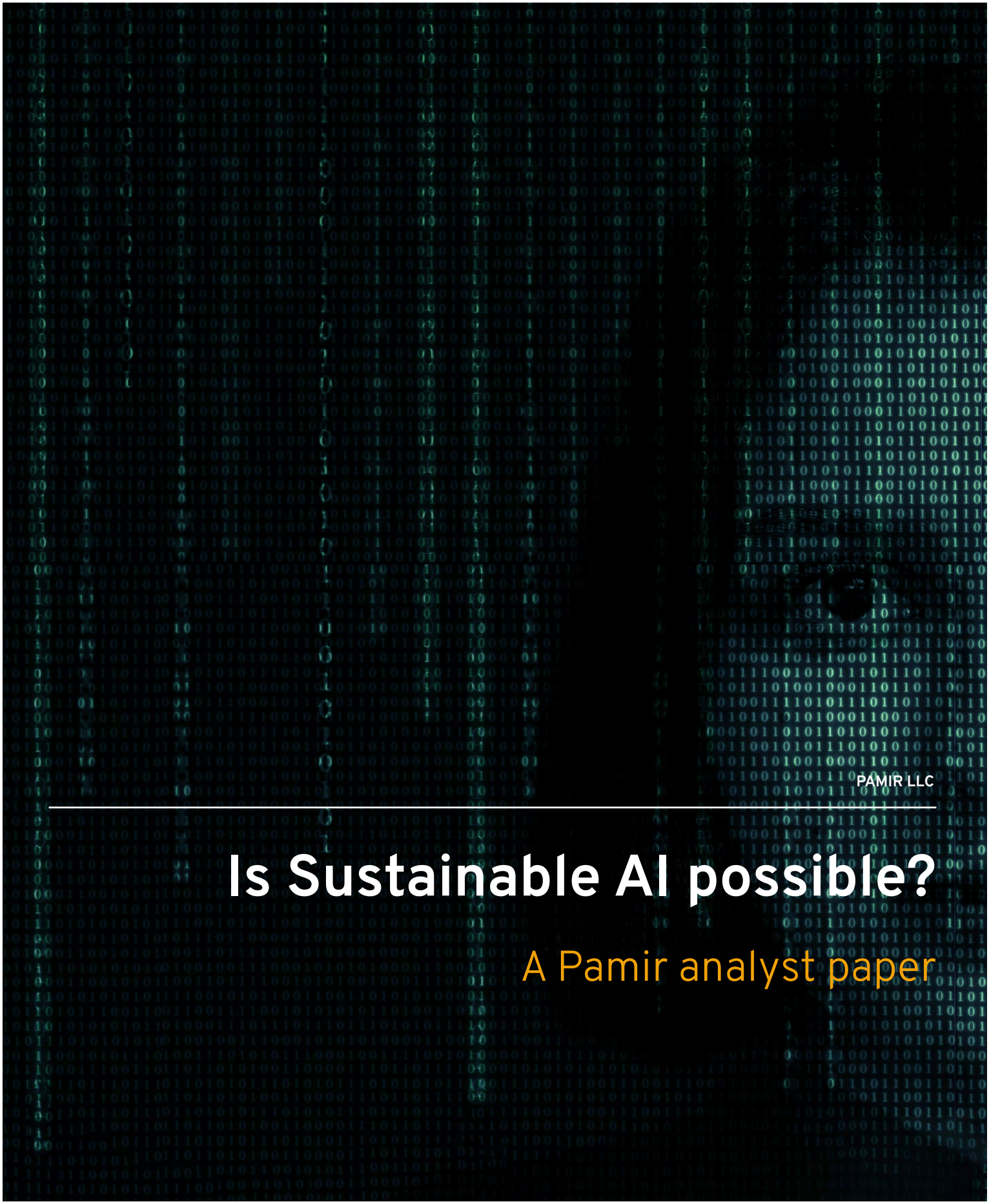
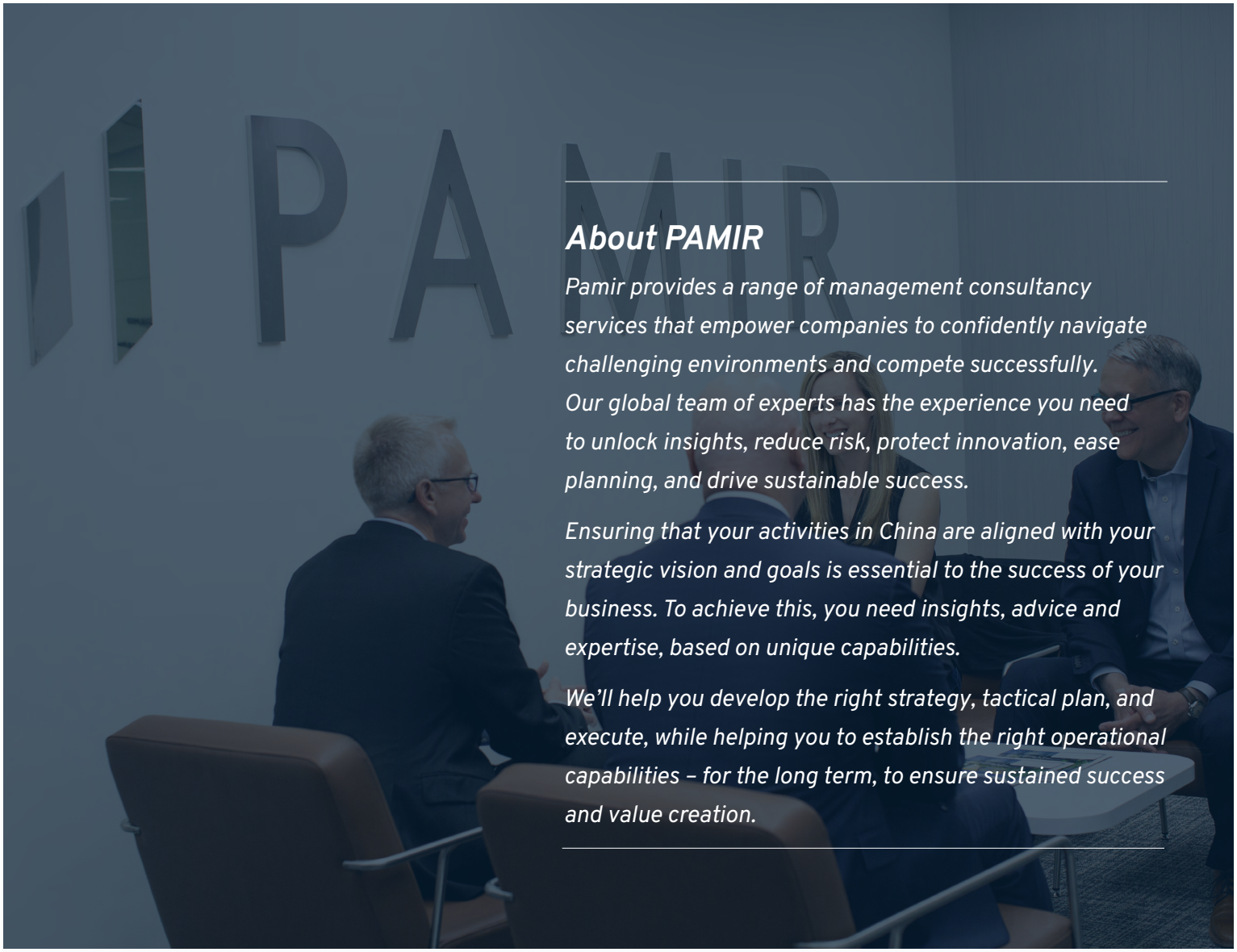


PAMIR LLC

Is Sustainable AI possible?

A Pamir analyst paper



About PAMIR

Pamir provides a range of management consultancy services that empower companies to confidently navigate challenging environments and compete successfully. Our global team of experts has the experience you need to unlock insights, reduce risk, protect innovation, ease planning, and drive sustainable success.

Ensuring that your activities in China are aligned with your strategic vision and goals is essential to the success of your business. To achieve this, you need insights, advice and expertise, based on unique capabilities.

We'll help you develop the right strategy, tactical plan, and execute, while helping you to establish the right operational capabilities – for the long term, to ensure sustained success and value creation.

Pamir Consulting LLC

8045 Leesburg Pike, Suite 220,
Vienna, Virginia
USA | 22182

Office: +1 703 783 5020
Mobile: +1 231 360 3408
Email: info@pamirllc.com
Website: www.pamirllc.com

A balancing act between AI computation power and environmental impact

Artificial Intelligence (AI) and Machine Learning (ML) are set to revolutionize technological innovation, transform industries and business operations, and create unforeseen new use cases and applications for consumers and enterprises alike. Its impact will be so widespread that some industry observers have likened it to the Industrial Revolution.

However, its adoption has sparked a critical dialogue about its environmental impact. AI requires vast energy-hungry, thirsty data centers to meet the computational demands of AI. AI can therefore be directly responsible for carbon emissions (when using non-renewable sources of energy), massive energy usage (at cost to other processing tasks), and the consumption of large volumes of fresh water required for cooling data centers and the semiconductors that drive them.

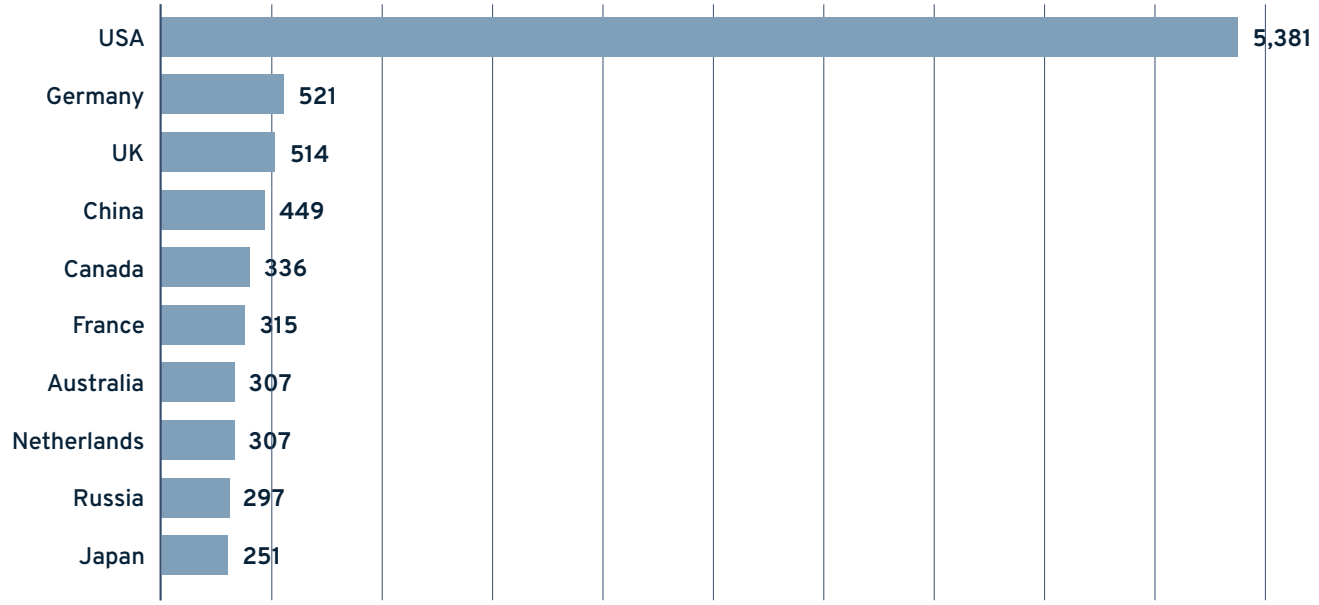
For example, ChatGPT – a generative-AI chatbot developed by U.S. AI specialist, OpenAI, and launched in November 2022 – has an estimated daily power usage almost equal to 180,000 U.S. households, while just a single ChatGPT conversation uses about 50cl of water[1]. Given that ChatGPT has around 180 million active users, it's plain to see the environmental demands of AI[2].

That's just from a consumer application – when it comes to Generative Pre-trained Transformer 3, or GPT-3 (a neural network machine learning model that uses internet data to learn and generate text, such as articles), researchers at the University of California found that the model consumes roughly 1 liter of water to generate 40-100 responses and 1 gigawatt-hour of electricity to train[3].

As technology companies invest heavily in embedding AI into everything from chatbots to climate modelling to self-driving vehicles, the impact on the environment must be considered to ensure that the benefits of the technology are not outweighed by its potential threat to the climate.

> *While the People's Republic of China has around 448 data centers, that number is continuing to rise amidst increased adoption of AI and supply chain transformation. China is one of the leading countries for hyperscalers. It is also a leading innovator, including the world's first underwater data center near Hainan Island, China, operated by Highlander.*

Figure 1: *Leading countries by number of data centers (March 2024)* ***Source:*** *Statista*



How many data centers in the world, by country

According to Statista, the U.S. has far more data centers than any other country in the world at 5,381 (as of March 2024), with data center revenue expected to reach $99.16 billion in 2024[4]. The U.S. has some of the largest data centers in the world and is the largest region for connectivity and cloud in the world. It is scaling up to meet the demands of AI.

The United States

Northern Virginia is home to some of the largest data centers, hosting technology companies such as AWS, Meta, Google, and Microsoft. Other important hubs include Phoenix, Dallas, Atlanta, Chicago, Silicon Valley, Portland, New York and New Jersey, Seattle, and Los Angeles.

Digital Realty, a co-location investment specialist, owns the most data centers in the United States with 124 facilities, followed by Equinix which operates 76 facilities[5].

But Meta Platforms owns the largest data center in the U.S. (in Oregon) at 4.6 million square feet. (Co-location service operators provide the building, cooling, power, and so on, but customers provide their own computational power i.e., servers.)

Figure 2: *Important data center hubs in the US*

Germany

Germany comes next with 522. Germany is the leading country in Europe with an estimated processing power size of 1,720 MW (megawatts) in 2023. Frankfurt, London, Amsterdam, Paris, and Dublin (FLAP-D) comprise an important data center hub (with Frankfurt having a true two-data center hub. But Berlin is growing rapidly as an essential cloud hub, with Google having launched a cloud region there in 2023, while Microsoft serves its Azure Germany North Europe region from the city. Berlin is also an important logistics gateway to Eastern Europe, which requires digital support.

United Kingdom

The UK is home to 517 data centers, but the market is undergoing significant growth. According to Statista, its value will grow from $14.64 billion in 2023 to $18.73 billion by 2028. The UK is also home to innovative spinoff companies such as Google DeepMind and Darktrace.

People's Republic of China

While the People's Republic of China has around 448 data centers, that number is continuing to rise amidst increased adoption of AI and supply chain transformation. China is one of the leading countries for hyperscalers. It is also a leading innovator, including the world's first underwater data center near Hainan Island, China, operated by Highlander. The center is submerged 35m under the surface and uses the sea as a natural cooling method. It can process more than four million high-definition images within 30 seconds – the equivalent of 60,000 traditional computers working simultaneously.

Canada

Canada has 335 data centers across the country, with growth from $3.32 billion in 2022 to $5.48 billion expected by 2028 – a CAGR of 8.71%. In November 2023, Microsoft announced it would spend $500 million to expand its data center

infrastructure in eastern Canada with multiple data centers in Quebec aimed at advancing its AI and cloud computing infrastructure in the region.

Other notable regions

France, and Paris in particular, is a well-established hub in Europe (314 data centers), with Digital Realty and Blackstone two of the largest operators. In February 2024, Equinix, wpd, and Schneider Electric signed an agreement to build new wind farms to further decarbonize energy for data centers. NTT has also invested in the region to grow its generative AI capabilities.

The Netherlands (299) has a number of leading companies investing in the country, such as Iron Mountain, Equinix, Digital Realty, and Google, with many aiming to utilize renewable energy. In Australia (306), companies such as ThreatLocker are building data centers to enhance cybersecurity, while Japan (218) is seeing rapid demand for data centers – again NTT has invested in the region to boost its Gen AI capabilities.

Russia (297) is keen to continue expanding its data center capability, with 70% of the country's centers located in Moscow, but that growth has been thwarted by its conflict with Ukraine. Ireland has also seen rapid growth with Meta, Google, and Microsoft having built data centers in the country, with more planned.

WHAT IS A DATA CENTER VS. A HYPERSCALER?

A data center is a large network of computing (servers), storage, and infrastructure components that can process and store huge amounts of data. Data centers can either be enterprise (often used by a single organization), co-location (which can be rented out or shared), managed service (offered as a service directly to customers), or cloud (which are distributed). They can support hundreds of servers; however, they may not be able to support the massive computational demands of AI and ML.

Hyperscale data centers, or hyperscalers, can support thousands of servers to meet the needs of AI (and other applications). They are a cloud-based, remote version of a data centers served by high-speed connectivity and offer massive computing power with elastic capacity. They can be individual hyperscalers or a network of data centers. Either way, they will be essential to support the needs of AI. However, there is concern over their huge energy and cooling requirements.

WHAT IS GENERATIVE AI (GEN AI)?

Where AI uses ML to make a prediction based on existing data (such as identifying a tumor within an X-ray based on how it has been trained), Generative AI (or Gen AI) uses existing data to create completely new content, such as text, imagery, audio, synthetic data. Essentially, it creates new data and objects that are similar to the content with which it was trained.

ChatGPT, for example, uses billions of parameters and has been trained on a huge amount of data, much of which is online.

But Gen AI can also be used to train a computer to recognize objects, i.e., computer vision, or to create new compounds based on what it 'understands' of existing chemical structures.

The sheer number of computations required, however, means that Gen AI uses much more energy than traditional AI. Hyperscalers generally use graphics processing units (GPU) – as opposed to traditional CPUs – which require 10–15 times more energy than a traditional CPU.

Hyperscale data centers, by country

Hyperscale data centers are required to meet the exceptional computational demands of AI and ML. As a result, the total capacity of hyperscale data centers has doubled in the past four years, and there were more than 1,000 large hyperscale data centers in operation worldwide at the end of 2023, according to Synergy Research Group[6]. This is up from 728 at the end of 2021, and Synergy predicts that there will be around 1,200 by 2026 to support Gen AI workloads.

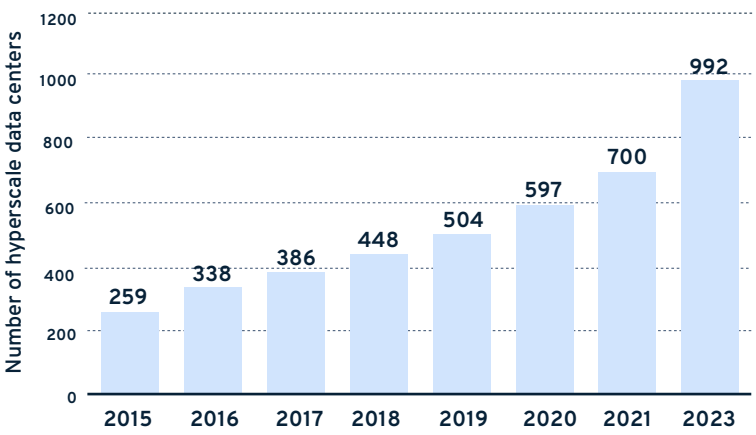
Synergy also forecasts that there will be a further 120-130 hyperscale data centers coming online within the next 5 years, driven largely by Gen AI requirements. Figure 3 below shows hyperscale data center growth.

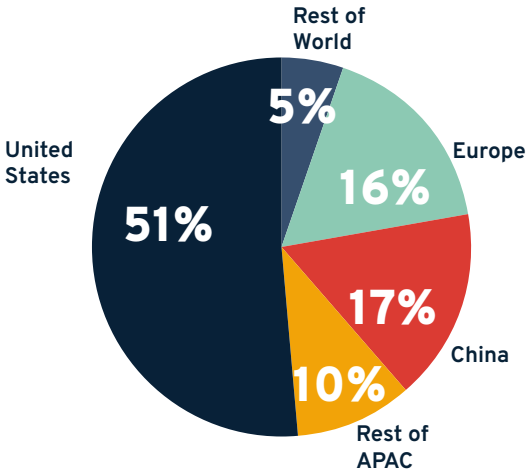
As shown in the graph, the U.S. accounts for 51% of the world's total hyperscale data center capacity – measured by megawatt (MW) – as at the end of 2023. Europe and China followed, with each of them representing around a third of the balance (17% and 16%, respectively). Ireland, Germany, Japan, Netherlands, and Australia are other notable regions.

In the U.S., the state of Virginia alone accounted for one-third of total capacity in the region, while globally Amazon AWS, Microsoft Azure, and Google Cloud Platform accounted for 60% of hyperscale data center capacity.

Meta, Alibaba (China), Tencent (China), Apple, and ByteDance (China) are all noted as other leading hyperscale operators. At the same time, there is a growth of multiple smaller hyperscale operators that are building facilities to serve capacity on a more local/regional level.

Figure 3: *Hyperscale data centers - growth and capacity* ***Source:*** *Synergy Research Group*





Concern over AI demand for energy and water

Energy usage

Calculations regarding the predicted energy usage of AI, specifically, are in their infancy, and tentative given the rapid, but yet unquantifiable, adoption of the technology.

However, there are a number of reports that have focused on data center (and hyperscaler) energy usage. For example, McKinsey & Company have estimated data center consumption by provider (Enterprises, Co-location, and Hyperscalers)[7] – see Figure 4.

A co-location service provides the building, cooling, power, bandwidth, and physical security for a customer to build their own data center i.e., servers and computational power.

The figure shows that this sector will comprise the majority of energy usage. However, the rise of hyperscalers is also notable.

The International Energy Agency (IEA) provides another snapshot in its 'Electricity 2024' report[8]. Figure 5 shows its predictions for data energy center consumption to 2026. It notes Denmark and Ireland as two rapidly growing regions for hyperscaler deployments.

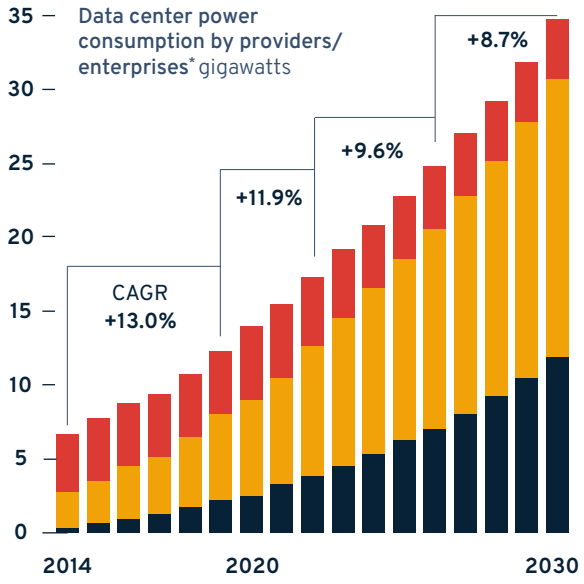
Notably, Ireland and Denmark are highlighted as they reflect the trends and concerns of the AI data center sector as a whole. For example, electricity demand from data centers in Ireland was 5.3 TWh in 2022, representing 17% of the country's total electricity consumption – equivalent to all its urban household electricity usage.

But, by 2026, the IEA predicts that AI applications will mean that data center energy requirements may reach a share of 32% of the country's total electricity demand. Ireland has one of the lowest corporate tax rates in the European Union (12.5%), which has led to Meta, Google, and Microsoft building new data centers in the country.

Likewise, Denmark (which hosts 34 data centers, half of them in Copenhagen), is seeing rapid data center expansion, which is expected to account for 6 TWh by 2026 (20% of the country's electricity demand). Denmark is also the hub for a new pan-European initiative, Net Zero Innovation Hub for Data Centers – which aims for innovation, pan-industry collaboration, and public engagement in an attempt to ensure that data center growth remains focused on helping countries to reach their net-zero carbon goals.

In the U.S., meanwhile, McKinsey provides a forecast for hyperscaler growth in the region. See Figure 6 (provided by The Economist)[9].

Figure 4: *Data center power consumption by provider* ***Source:*** *McKinsey & Company*



* Demand is measured by power consumption to reflect the number of servers a data center can house. Demand includes megawatts for storage, servers and networks.

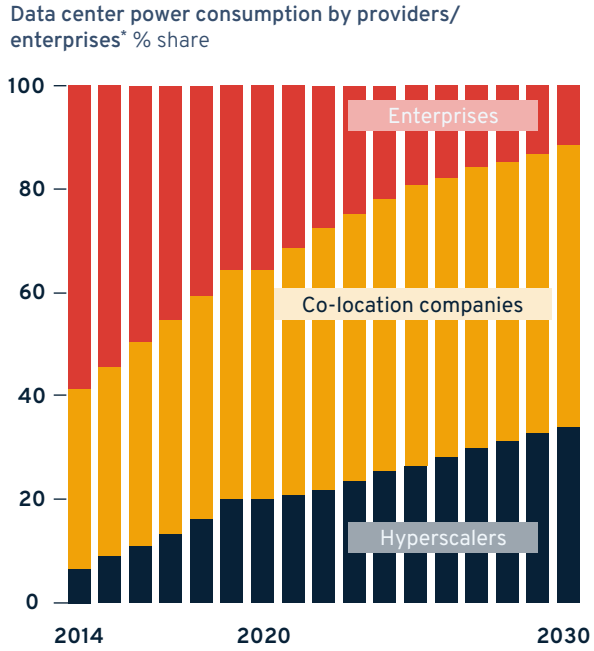
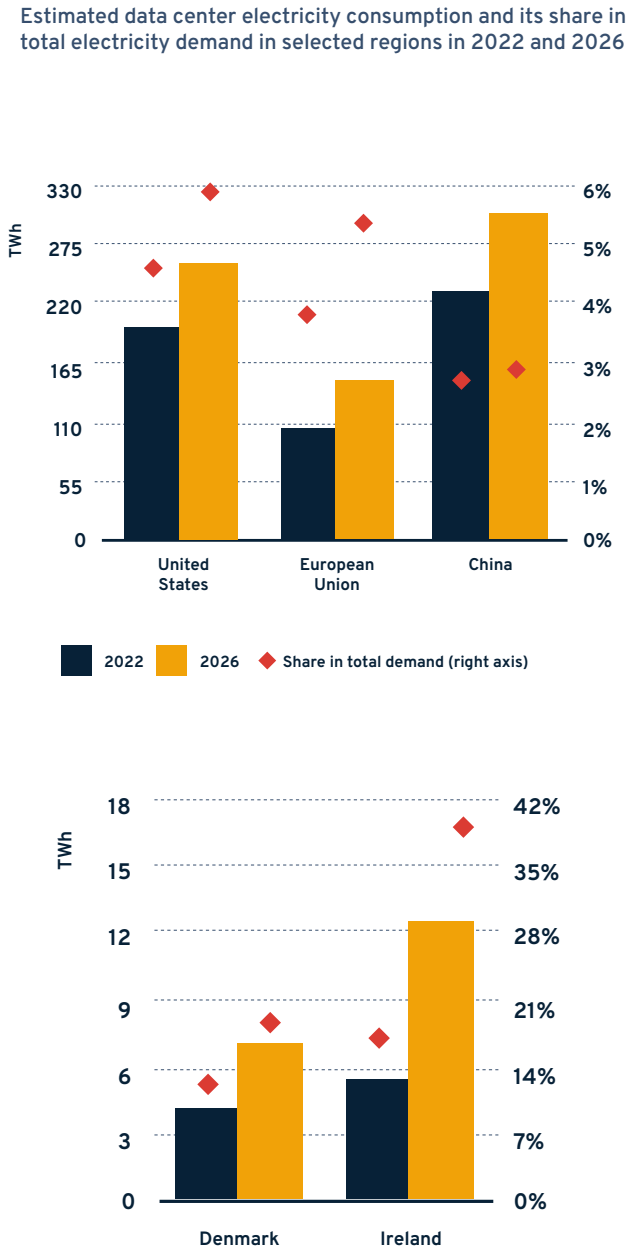


Figure 5: *Data center energy consumption to 2026* ***Source:*** *IEA*



Statistics for China, meanwhile, are hard to come by. However, China's State Grid Energy Research Institute (SGERI) expects electricity demand from data centers to double to 400 TWh by 2030, compared to 2020. Meanwhile, the IEA forecasts that electricity consumption from data centers in China will reach around 300 TWh by 2026 – slightly ahead of SGERI's forecast.

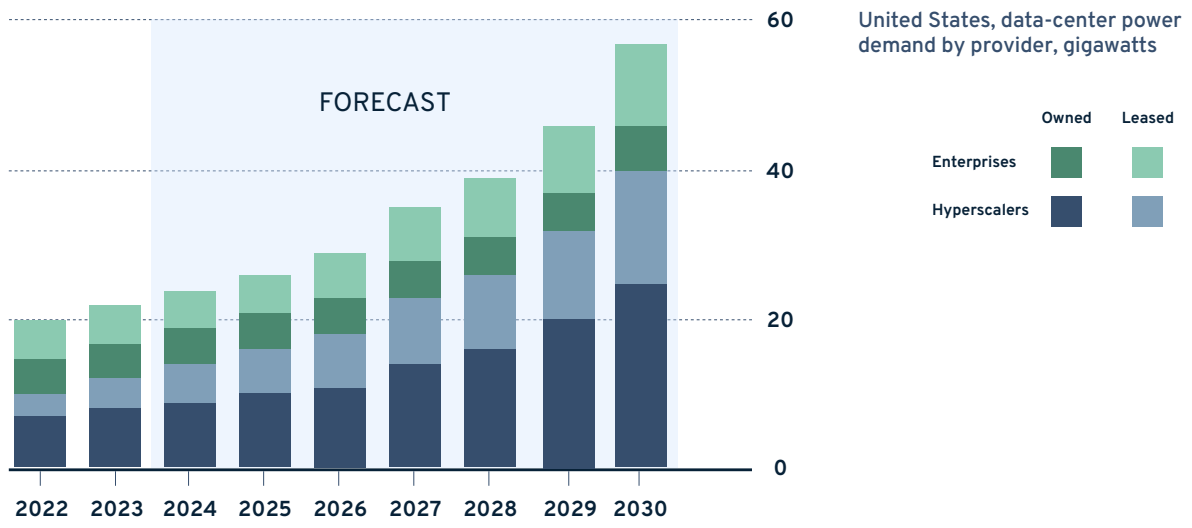
It is currently difficult to disassociate AI-based data center energy usage and non-AI usage. However, the Organization for Economic Cooperation and Development (OECD) AI Policy Observatory estimates that a typical large AI model consumers between 100-200 MWh of energy during training, which is around 10-20 times the average annual energy us of a family home in the U.S.

Water consumption

Data on the water consumption load of data centers are scarce, partly because currently technology companies are not required by legislation to reveal how much energy and water they use. But data centers require large volumes of water to not only cool large numbers of servers and processors but also to maintain a stable temperature so that IT equipment can run at an optimal level. This has led to multiple concerns, sometimes involving local authorities and water companies about the use of local freshwater for data center cooling.

For example, in 2022, the city government of The Dalles, Oregon, where Google runs three data centers, filed a lawsuit after public records showed that Google's data centers were using around one-quarter of the city's water supply[10]. While in February 2024, an environmental court in Chile partially reversed Google's permit to build a $200 million hyperscaler data center in Cerrillos, Santiago, due to

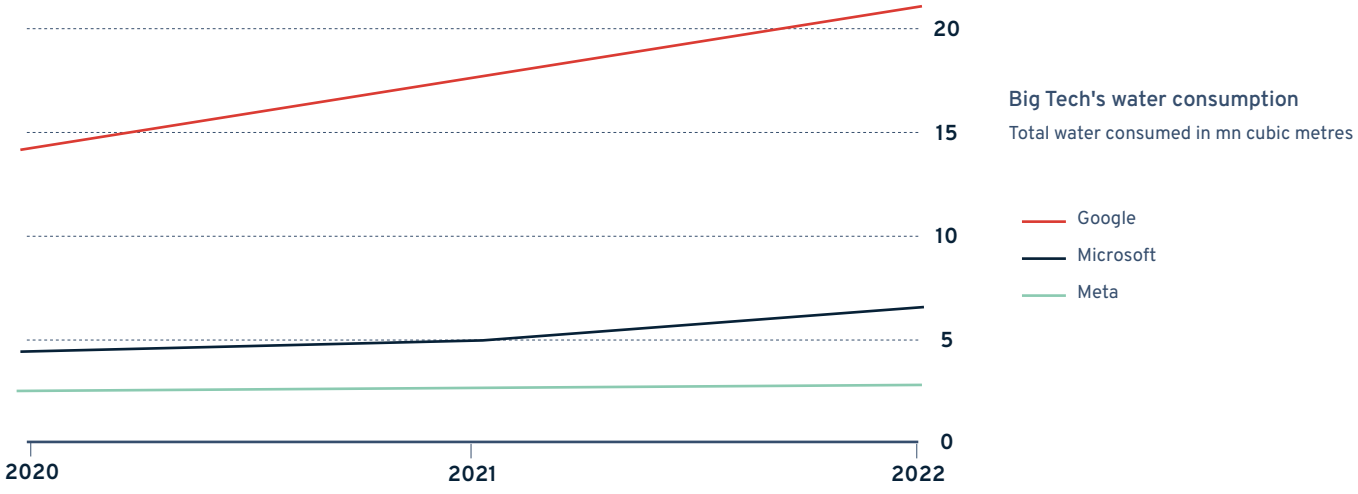
Figure 6: *Hyperscaler power demand in the US to 2030* ***Source:*** *McKinsey & Company*



concerns over climate change and over its use of the region's already strained reservoir.

Official company figures from 2022, show that Microsoft increased its water consumption by 34% on the previous year, while Google used 22% more in the same period, and Meta 3% more. See Figure 7.

Figure 7: *Microsoft, Google, and Meta increased data center water consumption, 2020-2022*
Source: *FT*



Arguably, this is one of AI's largest challenges – energy usage can be tempered to a certain extent through the use of renewable energy.

In fact, AI can be used to optimize energy usage, bringing further benefits. According to Water Resources Institute, at least 50% of the world's population (around 4 billion people) already live under highly water-stressed conditions for at least one month of the year[11].

It means that the need to reduce water consumption will require innovative solutions, including more efficient equipment, new cooling methods such as non-water-based coolants, and location-based solutions such as underwater data centers and centers based in cooler regions.

Lack of Transparency / Regulations

One of the problems for AI energy usage is that it's currently not possible for consumers of the technology to know how much energy different AI models or applications use, as there are no regulations governing transparency.

That means that technology providers are under no obligation to publish energy usage figures – at least in a consistent and comparable manner. Only Google has tried to calculate data center energy usage of AI only, stating that ML accounts for less than 15% of its data centers' energy use.

At the same time, most countries are just setting out on their journey for publishing AI rules and guidelines, but most of these relate to its use, rather than environmental sustainability. A few regions, including the United States, the EU, and the UK, have drafted preliminary guidelines but there is still little focus on energy and water usage and carbon emissions.

In May 2023, the European Parliament requested that environmental sustainability be added to the EU AI Act, listing it as a high-risk area. However, it so far only goes as far as to state: "Providers should also be encouraged to apply on a voluntary basis additional requirements related, for example, to environmental sustainability…"

Some governments and states have introduced data center moratoriums to give them more time to assess the environmental impact of AI data center growth and the regulations required. For example, in July 2022, Frankfurt announced that new data centers could only be built in certain areas of the city.

However, the aim is to limit the competition for space as "uncontrolled data centers can put up land prices, and crowd out other sectors."

It has also announced plans for a city-wide initiative to use the waste heat from data centers for other purposes. At the same time, data centers are expected to have to meet a "Blue Angel" efficiency stamp, but details for quality and resource use are still being developed.

In June 2022, Groton in Connecticut placed a one-year moratorium on large data centers to give the city time to consider whether to limit data centers to "light industrial zones and regulating data centers that are 10,000 square feet or larger."

Other jurisdictions previously imposed moratoriums, including Ireland (Dublin currently has a de facto moratorium on the building of new data centers in the city), and Amsterdam (which restricts new data centers from its central area).

Singapore also had a moratorium on new data center growth that has now lapsed, but the Infocomm Media Development Authority has recently issued new standards that allow an increase in data center operating temperatures in tropical zones, which could lead to a 2-5% reduction in the need for cooling. Of course, moratoriums are temporary – and not a long-term solution.

Federal Artificial Intelligence Environmental Impacts Act 2024

The U.S. has started the process for regulations around AI environmental sustainability. On 1 February 2024, the Federal Artificial Intelligence Environmental Impacts Act of 2024 was introduced to assess and mitigate the environmental impacts of AI technologies and to promote transparency and accountability in the development and use of AI systems.

Although a relatively short bill, the AI Environmental Impacts Act has four key provisions, which are the responsibility of the National Institute of Standards and Technology (NIST):

1. A study on the environmental impacts of AI and submission of a report to Congress.
2. The establishment of a consortium of stakeholders to investigate the environmental impacts of AI.
3. The establishment of a system for voluntary reporting by entities on the environmental impacts of artificial intelligence.
4. A report to Congress on findings surrounding the environmental impact of AI.

The Bill includes the requirement for the Administrator of the Environmental Protection Agency (EPA) to conduct a comprehensive study on the environmental impacts of AI, including energy and water consumption, pollution, and identifying local-scale impacts such as grid and water stress.

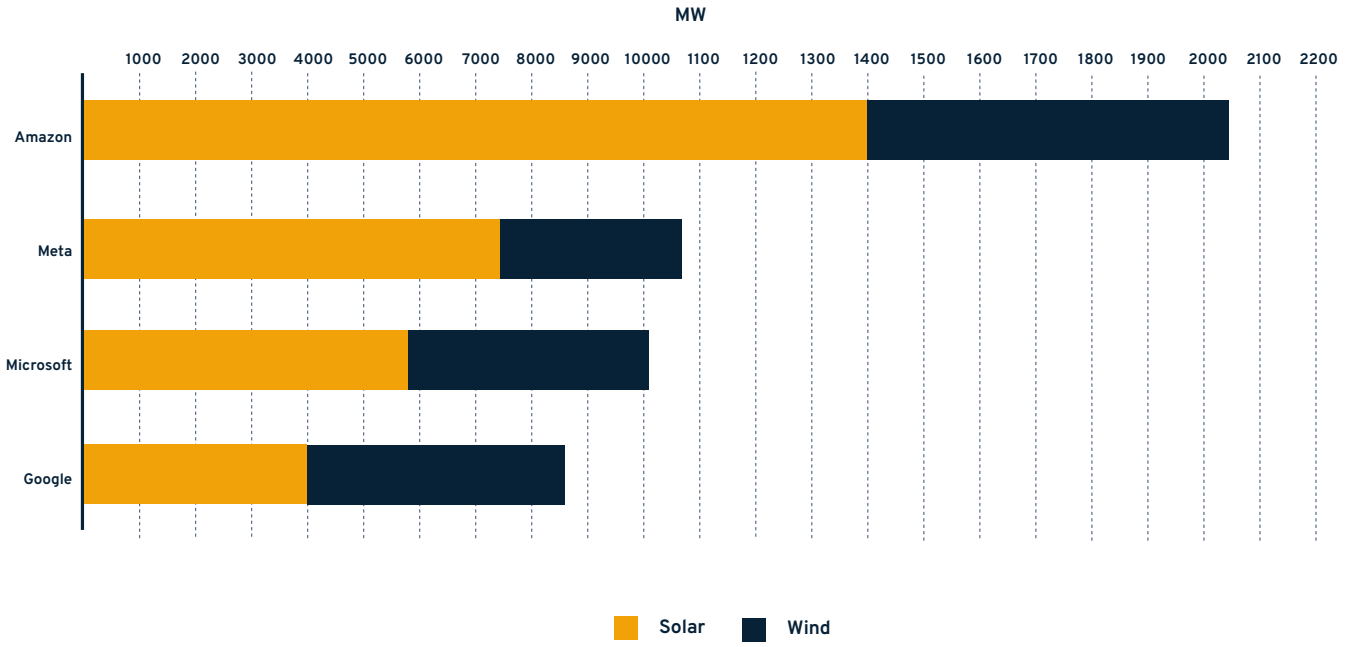
European Union

The EU has regulatory and voluntary schemes to improve energy efficiency at the component level (e.g. servers, data storage, heating, ventilation and air conditioning) such as ENERGY STAR and EU Ecodesign Regulations for servers and data storage products. Buildings-based data center energy efficiency certifications include the EU Code of Conduct on Data Centre Energy Efficiency, and BREEAM SD 5068 (in the UK). The Corporate Sustainability Reporting Directive (CSRD) was adopted in 2022 by the European Commission to come into force from 2024 onwards. It will require large organizations, including technology companies, to report sustainability indicators as well as energy and carbon emissions. The amended Energy Efficiency Directive (EED) will introduce energy and sustainability reporting requirements for data centers based in the European Union from May 2024. It will require data centers with installed capacity greater than 500 kW to report total energy consumption – including the share derived from renewable energy – water usage, and waste heat utilization.

China

In China, the government has called for average power use effectiveness (PUE) – calculated by dividing the total amount of power entering a data center by the power used to run the IT equipment within it – of 1.25 in the east and 1.2 in the west of the country as part of its Eastern Data and Western Computing project. Major cities have minimum PUE requirements for new data centers, including Beijing (1.4), Shanghai (1.3) and Shenzhen (no subsidies above 1.4).

Figure 8: *Power purchase agreements by Amazon, Microsoft, Meta and Google* ***Source:*** *IEA*



How to move to Sustainable AI?

Without a lack of current guidelines and regulations, much of the move towards Sustainable AI is being driven by the industry itself. There are a number of areas that can be addressed to reduce AI's environmental impact, such as:

- Carbon-Neutrality/Renewable Energy: Many industry players are committing to using 100% renewable energy by 2030. However, many smaller operators are setting out with CSR as a key focus from the start. The use of renewable energy will reduce the carbon footprint of data centers; however, this of course depends on availability and still needs to be used responsibly and shared for other purposes.
- Advanced Cooling Techniques/Water Management: Innovative cooling solutions, which can range from undersea data centers to the use of advanced materials that are more efficient at cooling than water, are being developed to reduce reliance on local freshwater supplies.
- High-efficiency Computing: The development of more efficient GPUs, processors, servers, and so on, is essential to not only reduce energy consumption, but also heat generation (which in turn requires less cooling).

Carbon Neutrality/Renewable Energy

Some of the largest players, including Google, Meta, Amazon, and Microsoft have committed to carbon neutrality by 2030, and many operators are major purchasers of renewable energy – while many do not consume renewable energy directly (due to a lack of availability), there is a commitment to power purchase agreements (PPAs).

Amazon, Microsoft, Meta, and Google are the four largest purchasers of corporate renewable energy PPAs, having contracted almost 50 GW to date, equal to the generation capacity of Sweden (see Figure 8).

Investment in renewable energy projects also protects data center operators from price volatility, while improving their brand reputation. Apple (2.8 TWh), Google (18.3 TWh), Meta (9.4 TWh) and Microsoft (13 TWh) purchased or generated enough renewable electricity to match 100% of their operational electricity consumption in 2021 (primarily in data centers).

However, renewable energy in a region may not match the energy profile of a local data center, and so it's often purchased (as a PPA) from a different region. But Google and Microsoft have set a 2030 target to source and match zero-carbon electricity on a 24/7 basis from each grid where demand is located – Iron Mountain has set a 2040 target.

Meanwhile, Chinese data center provider GDS has promised to achieve net-zero carbon emissions by 2030. The company has data centers across Southeast Asia as well as China, and has already switched 30 of its sites to renewable energy.

In January 2024, Tencent launched a new microgrid project at its Tianjin High-Tech Cloud Data Center, which it says generates enough solar energy to power 6,000 households. It installed 10.54MW of solar panels capable of producing 12 million KWh of electricity annually. To mitigate against fluctuations in supply, the data center operator has introduced filtering and suppression techniques, installed equipment to store excess energy, and is using an AI system to manage the energy demands of its facility.

Advanced cooling techniques/Water management

As noted, one of the main issues around data centers has been their use of water, usually taken from local freshwater supplies to cool equipment such as servers and GPUs/CPUs. This has led to concerns across the world about the amount of water that AI-driven hyperscalers, in particular, require given the number of regions on the planet facing water stress and shortages.

While some data centers use air cooling and evaporation to keep IT equipment at optimum operating efficiency, most have traditionally used water. However, several innovations are aimed at reducing the water usage of data centers for cooling.

Liquid cooling (non-water-based) is a relatively new technology that's slowly being introduced. Liquid-immersion cooling places the entire server equipment into dielectric fluid (either fluorochemicals or hydrocarbons) in a closed system. The fluid absorbs the heat emitted by the device, evaporates it and then condenses it, helping the device to cool down, much like a human body uses sweat to cool down.

A second liquid-cooling method uses non-flammable dielectric fluid, which is used to directly cool the processing chip or motherboard component (which generates the most heat) specifically, rather than immersing the entire system.

Other innovations include natural cooling methods, such as geothermal cooling, which uses the near-constant temperature below ground level. In data centers, geothermal cooling uses a closed-loop pipe system with water or another coolant that runs through vertical wells underground, filled with a heat-transferring liquid.

Iron Mountain's western Pennsylvania data center, Verne Global in Iceland and Green Mountain in Norway use geothermal cooling for their data centers.

Going a step further, China's HiCloud, a division of Highlander, announced in January 2024 that its new 35m deep underwater data center off the coast of Lingshui Li County, Hainan Province, was fully operational. Using the natural coolness of the sea, the company claims that the servers achieve an overall energy efficiency improvement of 40-60%. The data center mainly serves the recently constructed Hainan Free Trade Port.

AI itself is also being used to cool data centers. For example, Google has tested its DeepMind AI to monitor and optimize temperature and humidity levels to reduce heat output and cooling usage. The company claims to have reduced the cooling energy required by 40% in 18 months.

High-efficiency computing

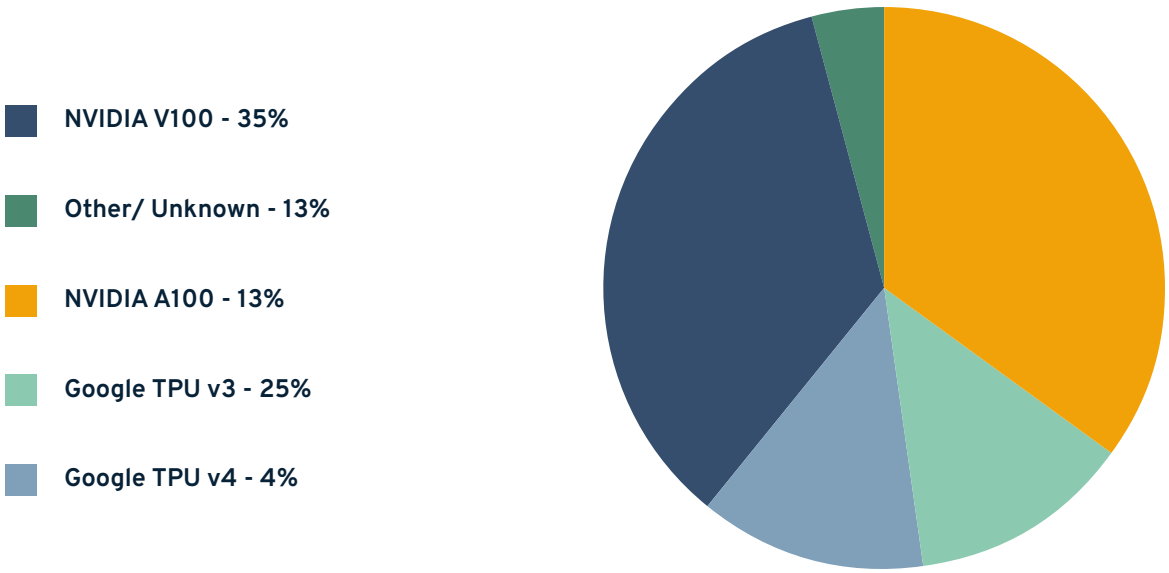
Improving the efficiency of the CPU/GPU semiconductors used within data centers is another area of focus for reducing the environmental impact of data centers and AI hyperscalers. Figure 9 shows the main chip providers for the training of AI models.

As shown in figure 9, Nvidia and Google are the leaders in AI-focused chips, however, while energy efficiency is of course a consideration, the main aim has been to boost computational power to drive AI and ML processing requirements. Nvidia holds a 65% market share in the data center AI chip market (as of 2023), up from 42% in 2022[12]. This drove a 239% increase in revenue, from $3.4 billion to $11.4 billion, during the period. Intel follows with a 22% market share.

Nvidia also dominates the AI accelerator market. AI accelerators, also known as neural processing units, are a class of specialized computer system designed specifically for AI and ML applications, including artificial neural networks. Nvidia accounted for 97% of the market in the $39 billion AI accelerator card market in 2023, up from 96% in 2022.

Nvidia's cutting-edge H100 AI GPUs are also (relatively) energy efficient. A H100 GPU, running at 61% annual utilization, consumes roughly 3,740 kilowatt-hours (kWh) of electricity annually – equivalent to the average American household. Its H200 – set for release in 2Q24 –is 50% more efficient than the H100 for both energy usage and total cost of ownership.

Figure 9: *Chips used to train large AI models, 2018-present* ***Source:*** *IEA*



But there is much innovation occurring in this space with future AI requirements, and sustainability, in mind. In April 2024, scientists at Tsinghua University in Beijing and the Beijing National Research Center for Information Science and Technology announced they had developed Taichi, a chip that uses light rather than electricity to perform advanced AI tasks. The researchers estimate it is around 1,000 times more energy efficient than a H100.

In the same month, Intel deployed the largest neuromorphic system ever made to Sandia National Laboratories, which is run by the U.S. Department of Energy's National Nuclear Security Administration. Neuromorphic computing is an entirely new kind of computation based on the human brain.

> *One of the challenges of achieving Sustainable AI is that there are few tools currently available for measuring its specific energy and water usage, and there is a lack of regulation and guidance around the building of new data centers and their consumption. It means that users of AI have no way of choosing appropriate environmentally friendly data center partners.*

Conclusions

The road to achieving Sustainable AI is complex and multifaceted, and it's important to remember – in the race for global supremacy of AI technology – the growing recognition of the problem, as well as the absolute necessity of reducing its impact. While many strides are being taken, the area is still very much in its infancy.

The proliferation of data centers has put increasing pressure on energy and water supplies, and the rapid growth of hyperscalers – which are required to provide the massive processing power required by AI and ML – has made the situation more serious. While renewable energy and PPAs can mitigate some of that energy consumption, the question becomes: How much of that energy should be dedicated to AI data centers?

At the same time, with many of the world's regions already water stressed, the increasing deployment of ever larger hyperscalers has caused local concerns about water supplies and as a result has even led to legal challenges.

One of the challenges of achieving Sustainable AI is that there are few tools currently available for measuring its specific energy and water usage, and there is a lack of regulation and guidance around the building of new data centers and their consumption. It means that users of AI have no way of choosing appropriate environmentally friendly data center partners.

For their part, the major data center players – such as Meta, Google, Microsoft, and Amazon – are playing some role in mitigating the footprint of their data centers, such as water management, more efficient equipment, and the purchase of PPAs. However, this is unlikely to be enough in the longer term.

Innovations such as liquid cooling, environmental cooling, energy-efficient chips and servers, and so on, are also needed to ensure the environmental sustainability of data centers. In truth, the solution lies in a multi-pronged approach encompassing renewable energy, optimized energy usage (ironically, through the use of AI), government regulation, industry standards, technological innovation, and a collaborative industry approach.

There is no doubt that AI offers multiple, significant opportunities for investment, innovation, and collaboration, but it must come with environmental sustainability at the forefront.

References

1. https://www.forbes.com/sites/cindygordon/2024/03/12/chatgpt-and-generative-ai-innovations-are-creating-sustainability-havoc/#:~:text=ChatGPT's%20daily%20power%20usage%20is,equivalent%20to%20one%20plastic%20bottle.
2. https://explodingtopics.com/blog/chatgpt-users#
3. https://www.economist.com/technology-quarterly/2024/01/29/data-centres-improved-greatly-in-energy-efficiency-as-they-grew-massively-larger#
4. https://www.statista.com/statistics/1228433/data-centers-worldwide-by-country/
5. https://dgtlinfra.com/united-states-data-centers/#:~:text=Houston%2C%20and%20Minneapolis.-,Who%20Owns%20the%20Most%20Data%20Centers%20in%20the%20United%20States,million%20net%20rentable%20square%20feet.
6. https://www.srgresearch.com/articles/hyperscale-data-centers-hit-the-thousand-mark-total-capacity-is-doubling-every-four-years
7. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/investing-in-the-rising-data-center-economy
8. https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf
9. https://www.economist.com/technology-quarterly/2024/01/29/data-centres-improved-greatly-in-energy-efficiency-as-they-grew-massively-larger
10. https://www.oregonlive.com/silicon-forest/2022/12/googles-water-use-is-soaring-in-the-dalles-records-show-with-two-more-data-centers-to-come.html
11. https://www.wri.org/insights/highest-water-stressed-countries#:~:text=And%20at%20least%2050%25%20of,jobs%2C%20food%20and%20energy%20security.
12. https://library.techinsights.com

Address

8045 Leesburg Pike,
Suite 220,
Vienna, Virginia
USA | 22182

Phone

Office: +1 703 783 5020
Mobile: +1 231 360 3408

Online

Email: info@pamirllc.com
Website: www.pamirllc.com