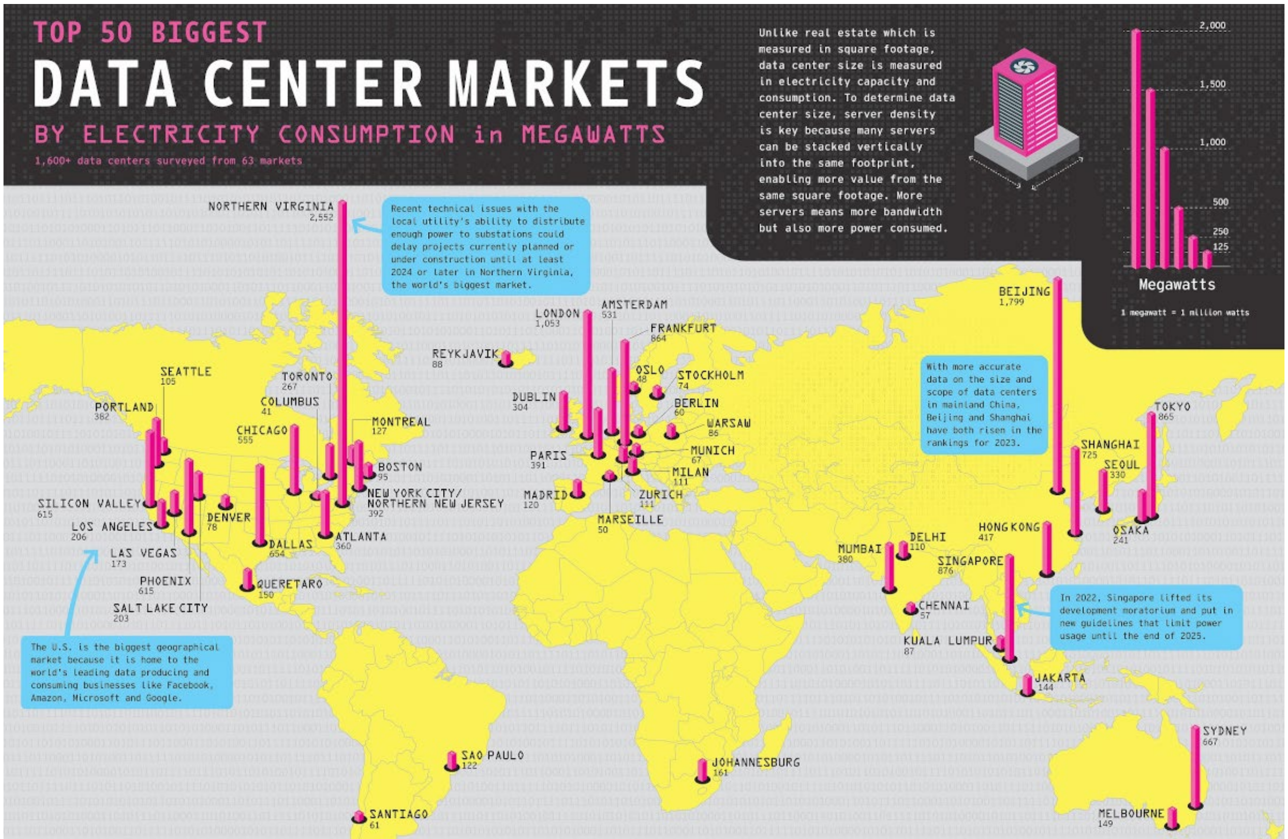
HITACHI

Technical brief

Powering data centers sustainably in an AI world

Discover how artificial intelligence can enhance power infrastructure

Data centers are getting bigger, denser, and more power-hungry



Our online lives depend on data centers, which play a central role in delivering the streaming content people crave, the apps we depend on, and the collaboration tools that increasingly drive business and commerce of every kind – it's all stored, managed and processed in data centers. With the advent of AI, the need for data processing has skyrocketed, and as a result data centers are getting bigger, denser, and more power-hungry than ever.

This has profound implications for power grids. Why? 'Hyperscale' data centers operated by big cloud providers like Amazon, Google, Microsoft and Facebook now consume hundreds of megawatts (MW) of power, and multi-gigawatt (GW) campuses are already being planned in the US. Projects are being developed across the country, often defined by the availability of abundant, affordable and clean electricity. The demand is less acute elsewhere in North America, at least for now, but Canada and Mexico are growing in popularity as data center locations. In Canada, access to plentiful hydroelectric power helps data center operators meet sustainability commitments, and Mexico is a popular nearshoring location with a rich talent pool. A few years of steady data center growth in these nations could severely strain the grid.

A surge in demand for artificial intelligence (AI), increased machine learning, and high-performance computing (HPC) workloads is accelerating these trends across North America. Omdia notes that AI applications accounted for only 10 percent of data center power capacity in 2020. Now it is already at nearly 30 percent and is projected to rise to 45 percent by 2030.[1] This is already showing up in rack density, which has risen from 7 kW per rack in 2021 to 12 kW today, according to AFCOM's State of the Data Center 2024 report.[2] Further, 25 percent of data centers report that they have racks in excess of 20 kW, with some reporting racks of more than 50 kW. As a result, the industry already accounts for almost 2 percent of total electricity usage in the US.[3] Some analysts believe it could climb as high as 7.5 percent by the end of the decade as more graphics processing units (GPUs) are added to serve the needs of HPC, generative AI (GenAI), and other demanding applications.

AI expansion and data center demand

Standing in sharp contrast: data center owners and operators are aggressively embracing sustainability initiatives. They seek to drastically lower their carbon footprints by adopting renewable energy and achieving higher levels of power efficiency. [4]

These competing vectors cannot be aligned unless underpinned by a modern, highly efficient power infrastructure, built around increasing volumes of sustainable energy, that reaches from the first point of utility connection to the server rack – one that harnesses the potential of AI to achieve the highest levels of energy efficiency.

AI crunches more numbers and sifts through larger blocks of data at a phenomenal pace. Serving an ever-larger user base with AI requires a great leap forward in processing power.

According to a Pure Storage report:

88% of those who adopted AI saw computing power rise dramatically.

74% confirmed that AI required or will require significant upgrades or a complete overhaul of their IT infrastructure.

10x the demand for generative AI is set to drive increase in data center electricity consumption by 2030, according to figures from Bloomberg Intelligence (BI).[6]

> "We are in the wild, wild west right now. There are no rules."
>
> **Joti Balani**
> AI expert and CEO of AI consultancy, FreshRiver.ai

The maximum power consumption of an Nvidia Graphics Processing Unit (GPU) has gone from 20 Watts to 120 W over the past decade.[7] Many facilities are switching from central processing units (CPUs) to GPUs to satisfy AI demand. Thus, rack density is destined to keep climbing steadily upward – and with it, the amount of power that needs to be fed into the data center, with some racks in use above 100 KW and designs up to 300 KW.

Generative AI (Gen AI) is exacerbating the situation. ChatGPT took only five days to reach a million users. It boasted almost 200 million users globally at its peak and now counts a steady 600 million visitors a month[8] (including 80 percent of the Fortune 500 and 43 percent of college students).[9] A single ChatGPT query can generate 100 times more carbon than a regular Google search. With 10 million queries per day,[10] estimates for daily power consumption amount to around 1 GWh each day, or enough to power 33,000 US households. Now factor in the training required for large language models (LLMs). It can require up to 10 gigawatt-hours (GWhs) to train a single LLM like ChatGPT-3[6], and there are scores of other gen AI applications out there and many Fortune 500 companies are building their own LLMs.[11]

Omdia estimates that data centers will need about 100 GW of additional power capacity to meet AI demand between 2024 and 2030.[12] That quantity of power will be difficult to obtain due to lack of generation, inadequate transmission capacity, lead times for electrical equipment and component supply chains that can stretch as long as 36 months, and an array of regulatory hurdles. Data centers will need a reliable partner to help them overcome these challenges.

Joti Balani, founder and managing director of Freshriver.ai, cautions that Pandora's box has just opened. Despite all the technological advances throughout humanity, we have never seen anything like AI. "This is not like the Web. This is not like mobile. It's not like cloud; it's a phenomenon permeating every organization relative to function, policies, and regulation."

She says that AI is causing two severe bottlenecks:

- A bottleneck inside the data center due to the sheer volume of storage capacity and computing power needed to cope with AI workloads.

- A bottleneck external to the data center due to the struggle to provide enough power supply for facility needs.

With almost half a trillion dollars in investment into AI startups over the past ten years and $11.3 billion more invested in AI in the US alone in the quarter following this study, the situation will only worsen.[13]

However, there are further factors to consider: the amount of power demanded by data centers used to be predictable and relatively static for the utilities supplying them. AI is disrupting the energy dynamics of the data center. Typical legacy server power supplies exhibit little usage change from idle to full load (no more than a 10 percent variation). As AI becomes more prevalent, data centers can expect fluctuating loads that ramp up and down regularly.

The unpredictability of AI demand is paralleled by the variability of supply due to the greater use of renewable resources, such as wind and solar. Consequently, there is a phenomenon known as the duck curve,[14] whereby solar resources drop off rapidly at points of highest demand. This can result in severe grid instability if not appropriately managed. Traditionally, fossil-fuel power plants could be brought online rapidly to fulfill unmet demand, but many of those sources are being retired. Those supplying power from the grid to the data center must manage these peaks and troughs effectively to guarantee supply reliability. Ensuring a steady flow of energy becomes a critical concern inside and outside the data center.

Now factor in the policy arena. There has been a backlash around data center power usage in certain areas due to rapidly growing demands placed on power grids. In some cities and regions, moratoriums have even been issued on building new data centers. Further, planned data centers often face a battle to find a location where they will have enough power. Many utilities are unable to provide the power requested by data centers. In some highly constrained grids, none may be on the horizon for the foreseeable future. Recently, the main siting priorities have been access to connectivity, location, water, land and power pricing, and availability. We're seeing power availability as the number one priority for data center operators.[15] Thus, several vectors are colliding: the need to increase the power supply to support AI versus a growing desire for sustainability and carbon neutrality.

Power modernization and energy efficiency

The solution lies in both a modernization of the power infrastructure and a redoubling of efforts to boost energy efficiency. This must be done inside and outside of the data center if AI is to be accommodated in alignment with sustainability goals.

Internally, advanced power settings must be deployed in server power supplies to match supply and demand closely. Advanced power management systems must become commonplace inside data centers. The good news is that progress has been made due to major advances in power usage effectiveness (PUE). PUE deals with how much of the power used in a facility is utilized for computing versus auxiliary cooling, lighting, etc. A PUE of 1 is the ideal. The Uptime Institute[16] gave an average for PUE worldwide of 1.56 in 2024. That's a considerable gain from 2007 when PUE scored 2.5. However, PUE averages seem to have stalled between 1.55 and 1.6 for several years despite regulatory pressure to bring it down further. It remains to be seen what the impact of AI will be on PUE.[17]

Facility owners and managers must extend their horizons beyond the four walls of the data center to rise to the challenge of supporting AI loads sustainably. They must turn their attention to the external power infrastructure feeding the data center. Utility interconnects need to be able to deal with big leaps up and down of intermittent power. Substations and other links to the data center must be able to orchestrate the flow of power derived from various energy resources and effectively deal with transients, reactive power shifts, and other electrical phenomena that negatively impact power quality. These needs are becoming even more acute due to an exponential increase in AI workloads.

What is different about AI data centers?

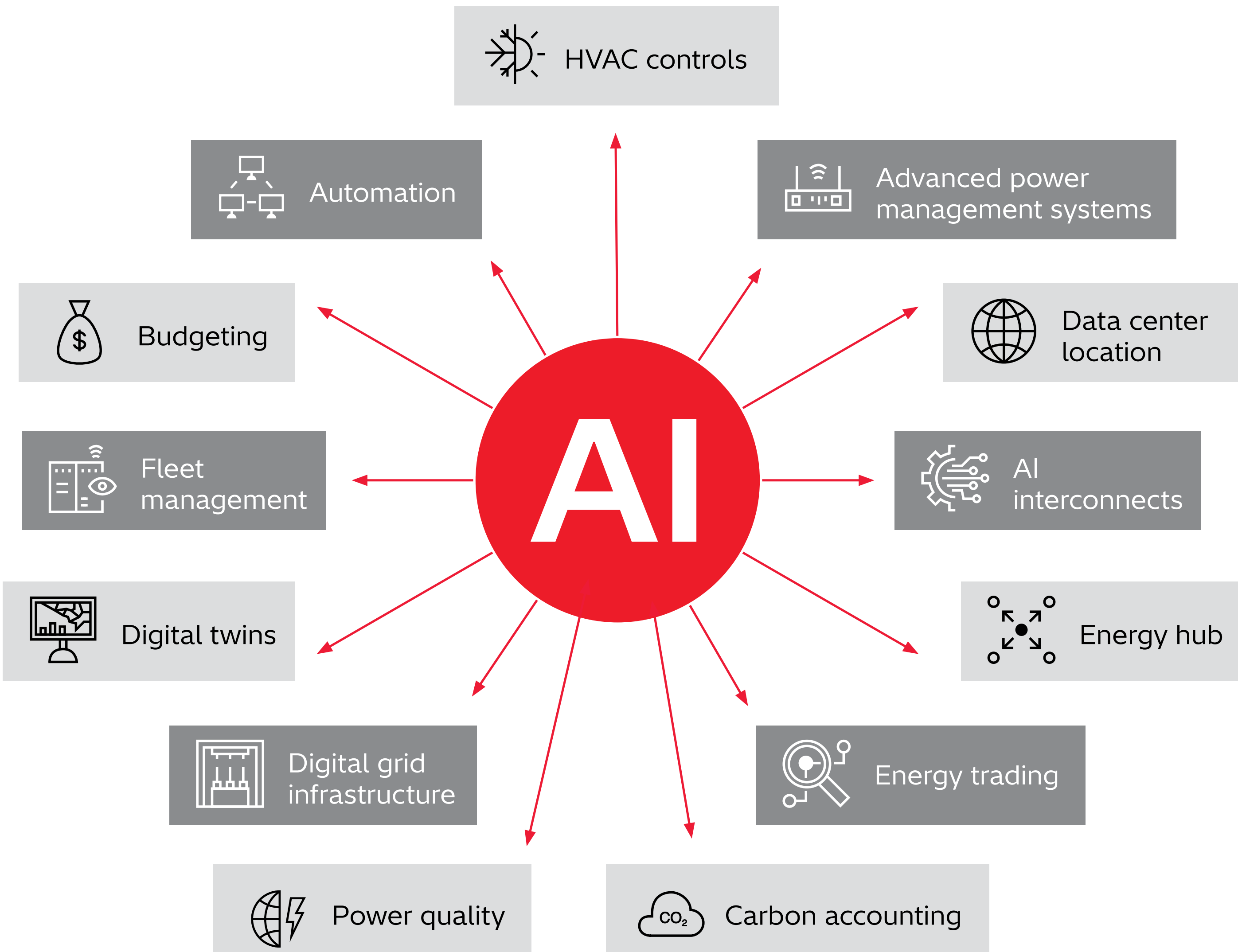
What sometimes isn't grasped is that the AI data center will be very different from typical data center operations. Traditionally, processing experiences a few spikes but has plenty of periods of relative inactivity. The data center is designed to cope with the peaks when they happen. Further, you might have a few servers working overtime while the rest of the infrastructure deals with light volumes of traffic.

The AI data center is an entirely different beast. Due to the heavy demands of LLMs and the amount of traffic generated by AI applications, the processors often work at a high load. The spikes are far higher than the typical data center, and the number of spikes is far greater. Thus, the entire data center can be operating flat out on a particular problem, and then suddenly the same workload can disappear. Such a spiky load profile can wreak havoc on the grid due to the large amount of power involved, the duration of the peaks, and their sudden cessation. Utilities can no longer view data centers as areas of predictable load. Instead, AI data centers are becoming potential areas of grid instability. Accordingly, the evolving landscape of AI data centers demands the deployment of grid stabilization and power quality equipment and systems, inside the data center, to ensure they remain good corporate citizens by minimizing disruptions on the local grid.

AI can enhance power infrastructure

AI poses a massive problem for data centers as its energy usage requirements misalign with ongoing sustainability efforts. Yet AI offers an opportunity to resolve apparent conflicts.

Here are just a few areas where AI can enhance data center power management and energy efficiency, both internally and externally.



HVAC controls

Hyperscalers have been experimenting for several years with the use of AI to improve the monitoring of data center cooling. Google, for example, reported cooling energy savings of 40 percent using AI. This was achieved without changing the computer floor's topology or physical layout.[18]

Advanced power management systems

By adding sensors throughout the data center, AI engines can interact with advanced power management systems to tightly control the flow of power, boost efficiency, and reduce consumption despite higher rack density. That necessitates digitalization across all facilities and IT systems to enable measurement accuracy and timely response to fast-evolving situations and emergencies. Fortunately, many data centers are in the midst of digitalization initiatives, including installing new sensors and upgrading from analog to digital sensors and digitized controls. Once completed, AI can be used to better regulate the provision of power, determine more precise setpoints, and optimize the flow of energy across the data center.

Data center location

Power management tools can harness AI to help data center planners pick the best location for new facilities. While proximity to internet backbones used to be a primary concern, the need for more power to address AI workloads has risen in priority. Modern tools can highlight ideal locations for new facilities where cheap, carbon-free power is available in abundance and where there are excellent fiber backbones in the vicinity of load centers.

AI interconnects

Given both the high-power usage requirements and the impact of new, more variable loads for data centers, a more collaborative approach is needed between the data center and the energy provider: a consumer/supplier relationship is no longer sufficient. Adding AI-based digital capabilities to high-voltage transmission and distribution systems could streamline the process of aligning data center demand with available renewable energy sources. Using this approach, data center operators could more easily access the clean power they need, and developers and utilities could find a market for unused supply, avoiding curtailments.

AI could potentially be introduced into interconnections between data centers and substations, whether owned by utilities or data center operators themselves. While this concept is new and unproven, AI systems can potentially bring greater alignment by feeding the required data to the utility and the data center to ensure the most efficient power usage.

Energy hub

Some data centers are transitioning to the establishment of their own energy hubs. These hubs can integrate different power sources seamlessly while offering grid protection and automation services. Emerging energy hubs armed with the latest grid technologies provide a future-proof solution as they can accept any type of power input, whether from solar, wind, natural gas, nuclear, hydroelectric, batteries, diesel generators, coal plants, biomass, or hydrogen fuel cells. Energy hubs can be configured to operate off-grid, in what's known as 'island mode,' or can be fully connected to the grid, seamlessly switching between these modes as needed. AI systems can help manage the transition from one energy source to another in real-time, avoiding surges and other damaging electrical repercussions. Energy Hubs that leverage AI have the potential to more effectively manage distributed and variable energy resources.

Energy trading

Utilities and energy providers increasingly regard grid stability as being of paramount importance. In some cases, they are beginning to require data centers to transition from being an energy consumer to a "prosumer," i.e., actively involved in the health of the grid by operating as a virtual power plant and user. In this new relationship, both parties can improve energy economics by buying and selling power to each other based on market conditions. If a data center has renewable generation, CHP, or other on-site assets, it may be able to sell excess energy to the grid at certain times while being able to obtain attractive rates for power from the grid due to the establishment of a mutually beneficial partnership with the utility. However, such arrangements demand real-time and same-day energy trading systems to cope with the incorporation of renewable energy, energy storage, and far-flung generation into data centers, as well as the exchange of excess energy resources from the data center to the grid. AI, then, will likely play a massive role in energy arbitrage. For example, a global enterprise with many data centers could use AI to ‘follow the sun’ to optimize the usage of renewable energy by shifting loads to locations that are powered 100 percent from renewable energy around the world all the time.

Carbon accounting

Data centers are increasingly subject to rules and regulations that require them to meet specific emissions, power usage, sustainability, and net-zero targets. Current tools only estimate these numbers. AI can add greater discernment, refinement, and accuracy to the conclusions reached by the latest power management systems as load shifts from one power source to another.

Power quality

Power quality is vitally needed to manage transients on the modern grid. Peaks and troughs in power frequency can be severely disruptive. Power management tools have been deployed on the grid for many years to minimize these variations. However, they are now needed inside the data center. Due to the variability of AI workloads, demand can suddenly surge and recede. With AI drawing so much power, swift and significant changes in data center power load can disrupt the grid. The latest AI-based tools can help data centers manage variable internal loads generated by AI applications.

Digital grid infrastructure

The latest batch of advanced transformers and other high-voltage equipment entering the market incorporate digital controls and asset management tools that enable them to utilize AI. This is a vital step in attaining sustainable and reliable data centers. After all, the failure of a 50 MVA transformer can take out an entire facility. Smarter transformers are now available that can smoothly switch loads from one power source to another without causing transients that can damage equipment. In addition, they can analyze dissolved gases in the liquid inside a transformer to determine which maintenance actions are needed, which components should be replaced, and at what point the equipment has reached its end-of-life. This moves transformer maintenance from a break-fix model to a predictive model.

Digital twins

By digitizing everything from the grid interconnect to the internal data center, AI can generate a digital twin that mirrors the entire power infrastructure; in fact, digital twins of substations are available today. Designers and managers can use these twins to run simulations of proposed changes to determine the best way forward in terms of cost, carbon footprint, and energy efficiency. Retrofit options can be run against what-if scenarios to find the best option. Designers can simulate many different designs to select the most ideal configuration.

Fleet management

Those managing multiple data centers can use AI to monitor transformers and other power infrastructure to customize operations for optimum usage. Lessons learned with one transformer or substation can be applied across the fleet. This leads to longer life and more efficiency.

Budgeting

It is one thing to advocate advanced AI-based tools and another to afford them. But AI can help by highlighting ways to improve heating, ventilation, and air conditioning (HVAC) loads, streamline power usage, enhance overall power visibility, and recommend easy and inexpensive upgrades. That way, a data center can transition gradually toward an AI-enabled future.

Automation

Data centers are continually asked to do more with less. The trend of fewer and fewer people in the data center will likely continue. Data center managers, therefore, need tools that help them automate key functions to get more done in less time. Automating the entire power infrastructure with AI can help the data center manager cope with ongoing demands.

Conclusion

The age of AI is here, and those operating data centers must prepare their power infrastructure for a future that relies on AI to enable the digital interactions that power our society, and can move us toward a sustainable energy future for all. This will require us to:

- Modernize internal data center power management systems.
- Digitize and modernize substations and support them with asset monitoring and management systems that can provide more accurate predictive maintenance and greater insights into process efficiency.
- Harness AI to optimize power, cooling, computing, and storage requirements within the data center.
- Adopt the latest digital and AI-enabled systems, grid infrastructure interconnects, and energy trading platforms to bring power reliably and affordably from the grid (or from renewable assets) to the data center.

That said, there is plenty of room for caution. The hype unleashed by AI vendors about the potential of this new technology is almost unprecedented. Data center owners and operators are advised to focus on distinct use cases that will provide material benefit in terms of revenue or time saved – and that can do so affordably.

Hitachi Energy provides advanced, AI-based power management solutions based on real-world experience with the top players in the data center space. The company has built and modernized the power grid for over 100 years. Spend time with our experts; we can help you work through the challenges inherent in aligning AI, power management, and data center sustainability.

References

(1) Data Center Building Report – 2023 Omdia Data Center Building Report – 2024 Omdia (informa.com)

(2) AFCOM's State of the Data Center report 2024 AFCOM - Advancing IT and Data Center Infrastructure Professionals

(3) Data Centers and Servers | Department of Energy

(4) Climate Neutral Data Centre Pact – The Green Deal need Green Infrastructure

(5) Drivers of Change: Meeting the Energy and Data Challenges of AI Adoption https://www.purestorage.com/docs.html?item=/type/pdf/subtype/doc/path/content/dam/pdf/en/analyst-reports/ar-meeting-energy-challenges-of-ai-adoption.pdf

(6) Solar, gas, and batteries race to power AI revolution | Insights | Bloomberg Professional Services

(7) TechSpot https://www.techspot.com/articles-info/2540/images/2022-10-06-image.png

(8) Number of ChatGPT Users (Oct 2024) (explodingtopics.com)

(9) 107 Up-to-Date ChatGPT Statistics & User Numbers [April 2024] (nerdynav.com)

(10) New ChatGPT + AI Statistics - (Updated August 2024) (mailbutler.io)

(11) Q&A: UW researcher discusses just how much energy ChatGPT uses | UW News washington.edu

(12) Data Center Building Report – 2023 Omdia Data Center Building Report – 2024 Omdia (informa.com)

(13) New ChatGPT + AI Statistics - (Updated August 2024) (mailbutler.io)

(14) Duck curve - EIA https://www.eia.gov/todayinenergy/detail.php?id=56880

(15) Power Is Key to Unlocking AI Data Center Growth (datacenterknowledge.com)

(16) Uptime Institute Global Data Center Survey https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2023

(17) Building a sustainable data centre: net zero carbon | RPS (rpsgroup.com)

(18) Google just gave control over data center cooling to an AI | MIT Technology Review https://www.technologyreview.com/2018/08/17/140987/google-just-gave-control-over-data-center-cooling-to-an-ai/

Hitachi Energy USA Inc.
901 Main Campus Drive
Raleigh, NC 27606, USA

Contact us:
US & Canada: +1 800-290-5290
Mexico: +52 800-681-6535
hitachienergy.com/contact-us

Learn more:
hitachienergy.com/markets/data-centers

