

INTELLIGENCE UPDATE

Gen AI power consumption surges higher faster



Daniel Bizo 28 May 2025

Uptime Intelligence published its first estimate on the global power needs of generative AI in March 2024 (see [Generative AI and global power consumption: high, but not that high](#)), and the infrastructure build-out to serve it was already unprecedented in scale and speed.

However, what has happened in the months since has exceeded all expectations, including Uptime Intelligence's GPU shipment projections. Nvidia's revenues exemplify this: the fabless silicon and systems vendor saw its data center business swell to more than \$115 billion in its financial year ending January 26, 2025. To put that into perspective, Intel's best year in data centers (2020) brought in \$26 billion, when it had a hegemonic position in server processors.

The past year warrants a revision to our generative AI power estimates. Yet, stratospheric spending levels have not brought greater clarity about the immediate future of generative AI — if anything, visibility has worsened as the stakes have risen. In discussions with executives at major suppliers, it is clear that no one really knows how much longer this investment exuberance can defy financial gravity in the absence of meaningful revenues from generative AI services for consumers and businesses.

Compounding this is a lack of clarity about what exactly is happening behind the headline numbers. For example, how much money and power are expended on training as opposed to inference; where and how service providers and end-user organizations are deploying generative AI-powered applications; and what all this means for infrastructure siting and design. Then, there is the likely growing spread of alternative AI systems not using Nvidia GPUs — such as those from AMD, Huawei and hyperscale cloud vendors — making it even more difficult to track market movements. The data center industry is deep in uncharted territory.

On a more positive note, the industry is developing a better understanding of some aspects of AI models and the specialized hardware they require. It now recognizes that AI training clusters exhibit highly volatile load profiles, which are inherent to the architecture of massively parallel training workloads (see [Erratic power profiles of AI clusters: the root causes](#)).

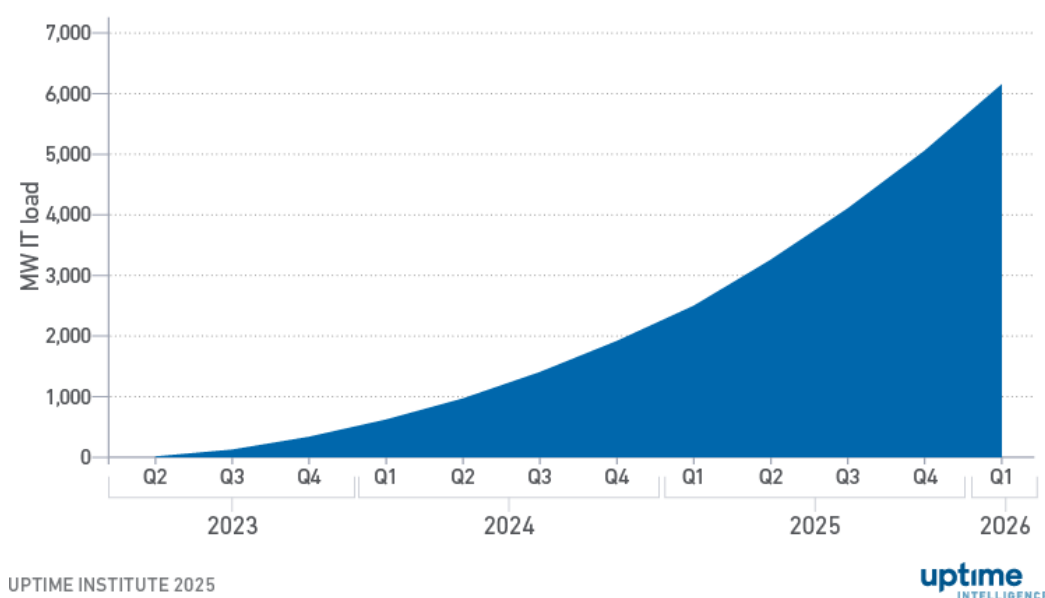
It is also clear that while GPUs and their server clusters are highly utilized, average utilization is nowhere near 100%, even for training — it is probably closer to 70% we think. Given the spare capacity at many cloud operators, this figure is likely to be generous still. Another factor to consider is that despite their computationally intense nature, AI workloads do not typically sustain server power close to their nameplate due to a variety of architectural and operational reasons. There are several factors playing into the lower estimated hardware power use, which a previous Uptime Intelligence report discusses (see [GPU utilization is a confusing metric](#)).

Consequently, the average power consumption figures for AI hardware, used in some power consumption models, have been exaggerated.

Offsetting some of the reductions is the observation that training is likely to dominate over inference as a share of all AI workloads for longer than previously thought. Infrastructure investments remain heavily skewed toward training more, better and larger generative models, even as inference energy escalates steeply with "reasoning" and "deep research" models. Precisely where this split lies, and how much infrastructure is deployed for inference, is perhaps the single greatest source of uncertainty in projecting future energy use associated with generative AI models.

Building on estimates of Nvidia's global GPU shipments, hardware specifications and operational assumptions (e.g., average system utilization and AI workload intensity) — Uptime Intelligence's updated model suggests that by the end of the first quarter of 2025, AI accounted for an additional 2.5 GW of (incremental) IT load globally (see **Figure 1**). This marks an increase from our previous projection of nearly 2 GW. This estimate reflects power consumption and should not be confused with design capacity.

Figure 1 Estimated power consumed by generative AI systems globally



The updated model projects a doubling of power consumption by the end of 2025, reaching 5 GW of IT load serving generative AI workloads. The corresponding annualized total data center power consumption is around 38 TWh, based on an average PUE of 1.3. This is better than the global average PUE of around 1.6, but assumes AI systems are predominantly housed in recent or brand-new facilities optimized for heat rejection efficiency.

By the start of 2026, the model estimates that generative AI training and inference systems will add about 1.1 GW of IT load per quarter globally — with an even higher facility capacity requirement to account for peak loads. According to data compiled by Goldman Sachs, the world typically adds between 2 GW and 3 GW of data center capacity each quarter. If these numbers are accurate, they indicate that AI systems will soon represent around half — or even more — of the demand for new data center supply.

Once again, the horizon of our projection is limited to about a year out due to high divergence between scenarios built on different assumptions about generative AI infrastructure. This

uncertainty is also reflected in the published estimates and forecasts of data center power demand. For example, a 2024 study by the Lawrence Berkeley National Laboratory on US data center power use shows an upper bound estimate for 2025 that is about 50% higher than the lower bound — a gap that widens to nearly 80% by 2028.

The researchers also note that lack of information disclosed by operators on data center infrastructure performance characteristics, including both IT and facilities, hinders the accuracy of these models. This has long been the case with data centers, but the introduction of generative AI makes accurately modeling data center power demand all the more challenging. Cautiousness of equipment vendors about excess manufacturing capacity means that while lead times have gradually improved, supply will continue to be tight for now.

ABOUT THE AUTHOR



Daniel Bizo

25 Aug 2026

Over the past 15 years, Daniel has covered the business and technology of enterprise IT and infrastructure in various roles, including industry analyst and advisor. His research includes sustainability, operations, and energy efficiency within the data center, on topics like emerging battery technologies, thermal operation guidelines, and processor chip technology.

dbizo@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. For over 30 years, the company has established industry-leading benchmarks for data center performance, resilience, sustainability, and efficiency. These benchmarks assure customers that their digital infrastructure will perform consistently with their business needs under any operating conditions. With more than 4,300 awards issued globally and over 1,100 active projects in 120+ countries, Uptime has helped tens of thousands of companies optimize critical IT assets while managing costs, resources, and efficiency.

The organization's Tier Standard is the IT industry's most trusted and adopted global standard for data center design, construction, and operations. Core offerings include Tier Certifications, Management & Operations reviews, AI Infrastructure Advisory, Data Center Risk Assessments, Energy & Sustainability Assessments, Cybersecurity Assessments & Ratings, and IT Advisory Services.

Uptime Education programs have been completed by more than 100,000 data center professionals, establishing Uptime as the Learning & Development partner of choice worldwide. Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, Singapore, and Madrid, and full-time Uptime professionals based in over thirty-four countries around the world.

For more information, visit www.uptimeinstitute.com