theLisboncouncil
Lisbon Council Research

Special Report
1/2024

Sustainable Computing for a Sustainable Planet

How Technological Innovation and Good Governance Can Help Unleash the Benefits of Artificial Intelligence Without Compromising the Green Transition

In less than two years, artificial intelligence (AI) has taken the world by storm, opening new possibilities for scientific progress, innovation and productivity gains. But alongside the awe, this new technological disruption has attracted loud concerns about its possible harms, notably in terms of fast-growing energy consumption. The base case prediction by the International Energy Agency expects data centres to double their electricity use by 2026 compared to 2022, when they were responsible for 2% of total electricity global consumption.[1]

The policy question is then straightforward: can the combined goals of scientific progress and the green transition be achieved? How can we make computing not only more powerful but also more sustainable?

The good news is that Europe, like other modern economies, has been able to decouple growth from carbon emission, to produce "more from less" as Massachussets Institute of Technology's Andrew McAfee titled his recent book. The same change – though to a much greater degree – is already happening in AI. Thanks first to Moore's law and now to new architectures, the progress of computing power has decoupled from energy consumption. Today's new processors deliver an order of magnitude more computing power with the same amount of energy – and the ratio keeps falling. This does not mean that the problem is solved; just as in many other aspects of the green transition, innovation needs to continue and even accelerate.[2]

> *'We can make computing not only more powerful but also more sustainable.'*

[1] This special report is the result of a six-month research project led by Lisbon Council Research and supported by NVIDIA Corporation, the leading provider of graphic processing units and software for AI. Throughout, NVIDIA Corporation served as a key knowledge partner, often answering questions from the curious authors. The authors would like to thank Giorgio Amati, Alex Borg, Ned Finkle, Dion Harris, Paul Hofheinz, Bea Longworth, Alberto Mittestainer, Cristina Moise, Joshua Parker, Anna Pizzamiglio and Liviu Știrbăț. The views expressed here are independent and any error of fact or judgement are the authors' sole responsibility. For the latest data points on data centres energy consumption, see International Energy Agency, *Electricity 2024* (Paris: IEA, 2024).

[2] On the crucial role of innovation for climate change, see Dirk Pilat, "Driving Innovation to Curb Climate Change: Recommendations for COP 27," *Lisbon Council Interactive Policy Brief*, 30, 2022; Andrew McAfeee, *More From Less* (New York: Simon & Schuster Paperbacks, 2019).

This policy brief investigates the most recent computing innovations, why they matter and what policy tools can be used to accelerate innovation and manage trade-offs. Among the key findings:

- While energy consumption for training large language models has increased substantially, hardware and software innovation such as accelerated computing is progressing so fast that the overall energy consumption is growing much less quickly than computing requirements and performances. This is particularly evident for the deployment of the models in specific contexts based on live data (the process known as inference), which is responsible for the vast majority of energy consumption.

- The adoption of AI is already generating energy efficiency benefits across all economic activities. Even if the predictions that data centres will soon account for 4% of global energy consumption become a reality, AI is having a major impact on reducing the remaining 96% of energy consumption. It also plays a crucial role in the innovation needed to address climate change (e.g., new materials for batteries) and contributes to solving many other human challenges (e.g., drug development). AI adoption accelerates scientific discovery across all scientific fields.[3]

- More transparency is needed on data centres' energy consumption and the recast energy efficiency directive (EED recast) tackles this head-on. However, the choice of power usage efficiency (PUE) as a key performance indicator, as established in the recently approved EED delegated act, is harmful as it does not account for data centres' output and performance. This will discourage green innovation and energy efficiency in general across the economy, both in and through AI, ultimately harming the fight against climate change.[4]

- Existing metrics are a work-in-progress. New, multi-parameter indicators that include computing performance should be developed in the context of global fora such as the G7 and the Organisation for Economic Co-operation and Development (OECD). The data provided through the EED recast reporting requirements should be treated as an important starting point feeding into this process.

Accelerated Computing, Accelerated Science

The nature of science is continuously evolving. In the last 20 years, virtually every field of science has developed a computational branch, so much so that scholars have put forward the idea of a fourth paradigm of scientific discovery.[5] Since 2019, the opportunities offered by machine learning and generative AI have radically accelerated science in many different fields, allowing unprecedented progress.

> *'The acceleration of science is not a goal* per se, *but it brings invaluable benefits.'*

This acceleration of science is not a goal per se, but it brings invaluable benefits. The COVID-19 vaccine is the clearest example. AI played a fundamental role in the development of the vaccine, which was developed and deployed in less than a year (the previous record for

[3] For an overview of the role of machine learning in climate change see David Rolnick et al., "Tackling Climate Change With Machine Learning," *arXiv*, 2019.

[4] See European Parliament and the Council of the European Union, *Directive (EU) 2023/1791 of the European Parliament and of the Council of 13 September 2023 on Energy Efficiency and Amending Regulation (EU) 2023/955 (Recast)*, 2023 and European Commission, *Commission Delegated Regulation of 14.3.2024 on the First Phase of the Establishment of a Common Union Rating Scheme for Data Centres*, C(2024) 1639, 2024.

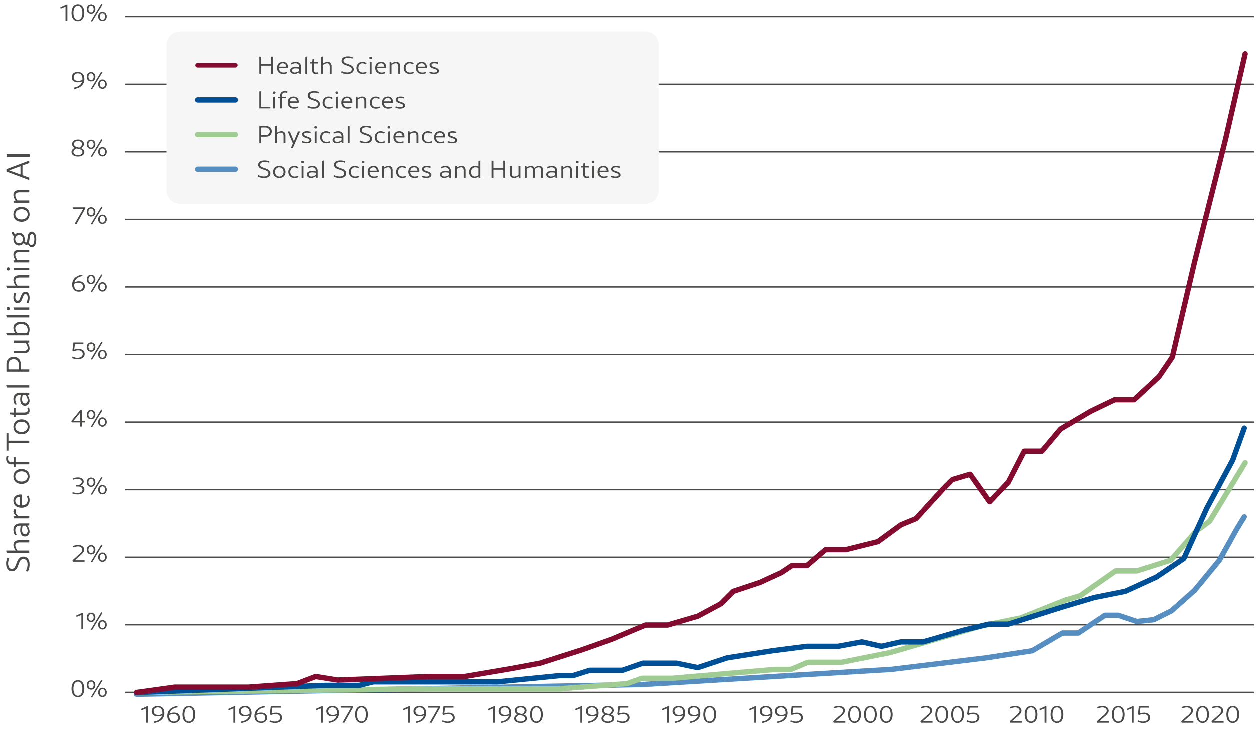
[5] See Tony Hey, Stewart Tansley and Kirstin Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Redmond: Microsoft Research, 2009); Jean-Claude Burgelman, David Osimo and Marc Bogdanowicz, "Science 2.0 (Change Will Happen….)," *First Monday*, 15(7), 2010.

the fastest vaccine was the mumps vaccine, which took about four years). This acceleration saved 1.4 million lives in Europe alone.[6] Faster weather predictions help save people from flooding, as do faster tsunami predictions. Faster time-to-market for new batteries reduces carbon emissions and their letal consequences. To put it briefly, accelerating science saves lives.

> *'Accelerating science saves lives.'*

This is why fast-growing adoption trends are visible across scientific fields, as shown in Figure 1.

Figure 1: Share of Publications on AI Application in Science over Total Publications per Field of Science[7]



Recent reports from both the European Commission and the OECD clearly spell out the advantages of AI in science: the pharmaceutical industry can reduce the time and cost involved in developing new drugs by analysing molecular structures, predicting drug interactions and identifying potential candidates for new medications. In materials science, AI can streamline the process of developing advanced materials for various applications by simulating material structures, predicting material behaviour and optimising material properties. In climate science, AI technologies are used in climate modelling, environmental monitoring and natural resource management to analyse complex environmental data, predict climate patterns and assess environmental risks. AI-driven insights help scientists better understand climate change and its impact on ecosystems.[8]

[6] The estimate is included in Ashwani Sharma, Tarun Virmani, Vipluv Pathak, Anjali Sharma, Kamla Pathak and Girish Kumar, "Artificial Intelligence-Based Data-Driven Strategy to Accelerate Research, Development, and Clinical Trials of COVID Vaccine," *BioMed Research International*, 2022; The WHO European Respiratory Surveillance Network, "Estimated Number of Lives Directly Saved by COVID-19 Vaccination Programs in the WHO European Region, December 2020 to March 2023," *medRxiv*, 2024.

[7] The source for the chart is Stefan Hajkowicz, Conrad Sanderson, Sarvnaz Karimi, Alexandra Bratanova and Claire Vaughtin, "Artificial Intelligence Adoption in the Physical Sciences, Natural Sciences, Life Sciences, Social Sciences and the Arts and Humanities: A Bibliometric Analysis of Research Publications from 1960-2021," *Technology in Society*, 74, 2023.

[8] See European Commission, Directorate-General for Research and Innovation, *AI in Science: Harnessing the Power of AI to Accelerate Discovery and Foster Innovation: Policy Brief* (Luxembourg: Publications Office of the European Union, 2023) and Organisation for Economic Co-operation and Development, *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research* (Paris: OECD Publishing, 2023).

The surge in AI adoption is expected to drive demand for computing power and energy. The most powerful supercomputers in 2023 have crossed the exascale threshold, and their performance has seen a 1,000-fold increase in 15 years.[9] The computing requirements for training machine learning models have accelerated exponentially and have recently been estimated to double every 5-6 months since 2010.[10]

'The surge in AI adoption is expected to drive demand for computing power and energy.'

However, the exponential growth of the computing power required for training models should not be taken as representative of the overall growth attributable to AI. The majority of AI energy consumption (from 60% to 80%, depending on the estimate) is attributed to running the models on live data (in technical terms, inference), rather than training the models. And recent studies clearly demonstrate that the energy consumption of inference is growing far more slowly than the energy consumption of training, and certainly not exponentially.[11]

The reason for this is simple. As computing demand increases, innovation in hardware and software is rapidly reducing energy consumption and improving energy efficiency.

European Union Performance in Science

According to the European Commission's recent assessment of the scientific performance of the European Union, the EU's major trading partners, such as China and the United States, have been outpacing the EU when it comes to innovation performance improvement in recent years. China has become the global leader in terms of the volume of scientific publications while the U.S. has retained its lead in terms of quality and impact. The European Union also lags behind its main international competitors in terms of the number of start-up and scale-up firms, and the number of unicorns in the EU is much lower than in the EU's main competitors.[12]

Grasping the opportunities of AI is crucial for European science and innovation to remain competitive globally. However, as the European Commission points out in a recent report, Europe is becoming less attractive for AI researchers: "The EU faces a significant brain drain problem: 20% of top EU AI researchers went to the U.S. for their graduate school, and a further 14% left for their post-graduate work." In addition, current data paint a worrying picture in terms of European scientists' adoption of AI. China has overtaken both the European Union and the United States in terms of both number and share of publications on AI applications in science.[13]

[9] Exascale refers to the capability of performing a billion billion (i.e., 10^18) calculations per second, commonly referred to as one exaflop. Exascale computing has the potential to enable breakthroughs in various scientific, engineering and societal domains by tackling complex problems and simulations that were not previously feasible due to computational limitations. This technology facilitates advanced simulations and modelling in fields such as climate science, astrophysics, materials science and biology. It enables researchers to explore complex phenomena and conduct large-scale simulations with unprecedented detail and accuracy. As of January 2024, only one exascale supercomputer exists, the U.S.-based Frontier.

[10] See Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn and Pablo Villalobos, "Computer Trends Across Three Eras of Machine Learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022.

[11] For these estimations, see Radosvet Desislavov, Fernando Martínez-Plumed and José Hernández-Orallo, "Trends in AI Inference Energy Consumption: Beyond the Performance-vs-Parameter Laws of Deep Learning," *Sustainable Computing: Informatics and Systems, 38, 2023;* Daniel Castro, *Rethinking Concerns About AI's Energy Use* (Center for Data Innovation, 2024).

[12] European Commission, Directorate-General for Research and Innovation, *Science, Research and Innovation Performance of the EU 2022: Building a Sustainable Future in Uncertain Times* (Luxembourg: Publications Office of the European Union, 2023).

[13] The data are included in European Commission, Directorate-General for Research and Innovation, *AI in Science: Harnessing the Power of AI to Accelerate Discovery and Foster Innovation: Policy Brief* (Luxembourg: Publications Office of the European Union, 2023); European Commission, Directorate-General for Research and Innovation, *Trends in the Use of AI in Science: A Bibliometric Analysis* (Luxembourg: Publications Office of the European Union, 2023).

The Advent of Accelerated Computing

The exponential progress in computing power is a long-standing trend encoded in Moore's law. In 1965, Intel co-founder Gordon Moore predicted that the number of transistors on a microchip would double approximately every two years due to shrinking transistor dimensions and other improvements, leading to an exponential increase in computing power and a decrease in cost per transistor. Moore's observation and extrapolation became a roadmap for the industry and the law held true for over half a century.[14]

Today, Moore's law remains valid not for the technical trajectory of computing but for the general idea of exponential growth of computing power. Computing power continues to progress exponentially, but fitting more transistors on chips is no longer possible. The current size of transistors in top-of-the-line microprocessors is around five nanometres across, which is smaller than most viruses. Shrinking transistor sizes has led to increased heat generation and power consumption, rather than the expected performance gains: "as electrons had to move faster and faster through silicon circuits that were smaller and smaller, the chips began to get too hot."[15]

Yet, new ways have been found to continue increasing computing power. First came parallel computing. Instead of relying solely on increasing the number of transistors on a single chip, parallel processing involves using multiple processors to handle computations simultaneously. Today's computers and smartphones typically hold multiple processors; the iPhone 14, for instance, has six core processing units.

Parallel computing has evolved into accelerated computing. In addition to running tasks in parallel across multiple processors, performance improvements can be achieved by adding specialised accelerators such as Graphics Processing Units (GPUs). While Central Processing Units (CPUs) have two to 64 cores, GPUs can have several thousand cores. Originally designed for managing the graphics of computer games, GPUs turned out to be extremely effective for AI model training tasks due to their unique architecture. Their ability to process many more computations simultaneously makes them particularly suited for applications requiring high computational throughput, such as graphics rendering, scientific simulations and machine learning tasks. GPUs can handle thousands of threads at once, significantly accelerating tasks that can be parallelised compared to traditional CPUs, which are optimised for sequential processing tasks. This specialisation in parallel processing enables GPUs to deliver substantial performance improvements in suitable computational tasks. A frequently used metaphor compares a CPU to a Ferrari and a GPU to a bus: to process one task quickly, a CPU works better, just like a Ferrari can quickly transport one or two people; a GPU on the other hand, takes more time to process one task, but is radically more efficient at processing many tasks, just as a bus is a more efficient way to transport 50 people as opposed to multiple trips in a Ferrari. Many supercomputers distribute workload between many GPUs and CPUs, depending

> *'Originally designed for managing the graphics of computer games, GPUs turned out to be extremely effective for AI model training tasks due to their architecture.'*

[14] See M. Mitchell Waldrop, "The Chips Are down for Moore's Law," *Nature News*, 530, 11 February 2016.

[15] See Tekla S. Perry, "Forget Moore's Law—Chipmakers Are More Worried About Heat and Power Issues," *IEEE Spectrum*, 11 October 2019.

on the task to be performed. For a concrete example, see the box on the Leonardo pre-exascale supercomputer on page 8.

When it comes to frontier applications like machine learning or deep learning, the performance of GPUs is an order of magnitude better than that of CPUs.[16] These performance advantages are particularly useful for the supercomputers running applications with the highest workload. Since 2010, accelerated computing has become the norm in top supercomputers; by 2023, nine of the top 10 supercomputers were using accelerators. Accelerators have somehow taken the baton from Moore's law in delivering exceptional performance.

> *'Among the top 500 supercomputers, those based on parallel computing have an average energy efficiency score of 22 gigaflops per watt consumed, against an average of four for those not using accelerated computing.'*

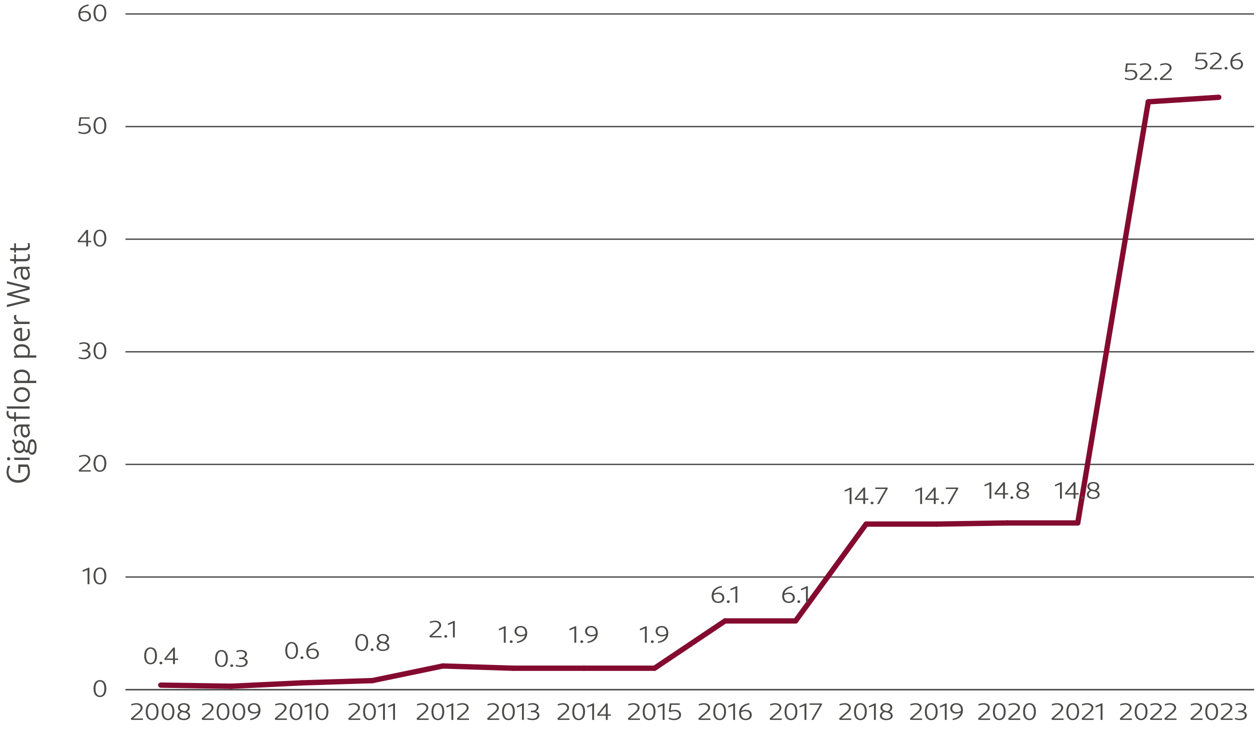
The good news is that similar improvements can be found in terms of energy efficiency. When considering the output, the differences are again an order of magnitude in favour of GPUs. Among the top 500 supercomputers, those based on accelerated instead of parallel computing have an average energy efficiency score of 22 gigaflops per watt consumed, against an average of four for those not using accelerated computing. Supercomputers based on accelerated computing typically outperform other supercomputers too, taking the top 48 places in the Green500 rankings.[17]

Once again, just as the increase in computing performance is a long-term evolution first achieved through shrinking transistors and then through new architecture, progress in energy efficiency is also a long-term trend. The energy efficiency of supercomputers has increased over 100-fold in the last 15 years, from 0.5 to 52.6 gigaflops per watt consumed, as the figure on page 7 shows.

[16] GPUs significantly outperform CPUs in deep learning tasks, offering higher throughput and more stable performance at a lower cost. In one detailed comparison conducted by Microsoft Azure, GPU clusters demonstrated superior throughput compared to CPU clusters for various deep learning models. Specifically, a single GPU cluster outperformed a 35-pod CPU cluster by at least 186%, and a 3-node GPU cluster outperformed the CPU cluster by 415% in terms of throughput. This improvement was even more pronounced for smaller networks, with a single-node GPU cluster performing up to 804% better than the CPU cluster for certain frameworks. See https://azure.microsoft.com/en-us/blog/gpus-vs-cpus-for-deployment-of-deep-learning-models/. Another practical example comparing TensorFlow 2 performance on CPUs versus GPUs revealed that training on a GPU was considerably faster, taking approximately 85% less time for training deep learning models. On a CPU, training epochs took around 480 seconds each, with high CPU utilisation, whereas on a GPU, the same training epochs took around 75 seconds each, significantly reducing the training time and using the GPU effectively while keeping CPU utilisation lower. See https://datamadness.github.io/TensorFlow2-CPU-vs-GPU.

[17] The Green500 metric uses the energy efficiency metric of gigaflops per watt, which is valuable insofar as it captures both input and performance. However, it has been often criticised as being excessively simplistic in measuring performance. See Michael Feldman, "TOP500 Meanderings: HPCG Gains Steam as Alternative Benchmark for Supercomputers," *TOP500.org*, 14 March 2018.

Figure 2: Energy Efficiency of the Fastest Supercomputer in Gigaflop per Watt



Source: Authors' elaboration of Top500.org data

Innovation in hardware and software is continuously and rapidly improving the energy efficiency of AI. Energy efficiency has become a unique selling point for chip producers, with newly released chips regularly slashing energy consumption metrics. The question is then: how do we create the optimal framework conditions to ensure that innovation in energy efficiency continues to evolve at least as fast as the demand for computing?[18]

> *'Energy efficiency has become a unique selling point for chip producers, with newly released chips regularly slashing energy consumption metrics.'*

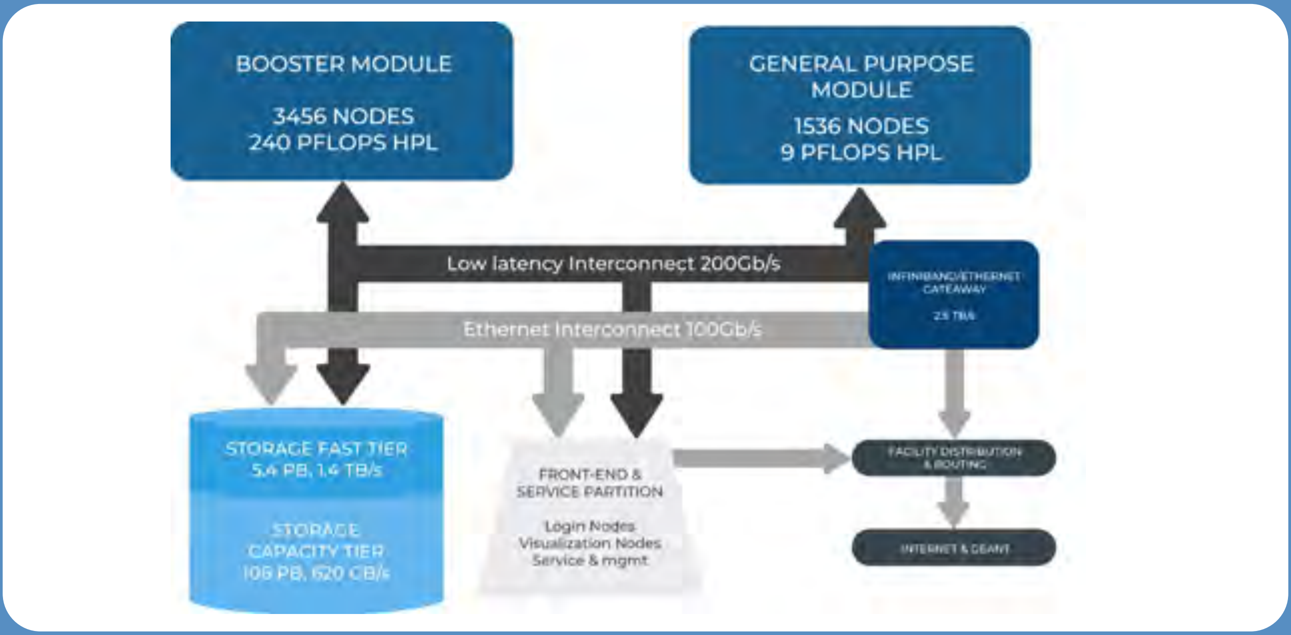
[18] For a comprehensive discussion on AI energy costs, see Daniel Castro, *Rethinking Concerns About AI's Energy Use* (Center for Data Innovation, 2024).

Accelerated Computing in Practice: the Case of Leonardo

Leonardo is one of the world's most powerful supercomputers, ranking sixth in the global top 500 ranking. The supercomputer is designed to serve as a research facility for a broad class of scientific investigations, offering a complete set of state-of-the-art hardware and software technologies. It is one of the few pre-exascale computers.

Supercomputers are not big computers with a powerful processor. To deliver pre-exascale performance, a complex, multi-layered architectural system with thousands of nodes and processors is needed – so-called accelerated computing. Accelerated computing involves breaking down a large computational problem into smaller, more manageable tasks that can be executed simultaneously across multiple processing units. In the case of Leonardo, this involves distributing computational workloads across many nodes divided into a booster partition with 3,456 nodes and 13,824 GPUs, and a data-centric partition with 1,536 nodes and 172,032 CPU cores. What makes Leonardo unique and innovative is not just the adoption of advanced chips, but an architecture that allows the best exploitation of the computing power of the different kinds of processors (GPUs and CPUs).

Architectural Overview of Leonardo



Source: Matteo Turisini, Giorgio Amati and Mirko Cestari, "LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI Applications," *arXiv*, 2023

The booster partition plays a critical role in enabling Leonardo to support parallel processing and provide advanced computational capabilities for a wide range of scientific research and engineering applications, with a focus on supporting both traditional high-performance computing (HPC) and emerging AI workloads. The data-centric partition, on the other hand, leverages the latest CPU technologies to fulfil a vast range of traditional HPC tasks such as large-scale data processing, e.g., for data mining, data warehousing and business intelligence.

But the novelty of Leonardo lies not only in its hardware. The software components are just as important. In addition to workload manager software that manages computational tasks, Leonardo provides a wide range of tools for developers. Not all algorithms can run on GPUs; they need to be developed specifically for this system and existing software needs to be ported to the GPU, representing a significant additional cost. The software ecosystem for Leonardo includes architecture-specific suites such as Intel OneAPI and NVIDIA HPC SDK, which provide optimised libraries for developing and running applications on the system.

This complex architecture is effective not only in terms of performance, but also in terms of energy efficiency. Leonardo has a PUE of 1.1. This means that the energy needed to cool down the system is only 10% of the power used to feed it. When considering energy efficiency expressed in gigaflops per watt using the Green500 methodology, Leonardo ranks 18th in the world with 32 gigaflops per watt.

As Leonardo is a relatively new system, there may not be many published results yet in terms of the output of applications run on the system. Some examples of the types of applications benefiting from Leonardo's high-performance computing capabilities include:

1. **Climate and weather modelling:** Leonardo is used to run large-scale climate and weather models, which are critical for understanding and predicting weather patterns, climate change and other environmental phenomena.

2. **Computational fluid dynamics (CFD):** Leonardo's capabilities in CFD simulations can be used to optimise the design of aircraft, automobiles and other complex systems. The output of these simulations can inform the design of more efficient and sustainable transportation systems.

3. **Life sciences:** Leonardo's capabilities in genomics, proteomics and drug discovery can be used to accelerate the development of new treatments and therapies for a wide range of diseases. The output of these applications could lead to new breakthroughs in medicine and healthcare.

4. **Materials science:** Leonardo's capabilities in simulating the behaviour of materials at the atomic and molecular level can be used to design new materials with specific properties for use in a variety of applications. The output of these simulations could lead to the development of new materials for use in energy storage, electronics and other fields.

Leonardo is an infrastructure open to researchers from academia, research institutes, public authorities and, most recently, industry. Researchers can apply for access to computing time through open calls for proposal. The access is mainly based on scientific merit and technical assessment for suitability to perform on Leonardo architectures to ensure the highest scientific reach of the selected project.

The Role of Metrics

In this context of accelerating trends and trade-offs, metrics are a crucial policy tool. They contribute to a better understanding of the problem, the mitigating measures and the opportunities, as well as to managing the trade-off between them. Alongside their importance for tracking progress, they can be used to create incentives and change behaviours. They should be selected not just because of their intrinsic validity, but also because of how they shape markets. So how can we identify the most appropriate metrics for energy efficiency in data centres?

When it comes to the energy efficiency of digital technologies, the most important policy measure is the EED recast, which includes a section dedicated to data centres. It provides data centres with strict reporting requirements on key performance indicators (KPIs). The follow-up EED delegated act identifies PUE as the key indicator for the energy efficiency of data centres and requires data centres to report their performance.[19]

'PUE does not consider output, so any improvement in the actual performance is simply not captured.'

Energy efficiency is typically calculated as a ratio of input to output, such as energy consumed for an outcome achieved. A real-life example is cars, where the KPI is litres of fuel consumed per 100 kilometres.

However, for data centres, it is often difficult to obtain a valid measurement of the output or service provided. Data centres perform very different tasks, such as computing, transmission and storage. The most used metric is PUE, which is simply the ratio of total energy consumption to the energy that is devoted purely to computing. A value of 1 means that all the energy is used for computing and not for accessory functions such as cooling. The best performer has a PUE of 1.01. The limitations are obvious: PUE does not consider output, so any improvement in the actual performance is simply not captured. For instance, using NVIDIA GPUs in weather forecasting can achieve 24 times more computing power than CPUs, with 127 gigawatt hours less energy consumption – looking only at PUE, that difference is not captured.[20]

PUE was a useful indicator when there were no major differences in output, as innovation was simply driven by ever more powerful CPUs, following Moore's law. Under these conditions, PUE was a good proxy for energy efficiency. However, the demand for greater computing power is no longer met by innovation at the CPU level, as described in the previous chapter. Innovation is now happening at the level of the computing architecture, for instance through accelerated computing.[21]

[19] See European Parliament, *Directive (EU) 2023/1791 of the European Parliament and of the Council of 13 September 2023 on Energy Efficiency and Amending Regulation (EU) 2023/955 (Recast)*, 2023 and European Commission, *Commission Delegated Regulation of 14.3.2024 on the First Phase of the Establishment of a Common Union Rating Scheme for Data Centres.* C(2024) 1639, 2024.

[20] For NVIDIA data, see https://resources.nvidia.com/en-us-energy-efficiency/faster-weather-prediction.

[21] The limitations of PUE have been known from its establishment in 2007 but are increasingly recognised as harming its validity today. See Andy Lawrence, "PUE: The Golden Metric Is Looking Rusty", *Uptime Institute Blog*, 02 November 2020.

How AI Helps the Energy Transition

Alongside all the concerns about its power consumption, AI is proving particularly effective in advancing the green energy transition. It does this in many ways, leveraging its ability to analyse large datasets, identify patterns and make predictions. As a recent article by the International Energy Agency puts it, "AI and energy are the new power couple."[22]

Energy efficiency: AI optimises energy use in various sectors, including buildings, manufacturing and transportation. By analysing data on energy consumption, AI can identify inefficiencies and suggest improvements, such as optimising heating, ventilation and air conditioning (HVAC) systems in buildings or enhancing logistics and routing in transportation to reduce fuel consumption.

While there is no clear-cut, comprehensive calculation many studies point to considerable reductions. For instance, a recent study estimates substantial savings: 35% energy cost savings in buildings; 25% energy savings by heating, ventilation and air conditioning equipment; 50% energy savings by artificial lighting systems; up to 70% savings in information transfer and communication power; a continuous output of 30% peak power from the renewable energy device to the microgrid; and 20% power demand reduction in the factory. In other words, while according to the International Energy Agency AI might increase energy consumption to reach 4% of total electricity consumption, the impact of AI on the "remaining 96%" of energy consumption is expected to clearly outstrip the direst AI energy consumption predictions. And even when it comes to data centres themselves, the adoption of AI in their energy management has already led to energy savings ranging from 9 to 40%.[23]

Renewable energy: AI improves the integration of renewable energy sources into the power grid by forecasting supply and demand as well as renewable energy availability (e.g., solar and wind power). This helps balance the grid and reduces reliance on fossil fuels by ensuring that excess energy generated from renewable sources can be stored and used when demand is high, thereby reducing waste and increasing the share of renewables in the energy mix.

But it goes beyond that: AI is crucial to new low-carbon sources of energy such as nuclear fusion, which has long been referred to as an "exascale" grand challenge. For instance, the Energy-oriented Centre of Excellence for Exascale HPC applications (EoCoE-III) project applies cutting-edge computational methods in its mission to foster the transition to decarbonised energy in Europe. Energy is a primary area of research for supercomputers.[24]

It's also worth pointing out that this is only part of the contribution that AI can make to fighting climate change. AI is instrumental to building better earth models, discovering new materials necessary for the green transition, or sequestering carbon from the atmosphere.[25]

[22] See International Energy Agency, "Why AI and Energy Are the New Power Couple," *IEA*, 02 November 2023.

[23] For these estimations, see Da-Sheng Lee, Yan-Tang Chen and Shih-Lung Chao, "Universal Workflow of Artificial Intelligence for Energy Saving," *Energy Reports*, 8, pp. 1602–33, 2022.

[24] See Rafael Mayo-Garcia and Edouard Audit, *Exascale, a Great Opportunity for Clean Energy Transition in Europe* (European Energy Research Alliance and Energy-oriented Center of Excellence, 2022).

[25] See Rolnick et al., Op. cit.

Policy Recommendations and Next Steps

Achieving the goal of net zero is not about reducing our standards of living, our hopes and ambitions, or our capability for digital innovation. As the European Commission points out, "the European Green Deal […] is a new growth strategy that aims to transform the EU into a fair and prosperous society, with a modern, resource-efficient and competitive economy where there are no net emissions of greenhouse gases in 2050 and where economic growth is decoupled from resource use."[26] This is about doing more with less. And the good news is that this is not only possible, it's already a reality: in OECD countries, there has been a visible decoupling of resource usage and output over the last 30 years.[27]

This is where the emergence of new technological solutions that reduce the carbon footprint of human activities (often called cleantech) plays a crucial role: reducing the trade-offs between economic growth and carbon emission through innovative products and services that can deliver more with less. The same pattern can be seen when it comes to energy-efficient data centres. Innovations such as accelerated computing enable radically more powerful performance that accelerates scientific discovery while reducing energy consumption. Policies need to create incentives that encourage this kind of innovation in Europe. This is precisely what the proposed "net zero industry act" aims to foster.[28]

To be clear, the intention here is not to argue that market innovation alone will solve the issue of energy efficiency. There is a real trade-off between net zero and the computing requirements of AI and there is no guarantee that market trends by themselves will solve it. Government intervention is probably needed, but it must tilt the playing field to encourage green innovation and create the right conditions for energy efficiency improvements.

'We need to ensure that the goals of different policy measures are aligned.'

We need to ensure that the goals of different policy measures are aligned. That Europe doesn't have one policy to promote the supply of energy-efficient, innovative products and services, and another that reduces demand for those products and services because of unsuitable metrics. Or one policy that promotes supercomputing and advanced data analytics to address climate change and increase energy efficiency and another that raises barriers to its use.

We also need to navigate the narrow path between degrowth and free market excesses, between techno-optimism and moral panic. There are benefits and there are costs. We need to qualify and quantify them, and act accordingly and proportionately.

This leads to four key conclusions:

1 AI energy consumption is certainly rising, but not exponentially. Innovations in hardware, software and architecture are bringing radical improvements not just in AI performance and computing power, but also in energy efficiency. Exponential growth in energy consumption is visible only in specific cases, such as the training of large language

[26] The quote is from the European Commission, *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: The European Green Deal*, COM(2019) 640, 12 November 2019.

[27] See Organisation for Economic Co-operation and Development, *Environment at a Glance Indicators* (Paris: OECD Publishing, 2023).

[28] European Commission, *Proposal for a Regulation of the European Parliament and of the Council on Establishing a Framework of Measures for Strengthening Europe's Net-Zero Technology Products Manufacturing Ecosystem (Net Zero Industry Act)*, C(2023) 161, 2023.

models over the last two years, but does not reflect the expected overall trends of AI energy consumption. The largest share of consumption comes from inference, which is not growing exponentially.

2 In terms of energy efficiency, AI generates benefits across all sectors that very likely outweigh the increase in consumption related to AI training and inference, even when using base case scenarios. For instance, the doubling of energy consumption by data centres by 2026 predicted by the International Energy Agency in the base scenario is much smaller than the expected energy savings from AI deployment across the sectors of the economy, based on the existing data points (see the box on AI and energy efficiency on page 11).

3 This doesn't mean that the market alone will solve these issues. Data is scarce, and there is a real and urgent need for greater transparency on energy consumption data related to AI and to data centres in general, as the EED recast aims to achieve, and to tilt the market further towards energy efficiency. But the focus on a single, limited metric is counterproductive. While PUE remains widely used and consistently available, it is no longer appropriate for the current technological reality as it ignores the output performance and only focuses on input. Put bluntly, this harms, rather than helps, the move to sustainable computing, reducing incentives for green innovation in computing power and ultimately making data centres become less energy efficient and less sustainable.

4 A multi-parameter framework based on multiple indicators and accounting for computing performance would be more effective. The data gathered through the EED reporting requirements should be closely analysed by the research community, industry and regulators to identify the right metrics for computing performance and energy efficiency. A permanent, structured effort should be put in place to develop and test new metrics, data points and compound indicators. This effort should be international in nature, within fora such as the G7 and the OECD. The current metrics defined in the delegated act should be treated as a starting point and a work in progress.[29]

If metrics such as PUE become the cornerstone for measuring energy efficiency, or even worse, for monitoring and enforcing it, it will ultimately lead to the adoption of sub-optimal solutions in data centres, more energy consumption, and will directly harm the already struggling European innovation ecosystem (see box on Europe's performance in science on page 4) by slowing down the adoption of AI across scientific disciplines and academic sectors. It will make both AI training and inference less efficient, more costly and more energy-consuming.

Humanity is facing dramatic challenges but continues to show astonishing ingenuity. This is not a time for quick fixes and ideological confrontation between free market and regulation, but for continuous tinkering and steering of market forces in line with the policy goals, and most importantly, with the survival and progress of humankind.

> *'Humanity is facing dramatic challenges but continues to show astonishing ingenuity.'*

[29] A recent report by the Joint Research Centre of the European Commission identifies nine such indicators. See Paulo Bertold, "Assessment Framework for Data Centres in the Context of Activity 8.1 in the Taxonomy Climate Delegated Act" (Ispra: European Commission, 2023). In addition, other performance oriented indicators include total processing performance used in the context of U.S. export controls and the mentioned Green500 energy efficiency indicator (see page 6).

References and Further Reading

McAfee, Andrew. *More from Less* (New York: Simon & Schuster Paperbacks, 2019)

Gilles, Babinet, *Green IA: L'intelligence Artificielle au Service du Climat* (Paris: Odile Jacob, 2024)

Castro, Daniel. *Rethinking Concerns About AI's Energy Use* (Center for Data Innovation, 2024)

Desislavov, Radosvet, Fernando Martínez-Plumed and José Hernández-Orallo. "Trends in AI Inference Energy Consumption: Beyond the Performance-vs-Parameter Laws of Deep Learning," *Sustainable Computing: Informatics and Systems*, 38, 2023

European Commission. *Proposal for a Regulation of the European Parliament and the Council on Establishing a Framework of Measures for Strengthening Europe's Net-Zero Technology Products Manufacturing Ecosystem* (Net Zero Industry Act), C(2023) 161, 2023

———. *Commission Delegated Regulation of 14.3.2024 on the First Phase of the Establishment of a Common Union Rating Scheme for Data Centres*, C(2024) 1639, 2024

European Commission, Directorate-General for Research and Innovation. *AI in Science: Harnessing the Power of AI to Accelerate Discovery and Foster Innovation: Policy Brief* (Luxembourg: Publications Office of the European Union, 2023)

———. *Science, Research and Innovation Performance of the EU 2022: Building a Sustainable Future in Uncertain Times* (Luxembourg: Publications Office of the European Union, 2022)

———. *Trends in the Use of AI in Science: A Bibliometric Analysis*. (Luxembourg: Publications Office of the European Union, 2023)

European Parliament and the Council. *Directive (EU) 2023/1791 of the European Parliament and of the Council of 13 September 2023 on Energy Efficiency and Amending Regulation (EU) 2023/955 (Recast)*, 2023

Hajkowicz, Stefan, Conrad Sanderson, Sarvnaz Karimi, Alexandra Bratanova and Claire Naughtin. "Artificial Intelligence Adoption in the Physical Sciences, Natural Sciences, Life Sciences, Social Sciences and the Arts and Humanities: A Bibliometric Analysis of Research Publications From 1960-2021," *Technology in Society*, 74, 2023

Hey, Tony, Stewart Tansley and Kirstin Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Redmond, Washington: Microsoft Research, 2009)

International Energy Agency, *Electricity 2024* (Paris: IEA, 2024)

———. "Why AI and Energy Are the New Power Couple," *IEA*, 2023

Lee, Da-Sheng, Yan-Tang Chen and Shih-Lung Chao. "Universal Workflow of Artificial Intelligence for Energy Saving," *Energy Reports*, 8, pp. 1602–33, 2022

Organisation for Economic Co-operation and Development. *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research* (Paris: OECD Publishing, 2023)

———, *Environment at a Glance Indicators* (Paris: OECD Publishing, 2023)

Pilat, Dirk. “Driving Innovation to Curb Climate Change: Recommendations for COP 27,” *Lisbon Council Interactive Policy Brief*, 30, 2022

Rolnick, David, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran et al. “Tackling Climate Change with Machine Learning,” *arXiv*, 2019

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn and Pablo Villalobos. “Compute Trends Across Three Eras of Machine Learning,” in *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022

Turisini, Matteo, Giorgio Amati and Mirko Cestari. “LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI Applications,” *arXiv*, 2023

Waldrop, M. Mitchell. “The Chips Are Down for Moore’s Law,” *Nature News*, 530, 11 February 2016

theLisboncouncil
Lisbon Council Research

About Lisbon Council Research

Lisbon Council Research is the scientific arm of the Lisbon Council for Economic Competitiveness and Social Renewal asbl, a Brussels-based non-profit association committed to making Europe “the most competitive and dynamic knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion,” as European leaders vowed to do in Lisbon, Portugal at a landmark 2000 summit. Lisbon Council Research conducts on-going research into Europe’s economic and social challenges.

Lead Authors:
David Osimo and Senan Khawaja

Published under the editorial responsibility of Lisbon Council Research sprl

Lisbon Council Research sprl
Résidence Palace
Rue de la Loi 155
1040 Brussels, Belgium