

Commentary

Artificial intelligence: Supply chain constraints and energy implications

Alex de Vries-Gao^{1,2,3,*}

¹Institute for Environmental Studies, VU, Amsterdam, the Netherlands

²Digiconomist, Almere, the Netherlands

³De Nederlandsche Bank, Amsterdam, the Netherlands

*Correspondence: alex@digiconomist.net

<https://doi.org/10.1016/j.joule.2025.101961>

Alex de Vries-Gao is a PhD candidate at the VU Amsterdam Institute for Environmental Studies and the founder of Digiconomist, a research company dedicated to exposing the unintended consequences of digital trends. His research focusses on the environmental impact of emerging technologies and has played a major role in the global discussion regarding the sustainability of blockchain technology and artificial intelligence.

Introduction

In 2023 and 2024, the rapid adoption of generative artificial intelligence (AI) applications—fueled by the launch of OpenAI's popular AI chatbot, ChatGPT—drew significant media attention to the increasing power demand of AI applications as a whole. As data center power demand rapidly rose to support these applications, tech companies such as Google faced a “power capacity crisis” in their efforts to expand data center capacity.¹ Despite this attention, it remains uncertain how the actual power demand of AI has developed over these years. Companies such as Microsoft and Google reported increasing electricity consumption and carbon emissions in their 2024 environmental reports, citing AI as the main driver of this growth. However, these companies only provided data center-wide metrics at best, making it impossible to distinguish between AI and other types of workloads. Google even described such a distinction as “not meaningful,” whereas in 2022, Google researchers still provided this type of information. At that time, Patterson et al. concluded that machine learning training and inference represented “10%–15% of Google's total energy use” from 2019 to 2021.² Aggregating such information across big tech companies would likely capture a significant share of global AI workloads and provide a solid starting point for assessing global AI power demand. However, even when it was released, this type of information was already exceptional, with Google being the only big tech company to

disclose such data.³ Now, Google has stopped providing these insights. With useful information regarding AI's power demand becoming increasingly scarce, academic research has repeatedly stressed the urgent need for better data.

However, comments in Google's 2024 environmental report suggest that better data are unlikely to be provided anytime soon. The European Union's AI Act touches on environmental sustainability but treats environmental disclosure mostly as voluntary. Members of the European Parliament who raised concerns over this voluntary nature—particularly in light of Google's comments—received only a muted response from the European Commission, as the authority merely stated it would evaluate the AI Act “by 2 years after the date of application.”⁴ The AI Act does require providers of “general-purpose AI models” to disclose the energy consumption for model training. However, inference accounted for most of Google's AI electricity costs from 2019 to 2021.² Given the mass adoption of AI over the past 2 years, inference is likely an even greater factor in the life cycle of an AI model today. As a result, this disclosure will provide, at best, only limited insights. Moreover, the AI Act's rules on general-purpose AI will not take effect until August 2025. Therefore, insights into the growing power demand of AI in recent years will need to be obtained through other means.

With hardware operators unwilling to publicly disclose details of their AI hardware electricity use, one alternative is to

analyze the AI hardware supply chain to estimate the production output of relevant devices. This output can then be combined with publicly available electricity consumption profiles to assess a potential range of total power demand. However, supply chain partners are typically bound by client confidentiality, meaning they generally do not disclose specific production output information either. At this point, assessing the development of AI power demand becomes nearly impossible without further regulation requiring any of the involved companies to provide more transparency. As a last resort, analyst estimates can be used to replace the missing inputs from the AI hardware supply chain, though this requires careful triangulation. Analyst estimates may be influenced by bias, and assessing their validity can be challenging due to the use of proprietary models and assumptions. Therefore, triangulation is necessary to improve reliability. This article will outline an approach to combine analyst estimates, earnings call transcripts, and device details to estimate AI hardware production, as well as the scale and trends in the subsequent power demand of these devices. By quantifying the potential electricity consumption of AI, the outcomes of this analysis provide a starting point for further investigation in an otherwise opaque industry, enabling stakeholders to better assess the trade-offs involved in AI development and deployment. While this analysis does not explore the full spectrum of costs or weigh them against AI's benefits, it sheds light

on a critical input for such an assessment. These insights can inform energy infrastructure planning and policy discussions on AI sustainability.

AI hardware supply constraints: The manufacturing bottleneck

Within the AI hardware supply chain, there is broad analyst consensus on one small but crucial piece of information that can be used to obtain insights into AI's growing power demand: the estimated chip-on-wafer-on-substrate (CoWoS) packaging capacity of Taiwan Semiconductor Manufacturing Company (TSMC), a key player in the manufacturing process of AI hardware. This packaging technology has been essential for AI accelerators (i.e., hardware designed specifically for AI workloads) in recent years. CoWoS enables the integration of processing units such as graphics processing units (GPUs) and high-bandwidth memory (HBM) in a single package, thereby reducing latency and increasing the rate at which data are read from or stored in memory. This is necessary to address a common computing problem known as the "memory wall," which refers to the observation that the rate of improvement in processor speeds has outpaced improvements in memory bandwidth.⁵ Even though this growing divergence was first observed over three decades ago, it has persisted until today, making memory the "primary bottleneck in AI applications."⁶ Over time, (generative) AI models have become progressively larger and more complex, driven by the link between model size and the ultimate performance of the model. In simple terms, a bigger model (both in terms of the number of model parameters and the size of the dataset used to train the model) tends to perform better.⁷ At the same time, this also translates to increasing demand for processing power and memory bandwidth, causing AI applications to encounter the memory wall.

As a result, all the devices that dominate the advanced AI accelerator landscape—including NVIDIA's Ampere, Hopper, and Blackwell series; AMD's Instinct series; and Google's Tensor Processing Units—now utilize HBM and the complementary CoWoS packaging technology. This, in turn, has made CoWoS capacity the biggest bottleneck for AI accelerator manufacturing. TSMC is the dominant

provider of CoWoS capacity (see [Data S1](#), sheet 1), responsible for packaging the chips of all the aforementioned device types. However, throughout 2023 and 2024, demand for TSMC's CoWoS capacity exceeded the company's supply. During TSMC's Q2 2023 earnings call, the company commented, "Especially for the CoWoS, we do have some very tight capacity—very hard to fulfill 100% of what customers needed," while adding, "We expect [this] tightness somewhat [to] be released in next year, probably toward the end of next year." During TSMC's Q3 and Q4 2024 earnings calls, the company subsequently confirmed the CoWoS capacity remained tight, stating, "Today's situation is our customers' demand far exceeds our ability to supply," and "We have very tight capacity and cannot even meet customers' need[s]." In the latter earnings call, TSMC also confirmed that CoWoS packaging was "highly concentrated with AI-related demand," stating, "Yes, today is all AI focused" (see [Data S1](#), sheet 2, for all relevant earnings call transcripts). An assessment of the limits of TSMC's CoWoS capacity therefore makes it possible to evaluate the maximum production output of advanced AI accelerators.

To make this assessment, it is crucial to first understand the CoWoS packaging process and chip manufacturing in general. Because chip production involves large-scale replication of a single design, it can be roughly compared to a more familiar printing process, such as business cards. Fabless companies (i.e., companies that outsource fabrication), such as NVIDIA, first create a chip design, which is then sent to foundries, such as TSMC. As a business card design would be printed on a large sheet of paper before individual cards are cut out, these chip designs are printed on round silicon wafers with a diameter of up to 300 mm, forming large-scale integrated circuits (LSIs). In a subsequent process, individual dice are cut and processed from these LSIs, which must be packaged afterward. With CoWoS packaging, the processor and memory dice are integrated into a single package. Even though there are multiple variants of CoWoS packaging technology, most advanced AI accelerators in recent years have used CoWoS with sil-

icon interposer (CoWoS-S). In this specific variant, processor and memory dice are vertically integrated on a single substrate using a monolithic silicon interposer. Like processor and memory dice, these interposers are printed on and cut from round silicon wafers. If the interposer dimensions used for AI accelerators are known, then it is possible to determine how many packages and devices using these packaged chips can be manufactured with a given CoWoS wafer capacity. In the context of business cards, this would be equivalent to trying to determine the number of business cards that can be printed for a certain capacity of paper sheets while knowing both the dimensions of the individual business cards and the paper sheets they are printed on.

AI accelerator production output

Analysts estimated that TSMC had a total CoWoS capacity of approximately 126,500 and 327,400 300 mm wafers in 2023 and 2024, respectively (see [Data S1](#), sheet 3). Notably, this capacity more than doubled from 2023 to 2024, an order of magnitude increase confirmed by TSMC itself. While TSMC did not disclose exact CoWoS packaging capacity figures, the company stated the following during its Q3 2024 earnings call "We work very hard and increase the capacity by about more than twice, more than two times as of this year compared with last year" (see [Data S1](#), sheet 2). Analysts also estimated that a majority of TSMC's CoWoS capacity was used by NVIDIA and AMD: the two companies together accounted for 52% and 58% of TSMC's CoWoS capacity in 2023 and 2024, respectively (see [Data S1](#), sheet 4). NVIDIA alone used an estimated 44% and 48% of TSMC's CoWoS capacity in 2023 and 2024, respectively, translating to approximately 55,283 and 158,059 packaging wafers during these years. AMD used 8% and 10% of TSMC's CoWoS capacity in 2023 and 2024, respectively, equating to 10,442 and 32,774 packaging wafers. During these years, the devices of these two companies almost exclusively utilized TSMC's CoWoS-S packaging technology (see [Data S1](#), sheet 6). While NVIDIA adopted CoWoS with local silicon interconnect (CoWoS-L) for its Blackwell generation, both NVIDIA and TSMC struggled to ramp up volume production of Blackwell

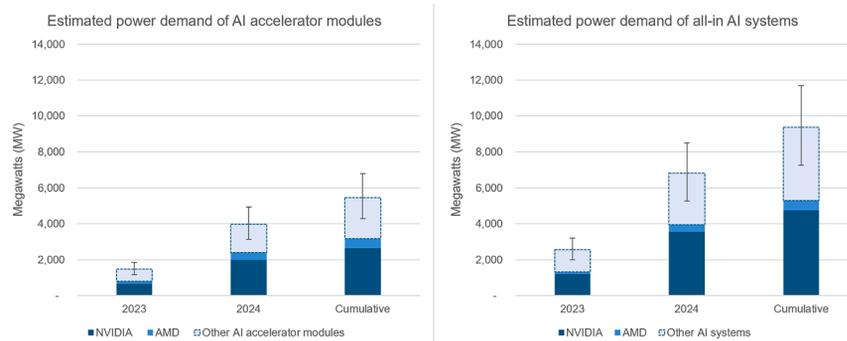


Figure 1. Estimated power demand of AI accelerator modules and AI systems manufactured in 2023 and 2024, along with their cumulative power demand by 2025

The power demand is estimated assuming a utilization rate of 65% and a PUE of 1.2, with error bars indicating the impact of varying PUE values between 1.1 and 1.3 and utilization rates varying between 55% and 75%.

devices in 2024.⁸ As a result, this CoWoS variant played a minor role in the total allocation of CoWoS capacity. It was estimated that NVIDIA's demand for CoWoS-L wafers in 2024 was at most 43,000 wafers in total (see [Data S1](#), sheet 6), representing 13% of TSMC's estimated total CoWoS capacity for that year. TSMC offers another CoWoS variant, CoWoS with silicon interposer and fan-out redistribution layer interposer (CoWoS-R),⁹ but this variant was not used by any of the common devices in the AI accelerator landscape and can therefore be excluded from the present analysis. Without these variants complicating the assessment, it is possible to determine how many of NVIDIA's and AMD's devices have been produced by first establishing how many individual devices could be manufactured from a single CoWoS wafer and then multiplying the estimated yield by the total allocated CoWoS capacity.

It can be assumed that a single CoWoS-S wafer can yield enough interposers to manufacture 28 Ampere or 29 Hopper devices (see [Data S1](#), sheet 5). AMD's devices use larger interposers, so a single CoWoS-S wafer can only yield 16 MI200 series devices and 12 MI300 series devices (see [Data S1](#), sheet 5). This calculation also assumes a perfect yield, but CoWoS-S yields have been estimated at over 99%.¹⁰ It can be assumed that the CoWoS capacity used by NVIDIA in 2023 was split equally between the Ampere and Hopper series (see [Data S1](#), sheet 6), leaving 27,642 packaging wafers for each generation. This means NVIDIA

could have produced $(27,642 \times 28 =)$ 773,976 Ampere and $(27,642 \times 29 =)$ 801,618 Hopper devices in 2023. In 2024, NVIDIA's Ampere series was discontinued, so it can be assumed the company used an estimated 115,059 packaging wafers for Hopper devices (158,059 minus the 43,000 CoWoS-L wafers for Blackwell devices). With this allocation NVIDIA could have produced $(115,059 \times 29 =)$ 3,336,771 units. Assuming AMD used its CoWoS capacity supply exclusively for MI200 series devices in 2023 and MI300 series devices in 2024, the company could have produced $(10,442 \times 16 =)$ 167,072 MI200 units in 2023 and $(32,774 \times 12 =)$ 393,288 MI300 units in 2024. The yield on the CoWoS-L wafers is more difficult to determine. CoWoS-L is a more complex packaging technology, and it is unclear what yield rates apply. Yield challenges were the primary reason NVIDIA struggled to produce Blackwell devices in 2024.⁸ Morgan Stanley analysts estimated that a single CoWoS-L wafer could result in 14 Blackwell devices (see [Data S1](#), sheet 5). However, this estimate suggests a similar yield to the Ampere and Hopper series, given that Blackwell devices require two processor dice instead of one. To remain conservative, and given the limited expected presence of Blackwell devices in NVIDIA's total device output, it can be assumed that a CoWoS-L wafer yields only seven Blackwell devices. With 43,000 CoWoS-L wafers, this would result in approximately 301,000 Blackwell devices.

Power demand

The power consumption profile of all these devices is very similar (see [Data S1](#), sheet 6). Within NVIDIA's Hopper series, the flagship H100 and H200 devices both have a thermal design power (TDP) of 700 W. This is also the starting point for NVIDIA's Blackwell devices, though these may have a TDP as high as 1,000 W. For AMD's MI300X, this figure is only slightly higher at 750 W. NVIDIA's older Ampere series, with the flagship A100 device, had a TDP of 400 W, while AMD's MI250X had a TDP of 500 W. However, these older devices likely represent a limited share of the total production output under consideration, as they were primarily relevant in 2023. Multiplying the TDP for each generation by the estimated production output of each device type reveals the total TDP of AI accelerator modules produced by NVIDIA and AMD in 2023 and 2024. The cumulative TDP of these devices is 3.8 GW (see [Data S1](#), sheet 6), meaning that without further production output in 2025, AI accelerator modules produced by NVIDIA and AMD alone could consume more electricity than a country such as Ireland in 2025. Moreover, several factors suggest that the total TDP of AI hardware will likely be significantly higher. First, the scope of this assessment is currently limited to NVIDIA and AMD devices, but these companies only used 57% of TSMC's combined total CoWoS capacity in 2023 and 2024. It is not possible to accurately assess the potential impact of the remaining 43%, as this capacity is primarily used by companies, such as Google's partner Broadcom, to manufacture Google's tensor processing units. These are custom, in-house solutions for which the product specifications have not been disclosed, making it impossible to determine their relevant dimensions or power usage. If power demand correlates with CoWoS capacity usage, the total TDP of AI accelerator modules may reach 6.7 GW rather than 3.8 GW. Another major factor is that the other additional components that will be used alongside these devices have not yet been considered. Typical AI systems, such as the DGX H100/H200 and DGX B200, have a TDP at least 79% higher than the TDP of the AI accelerator modules alone. Adding 79% on top of the previously estimated figures would

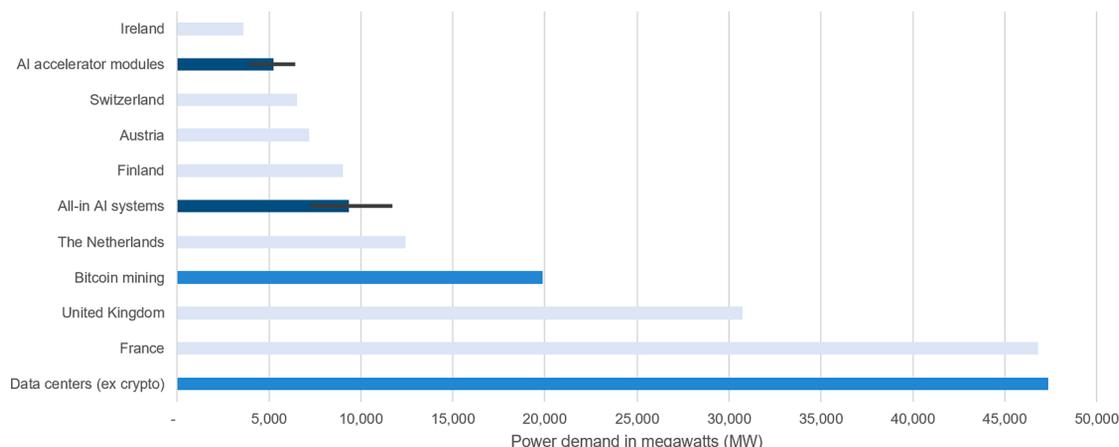


Figure 2. Scale of the estimated power demand of AI hardware

This figure illustrates the estimated power demand of AI hardware by 2025 compared to the power demand of Ireland (2023), Switzerland (2023), Austria (2023), Finland (2022), the Netherlands (2023), Bitcoin mining (March 2025), the United Kingdom (2023), France (2023), and total data center power demand (excluding cryptocurrency mining, 2024).

increase the total estimated TDP for NVIDIA and AMD AI systems to 6.8 GW, with the potential total TDP—including other AI systems—rising to 12.0 GW (see [Data S1](#), sheet 6).

The figures above do not account for utilization rates, which typically range from 60% to 70% for AI workloads¹¹ depending on factors such as whether the hardware is used for training or inference. Another important factor to consider is power usage effectiveness (PUE), which reflects the ratio of total data center facility power demand to IT equipment power demand. A significant portion of additional data center electricity consumption comes from cooling systems required to maintain optimal operating temperatures for IT equipment, though other facility requirements, such as lighting, are also included in the PUE value. The average global data center PUE is 1.56, but new regulations, such as the German Energy Efficiency Act, mandate that existing data centers achieve a PUE of 1.3 by 2030. Additionally, as of July 2026, new data centers will be required to have a PUE of at most 1.2 under the same regulation.¹² Assuming a utilization rate of 65% and a PUE of 1.2, the estimated power demand of AI accelerator modules produced by NVIDIA and AMD in 2023 and 2024 could reach 3.0 GW by 2025. Including other AI accelerator modules produced using TSMC’s CoWoS capacity, this figure could rise to 5.2 GW (see [Data S1](#), sheet 6). For AI systems, these

figures could increase further to 5.3 GW and 9.4 GW, respectively. [Figure 1](#) summarizes the potential power demand of AI accelerator modules and AI systems produced in 2023 and 2024, with error bars indicating the impact of varying PUE values between 1.1 and 1.3 and utilization rates varying between 55% and 75% (see [Data S1](#), sheet 7).

Over the full year of 2025, a power demand of 5.3–9.4 GW could result in 46–82 TWh of electricity consumption (again, without further production output in 2025). This is comparable to the annual electricity consumption of countries such as Switzerland, Austria, and Finland (see [Figure 2](#); [Data S1](#), sheet 6). As the International Energy Agency estimated that all data centers combined (excluding crypto mining) consumed 415 TWh of electricity in 2024, specialized AI hardware could already be representing 11%–20% of these figures. These outcomes are primarily sensitive to assumptions regarding utilization rates and PUE values ([Data S1](#), sheet 7), as illustrated in [Figure 3](#), which captures how variations in the different variables discussed in this article impact the final estimates.

Of course, power demand is set to continue expanding rapidly as the supply chain increases its production capacity while demand remains high. TSMC has already confirmed its target to double its CoWoS capacity again in 2025 (see [Data S1](#), sheet 2). This could mean the total power demand associated with devices pro-

duced using TSMC’s CoWoS capacity will also double from 2024 to 2025—just as it did from 2023 to 2024 ([Figure 1](#)), when TSMC similarly doubled its CoWoS capacity. At this rate, the cumulative power demand of AI accelerator modules produced in 2023, 2024, and 2025 could reach 12.8 GW by the end of 2025. For AI systems, this figure would rise to 23 GW, surpassing the electricity consumption of Bitcoin mining and approaching half of total data center electricity consumption (excluding crypto mining) in 2024. However, with the industry transitioning from CoWoS-S to CoWoS-L as the main packaging technology for AI accelerators, continued suboptimal yield rates for this new packaging technology may slow down both device production and the total power demand associated with these devices.¹³ Moreover, although demand for TSMC’s CoWoS capacity exceeded supply in both 2023 and 2024, it is not guaranteed that this trend will persist throughout 2025. Several factors could lead to a slowdown in AI hardware demand, such as waning enthusiasm for AI applications. Additionally, AI hardware may face new bottlenecks in the manufacturing and deployment process. While limited CoWoS capacity has constrained AI accelerator production and power demand over the past 2 years, export controls and sanctions driven by geopolitical tensions could introduce new disruptions in the AI hardware supply chain. Chinese companies have already

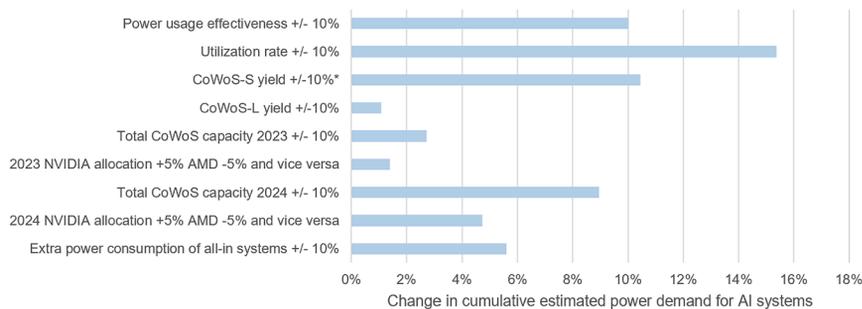


Figure 3. Sensitivity analysis for the estimated power demand of AI systems

This figure illustrates the percentage change in the estimated power demand of AI systems by 2025 by changing the assumptions used for making this estimate with a given percentage.

*The default assumption for CoWoS-S yield is 100%, so further increases are not possible.

faced restrictions on the type of AI hardware they can import, leading to the notable release of Chinese tech company DeepSeek’s R1 model. This large language model may achieve performance comparable to that of OpenAI’s ChatGPT, but it was claimed to do so using less advanced hardware and innovative software.¹⁴ These innovations can reduce the computational and energy costs of AI. At the same time, this does not necessarily change the “bigger is better” dynamic that has driven AI models to unprecedented sizes in recent years.⁷ Any positive effects on AI power demand as a result of efficiency gains may be negated by rebound effects, such as incentivizing greater use and the use of more computational resources to improve performance.¹⁵ Furthermore, multiple regions attempting to develop their own AI solutions may, paradoxically, increase overall AI hardware demand. Tech companies may also struggle to deploy AI hardware, given that Google already faced a “power capacity crisis” while attempting to expand data center capacity. For now, researchers will have to continue navigating limited data availability to determine what TSMC’s expanding CoWoS capacity means for the future power demand of AI.

Future research may also examine where AI hardware production output is ultimately deployed, as this is crucial for assessing the environmental impact of the electricity consumed by these devices. The characteristics of the relevant power grids will provide insights into the carbon and water intensity of the electricity generated to power AI hardware.

A significant portion of this hardware may end up in the United States, as OpenAI has partnered with several others in a joint venture called Stargate to invest up to \$500 billion over 4 years in new data center infrastructure across the country. There are early indications that these data centers could exacerbate dependence on fossil fuels: oil and gas company Crusoe has reportedly secured 4.5 GW of natural gas power capacity for AI data centers, with Stargate as one of its potential customers.¹⁶ However, while a growing reliance on fossil fuels threatens to undermine climate goals, effective policy responses first require urgent transparency.

DECLARATION OF INTERESTS

The author declares no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.joule.2025.101961>.

REFERENCES

- Kimball, S. (2024). Google says U.S. is facing a power capacity crisis in AI race against China. CNBC. February 12, 2025. <https://www.cnbc.com/2025/02/12/google-says-us-faces-power-capacity-crisis-in-ai-race-against-china.html>.
- Patterson, D., Gonzalez, J., Holzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D.R., Texier, M., and Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer* 55, 18–28. <https://doi.org/10.1109/MC.2022.3148714>.
- Masanet, E., Lei, N., and Koomey, J. (2024). To better understand AI’s growing energy use, analysts need a data revolution. *Joule* 8,

2427–2436. <https://doi.org/10.1016/j.joule.2024.07.018>.

- European Parliament (2024). Parliamentary question - P-001974/2024(ASW). https://www.europarl.europa.eu/doceo/document/P-10-2024-001974-ASW_EN.html.
- Wulf, W.A., and McKee, S.A. (1995). Hitting the memory wall: implications of the obvious. *SI-GARCH Comput. Archit. News* 23, 20–24. <https://doi.org/10.1145/216585.216588>.
- Gholami, A., Yao, Z., Kim, S., Hooper, C., Mahoney, M.W., and Keutzer, K. (2024). AI and Memory Wall. *IEEE Micro* 44, 33–39. <https://doi.org/10.1109/MM.2024.3373763>.
- Ananthaswamy, A. (2023). In AI, is bigger always better? *Nature* 615, 202–205. <https://doi.org/10.1038/d41586-023-00641-w>.
- Hsiao, J. (2024). Nvidia takes full Blackwell delay accountability, seeks to dispel tension with TSMC rumors. *Digitimes*. October 24, 2024. <https://www.digitimes.com/news/a20241024VL201/nvidia-tsmc-blackwell-supplier-production.html>.
- TSMC (2024). CoWoS. <https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/cowos.htm>.
- Karaahmetovic, V. (2024). NVIDIA 2025 GPU unit forecast raised at Mizuho. *Investing.com*. September 30, 2024. <https://www.investing.com/news/stock-market-news/nvidia-2025-gpu-unit-forecast-raised-at-mizuho-3640738>.
- TrendForce (2024). Datacenter GPUs May Have an Astonishingly Short Lifespan of Only 1 to 3 Years. <https://www.trendforce.com/news/2024/10/31/news-datacenter-gpus-may-have-an-astonishingly-short-lifespan-of-only-1-to-3-years/>.
- ICIS (2024). Data centres: Hungry for power. <https://www.icis.com/explore/resources/data-centres-hungry-for-power/>.
- Zuhair, M. (2025). NVIDIA’s CEO Jensen Huang Addresses CoWoS Order Cut Rumors; Claims Lower Figures Are Simply Due To The Switch Towards CoWoS-L. *WCCFTech*. March 3, 2025. <https://wccfttech.com/nvidia-ceo-jensen-huang-addresses-cowos-order-cut-rumors/>.
- Mills, S., and Whittle, R. (2025). DeepSeek: what you need to know about the Chinese firm disrupting the AI landscape. *The Conversation*. January 31, 2025. <https://theconversation.com/deepseek-what-you-need-to-know-about-the-chinese-firm-disrupting-the-ai-landscape-248621>.
- IEA (2025). Energy and AI. <https://iea.blob.core.windows.net/assets/dd7c2387-2f60-4b60-8c5f-6563b6aa1e4c/EnergyandAI.pdf>.
- Gooding, M. (2025). Crusoe secures 4.5GW of natural gas power for AI data centers. *Data Center Dynamics*. March 17, 2025. <https://www.datacenterdynamics.com/en/news/crusoe-secures-45gw-of-natural-gas-for-ai-data-centers-report/>.