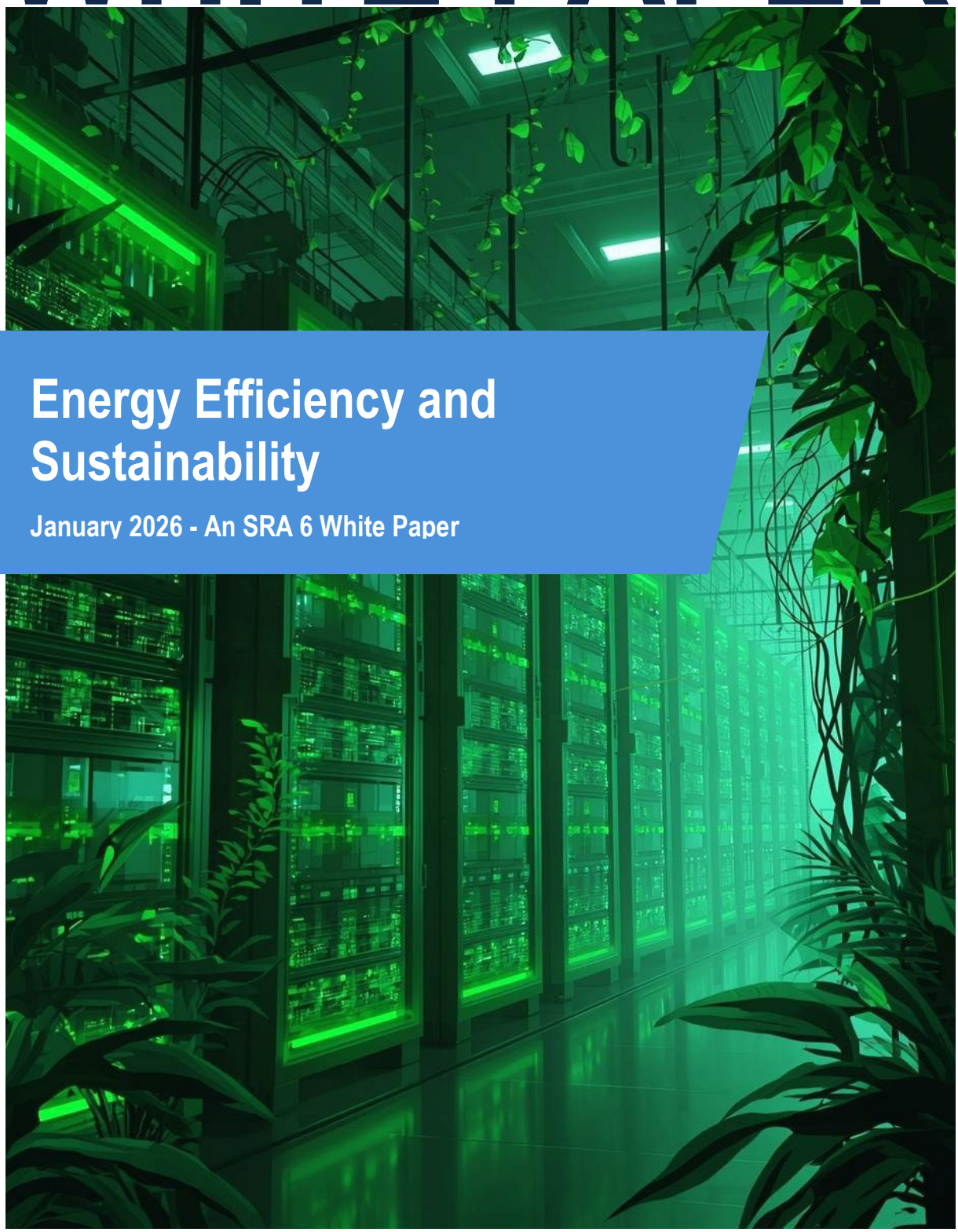
WHITE PAPER



Energy Efficiency and Sustainability

January 2026 - An SRA 6 White Paper

ETP 4 HPC

EUROPEAN TECHNOLOGY PLATFORM FOR HIGH PERFORMANCE COMPUTING

Key contributors

Andreas Wierse (SICOS BW GmbH)

Pierre Seroul (Eviden)

Ondrej Vysocky (IT4Innovations national super computing centre, VŠB – Technical University of Ostrava)

Table of contents

Executive Summary 3

Glossary of acronyms 4

Introduction 5

Trends and Outlook 6

Key Insights 7

1 Integrating Environmental Constraints and Opportunities into the Design and Operation of HPC and AI Systems 8

2 HPC hardware-software co-design 10

3 Application workflows: Software and Data Stack 11

4 Optimizing HPC Lifecycle and Sustainability 13

5 Raising Awareness and Empowering Stakeholders 16

Post Exascale Vision 18

Key R&I Recommendations 18

Conclusions 20

Contributors 21

References 22

Executive Summary

This is a white paper released as part of the ETP4HPC's Strategic Research Agenda 6.

Energy efficiency and sustainability (EE&S) have become central challenges for modern High-Performance Computing (HPC) and AI infrastructures. While energy savings are often viewed as the main issue, sustainability also includes ecological, economic, and societal dimensions. A full life-cycle perspective—from planning and procurement to operation and decommissioning—is necessary to avoid rebound effects such as the Jevons paradox[1].

HPC and AI systems increasingly interact with their environment. Their multi-megawatt, highly variable power demand influences grid stability and requires better forecasting and closer coordination with energy providers. Future operations must incorporate energy- and carbon-aware scheduling, while heat reuse concepts and standardized sustainability metrics (PUE, WUE, ERE) become integral to system design and evaluation.

Growing hardware heterogeneity (CPUs, GPUs, accelerators, and emerging neuromorphic or quantum devices) offers efficiency potential that is still underutilized due to insufficient software optimization. Progress requires stronger hardware–software co-design, greater support for developers in exploiting advanced architectural features, and systematic use of monitoring data to identify inefficient applications. European semiconductor initiatives further enable alignment between hardware and computational requirements.

Software, algorithms, and workflows are equally important. Energy-efficient algorithms, data reduction techniques (compression, filtering, deduplication), improved I/O strategies, and optimized workflows can significantly reduce resource usage. Strengthening reproducibility and robust experiment packaging across diverse systems helps prevent wasted compute time and supports sustainable software lifecycles.

Digital twins (DTs) are emerging as key tools for lifecycle management. They support planning, operational optimization, predictive analysis, and end-of-life decisions. Effective DTs require harmonized monitoring, standard metrics, and shared methodologies. A European knowledge initiative could accelerate adoption and ensure consistent practices across HPC sites.

Raising awareness and empowering stakeholders—developers, users, operators, vendors, and funding bodies—is essential. Transparent monitoring infrastructures and vendor-agnostic sustainability indicators enable informed decision-making. User feedback mechanisms (dashboards, reports, incentives) can encourage more energy-conscious behaviour. Funding agencies, particularly the European Commission, should embed sustainability metrics, full life-cycle assessments [ISO 14040/14044], and continuous reporting requirements into procurement and project evaluations.

In the post-exascale era, simple performance scaling through increased energy use is no longer viable. With growing AI workloads and the need to reduce energy demand and CO_2 emissions, coordinated progress across hardware, software, operations, infrastructure, and user behaviour is imperative. Only a combined effort by all stakeholders will enable high-performing HPC and AI systems to meet future needs while aligning with long-term sustainability goals.

[1] https://en.wikipedia.org/wiki/Jevons_paradox

Glossary of acronyms

AC: Alternating Current

AI: Artificial Intelligence

DC: Direct Current

DT: Digital Twin

EC: European Commission

EE&S: Energy Efficiency and Sustainability

EPEAT: Electronic Product Environmental Assessment Tool

ERE: Energy Reuse Effectiveness

FP64: Double-precision floating-point format (64 bits)

HBM: High Bandwidth Memory

HPC: High Performance Computing

ICT: Information and Communication Technology

KPI: Key Performance Indicator

LCA: Life Cycle Analysis

NUMA: Non-Uniform Memory Architecture

OODA: Observe, Orient, Decide, and Act (feedback loop)

OS: Operating System

PEF: Product Environmental Footprint

PUE: Power Usage Effectiveness

RJMS: Resource and Jobs Management System

RoHS: Restriction of Hazardous Substances

TCO: Total Cost of Ownership

TRL: Technology Readiness Level

WEEE: Waste Electrical and Electronic Equipment

WUA: Water Usage Effectiveness



Introduction

In the title of this White Paper, energy efficiency is quoted before sustainability. And somehow, in supercomputing, saving energy seems to be the most imminent challenge. It must be noted that sustainability is a lot more than just saving energy. Its widely accepted definition includes three areas: **ecological** (which often is interpreted as being the whole of sustainability), **economic** (which is actually obvious, given the current discussions around volatile energy prices), and finally **societal** (which is clearly the least commonly noted aspect). In this White Paper, we will attempt to strike a balance, which is essential for success.

Only if an overall perspective is taken, effects like the Jevons paradox can be avoided or at least reduced: if only plain energy consumption counts, improvements there might lead to higher usage because the users are more relaxed once they have the feeling that efficiency has increased. However, we need users and every other stakeholder to keep their attention high and not rest on the laurels they presumably earned. Here, the third pillar of sustainability comes into play: societal elements may well be related to ethics and will help to keep the overall goal in mind.

What shall also be considered in an overall view is the effect that the results of the computations have and how this relates to resource usage. Of course, there are areas where this is rather difficult to define, but wherever it is possible, it should be done. The users should not only be aware of the energy efficiency of their software, but they should also understand the effects their results have beyond the direct HPC environment.

Both the immediate efficiency improvements and the real-world effects of computation results should also be taken into account by funding agencies, regardless of whether they pay for the systems, the buildings, the staff, or the research and software development. Rules set by funding agencies have an immediate impact on the awareness of energy efficiency and sustainability topics, and they are a strong tool when it comes to adjusting the HPC ecosystem to be fit for the upcoming challenges, some of which are already visible today.



Trends and Outlook

We have identified a number of trends that we also used to structure the document. The topics addressed span from process-oriented and organizational issues, through hardware and low-level OS aspects, to higher-level software, data management, and stakeholder awareness (including users and financiers). You will find these trends mapped directly to the sections:

1. **Monitoring & Optimisation, Middleware/power management (RJMS), Digital Twins**

In order to be able to measure energy efficiency and sustainability, it is mandatory to be able to monitor the systems and their surrounding infrastructure. This has to go beyond simply reading hardware sensors and metres. Especially resource and job management systems can play an important role here. In the ideal world, a full digital twin of such a system would enable the centre to gain a thorough understanding of it well before the purchase decision is made.

2. **Integrating environmental context for HPC full life cycle optimisation**

The typical environmental context focuses on energy consumption and temperature issues during the operational phase. While it is still possible to improve the understanding in this phase (which, of course, should be done), it is also really important to widen the perspective to the whole life cycle: from the planning and building of the system to the destruction and recycling of it.

3. **Architecture of the HPC system (computing hardware & OS/system)**

Already visible today is a movement that goes from the general-purpose CPU that followed the vector era to a much more diversified setup, including GPUs, but also targeting upcoming technologies like neuromorphic or quantum computing. In such a diverse setup, the architecture of a system and its operating system play a very important role when it comes to improving EE&S.

4. **Application workflows: Data and Software stack**

For a long time, the effect of efficient software and smart data handling has been underestimated. This is clearly related to the fact that the computation time cost was the primary factor when addressing efficiency. However, the effect of longer compute times, which at the same time leads to less resource usage overall, should be ranked significantly higher than it is today.

5. **Energy- and power-aware users and funding agencies**

Finally, it is crucial that the users as well as the funding bodies are aware of their influence on the EE&S aspects of running an HPC system. If the users do not see their own benefit in improving efficiency, or are not provided with pertinent leverages, the full potential for sustainability of a given system is not going to be tapped. Also, the leverage that funding agencies possess, when outlining the boundary conditions for their funding, cannot be underestimated.

We are seeing a slow but steady growth of the awareness of all of these EE&S aspects, but it is important to support this development. If funding agencies and HPC providers alike take EE&S into account as soon as they start thinking about a new system, the chances are high that significant progress will be made. However, in order to achieve this, it is very important to reach as many of those stakeholders as possible, identify those that are the most advanced and that are willing to follow this path, and build the examples and success stories that will convince others that this is the right direction.

Key Insights

- An increase in overall power consumption in HPC and Cloud services is expected (rise of AI)
- A reduction of "real-world" resource usage by using HPC simulations (optimisation of processes) must be integrated into an overall examination
- The whole life cycle of systems and centres needs to be addressed, including software, computer-hall re-usage, etc.
- Continuous monitoring of the usage of the system resources should be established, and reasonable power management capabilities should be made available
- Users and administrators of HPC systems should become energy-aware, a path to energy/CO_2 emissions accounting and relevant leverages must be developed
- Power monitoring systems should be harmonised, and standard metrics must be defined that reliably reflect the "real" power consumption
- An HPC system digital twin could be very helpful: it would allow a more realistic prediction of power consumption
- Hardware will be increasingly heterogeneous; energy consumption should be considered when co-design is applied or a platform is chosen
- Modular construction of HPC systems seems to be becoming more popular or even inevitable; construction of computer rooms should reflect that (simplified, adaptable, ...)
- Improvements in efficient DC cooling infrastructure are needed (warm water cooling, heat re-use, elimination of water evaporation, etc.)

1 Integrating Environmental Constraints and Opportunities into the Design and Operation of HPC and AI Systems

Modern computing systems are increasingly interconnected with their surrounding environment from multiple perspectives. They impact their environment through energy consumption and heat generation, while simultaneously being constrained by factors such as electricity grid limitations, ambient temperature, and fluctuating electricity prices.

High-Performance Computing (HPC) systems and AI factories are now integral components of the power grid, with their power demands reaching several megawatts [Top500.org]. Their consumption varies significantly over time [Benhari], driven by workload intensity and the nature of the applications running — a phenomenon further amplified by the growing heterogeneity of processing units and accelerators. Such variations can affect grid stability, creating a need for new communication mechanisms between system operators and grid providers [International Energy Agency]. For these mechanisms to be effective, reliable forecasting of future power consumption is essential.

Energy consumption, however, is not the only environmental concern. During operation, HPC and AI systems interface with multiple utility flows, primarily electricity, but also heat and water. As server cooling remains critical, improving cooling efficiency is paramount. This requires enhancing heat extraction methods. Free cooling [Daraghmeh and Wang] should be utilized whenever the external ambient temperature is lower than the internal temperature of the computing environment. For upcoming systems, established techniques such as direct liquid cooling and hot water cooling should be more broadly adopted, while emerging methods like two-phase (diphasic) cooling need to be industrialized [Chen]. Furthermore, selecting electronic components based on their thermal characteristics — enabling operation at higher temperatures without degrading performance or lifespan — can further reduce system Power Usage Effectiveness (PUE) and water consumption.

Additionally, when evaluating proposals for the deployment of new computational centres, funding agencies should emphasize the requirement to recover and repurpose waste heat — for example, for heating nearby buildings or providing sanitary hot water [Yuan] [Ebrahimi]. The effectiveness of these approaches should be measured using standardized metrics, not limited to PUE [Avelar] but also including Water Usage Effectiveness (WUE [Patterson]) and Energy Reuse Effectiveness (ERE [Patterson]).

Conversely, HPC systems are increasingly dependent on their environment. In the event of a grid failure, systems must be capable of limiting power consumption while minimizing the impact on production. With intermittent renewable energy sources playing a growing role in electricity generation [“Renewable energy statistics”], grid feedback is no longer confined to rare events but may become a frequent operational consideration [Smith] [McKinsey & Company]. Rising electricity prices in Europe and the need to avoid peak consumption periods further highlight this challenge. As a result, system operators must anticipate and manage these constraints through specialized tools designed to control hardware power consumption (including accelerators), optimize workload prioritization, and mitigate job duration variability.

Beyond energy consumption, the carbon footprint of these infrastructures is becoming a major societal concern [OECD]. It is increasingly necessary to provide users with real-time feedback on energy consumption and to incorporate the carbon intensity of electricity into job scheduling — evolving from energy-aware to carbon-aware scheduling. Leveraging system and application heterogeneity, along with differences in local carbon intensity, a further

extension of this concept lies in the federation of HPC systems, enabling jobs to be executed in the most carbon-efficient location [Li]. Provided data management is properly handled, this approach enables better resource sharing and reduces unnecessary infrastructure investments.

Achieving these goals requires several building blocks. Monitoring of energy, power, and carbon impact at multiple levels — both in time and across system components — is essential. Communication between the Resource and Job Management Systems (RJMS) and the grid operators must be formalized. Users also require new capabilities to express their preferences regarding energy, carbon, and performance trade-offs. These metrics must cover the entire scope of HPC centres and AI factories, including not only computing systems but also cooling infrastructures and buildings. In addition, a full Life Cycle Analysis (LCA), compliant with ISO 14040 standards [ISO], should be conducted — encompassing the design, operation, recycling, and disposal phases of all system components.

Reducing the overall environmental footprint requires action across each stage of this life cycle. While operational phase optimizations have been discussed, manufacturing and end-of-life impacts must also be addressed. Electronic components should comply with European regulations (e.g., PEF 3.0 [Fazio], WEEE [European Commission], RoHS) and adhere to recognized ecolabels (e.g., EPEAT [Global Electronics Council], TCO [TCO certified], Energy Star [Energy Star]). Cooling systems and buildings must be appropriately dimensioned relative to computational capacity. Given the evolving nature of most HPC centres and AI factories, modular or containerized infrastructure designs should be considered to enable flexible adaptation of cooling and building resources in line with computational demand fluctuations. JUPITER, the first European Exascale system, with its Modular Data Centre technology [Eviden], is a good example of such infrastructures.



2 HPC hardware-software co-design

Over the past decade, the energy efficiency of computing hardware has significantly improved. In 2014, the leading system on the Green500 [Top500.org] list achieved 5.3 GFLOPs/W. In contrast, today's most efficient system, JEDI, delivers an impressive 72.7 GFLOPs/W. It is important to note, however, that such levels of efficiency are typically attained using the HPL benchmark, which is highly optimized for the specific hardware architecture. In real-world scenarios, many applications still lack optimizations for hardware acceleration. As a result, they rely solely on CPU efficiency, which can only be harnessed through deliberate efforts by both developers and users. Continuing and even increasing support for HPC software developers is necessary to help them rewrite their applications using all possible accelerators, including but not limited to FPGAs or quantum machines.

Understanding the underlying hardware is imperative to achieving peak computational performance and energy efficiency. To exploit the full potential of modern architectures, several key aspects must be addressed, such as:

- **Increasing core counts**: Modern CPUs offer a growing number of cores, while to utilize the chip's memory bandwidth, it is not necessary to use them all. Thus, full core utilization is not always necessary or beneficial for performance and energy efficiency.
- **Chiplet-based architectures and NUMA domains**: The trend toward chiplet designs introduces additional NUMA domains. This demands improved thread affinity management and memory allocation strategies to minimize data movement and latency.
- **High Bandwidth Memory (HBM)**: HBM provides significantly higher throughput than conventional DDR memory. Nonetheless, its peak bandwidth is achievable only under stream access patterns. Developers must identify memory allocations suitable for HBM and adapt data access patterns accordingly to benefit from the performance gains. [Jun et al.]
- **High-end instruction set support**: Both CPUs and GPUs support various instruction sets. To fully harness hardware capabilities, developers must target high-level vector and matrix instruction sets. Moreover, the recent rise of AI has started to lead to limited hardware support for FP64. Thus, exploration of mixed-precision computation usage might be necessary [Kashi et al.].

Despite the increasing complexity of the computing hardware, most software users are not incentivized to optimize the efficiency of the code they execute. They typically depend on developers to implement such enhancements. In many scientific fields, applications still rely on legacy software developed years ago—software that is inherently incompatible with modern hardware optimizations [Amaral and al.].

This situation must not be passively accepted. Allowing outdated software to persist on cutting-edge systems will only exacerbate inefficiencies. Instead, proactive investment in the development of modern replacement applications is necessary. While the Technology Readiness Level (TRL) of such software may currently be low, this should not hinder funding and support for its development.

A datacentre resource usage monitoring is the critical first step on the path to optimise system usage. Without granular visibility, it is impossible to determine which software is being executed and how efficiently it utilizes system resources [“Software Environment - Tools”]. Once this insight is obtained, datacentres can direct training and advanced user support toward individuals or groups that underutilize hardware, thereby driving more efficient system usage overall.

Europe is investing in semiconductor development to enhance its sovereignty in this strategic autonomy in this critical domain [“European Processor Initiative”]. As well as software must evolve to match advances in hardware, the hardware itself should be tailored to meet the specific requirements of target applications. Many HPC-focused Horizon projects have embraced hardware-software co-design tasks [Kulkarni et al.]. However, the influence of software feedback on chip design should be further reinforced, particularly within the context of EuroHPC Centres of Excellence (CoE) projects.



3 Application workflows: Software and Data Stack

In addition to ongoing research on the hardware side, energy efficiency in HPC must be significantly enhanced through software. Several research avenues can be pursued in this context.

One approach focuses on designing energy-efficient algorithms that minimize computational load and energy consumption in HPC systems. This includes developing algorithms that reduce computational complexity and creating approximation algorithms that provide near-optimal solutions with substantially less computational effort. Particularly relevant for HPC are parallel algorithms that efficiently distribute workloads across multiple processors. This application-layer consideration primarily concerns end-users of HPC systems. To encourage adoption, it is essential to provide user-friendly software libraries implementing these algorithms and to highlight not only their computational benefits but also their impact on power consumption.

Beyond algorithm research, another key aspect of improving energy efficiency involves optimizing data management and data movement in HPC systems based on specific use cases [Khelilli]. Data reduction techniques such as compression, deduplication, and filtering are crucial for saving energy during storage and transmission. Both lossless and lossy data compression methods can significantly reduce data size, thus conserving storage space and energy. Deduplication strategies, which eliminate redundant data copies, are especially useful in environments with large datasets. Data filtering methods that remove unnecessary or irrelevant data before processing can reduce the volume of data that needs to be processed, thereby saving computational resources and energy. Closely related is the efficient performance of I/O operations on data. Techniques like asynchronous I/O, which allow overlapping computation and I/O activities, can minimize idle times and energy waste [Jin et al.] [Tseng et al.]. Buffering and caching reduce disk access, as accessing data from memory is far more energy efficient. Parallel I/O strategies that distribute operations across multiple nodes or devices enhance throughput and reduce the time and energy required for data transfer; power could for example be a new KPI for I/O and storage. Improving monitoring infrastructures could be an interesting topic in the near future.

If optimizing algorithms and data management is not feasible or needs to be supplemented, energy efficiency in HPC can also be improved by optimizing application workflows. Effective workflow design involves breaking down tasks into smaller, manageable sub-tasks and optimizing their sequence. Dynamic resource allocation adjusts resources based on current workload demands, preventing both underutilization and over-provisioning [Carretera et al.]. Automation tools and frameworks can streamline workflows, reduce manual intervention, and optimize resource usage in real time. Additionally, by quantifying the sustainability or energy efficiency of resource allocation, one can enhance standard dynamic resource allocation through power-aware job scheduling. Scheduling algorithms that prioritize jobs based on their power consumption profiles can significantly reduce energy use. Jobs can be scheduled to run during off-peak energy hours or distributed to nodes with lower power usage [Menear et al.]. Implementing energy profiling techniques for different jobs and workloads provides data for informed scheduling decisions, balancing performance and energy efficiency. Multi-objective scheduling algorithms that consider both performance and energy consumption aim to strike a balance between minimizing energy usage and maintaining acceptable performance levels.

However, the effort is not limited to the software itself. Reproducibility and, more generally, packaging and management of large-scale experiments should be strengthened. Long-duration experiments must be reliable to prevent wasting computing resources. Currently, no packaging software is reliable enough to provide such an environment, allowing experiments to run swiftly on different architectures, different supercomputers, and at different scales. Such an environment must be augmented with an open diffusion policy, taking into account reusability, to optimize the overall software life cycle.

On the systems software side, a EuroHPC funded project [SEANERGYS] is already starting to address the key aspects of Monitoring, AI based telemetry and system data analytics and Dynamic Resource Management for energy efficiency.



4 Optimizing HPC Lifecycle and Sustainability

First defined in 2010 by NASA, the digital twin (DT) concept usually refers to virtual representations of a physical entity or system. Applied to the field of HPC, it could help predict bottlenecks, risk of malfunction, adjust necessary maintenance intervals, and gain knowledge on the potential of possible optimizations. We recognise several levels of digital twins. The base is to have a 3D model of a datahall and its residing computing resource. The next step is to feed the model with continuous monitoring data. Such a database can be used to create a model to predict future behaviour. Finally, using simulations, the resource can be optimized already at the moment of system design, or later to optimize its operation. Another example is NVIDIA Air to create digital twins of the network fabric, see NVIDIA Air [NVIDIA].



Obviously, to be able to impact the system construction parameters, the system digital twin must be created at the early stage of the system procurement. Efficiency and consumption predictions already delivered by a vendor of different parts can help in ensuring that the datacentre's cooling and power infrastructure is scaling sufficiently to stay within the envisaged Total Cost of Ownership (TCO) and power envelopes already in the procurement phase. In a time when the five-year maintenance of any HPC system by far exceeds the initial cost of acquisition, power usage, and TCO are aspects of high importance in the dialogue with the funding agencies. They can ensure that funding is applied effectively and in accordance with the very definition of sustainability.

Monitoring and the lessons learned from the resulting virtual representation of an existing HPC resource not only increase the optimized usage of the system, prolonging its productive life that way, but also avoid single-use-only and trash-everything-at-once scenarios. Later on, it can provide additional review material for the final decision on when to shut down a system for good, besides the criteria of lack of hardware or software compatibility, or manpower expertise.

The digital twin can also give a better idea of the efficiency and sustainability of the HPC system. It should be able to answer questions such as:

- What are the spatial installation limitations of a new system in the data hall?
- Does the network experience conjunctions? If yes, when and where does the problem arise?
- Does the system show non-standard behaviour?
- Is the cooling system properly configured to avoid over- and under-cooling?

And many more complex questions that are hard to answer without understanding the context of each part of the complex system as an HPC/AI cluster. In general, the DT should not only describe the current status of the system but, most importantly, predict its future status and provide feedback for optimization. An increased understanding of where optimization efforts could lead to the most gain (as provided by a good digital twin) will lead to a continued improvement of infrastructure and supercomputer power efficiency. Showing where power and, thereby, money get regularly lost is highly relevant when developing cost models while operating a datacentre.

The ability to simulate possible what-if scenarios may also increase the willingness of some sites to experiment with deploying some of the dynamic resource management solutions. These DRSMs (job scheduling, meta-scheduling, job moldability & malleability, power management, etc.) often behave as a black box,

which makes them administrator unfriendly, while their positive impact can be huge. Understanding the behaviour through simulations may change the situation.

Within the industry, the DT concept is already getting standardized (IPC-2551, ISO/IEC 30173, ISO 23247). Concerning HPC systems, yet a non-negligible amount of work must be done. Datacentre monitoring systems should be harmonised across sites, standard metrics must be defined that reliably reflect all aspects of resource usage, including but not limited to power, water usage, CO_2 production, heat management, computing hardware workload describing metrics, or network communication. The international initiative ExaDigit, led by Oak Ridge National Laboratory, starts with defining such standards for a datacentre digital twin to be used by contributing partners. Knowledge in constructing and using DT is still growing, and we are learning by doing. A European knowledge exchange initiative on digital twins within HPC, followed by a consequent train-the-trainers programme, might be desirable.

Looking at digital twins for HPC systems as a whole, we can conclude that we consider digital twins as an excellent lifecycle management tool that allows for the emphasis on an end-to-end approach of the represented system. With a focus on best trade-offs instead of only one single issue at a time, it can help decision-makers optimize for economic and ecological impact. Insights gained through the DT process are essential to support communication between all relevant players (vendors, resource providers, policymakers, users, and software developers) within an HPC ecosystem so they can understand their own role on a combined path toward increased sustainability.

5 Raising Awareness and Empowering Stakeholders

Achieving significant progress in energy efficiency and sustainability (EE&S) within HPC and AI systems inevitably requires that all key stakeholders — especially those influencing procurement decisions and system usage — develop a comprehensive understanding of the energy and environmental footprint of these infrastructures. This awareness must extend beyond direct energy consumption to also include indirect or second-order effects, such as the Jevons paradox: when a system becomes more efficient, users may increase their consumption, offsetting part or all of the gains achieved — a phenomenon known as the rebound effect [York and McGee] [Alcott et al.].

Among these critical stakeholders are application developers, end-users, and funding bodies, who collectively shape system usage patterns and procurement strategies.

Historically, improvements in hardware efficiency have been a key driver of ICT expansion over the past decades [Koomey et al.]. Continued investment in this domain remains essential — targeting both computing electronics and cooling technologies. However, these hardware efficiency gains have predominantly been leveraged to scale systems up, increasing their performance but also their overall energy consumption. Moving forward, it is clear that additional levers beyond hardware efficiency must be activated [AFNOR Editions] [Mistral AI].

A fundamental step in this direction is to provide stakeholders with a holistic view of the environmental impact of their actions throughout the entire life cycle of HPC systems. This perspective should not only cover the operational phase but also the upstream and downstream stages: resource extraction, manufacturing, system integration, infrastructure construction, system decommissioning, disassembly, and recycling. Some approaches even consider software development activities as part of this comprehensive environmental assessment. This methodology, known as Life Cycle Assessment (LCA), is formally defined by ISO standards 14040 and 14044 and should be systematically applied.

To support this approach, it is essential that users, developers, and system administrators have accurate and transparent insights into their hardware and software usage. Reliable monitoring infrastructures are therefore a critical component. Monitoring should go beyond basic energy and power consumption metrics to also capture all relevant data needed to 1) construct meaningful EE&S Key Performance Indicators (KPIs) and 2) provide actionable insights to improve them. Wherever possible, such metrics should be standardized and vendor-agnostic [resilio] [LCIE DEPARTMENT CODDE], ideally integrated within open-source and normative frameworks to ensure comparability and interoperability [SEANERGYS] [ALUMET] [PowerAPI].

Based on these monitoring systems, administrators should be able to derive comprehensive indicators for their infrastructure, including Power Usage Effectiveness (PUE), Total Cost of Ownership (TCO), carbon emissions, Water Usage Effectiveness (WUE), and Energy Reuse Effectiveness (ERE). Such indicators foster a constructive benchmarking environment, enabling infrastructure operators to compare their sustainability performance and identify areas for improvement.

Application developers would also benefit from this monitoring capability by identifying bottlenecks or inefficiencies in their code, allowing them to optimize applications for higher performance and energy efficiency. Similarly, end-users could leverage this information to better understand the impact of their workloads, enabling them to configure and schedule jobs in a more responsible and efficient manner.

Providing direct feedback to users is, in general, a powerful mechanism for driving energy efficiency improvements. This feedback can take various forms, such as detailed job reports, dashboards, or graphical interfaces displaying KPIs and scoring mechanisms [SEANERGYS].

Users can only take action towards optimization if they are first made aware of their potential for improvement, supported by accurate and reliable information on energy consumption and system behaviour.

Beyond awareness, it is essential to assist users in adopting more energy-conscious behaviours. This can be achieved by proposing optimized usage scenarios that balance performance and energy consumption based on the operational context [“Workload resampling for performance evaluation of parallel job schedulers”] [IEEE Staff] [Cerf]. These scenarios could be generated using digital twins or frugal AI algorithms — ensuring that the energy consumption of these auxiliary systems remains lower than the savings they enable. Guidance can be provided through predictions, recommendations, nudges, or incentives. Recommendations may simply highlight the most efficient or highest-performance configurations. Nudges could encourage responsible behaviour through gamification or peer comparison. Incentives might involve mechanisms such as job queue prioritization or energy-aware credits to reward users who adopt more efficient practices.

Finally, energy efficiency and sustainability must become central criteria for funding agencies when evaluating proposals for HPC systems or AI factories acquired with public or institutional funds. These agencies hold significant influence and can act as major drivers of change. In procurement processes, they should assign substantial weight to sustainability indicators — including expected PUE, WUE, energy reuse strategies, carbon footprint, system lifetime, and complete LCA assessments — while requiring transparent reporting and compliance with ISO standards. Additionally, software environments delivered with these systems should provide built-in mechanisms to continuously report on these sustainability metrics throughout the system’s operational lifespan.



Post Exascale Vision

The aspect of energy efficiency and sustainability will become even more important in the post-exascale era. Simple math easily shows that a plain scaling will not be possible. This is, of course, not new, but the order of magnitude that has been reached leaves little room for plain scaling. Especially when taking into account that the growth of the need for AI computing resources is tremendous since its applications reach much further than those of classic HPC. In combination with the challenge to reduce not only the CO_2 footprint but the need for energy in general there is a significant necessity to make progress on several elements of the whole ecosystem as well as over the complete life cycle. This will not be successful while only focusing on one or two of these aspects.

It is, therefore, crucial that progress will be made on several of these elements. At the same time, this needs to be linked to a holistic perspective that sets the different measures into an overall context. The combination of these parts will lead to much more than just the sum of it. Aspects that have been neglected in the past (e.g. ecologically optimised programming in order to, for example, consume less energy for a given task, even if it might take a little more compute time) should be ranked higher in the priority list. Especially in the rapidly evolving field of AI computations, it might be a good idea to establish this as an important goal in addition to those already existing. This is even more important since the growth of speed per invested Euro in the TOP500 list is far from the gains that could be reached in the past.



Key R&I Recommendations

- Improvements in efficient data centre cooling infrastructure should be developed (warm water cooling, heat reuse, elimination of water evaporation, etc.)
- Power monitoring systems should be harmonised, a standard metric must be defined that reliably reflects the “real“ power consumption
- All relevant stakeholders should start working on an HPC/AI system digital twin that provides the basis for a more realistic prediction of power consumption
- Continuous monitoring of the efficient usage of the system resources should be established, and full-stack (users, applications, RJMS, …) power management capabilities should be made available
- AI should be used to optimise HPC sustainability through dynamic scheduling, automated hardware/software configuration, and OODA feedback loops. These approaches enable real-time energy efficiency improvements, reduce operational waste, and provide data-driven insights for continuous system optimisation.
- Project funding should prioritise initiatives like dedicated energy-efficiency calls, standardised energy/CO_2 accounting frameworks, life-cycle assessments of HPC/AI systems, and interdisciplinary research. This includes developing digital twins and integrating social/legal sciences to drive systemic and behavioural change toward sustainability.

Conclusions

While the efficiency of systems when it comes to the usage of energy has been of relevance for a while, increasingly so when the energy costs briefly skyrocketed in 2022, at the same time, most of the attention has also been limited to it. It is extremely important to widen the focus, both in the temporal scale spanning from the earliest beginning (planning) to the final (destruction) and recycling phase, but also beyond pure energy usage. Sustainability is more than just the electricity meter and the related bill.

It is mandatory to include application and software-oriented aspects that go beyond the hardware, while, at the same time, the stakeholders' awareness of EE&S around the usage of HPC and AI systems must be raised.

The EC, as a major funding authority, obviously plays an important role in this, but the challenge is so big that others, like the member states, but also vendors and large users, should be approached as well. A combined effort is necessary to bring the importance of EE&S to the table. All stakeholders must be convinced that it is in their own genuine interest to invest effort here. This document's goal is to help the decision makers identify the areas where they should put the most emphasis.



Contributors

Michaela Barth, KTH

Marta Chinnici, ENEA

Georges Da Costa, University of Toulouse, IRIT

Esa Heiskanen, CSC

Sarah Neuwirth, Universität Mainz

Frédéric Pariente, NVIDIA

Gert Svensson, KTH

Chiara Vercellino, Fondazione Links

References

AFNOR Editions. “Référentiel général pour l'IA frugale - Mesurer et réduire l'impact environnemental de l'IA.” afnor Editions, 06 2024, https://www.boutique.afnor.org/fr-fr/norme/afnor-spec-2314/referentiel-general-pour-lia-frugale-mesurer-et-reduire-limpact-environneme/fa208976/421140#. Accessed 10 12 2025.

Alcott, Blake, et al. The Jevons Paradox and the Myth of Resource Efficiency Improvements. Taylor & Francis Group, 2012. Taylor & Francis, https://www.taylorfrancis.com/books/mono/10.4324/9781849773102/jevons-paradox-myth-resource-efficiency-improvements-john-polimeni-kozo-mayumi-blake-alcott-mario-giampietro. Accessed 10 December 2025.

ALUMET. “Measure your Energy Consumption.” Alumet, 2025, https://alumet.dev/. Accessed 10 12 2025.

Amaral, Vasco, and et. al. “Programming languages for data-Intensive HPC applications: A systematic mapping study.” Parallel Computing, vol. 91, no. n/a, 2019, n/a. ScienceDirect, https://www.sciencedirect.com/science/article/abs/pii/S0167819119301759. Accessed 19 12 2025.

Avelar, Victor et al. “PUE™: A COMPREHENSIVE EXAMINATION OF THE METRIC.” PUE™: A COMPREHENSIVE EXAMINATION OF THE METRIC, 2012, https://www.thegreengrid.org/en/resources/library-and-tools/20-PUE:-A-Comprehensive-Examination-of-the-Metric. Accessed 10 12 2025.

Benhari, Abdessalam et al. “Green HPC: An analysis of the domain based on Top500.” Green HPC: An analysis of the domain based on Top500, 2024, https://arxiv.org/abs/2403.17466. Accessed 10 12 2025.

Carretera, Jesus, et al. “Dynamic Resources Management for Exascale in ADMIRE Framework.” Procedia Computer Science, vol. 267, no. n/a, 2025, pp. 165-175. ScienceDirect, https://www.sciencedirect.com/science/article/pii/S1877050925025852. Accessed 19 12 2025.

Cerf, Sophie. “Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach.” web.cels.anl.gov/, 2021, https://web.cels.anl.gov/projects/argo/publications/2021-europar-cerf.pdf. Accessed 10 12 2025.

Chen, Hongwei et al. “Current Status and Challenges for Liquid-Cooled Data Centers.” Frontiers in Energy Research, vol. 10, no. ???, 2022, p. 10, https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2022.952680/full. Accessed 10 12 2025.

D'Agostino, Daniele, et al. “Hardware and Software Solutions for Energy-Efficient Computing in Scientific Programming.” Scientific Programming in the Fog and Edge Computing Era, vol. Special Issie, no. 1, 2021, n/a. Wiley Online Library, https://onlinelibrary.wiley.com/doi/10.1155/2021/5514284. Accessed 19 12 2025.

Daraghmeh, Hafiz M., and Chi-Chuan Wang. “A review of current status of free cooling in datacenters.” Applied Thermal Engineering, vol. 114, no. ???, 2017, p. 16. Science direct, https://www.sciencedirect.com/science/article/abs/pii/S135943111632405X. Accessed 10 12 2025.

Deep Projects. “LLview.” LLview, 2025, https://deep-projects.eu/modular-supercomputing/software/monitoring/llview/. Accessed 10 12 2025.

Ebrahimi, Khosrow. “A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities.” Renewable and Sustainable Energy Reviews, vol. 31, no. ???, 2014, pp. 622-638. Renewable and Sustainable Energy Reviews, https://files.ifi.uzh.ch/hilty/t/Literature_by_RQs/RQ%20104/2014_Ebrahimi_Jones_Fleischer_A_Review_of_data_center_cooling_technology.pdf. Accessed 10 12 2025.

Energy Star. “The simple choice for saving energy.” Energy Star, 2025, https://www.energystar.gov/. Accessed 10 12 2025.

European Commission. “Environmental Footprint Methods - Calculating the environmental impact of products and services.” Green Forum European Commission, 2024, https://green-forum.ec.europa.eu/green-business/environmental-footprint-methods_en. Accessed 10 12 2025.

“European Processor Initiative.” European Processor Initiative, 2025, https://www.european-processor-initiative.eu/. Accessed 19 12 2025.

Eviden. “Eviden provides the modular data center to host Europe's first exascale supercomputer.” Eviden, 2024, https://eviden.com/insights/press-releases/eviden-provides-the-modular-data-center-to-host-europes-first-exascale-supercomputer/. Accessed 10 12 2025.

Fazio, S. et al. “Guide for EF compliant data sets.” JRC, 2020, https://eplca.jrc.ec.europa.eu/permalink/Guide_EF_DATA.pdf. Accessed 10 12 2025.

Feng, Wu-chun. “The Green500 List: Encouraging Sustainable Supercomputing.” Computer, vol. 40, no. no. 12, 2007, p. 6. IEEE explore, https://ieeexplore.ieee.org/document/4404810. Accessed 10 12 2025.

Global Electronics Council. EPEAT® is the premier global ecolabel for electronics and technology products. 2025. EPEAT® is the premier global ecolabel for electronics and technology products., https://epeat.net/. Accessed 10 12 2025.

IEEE Staff. 2022 IEEE Conference on Control Technology and Applications (CCTA). IEEE, 2022. Adaptive Power Control for Sober High-Performance Computing, https://ieeexplore.ieee.org/abstract/document/9966115/references#references. Accessed 10 December 2025.

International Energy Agency. Energy and AI. Energy and AI, https://iea.blob.core.windows.net/assets/601eaec9-ba91-4623-819b-4ded331ec9e8/EnergyandAI.pdf. Accessed 10 12 2025.

ISO. “ISO 14040:2006 Management environnemental — Analyse du cycle de vie — Principes et cadre.” ISO, 2006, https://www.iso.org/fr/standard/37456.html. Accessed 10 12 2025.

Jin, Sian, et al. “Concealing Compression-accelerated I/O for HPC Applications through In Situ Task Scheduling.” HAL thesis, 2024, https://inria.hal.science/hal-04225758v1. Accessed 19 12 2025.

JRC. “Energy Consumption in Data Centres and Broadband Communication Networks in the EU.” JRC Publications Repository, 16 02 2024, https://publications.jrc.ec.europa.eu/repository/handle/JRC135926. Accessed 10 12 2025.

Jun, Hongshin, et al. “HBM (High Bandwidth Memory) DRAM Technology and Architecture.” 2017 IEEE International Memory Workshop (IMW), Monterey, CA, USA, vol. 2017, no. n/a, 2017, pp. 1-4. IEEE Xplore, https://ieeexplore.ieee.org/abstract/document/7939084. Accessed 19 12 2025.

Kashi, Aditya, et al. “Mixed-precision numerics in scientific applications: survey and perspectives.” Arxiv.org, 2024, https://arxiv.org/html/2412.19322v1. Accessed 19 12 2025.

Khelilli, Adrian. “File placement prediction in a hierarchical storage system through lifecycle exploitation for HPC.” HAL theses, 2025, https://theses.hal.science/tel-05387620. Accessed 19 12 2025.

Koomey, Jonathan, et al. “Implications of Historical Trends in the Electrical Efficiency of Computing.” IEEE Annals of the History of Computing, vol. Volume 33, no. Issue: 3, 2011, pp. 46-54. *IEEE explore*, https://ieeexplore.ieee.org/document/5440129. Accessed 10 12 2025.

Kulkarni, Kajol, et al. “Harvesting energy consumption on European HPC systems: Sharing Experience from the CEEC project.” Arxiv.org, 2025, https://arxiv.org/html/2511.03029v1. Accessed 19 12 2025.

LCIE DEPARTMENT CODDE. “NEGAOCTET.” DIGITAL POLLUTION, 2025, https://codde.fr/en/our-software/negaoctet. Accessed 10 12 2025.

Li, Baolin et al. “Toward Sustainable HPC: Carbon Footprint Estimation and Environmental Implications of HPC Systems.” Cornell University, 19 11 2023, https://arxiv.org/abs/2306.13177. Accessed 10 12 2025.

McKinsey & Company. “How grid operators can integrate the coming wave of renewable energy.” McKinsey, 2024, https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/how-grid-operators-can-integrate-the-coming-wave-of-renewable-energy. Accessed 10 12 2025.

Menear, Kevin, et al. “Energy-Aware HPC Scheduling with LLM-Based Power Prediction.” SC Workshops '25: Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, vol. 2025, no. n/a, 2025, pp. 1997-2006. ACM digital library, https://dl.acm.org/doi/10.1145/3731599.3767563. Accessed 19 12 2025.

Mistral AI. “Our contribution to a global environmental standard for AI.” Mistral, 22 07 2025, https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai. Accessed 10 12 2025.

NVIDIA. “NVIDIA Air - Networking.” NVIDIA, 2022, https://www.nvidia.com/en-us/networking/ethernet-switching/air/. Accessed 19 12 2025.

OECD. “Measuring the environmental impacts of artificial intelligence compute and applications.” OECD Publications, 15 11 2022, https://www.oecd.org/en/publications/measuring-the-environmental-impacts-of-artificial-intelligence-compute-and-applications_7babf571-en.html. Accessed 10 12 2025.

Patterson, Michael et al. “WATER USAGE EFFECTIVENESS (WUE™): A GREEN GRID DATA CENTER SUSTAINABILITY METRIC.” WATER USAGE EFFECTIVENESS (WUE™): A GREEN GRID DATA CENTER SUSTAINABILITY METRIC, 2011, https://airatwork.com/wp-content/uploads/The-Green-Grid-White-Paper-35-WUE-Usage-Guidelines.pdf. Accessed 10 12 2025.

Patterson, Mike et al. “ERE: A Metric for Measuring the Benefit of Reuse Energy From a Data Center.” WP# 29 - ERE: A Metric for Measuring the Benefit of Reuse Energy From a Data Center, 2010, https://datacenters.lbl.gov/sites/default/files/EREmetric_GreenGrid.pdf. Accessed 10 12 2025.

PowerAPI. “PowerAPI.” Measuring power consumption of your applications., 2025, https://powerapi.org/. Accessed 10 12 2025.

“Renewable energy statistics.” Statistics Explained, 2024, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics. Accessed 10 12 2025.

resilio. “Towards sustainable digitalization.” resilio, 2025, https://resilio-solutions.com/en. Accessed 10 12 2025.

SEANERGYS. “SEANERGYS Energy Efficient EXASCALE.” Deep Projects, 2025, https://deep-projects.eu/project/seanergys/. Accessed 10 12 2025.

Sencan, Efe, et al. “Analyzing GPU Utilization in HPC Workloads: Insights from Large-Scale Systems.” PEARC '25: Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration, vol. n/a, no. Article No. 14, 2025, pp. 1-8. ACM Digital Library, https://dl.acm.org/doi/10.1145/3708035.3736010. Accessed 19 12 2025.

Smith, Oliver et al. “The effect of renewable energy incorporation on power grid stability and resilience.” ResearchGate, 2022, https://www.researchgate.net/publication/358984174_The_effect_of_renewable_energy_incorporation_on_power_grid_stability_and_resilience. Accessed 10 12 2025.

Suarez, Estela, et al. “Energy efficiency trends in HPC: what high-energy and astrophysicists need to know.” Frontiers in Physics, vol. Volume 13, no. issue, 2025, ?? Frontiers in Physics, file:///Users/sicos_aw/Downloads/Energy_efficiency_trends_in_HPC_what_high-energy_a.pdf . Accessed 10 12 2025.

“Software Environment - Tools.” EuPEX, 2025, https://eupex.eu/sw-environment/sw-environment-tools/ . Accessed 19 12 2025.

Solórzano, Ana Luisa Veroneze, et al. “Toward Sustainable HPC: In-Production Deployment of Incentive-Based Power Efficiency Mechanism on the Fugaku Supercomputer.” IEEE explore, 2024, https://ieeexplore.ieee.org/document/10793214. Accessed 10 12 2025.

TCO certified. “Together toward sustainable IT.” TCO certified, 2025, https://tcocertified.com/. Accessed 10 12 2025.

Top500.org. “Green500.” Top500 The List, 2025, https://top500.org/lists/green500/. Accessed 10 12 2025.

Tseng, Shu-Mei, et al. “Demystifying asynchronous I/O Interference in HPC applications.” The International Journal of High Performance Computing Applications, vol. 35, no. 4, 2021, pp. 391-412. SageJournals, https://journals.sagepub.com/doi/10.1177/10943420211016511. Accessed 19 12 2025.

“Workload resampling for performance evaluation of parallel job schedulers.” Semantic Scholar, 21 04 2013, https://www.semanticscholar.org/paper/Workload-resampling-for-performance-evaluation-of-Zakay-Feitelson/1b96a384332596d2a42056b5aba34840d09ba7fd. Accessed 10 12 2025.

York, Richard, and Julius Alexander McGee. “Understanding the Jevons paradox.” Environmental Sociology, vol. 2, no. Issue 1, 2015, pp. 77-87. Taylor & Francis Online, https://www.tandfonline.com/doi/full/10.1080/23251042.2015.1106060#. Accessed 10 12 2025.

Yuan, Xiaoleu et al. “Waste heat recoveries in data centers: A review.” Renew*able and Sustainable Energy Reviews*, vol. 188, no. ???, 2023, p. 17, https://pdf.sciencedirectassets.com/271969/1-s2.0-S1364032123X0015X/1-s2.0-S1364032123006342/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEBEaCXVzLWVhc3QtMSJHMEUCIHmMSv4R8cVEmW3rKiUv82cCdL6dh64Io3GUMOB0ePd8AiEAzEQE9uhUep6wI2eTD6YfGxo3vohKVAnN3rfKrBe%2BfQA. Accessed 10 12 2025.



Cite as: Wierse A. Seroul P., Vysocky O. et al., ETP4HPC SRA6 White Paper – Energy Efficiency and Sustainability, 2026, ETP4HPC. https://doi.org/10.5281/zenodo.18350791

DOI: 10.5281/zenodo.18350791

© ETP4HPC 2026