Atlantic Vantage Point

avpcap.com

The Environmental Impact of AI: How Can We Balance Innovation and Sustainability?

Camille Périssère

Bringing perspectives and investing in tech
from both sides of the Atlantic

Introduction

In January 2025, newly elected US President Donald Trump unveiled "Stargate," a $500 billion initiative poised to become the most extensive AI infrastructure project in history. Backed by industry giants OpenAI, SoftBank, and Oracle, Stargate aims to solidify America's leadership in artificial intelligence by constructing vast data centers and accompanying energy systems necessary to power next-generation AI technologies.

While this ambitious endeavour underscores the strategic (and political) importance of AI, it also raises significant environmental concerns. Data centers are notorious for their substantial energy consumption, land use, and water requirements. Moreover, the rapid advancement of AI hardware accelerates electronic waste and intensifies the demand for rare raw materials. Notably, GenAI models are particularly resource intensive.

According to estimates, the initial training phase of GPT-4 would have emitted between 1,200 and 15,000 tons of CO_2, depending on whether the model was trained in a data center in Canada East (the lowest-carbon Azure region[1] thanks to its predominantly hydroelectric grid) or on grid electricity in California, where natural gas still makes up a significant portion of the energy mix. Under the worst-case assumption, this would be comparable to the yearly energy consumption of 2,000+ US homes[2]. Moreover, the electricity consumption from training GPT-4 may be approximately 40 to 48 times higher than that required to train GPT-3, even though GPT-4's total parameter count is believed to be only about 10 times greater. If these numbers seem staggering, they pale in comparison to the environmental impact of other stages in an AI model's life cycle - particularly the deployment (inference) phase.

Despite these challenges, GenAI also offers promising solutions to combat climate change. In agriculture, GenAI enhances precision farming by analyzing extensive datasets to optimize resource utilization. In the energy sector, companies like Google have employed AI to optimize data center operations, achieving significant reductions in energy consumption. DeepMind's AI system, for instance, reduced cooling energy usage in Google's data centers by up to 40%, translating to a 15% decrease in overall energy usage[3]. Furthermore, GenAI contributes to more accurate climate modeling, aiding in better understanding and adaptation to climate change. Notable companies in this space include Mitiga Solutions, Eoliann, or Jua AI - a Swiss startup that has developed a proprietary large-scale model, known as a "Large Physics Model," trained on petabytes of raw data to simulate atmospheric and environmental dynamics. These examples only scratch the surface of what GenAI can contribute to the fight against climate change - a topic substantial enough to deserve a white paper of its own.

1. Assuming GPT-4 was trained in an Azure data center because of OpenAI's partnership with Microsoft (https://medium.com/data-science/the-carbon-footprint-of-gpt-4-d6c676eb21ae)
2. https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator#results
3. https://quantumzeitgeist.com/deepmind-ai-cuts-google-data-center-cooling-bill-by-40-revolutionizing-energy-efficiency/#google_vignette

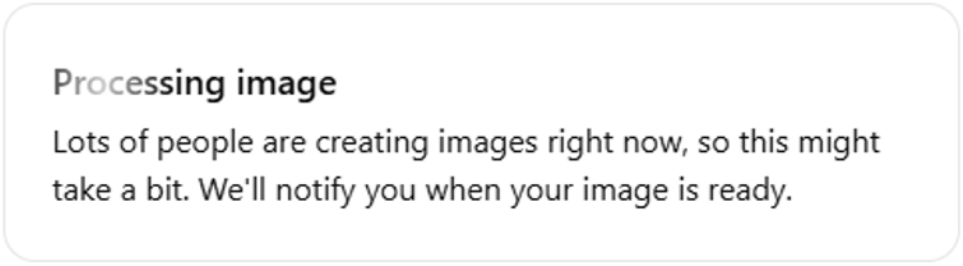
Nevertheless, accurately measuring the environmental impact of AI remains a complex task. As Golestan Sally Radwan, Chief Digital Officer of the United Nations Environment Programme, aptly stated: *"We need to make sure the net effect of AI on the planet is positive before we deploy the technology at scale."* This underscores the imperative to scrutinize not only the environmental footprint of AI but also to explore and implement solutions than can support a greener computing future.

Generative AI's Environmental Cost

While generative AI offers promising advancements, it also comes with a significant environmental cost. The development and operation of large-scale AI models - particularly the large language models behind GenAI - require immense energy for the data processing, model training and inference (which is the process of using a trained model to make predictions on new data).

Indeed, these models, often consisting of billions of parameters, like OpenAI's GPT-4, demand extraordinary computational resources. The electricity needed for such training processes leads to considerable carbon dioxide emissions and adds significant pressure to the power grid.

And the energy consumption doesn't end once a model is trained. Running these models in everyday applications, scaling them to millions of users, and continually refining their performance require ongoing and substantial energy input. In fact, a single query made through ChatGPT can use up to ten times more electricity than a standard Google search. Even generating a seemingly simple meme or "starter pack" image via ChatGPT could consume between two to five litres of water - you might think twice before cracking a joke to your colleague about that far-fetched career move.



Let's dive a bit deeper to understand why these processes are problematic for the environment.

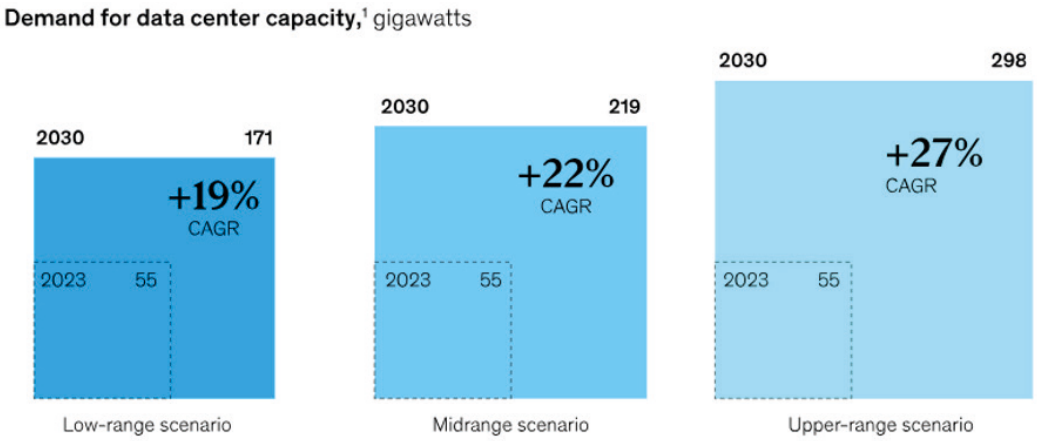
The Power-Hungry Nature of Generative AI

This rapid acceleration of generative AI adoption has obviously triggered a surge in demand for power-hungry data centers. And even though it's called "cloud computing," data centers very much exist in the physical world. They require vast amounts of electricity as well as water, space, and scarce resources, with both direct and indirect effects on biodiversity.

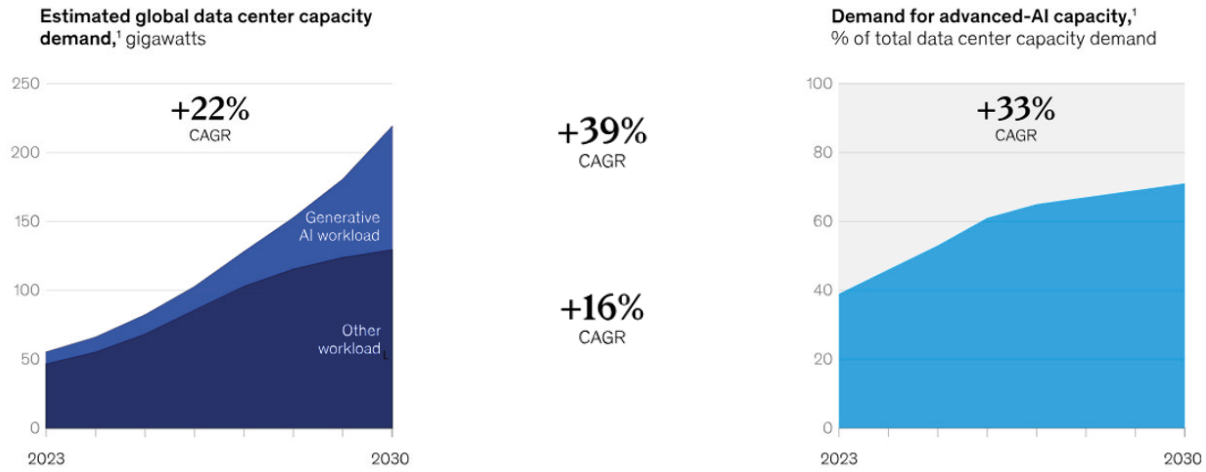
A data center is a temperature-controlled building that houses computing infrastructure, such as servers, data storage drives, and network equipment. Amazon, a leading cloud provider, operates over 100 data centers globally, each typically containing around 50,000 servers used to power its cloud computing services.

Despite some cloud providers' claims of utilizing renewable energy (such as Amazon[4]), many of these centres still largely depend on fossil fuels, compounding the technology's carbon footprint.

Global demand for data centyer capacity could more than triple by 2030.



AI (and especially GenAI) is the key driver of growth in demand for data center capacity



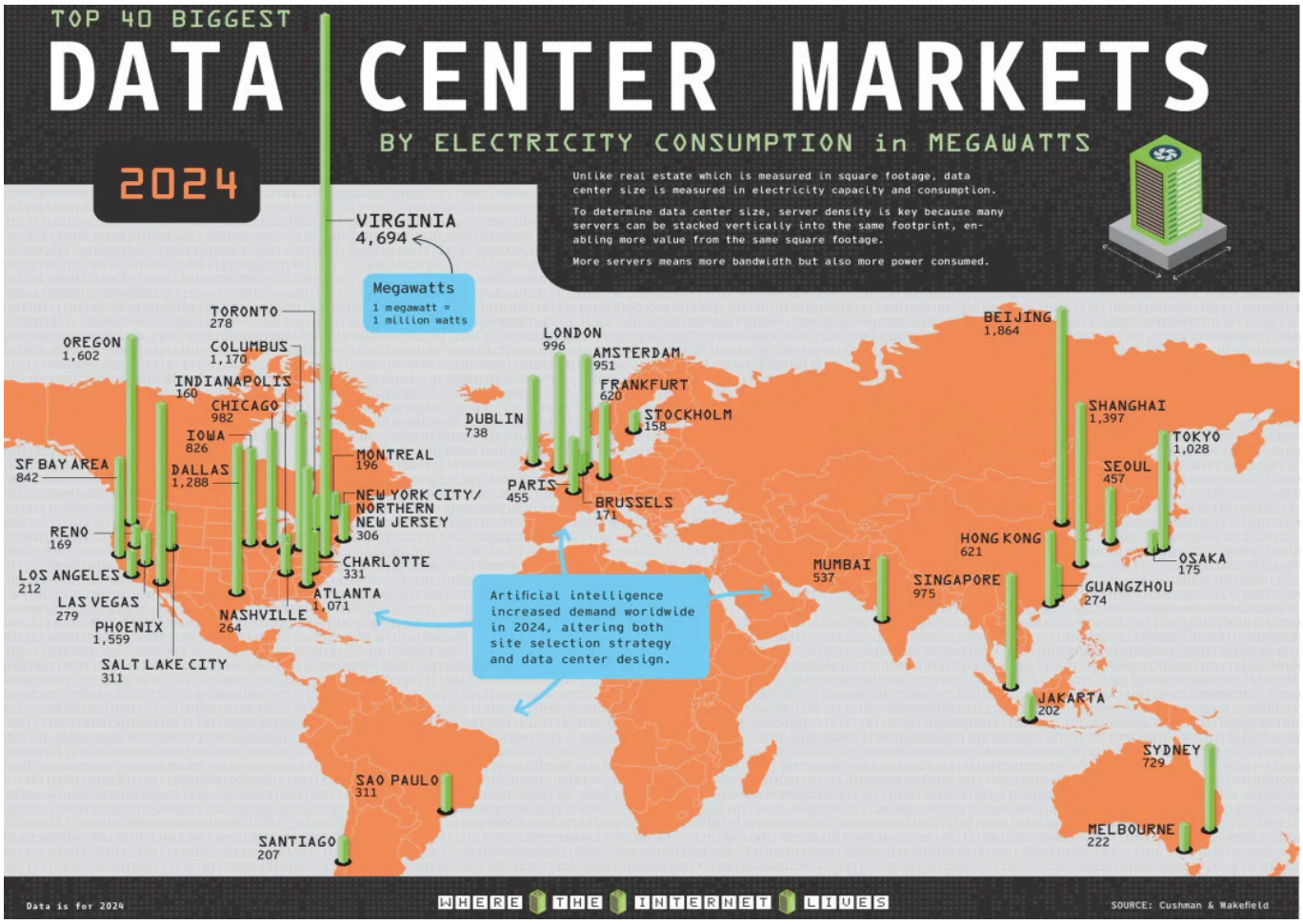
Source: McKinsey Data Center Demand model

4. https://www.datacenters.com/news/amazon-s-100-billion-data-center-expansion

Additionally, in some countries, concern about the pressure data centers exert on electricity grids as well as the impact on national climate targets have brought a complete halt to the building of new ones. Ireland, for instance, has stopped issuing new grid connections to data centers in the Dublin area until 2028 due to concerns about grid pressure and climate goals. Ireland's transmission system operator estimates that data centers will account for 28% of the country's power use by 2031.

Traditionally, data centers have been built near population centers to reduce latency, often leveraging colocation strategies - where multiple providers form a cluster to enhance resilience and prevent outages. A notable example is Telehouse Paris Voltaire, one of the largest colocation facilities in Paris, hosting over 100 carriers and serving as a key hub for 80% of France's live internet traffic[5]. However, when training AI models, low latency and network redundancy are less critical than during inference, when the model is actively serving users. As a result, AI training data centers are increasingly being located in more remote regions where power grids are less strained. Still, limited transmission infrastructure in these areas poses a growing risk of supply constraints as demand continues to increase.

Biggest Data Center Markets by Megawatts



Source: Voronoi by Visual Capitalist[6]

5. https://www.telehouse.fr/connectivite-data-center-telehouse/france/paris/telehouse-paris-voltaire/
6. https://www.voronoiapp.com/markets/Biggest-Data-Center-Markets-by-Megawatts-3006

The US remains the biggest market by far with 5,426 data centers as of April 2025[7], hosting the biggest data producing and consuming businesses. In terms of energy consumption, the Virginia market nearly doubled in one year. A titan project like Stargate raises legitimate concerns about the additional pressure made on electric grid, as well as the significant environmental impact.

GenAI's Hidden Water Consumption

Water usage is another concern. AI infrastructure consumes vast amounts of water to cool down data centres, with global AI-related operations projected to use six times more water than Denmark. This water demand is particularly alarming in regions already facing water scarcity, where the diversion of resources for cooling purposes could exacerbate existing shortages and strain local ecosystems. Moreover, the environmental impact of wastewater produced by these cooling processes, which often contains chemicals or pollutants, adds another layer of concern.

Rare Earth Reliance: GenAI's Material Footprint and Geopolitical Risks

Beyond energy use, generative AI accelerates the need for high-performance hardware such as GPUs and other specialized processors. These chips rely on a variety of critical raw materials, including rare earth elements (REE) like neodymium, dysprosium, terbium, and others. These elements are essential for manufacturing permanent magnets, high-efficiency cooling systems, and precision components used in advanced computing and data center infrastructure.

However, the extraction and processing of rare earth elements come with significant environmental costs. Mining operations are often concentrated in a few regions, particularly in China, which controls a large portion of global REE supply. The extraction process typically involves open-pit mining, chemical separation, and radioactive waste generation, leading to soil and water contamination, deforestation, and greenhouse gas emissions. Additionally, rare earth mining has been associated with poor labor conditions and geopolitical risks, raising concerns about ethical sourcing and supply chain stability.

China produces 60% of all REE used as components in high technology devices

China's Share	Extraction	Processing
Copper	8%	40%
Nickel	5%	35%
Cobalt	1.5%	65%
Rare Earths	60%	87%
Lithium	13%	58%

But even more than extraction, China is the dominant economy when it comes to processing operations.

Source: Visual Capitalist[8]

7. https://www.statista.com/statistics/1228433/data-centers-worldwide-by-country/
8. https://elements.visualcapitalist.com/visualizing-chinas-dominance-in-clean-energy-metals/?utm_source=chatgpt.com

As generative AI becomes more pervasive, the demand for REEs is expected to increase sharply, placing additional strain on already sensitive ecosystems and supply chains. This makes the environmental footprint of AI hardware not just an energy issue, but also a materials issue — not to mention the geopolitical risks posed by our dependence on China.

The E-Waste Ripple Effect of GenAI

The relentless pace of AI innovation also contributes to mounting electronic waste. As newer, more powerful devices replace outdated hardware, this cycle of continual upgrades perpetuates a linear, unsustainable economic model, further straining ecological systems.

"*When we think about the environmental impact of generative AI, it is not just the electricity you consume when you plug the computer in. There are much broader consequences that go out to a system level and persist based on actions that we take,*" says Elsa A. Olivetti, professor in the Department of Materials Science and Engineering and the lead of the Decarbonization Mission of MIT's new Climate Project.

Finding the Right Balance: Shaping the Future with Generative AI

In this chapter, we explore solutions and technologies that support a greener computing future. This list will continue to evolve as new techniques emerge daily to enhance AI energy efficiency and reduce the carbon footprint of computing.

Powering Data Centers with Renewable Energy and Carbon Removal Solutions

Shifting AI processing to data centres powered by renewable energy can be considered as the most critical step for a project like Stargate. In fact, to meet substantial energy requirements, several dedicated energy solutions are planned[9] such as:

- Solar energy and battery storage: SB Energy (SoftBank) will develop solar energy systems with battery storage (own off-grid power using behind-the-meter solutions
- Small Modular Nuclear Reactors (SMRs): Stargate plans to deploy SMRs used for establishing a stable baseload power supply and minimizing transmission losses[10]
- Carbon Capture: The current energy infrastructure will continue to rely on natural gas as a primary component. However, the implementation of Carbon Capture, Utilization, and Storage (CCUS) systems enables direct carbon dioxide emission capture at the source while supporting global greenhouse gas emission reduction initiatives

9. https://www.datacenterfrontier.com/machine-learning/article/55272808/from-billions-to-trillions-data-centers-new-scale-of-investment
10. https://www.certrec.com/blog/energy-demands-for-openai-stargate-project/?utm_source=chatgpt.com

In France, President Emmanuel Macron has announced a €109 billion investment in AI infrastructure to "Make France an AI powerhouse" - around five times less than Stargate, which is proportional to the population. Through this major investment plan, France aims to host key global infrastructure and help reduce the sector's overall environmental footprint, thanks to its low-carbon energy mix[11]. In fact, France benefits from several key assets that make its territory attractive for the development of dedicated AI infrastructure:

- A plentiful supply of decarbonized energy - 95% of the electricity produced is already low-carbon, primarily from nuclear power - combined with competitive and stable electricity prices;
- A strategic geographical position, with two-thirds of the EU's subsea cables landing in France;
- Land well-suited for data center projects, including 35 pre-identified sites across mainland France.

Innovating towards More Sustainable Semiconductor Materials

While silicon has been the cornerstone of semiconductor technology for decades, it is now approaching its physical limitations. It falls short in delivering the power efficiency, thermal management, switching speed, and voltage tolerance that next-generation data centres demand.

That's why the industry is exploring alternative materials like Gallium Nitride (GaN) and Silicon Carbide (SiC), which offer better performance in this area. GaN devices can provide 3x the power or charging speed of silicon[12], achieve as much as 40 % energy savings, and occupy roughly half the size and weight of their silicon counterparts. SiC, with its superior thermal conductivity and optimized performance at lower switching frequencies, is better suited for the high-power, high-voltage applications found at data-center perimeters. As a result, a hybrid architecture—SiC handling incoming medium-voltage conversion and GaN managing rack- and board-level DC-DC stages—is rapidly becoming the industry standard.

Beyond performance, this shift offers significant environmental benefits: GaN manufacturing emits 10x less CO_2 than silicon production and replacing legacy silicon with high-efficiency GaN in data-center power systems could slash electricity consumption by up to 10 %, yielding a substantial reduction in global carbon emissions.

11. https://www.elysee.fr/admin/upload/default/0001/17/d9c1462e7337d353f918aac7d654b896b77c5349.pdf
12. https://navitassemi.com/sustainability-and-gan/

Reducing Data Footprint

Effective data management is all about strategically utilizing storage space. By reducing your data footprint, you can minimize the number of storage systems required - reducing both cost and environmental impact.

There are several ways to reduce your data footprint[13]:

- Data deduplication removes redundant copies of files, freeing up storage space.
- Data compression reduces the size of stored data by identifying patterns and replacing them with shorter codes.
- Data tiering relocates infrequently accessed data from high-performance, expensive storage to more cost-effective alternatives.
- Archiving moves rarely used data offline, keeping it accessible when needed without occupying primary storage.

Deduplication and compression significantly shrink your storage footprint while maintaining data integrity. By compressing data, you can store more within the same infrastructure, delaying the need for additional storage expansion. Additionally, compressing data before transmission optimizes network bandwidth, leading to faster transfers, reduced latency, and improved overall efficiency. Ultimately, these optimizations lower storage costs, enhance performance, and contribute to a more sustainable and cost-effective data center.

Data infrastructure technologies remain one our main investment focus at AVP. Notable companies is this space include established players such as Cribl or Collibra, as well as earlier-stage companies like Onum, a Spanish startup founded in 2022 that reduces data analytics spend by an average of 50% by eliminating incomplete and duplicate data.

Optimizing Hardware Infrastructure

Virtualization / Containerization

Another way to reduce carbon emissions is to shrink the hardware footprint required to run your systems by deploying virtualization and containerization technologies. This reduces the need for additional hardware and minimizes energy consumption.

- Virtualization lets multiple virtual machines (VMs) share a single physical server, dynamically allocating CPU, memory, and storage to each VM. This consolidation boosts efficiency, simplifies management, and scales more smoothly.
- Containerization takes efficiency further by packaging applications in lightweight, self-contained environments. Because containers use far fewer resources than full VMs, you can run more of them on the same hardware, speeding up deployment, scaling, and infrastructure optimization.

13. Decarbonizing the Data Center for Dummies

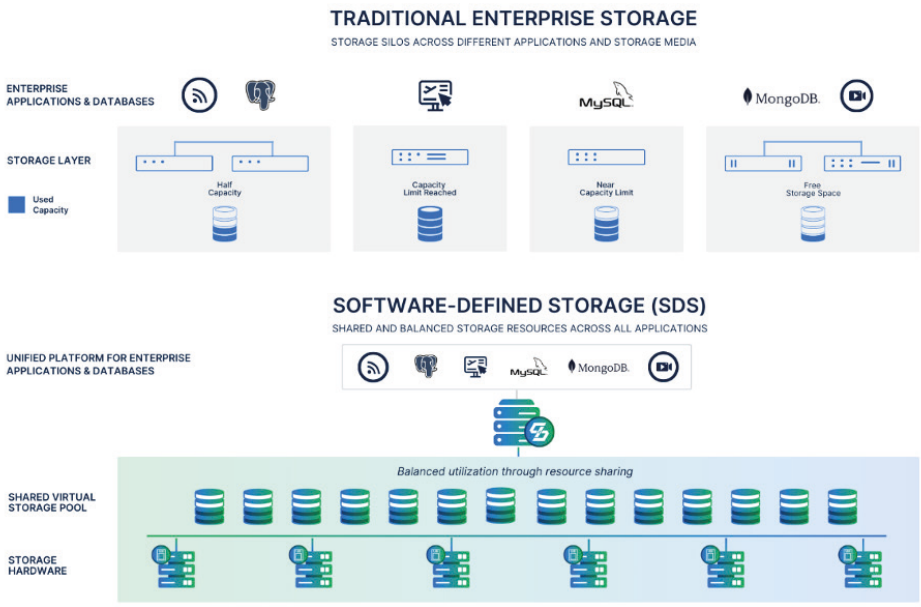
To effectively manage containerized applications, orchestration platforms - most notably the open-source standard Kubernetes - automate deployment, scaling, and monitoring to ensure seamless operation and optimal resource utilization. Because Kubernetes itself can be complex, it typically requires complementary tools to secure and manage cluster deployments. At AVP, we dedicate significant effort to evaluating SaaS solutions - such as the US-based Rafay - that simplify and streamline sophisticated Kubernetes environments.

By adopting containerization, businesses can significantly reduce their carbon footprint while improving performance and scalability. This makes containerization a key technology in the push for sustainable computing and eco-friendly IT infrastructure.

Software-Defined Storage (SDS)

SDS is another technique used in data storage management separating the software that handles tasks like storing, protecting, and organizing data from the actual physical hardware. SDS solutions use a layer of software to hide the differences between hardware types, making everything work together smoothly. Unlike hypervisors, which split one server into many virtual machines, SDS brings together different storage devices into one system that's easy to manage from a central place.

In the same manner Kubernetes does with compute, SDS dynamically provision, scale and manage storage. It can play a role in reducing energy consumption and a system's carbon footprint, especially at scale, in several ways. First, it improves resource efficiency by pooling and optimizing storage, which reduces the need for excess hardware. SDS also extends the life of existing hardware by allowing older and newer devices to work together, helping to cut down on electronic waste. It intelligently places data based on how often it's accessed, which ensures that energy-intensive storage is only used when necessary. SDS supports hybrid and cloud-based storage models, enabling organizations to shift data to more energy-efficient environments. SDS also allows hyperconverged deployments, where compute and storage share the same infrastructure for the highest resource efficiency—something traditional storage systems can't support. Finally, its automation capabilities allow systems to scale only when needed and power down idle resources, saving even more energy.



Source: Simplyblock

A promising company in the space is Simplyblock, a Berlin-based company which develops the next generation software-defined storage by handling high-throughput workloads by dynamically balancing performance, latency, and cost.

Using GenAI Selectively and Purposefully

Businesses should thoughtfully assess which AI technique is most appropriate for a given task, rather than defaulting to generative models. While Generative AI is powerful - particularly for tasks like code generation, content creation, and research automation - it is also among the most energy-intensive forms of AI. In many cases, a simpler predictive model can deliver sufficient results with a far lower environmental footprint.

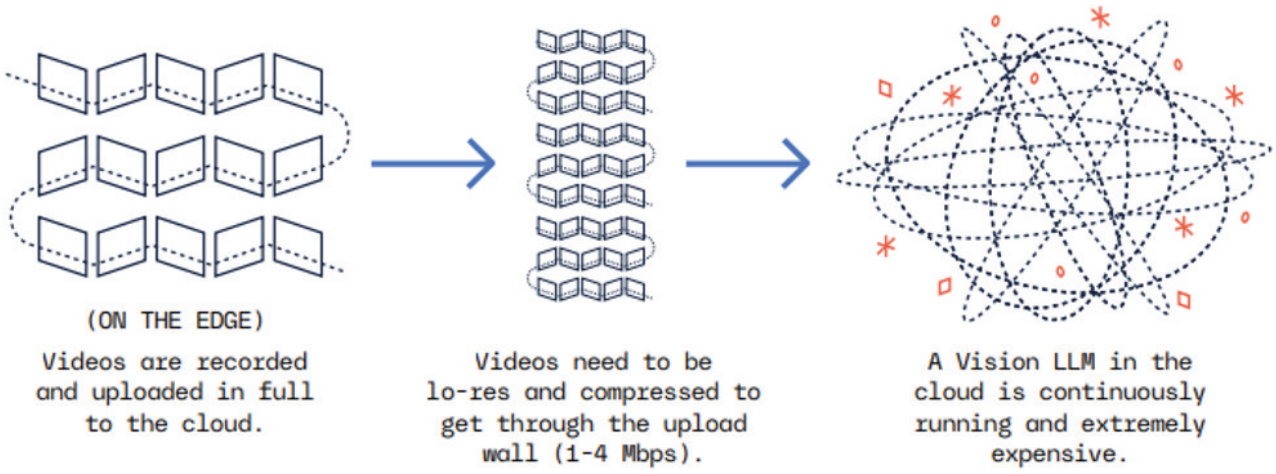
A promising approach to making GenAI more sustainable is through Edge GenAI, which deploys models directly on local devices such as smartphones, sensors, or industrial gateways. This reduces reliance on large, energy-hungry cloud data centers and minimizes the energy costs associated with transmitting data over networks. Because edge-deployed models are typically smaller and optimized, they enable faster, low-power processing and real-time inference. This is especially beneficial in sectors like smart agriculture, energy grids, and manufacturing, where local decision-making helps reduce waste and improve efficiency. Additionally, by extending the life of existing hardware and avoiding constant upgrades, Edge GenAI can help lower electronic waste and hardware-related emissions.

An example of innovation in this space is Plumerai, a company developing ultra-compact AI models for edge-based video intelligence. Their technology powers smart cameras in applications such as home

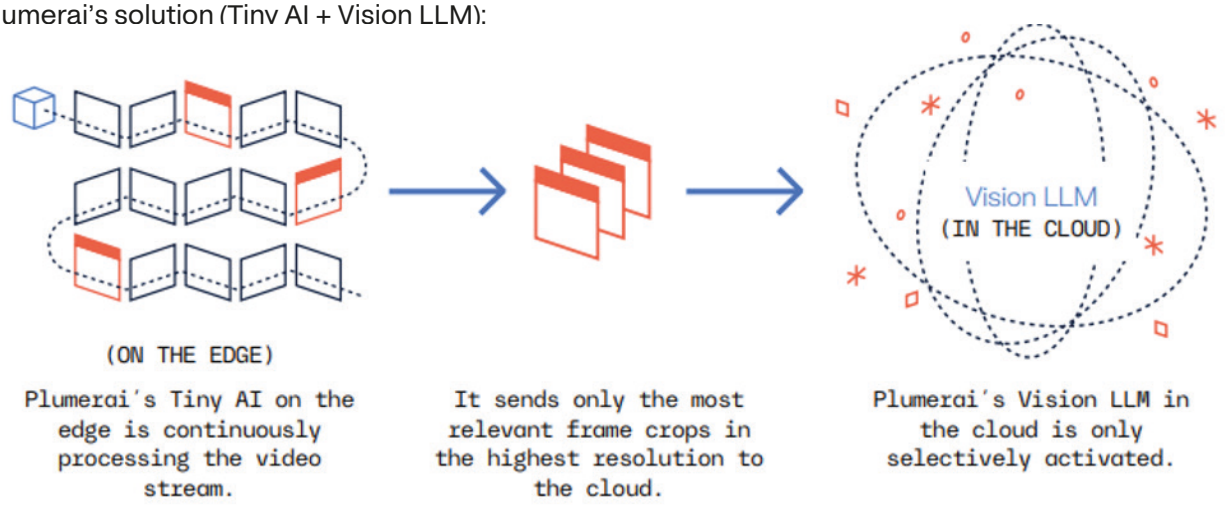
security, enterprise monitoring, and retail analytics. By embedding intelligence directly into the device, Plumerai's models reduce the need to constantly offload video data to the cloud, avoiding the energy costs of querying large vision models. Their system processes the video stream on-device, sending only the most relevant, high-resolution frame crops to the cloud for further analysis. The Plumerai Vision LLM, hosted in the cloud, is activated only when necessary, ensuring intelligent use of compute resources. This hybrid edge-cloud approach not only avoids continuous data transfer but also achieves high accuracy and ultra-fast performance, as the models operate on uncompressed local data.

Plumerai Tiny AI directing cloud-based Vision LLMs

Cloud-based intelligence video solutions:



vs. Plumerai's solution (Tiny AI + Vision LLM):



Source: Simplyblock

In short, by using GenAI selectively and shifting intelligence closer to where data is generated, Edge GenAI offers a concrete and efficient path to reducing AI's environmental impact.

Conclusion

As we conclude this white paper, it's essential to recognize that mitigating AI's environmental impact requires exploring a diverse array of green computing technologies beyond those discussed herein. Among these, quantum computing stands out for its potential to revolutionize energy efficiency in computational processes.

Quantum systems can tackle complex optimization problems, such as energy grid management, climate modeling, and supply chain logistics, with significantly less energy than classical supercomputers. This efficiency arises from quantum computing's ability to process vast combinations simultaneously, potentially reducing the carbon footprint associated with large-scale computations.

In alignment with this vision, AVP has invested in Alice & Bob, a French startup pioneering fault-tolerant quantum computing. Their innovative approach aims to develop the world's first error-resistant quantum computer by 2030, unlocking energy-efficient architectures that minimize computational overhead. By supporting such advancements, we at AVP are committed to fostering technologies that not only propel AI forward but also ensure its sustainability for the planet.

Beyond environmental considerations, the geopolitical and strategic implications of initiatives like Stargate are significant - particularly concerning AI dominance and its tangible impact on defence and physical security. It's almost certain that other global powers will respond, with similar initiatives likely to emerge in Europe and China.

avpcap.com

Atlantic Vantage Point

Bringing perspectives and investing in tech
from both sides of the Atlantic

New York Paris London