Can AI solve its sustainability problem?

INCREASING WORKLOADS AND THE GROWTH OF HPC AND AI SYSTEMS ARE CREATING A CARBON CRISIS FOR RESEARCH, WRITES EUGENIA BAHIT

With the proliferation of artificial intelligence (AI) and machine learning (ML), the high-performance computing industry's increasing workloads are contributing to environmental damage.

It may be confusing that cutting-edge technology such as AI, which helps humanity in so many fields, can have a negative impact and put an entire industry in check. Still, the increase in AI workloads is becoming a problem that needs urgent attention.

What causes increasing AI/ML workloads?

Artificial intelligence is advancing fast, and workloads are becoming bigger due to data volume and algorithm complexity.

Developing an intelligent machine requires the ability to make inferences that involve applying logical rules to deduce a possible conclusion. Current ML models do this by storing and classifying astronomical volumes of data, which is one of the factors responsible for large workloads.

The complexity of deep learning models – used to understand the data and apply the logical rules – typically requires recursion to process millions of parameters, demanding lots of resources. Therefore, the more complex the algorithms, the greater the workload.

Optimising the ML workflows to reduce workloads is not easy. According to Dion Harris[1], Head of Product Marketing for Accelerated Computing in HPC, AI, and Quantum Computing at Nvidia, some HPC users may not have the specific skills to optimise AI/ML workflows, and AI/ML experts may not have the experience to optimise them on HPC systems. Harris sees a gap between researchers with applied AI backgrounds and those with scientific computing knowledge.

"The rate at which AI is evolving requires deep expertise in data extraction, data preparation, AI training and inference, across a growing range of AI foundational models and techniques," he says. As a result, gaining experience lags behind the pace of AI evolution. Additionally, Ying-Chih Yang[2], CTO at SiPearl – a European company working to develop a low-power exascale microprocessor – thinks the workflow optimisation is also problematic because the addition of heterogeneous computing elements, such as GPUs or AI accelerators, improves performance on workloads while increasing software complexity.

Why are increasing workloads so critical?

From a business standpoint, the shortest answer is "money", as increased workloads lead to longer processing times, higher power consumption and, consequently, higher costs. However, a short answer is insufficient to become aware of the real problem.

The increase in workloads has a snowball effect from more than one perspective. When workloads increase processing time, the hardware has to work harder, causing a significant temperature rise, which demands more cooling. This not only increases costs but also leads to a larger carbon footprint – a direct result of greenhouse gas emissions. Therefore, the increasing AI workloads should not only be considered from a technical perspective. They ought to be managed to curb the environmental damage and ecological deterioration that is impairing the globe and primarily impacting the most underprivileged populations[3].

metamorworks/Shutterstock.com

"The carbon footprint needs to be viewed systematically, including development and transport of materials, manufacturing and transport of components, system design, energy sources, and reuse"

Yang emphasises the importance of viewing greenhouse gas emissions as a whole by pointing out that "the carbon footprint needs to be viewed systematically, including development and transport of materials, manufacturing and transport of components, system design, energy sources, and reuse".

Reducing the carbon footprint of AI workloads is a shared responsibility encompassing low-power hardware production, workflow optimisation, renewable energy sources, and a holistic approach covering the three greenhouse gas emissions scopes: direct, indirect, and indirect third-party emissions.

How does AI affect the environment?

The environmental impact of AI has been studied in recent years.

In 2019, Alexandre Lacoste and a group of Canadian scientists presented a Machine Learning Emission Calculator for cloud-based platforms[3]. They found that choosing a provider based on location offered a significant difference. The same model tested on three providers resulted in a carbon footprint 50 times larger in South Africa and 45 times bigger in China than in Canada. They also mentioned that GPUs were 10 times more effective than CPUs and 4 to 8 times less effective than TPUs. In an interview with journalist Payal Dhar, Lacoste expressed that using renewable energy for training ML models was the most significant change possible[4].

After the interview with Lacoste, Dhar wrote an article summarising studies since 2018. The report, titled "The carbon impact of artificial intelligence" and published in *Nature Machine Intelligence*[4], highlighted a finding showing that training a single ML model emitted the equivalent CO_2 of 125 round-trip flights between New York and Beijing.

That same year, a study by Simon Portegies Zwart[5] found that Python, the most popular programming language in astrophysics, produces the most CO_2. The paper concludes that using GPUs in computational astrophysics can significantly reduce the carbon footprint and emphasises using more efficient languages, such as Julia or Rust.

In 2021, Loïc Lannelongue wrote the article "Carbon footprint, the (not so) hidden cost of high-performance computing,"[6] which exposed numerous findings on the environmental impact of AI. Lannelongue, a principal contributor to several studies on this matter, suggested that the lack of data could be the primary reason this issue was not considered important within the industry.

Towards the end of 2021, an editorial note from Tan Yigitcanlar, published in the MDPI *Sustainability* journal, summarised the Green AI approach. He explains how it goes beyond the transition to renewable →

"The main differentiator of LUMI is the use of 100% hydroelectric power"

→ energy sources and its importance as public policy and its reliance on effective government regulations for success[7].

In 2022, Lannelongue participated in a study analysing "The Carbon Footprint of Bioinformatics"[8], concluding that: (a) cloud-based solutions reduce overhead power consumption; (b) the relationship between carbon footprint and the number of cores is not linear; (c) power consumption depends on the allocated memory; and (d) in some cases, the use of GPUs reduces the execution time, but multiplies the carbon footprint.

Something is wrong

Studies seem to agree that cloud-based HPC is the most energy-efficient. Reading the sustainability reports of cloud-based HPC owners, everything seems to be on the right track. Most aim to reach net zero emissions by 2030 and similar goals.

However, although the HPC era began in the 1960s[14], it was not until the end of the 1990s when the first terascale supercomputer appeared and towards the end of the 2000s that the first petascale supercomputer saw the light.

Long before the current exascale era, the world already knew about the severe consequences of fossil fuel burning and the necessity of renewable energy sources. The United Nations Framework Convention on Climate Change was adopted and signed by 154 countries in 1992. The damages of greenhouse gas emissions were already in the public domain when the IBM Roadrunner astonished the world.

However, if the world already had this knowledge at the end of the 1990s, why do the sustainability plans of cloud-based AI companies set 2030 as the goal for reaching carbon neutrality or net zero? A possible answer could be found in an article[9] by Jay Boisseau – HPC & AI Technology Strategist at Dell – where he gives three reasons for energy reduction: a reliance on costly fossil fuels, the environmental impact, and government pressure.

How can AI negatively impact humanity?

In January 2023, *The Lancet* published an article titled "Climate Change, Health, and Discrimination: action towards racial justice" [10], which mentions that the Global North is responsible for 92% of historical CO_2 emissions despite representing only 14% of the global population. It makes sense, since the Global North was the first to industrialise.

However, the priorities and reasons driving the industrialised world seem more aligned with individualistic and economic interests rather than the common good. Therefore, it is worth wondering what the priorities and reasons would be for reducing the carbon footprint and power consumption if the derived costs of fossil fuel burning were not so high. Fortunately, they are. So those who suffer the most from the product of global warming – namely, the Global South, which includes low-income countries and black, brown, and indigenous populations – could potentially avoid some of the terrible effects of an economy which is intrinsically supremacist[12] and unable to act motivated solely for the common good. Unfortunately, the Global South is already experiencing the impacts of global warming[10] due to the apathy of a world economy that has been inherently racist[11] [12] for decades.

What is the HPC industry doing to face the problem?

Each organisation works for sustainability from different perspectives.

At Nvidia, Harris says it is addressing the new AI era challenges on multiple levels. They focus on helping researchers by providing training, tools, and libraries – such as TensorRT – to optimise AI/ML workflows and pre-trained models and GPU-optimised frameworks. However, they are also working on energy-efficient hardware. Harris says that towards the end of 2023, they will start distributing the Grace CPU and Grace-Hopper CPU-GPU, enabling existing data centres to maintain

Doidam 10/Shutterstock.com

"SiPearl's product roadmap targets reducing the carbon footprint for workloads"

performance while doubling energy efficiency. They also presented cuLitho, a library for NVIDIA Hopper GPU systems that optimises the inverse lithography technology processes – used to develop nanoscale semiconductors – reducing power consumption to up to six times less than current CPU systems.

Intel, AMD, and ARM, among others, are working on more efficient hardware, while organisations such as SiPearl and LUMI are going even further.

LUMI: The cleanest, most effective future experience at the present

"HPC is energy-intensive and, therefore, has a high CO_2 footprint if operated in the wrong manner," says Dr Pekka Manninen[15], Director of Science and Technology at CSC, the Finnish centre where LUMI (one of the pan-European pre-exascale supercomputers) is located.

LUMI, a high-performance supercomputer with a sustained computing power of 375 petaflops, is considered one of the most advanced centres in terms of sustainability. Its entire power consumption comes from renewable energy sources, and the waste heat generated by the centre is used by the Finnish energy company Loiste, contributing nearly 20% to the district heating system. "By operating the HPC systems with CO_2-free energy, the game changes. The main differentiator of LUMI is the use of 100% hydroelectric power," says Manninen.

Based on an AMD Instinct GPU architecture, LUMI is the fastest HPC in Europe according to the TOP500 list and the third worldwide[13]. "The tenders were evaluated in terms of their technical value, performance, supply chain considerations, and EU-added value. We valued price-performance and energy efficiency a lot, which gave the edge to AMD," says Manninen.

When it comes to sustainability, he is direct: "No HPC facility should consider any other choice than 100% CO_2-free electricity. That is the leading factor in the sustainability of an HPC installation. The rest, like providing users with information about their jobs, tinkering with energy-efficient algorithms, scheduling, etc., we have seen over the years, is just 'peanuts', and the whole equation of end-to-end sustainability is dominated by the choice of power sources."

Like Yang, Manninen thinks sustainability should be seen as a whole, building an end-to-end systematic evaluation methodology that considers the system supply chain to its decommissioning.

SiPearl: the energy-efficiency microprocessor promise

Founded in 2019 and headquartered in France, SiPearl is working to produce a low-power, high-performance microprocessor. The first generation, Rhea, aims to be the first worldwide HPC-dedicated processor designed to support any third-party accelerator and deliver energy-efficient computing at scale, explains Yang.

"SiPearl's product roadmap targets reducing the carbon footprint for workloads. When available, the processor should significantly reduce the carbon footprint by up to 30% and 50%," he says.

Rhea is based on a low-power optimised Arm Neoverse architecture, which, according to Yang, "offers an ideal choice for performance per watt and energy efficiency for scalar and vector processing."

Yang further points out that SiPearl is working with Arm to integrate the network on the chip to interconnect the compute elements. He highlights the importance of the comprehensive Arm plan to produce future processor cores for seamless software migration, facilitating consistent improvements over time.

What's next?

No one can doubt the advantages of AI when every day, the media leave us astonished with technological innovations. But the world lives two parallel realities. One is technologically perfect, filled with unimaginable advancements, where large companies invest billions in increasingly sophisticated technology. And then, there is the other reality. One in which the waste ends up, and the nations that were once colonies of the former still suffer the consequences of inequality. It is a reality where energy crises force entire populations to live in inhumane conditions.

Carbon neutrality and net zero are preferable to any greenhouse gas emissions. However, they are not the ultimate solution. The workloads required to train deep learning models are growing daily, increasing power consumption. The only viable option is a joint effort to transition to renewable energy sources, build a systematic methodology for comprehensively evaluating sustainability, develop energy-efficient hardware, and optimise AI/ML workflows without neglecting the fundamental role of states, where nations commit to creating clear public policies that include minorities and underprivileged populations in decision-making.

References:

[1] www.linkedin.com/in/dionharris
[2] https://sipearl.com/en/ying-chih-yang
[3] https://arxiv.org/abs/1910.09700
[4] www.nature.com/articles/s42256-020-0219-9
[5] www.nature.com/articles/s41550-020-1208-y
[6] https://academic.oup.com/itnow/article-abstract/63/4/12/6503631
[7] www.mdpi.com/2071-1050/13/24/13508
[8] https://academic.oup.com/mbe/article/39/3/msac034/6526403
[9] www.cio.com/article/418812/whats-new-and-whats-next-in-2023-for-hpc.html
[10] https://pubmed.ncbi.nlm.nih.gov/36343651/
[11] https://bbc.in/3NRG5QA
[12] https://archive.org/details/statement-on-systemic-and-pervasive-racism-within-the-environmental-field
[13] www.top500.org/system/180048/
[14] https://digital-library.theiet.org/content/journals/10.1049/cce_19970107
[15] www.linkedin.com/in/pekkamanninen