grid carbon intensity as a signal [173].

Energy consumption has been a priority for the silicon chips industry in the past years. Particularly, the adoption of RISC-based architectures such as ARM and ongoing with RISC-V allowed a considerable increase in performance per watt. However, the energy efficiency in supercomputers has been greatly improved at the expense of greater system heterogeneity with the generalization of accelerators (e.g., GPU) and the increasing ratio of GPUs per node. This led to GPUs being commonly used not only to speed up the training of machine learning algorithms, but also to execute large, parallel processing jobs for a broad spectrum of scientific and engineering applications.

The key findings from this study can have relevant implications for different areas. From an engineering perspective, the design of new HPC centers must account for efficiency from the building level (power architecture and supply), to the sizing and arrangement of the energy (generation, storage, cooling/heating) and IT infrastructure, down to choices at component-level that can enhance environmental performance, taking advantage of ongoing advances in computing architectures. From an energy systems perspective, HPC systems pose challenges in their integration in the electrical grid due to large power consumption and fluctuations, but also offer opportunities to foster its resilience and decarbonization, for example by using distributed energy resources and depicting scientific computing workloads as controllable loads to provide flexibility. Regarding the decarbonization of the HPC sector, there is a need for clearer metrics and concrete targets, which requires changes in policy and regulation, namely to clarify what key performance indicators HPC centers must disclose, and to improve transparency in this process. Finally, from a management perspective, service providers must balance performance with environmental concerns, leading decision-making and investments towards sustainable solutions within their data centers, adopting sustainable business models, and developing novel mechanisms to cooperate with HPC users and electrical grid operators for further decarbonization.

This study was limited to the available information on state-of-the-art supercomputers. However, this limitation should not affect the main findings, as the study focuses only on perceiving the general trends in the sector. Further research could include more quantitative analysis to compare the top supercomputers currently operational, focusing on their investments in sustainable solutions and their environmental performance. Since the service providers do not disclose some performance indicators, this type of analysis is limited. Moreover, future lines of research could include analyzing if the climate change and, more specifically, carbon emissions targets set for the ICT and data center sector are aligned with the current efforts from HPC centers, and if the identified strategies for decarbonization are enough to fulfill those targets.

CRediT authorship contribution statement

C.A. Silva: Conceptualization, Writing – original draft. **R. Vilaça:** Conceptualization, Writing. **A. Pereira:** Conceptualization, Writing. **R.J. Bessa:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is financed by *Fundo Ambiental* (FA) within project Sustainable HPC (*Computação de Elevado Desempenho Sustentável*).

References

[1] Sterling T, Anderson M, Brodowicz M. Introduction. In: High performance computing. Elsevier; 2018, p. 1–42.

[2] Jones N. How to stop data centres from gobbling up the world's electricity. Nature 2018;561(7722):163–6.

[3] Markets and Markets. High performance computing market by component, computation type, industry, deployment, serve price band, verticals & region - 2007. 2023, https://www.marketsandmarkets.com/Market-Reports/Quantum-High-Performance-Computing-Market-631.html. [Accessed 18 February 2023].

[4] Oró E, Depoorter V, Garcia A, Salom J. Energy efficiency and renewable energy integration in data centres. Strategies and modelling review. Renew Sustain Energy Rev 2015;42:429–45.

[5] Muhammed T, Mehmood R, Albeshri A, Alsolami F. HPC-smart infrastructures: A review and outlook on performance analysis methods and tools. In: Mehmood R, See S, Katib I, Chlamtac I, editors. Smart infrastructure and applications. Cham: Springer International Publishing; 2020, p. 427–51, Series Title: EAI/Springer Innovations in Communication and Computing.

[6] Andrae A, Edler T. On global electricity usage of communication technology: Trends to 2030. Challenges 2015;6(1):117–57.

[7] Manganelli M, Soldati A, Martirano L, Ramakrishna S. Strategies for improving the sustainability of data centers via energy mix, energy conservation, and circular energy. Sustainability 2021;13(11):6114.

[8] Chen W. The demands and challenges of exascale computing: an interview with Zuoning Chen. Natl Sci Rev 2016;3(1):64–7.

[9] top500org. November 2022 | TOP500. 2022, https://www.top500.org/lists/top500/2022/11/. [Accessed 22 February 2023].

[10] top500org. November 2022 | Green500. 2022, https://www.top500.org/lists/green500/2022/11/. [Accessed 22 February 2023].

[11] Milojicic D, Faraboschi P, Dube N, Roweth D. Future of HPC: Diversifying heterogeneity. In: 2021 Design, automation & test in Europe conference & exhibition. Grenoble, France: IEEE; 2021, p. 276–81.

[12] Cardwell SG, Vineyard C, Severa W, Chance FS, Rothganger F, Wang F, et al. Truly heterogeneous HPC: Co-design to achieve what science needs from HPC. In: Nichols J, Verastegui B, Maccabe A, Hernandez O, Parete-Koon S, Ahearn T, editors. Driving scientific and engineering discoveries through the convergence of HPC, Big Data and AI, vol. 1315. Cham: Springer International Publishing; 2020, p. 349–65, Series Title: Communications in Computer and Information Science.

[13] Botín-Sanabria DM, Mihaita A-S, Peimbert-García RE, Ramírez-Moreno MA, Ramírez-Mendoza RA, Lozoya-Santos JdJ. Digital twin technology challenges and applications: A comprehensive review. Remote Sens 2022;14(6).

[14] Krishnasamy E, Varrette S, Mucciardi M. Edge computing: An overview of framework and applications. External report, Partnership for Advanced Computing in Europe (PRACE); 2020.

[15] ETP4HPC. ETP4HPC strategic research agenda (SRA) European HPC Research priorities 2023–2027. External report, European Technology Platform (ETP) for High-Performance Computing (HPC); 2022.

[16] Nafus D, Schooler EM, Burch KA. Carbon-responsive computing: Changing the nexus between energy and computing. Energies 2021;14(21):6917.

[17] Czarnul P, Proficz J, Krzywaniak A. Energy-aware high-performance computing: Survey of state-of-the-art tools, techniques, and environments. Sci Program 2019;2019:1–19.

[18] D'Agostino D, Merelli I, Aldinucci M, Cesini D. Hardware and software solutions for energy-efficient computing in scientific programming. In: Mateos C, editor. Sci Program 2021;2021:5514284, Publisher: Hindawi.

[19] Cao Z, Zhou X, Hu H, Wang Z, Wen Y. Towards a systematic survey for carbon neutral data centers. IEEE Commun Surv Tutor 2022;1.

[20] Rostirolla G, Grange L, Minh-Thuyen T, Stolf P, Pierson J, Da Costa G, et al. A survey of challenges and solutions for the integration of renewable energy in datacenters. Renew Sustain Energy Rev 2022;155:111787.

[21] Wilde T, Auweter A, Shoukourian H. The 4 Pillar Framework for energy efficient HPC data centers. Comput Sci Res Dev 2014;29(3–4):241–51.

[22] Hussain SM, Wahid A, Shah MA, Akhunzada A, Khan F, Amin Nu, et al. Seven pillars to achieve energy efficiency in high-performance computing data centers. In: Jan MA, Khan F, Alam M, editors. Recent trends and advances in wireless and IoT-enabled networks. Cham: Springer International Publishing; 2019, p. 93–105, Series Title: EAI/Springer Innovations in Communication and Computing.

[23] Shehabi, Arman, Smith SJ, Sartor DA, Brown RE, Herrlin M, et al. United States data center energy usage report. Tech. rep. LBNL-1005775, Berkeley National Laboratory; 2016.

[24] Auweter A, Bode A, Brehm M, Huber H, Kranzlmüller D. Principles of energy efficiency in high performance computing. In: Kranzlmüller D, Toja AM, editors. Information and communication on technology for the fight against global warming, vol. 6868. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011, p. 18–25, Series Title: Lecture Notes in Computer Science.

[25] Strevell M, Lambiaso D, Brendamour A, Squillo T. Designing an energy-efficient HPC supercomputing center. In: Proceedings of the 48th international conference on parallel processing: Workshops. Kyoto Japan: ACM; 2019, p. 1–8.

[26] Conficoni C, Bartolini A, Tilli A, Cavazzoni C, Benini L. HPC cooling: A flexible modeling tool for effective design and management. IEEE Trans Sustain Comput 2021;6(3):441–55.

[27] Ebrahimi K, Jones GF, Fleischer AS. A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. Renew Sustain Energy Rev 2014;31:622–38.

[28] Nonaka J, Hanawa T, Shoji F. Analysis of cooling water temperature impact on computing performance and energy consumption. In: 2020 IEEE international conference on cluster computing. 2020, p. 169–75.

[29] Yuan X, Zhou X, Pan Y, Kosonen R, Cai H, Gao Y, et al. Phase change cooling in data centers: a review. Energy Build 2021;236:110764.

[30] Sridhar A, Styslinger S, Duron C, Bhavnani SH, Knight RW, Harris D, et al. Cooling of high-performance server modules using direct immersion. In: Volume 2: Heat transfer enhancement for practical applications; Fire and combustion; multi-phase systems; Heat transfer in electronic equipment; Low temperature heat transfer; Computational heat transfer. Rio Grande, Puerto Rico, USA: American Society of Mechanical Engineers; 2012, p. 759–65.

[31] Dug Technology. Data centre immersion cooling system. 2023, https://dug.com/dug-cool. [Accessed 22 February 2023].

[32] Green Revolution Cooling. High-performance computing - Green Revolution Cooling. 2023, https://www.grcooling.com. [Accessed 22 February 2023].

[33] Lim S-Y, Chang H-J. Airflow management analysis to suppress data center hot spots. Build Environ 2021;197:107843.

[34] Ljungdahl V, Jradi M, Veje C. A decision support model for waste heat recovery systems design in Data Center and High-Performance Computing clusters utilizing liquid cooling and Phase Change Materials. Appl Therm Eng 2022;201:117671.

[35] Huang P, Copertaro B, Zhang X, Shen J, Löfgren I, Rönnelid M, et al. A review of data centers as prosumers in district energy systems: renewable energy integration and waste heat reuse for district heating. Appl Energy 2020;258:114109.

[36] CSC – IT Center for Science. Lumi will be here in one year. 2020, https://www.csc.fi/en/-/lumi-will-be-here-in-one-year. [Accessed 22 February 2023].

[37] Shin W, Oles V, Karimi AM, Ellis JA, Wang F. Revealing power, energy and thermal dynamics of a 200PF pre-exascale supercomputer. In: Proceedings of the international conference for high performance computing, networking, storage and analysis. New York, NY, USA: Association for Computing Machinery; 2021.

[38] Stewart GL, Koenig GA, Liu J, Clausen A, Klingert S, Bates N. Grid accommodation of dynamic HPC demand. In: Proceedings of the 48th international conference on parallel processing: Workshops. Association for Computing Machinery; 2019.

[39] Krein PT. Data center challenges and their power electronics. CPSS Trans Power Electron Appl 2017;2(1):39–46.

[40] Chen Y, Shi K, Chen M, Xu D. Data center power supply systems: from grid edge to point-of-load. IEEE J Emerg Sel Top Power Electron 2023;11(3):2441–56.

[41] Pospieszny M. Electricity in HPC centres. Tech. rep., PRACE; 2017.

[42] Liu L, Zhang Q, Zhai ZJ, Yue C, Ma X. State-of-the-art on thermal energy storage technologies in data center. Energy Build 2020;226:110345.

[43] Kurtz J, Hovsapian R. ARIES advanced research on integrated energy systems research plan. Tech. rep., United States: National Renewable Energy Lab; 2021.

[44] Minho Advanced Computing Center. MACC welcomes the sustainable HPC project. 2023, https://macc.fccn.pt/neweventsarchive/01. [Accessed 22 February 2023].

[45] Zimmermann S, Meijer I, Tiwari MK, Paredes S, Michel B, Poulikakos D. Aquasar: A hot water cooled data center with direct energy reuse. Energy 2012;43(1):237–45, 2nd International Meeting on Cleaner Combustion (CM0901-Detailed Chemical Models for Cleaner Combustion).

[46] Van HN, Tran FD, Menaud J-M. Performance and power management for cloud infrastructures. In: 2010 IEEE 3rd international conference on cloud computing. 2010, p. 329–36.

[47] Liu N, Li Z, Xu J, Xu Z, Lin S, Qiu Q, et al. A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning. In: 2017 IEEE 37th international conference on distributed computing systems. 2017, p. 372–82.

[48] Abu Ahmad W, Bartolini A, Beneventi F, Benini L, Borghesi A, Cicala M, et al. Design of an energy aware petaflops class high performance cluster based on power architecture. In: 2017 IEEE international parallel and distributed processing symposium workshops. 2017, p. 964–73.

[49] Labasan S. Energy-efficient and power-constrained techniques for exascale computing area exam position paper. 2016.

[50] Netti A, Shin W, Ott M, Wilde T, Bates N. A conceptual framework for HPC operational data analytics. In: 2021 IEEE international conference on cluster computing. Portland, OR, USA: IEEE; 2021, p. 596–603.

[51] Cai C, Wang L, Khan SU, Tao J. Energy-aware high performance computing: A taxonomy study. In: 2011 IEEE 17th international conference on parallel and distributed systems. Tainan, Taiwan: IEEE; 2011, p. 953–8.

[52] Goiri I, Haque ME, Le K, Beauchea R, Nguyen TD, Guitart J, et al. Matching renewable energy supply and demand in green datacenters. Ad Hoc Netw 2015;25:520–34.

[53] Kassab A, Nicod J-M, Philippe L, Rehn-Sonigo V. Green power aware approaches for scheduling independent tasks on a multi-core machine. Sustain Comput Inform Syst 2021;31:100590.

[54] Aikema D, Kiddle C, Simmonds R. Energy-cost-aware scheduling of HPC workloads. In: 2011 IEEE international symposium on a world of wireless, mobile and multimedia networks. 2011, p. 1–7.

[55] Georgiou Y, Cadeau T, Glesser D, Auble D, Jette M, Hautreux M. Energy accounting and control with SLURM resource and job management system. In: Chatterjee M, Cao J-n, Kothapalli K, Rajsbaum S, editors. Distributed computing and networking. Lecture notes in computer science, Berlin, Heidelberg: Springer; 2014, p. 96–118.

[56] Eastep J, Sylvester S, Cantalupo C, Geltz B, Ardanaz F, Al-Rawi A, et al. Global extensible open power manager: A vehicle for HPC community collaboration on co-designed energy management solutions. In: High performance computing: 32nd international conference, isc high performance 2017, Frankfurt, Germany, June 18–22, 2017, proceedings. Berlin, Heidelberg: Springer-Verlag; 2017, p. 394–412.

[57] Corbalan J, Alonso L, Aneas J, Brochard L. Energy optimization and analysis with EAR. In: 2020 IEEE international conference on cluster computing. 2020, p. 464–72.

[58] Goiri I, Katsak W, Le K, Nguyen TD, Bianchini R. Parasol and GreenSwitch: managing datacenters powered by renewable energy. In: Proceedings of the eighteenth international conference on architectural support for programming languages and operating systems. Houston, Texas, USA: ACM Press; 2013, p. 51.

[59] Pierson J-M, Stolf P, Sun H, Casanova H. MILP formulations for spatio-temporal thermal-aware scheduling in Cloud and HPC datacenters. Cluster Comput 2020;23(2):421–39.

[60] Li Y, Wang X, Luo P, Pan Q. Thermal-aware hybrid workload management in a green datacenter towards renewable energy utilization. Energies 2019;12(8):1494.

[61] Madon M, Pierson J-M. Integrating pre-cooling of data center operated with renewable energies. In: 2020 International conferences on internet of things (IThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData) and IEEE congress on cybermatics. Rhodes, Greece: IEEE; 2020, p. 332–41.

[62] Zhou Z, Liu F, Xu Y, Zou R, Xu H, Lui JC, et al. Carbon-aware load balancing for geo-distributed cloud services. In: 2013 IEEE 21st international symposium on modelling, analysis and simulation of computer and telecommunication systems. San Francisco, CA, USA: IEEE; 2013, p. 232–41.

[63] Ding Z, Xie L, Lu Y, Wang P, Xia S. Emission-aware stochastic resource planning scheme for data center microgrid considering batch workload scheduling and risk management. IEEE Trans Ind Appl 2018;54(6):5599–608.

[64] Radovanovic A, Koningstein R, Schneider I, Chen B, Duarte A, Roy B, et al. Carbon-aware computing for datacenters. IEEE Trans Power Syst 2022;1.

[65] Wiesner P, Behnke I, Scheinert D, Gontarska K, Thamsen L. Let's wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud. In: Proceedings of the 22nd international middleware conference. Québec city Canada: ACM; 2021, p. 260–72.

[66] Acun B, Lee B, Maeng K, Chakkaravarthy M, Gupta U, Brooks D, et al. A holistic approach for designing carbon aware datacenters. 2022, arXiv:2201.10036.

[67] Saurav SK, Benedict S. A taxonomy and survey on energy-aware scientific workflows scheduling in large-scale heterogeneous architecture. In: 2021 6th international conference on inventive computation technologies. 2021, p. 820–6.

[68] Valter H, Karlsson A, Pericàs M. Energy-efficiency evaluation of OpenMP loop transformations and runtime constructs. 2022, arXiv.

[69] Shankar S, Reuther A. Trends in energy estimates for computing in AI/Machine learning accelerators, supercomputers, and compute-intensive applications. In: 2022 IEEE high performance extreme computing conference. IEEE; 2022.

[70] Dutot P-F, Mercier M, Poquet M, Richard O. Batsim: A realistic language-independent resources and jobs management systems simulator. In: Desai N, Cirne W, editors. Job scheduling strategies for parallel processing. Cham: Springer International Publishing; 2017, p. 178–97.

[71] Aksar B, Schwaller B, Aaziz O, Leung VJ, Brandt J, Egele M, et al. E2EWatch: An end-to-end anomaly diagnosis framework for production HPC systems. In: Sousa L, Roma N, Tomás P, editors. Euro-Par 2021: Parallel processing. Cham: Springer International Publishing; 2021, p. 70–85.

[72] Kurowski K, Oleksiak A, Piątek W, Piontek T, Przybyszewski A, Węglarz J. DCworms – A tool for simulation of energy efficiency in distributed computing infrastructures. Simul Model Pract Theory 2013;39:135–51, S.I.Energy efficiency in grids and clouds.

[73] Zhang Z, Lang M, Pakin S, Fu S. Tracsim: Simulating and scheduling trapped power capacity to maximize machine room throughput. Parallel Comput 2016;57:108–24.

[74] Casanova H, Giersch A, Legrand A, Quinson M, Suter F. Versatile, scalable, and accurate simulation of distributed applications and platforms. J Parallel Distrib Comput 2014;74(10):2899–917.

[75] Heinrich FC, Cornebize T, Degomme A, Legrand A, Carpen-Amarie A, Hunold S, et al. Predicting the energy-consumption of MPI applications at scale using only a single node. In: 2017 IEEE international conference on cluster computing. 2017, p. 92–102.

[76] Coleman T, Casanova H, Gwartney T, da Silva RF. Evaluating energy-aware scheduling algorithms for I/O-intensive scientific workflows. In: Paszynski M, Kranzlmüller D, Krzhizhanovskaya VV, Dongarra JJ, Sloot PMA, editors. Computational science – ICCS 2021. Lecture notes in computer science, Cham: Springer International Publishing; 2021, p. 183–97.

[77] Poquet M. Simulation approach for resource management [Ph.D. thesis], Université Grenoble Alpes; 2017.

[78] RISC-V. RISC-V board of directors. 2023, https://riscv.org/about/board-of-directors/. [Accessed 22 February 2023].

[79] Li P. Sifive intelligence X280 as AI compute host: Google datacenter case study. 2022, SiFive, https://www.sifive.com/blog/sifive-intelligence-x280-as-ai-compute-host-google. [Accessed 22 February 2023].

[80] Almeida F, Assunção MD, Barbosa J, Blanco V, Brandic I, Da Costa G, et al. Energy monitoring as an essential building block towards sustainable ultrascale systems. Sustain Comput Inform Syst 2018;17:27–42.

[81] Li K. Improving multicore server performance and reducing energy consumption by workload dependent dynamic power management. IEEE Trans Cloud Comput 2016;4(2):122–37.

[82] Chadha M, Gerndt M. Modelling DVFS and UFS for region-based energy aware tuning of HPC applications. In: 2019 IEEE international parallel and distributed processing symposium. 2019, p. 805–14.

[83] Schone R, Ilsche T, Bielert M, Velten M, Schmidl M, Hackenberg D. Energy efficiency aspects of the AMD zen 2 architecture. In: 2021 IEEE international conference on cluster computing. IEEE; 2021.

[84] Majumdar A, Piga L, Paul I, Greathouse JL, Huang W, Albonesi DH. Dynamic GPGPU power management using adaptive model predictive control. In: 2017 IEEE international symposium on high performance computer architecture. 2017, p. 613–24.

[85] Krzywaniak A, Czarnul P, Proficz J. DEPO: A dynamic energy-performance optimizer tool for automatic power capping for energy efficient high-performance computing. Softw - Pract Exp 2022;52:2598–634.

[86] Krzywaniak A, Czarnul P, Proficz J. Dynamic GPU power capping with online performance tracing for energy efficient GPU computing using DEPO tool. Future Gener Comput Syst 2023;145:396–414.

[87] Kodama Y, Odajima T, Arima E, Sato M. Evaluation of power management control on the supercomputer fugaku. In: 2020 IEEE international conference on cluster computing. 2020, p. 484–93.

[88] Pereira R, Couto M, Ribeiro F, Rua R, Cunha J, Fernandes JP, et al. Ranking programming languages by energy efficiency. Sci Comput Prog 2021;205:102609.

[89] Portegies Zwart S. The ecological impact of high-performance computing in astrophysics. Nat Astron 2020;4(9):819–22.

[90] Augier P, Bolz-Tereick CF, Guelton S, Mohanan AV. Reducing the ecological impact of computing through education and Python compilers. Nat Astron 2021;5(4):334–5.

[91] Perkel JM. Julia: come for the syntax, stay for the speed. Nature 2019;572(7768):141+, 141.

[92] Chen J, Manivannan M, Abduljabbar M, Pericàs M. ERASE: Energy efficient task mapping and resource management for work stealing runtimes. ACM Trans Archit Code Optim 2022;19(2).

[93] Moraru M, Warnet M, Loiseau J, Ramakrishnaiah V, Prajapati N, Lim H, et al. Transformations for energy efficient accelerated chain matrix multiplication (TEE-ACM 2). 2022, Supercomputing, Poster.

[94] Müller A, Deconinck W, Kühnlein C, Mengaldo G, Lange M, Wedi N, et al. The ESCAPE project: Energy-efficient scalable algorithms for weather prediction at exascale. Geosci Model Dev 2019;12(10):4425–41.

[95] Szustak L, Wyrzykowski R, Kuczynski L, Olas T. Architectural adaptation and performance-energy optimization for CFD application on AMD EPYC Rome. IEEE Trans Parallel Distrib Syst 2021;32:2852–66.

[96] Jiang J, Du J, Huang D-E, Chen Z, Lu Y, Liao X. Full-stack optimizing transformer inference on ARM many-core CPU. IEEE Trans Parallel Distrib Syst 2023;34:2221–35.

[97] Chowdhury A, Kumaraswamy M, Gerndt M. READEX tool suite for energy-efficiency tuning of HPC applications. In: Proceedings of the 2017 workshop on software engineering methods for parallel and high performance applications. SEM4HPC '17, New York, NY, USA: Association for Computing Machinery; 2017, p. 11–2.

[98] Marjanović V, Gracia J, Glass CW. Performance modeling of the HPCG benchmark. In: Jarvis SA, Wright SA, Hammond SD, editors. High performance computing systems. performance modeling, benchmarking, and simulation. Cham: Springer International Publishing; 2015, p. 172–92.

[99] Roberts SI, Wright SA, Fahmy SA, Jarvis SA. Metrics for energy-aware software optimisation. In: Kunkel JM, Yokota R, Balaji P, Keyes D, editors. High performance computing. Cham: Springer International Publishing; 2017, p. 413–30.

[100] Gupta U, Kim YG, Lee S, Tse J, Lee H-HS, Wei G-Y, et al. Chasing carbon: The elusive environmental footprint of computing. In: 2021 IEEE international symposium on high-performance computer architecture. Seoul, Korea (South): IEEE; 2021, p. 854–67.

[101] Kubert R, Wesner S. Using service level agreements in a high-performance computing environment. Scalable Comput Pract Exp 2011;12(2):164–78.

[102] Gantikow H, Reich C, Knahl M, Clarke N. A taxonomy for HPC-aware cloud computing. Sl: sn 2015;57–62.

[103] Haque ME, Le K, Goiri I, Bianchini R, Nguyen TD. Providing green SLAs in high performance computing clouds. In: 2013 International green computing conference proceedings. Arlington, VA, USA: IEEE; 2013, p. 1–11.

[104] Hasan MS, Kouki Y, Ledoux T, Pazat J-L. Exploiting renewable sources: When green SLA becomes a possible reality in cloud computing. IEEE Trans Cloud Comput 2017;5(2):249–62.

[105] Netto MAS, Calheiros RN, Rodrigues ER, Cunha RLF, Buyya R. HPC cloud for scientific and business applications: taxonomy, vision, and research challenges. ACM Comput Surv 2019;51(1):1–29.

[106] Ligozat A-L, Névéol A, Daly B, Frenoux E. Ten simple rules to make your research more sustainable. In: Schwartz R, editor. PLoS Comput Biol 2020;16(9):e1008148.

[107] Govaart G, Hofmann SM, Medawar E. The sustainability argument for open science. Preprint, Open Science Framework; 2021.

[108] Mayo-Garcia R, Audit E. Exascale, a great opportunity for clean energy transition in Europe. Tech. rep., European Energy Research Alliance; 2022.

[109] Huerta EA, Khan A, Davis E, Bushell C, Gropp WD, Katz DS, et al. Convergence of artificial intelligence and high performance computing on NSF-supported cyberinfrastructure. J Big Data 2020;7(1):88.

[110] Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI. Commun ACM 2020;63(12):54–63.

[111] Patterson D, Gonzalez J, Holzle U, Le Q, Liang C, Munguia L-M, et al. The carbon footprint of machine learning training will plateau, then shrink. Computer 2022;55(7):18–28.

[112] Borghesi A, Bartolini A, Milano M, Benini L. Pricing schemes for energy-efficient HPC systems: Design and exploration. Int J High Perform Comput Appl 2019;33(4):716–34.

[113] Georgiou Y, Glesser D, Rzadca K, Trystram D. A scheduler-level incentive mechanism for energy efficiency in HPC. In: CCGrid 2015 - 15th IEEE/ACM international symposium on cluster, cloud and grid computing. Shenzhen, China; 2015, p. 617–26.

[114] Lannelongue L, Grealey J, Inouye M. Green algorithms: Quantifying the carbon footprint of computation. Adv Sci 2021;8(12):2100707.

[115] European Comission. 2030 Digital compass: The European way for the digital decade. Tech. rep., European Comission; 2021.

[116] Council of the European Union. Council regulation on establishing the european high performance computing joint undertaking and repealing regulation (EU) 2018/1488. Tech. rep., Council of the European Union; 2018.

[117] PRACE. Partnership for advanced computing in Europe. 2023, https://prace-ri.eu. [Accessed 22 February 2023].

[118] ETP4HPC. European technology platform (ETP) for high-performance computing (HPC). 2023, https://etp4hpc.eu/. [Accessed 22 February 2023].

[119] EE HPC WG. EE HPC WG Home Site. 2023, https://eehpcwg.lbl.gov/. [Accessed 22 February 2023].

[120] Acton M, Bertoldi P, Booth J. 2021 Best practice guidelines for the EU code of conduct on data centre energy efficiency. Tech. Rep. JRC123653, Joint Research Centre, European Union; 2021.

[121] Center of Expertise for Energy Efficiency in Data Centers. Master list of energy efficiency actions. Tech. rep., Lawrence Berkeley National Laboratory; 2016.

[122] Center on Regulation in Europe. Data centres & the grid – Greening ICT in Europe. Report, CERRE; 2021.

[123] Dietrich JM, Lawrence A. Navigating regulations and standards. Tech. Rep. UI Intelligence Report 68, Uptime Institute; 2022.

[124] EEP. European exascale projects. 2023, http://www.exascale-projects.eu/. [Accessed 22 February 2023].

[125] Schulz M, Jana S, Brink S, Sakamoto R. HPC PowerStack: Community-driven collaboration on power-aware system stack.

[126] European Comission. Framework Partnership Agreement (FPA) for developing a large-scale European initiative for High Performance Computing (HPC) ecosystem based on RISC-V.

[127] HEROES. HEROES - Hybrid eco responsible optimized European solution. 2023, https://heroes-project.eu/. [Accessed 22 February 2023].

[128] Armejach A, Brank B, Cortina J, Dolique F, Hayes T, Ho N, et al. Mont-blanc 2020: Towards scalable and power efficient European HPC processors. In: 2021 Design, automation & test in Europe conference & exhibition. 2021, p. 136–41.

[129] Wedi N, Bauer P, Mueller A, Deconinck W. Energy-efficient scalable algorithms for weather prediction at exascale (ESCAPE). In: 18th workshop on high performance computing in meteorology. 2018.

[130] Silvano C, Palermo G, Agosta G, Ashouri AH, Gadioli D, Cherubin S, et al. Autotuning and adaptivity in energy efficient HPC systems: The ANTAREX toolbox. In: Proceedings of the 15th ACM international conference on computing frontiers. New York, NY, USA: Association for Computing Machinery; 2018, p. 270–5.

[131] ECOSCALE. ECOSCALE. 2023, https://cordis.europa.eu/project/id/671632. [Accessed 22 February 2023].
[132] ADEPT. ADEPT. 2023, https://cordis.europa.eu/project/id/610490. [Accessed 22 February 2023].
[133] EXA2GREEN. Exa2Green. 2023, http://exa2green-project.eu/. [Accessed 22 February 2023].
[134] European Investment Bank. Financing the future of supercomputing: How to increase investments in high performance computing in Europe. LU: Publications Office; 2018.
[135] Amazon Web Services. AWS HPC. 2023, https://aws.amazon.com/hpc/. [Accessed 22 February 2023].
[136] Microsoft. Azure HPC. 2023, https://azure.microsoft.com/solutions/high-performance-computing. [Accessed 22 February 2023].
[137] iExec. iExec. 2023, https://iex.ec. [Accessed 22 February 2023].
[138] Hypernet Labs. Galileo. 2023, https://hypernetlabs.io/galileo/. [Accessed 22 February 2023].
[139] Eurich M, Calleja P, Boutellier R. Business models of high performance computing centres in higher education in Europe. J Comput Higher Educ 2013;25(3):166–81.
[140] Lannelongue L, Grealey J, Bateman A, Inouye M. Ten simple rules to make your computing more environmentally sustainable. In: Schwartz R, editor. PLoS Comput Biol 2021;17(9):e1009324.
[141] Borealis. Borealis data center. 2023, https://bdc.is/. [Accessed 22 February 2023].
[142] LANCIUM. LANCIUM compute. 2023, https://portal.lancium.com/. [Accessed 22 February 2023].
[143] Klingert S. Mapping data centre business types with power management strategies to identify demand response candidates. In: Proceedings of the ninth international conference on future energy systems. Karlsruhe Germany: ACM; 2018, p. 492–8.
[144] Ahmed K. Energy demand response for high-performance computing systems [Doctor of philosophy computer science], Florida International University; 2018.
[145] Klingert S, Szilvas S. Spinning gold from straw - evaluating the flexibility of data centres on power markets. Energy Inf 2020;3(1):7.
[146] Wilson DC, Paschalidis IC, Coskun AK. Site-wide HPC data center demand response. In: 2022 IEEE high performance extreme computing conference. 2022, p. 1–7.
[147] Cioara T, Anghel I, Salomie I, Antal M, Pop C, Bertoncini M, et al. Exploiting data centres energy flexibility in smart cities: Business scenarios. Inform Sci 2019;476:392–412.
[148] Bates N, Ghatikar G, Abdulla G, Koenig GA, Bhalachandra S, Sheikhalishahi M, et al. Electrical grid and supercomputing centers: An investigative analysis of emerging opportunities and challenges. Informatik-Spektrum 2015;38(2):111–27.
[149] Clausen A, Koenig G, Klingert S, Ghatikar G, Schwartz PM, Bates N. An analysis of contracts and relationships between supercomputing centers and electricity service providers. In: Proceedings of the 48th international conference on parallel processing: Workshops. Kyoto Japan: ACM; 2019, p. 1–8.
[150] Lancium. Lancium. 2023, https://lancium.com/. [Accessed 22 February 2023].
[151] Baumann C. Data centers of the future require microgrids. Mission Crit https://www.missioncriticalmagazine.com/articles/93239-data-centers-of-the-future-require-microgrids. [Accessed 22 February 2023].
[152] European CommissionJoint Research Centre. Energy communities: An overview of energy and social innovation. LU: Publications Office; 2020.
[153] Telia. Telia Helsinki data center. 2023, https://www.telia.fi/business/data-center-services/data-centers/helsinki-data-center. [Accessed 22 February 2023].
[154] Lake Parime. Lake parime. 2023, https://lakeparime.com/. [Accessed 22 February 2023].
[155] Soluna. Soluna. 2023, https://www.solunacomputing.com/. [Accessed 22 February 2023].
[156] Zhang X, Turiansky ME, Van de Walle CG. All-inorganic halide perovskites as candidates for efficient solar cells. Cell Rep Phys Sci 2021;2(10):100604.
[157] Liu H, Zhu Z, Yan Q, Yu S, He X, Chen Y, et al. A disordered rock salt anode for fast-charging lithium-ion batteries. Nature 2020;585(7823):63–7.
[158] GAIA-X. GAIA-X: A federated data infrastrucure for Europe. 2023, https://www.data-infrastructure.eu. [Accessed 22 February 2023].
[159] Rojas E, Meneses E, Jones T, Maxwell D. Analyzing a five-year failure record of a leadership-class supercomputer. In: 2019 31st international symposium on computer architecture and high performance computing. 2019, p. 196–203.
[160] Li B, Roy RB, Wang D, Samsi S, Gadepally V, Tiwari D. Toward sustainable HPC: Carbon footprint estimation and environmental implications of HPC systems. 2023, arXiv.
[161] Van de Voort T, Zavrel V, Torren Galdiz I, Hensen J. Analysis of performance metrics for data center efficiency – should the Power Utilization Effectiveness PUE still be used as the main indicator? (Part 1). Tech. rep., Federation of European Heating, Ventilation and Air Conditioning Associations (REHVA); 2017.
[162] Van de Voort T, Zavrel V, Torren Galdiz I, Hensen J. Analysis of performance metrics for data center efficiency – should the Power Utilization Effectiveness PUE still be used as the main indicator? (Part 2). Tech. rep., Federation of European Heating, Ventilation and Air Conditioning Associations (REHVA); 2017.
[163] Bizo D, Ascierto R, Lawrence A, Davis J. Uptime Institute global data center survey 2021. Tech. rep. UI intelligence report 51, Uptime Institute; 2021.
[164] Reddy VD, Setz B, Rao GSV, Gangadharan G, Aiello M. Metrics for sustainable data centers. IEEE Trans Sustain Comput 2017;2(3):290–303.
[165] Levy M, Raviv D. An overview of data center metrics and a novel approach for a new family of metrics. Adv Sci Technol Eng Syst J 2018;3(2):238–51.
[166] Lykou G, Mentzelioti D, Gritzalis D. A new methodology toward effectively assessing data center sustainability. Comput Secur 2018;76:327–40.
[167] Yang X, Zhou Z, Wallace S, Lan Z, Tang W, Coghlan S, et al. Integrating dynamic pricing of electricity into energy aware scheduling for HPC systems. In: Proceedings of the international conference on high performance computing, networking, storage and analysis. Denver Colorado: ACM; 2013, p. 1–11.
[168] Bourassa N, Johnson W, Broughton J, Carter DM, Joy S, Vitti R, et al. Operational data analytics: optimizing the national energy research scientific computing center cooling systems. In: Proceedings of the 48th international conference on parallel processing: Workshops. ICPP 2019, New York, NY, USA: Association for Computing Machinery; 2019.
[169] Terai M, Yamamoto K, Miura S, Shoji F. An operational data collecting and monitoring platform for Fugaku: System overviews and case studies in the prelaunch service period. In: Jagode H, Anzt H, Ltaief H, Luszczek P, editors. High performance computing. Cham: Springer International Publishing; 2021, p. 365–77.
[170] Netti A, Müller M, Guillen C, Ott M, Tafani D, Ozer G, et al. DCDB wintermute: Enabling online and holistic operational data analytics on HPC systems. In: Proceedings of the 29th international symposium on high-performance parallel and distributed computing. New York, NY, USA: Association for Computing Machinery; 2020, p. 101–12.
[171] Netti A, Müller M, Auweter A, Guillen C, Ott M, Tafani D, et al. From facility to application sensor data: modular, continuous and holistic monitoring with DCDB. In: Proceedings of the international conference for high performance computing, networking, storage and analysis. New York, NY, USA: Association for Computing Machinery; 2019.
[172] Peckham O. At SC22, carbon emissions and energy costs eclipsed hardware efficiency. 2022, HPCwire, https://www.hpcwire.com/2022/12/02/at-sc22-carbon-emissions-and-energy-costs-eclipsed-hardware-efficiency/. [Accessed 18 February 2023].
[173] Koningstein R. We now do more computing where there's cleaner energy. 2021, https://blog.google/outreach-initiatives/sustainability/carbon-aware-computing-location/. [Accessed 22 February 2023].